IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Depersonalized Cross-Subject Vigilance Estimation with Adversarial Domain Generalization

Bo-Qun Ma[1], He Li[1], Yun Luo[1], Bao-Liang Lu[1,2,3,*]
[1]Center for Brain-like Computing and Machine Intelligence
Department of Computer Science and Engineering
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
[3]Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China

***Abstract*—Subject variability is a major obstacle to vigilance estimation. The conventional subject-specific models fail to perform well on unknown subjects. The existing studies mainly focus on domain adaptation utilizing labeled/unlabeled subject-specific data. However, it is still expensive and inconvenient to collect task-specific data from unknown subjects in some real-world applications. In this paper, we introduce domain generalization methods for building vigilance estimation models without requiring any information from the unknown subjects. We first generalize the structure of Domain Adversarial Neural Network (DANN) into Domain Generalization (DG-DANN), and then propose a novel adversarial structure called Domain Residual Network (DResNet). We compare a popular domain generalization method, Domain-Invariant Component Analysis (DICA), with our proposed approach. In terms of the estimation accuracy and generalization ability, we designed two different settings for evaluation experiments on a public dataset called SEED-VIG. Experimental results indicate that our new model achieves comparable accuracy but more stable performance without using additional information from the unknown subjects in comparison with the state-of-the-art domain adaptation methods. Furthermore, domain generalization models also perform well on the tasks with multiple unknown subjects.**

***Index Terms*—domain generalization, adversarial network, electroencephalography (EEG), electrooculography (EOG), vigilance estimation**

I. Introduction

With the rapid increase of vehicles, driving safety has become a crucial issue. Fatigue driving is reported to be one of the most prominent causes of traffic accidents [1]. People usually fail to drive safely due to lack of vigilance. Therefore, it is of great importance to estimate drivers' vigilance in real-time.

In the past decades, various kinds of signal modalities have been exploited to estimate drivers' vigilance [2]. As a signal which directly reflects brain activity, electroencephalography (EEG) has been demonstrated to be a reliable and promising indicator of human mental state [3] [4]. Recently, a close relationship and complementary characteristics between EEG and eye movement data have been found in emotion recognition [5]. Thus, multimodal approaches have been employed to further improve the performance of the vigilance estimation models [6] [7].

*Corresponding author: Bao-Liang Lu (bllu@sjtu.edu.cn)

In Brain-Computer Interface (BCI) systems, subject variability is one of the major problems for real-world applications [8]. Due to the subject variability, conventional models trained on the data from one subject suffer from compromised performance when applied to estimate the vigilance level of another subject. There are the following several causes for subject variability in EEG and EOG data: a) individual differences in human brain functional and anatomical connection [9]; b) misregistration during data collection resulting from different skull shapes across subjects; c) changes of environment and sensors' state in different experiment sessions and days; and d) variety of subjects' mental state, emotional condition and task-irrelevant brain activity disturbance. Collecting labeled EEG data is extremely expensive and time-consuming. Therefore, it is necessary to explore an efficient approach to reduce the subject dependency of the BCI systems. In previous studies, efforts have been made to address this problem. The methods for tackling subject variability can be classified into two categories, the subject-variant approach and the subject-invariant approach. Specifically, subject-variant approaches calibrate the pre-trained models with additional data from the test subject, while subject-invariant approaches investigate robust models which can perform well on other subjects.

Currently, transfer learning has attracted the attention of many researchers [10] [11], which can transfer the knowledge learned from the existing data to new application circumstances. Thus, transfer learning has a high potential in developing the generalized BCI systems. When training data and test data are sampled from different distributions, transfer learning methods are considered to be an appropriate choice in comparison with traditional machine learning approaches. There are two popular branches of transfer learning: domain adaptation and domain generalization. Training with labeled source domain data and unlabeled target domain data, domain adaptation methods focus on model enhancement on the target domain. As another branch of transfer learning, domain generalization [12] considers applying knowledge extracted from multiple related domains to other previously unseen domains. Domain adaptation methods are effective in the circumstances where target domain information is available. However, when dealing with unknown domains without extra information, domain generalization methods are more suitable.

978-1-7281-2009-6/$31.00 paper N-19827.pdf

Many studies have demonstrated that domain adaptation methods are effective in BCI systems [13] [14]. Among these domain adaptation methods, significant performance improvement has been achieved by using deep adversarial models such as Deep Adversarial Neural Network (DANN) [15] [16]. However, a typical characteristic of EEG data is that each individual is regarded as an independent data domain. In order to train an EEG-based model, domain adaptation methods need to collect task-related EEG data from the specific subject, which is quite costly and inconvenient. Another challenging problem is the poor generalization ability of this well-trained model, which is designed for the specific subject and cannot exhibit excellent performance for other new subjects. As a result, domain adaptation approaches are only suitable to the circumstances of personalizing models for specific subjects, where we have to recollect data and retrain a new model whenever there are new subjects to be estimated. Hence, for a BCI system which is applicable for unknown users, i.e., a depersonalized BCI system, domain adaptation methods may become inefficient.

Ideally, a generalized method is supposed to address this problem, which leads to the topic of domain generalization. Domain generalization methods in BCI systems aim at robust performance on unknown subjects with only one depersonalized model. From the perspective of transfer learning, the domain adaptation and domain generalization methods usually correspond to subject-variant and subject-invariant approaches in BCI systems, respectively.

In this paper, we aim to apply domain generalization methods to dealing with depersonalized cross-subject vigilance estimation problem without requiring any information from the new subjects. The main contributions of this paper are the following three aspects. Firstly, to the best of our knowledge, this work is the first to accomplish the depersonalized vigilance estimation task with domain generalization models. Additionally, we first extended DANN to the domain generalization situation, and then proposed a new deep adversarial model called Domain Residual Network (DResNet) by introducing domain residual components, which are similar to the structure of the ResNet [17]. Finally, we applied a conventional domain generalization method to this problem. According to the experimental results on a public multimodal dataset, domain generalization methods can reach a comparable accuracy as the state-of-the-art domain adaptation methods, and have more advantages in stability when dealing with the subject variability problem. Furthermore, domain generalization methods have the capability of conducting vigilance estimation over multiple new subjects with only one well-trained model.

The rest of this paper is organized as follows. Section II gives a brief review about the related work on subject variability and domain generalization. Section III introduces several domain generalization methods we used in this paper. Section IV describes the experiment setup and the dataset. Section V discusses the experimental results. Finally, we conclude our work in Section VI.

II. Related Work

A. Traditional Solutions to Subject Variability in BCI Systems

In recent years, efforts have been made to diminish the influence of subject variability in BCI systems. For the subject-variant approach, a model is firstly pre-trained on the training data from the existing subjects and then a small amount of calibration data from the new subject is used to tune the pre-trained model. Devlaminck [18] provided a multi-subject approach which can reduce the demand of calibration data from the new subject. Kang and Choi [19] used variational inference to learn a shared latent subspace of spatial patterns across subjects. Morioka [8] developed a model which utilized the resting-state data instead of task-based data to tune the model. These methods make it possible for BCI systems to perform well with the calibration data. However, the calibration session needs to be repeated every time before using the BCI systems, which is rather inefficient. Other researchers focus on domain adaptation approaches [14], which utilize the extra information extracted from the unlabeled data of the target subject. Zheng and Lu [13] proposed personalized EEG-based affective models with conventional transfer learning approaches. Li *et al.* [16] further improved the accuracy by using deep adversarial networks.

In contrast, subject-invariant approach aims at designing a robust BCI model which can diminish the subject variability without requiring any calibration data from the new subjects. Fazli *et al.* [20] applied quadratic regression to sparse the ensemble of subject-specific temporal and spatial filter classifiers. Reuderink *et al.* [21] presented a second-order baselining procedure to reduce individual difference. Tu and Sun [22] proposed a subject transfer framework for EEG classification, which can achieve positive knowledge transfer. Samek *et al.* [23] formulated the common spatial pattern algorithm as a divergence maximization problem and provided a subject-invariant framework. Although these methods eliminate the need of calibration, they still suffer from the compromised accuracy.

B. Domain Generalization

Domain generalization can be addressed in three ways. To start with, all the training domains are used to find a domain invariant representation space. Therefore, we can project data from different domains into the common space and learn a general model. Muandet *et al.* [24] proposed a kernel-based optimization algorithm, Domain-Invariant Component Analysis (DICA), to learn a transformation for data from different domains. Focus on cross-domain object recognition, Ghifary *et al.* [25] developed a multi-task autoencoder that attains good generalization performance.

Another intuitive idea is to regulate model weights by exploiting the information from training domains. Khosla *et al.* [26] divided the weights of the classifier into two parts: (1) visual world weights that are common to all domains, and (2) bias weights corresponding to each domain. In this way, more generalized weights and models can be obtained by explicitly

declaring bias weights. Fang *et al.* [27] proposed a metric-based learning algorithm, which has good generalization ability due to the less biased distance metric.

Finally, better generalization performance can be achieved by predicting which known domain is the most similar to the test domain. Xu *et al.* [28] added a nuclear-norm regularizer to an exemplar-SVM to calculate the similarity of positive samples.

In our work, specific solutions to the subject variability problem in depersonalized vigilance estimation have been designed based on the first and second ideas of domain generalization methods mentioned above.

III. Methods

In this section, we first give a description about the domain generalization problem and then elaborate the models that we adopt to deal with the depersonalized vigilance estimation problem in this paper.

Let $\mathcal{X}$ denote the input space, and $\mathcal{Y}$ denote the output space. The set of all joint distributions on $\mathcal{X} \times \mathcal{Y}$ is $\mathbb{P}_{XY}$. Here, we assume the elements of $\mathbb{P}_{XY}$ are observed from a distribution $\boldsymbol{P}$. Then we can define a domain $D_i = \{X_i, Y_i\}$, where $X_i, Y_i = \{(x_1, y_1), (x_2, y_2), ..., (x_{n_i}, y_{n_i})\}$ are n_i samples from the joint distribution $P^i_{XY} \in \mathbb{P}_{XY}$. Respectively, the marginal probability distribution P^i_X and the conditional probability distribution $P^i_{Y|X}$ of domain D_i can be obtained.

In the domain generalization problem, usually there are samples S from k different domains $D_1, D_2, ..., D_k$. We assume the marginal distribution varies among different domains, while the conditional distribution remains stable, i.e., $P^i_X \neq P^j_X, P^i_{Y|X} \approx P^j_{Y|X}$ when $i \neq j$. Given the samples S, our goal is to find a function $f : \mathcal{X} \to \mathcal{Y}$ which can maintain $P_{Y|X}$ despite the changes of P_X. Thus, f can generalize well on test data from any previously unseen domain $D_t = \{X_t\}$, where X_t are sampled from the unknown distribution P^t_X [12]. The domain generalization problem is depicted in Fig. 1.

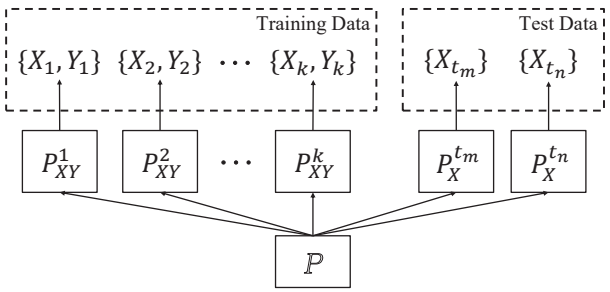


Fig. 1. Illustration of the domain generalization problem.

A. Domain-Invariant Component Analysis (DICA)

DICA defines a distance called distributional variance which measures the similarity of different domains and tries to find a transformation to a low-dimensional subspace that minimizes the distributional variance [24]. In a reproducing kernel Hilbert space (RKHS), distributions can be represented as elements using the mean map function:

$$\mu : \mathbb{P}_x \to \mathcal{H} : P \mapsto \int_{\mathcal{X}} k(x, \cdot) dP(x) =: \mu_P \tag{1}$$

As mentioned above, given input space $\mathcal{X}$, the set of distributions can be denoted as $\mathbb{P} = \{P^1, P^2, ..., P^k\}$, where $\mathbb{P}$ is drawn according to a distribution $\boldsymbol{P}$. Based on the set $\mathbb{P}$, $k \times k$ Gram matrix G is defined as the inner product of the kernel mean maps for P^i and P^j with entries:

$$G_{ij} := \langle \mu_{P^i}, \mu_{P^j} \rangle_{\mathcal{H}} = \iint k(x, z) dP^i(x) dP^j(z) \tag{2}$$

where $i, j = 1, 2, ..., k$.

To measure the divergence between different distributions, the definition of distributional variance with the Gram matrix G is formulated as:

$$\mathbb{V}_{\mathcal{H}}(\mathcal{P}) := \frac{1}{k} tr(\Sigma) = \frac{1}{k} tr(G) - \frac{1}{k^2} \sum_{i,j=1}^{k} G_{ij} \tag{3}$$

where $\mathcal{P}$ denotes a probability distribution on $\mathcal{H}$ with $\mathcal{P}(\mu_{P^i}) = \frac{1}{k}$ and $\Sigma := G - \mathbf{1}_k G - G\mathbf{1}_k + \mathbf{1}_k G \mathbf{1}_k$ is the co-variance operator of $\mathcal{P}$.

Through an orthogonal transform $\mathcal{B}$, DICA finds a domain-invariant m-dimensional subspace where distributional variance across different domains can be minimized. Specifically, let $\mathcal{S} = \{(x^i_m, y^i_m)^{n_i}_{m=1}\}^k_{i=1}$ be the data samples from k domains. For brevity, we denote $\mathcal{S}$ as $\{(x_m, y_m)\}^n_{m=1}$, where $n = \sum^k_{i=1} n_i$. In order to estimate the distributional variance of $\mathcal{S}$, the kernel matrix is defined as

$$K = \begin{bmatrix} K_{1,1} & \cdots & K_{1,k} \\ \vdots & \ddots & \vdots \\ K_{k,1} & \cdots & K_{k,k} \end{bmatrix} \in \mathbb{R}^{n \times n} \tag{4}$$

where $[K_{i,j}]_{m,n} = k(x^i_m, x^j_n)$ is the Gram matrix between domain i and domain j. According to Eq. (3), we can calculate the coefficient matrix Q, where $Q_{i,j} = (k-1)/(k^2 n_i^2)$ if $i = j$, else $-1/(k^2 n_i n_j)$ is in the shape of $\mathbb{R}^{n_i \times n_j}$. Assume $b_m = \sum^n_{i=1} \beta^i_m \phi(x_i) = \Phi_x \boldsymbol{\beta}_m$ is the m-th basis function of $\mathcal{B}$, where $\Phi_x = [\phi(x_1), \phi(x_2), ..., \phi(x_n)]$ and $\boldsymbol{\beta}_m$ is a coefficient vector of n-dimensional. The kernel matrix $\tilde{K}$ for S after transformation $\mathcal{B}$ is:

$$\tilde{K} := (b^T_m \Phi_x)^T b^T_m \Phi_x = KBB^T K \tag{5}$$

where $B = [\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_m}]$. Thus, the empirical distributional variance can be calculated as:

$$\hat{\mathbb{V}}_{\mathcal{H}}(\mathcal{BS}) = tr(\tilde{K}Q) = tr(B^T KQKB) \tag{6}$$

Besides minimizing the distributional variance, DICA also focus on preserving the functional relationship between input data X and the corresponding label Y. Suppose $\Phi_y = [\varphi(y_1), ..., \varphi(y_n)]$ and $L = \Phi^T_y \Phi_y$, the objective function of DICA is:

$$\max_{B \in \mathbb{R}^{n \times m}} \frac{\frac{1}{n} tr(B^T L (L + n\epsilon I_n)^{-1} K^2 B)}{tr(B^T KQKB + BKB)} \tag{7}$$

where ϵ is a kernel regularizer. The unsupervised DICA (UDICA) is a special case of DICA, where $L = \frac{1}{n}I$ and $\epsilon \to 0$. For further details, readers are recommended to refer to [24].

B. Domain Generalization on Domain Adversial Neural Network (DG-DANN)

DANN is a deep domain adaptation model [15], which is trained by labeled data from the source domain and unlabeled data from the target domain. Specifically, DANN contains three components, including feature extractor G_f, label predictor G_y, and domain classifier G_d. For simplicity, we only consider the case where there is only one layer in each component.

The feature extractor G_f maps the p-dimensional input features to a new d-dimensional space by learning a function $G_f : \mathcal{X} \to \mathbb{R}^d$. Hence, new features can be extracted from inputs through G_f with an activation function f and parameters $\theta_f = \{W_f, b_f\} \in \mathbb{R}^{d\times p} \times \mathbb{R}^d$ by calculating:

$$G_f(x;\theta_f) = f(W_f x + b_f) \tag{8}$$

The label predictor G_y predicts the label of inputs through function: $G_y(G_f(X);\theta_y)$. The prediction loss of a sample (x_i, y_i) is defined as $L_y(\hat{y}_i, y_i)$ with the prediction $\hat{y}_i$.

In DANN, the domain classifier G_d is a binary classifier: $G_d(G_f(X);\theta_d)$ since the inputs come from either the source domain or the target domain in the topic of domain adaptation. Thus, we can obtain the domain prediction $\hat{d}_i$ of a sample x_i. According to [15], the loss of G_d is defined as:

$$\begin{aligned} L_d(G_d(G_f(x_i)), d_i) = d_i \log \frac{1}{G_d(G_f(x_i))} \\ + (1-d_i)\log\frac{1}{1-G_d(G_f(x_i))} \end{aligned} \tag{9}$$

with a sample (x_i, d_i), where d_i is the binary domain label for sample x_i. For brevity, we denote the loss as $L_d(\hat{d}_i, d_i)$.

Assume there are N samples including n labeled source domain samples and n' unlabeled target domain samples. We can formulate the loss function of the three components together as:

$$\begin{aligned} E(\theta_f,\theta_y,\theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y(\hat{y}_i, y_i) \\ - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} L_d(\hat{d}_i, d_i) + \frac{1}{n'}\sum_{i=n+1}^{N} L_d(\hat{d}_i, d_i)\right) \end{aligned} \tag{10}$$

The optimization is organized as:

$$\begin{aligned} (\hat{\theta_f}, \hat{\theta_y}) = arg\min_{\theta_f,\theta_y} E(\theta_f, \theta_y, \hat{\theta_d}) \\ (\hat{\theta_d}) = arg\max_{\theta_d} E(\hat{\theta_f}, \hat{\theta_y}, \theta_d) \end{aligned} \tag{11}$$

which is integrated by a special-designed layer called Gradient Reversal Layer (GRL) between G_f and G_d [15]. After the optimization, G_d becomes a good domain classifier and G_y should perform well for the features extracted by G_f. Meanwhile, G_f is supposed to find a mapping to the feature space where task-related knowledge are reserved and most of the domain-variant information are excluded.

In this paper, we focus on the domain generalization situation, and extend DANN to domain generalization. Assume we can sample data from k different known domains, and there are N samples (x_i, y_i, d_i) in total, including n_k samples from each domain D_k. Following the idea of finding a domain-invariant feature space, we can preserve the feature extractor G_f and the label predictor G_y, and generalize the domain classifier G_d as a k-class domain classifier, where the parameters become $\theta_d = \{W_d, b_d\} \in \mathbb{R}^{k\times d} \times \mathbb{R}^k$. Accordingly, the loss of G_d can be modified with respect to the correct domain label d_i as:

$$L_d(G_d(G_f(x_i)), d_i) = \log\frac{1}{G_d(G_f(x_i))_{d_i}} \tag{12}$$

Additionally, the loss function of the Domain Generalization version of DANN (DG-DANN) is formulated as:

$$E(\theta_f,\theta_y,\theta_d) = \frac{1}{N}\sum_{i=1}^{N} L_y(\hat{y}_i, y_i) - \lambda\frac{1}{N}\sum_{i=1}^{N} L_d(\hat{d}_i, d_i) \tag{13}$$

The architecture of DG-DANN for regression tasks is illustrated in Fig. 2. After the optimization, facilitated by the well-trained domain classifier G_d, the feature extractor G_f is supposed to achieve the target of finding the domain-invariant feature space, where the loss function could be minimized.

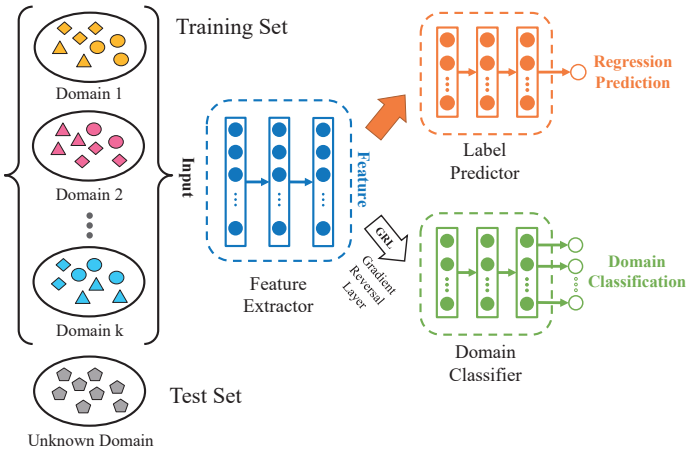


Fig. 2. The architecture of DG-DANN for regression tasks. Colors and shapes for each domain in the training set mean different domain shifts and labels. For the data in the test set, the domain shifts and the labels are unknown.

C. Domain Residual Network (DResNet)

Another intuitive idea for domain generalization is to regulate model parameters using the information of the training domains. As mentioned above, each domain D_i corresponding to a joint distribution P^i_{XY} is observed from the distribution $\boldsymbol{P}$, which represents the common space. Hence, domain shifts are biases during the observation. Inspired by [26], we assume that the bias of a known domain can be modeled with parameters and the model parameters for common space can be trained

together with bias parameters in a joint manner. Since the domain bias parameters are similar to the residual part of the ResNet, we named this model as Domain Residual Network (DResNet).

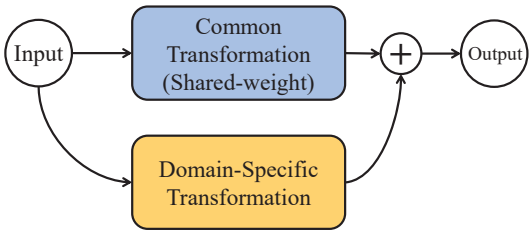


Fig. 3. One layer of feature extractor G_f in DResNet for a specific domain

The proposed DResNet focuses on the feature extractor G_f as defined in the DG-DANN model. According to [26], improvement on generalization ability of the model can be achieved though training a set of parameters θ_f^c in G_f for the common space $\boldsymbol{P}$ by undoing the domain biases. For each known domain D_i, the bias can be explicitly defined by parameters $\theta_f^{\delta_i}$. θ_f^c and $\theta_f^{\delta_i}$ are related by the equation:

$$\theta_f^i = \theta_f^c + \theta_f^{\delta_i} = \{W_f^c + W_f^{\delta_i}, b_f^c + b_f^{\delta_i}\} \tag{14}$$

Consequently, for an input x from domain i, the feature extractor G_f for DResNet is organized as:

$$G_f(x; \theta_f^i) = f\left((W_f^c x + b_f^c) + (W_f^{\delta_i} x + b_f^{\delta_i})\right) \tag{15}$$

As shown in Fig. 3, the feature extractor is divided into two parts: a shared-weight common part which is same for all domains and a domain-specific part where the parameters are unique for each known domain D_i.

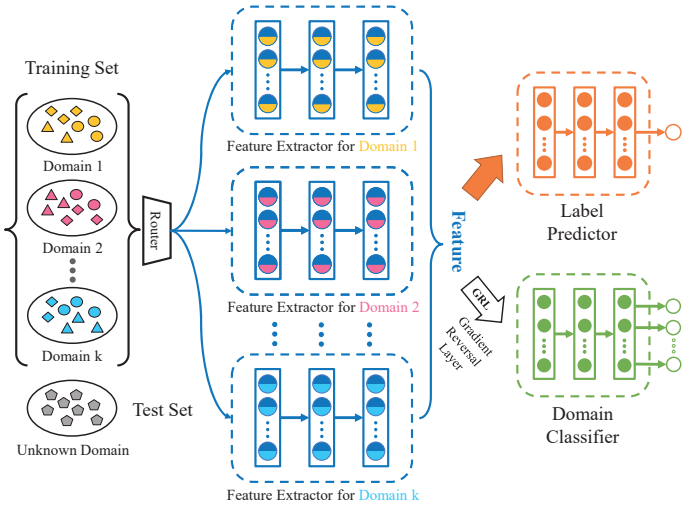


Fig. 4. The architecture of our proposed DResNet for regression tasks. The different colors and shapes for each domain in the training set represent the domain discrepancy and the label for each sample, respectively. For each layer in the feature extractor, the dark blue component depicts the shared-weight part, and the other domain-specific part is colored according to the domain. For the test set, the domain distribution and the labels are unknown.

During the back propagation, for each sample (x_i, y_i, d_i), the gradient only updates the parameters in the common part and the i-th domain-specific part. We can formulate the optimization as:

$$\begin{aligned}(\hat{\theta_f^i}, \hat{\theta_y}) &= arg\min_{\theta_f, \theta_y} E(\theta_f^i, \theta_y, \hat{\theta_d}) \\ (\hat{\theta_d}) &= arg\max_{\theta_d} E(\hat{\theta_f^i}, \hat{\theta_y}, \theta_d)\end{aligned} \tag{16}$$

When DResNet predicts the labels for an unknown test domain, it only activate the common part of the G_f and the label predictor G_y. The whole structure of DResNet is illustrated in Fig. 4

IV. Experiments

A. The SEED-VIG Dataset

SEED-VIG is a public multimodal vigilance estimation dataset[1]. Driving experiments were conducted based on an established simulation system. During the experiment, subjects were required to drive in a real vehicle placed in the lab. Monitored by a software, their motions including stepping on the accelerator, braking and turning the steering wheel are reflected on a large LCD screen which displays the animation of simulated road situation.

Twenty-three healthy participants with normal or corrected-to-normal vision volunteered to participate in the experiment. In order to make it easier for subjects to get tired during the simulated driving, all experiments are performed after lunch or at night. Each subject drove for two hours, during which both the forehead EEG and EOG signals were recorded using Neuroscan system with a sampling rate of 1000 Hz. The data were labeled by the percentage of eye closure (PERCLOS) recorded by the eye tracking glasses [29], which ranges from 0 (high vigilance level) to 1 (low vigilance level).

B. Preprocessing and Feature Extraction

The EEG data were further downsampled to 125 Hz and segmented by a 8-second time window without overlapping. Hence, there are 885 features for each subject. Since EOG and EEG data were recorded simultaneously thus mixed together, independent component analysis (ICA) was applied to find the EEG component from the raw data. The differential entropy (DE) features [3] were then extracted in every 2 Hz band from 1 Hz to 50 Hz for all of the 4 electrodes set on the forehead. Therefore, the dimension of EEG features is 100.

As for the EOG features, the raw data was firstly preprocessed by the ICA method to attain the EOG component, including the vertical EOG (VEO) and the horizontal EOG (HEO). Then after performing the Mexican hat wavelet transform, we can extract 36 eye movement features including the rate and amplitude of blink, saccade and fixation with peak detection. Concatenating the EOG features with the EEG features, we finally obtained 885 samples with 136 multimodal features for each subject. In order to remove the artifacts as possible, we applied the moving average method to further smoothen the features with a window size of 30.

[1]http://bcmi.sjtu.edu.cn/seed/download.html

C. Evaluation Details

In this paper, the DG methods are evaluated from the perspectives of both estimation accuracy and generalization ability. Thus, we adopt two settings for the evaluation. Firstly, leave-one-subject-out cross validation was applied to compare the estimation performance of the DG methods with the DA algorithms. During each validation, we set the data from one subject as the unknown test domain and used the data from the other 22 subjects as the known training domains for the DG methods. Whereas for the DA method, we set one test subject as the target domain and regarded the other 22 subjects together as one source domain. The other setting for evaluation is to randomly select one third of the subjects (8 people) as the unknown test domains to be estimated by the DG model trained with the other two thirds of the subjects' data (15 people). In aim of maintaining the same test granularity as the first evaluation setting, we randomly selected 23 sets of the test subjects and measured the performance on each test domain. By estimating on different test domains using the same well-trained DG model, we can evaluate the generalization ability of the DG methods.

We used the linear support vector regression (SVR) algorithm [30] as the baseline method in both settings. For the DG methods, we evaluated Domain-Invariant Component Analysis (DICA) [24], Domain Generalization on DANN (DG-DANN) and Domain Residual Network (DResNet). Specifically, for the first evaluation setting, we compared the DG methods with shallow traditional algorithms, e.g., Transfer Component Analysis (TCA) [31], Maximum Independence Domain Adaptation(MIDA) [32], as well as the deep DA models, such as Domain Adversarial NeuroNetwork (DANN) [15] and Adversarial Discriminative Domain Adaptation (ADDA) [33].

The estimation accuracy of is evaluated by the Pearson's correlation coefficient (PCC) and the root-mean-square-error (RMSE). Higher PCCs represent higher similarity between our prediction and the ground truth, and lower RMSEs represent more accurate predictions. For the dimension reduction methods, the subspace dimension was selected in the range of $\{10, 20, ..., 120\}$. For the linear SVR, the parameters C and ϵ were searched from $\{2^n | n \in \{-10, ..., 10\}\}$ and $\{10^n | n \in \{-5, ..., -1\}$, respectively. For the deep models, we applied Adam optimizer and applied random search strategy for parameter tuning. The search space for learning rate and the hyper parameter λ for GRL were set as $\{2^n \times 10^{-4} | n \in [-10, 10]\}$ and $\{10^n | n \in [-5, -1]\}$.

V. Results and Discussion

A. Leave-one-subject-out Evaluation

The estimation accuracies of the DG and DA models are compared using the leave-one-subject-out settings. Table I presents the average (Avg) and standard deviation (Std) of PCC and RMSE across different approaches. Domain generalization methods yield similar but more stable accuracy performance comparing with domain adaptation methods. Specifically, for deep domain generalization methods, DResNet and DG-DANN achieve much more stable performance with the lowest Std value in both PCC and RMSE. The results demonstrate that the DG-DANN and DResNet are comparable to the deep DA models in terms of estimation accuracy and outperform DICA and the shallow DA methods. Besides, DResNet performs slightly better than DG-DANN, with the average PCC of 0.8440. For shallow approaches, we can observe that DICA performs better than TCA and MIDA in terms of the PCC with a lower Std value, which indicates that the DG methods are more stable than the DA models.

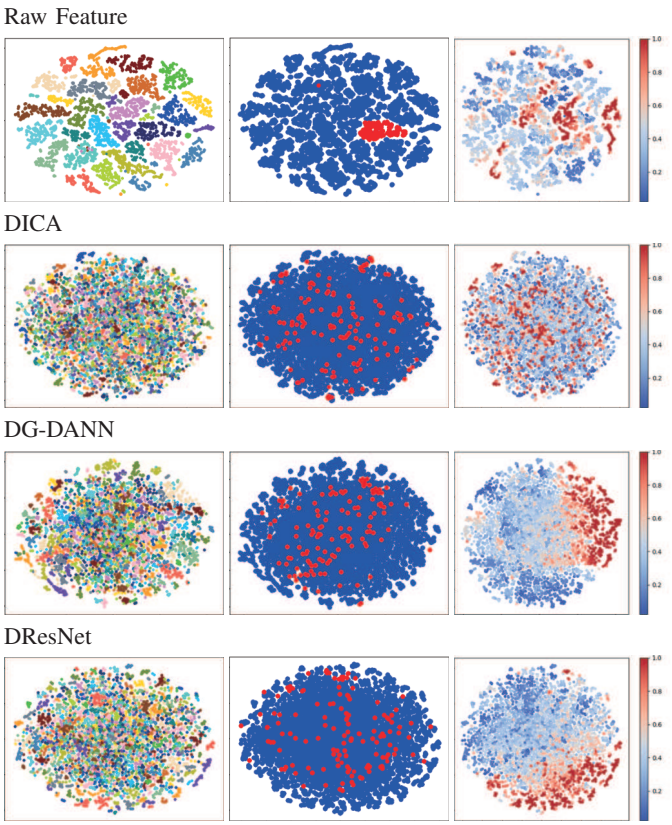
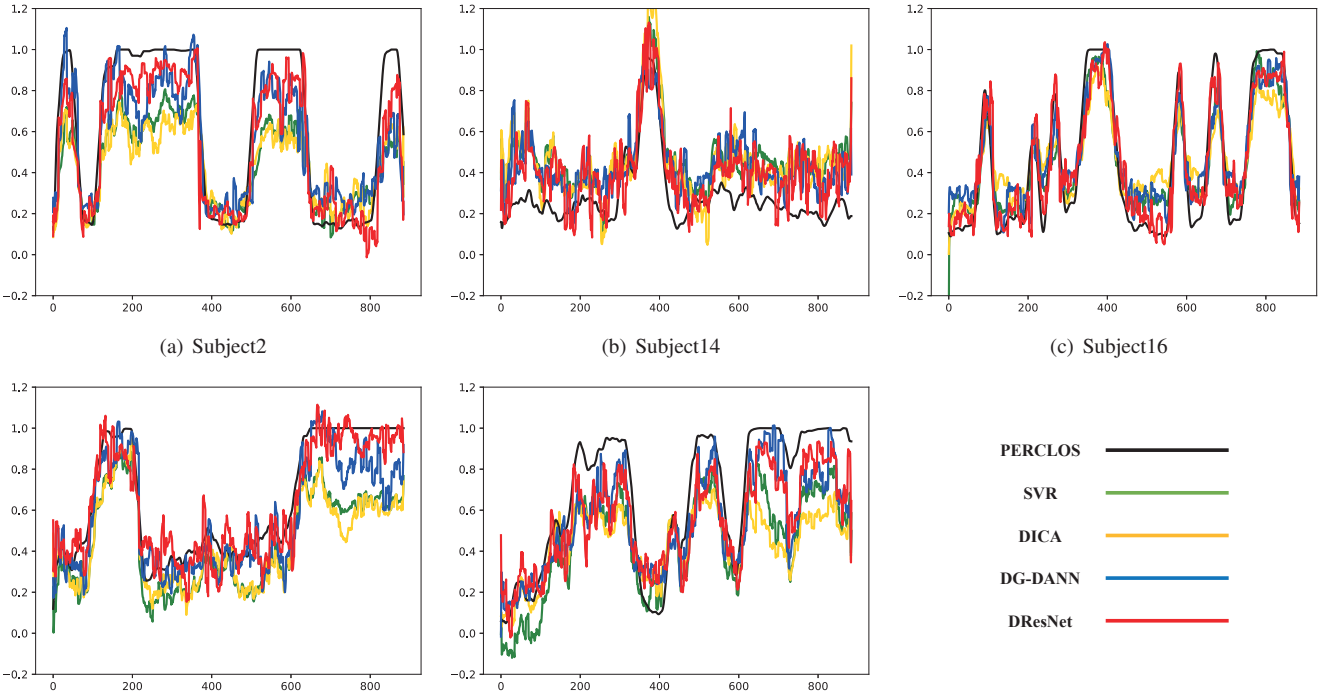


Fig. 5. Domain generalization feature visualization.

We give further analysis of the features extracted from the DG models with t-SNE [34] visualization shown in Fig. 5. The first column of figures displays the features from the training set and the test set which are colored according to their domains. In the second column, we marked all the training data as blue points and the test data as red points. The last column shows these features according to their labels. The four rows represent the features from the original data, DICA, DG-DANN and DResNet, respectively. The first row describes the phenomenon of domain shift in the raw data of SEED-VIG dataset. The other three rows indicate that all the features from different domains are mixed evenly after applying the DG approaches, which verifies that the DG models can reduce the subject variability and map the biased features to approximately the same distribution. Thus, the model trained with the known domains can perform considerably well when estimating data from the unknown domains. Furthermore, benefiting from the unified training, the feature extractors of

TABLE I
RESULTS OF LEAVE-ONE-SUBJECT-OUT EVALUATION

		Baseline	Domain Adaptation Methods				Domain Generalization Methods		
		SVR	TCA	MIDA	DANN	ADDA	DICA	DG-DANN	DResNet
PCC	Avg	0.7606	0.7786	0.7858	0.8402	**0.8442**	0.7733	0.8320	0.8440
	Std	0.2314	0.2152	0.1900	0.1535	0.1336	0.1382	0.1000	**0.0935**
RMSE	Avg	0.1689	0.1596	0.1840	0.1427	**0.1405**	0.2007	0.1470	0.1420
	Std	0.0673	0.0544	0.0753	0.0588	0.0514	0.0674	0.0444	**0.0402**



(a) Subject2 (b) Subject14 (c) Subject16

(d) Subject19 (e) Subject22 (f) Legend

Fig. 6. The vigilance prediction of five subjects with different methods.

DG-DANN and DResNet can be trained with both domain and label informations. Therefore, label prediction is much easier with features extracted from the deep DG models which can account for the obvious trends of the label changes in the third column of DG-DANN and DResNet features in Fig. 5.

The estimation results of five subjects are depicted in Fig. 6. We present all the predictions of the three DG methods and compare them with the true label (PERCLOS) and the baseline model (SVR). The DG methods yield the similar prediction trend as the true label. Moreover, deep DG models attain higher accuracy for most of the subjects.

B. Multiple-random-subject-out Evaluation

For the large-scale BCI applications in practice, it is of great importance for models to generalize well on multiple unknown subjects. In order to evaluate the generalization ability of the DG methods, we randomly selected 23 different sets of 8 subjects as the unknown domains, and then gave vigilance estimation for each subject based on the DG models trained on the other 15 subjects. It should be noted that the DA approaches are inefficient for this estimation strategy since we need to train several different models for each test set as there are multiple target domains. Table II summarizes the results of the DG approaches and the baseline SVR model in the multiple-random-subject-out evaluation. We can see that the performance of each method declines slightly due to the decreased size of the training set. Nevertheless, since there are 15 domains in the training set, the DG methods can leverage the sufficient domain information to overcome the subject variability and perform well on the 8-subject test set. In addition, DG-DANN and DResNet are less influenced and still outperform other approaches robustly and DResNet gives the best performance with the average PCC of 0.8386 and the average RMSE of 0.1569. These results demonstrate that DG methods are effective when addressing the depersonalized cross-subject vigilance estimation problem.

TABLE II
EVALUATION RESULTS OF GENERALIZATION PERFORMANCE

		Baseline	DICA	DG-DANN	DResNet
PCC	Avg	0.7499	0.7719	0.8294	**0.8386**
	Std	0.1980	0.1841	0.1541	**0.1532**
RMSE	Avg	0.2068	0.1735	0.1604	**0.1569**
	Std	0.0587	0.0468	0.0782	**0.0735**

VI. Conclusion

In this paper, we have studied the depersonalized cross-subject vigilance estimation problem, which aims at effective regression models for unknown subjects. To reduce the subject variability, we have introduced the idea of domain generalization, where models can be trained without any information from the unknown test subject. We have proposed two novel deep adversarial DG models and adopted a popular conventional DG method. Based on two intuitive ideas for DG, we have extended DANN to DG and have proposed a novel structure called DResNet. In our methods, subject variability can be diminished and information from both label and domain can be utilized to further improve the performance. Evaluations under two different settings on a public dataset have indicated that our proposed DG methods are effective for reducing subject variability on depersonalized cross-subject vigilance estimation problem.

Acknowledgment

This work was supported in part by the grants from the National Key Research and Development Program of China (Grant No. 2017YFB1002501), the National Natural Science Foundation of China (Grant No. 61673266), and the Fundamental Research Funds for the Central Universities.

References

[1] G. Zhang, K. K. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Analysis & Prevention*, vol. 87, pp. 34–42, 2016.

[2] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: a review," *Sensors*, vol. 12, no. 12, pp. 16 937–16 953, 2012.

[3] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[4] C.-T. Lin, R.-C. Wu, S.-F. Liang, W.-H. Chao, Y.-J. Chen, and T.-P. Jung, "EEG-based drowsiness estimation for safety driving using independent component analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 12, pp. 2726–2738, 2005.

[5] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Proceedings of the Twenty-Forth International Joint Conference on Artificial Intelligence*, vol. 15, 2015, pp. 1170–1176.

[6] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[7] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *Journal of Neural Engineering*, vol. 14, no. 2, p. 026017, 2017.

[8] H. Morioka, A. Kanemura, J.-i. Hirayama, M. Shikauchi, T. Ogawa, S. Ikeda, M. Kawanabe, and S. Ishii, "Learning a common dictionary for subject-transfer decoding with resting calibration," *NeuroImage*, vol. 111, pp. 167–178, 2015.

[9] J. Gu and R. Kanai, "What contributes to individual differences in brain structure?" *Frontiers in Human Neuroscience*, vol. 8, p. 262, 2014.

[10] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[11] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[12] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in Neural Information Processing Systems*, 2011, pp. 2178–2186.

[13] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2732–2738.

[14] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[16] H. Li, W.-L. Zheng, and B.-L. Lu, "Multimodal vigilance estimation with adversarial domain adaptation networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–6.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multisubject learning for common spatial patterns in motor-imagery BCI," *Computational Intelligence and Neuroscience*, vol. 2011, p. 8, 2011.

[19] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Networks*, vol. 57, pp. 39–50, 2014.

[20] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural Networks*, vol. 22, no. 9, pp. 1305–1312, 2009.

[21] B. Reuderink, J. Farquhar, M. Poel, and A. Nijholt, "A subject-independent brain-computer interface based on smoothed, second-order baselining," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4600–4604.

[22] W. Tu and S. Sun, "A subject transfer framework for EEG classification," *Neurocomputing*, vol. 82, pp. 109–116, 2012.

[23] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[24] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, 2013, pp. 10–18.

[25] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proceedings of the IEEE international Conference on Computer Vision*, 2015, pp. 2551–2559.

[26] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *European Conference on Computer Vision*. Springer, 2012, pp. 158–171.

[27] C. Fang, Y. Xu, and D. 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*, 2013, pp. 1657–1664.

[28] Z. Xu, W. Li, L. Niu, and D. Xu, "Exploiting low-rank structure from latent domains for domain generalization," in *European Conference on Computer Vision*. Springer, 2014, pp. 628–643.

[29] X.-Y. Gao, Y.-F. Zhang, W.-L. Zheng, and B.-L. Lu, "Evaluating driving fatigue detection algorithms using eye tracking glasses," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2015, pp. 767–770.

[30] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[31] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[32] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 288–299, 2018.

[33] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 4.

[34] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.