Domain Generalized RPPG Network: Disentangled Feature Learning with Domain Permutation and Domain Augmentation

Wei-Hao Chung, Cheng-Ju Hsieh, Sheng-Hung Liu, and
Chiou-Ting Hsu[0000−0001−8857−2481]

National Tsing Hua University, Hsinchu, Taiwan
godofyax@gmail.com, peter55180831@gmail.com, oscar545407@gmail.com, and
cthsu@cs.nthu.edu.tw

Abstract. Remote photoplethysmography (rPPG) offers a contactless method for monitoring physiological signals from facial videos. Existing learning-based methods, although work effectively on intra-dataset scenarios, degrade severely on cross-dataset testing. In this paper, we address the cross-dataset testing as a domain generalization problem and propose a novel DG-rPPGNet to learn a domain generalized rPPG estimator. To this end, we develop a feature disentangled learning framework to disentangle rPPG, identity, and domain features from input facial videos. Next, we propose a domain permutation strategy to further constrain the disentangled rPPG features to be invariant to different domains. Finally, we design a novel adversarial domain augmentation strategy to enlarge the domain sphere of DG-rPPGNet. Our experimental results show that DG-rPPGNet outperforms other rPPG estimation methods in many cross-domain settings on UBFC-rPPG, PURE, COHFACE, and VIPL-HR datasets.

1 Introduction

Since the outbreak of new epidemics, remote estimation of human physiological states has attracted enormous attention. Remote Photoplethysmography (rPPG), which analyzes the blood volume changes in optical information of facial videos, is particularly useful in remote heart rate (HR) estimation. Earlier methods [1–8] usually adopted different prior assumptions to directly analyze the chromaticity of faces. Recent deep learning-based methods [9–19], through either multi-stage or end-to-end training networks, have achieved significant breakthroughs in rPPG estimation.

Although existing learning-based methods performed satisfactorily in intra-dataset testing, their cross-dataset testing performance tends to degrade severely. This cross-dataset or cross-domain issue is especially critical in rPPG estimation, because different rPPG datasets were recorded using their own equipment, under different environments or lighting conditions and thus exhibit a broad diversity. For example, the videos in UBFC-rPPG dataset [20] were recorded at 30 fps in a well-lighted environment; whereas the videos in COHFACE dataset [21] were recorded at 20 fps under two illumination conditions. The PURE dataset [22] even includes different motion settings when recording the videos. Therefore, if the training and testing data are from different datasets, the model trained in one dataset usually fails to generalize to another one.

We address the cross-dataset testing issue as a domain generalization (DG) problem and assume different "domains" refer to different characteristics (e.g., illumination conditions or photographic equipment) in the rPPG benchmarks. Domain generalization has been developed to facilitate the model to unseen domains at the inference time. Previous methods [23–32] have shown the effectiveness of DG on many classification tasks. However, many of these DG mechanisms, such as contrastive loss or triplet loss, are designed for classification problems and are inapplicable to the regression problem of rPPG estimation. Moreover, because rPPG signals are extremely vulnerable in comparison with general video content, any transformation across different domains (e.g., video-to-video translation, illumination modification, and noise perturbation) will substantially diminish the delicate rPPG signals.

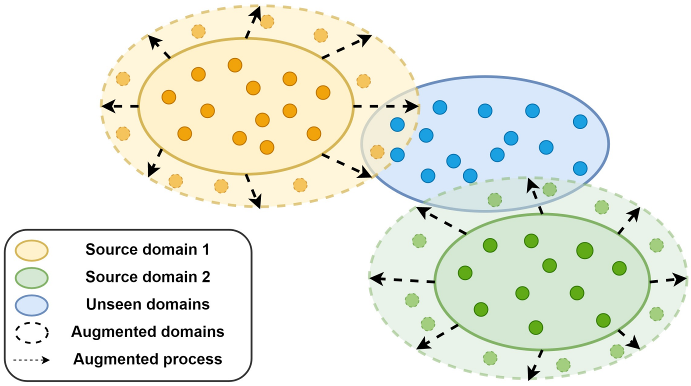


Fig. 1: Illustration of the proposed domain augmentation.

In this paper, we propose a novel Domain Generalized rPPG Network (DG-rPPGNet) via disentangled feature learning to address the domain generalization problem in rPPG estimation. Through the feature disentanglement framework, we first disentangle the rPPG, identity (ID), and domain features from the input data. Next, we develop two novel strategies, including domain permutation and domain augmentation, to cope with the disentangled feature learning. We devise a domain permutation strategy to ensure that the disentangled rPPG features are also invariant to different source domains. In addition, to generalize the model towards unseen domains, we propose a learnable domain augmentation strategy to enlarge the domain sphere during the model training. As illustrated in Figure 1, the proposed domain augmentation aims to generate "adversarial domains", which maximally degrade the prediction accuracy of the rPPG estimator, to offer the model with information outside the source domain boundaries.

Our contributions are summarized below:

1) We propose a novel end-to-end training network DG-rPPGNet for rPPG estimation. To the best of our knowledge, this is the first work focusing on domain generalization issue in rPPG estimation.
2) We devise a disentangled feature learning framework, cooperated with domain permutation and domain augmentation, to significantly increase the generalization capability of rPPG estimation on unseen domains.
3) Experimental results on UBFC-rPPG, PURE, COHFACE, and VIPL-HR datasets show that the proposed DG-rPPGNet outperforms other rPPG estimation methods in cross-domain testing.

2 Related Work

2.1 Remote Photoplethysmography Estimation

Earlier methods [1–8] adopted different assumptions to design hand-crafted methods for rPPG estimations and usually do not generalize well to videos recorded in less-controlled scenarios. The learning-based methods [9–19], either through multi-stage processing or end-to-end training, benefit from the labeled data and largely improve the estimation performance over traditional methods. For example, in [14], a Dual-GAN framework is proposed to learn a noise-resistant mapping from the pre-processed spatial-temporal maps into the ground truth blood volume pulse (BVP) signals. In [18], a network is proposed to enhance highly compressed videos and to recover rPPG signals from the enhanced videos. In [10], a multi-task framework is developed to augment the rPPG dataset and to predict rPPG signals simultaneously. In [19], a video transformer is proposed to adaptively aggregate both local and global spatio-temporal features to enhance rPPG representation. Nevertheless, these learning-based methods mostly focus on improving intra-dataset performance but rarely concern the domain generalization issue.

2.2 Feature Disentanglement

Disentangled feature learning aims to separate the informative (or explanatory) variations from multifactorial data and has been extensively studied for learning task-specific feature representation in many computer vision tasks. For example, disentangled representation learning has been included in detecting face presentation attacks [33] or in unsupervised cross-domain adaptation [34]. In [35], a cross-verified feature disentangling strategy is proposed for remote physiological measurement. By disentangling physiological features from non-physiological features, the authors in [35] improved the robustness of physiological measurements from disentangled features.

2.3 Domain Generalization

Domain generalization aims to learn a representation or a model from multiple source domains and to generalize to unseen target domains. Because the target domains are unseen in the training stage, there is no way to match the source-to-target distributions or to minimize the cross-domain shift when developing the model. Therefore, most domain generalization methods either focus on learning domain-invariant representation or design different augmentation strategies to enlarge the source domains. For example, to address the DG issue in object detection, the authors in [27] designed a disentangled network to learn domain-invariant representation on both the image and instance levels. In [28], to tackle the semantic segmentation issue in real-world autonomous driving scenarios, the authors proposed using domain randomization and pyramid consistency to learn generalized representation. In [30], the authors proposed to augment perturbed images to enable the image classifier to generalize to unseen domains.

3 Proposed Method

3.1 Problem Statement and Overview of DG-rPPGNet

In the domain generalization setting, we are given a set of M source domains $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_M\}$ but have no access to the target domain $\mathcal{T}$ during the training stage. Let $\mathcal{S}_i = \{(x_j, s_j, y_j^{id}, y_j^{domain})\}_{j=1}^{N_i}$ denote the ith source domain, and x_j, s_j, y_j^{id}, and y_j^{domain} denote the facial video, the ground truth PPG signal, the subject ID label, and the domain label, respectively. Assuming that the unseen target domain $\mathcal{T}$ has very different distribution with the source domains, our goal is to learn a robust and well-generalized rPPG estimator to correctly predict the rPPG signals for the facial videos from any unseen target domain.

In this paper, we propose a novel DG-rPPGNet to tackle the domain generalization problem in rPPG estimation from three aspects. First, we develop a disentangled feature learning framework to disentangle rPPG-relevant features from other domain-dependent variations. Second, we devise a domain permutation strategy to ensure that the disentangled rPPG features are invariant to different source domains. Finally, we design a learnable domain augmentation strategy to augment the source domains to enable DG-rPPGNet to generalize to unseen domains.

Figure 2 illustrates the proposed DG-rPPGNet, which includes a global feature encoder F, three extractors (i.e., S_{rPPG}, S_{id}, and S_{domain}) for feature disentanglement, two decoders (i.e., $D_{feature}$ and D_{video}), two rPPG estimators (i.e., E_{rPPG}^{global} and E_{rPPG}^{disent}), one ID classifier C_{id}, and one domain classifier C_{domain}. All these components in DG-rPPGNet are jointly trained using the total loss defined in Sec. 3.5.

3.2 Disentangled Feature Learning

In this subsection, we describe the disentangled feature learning in DG-rPPGNet and the corresponding loss terms. For each input facial video x, we first obtain

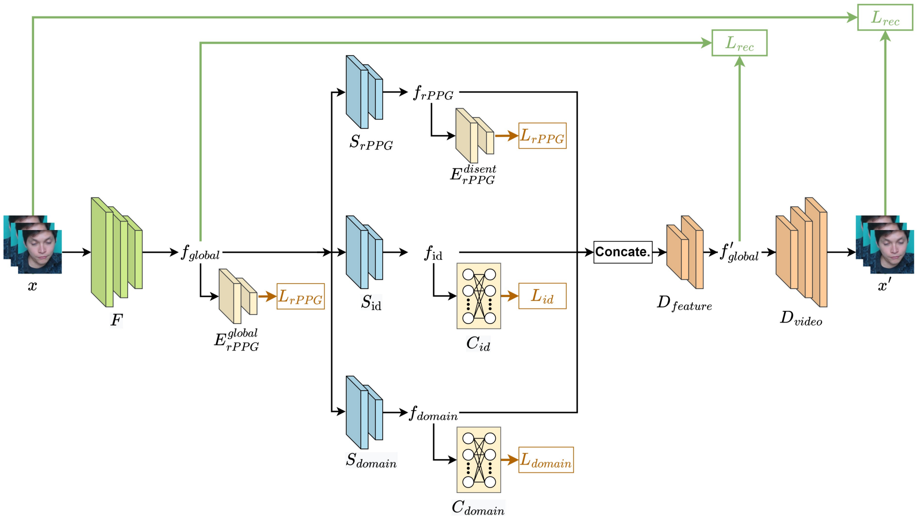


Fig. 2: The proposed DG-rPPGNet, consisting of a global feature encoder F, a rPPG extractor S_{rPPG}, an ID extractor S_{id}, a domain extractor S_{domain}, a feature decoder $D_{feature}$, a video decoder D_{video}, two rPPG estimators E_{rPPG}^{global} and E_{rPPG}^{disent}, an ID classifier C_{id}, and a domain classifier C_{domain}.

its global feature f_{global} using the global feature encoder F by,

$$f_{global} = F(x). \tag{1}$$

Next, we use three extractors S_{rPPG}, S_{id}, and S_{domain} to disentangle the rPPG feature f_{rPPG}, the ID feature f_{id}, and the domain feature f_{domain} from f_{global}, respectively:

$$\begin{aligned} f_{rPPG} &= S_{rPPG}(f_{global}), & (2)\\ f_{id} &= S_{id}(f_{global}), \text{ and} & (3)\\ f_{domain} &= S_{domain}(f_{global}). & (4) \end{aligned}$$

To constrain the disentangled feature learning in DG-rPPGNet, we define three prediction consistent losses for each of the features f_{rPPG}, f_{id}, and f_{domain}, and one reconstruction loss to jointly train the model.

The prediction consistent losses are defined by minimizing the prediction losses of the corresponding labels by,

$$\begin{aligned} L_{rPPG}^{disent} &= L_{np}(E_{rPPG}^{global}(f_{global}), s) \\ &\quad + L_{np}(E_{rPPG}^{disent}(f_{rPPG}), s), & (5)\\ L_{id}^{disent} &= CE(C_{id}(f_{id}), y^{id}), \text{ and} & (6)\\ L_{domain}^{disent} &= CE(C_{domain}(f_{domain}), y^{domain}), & (7) \end{aligned}$$

where L_{np} is the negative Pearson correlation between the predicted signal s' and the ground truth signal s:

$$L_{np}(s', s) = 1 - \frac{(s-\bar{s})^t(s'-\bar{s'})}{\sqrt{(s-\bar{s})^t(s-\bar{s})}\sqrt{(s'-\bar{s'})^t(s'-\bar{s'})}}, \tag{8}$$

and $CE(\cdot)$ denotes the cross-entropy loss. Note that, in Equation (5), because rPPG signals are more vulnerable than the other two, we additionally include $E^{global}_{rPPG}(f_{global})$ to constrain the rPPG consistent loss. Since both f_{global} and f_{rPPG} capture the same rPPG signal, the double constraints in Equation (5) not only consolidate the rPPG feature disentanglement but also accelerate the model convergence.

We define the reconstruction loss L^{disent}_{rec} by enforcing the decoder $D = D_{video} \circ D_{feature}$ to reconstruct the input video in both the global feature space f_{global} and the color space x by,

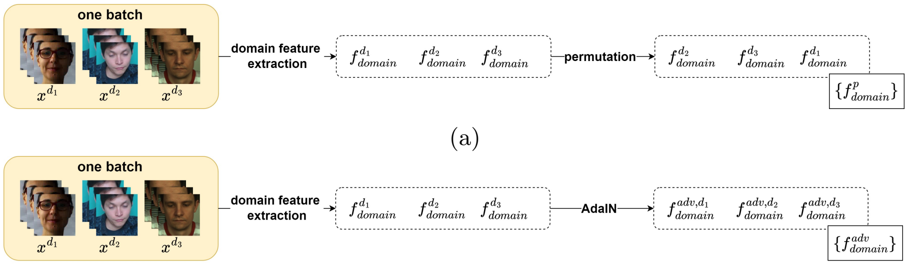
$$L^{disent}_{rec} = ||f_{global} - f'_{global}||_1 + ||x - x'||_1, \tag{9}$$

where

$$f'_{global} = D_{feature}(f_{rPPG}, f_{id}, f_{domain}), \text{ and} \tag{10}$$

$$x' = D_{video}(f'_{global}). \tag{11}$$

3.3 Domain Permutation for Domain-Invariant Feature Learning



(a)

(b)

Fig. 3: Illustrations of the (a) domain permutation; and (b) domain augmentation.

In Equations (5) (6) and (7), although we constrain DG-rPPGNet to extract f_{rPPG}, f_{id}, and f_{domain} from f_{global}, there is no guarantee that these features are successfully disentangled. Specifically, our major concern is that the extracted rPPG feature f_{rPPG} should capture not only rPPG-relevant information but also be invariant to different domains. In other words, we expect that

f_{rPPG} should contain little or no domain-dependent variations. Therefore, in this subsection, we devise a novel domain permutation strategy to consolidate the feature disentanglement and further encourage S_{rPPG} to focus on extracting domain-invariant rPPG features.

Given one batch of input facial videos, we first extract their global features by Equation (1) and then extract the rPPG features, ID features, and domain features by Equations (2) (3) and (4). Next, we randomly permute the locations of domain features $\{f_{domain}\}$ within this batch and have the permuted domain features $\{f^p_{domain}\}$ by,

$$\{f^p_{domain}\} = Permute(\{f_{domain}\}), \tag{12}$$

where $Permute(\cdot)$ is a random permutation operation.

An example is given in Figure 3 (a), where the input batch consists of three videos x^{d_1}, x^{d_2}, and x^{d_3} sampled from different domains and their original domain features are $f^{d_1}_{domain}$, $f^{d_2}_{domain}$, and $f^{d_3}_{domain}$, respectively. After random permutation, the three videos have their new domain features as $f^{d_2}_{domain}$, $f^{d_3}_{domain}$, and $f^{d_1}_{domain}$, respectively. Our rationale is that, if the disentangled rPPG feature f_{rPPG} and ID feature f_{id} are indeed invariant to different domains, then the global feature f^p_{global} reconstructed using the permuted domain features, i.e.,

$$f^p_{global} = D_{feature}(f_{rPPG}, f_{id}, f^p_{domain}), \tag{13}$$

should carry the same rPPG feature and ID feature as the original one f_{global}.

Next, we reconstruct a video by decoding f^p_{global} and then encode this reconstructed video to obtain its global feature $f^{'p}_{global}$ by,

$$f^{'p}_{global} = F(D_{video}(f^p_{global})). \tag{14}$$

Finally, we extract the three features f^p_{rPPG}, f^p_{id}, and $f^{'p}_{domain}$ from $f^{'p}_{global}$ by,

$$f^p_{rPPG} = S_{rPPG}(f^{'p}_{global}), \tag{15}$$

$$f^p_{id} = S_{id}(f^{'p}_{global}), \text{ and} \tag{16}$$

$$f^{'p}_{domain} = S_{domain}(f^{'p}_{global}). \tag{17}$$

Similar to Equations (5) (6) and (7), we define the prediction consistent losses to constrain f^p_{rPPG}, f^p_{id}, and $f^{'p}_{domain}$ by,

$$\begin{aligned} L^p_{rPPG} &= L_{np}(E^{global}_{rPPG}(f^{'p}_{global}), s) \\ &\quad + L_{np}(E^{disent}_{rPPG}(f^p_{rPPG}), s), \end{aligned} \tag{18}$$

$$L^p_{id} = CE(C_{id}(f^p_{id}), y^{id}), \text{ and} \tag{19}$$

$$L^p_{domain} = CE(C_{domain}(f^{'p}_{domain}), Permute(y^{domain})), \tag{20}$$

where $Permute(y^{domain})$ is the ground truth domain label of $f^{'p}_{domain}$.

We also define a reconstruction loss L^p_{rec} to constrain that (1) the rPPG and ID features should remain unchanged, before and after the domain permutation; and (2) the permuted domain features and the global features should remain the same after the decoding and re-encoding steps. We thus formulate the reconstruction loss L^p_{rec} in the feature spaces by,

$$\begin{aligned} L^p_{rec} = &||f_{rPPG} - f^p_{rPPG}||_1 + ||f_{id} - f^p_{id}||_1 \\ &+ ||f^p_{domain} - f^{'p}_{domain}||_1 + ||f^p_{global} - f^{'p}_{global}||_1. \end{aligned} \tag{21}$$

3.4 Domain Augmentation via AdaIN

In Sec. 3.2 and Sec. 3.3, the disentangled feature learning and domain permutation involve only the set of source domains S in the model training but are oblivious to any external domains. Because data augmentation is widely adopted to alleviate the data shortage, we adopt the idea and design a domain augmentation strategy to enlarge the sphere of source domains.

Unlike most generic augmentation methods, the proposed domain augmentation has two specific goals. First, the augmented domains should well preserve discriminative information in the original source domains S; and second, they should offer diverse characteristics different from those in S so as to simulate the unseen domains. To balance the two competing goals, we propose (1) using AdaIN [36] to generate the augmented domains by transforming the style of S without changing their discriminative content, and then (2) enforcing the augmented domains to act as the adversaries of S so as to expand the sphere of S. In the second part, we adopt the idea of adversary examples [37] and define "adversarial domains" as the domains generated to mislead the rPPG estimators and to offer unseen information to the model.

We formulate the proposed data augmentation as an adversary domain learning problem in the parameter space of AdaIN. Given one batch of input facial videos, we extract their global features by Equation (1) and then extract the rPPG features f_{rPPG}, ID features f_{id}, and domain features f_{domain} by Equations (2) (3) and (4). respectively. Next, we use AdaIN [36] to transform the domain feature from f_{domain} to f^{adv}_{domain} by,

$$f^{adv}_{domain} = AdaIN(f_{domain}, \alpha, \beta), \tag{22}$$

where

$$AdaIN(f, \alpha, \beta) = \alpha \cdot \frac{f - \mu_f}{\sigma_f} + \beta, \tag{23}$$

α and β are two learnable parameters; μ_f and σ_f are the mean and standard deviation of the feature map f, respectively. An example is shown in Figure 3 (b), where the domain features $f^{d_i}_{domain}(i = 1, 2, 3)$ of x^{d_i} are transformed by AdaIN into f^{adv,d_i}_{domain}.

To ensure f^{adv}_{domain} behaves like adversarial domains, we constrain the two parameters α and β by "maximizing" the prediction losses L_{np} of the two rPPG estimators E^{global}_{rPPG} and E^{disent}_{rPPG} by,

$$\begin{aligned}(\alpha, \beta) = \underset{\alpha,\beta}{argmax}\ & L_{np}(E^{global}_{rPPG}(f^{'adv}_{global}), s) \\ & +L_{np}(E^{disent}_{rPPG}(f^{adv}_{rPPG}), s), \end{aligned} \tag{24}$$

where $f^{'adv}_{global}$ and f^{adv}_{rPPG} are the re-encoded global feature and the extracted rPPG feature of $f^{'adv}_{global}$ obtained by:

$$f^{'adv}_{global} = F(D_{video}(f^{adv}_{global})),\ \text{and} \tag{25}$$

$$f^{adv}_{rPPG} = S_{rPPG}(f^{'adv}_{global}), \tag{26}$$

and $f^{'adv}_{global}$ is re-encoded from the reconstructed global feature f^{adv}_{global} by,

$$f^{adv}_{global} = D_{feature}(f_{rPPG}, f_{id}, f^{adv}_{domain}). \tag{27}$$

To facilitate the implementation of Equation (24), we use the gradient reversal layer (GRL) [38] to flip the gradients (1) between F and E^{global}_{rPPG}, and (2) between S_{rPPG} and E^{disent}_{rPPG}. Hence, we reformulate Equation (24) by,

$$\begin{aligned}(\alpha, \beta) = \underset{\alpha,\beta}{argmin}\ & L_{np}(E^{global}_{rPPG}(GRL(f^{'adv}_{global})), s) \\ & +L_{np}(E^{disent}_{rPPG}(GRL(f^{adv}_{rPPG})), s). \end{aligned} \tag{28}$$

Finally, we extract the other two features f^{adv}_{id} and f^{adv}_{domain} from $f^{'adv}_{global}$ by,

$$f^{adv}_{id} = S_{id}(f^{'adv}_{global}),\ \text{and} \tag{29}$$

$$f^{'adv}_{domain} = S_{domain}(f^{'adv}_{global}). \tag{30}$$

Similar to Equations (18) (19) and (21), we again impose the prediction consistent losses L^{adv}_{rPPG}, L^{adv}_{id} and the reconstruction loss L^{adv}_{rec} to constrain the model learning by,

$$\begin{aligned} L^{adv}_{rPPG} &= L_{np}(E^{global}_{rPPG}(f^{'adv}_{global}), s) \\ &\quad + L_{np}(E^{disent}_{rPPG}(f^{adv}_{rPPG}), s), \end{aligned} \tag{31}$$

$$L^{adv}_{id} = CE(C_{id}(f^{adv}_{id}), y^{id}),\ \text{and} \tag{32}$$

$$\begin{aligned} L^{adv}_{rec} &= ||f_{rPPG} - f^{adv}_{rPPG}||_1 + ||f_{id} - f^{adv}_{id}||_1 \\ &\quad + ||f^{adv}_{domain} - f^{'adv}_{domain}||_1 + ||f^{adv}_{global} - f^{'adv}_{global}||_1. \end{aligned} \tag{33}$$

Here, we do not include the domain prediction consistent loss, because there exist no ground truth labels for the adversarial domains.

3.5 Loss Function

Finally, we include the feature disentanglement loss L_{total}^{disent}, the domain permutation loss L_{total}^{p}, and the domain augmentation loss L_{total}^{adv} to define the total loss L_{total}:

$$L_{total} = L_{total}^{disent} + L_{total}^{p} + L_{total}^{adv}, \tag{34}$$

where

$$L_{total}^{disent} = \lambda_1 L_{rPPG}^{disent} + \lambda_2 L_{id}^{disent} + \lambda_3 L_{domain}^{disent} + L_{rec}^{disent}, \tag{35}$$

$$L_{total}^{p} = \lambda_1 L_{rPPG}^{p} + \lambda_2 L_{id}^{p} + \lambda_3 L_{domain}^{p} + L_{rec}^{p}, \text{ and} \tag{36}$$

$$L_{total}^{adv} = \lambda_1 L_{rPPG}^{adv} + \lambda_2 L_{id}^{adv} + L_{rec}^{adv}, \tag{37}$$

and λ_1, λ_2, and λ_3 are hyper-parameters and are empirically set as 0.01 in all our experiments.

3.6 Inference Stage

In the inference stage, we include only the global feature encoder F and the global rPPG estimator E_{rPPG}^{global} to predict the rPPG signals of the test facial videos. We do not include the rPPG extractor S_{rPPG} and the rPPG estimator E_{rPPG}^{disent} during the inference stage, because they are trained only on source domains and may not well disentangle the features on unseen domains.

4 Experiments

4.1 Datasets and Cross-Domain Setting

The UBFC-rPPG dataset [20] consists of 42 RGB videos from 42 subjects; i.e., each subject contributes one single video. The videos were recorded by Logitech C920 HD Pro at 30 fps with resolution of 640 × 480 pixels in uncompressed 8-bit format. The PPG signals and corresponding heart rates were collected by CMS50E transmissive pulse oximeter. We follow the setting in [14] to split the dataset into the training and testing sets with videos from 30 and 12 subjects, respectively.

The PURE dataset [22] consists of 60 RGB videos from 10 subjects. Each subject performs 6 different activities, including (1) sitting still and looking directly at the camera, (2) talking, (3) slowly moving the head parallel to the camera, (4) quickly moving the head, (5) rotating the head with 20° angles, and (6) rotating the head with 35° angles. All videos were recorded using an eco274CVGE camera at 30 fps and with resolution of 640 × 480 pixels. The PPG signals were captured by using Pulox CMS50E finger clip pulse oximeter with sampling rate of 60 Hz. To align with the videos, the PPG signals are reduced to 30 Hz. We follow the setting in [10] to split the dataset into the training and testing sets with videos from 7 and 3 subjects, respectively.

The COHFACE dataset [21] consists of 160 one-minute-long sequence RGB videos from 40 subjects. The videos were recorded under two illumination conditions, including (1) a well-lighted environment, and (2) a natural light environment. All videos were recorded using Logitech HD C525 at 20 fps with resolution of 640x480 pixels. The PPG signals were taken by a contact blood volume pulse sensor model SA9308M. We follow the setting in [10] to split the dataset into the training and testing sets with videos from 24 and 16 subjects, respectively.

The VIPL-HR dataset [39] contains 2378 RGB videos from 107 subjects. The dataset was recorded using 3 different devices under 9 scenarios. We follow the setting in [16] and use a subject-exclusive 5-fold cross-validation protocol on VIPL-HR. In addition, because the facial videos and PPG signals have different sampling rates, we resample the PPG signals to match the corresponding video frames by linear interpolation.

Cross-Domain Setting. When experimenting on these datasets, we consider each dataset refers to one domain, except COHFACE, which is considered as two domains and each one refers to either the well-lighted or natural light settings. In Sec. 4.4 and 4.5, we adopt two experimental settings by randomly choosing two datasets from UBFC-rPPG, PURE, and COHFACE to form the set of source domains and then testing on (1) the remaining one (that is, we have three cross-domain settings: "P+C→U", "U+C→P", and "U+P→C") in Sec. 4.4 and 4.5; and (2) the VIPL-HR dataset in Sec. 4.5.

4.2 Implementation Details

The architectures of the global feature encoder F, the extractor S, the decoder D, the rPPG estimator E, and the classifier C in DG-rPPGNet are given in the supplementary file. We train DG-rPPGNet in two stages. We first train the disentangled feature learning model with domain permutation for 300 epochs and then fine-tune the model with domain augmentation for 100 epochs. We train the model using Nvidia RTX 2080 and RTX 3080 with one sample from each domain in one batch, and use Adam optimizer with the learning rate of 0.0002. For all the facial videos, we use [40] to detect face landmarks, crop the coarse face area and resize the cropped areas into 80×80 pixels. In each epoch, we randomly sample 60 consecutive frames from the training videos of each domain to train DG-rPPGNet.

4.3 Evaluation Metrics

To assess the performance of DG-rPPGNet on rPPG estimation, we follow [10] to derive heart rate (HR) from the predicted rPPG signals and then evaluate the results in terms of the following metrics: (1) Mean absolute error (MAE), (2) Root mean square error (RMSE), and (3) Pearson correlation coefficient (R).

4.4 Ablation Study

We conduct ablation studies on three cross-domain settings, including "P+C→U", "U+C→P", and "U+P→C". In Table 1, L^{disent}_{total}, L^{p}_{total}, and L^{adv}_{total} indicate that we include the corresponding losses as defined in Equations (35) (36) and (37), respectively, to train DG-rPPGNet.

We first evaluate the effectiveness of domain permutation. When including $L^{disent}_{total} + L^{p}_{total}$ in the model training, we significantly improve the performance over using L^{disent}_{total} alone by reducing MAE about 85% and RMSE about 86% in "P+C→U". However, because the proposed domain permutation focuses on learning domain-invariant features within the source domains, the model still lacks the ability to generalize to unseen domains. Nevertheless, although we achieve no improvement in "U+C→P" and "U+P→C" with $L^{disent}_{total} + L^{p}_{total}$, we see that the setting $L^{disent}_{total} + L^{p}_{total} + L^{adv}_{total}$ significantly outperforms $L^{disent}_{total} + L^{adv}_{total}$ when we further include domain augmentation in DG-rPPGNet. These results show that the proposed domain permutation works cooperatively with domain augmentation to support the model to learn domain-invariant and rPPG-discriminative features in the augmented domains. Finally, when including L^{adv}_{total} in DG-rPPGNet, we see significant performance improvement with reduced MAE (about 81%, 28%, and 3%; and about 92%, 50% and 42%) and RMSE (about 80%, 29%, and 6%; and about 89%, 54% and 41%), without and with L^{p}_{total}, respectively. These results verify that the proposed domain augmentation substantially enlarges the domain sphere and enables the model to generalize to unseen domains.

Table 1: Ablation study

Loss terms			P+C→U			U+C→P			U+P→C		
L^{disent}_{total}	L^{p}_{total}	L^{adv}_{total}	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑
✓			7.74	12.34	0.40	6.14	10.26	0.56	12.54	15.30	0.07
✓	✓		1.17	1.71	0.83	6.53	11.96	0.47	14.33	16.93	0.08
✓		✓	1.46	2.46	0.78	4.36	7.19	0.70	12.18	14.35	**0.39**
✓	✓	✓	**0.63**	**1.35**	**0.88**	**3.02**	**4.69**	**0.88**	**7.19**	**8.99**	0.30

In Table 2, we evaluate the proposed domain augmentation by comparing with random domain augmentation. We simulate the random domain augmentation by replacing the parameters α and β in Equation (28) with values randomly sampled from standard Gaussian distribution. The results in Table 2 show that the proposed method outperforms the random augmentation with reduced MAE (by about 38%, 50% and 33%) and RMSE (by about 18%, 56% and 29%) and verify its effectiveness on domain generalization.

Table 2: Evaluation of Domain Augmentation

Augmentation	P+C→U			U+C→P			U+P→C		
	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑
Random	1.02	1.65	0.82	6.08	10.54	0.56	10.80	12.71	0.25
Adversarial	**0.63**	**1.35**	**0.88**	**3.02**	**4.69**	**0.88**	**7.19**	**8.99**	**0.30**

4.5 Results and Comparison

We compare our results on the three settings: "P+C→U", "U+C→P", and "U+P→C", with previous methods [1–5, 10, 14] in Tables 3, 4, and 5, respectively. Note that, the methods [1–5] (marked with †) are not learning-based methods and thus have neither training data nor cross-domain issue. The other learning-based methods (marked with ∗) all adopt different cross-domain settings from ours. Although there exist no results reported under the same cross-domain settings as ours for a fair comparison, we include their results here to assess the relative testing performance on these rPPG datasets.

Table 3: Cross-domain test on "P+C→U"

Method	MAE↓	RMSE↓
GREEN† [2]	8.29	15.82
ICA† [3]	4.39	11.60
POS† [4]	3.52	8.38
CHROM† [1]	3.10	6.84
Multi-task* [10]	1.06	2.70
Dual-GAN* [14]	0.74	**1.02**
DG-rPPGNet	**0.63**	1.35

Table 4: Cross-domain test on "U+C→P"

Method	MAE↓	RMSE↓
LiCVPR† [5]	28.22	30.96
POS† [4]	22.25	30.20
ICA† [3]	15.23	21.25
GREE† [2]	9.03	13.92
CHROM† [1]	3.82	6.8
Multi-task* [10]	4.24	6.44
DG-rPPGNet	**3.02**	**4.69**

In Table 3, we show our results on "P+C→U" and compare with previous methods with testing results on UBFC-rPPG. We cite the performance of the four non-learning-based methods, i.e., GREEN [2], ICA [3], POS [4], and CHROM [1], from [9]. The two methods, Multi-task [10] and Dual-GAN [14], are trained on PURE dataset; and their cross-domain setting is considered as "P→U". Table 3 shows that DG-rPPGNet achieves the best performance with MAE 0.63 even without involving UBFC-rPPG in the training stage.

In Table 4, we show our results on "U+C→P" and compare with previous methods with testing results on PURE. We cite the result of LiCVPR [5] from [16], and use an open source toolbox [41] to obtain the results of GREEN [2], POS [4], and ICA [3] on PURE. The model of Multi-task [10] is trained on UBFC-rPPG alone, i.e., its cross-domain setting here is "U→P". We again show that DG-rPPGNet outperforms the other methods with MAE 3.02 and RMSE 4.69.

In Table 5, we show our results on “U+P→C” and compare with previous methods with testing results on COHFACE. Again, we use the open source toolbox [41] to obtain the results of POS [4], ICA [3], and GREEN [2] on COHFACE. The results once again show that DG-rPPGNet outperforms the other methods with MAE 7.19 and RMSE 8.99.

Table 5: Cross-domain test on “U+P→C”

Method	MAE↓	RMSE↓
POS† [4]	19.86	24.57
LiCVPR† [5]	19.98	25.59
ICA† [3]	14.27	19.28
GREEN† [2]	10.94	16.72
CHROM† [1]	7.8	12.45
DG-rPPGNet	**7.19**	**8.99**

Table 6: Cross-domain test on VIPL-HR

Method	MAE↓	RMSE↓
Averaged GT	22.21	26.70
DG-rPPGNet (U+P)	18.38	18.86
DG-rPPGNet (U+C)	18.23	18.81
DG-rPPGNet (P+C)	15.95	17.47

Finally, in Table 6, we evaluate the generalization capability of the proposed DG-rPPGNet. We use different combinations of the three small-scale datasets PURE, COHFACE, and UBFC-rPPG as the source domains and then test on the large-scale dataset VIPL-HR [39]. Because there exist no similar experimental results, we show the averaged ground truth signals (marked by “Averaged GT”) of VIPL-HR as the baseline results for comparison. The results show that, even trained on small-scale datasets, the proposed DG-rPPGNet substantially exceeds the baseline results and reduces MAE (by about 17%, 18%, and 28%) and RMSE (by about 29%, 30%, and 35%).

5 Conclusion

In this paper, we propose a DG-rPPGNet to address the domain generalization issue in rPPG estimation. The proposed DG-rPPGNet includes (1) a feature disentangled learning framework to extract rPPG, ID, and domain features from facial videos; (2) a novel domain permutation strategy to constrain the domain invariant property of rPPG features; and (3) an adversarial domain augmentation strategy to increase the domain generalization capability. Experimental results on UBFC-rPPG, PURE, COHFACE, and VIPL-HR datasets show that the proposed DG-rPPGNet outperforms other rPPG estimation methods in many cross-domain testings.

References

1. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering **60** (2013) 2878–2886
2. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Optics express **16** (2008) 21434–21445
3. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express **18** (2010) 10762–10774
4. Wang, W., Den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote ppg. IEEE Transactions on Biomedical Engineering **64** (2016) 1479–1491
5. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 4264–4271
6. Wang, W., Stuijk, S., De Haan, G.: A novel algorithm for remote photoplethysmography: Spatial subspace rotation. IEEE transactions on biomedical engineering **63** (2015) 1974–1984
7. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering **60** (2013) 2878–2886
8. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE transactions on biomedical engineering **58** (2010) 7–11
9. Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., Chen, X.: Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE Journal of Biomedical and Health Informatics **25** (2021) 1373–1384
10. Tsou, Y.Y., Lee, Y.A., Hsu, C.T.: Multi-task learning for simultaneous video generation and remote photoplethysmography estimation. In: Proceedings of the Asian Conference on Computer Vision. (2020)
11. Bousefsaf, F., Pruski, A., Maaoui, C.: 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. Applied Sciences **9** (2019) 4364
12. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 349–365
13. Lee, E., Chen, E., Lee, C.Y.: Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In: European Conference on Computer Vision, Springer (2020) 392–409
14. Lu, H., Han, H., Zhou, S.K.: Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12404–12413
15. Niu, X., Han, H., Shan, S., Chen, X.: Synrhythm: Learning a deep heart rate estimator from general to specific. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE (2018) 3580–3585
16. Špetlík, R., Franc, V., Matas, J.: Visual heart rate estimation with convolutional neural network. In: Proceedings of the british machine vision conference, Newcastle, UK. (2018) 3–6
17. Tsou, Y.Y., Lee, Y.A., Hsu, C.T., Chang, S.H.: Siamese-rppg network: Remote photoplethysmography signal estimation from face videos. In: Proceedings of the 35th annual ACM symposium on applied computing. (2020) 2066–2073

18. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 151–160
19. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P.H., Zhao, G.: Physformer: facial video-based physiological measurement with temporal difference transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 4186–4196
20. Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters **124** (2019) 82–90
21. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. arXiv preprint arXiv:1709.00962 (2017)
22. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, IEEE (2014) 1056–1062
23. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: Metareg: Towards domain generalization using meta-regularization. Advances in neural information processing systems **31** (2018)
24. Li, Y., Yang, Y., Zhou, W., Hospedales, T.: Feature-critic networks for heterogeneous domain generalization. In: International Conference on Machine Learning, PMLR (2019) 3915–3924
25. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 834–843
26. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE international conference on computer vision. (2017) 5542–5550
27. Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., Cai, J.: Domain-invariant disentangled network for generalizable object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 8771–8780
28. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 2100–2110
29. Li, L., Gao, K., Cao, J., Huang, Z., Weng, Y., Mi, X., Yu, Z., Li, X., Xia, B.: Progressive domain expansion network for single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 224–233
30. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 13025–13032
31. Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745 (2018)
32. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: Self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 9619–9628
33. Wang, G., Han, H., Shan, S., Chen, X.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6678–6687

34. Lee, S., Cho, S., Im, S.: Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 15252–15261
35. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: European Conference on Computer Vision, Springer (2020) 295–310
36. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 4401–4410
37. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
38. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning, PMLR (2015) 1180–1189
39. Niu, X., Han, H., Shan, S., Chen, X.: Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In: Asian Conference on Computer Vision, Springer (2018) 562–576
40. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision. (2017)
41. McDuff, D., Blackford, E.: iphys: An open non-contact imaging-based physiological measurement toolbox. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE (2019) 6521–6524