Dual-bridging with Adversarial Noise Generation for Domain Adaptive rPPG Estimation

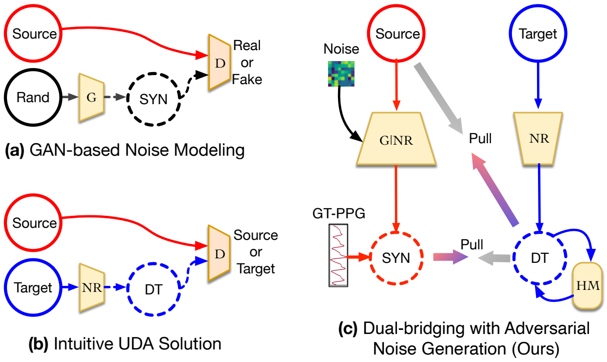
Jingda Du*, Si-Qi Liu*, Bochao Zhang, Pong C. Yuen
Department of Computer Science, Hong Kong Baptist University, Hong Kong
csjddu, siqiliu, csbczhang, pcyuen @comp.hkbu.edu.hk

Abstract

The remote photoplethysmography (rPPG) technique can estimate pulse-related metrics (e.g. heart rate and respiratory rate) from facial videos and has a high potential for health monitoring. The latest deep rPPG methods can model in-distribution noise due to head motion, video compression, etc., and estimate high-quality rPPG signals under similar scenarios. However, deep rPPG models may not generalize well to the target test domain with unseen noise and distortions. In this paper, to improve the generalization ability of rPPG models, we propose a dual-bridging network to reduce the domain discrepancy by aligning intermediate domains and synthesizing the target noise in the source domain for better noise reduction. To comprehensively explore the target domain noise, we propose a novel adversarial noise generation in which the noise generator indirectly competes with the noise reducer. To further improve the robustness of the noise reducer, we propose hard noise pattern mining to encourage the generator to learn hard noise patterns contained in the target domain features. We evaluated the proposed method on three public datasets with different types of interferences. Under different cross-domain scenarios, the comprehensive results show the effectiveness of our method.

1. Introduction

With the development of rPPG technology, physiological metrics such as heart rate [27], heart rate variability [34], respiratory rate [21] can also be estimated from facial videos. Deep learning-based rPPG methods overcome non-physiological intensity variations [30, 49] and model noise in training samples [24, 28]. Despite the high accuracy under intra-dataset evaluations, the deep rPPG models may not be able to generalize well to unseen interferences in the test domain. The domain gap is mainly from unseen non-physiological interferences such as lighting conditions, camera sensors, video compression algorithms, facial expressions, *etc*. They can induce distortions in estimated rPPG signals and reduce both the accuracy and the reliability of pulse-related metrics estimation. Considering it is hard to cover all interferences during the training stage, to improve the usability of rPPG in realistic applications, one main challenge is how to boost the generalizability of rPPG models to unseen scenarios.



(a) GAN-based Noise Modeling

(b) Intuitive UDA Solution

(c) Dual-bridging with Adversarial Noise Generation (Ours)

Figure 1. The comparison between (a) typical intra-dataset adversarial rPPG noise modeling, (b) an intuitive UDA framework for rPPG feature alignment, and (c) our proposed dual-bridging network with adversarial noise modeling and hard noise pattern mining (HM). Here SYN denotes synthetic data, DT is for the denoised target domain, G, D, and NR denotes the generator, domain classifier, and noise reducer, respectively.

In recent research of rPPG, both deep learning-based frameworks and mechanisms [45, 46, 49] are proposed to overcome the non-physiological intensity variations. GAN-based disentanglement learning has also been adopted to reduce the noise from pseudo [28] or synthesized [24] noisy features. We summarize this approach in Figure 1 (a) where a discriminator is employed to distinguish the generated feature (SYN in figure) from the original one. These methods can perform well under intra-dataset evaluation settings since the in-distribution noise patterns are thoroughly investigated with a large number of adversarial learning iterations. However, they may fail when encountering unseen

[*] Equal contribution

domains in real application scenarios since noise patterns may be different from the ones of training data.

Intuitively, the unsupervised domain adaptation (UDA) technique can help in bridging the gap between source and target domain [7, 14, 17, 44]. As shown in Figure 1 (b), a noise reducer module NR that aims to obtain noise-free domain-invariant representations can be learned by fighting against the domain classifier D. However, this intuitive solution may not work well since the domain classification may not be able to give sufficient information for NR to identify whether the feature components are noise or physiological information. Directly aligning the rPPG features from different domains may end up distorting the physiological information since they are from different subjects. The ground-truth PPG (GT-PPG) signal with detailed waveform information helps preserve the physiological information and can provide much more informative guidance with the regression task. However, GT-PPG is available in the source domain but not the target domain. How to leverage the source domain GT-PPG to train NR to be robust to the noise from the target domain is the key issue to be solved in this work. To achieve it we propose the dual bridging noise modeling network as shown in Figure 1 (c). The first bridging works as high-level guidance where the denoised target domain feature is adversarially pulled to the source domain feature (as Figure 1 (b)). On top of it, the second bridging aims to help synthesize the target domain noise and inject it into the source domain denoised feature so that the GT-PPG regression can help finetune the NR for better robustness in the target domain. An adversarial noise generation module ($G|NR$) is designed where the generator is conditioned on the NR so that it keeps on overcoming the complex noise pattern that can hardly be solved in the first bridging. With the high-level guidance (first bridging) and detailed signal regression (second bridging), the NR can handle the target domain noise better and therefore improve the accuracy of rPPG estimation in the target domain. To further discover the remaining noise vestige, we build a hard noise pattern mining mechanism to squeeze the unsolved local noise pattern from the denoised target feature so that $G|NR$ can thoroughly synthesize it.

In sum, the contributions of this work are: (1) A dual-bridging noise modeling network that adapts target domain noise in a coarse-to-fine manner. (2) An adversarial noise generation mechanism to progressively synthesize and inject the hard target domain noisy features into the source domain while keeping the physiological information. (3) A hard noise pattern mining mechanism to further explore the target domain noise patterns with larger variations. We evaluated the proposed method on three public datasets with various types of interferences including facial motion and expression, video compression, skin tone, and heartbeat ranges. Under different cross-domain scenarios, the comprehensive results show the effectiveness of our method.

2. Related Work

2.1. Remote physiological estimation

Traditional rPPG methods [5, 16, 32, 39, 40, 41] estimate pulse signal from facial videos by extracting and modeling the detailed heartbeat-caused skin color variation. To estimate pulse-induced intensity variations with more details from facial videos, deep learning-based networks and mechanisms have been proposed. With facial video input, various spatial-temporal neural networks have been developed [4, 20, 22, 23, 37, 48, 49]. To evaluate the effectiveness, rPPG datasets with various interferences including head motion [36], facial expression [15, 39], video compression [9], and skin color [42] are constructed. In addition, rPPG estimation can be conducted from pre-processed representations like normalized difference [3, 18, 21, 30], spatial-temporal map [24, 28]. To further improve robustness, self-adaptive [3, 21] and background-guided [19, 30] attention mechanisms have been proposed to emphasize important facial regions in physiological representation. Furthermore, [28] proposed the cross-verified strategy to disentangle noise and physiological representation and [24] proposed to model in-distribution noise for learning noise-resistant physiological representation. [25, 26] proposed to synthesize diverse facial videos for rPPG estimation. Unsupervised learning-based methods [8, 37] are also proposed to learn rPPG estimation from unlabeled facial videos. However, the previously proposed methods are prone to experience performance drops in cross-dataset evaluation, where the target domain may contain unseen noise. This is due to unseen noise patterns that may not be fully overcome, and the physiological representation may be polluted and cause distortions in the rPPG signals.

2.2. Unsupervised domain adaptation methods

In recent research on computer vision, many unsupervised domain adaptation methods have been proposed to learn domain-invariant representation. [7] proposed to suppress domain-specific information with domain label guided adversarial training. [33] proposed a bi-directional generation framework that can map target domain samples to the source domain and preserve semantic information for classification. [14] proposed to use adversarial domain training and distributional feature alignment guided by maximum density divergence to achieve adaptation. [43] considered that domain adversarial training may not be coordinated with the main task and proposed using meta-learning to improve the effectiveness of feature alignment. Unlike common computer vision tasks, feature alignment in the rPPG estimation requires more detailed information which cannot be fully provided by domain labels. There-

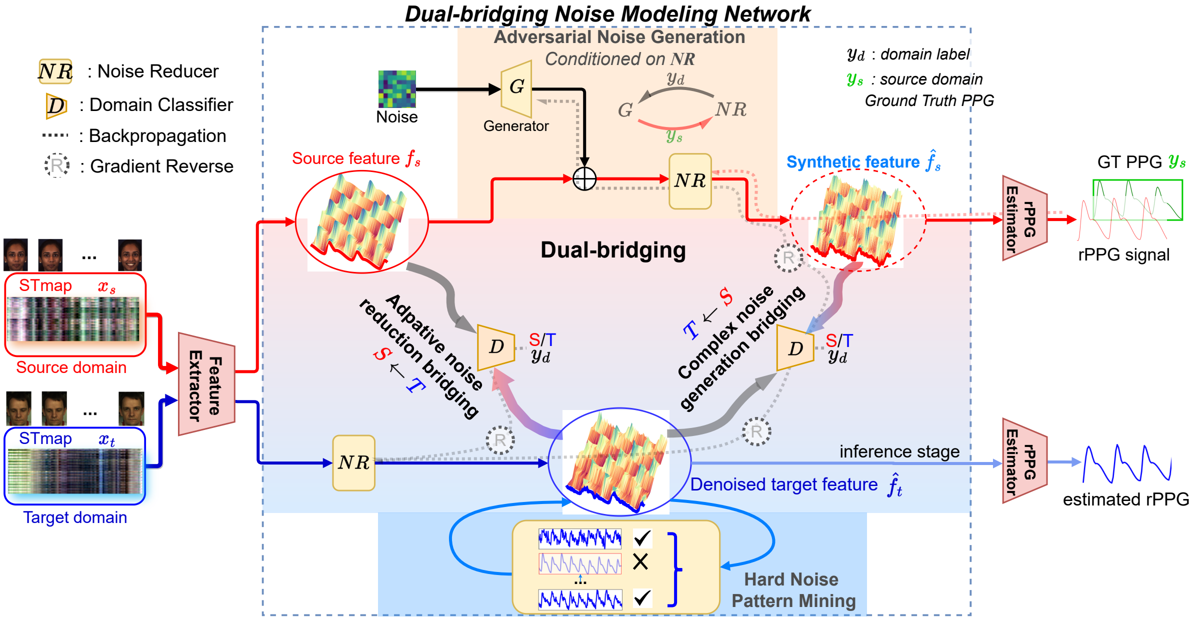


Figure 2. Framework of the dual-bridging noise modeling network. The dual-bridging can reduce the domain discrepancy in the feature space by training the noise reducer with both high-level guidance from domain classification and detailed waveform guidance from synthetic features regression. The adversarial noise generation forces the generator G to compete with the noise reducer NR and synthesize noisy features with unsolved noise in the target domain. The hard noise pattern mining removes high-quality temporal features and encourages the generator to further explore hard noise patterns in the target domain.

fore, we propose the dual-bridging noise modeling network to take advantage of both domain labels and the source domain ground truth to guide the rPPG feature alignment.

3. Proposed Method

The overview of the proposed dual-bridging noise modeling network is shown in Figure 2. Given the input face video from the source and target domain, we construct the STMap (details are in section 4.2) as the input of the feature extractor and then perform dual-bridging noise modeling on the middle representation. The first bridging provides high-level guidance where the source feature f_j^s and the denoised target feature f_i^t are pulled together with the domain classifier. The synthesized target domain feature f_j^s is obtained from the f_i^t through the adversarial noise generation module. The second bridge aims to pull f_j^s to f_i^t so that the target domain noise can be overcome by NR with the guidance of PPG y_s (regression) The generator in the adversarial noise generation module is conditioned on the NR so that it keeps on overcoming the complex noise pattern that can hardly be solved in the first bridging. With the dual-bridging noise modeling that contains both high-level (first bridge) and detailed waveform guidance (second bridge), the NR can handle the target domain noise better. Finally, NR works with the feature extractor and rPPG estimator to do the rPPG estimation in the target domain.

3.1. Dual-bridging

With respect to the feature extractor Θ, we define source domain features as $F_s = \{f_j^s\}_{j=1}^M$ and target domain features as $F_t = \{f_i^t\}_{i=1}^N$, where $f_j^s = \Theta(x_j^s)$ and $f_i^t = \Theta(x_i^t)$. Since the rPPG patterns are similar between different participants, the pre-trained feature extractor can preserve enough physiological information in F_t and noise modeling is the key to noise reduction of the target domain features. Detailed information for guidance is necessary to overcome the complex noise in the target domain features.

To reduce the noise contained in F_t, we propose the adaptive noise reduction bridging to reduce the noise contained in target domain features supervised by the domain classifier. In this bridging, we design the noise reducer with the noise estimation function Φ in a residual way as follows:

$$\hat{f}_i^t = f_i^t - \Phi(f_i^t) \tag{1}$$

where $\hat{f}_i^t$ is a denoised physiological feature and we define $\hat{F}_t = \{\hat{f}_i^t\}_{i=1}^N$ as the denoised target domain. The direct estimation of high-quality physiological features could be hard, and we choose to estimate the noise and use the residual structure to reduce the noise in the target domain features. We build an adversarial training between the noise reducer Φ and the domain classifier D as follows:

$$\underset{D}{max}\,\underset{\Phi}{min}\, L_{NR} = \mathbb{E}[\|1 - D(f_j^s)\|_2] + \mathbb{E}[\|D(\hat{f}_i^t)\|_2] \tag{2}$$

With high-level guidance from domain classification, some simple noise patterns can be overcome, and less noisy features can be aligned. However, high-level guidance may not overcome complex noise patterns without detailed waveform information for guidance. And the brute-force feature alignment may even change the physiological information of rPPG features when complex noise patterns are encountered. The phenomenon can be reflected by our ablation study in section 4.3.

To involve detailed information for supervision, we propose the complex noise generation bridging to generate target domain noise and synthesize noisy features to guide the noise reducer. From the source domain direction, we estimate the noise patterns in the target domain by generating noise and synthesizing noisy features as follows:

$$\overline{f}_j^s = f_j^s + \Psi(n_k) \tag{3}$$

where $\overline{f}_j^s$ is a raw noisy feature, n_k is random noise, and Ψ is the generator. And we define the synthetic feature domain $\hat{F}_s = \{\hat{f}_j^s\}_{j=1}^M$ as follows:

$$\hat{f}_j^s = \overline{f}_j^s - \Phi(\overline{f}_j^s) \tag{4}$$

To control the range of noise generation, we build the adversarial training between the noise generator and the domain classifier with respect to the denoised target domain as follows:

$$\begin{aligned} \underset{D}{max}\,\underset{\Psi}{min}\, L_G = \mathbb{E}[\|1 - D(f_j^s)\|_2] + \mathbb{E}[\|D(\overline{f}_j^s)\|_2] \\ + \mathbb{E}[\|D(\hat{f}_i^t)\|_2] \end{aligned} \tag{5}$$

In this bridging, the domain classifier can guide the generator to learn target domain noise patterns and synthesize noisy features with preserved physiological information. Thus, the source domain PPG signals y_j^s can be utilized to guide the noise reducer to overcome generated complex noise patterns as follows:

$$L_{SYN} = 1 - r(E(\hat{f}_j^s), y_j^s) \tag{6}$$

E is the rPPG estimator, and r is the Pearson correlation function. With this bridging, the physiological information can be preserved in the noise reduction and the other bridging can align more target domain features under large domain discrepancy.

3.2. Adversarial noise generation

With the guidance from the domain classifier, part of the target domain noise patterns can be generated. However, in the adversarial training process, the generator may not fully explore target domain noise patterns. To comprehensively explore the target domain noise for rPPG feature alignment, we propose a novel adversarial noise generation. As shown in Figure 2, the generator is conditioned on the noise reducer and forced to explore the unsolved noise contained in the target domain features. The noise reducer is trained with the synthetic features and corresponding PPG signals to overcome the generated noise. The adversarial correlation between the generator and the noise reducer is shown as follows:

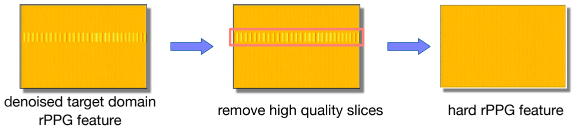

denoised target domain rPPG feature
remove high quality slices
hard rPPG feature

Figure 3. Illustration of hard noise pattern mining.

$$\underset{\Psi}{max}\,\underset{\Phi}{min}\, L_{ANG} = \mathbb{E}[\|1-D(\hat{f}_j^s)\|_2] + \mathbb{E}[1-r(E(\hat{f}_j^s), y_j^s)] \tag{7}$$

This adversarial training is based on two different optimization functions, but not the gradient reverse layer. Therefore, the novel noise patterns are not generated based on adversarial attacks but are searched from unsolved noise patterns in the target domain. With our adversarial noise generation, the dual-bridging structure can progressively generate target domain noise patterns, and the noise reduction ability can be significantly improved.

3.3. Hard noise pattern mining

To further improve the robustness of the noise reducer, we propose the hard noise pattern mining as illustrated in Figure 3. Considering the redundancy of rPPG estimation, features with one or several high-quality temporal features can contribute to accurate estimation. Thus, the noise reducer may just learn to overcome relatively simple noise patterns and some hard noise patterns may not be learned. To further explore the hard noise patterns, we need to reduce the shortcuts for noise reduction and generate hard noise patterns for training. In our hard noise pattern mining module, we identify high-quality temporal features with respect to the morphology consistency M which is defined as follows:

$$M = \max_{t=1}^{n} r(g_{0:T}, g_{t:t+T}) \tag{8}$$

r is the Pearson correlation, g is a temporal feature, t is the sliding start point, and T is the time length of morphology consistency evaluation. With respect to the morphology consistency, we define temporal features with significant periodicity as high-quality features. After identifying high-quality temporal features, we can remove the corresponding feature slice and preserve hard slices for noise generation. Correspondingly, more hard noise patterns will be generated and the noise reduction ability can be further improved for rPPG feature alignment.

The detailed optimization process of our proposed

Algorithm 1 The optimization process for our dual-bridging algorithm.

```
Input: Noise generator $\Psi$, noise reducer $\Phi$ and domain classifier $D$, total epochs $L$, mini-batch $B$, source domain samples $x_j^s$, source domain labels $y_j^s$, target domain samples $x_i^t$.
  for $l = 1$ to $L$ do
    for $b = 1$ to $B$ do
      Obtain $f_j^s$ and $f_i^t$ with pre-trained feature extractor from $x_j^s$ and $x_i^t$
      Obtain $\hat{f}_i^t$ with noise reducer
      Update noise reducer $\Phi$ and domain classifier $D$ with $L_{DE}$
      Generate $\overline{f}_j^s$ and $\hat{f}_j^s$ with updated noise reducer and noise generator
      Update noise generator $\Psi$ with $L_G$ and $L_R$
      Update noise reducer $\Phi$ with $L_{SYN}$
    end for
  end forOutput: Learned noise reducer $\Phi$.
```

method is shown in Algorithm 1. After training the dual-bridging algorithm, we can estimate rPPG signals in the target domain as follows:

$$\hat{y}_i^t = E(\Theta(x_i^t) - \Phi(\Theta(x_i^t))) \tag{9}$$

$\hat{y}_i^t$ is the estimated rPPG signal.

4. Experiments

We conduct experiments on three publicly available datasets with different types of interference to evaluate the accuracy and generalization ability of our dual-bridging noise modeling network. Three protocols that simulate the typical domain discrepancies in real application scenarios are adopted, *i.e.*, 1) task-independent evaluation, 2) participant-independent evaluation, and 3) cross-dataset evaluation. Mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (r) are adopted to evaluate the performance of average heart rate based on the estimated rPPG signal. The unit for MAE and RMSE is beats per minute (bpm).

4.1. Datasets and Baselines

We evaluate the proposed method on the following datasets, which cover different types/levels of head motion, facial expression, compression artifacts, skin color, and also heartbeat ranges.

PURE [36] dataset contains facial videos of 10 Caucasian participants (8 men, 2 women) in six different behavior modes with head motion and facial expression. The videos are uncompressed and captured at 30fps in 640 × 480.Synchronized GT-PPG signals are captured at 60 Hz using pulse oximeter pulox CMS50E.

MMSE-HR [39] datasets contains 102 facial videos captured from 40 participants of different ethnicities with different emotions. The emotion-related spontaneous facial expressions can also induce facial intensity variations. The video frames are captured at 25fps in 1040×1392 and compressed in JPG format, and their frame rate is 25Hz. GT blood volume signals are captured by the Biopac MP150.

UBFC-rPPG [1] dataset contains 42 facial videos captured under a human-computer interaction scene which causes head motions and larger heart rate range (63-145 bpm). Videos are recorded by the Logitech C920 Webcam at 30fps in 640 × 480. Synchronized GT-PPG signals are captured by CMS50E.

Baselines We compared the proposed method with traditional rPPG methods [5,16,32,39,40,41] and the latest deep rPPG models [3,4,13,21,22,23,24,27,28,29,34,37,45,47,49] to show the generalization ability. We also compared the proposed method with popular UDA methods [7,17] to show the effectiveness of rPPG feature alignment.

4.2. Implementation details

Our method is implemented with PyTorch and trained on an NVIDIA Tesla V100. We align the facial regions with four landmarks using the perspective transformation and generate the STmap [27] as input. We resample all STmaps and the corresponding labels at 30Hz. The frame number for each STmap is 300 and adjacent STmaps have 270 overlapped frames. In experiments, we pre-train the 2D CNN model on the source domain for rPPG estimation. The threshold of morphology consistency for hard noise pattern mining is 0.82. In the domain adaptation training stage, we set the batch size to 64 and adopt the Adam optimizer [11] for training. The learning rate for the noise pattern generator and noise reducer is 1e-4. And the learning rate for the domain classifier is 3e-5. The weight decay is 1e-4 for all optimizers. We train our domain adaptation algorithm for 200 epochs in the task-independent and PURE to MMSE-HR evaluations, and 2000 epochs in the MMSE-HR to PURE evaluation. We select the weights of the noise reducer for the test according to the estimation loss of synthesized samples. Under certain domain discrepancies, the details of the rPPG signals may not be preserved.

4.3. Task-independent evaluation

We first evaluate the proposed method in the task-independent scenario, which has a smaller domain discrepancy. The MMSE-HR collected facial videos under different activities and we regard each activity as a task. Within MMSE-HR, the unseen non-physiological intensity variations are mainly induced by different head motion and facial expression changes In this evaluation, we set one task as the target domain and the other tasks as the source domain. From the experimental results in Table 1, we can find that even under the intra-dataset scenario, the domain

Table 1. Experimental results for task-independent evaluation on MMSE-HR.

Method	MAE	RMSE	r
Li2014 [16]	-	19.95	0.38
CHROM [5]	-	13.97	0.55
Tulyakov2016 [39]	-	11.37	0.71
ST-Attention* [29]	-	10.10	0.64
RhythmNet [27]	-	5.03	0.86
CVD* [28]	-	6.04	0.84
PhysNet [47]	-	13.25	0.44
DeepPhys [3]	4.43	9.98	0.80
TS-CAN [21]	3.85	7.21	0.86
AutoHR [45]	-	5.87	0.89
BVPNet [4]	-	7.47	0.79
Federated2022 [23]	2.99	2.42	0.79
EfficientPhys-C [22]	2.91	5.43	0.92
EfficientPhys-T1 [22]	3.48	7.21	0.86
PhysFormer* [49]	2.84	5.36	0.92
ERM [12]	1.30	2.58	**0.99**
DANN [7]	1.24	2.71	**0.99**
CST [17]	1.20	2.42	**0.99**
Ours	**0.85**	**2.05**	**0.99**

* Trained on VIPL-HR datasets due to the large model-scale

Table 2. Experimental results for participant-independent evaluation on UBFC-rPPG dataset.

Method	MAE	RMSE	r
GREEN [40]	4.47	11.6	0.842
ICA [32]	3.51	8.64	0.908
CHROM [5]	3.44	4.61	0.968
POS [41]	2.44	6.61	0.936
CK [35]	2.29	3.80	0.981
Frédéric [2]	5.45	8.64	-
HeartTrack [31]	2.41	3.37	0.983
ETA-rPPGNet [10]	1.46	3.97	0.93
DAE [34]	1.48	2.49	0.97
PulseGAN [34]	1.19	2.10	0.98
Meta-rPPG [13]	5.97	7.42	0.53
CVD [28]	2.19	3.12	0.99
Gideon2021 [8]	3.6	4.6	0.95
Federated2022 [23]	2.00	4.38	0.93
Dual-GAN [24]	0.44	0.67	0.99
ContrastPhys [37]	0.64	1.00	0.99
ERM [12]	0.75	1.84	**0.99**
DANN [7]	0.58	1.19	**0.99**
CST [17]	0.41	1.04	**0.99**
Ours	**0.16**	**0.57**	**0.99**

discrepancy can decrease the performance of deep rPPG models. And under small domain discrepancies, common UDA methods are effective and can overcome certain noise patterns in the target domain. Our method can effectively reduce the unseen non-physiological intensity variations in feature space through noise reduction of rPPG features and achieves state-of-the-art performance.

4.4. Participant-independent evaluation.

The domain discrepancy can also come from differences between participants like head motions under human-computer interaction scenarios. Caucasian and Asian participants are included and we evaluate this participant-independent setting with the UBFC-rPPG dataset. We follow the protocol in [24,34] to select 30 participants for training and use the rest 12 participants for testing. As the experimental results shown in Table 2, the domain discrepancy is small with respect to deep rPPG models and pre-trained rPPG models also have low error. Compared to common UDA methods, our proposed method can provide more detailed information for noise reduction supervision and more noise patterns can be overcome in the feature alignment.

4.5. Cross-dataset evaluation

Our method is also evaluated in the cross-dataset setting with MMSE-HR and PURE. The selected datasets have differences in aspects such as head motion, From the view of the rPPG estimation, the MMSE-HR dataset contains more challenging scenarios (video compression, facial expression changes) than the PURE dataset. A challenging source domain could be beneficial to generalization ability and reduce the difficulty of the adaptation process. We provide details of the cross-dataset evaluations in the following paragraphs.

From MMSE-HR to PURE. In this experiment, we select the MMSE-HR dataset as the source domain and the PURE dataset as the target domain. Since the source domain is more challenging than the target domain , the adaptation process could be relatively easy. The experimental results are shown in Table 3. We can find that DANN cannot improve the generalization ability of rPPG models and only high-level guidance cannot align noisy features to high-quality features. While the cycle self-training strategy can provide certain details of the rPPG waveform and is more effective in reducing the domain discrepancy in rPPG estimation. Our method can progressively generate noise patterns in the target domain and synthesize noisy samples to provide detailed information from the source domain labels. Our method achieves state-of-the-art performance in this UDA scenario, and the low RMSE shows that our method is effective in reducing strong outliers. We provide some examples in Figure 3 (a) in which the noise reduction process can remove fake peaks in one period. After removing fake peaks, the periodicity of rPPG signal can be improved and the heart rate estimation can be more accurate.

From PURE to MMSE-HR. In this experiment, we select the PURE dataset that has fewer facial videos with mild

Table 3. Experimental results for cross-dataset evaluation between PURE and MMSE-HR datasets.

	MMSE-HR → PURE			PURE → MMSE-HR		
Method	MAE	RMSE	r	MAE	RMSE	r
CHROM [5]	3.25	12.92	0.84	5.72	12.69	0.58
POS [41]	2.83	12.49	0.85	4.98	13.11	0.53
CVD [28]	2.75	3.98	0.98	4.08	7.03	0.84
ERM [12]	2.49	8.48	0.93	2.59	5.44	0.96
DANN [7]	2.69	6.97	0.95	2.84	7.65	0.93
CST [17]	1.27	2.96	**0.99**	2.32	5.97	0.96
EfficientT1 [22]	-	-	-	3.04	5.91	0.92
PhysFormer [49]	-	-	-	2.84	5.36	0.92
Synthetic [25]	-	-	-	2.26	3.70	0.97
Ours	**1.10**	**1.67**	**0.99**	**1.71**	**3.72**	**0.98**

Table 4. Experimental results for cross-dataset evaluation from PURE to UBFC-rPPG.

	PURE → UBFC-rPPG		
Method	MAE	RMSE	r
CHROM [5]	3.10	6.84	0.93
POS [41]	3.52	6.84	0.90
DAE [34]	2.70	5.17	0.96
PulseGAN [34]	2.09	4.42	0.97
Siamese-rPPG [38]	1.29	8.73	-
Dual-GAN [24]	0.74	**1.02**	**0.99**
ERM [12]	1.50	3.36	0.96
DANN [7]	3.24	6.13	0.96
CST [17]	2.24	4.73	0.97
Ours	**0.58**	1.11	**0.99**

interferences (like slow head rotation, translation) as the source Compared to the MMSE→PURE setting, the domain gap in PURE→MMSE is harder to overcome since the noise patterns in the target domain are more challenging. In this experiment, our method is also compared with two rPPG methods [22, 49] which are pre-trained in the VIPL-HR dataset with more facial videos and scenarios. We also compare with [25] which uses synthetic facial videos and the AFRL dataset [6] for training. As shown in Table 3, in this challenging setting, we find that all baseline methods have lower performance compared with the other cross-dataset settings. Under this challenging setting, our dual-bridging with adversarial noise generation algorithm can effectively explore the specific noise patterns from the target domain. Additionally, by synthesizing noisy features similar to target domain features, the noise reduction ability can be dramatically improved to reduce the domain discrepancy. And our method can achieve state-of-the-art performance under this cross-dataset validation. We also provide some examples in Figure 3 (b) and find that noise reduction can improve the quality of rPPG signals. Especially some noisy rPPG signals show low periodicity, after the noise reduction, clear peaks can be observed. In this case, the intuitive UDA solution DANN also may not be able to learn domain-invariant representation with domain labels only for supervision. The cycle self-training strategy can only adaptively overcome a part of the domain discrepancy. Whereas, the EfficientT1, PhysFormer, and Synthetic methods pre-trained on larger datasets show better robustness.

From PURE to UBFC-rPPG. In this experiment, we select the PURE dataset with head motion as the source domain and the UBFC-rPPG dataset with compression and a wider range of heart rate as the target domain. The experimental results are shown in Table 4. The compression enlarges the domain discrepancy for rPPG estimation because some information is lost during compression and facial intensities are changed. Our proposed method can simulate the compression-induced noise and provide detailed information to effectively align rPPG features. In this setting, the pre-trained rPPG models in [34, 38] cannot handle the relatively larger domain gap (caused by the video compression) well. For common UDA methods, the high-level guidance from domain classification cannot fully guide the rPPG feature alignment and noise is still preserved in rPPG features.

In this cross-dataset evaluation, we show the domain adaptation process in rPPG estimation under various domain discrepancies. Experimental results show that adapting deep rPPG models to less challenging scenarios, common UDA methods are effective. However, to adapt deep rPPG models to more challenging scenarios, more detailed information about the rPPG waveform is desired to reduce the domain discrepancy. And our method can provide both noise patterns and waveform information with the synthesized samples for noise reduction training. We can find that the domain discrepancy has been reduced after noise reduction, as shown in Figure 4, especially for the harder cases PURE→MMSE-HR. Compared to collecting large-scale dataset to improve the generalization ability of deep rPPG models, our dual-bridging noise modeling network is effective and convenient.

Ablation study. We conduct ablation study with cross-dataset settings to show the effectiveness of each bridge in our method as shown in Table 5. We first show the experimental result of the pre-train model. The NR ↔ D represents that a noise reducer and a domain classifier are adversarially trained to reduce the domain discrepancy. Denoised target domain features are used to estimate rPPG signals. The noise reducer can adaptively overcome some noise patterns under a small domain discrepancy. When the domain discrepancy is large, the domain labels may not provide enough detailed information for training the noise reducer. To prevent the noise reducer from generating random phys-

Table 5. Experimental results for ablation study.

	MMSE-HR → PURE			PURE → MMSE-HR		
Method	MAE	RMSE	r	MAE	RMSE	r
Pre-trained rPPG model	2.96	9.57	0.91	2.09	3.92	**0.98**
+ NR ↔ D	2.13	7.14	0.95	20.27	26.34	0.58
+ NR ↔ D + G ↔ D	1.66	3.66	0.98	1.88	3.85	**0.98**
+ NR ↔ D + G ↔ D + G|NR ↔ D	1.13	1.71	**0.99**	**1.71**	**3.72**	**0.98**
+ NR ↔ D + G ↔ D + G|NR ↔ D + HM	**1.10**	**1.67**	**0.99**	1.88	4.16	**0.98**

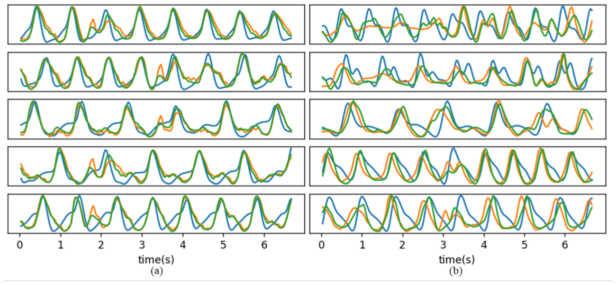


Figure 4. Visualization of rPPG estimation before (orange) and after (green) noise reduction compared with GT-PPG (blue). (a) contains examples of MMSE-HR→PURE evaluation and (b) is of PURE→MMSE-HR.

iological features under adversarial training, the strong regression supervision from rPPG waveform is necessary. For NR↔ D + G ↔ D, the noise generator is not conditioned on the noise reducer and can also synthesize noisy samples to train the noise reducer. The synthesized noisy samples prevent the noise reducer from generating random physiological features with detailed guidance from PPG signals. For NR ↔ D + G ↔ D + G|NR ↔ D, *i.e.*, dual-bridging with adversarial noise generation, the noise patterns in the target domain can be progressively generated to improve the denoising ability. The hard noise pattern mining can improve the robustness of the noise reducer when there is less interference in the target domain However, with strong interference in the target domain, too many hard noise patterns may distract the noise reducer from overcoming originally generated noise patterns and we saw a decrease of performance in the PURE → MMSE-HR.

4.6. Discussion

The generalization problem is a key challenge in applying rPPG techniques to real-world health monitoring applications. Especially for deep learning-based rPPG methods, the training data contain certain biases that may prevent rPPG models from learning domain-invariant physiological representations. Results show that generating noise patterns in the target domain and synthesizing samples for noise reduction can effectively reduce the domain discrepancy thereby improving the quality of rPPG signals.

We also found the effectiveness of using domain labels to help domain-invariant physiological representation learning when domain discrepancy is small. However, when domain discrepancy increases, the gain of feature alignment in such adversarial training tends to be smaller. Without target domain labels for noise reduction, generating noise patterns similar to the target domain is easier than directly denoising target domain features under a large domain discrepancy. In the noise generation procedure, some out-of-distribution noise patterns can also be generated and can further improve the generalization ability for unseen interferences.

Although our dual-bridging adaptation can work on most cases and achieve state-of-the-art, we still have limitations on handling hard cases with very large domain discrepancies (e.g., large head motion + facial expression + video compression in Table 4). Such rarely happened combined artifacts and can be regarded as a long-tailed distribution (of artifacts) problem. How to pay more attention to those rare but hard cases could be a key to solve it in the future.

5. Conclusion

Recent deep learning-based rPPG methods can learn to overcome non-physiological intensity variations from data, but may also involve bias induced by data and hinder the generalization ability. In this paper, we proposed a novel dual-bridging noise modeling network to reduce the domain discrepancy in rPPG estimation. With the proposed tri-adversarial optimization, the noise generator can progressively explore the noise patterns in the target domain and the coarse-to-fine guidance can effectively improve the noise reduction ability. In addition, comprehensive evaluations of different domain discrepancies show the effectiveness of our method. In future work, we may consider using unlabeled and synthesised [42] facial videos from multiple domains with various external interferences to learn domain-invariant physiological representations.

6. Acknowledgement

This project is partially supported by Hong Kong RGC General Research Fund RGC/HKBU12200820.

References

[1] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 5

[2] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9(20):4364, 2019. 6

[3] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *ECCV*, pages 349–365, 2018. 2, 5, 6

[4] Abhijit Das, Hao Lu, Hu Han, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation. In *FG*, 2021. 2, 5, 6

[5] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 5, 6, 7

[6] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 1462–1469. IEEE, 2014. 7

[7] 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. 2, 5, 6, 7

[8] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *ICCV*, 2021. 2, 6

[9] Guillaume Heusch, Andre Anjos, and Sebastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017. 2

[10] Min Hu, Fei Qian, Dong Guo, Xiaohua Wang, Lei He, and Fuji Ren. Eta-rppgnet: Effective time-domain attention network for remote heart rate measurement. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. 6

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[12] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011. 6, 7

[13] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *ECCV*, 2020. 5, 6

[14] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3918–3930, 2020. 2

[15] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 242–249. IEEE, 2018. 2

[16] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *CVPR*, 2014. 2, 5, 6

[17] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *NeuIPS*, 34:22968–22981, 2021. 2, 5, 6, 7

[18] Si-Qi Liu, Xiangyuan Lan, and Pong C Yuen. Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:2683–2696, 2021. 2

[19] Si-Qi Liu, Xiangyuan Lan, and Pong C Yuen. Learning temporal similarity of remote photoplethysmography for fast 3d mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security (TIFS)*, 17:3195–3210, 2022. 2

[20] Si-Qi Liu and Pong C Yuen. A general remote photoplethysmography estimator with spatiotemporal convolutional network. In *FG*, pages 481–488. IEEE, 2020. 2

[21] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *NeurIPS*, 33:19400–19411, 2020. 1, 2, 5, 6

[22] Xin Liu, Brian L Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement. *arXiv preprint arXiv:2110.04447*, 2021. 2, 5, 6, 7

[23] Xin Liu, Mingchuan Zhang, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Federated remote physiological measurement with imperfect data. In *CVPRW*, 2022. 2, 5, 6

[24] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *CVPR*, pages 12404–12413, 2021. 1, 2, 5, 6, 7

[25] Daniel Mcduff, Javier Hernandez, Xin Liu, Erroll Wood, and Tadas Baltrusaitis. Using high-fidelity avatars to advance camera-based cardiac pulse measurement. *IEEE Transactions on Biomedical Engineering*, 2022. 2, 7

[26] Daniel McDuff, Miah Wander, Xin Liu, Brian L Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. Scamps: Synthetics for camera measurement of physiological signals. *arXiv preprint arXiv:2206.04197*, 2022. 2

[27] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 1, 5, 6

[28] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *ECCV*, pages 295–310. Springer, 2020. 1, 2, 5, 6, 7

[29] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *FG*, pages 1–8. IEEE, 2019. 5, 6

[30] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *ICCV*, pages 4955–4964, 2021. 1, 2

[31] Olga Perepelkina, Mikhail Artemyev, Marina Churikova, and Mikhail Grinenko. Hearttrack: Convolutional neural network for remote video-based heart rate monitoring. In *CVPRW*, 2020. 6

[32] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, 2010. 2, 5, 6

[33] Paolo Russo, Fabio M Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *CVPR*, pages 8099–8108, 2018. 2

[34] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021. 1, 5, 6, 7

[35] Rencheng Song, Senle Zhang, Juan Cheng, Chang Li, and Xun Chen. New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method. *Computers in biology and medicine*, 116:103535, 2020. 6

[36] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 2, 5

[37] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *ECCV*, 2022. 2, 5, 6

[38] Yun-Yun Tsou, Yi-An Lee, Chiou-Ting Hsu, and Shang-Hung Chang. Siamese-rppg network: Remote photoplethysmography signal estimation from face videos. In *Proceedings of the 35th annual ACM symposium on applied computing*, pages 2066–2073, 2020. 7

[39] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *CVPR*, pages 2396–2404, 2016. 2, 5, 6

[40] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2, 5, 6

[41] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 2, 5, 6, 7

[42] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20587–20596, 2022. 2, 8

[43] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In *CVPR*, pages 16643–16653, 2021. 2

[44] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Toalign: Task-oriented alignment for unsupervised domain adaptation. *NeurIPS*, 34, 2021. 2

[45] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020. 1, 5, 6

[46] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 1

[47] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *BMVC*, 2019. 5, 6

[48] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *CVPR*, pages 151–160, 2019. 2

[49] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: facial video-based physiological measurement with temporal difference transformer. In *CVPR*, pages 4186–4196, 2022. 1, 2, 5, 6, 7