SFDA-rPPG: Source-Free Domain Adaptive Remote Physiological Measurement With Spatiotemporal Consistency

Yiping Xie, Zitong Yu, *Senior Member, IEEE*, Bingjie Wu, Weicheng Xie, *Member, IEEE*, and Linlin Shen, *Senior Member, IEEE*

***Abstract*—Remote photoplethysmography (rPPG) is a noncontact method that uses facial video to predict changes in blood volume, enabling physiological metrics. Traditional rPPG models often struggle with poor generalization capacity in unseen domains. Current solutions to this problem are to improve its generalization in the target domain through domain generalization (DG) or domain adaptation (DA) techniques. However, traditional DA methods usually require access to both source-domain data and target-domain data, which cannot be implemented in scenarios with limited access to source data due to privacy issues. In this article, we propose the first source-free DA benchmark for rPPG (SFDA-rPPG) measurement, which overcomes these limitations by enabling effective DA without access to source-domain data. Our framework incorporates a three-branch spatiotemporal consistency network (TSTC-Net) to enhance feature consistency across domains. Furthermore, we propose a new rPPG distribution alignment loss based on the frequency-domain Wasserstein distance (FWD), which leverages optimal transport to align power spectrum distributions across domains effectively and further enforces the alignment of the three branches. Extensive cross-domain experiments and ablation studies demonstrate the effectiveness of our proposed method in source-free DA (SFDA) settings. Our findings highlight the significant contribution of the proposed FWD loss for distributional alignment, providing a valuable reference for future research and applications. The codes are available at https://github.com/XieYiping66/SFDA-rPPG**

Received 2 April 2025; revised 29 August 2025; accepted 8 October 2025. Date of publication 20 November 2025; date of current version 4 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62276170, Grant 62306061, Grant 62576076, and Grant 62576216; in part by the Open Fund of National Engineering Laboratory for Big Data System Computing Technology under Grant SZU-BDSC-OF2024-02; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515140037; and in part by Guangdong Provincial Key Laboratory under Grant 2023B1212060076. The Associate Editor coordinating the review process was Dr. Xun Chen. *(Corresponding authors: Zitong Yu; Weicheng Xie.)*

Yiping Xie is with the School of Computer Science and Software Engineering and Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China, and also with the School of Computing and Information Technology, Great Bay University, Dongguan 523000, China.

Zitong Yu is with the School of Computing and Information Technology, Great Bay University, Dongguan 523000, China, and also with Guangdong Provincial Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China (e-mail: zitong.yu@ieee.org).

Bingjie Wu is with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632.

Weicheng Xie and Linlin Shen are with the School of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: wcxie@szu.edu.cn).

Digital Object Identifier 10.1109/TIM.2025.3635333

***Index Terms*—Consistency learning, remote heart rate measurement, remote photoplethysmography (rPPG), Wasserstein distance (WD).**

I. Introduction

THROUGH the change of blood volume in the optical information of facial video, remote photoplethysmography (rPPG) technology can estimate physiological metrics such as heart rate [1], [2], [3], blood pressure [4], and so on. Early methods estimate the rPPG signal by subtle color changes [5], [6], [7], [8], [9], [10], [11] in the face region, which are extracted from frames by using face detection methods such as the MTCNN [12]. However, these methods need to manually set the region of interest and some filtering operations and are easily influenced by factors that have an impact on skin color, such as illumination changes, and lack robustness.

With the advancement of deep learning, computer vision techniques have demonstrated remarkable versatility in healthcare applications. In dental imaging, for instance, methods for 3-D tooth instance segmentation [13] and video-based orthodontic treatment monitoring [14], [15] address similar challenges in medical data processing. Similarly, deep learning-based rPPG methods have shown the potential to mitigate the impact of nonphysiological factors [16], [17], [18], [19], [20], [21]. Notably, approaches leveraging 2D-CNNs [21] and 3D-CNNs [20] have been employed to optimize training and capture richer rPPG representations. Besides, Sun et al. proposed an unsupervised contrastive learning framework using the spatiotemporal similarity of rPPG to train the model to learn richer rPPG knowledge without access to labels. However, they may fail when dealing with unseen domains or encountering different domains with domain gaps, which are common in real application scenarios, such as domain gaps caused by different datasets collection methods. How to achieve better cross-domain performance is one of the current hot topics in rPPG research.

Currently, the domain adaptation (DA) [22] and the domain generalization (DG) [24], [25] technologies are helpful to improve the domain gap problem caused by the change of nonphysiological factors. However, due to the unclear example-specific differences, this solution may not work well directly for rPPG measurement [26]. While many DA methods rely on clustering to create pseudo-labels, this approach is

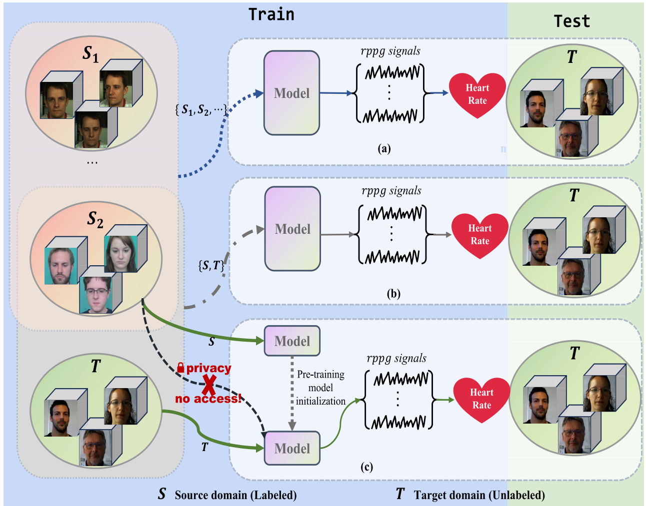

Fig. 1. Overview of cross-domain rPPG measurement methods. (a) Multisource DG [22] utilizes labeled data from multiple source domains. (b) Unsupervised DA [23] leverages labeled data from both the source and target domains. (c) SFDA utilizes a pretrained model from the source domain and unlabeled data from the target domain.

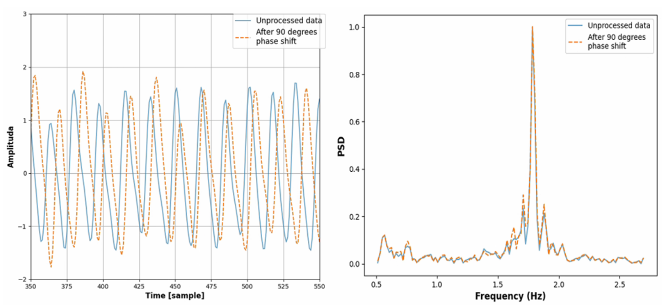

Fig. 2. Comparison of time-domain and frequency-domain stability under 90° phase shift. Time-domain signals showing significant visual differences (left). Frequency-domain PSD showing preserved spectral characteristics (right).

less effective for rPPG signals since heart rate data cannot be easily clustered into distinct categories. In the recent research of rPPG, Chung et al. [27] improved rPPG estimation in the unseen domain by using the feature learning framework of domain replacement and domain expansion. Du et al. [23] aligned the middle domain with a dual-bridge network and synthesized the target noise in the source domain to reduce the domain difference. However, these methods all need to use source-domain data, leading to privacy problems. That is, both source-domain and target-domain data are used for training. But is there a way to adapt a model on a target domain based solely on previous knowledge, that is, a model trained on the source domain, without accessing the source-domain data during adaptation? This approach can effectively protect the privacy of source-domain data, since only the model learned in the source domain needs to be used during adaptation. Therefore, we introduce the source-free DA (SFDA) paradigm in this article.

As shown in Fig. 1, we illustrate the essential differences between different cross-domain methods. DG-based rPPG methods [22] seek to enhance model generalizability by leveraging information from multiple source domains, as depicted in Fig. 1(a). The DG method uses multiple source-domain data for learning at the same time to improve the generalization of the model. However, it does not use the data of the target domain, but only learns the general knowledge in different fields. Traditional unsupervised DA-based rPPG aims to minimize the discrepancy between source and target domains by examining relationships between their respective datasets, as illustrated in Fig. 1(b). Ordinary DA methods use both source-domain data and target-domain data and cannot protect the privacy of either domain's data. These methods inherently utilize source-domain data during training, raising privacy concerns for facial data in rPPG applications. Conversely, our approach employs source-free domain adaptation (SFDA) based methods, as depicted in Fig. 1(c), which relies solely on a pretrained model from source-domain data and incorporates the target domain's unlabeled data for DA. At the same time, our method adapts the knowledge learned by the model trained on the source-domain data to another unlabeled domain, which is the same as the idea of people gradually learning in the real world, such as using mathematical knowledge to solve physics.

Our SFDA benchmark for the rPPG (SFDA-rPPG) framework integrates a three-branch spatiotemporal consistency network (TSTC-Net) to improve feature consistency across different domains. In addition, we propose a globally distributed alignment loss that is more appropriate for the rPPG task. Previous work [28] used Wasserstein distance (WD) loss in the time domain to analyze PPG peaks. However, time-domain rPPG signals often contain noise that can be better identified and separated in the frequency domain. Therefore, theoretically, frequency-domain Wasserstein distance (FWD) focuses more on the power spectral density (PSD) distribution across different frequency bands, and these distributions are more stable compared to rPPG signals in the time domain. As demonstrated in Fig. 2, phase-shifted signals maintain consistent spectral characteristics in the frequency domain while showing significant variations in the time domain, confirming the superior stability of frequency-domain representations for rPPG analysis. Thus, we design the FWD loss for effective rPPG supervision across different domains. To verify its effectiveness, we also compare the performance of time-domain WD and frequency-domain WD in Section IV-G4.

Our main contributions are summarized as follows.

1) We propose a novel SFDA-rPPG framework, incorporating a TSTC-Net to enhance rPPG measurement across domains. To the best of our knowledge, this is the first application of an SFDA approach in the rPPG field, ensuring both robust performance and improved privacy and security.
2) We propose a globally aligned FWD loss for rPPG distribution learning, which captures more stable PSD distributions across frequency bands.
3) We conduct comprehensive cross-dataset experiments to demonstrate the exceptional performance of SFDA-rPPG along with the efficacy of FWD loss.

II. Related Work

A. RPPG Measurement

As the rPPG signal change is very subtle, it is difficult to directly extract from videos. Some traditional methods

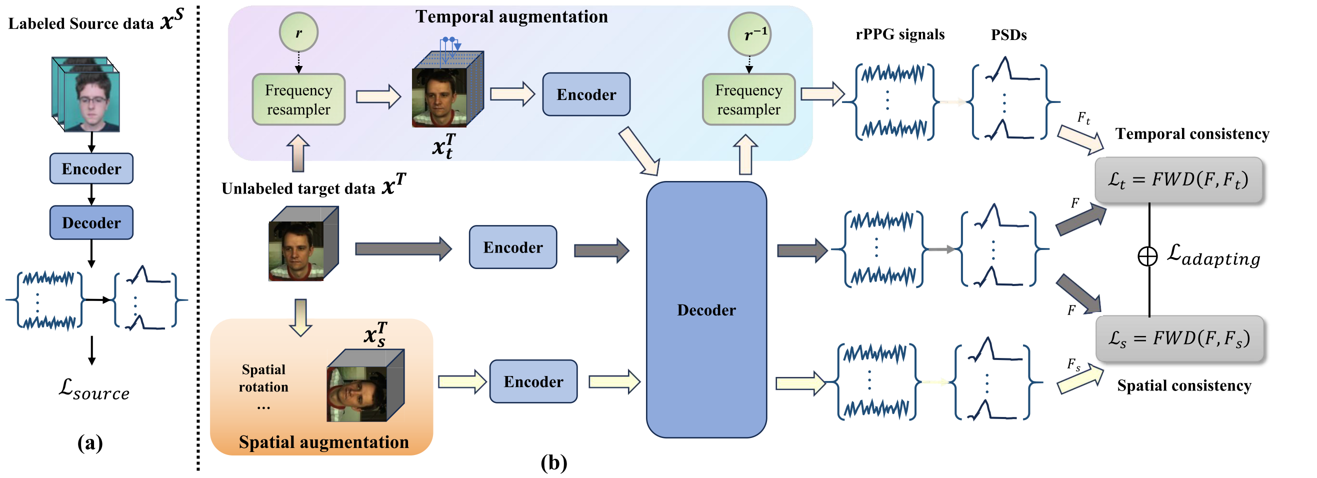

Fig. 3. Overview of SFDA-rPPG: A two-stage framework consisting of (a) supervised pretraining stage and (b) SFDA stage. First, the pretraining stage involves pretraining the model with labeled source-domain data. Second, the adaptive stage utilizes a spatiotemporal branching structure for consistency learning, enabling the source model to adapt more effectively to the unlabeled target-domain data. The encoder and decoder are disentangled from PhysNet [20] to facilitate improved representation learning.

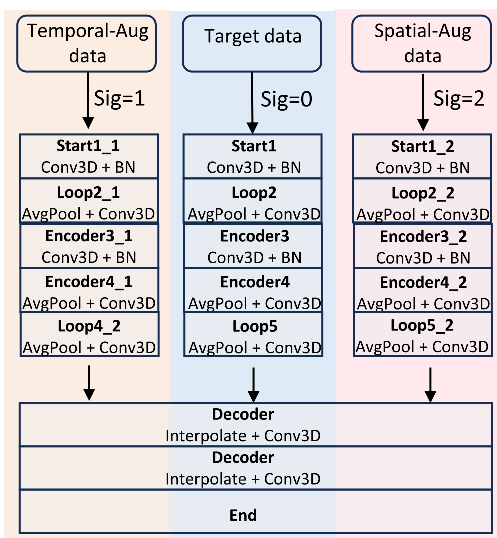

Fig. 4. Architecture overview of the SFDA-rPPG model. The framework consists of three parallel encoder branches (sig = 0, 1, 2) processing target data, temporal-augmented data, and spatial-augmented data, respectively, followed by a shared decoder for rPPG signal reconstruction. The encoder and the decoder are disentangled from PhysNet [20].

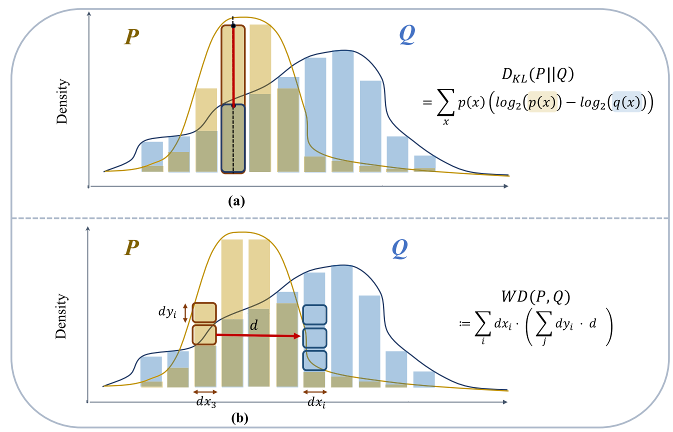

Fig. 5. Distribution alignment comparison diagram of (a) KL divergence and (b) WD. P and Q denote two distinct distributions, representing the distribution of two rPPG's power spectral densities after undergoing Softmax transformation. (a) Vertical probability dependence of KL divergence, where $p(x)$ and $q(x)$ represent the densities in the distributions P and Q. (b) Global advantages of the lateral optimal transmission of WD, where d represents the distance to be moved in the adaptation process.

based on face and mathematical model analysis have been proposed. For example, Balakrishnan et al. [6] used ambient light to recover rPPG from the facial skin area for the first time. Then, PCA [6], ICA [7], and other methods are used to decompose the original time signal to obtain the rPPG signal, and the adaptive matrix [8] is used to reduce noise and capture consistent clues related to rPPG. Furthermore, color subspace transformation methods such as CHROM [9] based on chromaticity and POS [10] based on orthogonal projection plane of skin color are used to generate HR waveform. Later, Park et al. [11] proposed using a skin boundary filter and mask filter to improve the rPPG performance.

However, the illumination conditions and motions are complicated, making manual traditional methods difficult to implement. Therefore, many deep learning methods have been proposed, including supervised and unsupervised methods. For supervised methods, Spetlík et al. [21] measured rPPG signals through 2D-CNN for the first time and conducted end-to-end training through alternating optimization. And Yu et al. proposed an end-to-end spatiotemporal network (PhysNet [20]), which considers the temporal context in facial video, and the first Transformer framework (TranRPPG [19]), which enhances automatic rPPG feature representation on the 3-D mask face task. For unsupervised methods, the temporal and spatial characteristics of rPPG are usually used for learning. For example, the contrastive learning framework [29] proposed by Sun and Li explored the spatial and temporal similarities of rPPG. A 3DCNN model is used to generate multiple rPPG signals from each video at different spatiotemporal locations, and the rPPG signals from the same video are pulled together, while the rPPG signals from different videos are pushed apart.

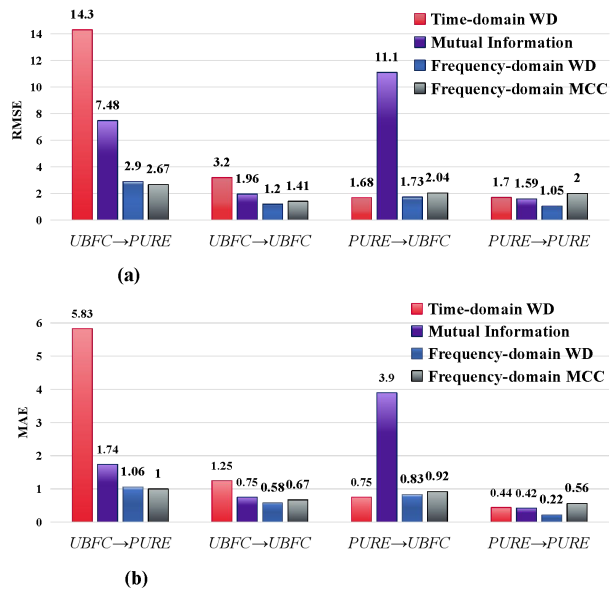

Fig. 6. Results of models supervised with time-domain and frequency-domain losses on PhysNet [20]. Datasets before the "arrow" refer to the training datasets, while those after the "arrow" refer to the testing datasets. For example, UBFC→PURE means training on UBFC and testing on PURE. (a) RMSE comparison between different losses. (b) MAE comparison between different losses.

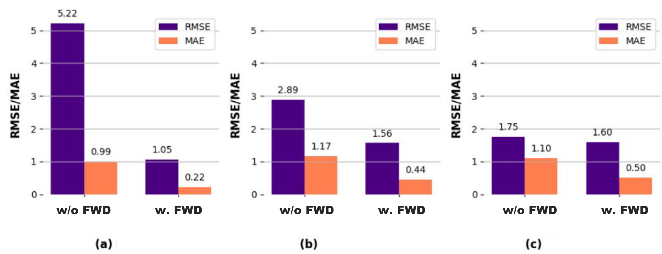

Fig. 7. Generalization evaluation of FWD loss on other backbones.

These methods can achieve acceptable performance on a single dataset.

Nowadays, researchers are increasingly focusing on cross-dataset evaluations across different datasets. They refer to data from different datasets as different domains because different rPPG data have different lighting conditions and other realistic influencing factors. In the recent research of rPPG, Chung et al. [27] improved rPPG estimation in the unseen domain by using the feature learning framework of domain replacement and domain expansion. Du et al. [23] aligned the middle domain with a dual-bridge network and synthesized the target noise in the source domain to reduce the domain difference. Despite improved performance in cross-domain situations, the performance is still far from satisfactory.

B. Source-Free DA

In terms of SFDA, research has made significant progress in recent years. The goal of SFDA is to adapt a model trained on the source domain by using only the target-domain data without accessing the source-domain data. This has important implications in terms of privacy protection and data transmission costs. The original SFDA methods mostly focused on adjusting model parameters through pseudo-label generation of target-domain data. For example, Liang et al. [30] proposed a passive domain transfer method that uses pseudo-labels for DA to improve adaptability by maximizing the entropy of target-domain data. Another direction is to perform passive domain transfer through generative adversarial networks (GANs). Kundu et al. [31] proposed a GAN-based SFDA method to improve the domain adaptability of the model through adversarial generation and training. In the video field, Li et al. [32] proposed a source-free video DA method that utilizes space–time–history consistency learning. This method effectively utilizes the spatiotemporal consistency of the target-domain video data and significantly improves adaptive performance on video data. Furthermore, recent research has also explored implementing SFDA through feature alignment. For example, Xia et al. [33] proposed a source-free domain transfer method for feature alignment by maximizing the consistency of source and target-domain features, which demonstrated excellent performance on multiple benchmark datasets. They all achieve effective adaptation to the target domain without the need for source-domain data through different technical means. This provides new solutions for scenarios that face data privacy or data acquisition difficulties in practical applications, which inspires us to explore the source-free domain adaptive rPPG measurement.

III. Methodology

Inspired by the work [32], we introduce the spatiotemporal consistency network for SFDA-rPPG. As illustrated in Fig. 3, this network leverages the learned model weights H_w from the source-domain and the unlabeled target data D_T to facilitate DA. Central to our method is the master encoder E trained within the model, alongside the temporal and spatial branch encoders E_t and E_s, and the decoder D. Overall, our approach unfolds in two main stages.

1) During the source-domain pretraining phase, the labeled source-domain data guides the model training to acquire the model weights H_w, where $H_w = \mathcal{D}_w \circ E_w$ with $\mathcal{D}_w$ and E_w representing the pretrained encoder and decoder weights of the model.
2) Subsequently, in the target DA phase, we adopt a TSTC-Net. Initially, the pretrained model's encoders and decoders serve as starting parameters, which are then fine-tuned on the target data to train the target encoders E, E_s, E_t, and decoder $\mathcal{D}$. We also provide the detailed algorithm of the adaptation stage in Algorithm 1.

A. Notations

Similar to generic DA/generalization tasks, we employ $D_S = \{(X_i^S, Y_i^S)\}_{i=1}^{n_S}$ to define the source-domain dataset, where n_S signifies the number of labeled source-domain data. Here, $X_i^S \in \mathbb{R}^{t\times w\times h\times 3}$ represents a video sample, and the label $Y_i^S = [y_1, y_2, \ldots, y_n]$, where $y_j \in \mathbb{R}$ embodies the rPPG signal. For the target domain, we let $D_T = {X_i^T}_{i=1}^{n_T}$ signify the unlabeled dataset, where n_T represents the number of samples and $X_i^T \in \mathbb{R}^{t\times w\times h\times 3}$.

Algorithm 1 SFDA-rPPG Adapting Process

Input: unlabeled target data $\{X_i^T\}_{i=1}^{n_T}$, pretrained model weights $H_w = \mathcal{D}_w \circ E_w$, and total epoch number e
Initialize encoders E, E_s, E_t with pretraining weights E_w, and initialize decoder $\mathcal{D}$ with $\mathcal{D}_w$;
for $i = 1$ to e **do**
    Get target mini-batch X_i^T.
    `# Spatial augmentation.`
    Get the spatial augmentation version X_s^T with Eq.(1).
    Calculate spatial prediction signals $f_s = \mathcal{D}(E_s(x_s^T))$ and its PSD F_s.
    `# Temporal augmentation.`
    Generate random number $r \in [0.66, 0.80]$ (refer to Fig. 8).
    Calculate the temporal prediction signals f_t by frequency resampler with Eq.(2).
    Transform f_t into frequency domain to get its PSD F_t.
    `# Obtain original data.`
    Calculate the original data prediction signals $f = \mathcal{D}(E(x_i^T)$ and its PSD F.
    Calculate FWD consistency loss $\mathcal{L}_{adapting}$ according to Eq.(7).
    Update E,E_s,E_t,$\mathcal{D}$ with $\mathcal{L}_{adapting}$.
end for

B. Three-Branch Spatiotemporal Consistency Network

To encourage the model to acquire stable predictive capabilities from the spatial and temporal transformations, in target DA, we partition the data into three distinct streams: 1) unlabeled target data; 2) spatial augmentation data; and 3) temporal augmentation data. We posit that an effective model should exhibit robustness across identical samples, yielding consistent outcomes for both the original sample and its temporally and spatially augmented variants.

As illustrated in Fig. 4, our architecture employs three parallel encoder branches with identical structures but independent parameters. The target data branch (sig = 0) uses encoder E to learn comprehensive representations, while the spatial and temporal augmentation branches (sig = 1, sig = 2) use encoders E_s and E_t to, respectively, acquire spatial and temporal specific representations. To prevent the model from learning ambiguous feature representations, we consistently align the outputs of the temporal and spatial branches with those of the unaltered target data branch. Subsequently, a shared decoder is utilized to learn task-specific feature decoding, accepting input from any of the three encoder branches through a unified interface controlled by the sig parameter. During training, all three branches contribute to robust feature learning, while during inference, only the target data branch is utilized for computational efficiency.

1) Spatial Augmentation: Following [39], we randomly apply one of six distinct spatial transformations to the video sample frames. Specifically, we establish the augmentation method set AU_S, including six spatial transformations: image rotations (0°, 90°, 180°, and 270°), and horizontal and vertical flips. We randomly choose an augmentation from the set and implement it on the current video sample frame sequence denoted as $X_i^T = \{x_1, x_2, x_3, \ldots\}$, representing the ith sequence of video frames in the target dataset

$$X_s^T = \{\mathrm{AU}_{s,j}(x_1), \mathrm{AU}_{s,j}(x_2), \mathrm{AU}_{s,j}(x_3), \ldots\} \tag{1}$$

where $\mathrm{AU}_{s,j}$ is the jth spatial augmentation randomly selected in AU_s and j belongs to the interval [1, 6]. Then, we get the predicted blood volume pulse (BVP) signal through $f_s = \mathcal{D}(E_s(X_s^T))$ and its PSD F_s.

2) Temporal Augmentation: The method of directly masking off some frame sequences for time augmentation is not effective [40] in the field of rPPG. Instead, we employ the frequency resampler introduced in [41] to generate our temporal augmented samples. In a similar way, we randomly select the resampling factor r within the range of [0.66, 0.80]. The difference is that we do not need to generate negative sample pairs, so we utilize the resampling mechanisms twice in succession directly. Initially, the sample is resampled by the r resampling factor to generate a sample with a higher heart rate. Then, to generate the prediction signal of the temporal augmented samples with the required scale, the signal value obtained after model estimation directly uses the inverse of r as a new factor to resample the result back to the original scale. This process adheres to the following formula:

$$f_t = \mathrm{FR}_{r^{-1}}\left(\mathcal{D}\left(E_t\left(\mathrm{FR}_r\left(X_i^t\right)\right)\right)\right) \tag{2}$$

where FR_r is a frequency resampler with r as the resampling factor and $\mathrm{FR}_{r^{-1}}$ is a frequency resampler with the factor of the inverse of r. This specific method will directly obtain the rPPG signal f_t of the temporal augmentation branch, and then its PSD F_t can be obtained.

C. FWD Loss

In [42], Kullback–Leibler (KL) divergence [43] is used as the label distribution loss to align the output signal with the Gaussian distribution of heart rate label. Nonetheless, as depicted in Fig. 5(a), KL divergence primarily emphasizes point-to-point probability deviation, overlooking the structural information encompassing the entire distribution. Conversely, the WD [44] offers a more comprehensive measure, addressing not only the deviation between individual probabilities, but also capturing differences across the overall distributions, as depicted in Fig. 5(b). Thus, this method can more accurately delineate the actual disparity among the frequency general density distributions associated with rPPG signals.

The WD, also known as the Earth mover's distance [45], provides a principled approach to measure the distributional discrepancy between probability measures. Rooted in optimal transport theory, it offers a geometrically meaningful metric for comparing probability distributions by measuring the minimum cost required to transform one distribution into another. Unlike other divergence measures, such as KL divergence, the WD respects the underlying metric structure of the space and provides a true distance metric. For two continuous probability distributions P and Q, WD can be defined as

$$\mathrm{WD}(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \int_{X \times Y} |x - y| \, d\gamma(x, y) \tag{3}$$

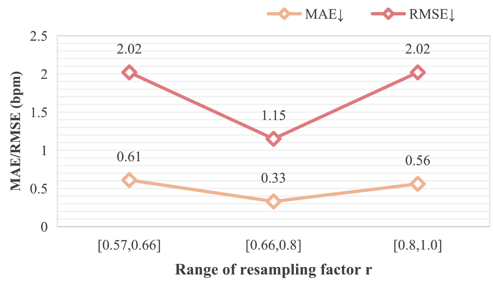

Fig. 8. Sensitivity analysis of resampling factor r from UBFC-rPPG [51] to PURE [50].

where $\Pi(P, Q)$ represents the set of all possible transmission plans to convert P to Q, γ is one of the specific transmission plans, $|x - y|$ represents the cost of moving from point x to point y in space, and the inf represents the minimum value taken out of all possible transmission plans.

For rPPG signals, time-domain WD [28] calculation can be performed directly to leverage the timing of systolic peaks from contact PPG as labels for model training. However, rPPG signals in the time domain contain too much noise interference, while frequency domain analysis can more easily remove some noise interference by filtering and eliminate the interference of phase information. Specifically, we compare the results of time-domain WD and frequency-domain WD in Fig. 6. So, we can get the PSD of rPPG signals to calculate a more robust alignment loss, and we refer to it as the FWD loss. We can express the PSDs of two rPPG signals as P and Q. For these point set distributions composed of real numbers, the transmission plan can also be expressed as

$$\mathrm{FWD}(P, Q) = \min_{T\in\Pi} \sum_{i,j} T_{ij} \cdot d\left(p_i, q_j\right) \tag{4}$$

where T is a transportation plan matrix, where $T_{i,j}$ represents the probability quality of transportation from point p_i to q_j. $d(p_i, q_j)$ is the distance between p_i and q_j. In the case of a 1-D signal, we simplify the distance to absolute distance. Π is the set of all possible transportation plans that satisfy the marginal distribution constraints from P to Q.

Specifically, we use the cumulative distribution function cdf [49] to find the effort needed for this 1-D most suitable transportation plan. Although there may be numerous optimal plans, there exists only one optimal result. A fundamental property of 1-D optimal transport is that the optimal transport plan has a monotonic structure: probability mass is transported from lower to higher quantiles in order. This means that for discrete distributions, we can directly compute the transport cost by comparing the cumulative distributions at corresponding frequency bins, since the optimal plan transports mass between points with the same cumulative probability. This property allows us to avoid solving the complex optimization problem in 4 and instead use the equivalent cdf-based computation. Therefore, the transmission FWD can be expressed as

$$\mathrm{FWD}(P, Q) = \sum_i \left|\mathrm{cdf}_i(P) - \mathrm{cdf}_i(Q)\right| \tag{5}$$

where cdf_i represents the ith value after distribution accumulation. The comparative results between FWD and other losses are presented in Table I and Fig. 6. Additionally, we demonstrate its applicability across different frameworks in Fig. 7.

D. Overall Training and Prediction Procedure

In the pretraining stage, we adopt a supervised learning method. After the source-domain data passes through PhysNet [20], the predicted rPPG signal f_{source} as well as its PSD F_{source} can be obtained. Finally, we use FWD loss to align it with the PSD F_{label} of the label

$$\mathcal{L}_{\text{source}} = \mathrm{FWD}\left(\mathrm{Softmax}\left(F_{\text{source}}\right), \mathrm{Softmax}\left(F_{\text{label}}\right)\right). \tag{6}$$

We use $\mathcal{L}_{\text{source}}$ to get the pretraining model weight H_w in the source domain, and initialize the encoders E, E_s, E_t, and decoder $\mathcal{D}$.

In the adaptive stage, we adopt a three-branch structure. The middle branch directly uses the original data to get the rPPG signal $f = \mathcal{D}(E(X_i^T))$ through the model, and then gets its PSD F. We use FWD loss to adapt our model

$$\begin{aligned}\mathcal{L}_{\text{adapting}} = {} & \mathrm{FWD}\left(\mathrm{Softmax}\left(F\right), \mathrm{Softmax}\left(F_s\right)\right) \\ & + \mathrm{FWD}\left(\mathrm{Softmax}\left(F\right), \mathrm{Softmax}\left(F_t\right)\right).\end{aligned} \tag{7}$$

During inference, we do not use all three branches simultaneously. We only use the middle branch, as it is the core branch containing the original dataset, which includes the knowledge obtained through consistency learning with the two augmented branches. So, we get the predicted rPPG signal through $f_s = \mathcal{D}(E_s(X_s^T))$.

IV. Experiment

The proposed Semi-rPPG method has been evaluated on four public datasets of PURE [50], UBFC-rPPG [51], and COHFACE [52]. We test the intradomain and cross-domain performance between UBFC-rPPG and PURE. In addition, we test the performance of UBFC-rPPG and PURE to COHFACE because COHFACE is a more realistic and complex scenario than UBFC-rPPG and PURE. Besides, we evaluate the applicability and effectiveness of our FWD and TSTC-Net on three different backbones.

A. Datasets and Metrics

1) Datasets: The proposed SFDA-rPPG method has been evaluated on three public datasets of PURE [50], UBFC-rPPG [51], COHFACE [52], and MMPD [53].

PURE [50] dataset consists of ten subjects, and each subject was recorded with a 1-min video under six scenarios. The videos were captured at a frame rate of 30 Hz with a cropped resolution of 640 × 480 pixels.

UBFC-rPPG [51] dataset has 42 videos from 42 subjects. The video was captured at 30fps with a resolution of 640 × 480 in uncompressed RGB format.

COHFACE [52] has 160 1-min videos of 40 subjects. The heart rate and breathing rate of the recorded subjects are

TABLE I

COMPARISON OF LOSS ABLATION ON PHYSNET [20]. WE USE NEGATIVE PEARSON LOSS [54] BY DEFAULT, COMBINED WITH THE LOSS MARKED UNDER THE METHOD. THE RESULTS SHOWN IN LIGHT GRAY BACKGROUND REPRESENT THE CROSS-DOMAIN TESTING BETWEEN UBFC AND PURE, WHERE OTHERS ARE INTRADOMAIN TESTING. BEST RESULTS ARE MARKED IN BOLD

Method			Train data	Test data					
				UBFC			PURE		
Frequency	KL	FWD		MAE	RMSE	R	MAE	RMSE	R
✓	✗	✗	UBFC	0.83	1.53	0.99	1.50	3.87	0.96
			PURE	3.33	12.33	0.46	0.56	1.76	0.99
✗	✓	✗	UBFC	0.75	1.47	0.99	4.22	12.86	0.54
			PURE	0.92	1.78	0.99	0.61	2.03	0.98
✗	✗	✓	UBFC	**0.58**	**1.22**	**0.99**	**1.06**	**2.93**	**0.97**
			PURE	**0.83**	**1.73**	**0.99**	**0.22**	**1.05**	**0.99**

synchronized with the videos. The videos have been recorded at a resolution of 640×480 pixels and a frame rate of 20 Hz.

MMPD [53] dataset consists of 660 1-min videos with a resolution of 320×240 and a frame rate of 30 Hz from 33 subjects with diverse skin types and activities under four lighting conditions. MMPD provides a compressed version dataset named miniMMPD. We use the mini-MMPD version in our experiments.

2) Metrics: Following the methodology outlined in [20] and [29] in the field, this study employs mean absolute error (MAE), root-mean-square error (RMSE), and Pearson correlation coefficient (R) as evaluation metrics. Notably, for MAE and RMSE, lower values indicate reduced error margins, whereas for R, values approaching 1.0 signify diminished error. Among them, both MAE and RMSE are measured in terms of beats per minute (bpm). And for convenience, these units will not be repeated in the subsequent tables or analyses.

B. Experimental Setup

In this investigation, the MTCNN algorithm [12] is employed for detecting facial regions and subsequent cropping within video frames. During the pretraining process, face clips represented as $3 \times 300 \times 128 \times 128$ (channels $\times$ time $\times$ height $\times$ width) matrices are utilized to pretrain the model and acquire initial weights. Following this, the pretrained architecture is disassembled to initialize encoders and decoders specific to the target domain. For the primary training phase, inputs comprised $3 \times 210 \times 128 \times 128$ (channels $\times$ time $\times$ height $\times$ width) face clip matrices. Training is conducted utilizing the AdamW optimizer [46] across both pretraining and training stages, across 30 epochs on four NVIDIA TITAN X GPUs, utilizing a learning rate of e-5. In the evaluation phase, following the protocol described in [29], each video is segmented into 30-s intervals with heart rate measurements obtained for each segment.

C. Cross-Dataset Testing Between UBFC and PURE

As shown in Table II, we conduct cross-domain experiments using the UBFC and PURE datasets to assess the cross-domain performance of our method. Specifically, before conducting evaluations on both datasets, we separately pretrain our model on one of them with a label and then adapt it to another dataset without a label. Additionally, we train the SHOT [30] and CoWA [38] models of SFDA for cross-domain performance assessment. We also show the cross-domain performance after training on the source domain on traditional supervised approaches [38]. From the results, our method is the best in cross-domain. Under the training of the unlabeled target domain, the MAE of the PURE dataset can reach 0.33, which is very close to the MAE of supervised PURE learning under PhysNet. The results demonstrate that our approach outperforms most of the methods, displaying superior cross-domain performance than the traditional SFDA clustering method.

D. Results From PURE/UBFC to COHFACE

We utilize the UBFC and PURE datasets and attempt to adapt them to the more complex COHFACE dataset, which comprises facial videos captured under wild lighting conditions. We compare the performance of other source-free domain adaptation methods [30], [38] of cross-domaining and the results of direct cross-domaining in supervised ways. As shown in Table III, our findings reveal that although the performance of transferring from both datasets to COHFACE is suboptimal, it yields reduced errors compared to prior methods and achieves some enhancements. However, due to the complex collection environment of COHFACE, it is difficult to learn better knowledge from datasets like UBFC and PURE, which are collected in experimental environments.

E. One-Shot Evaluations From PURE/UBFC to MMPD

We further evaluate our method's adaptability to challenging real-world conditions by examining cross-domain performance from controlled environments to more complex scenarios. Specifically, we utilize the UBFC and PURE datasets and attempt to adapt them to the more complex MMPD dataset, which comprises facial videos captured under wild lighting conditions. To rigorously test our approach, we designed a one-shot adaptation scenario where we randomly selected just one subject from the first 80% of the MMPD dataset for training, then evaluated performance on the remaining 20% of subjects. As shown in Table IV, our findings reveal that when adapting from UBFC to MMPD, our method achieves the best results with an MAE of 12.31 bpm, outperforming all baseline methods. For the PURE to MMPD adaptation, our approach remains competitive despite the challenging domain shift. These results demonstrate our method's capacity for generalization even with minimal target-domain data, though the varying performance between adaptation scenarios highlights how different source domains may transfer differently to more diverse target conditions.

F. One-Shot Evaluations Between UBFC and PURE

To ascertain the robust generalization capabilities under scenarios with limited proprietary data, we evaluate our method for one-shot training. Specifically, we employed all the data

TABLE II

CROSS-DATASET TESTING BETWEEN UBFC AND PURE. THE RESULTS SHOWN IN LIGHT GRAY BACKGROUND REPRESENT THE CROSS-DOMAIN TESTING BETWEEN UBFC AND PURE, WHERE OTHERS ARE INTRADOMAIN TESTING. BEST RESULTS ARE MARKED IN BOLD

Types	Method	Source data	Target data	Test data					
				UBFC			PURE		
				MAE	RMSE	R	MAE	RMSE	R
Traditional	CHROM [5]	UBFC	-	2.36	9.23	0.87	-	-	-
		PURE	-	3.44	4.61	0.97	0.75	2.23	0.99
	POS [10]	UBFC	-	2.11	9.11	0.87	-	-	-
		PURE	-	2.44	6.61	0.94	0.80	4.11	0.98
Supervised	PhysNet [20]	UBFC	-	0.75	1.47	0.99	3.81	-	0.87
		PURE	-	7.02	-	0.60	0.99	5.22	0.93
	TS-CAN+ [34]	UBFC	-	0.98	2.68	0.97	-	-	-
		PURE	-	-	-	-	1.80	3.45	0.99
	Dual-GAN [35]	UBFC	-	0.44	0.67	0.99	-	-	-
		PURE	-	0.74	1.02	0.99	0.82	1.31	0.99
	PulseGAN [36]	UBFC	-	1.19	2.10	0.98	-	-	-
		PURE	-	2.09	4.42	0.97	-	-	-
	Contrast-Phys+ (100%) [37]	UBFC	-	0.50	1.15	0.99	0.89	2.26	0.98
		PURE	-	0.61	2.02	1.98	0.44	1.49	0.99
Source-free DA	SHOT [30]	UBFC	PURE	0.58	1.22	0.99	1.17	2.89	0.97
		PURE	UBFC	0.58	1.22	0.99	0.83	2.60	0.98
	CoWA [38]	UBFC	PURE	0.75	1.47	0.99	1.22	2.98	0.97
		PURE	UBFC	0.83	1.53	0.99	0.50	1.73	0.99
	SFDA-rPPG (Ours)	UBFC	PURE	10.92	22.30	0.29	**0.33**	**1.15**	**0.99**
		PURE	UBFC	**0.41**	**1.08**	**0.99**	0.56	1.76	0.99

TABLE III

CROSS-DATASET TESTING FROM UBFC OR PURE TO COHFACE. BEST RESULTS ARE MARKED IN BOLD

Types	Method	UBFC → COHFACE			PURE → COHFACE		
		MAE	RMSE	R	MAE	RMSE	R
Supervised	EfficientPhys [46]	27.22	29.27	0.01	27.34	29.57	-
	PhysFormer [42]	21.27	26.84	0.03	17.74	20.52	-
	PhysNet [20]	12.62	19.94	**0.23**	13.96	20.36	0.002
SFDA	SHOT [30]	16.78	20.19	0.08	18.91	21.96	0.01
	CoWA [38]	23.91	27.63	-	15.03	18.43	0.08
	SFDA-rPPG (Ours)	**12.38**	**15.28**	0.04	**12.69**	**15.80**	**0.15**

TABLE IV

ONE-SHOT EVALUATIONS FROM UBFC OR PURE TO MMPD. BEST RESULTS ARE MARKED IN BOLD

Method	UBFC → MMPD			PURE → MMPD		
	MAE	RMSE	R	MAE	RMSE	R
TS-CAN [47]	14.01	21.04	**0.24**	**13.94**	**21.94**	**0.20**
PhysNet [20]	16.32	24.73	0.08	18.41	26.63	-0.05
DeepPhys [48]	17.50	25.00	0.05	16.92	24.61	0.05
EfficientPhys [46]	13.78	20.43	0.12	14.03	21.62	0.17
SFDA-rPPG (Ours)	**12.38**	**15.28**	0.04	14.73	22.05	0.13

from the source domain for pretraining and utilized only a single unlabeled video from the target domain for adaptation. Subsequently, we conduct testing using UBFC and PURE datasets. As shown in Table V, we compare the results of SFDA's methods for one-shot training in the target domain and the outcomes of the direct cross-domain method using PhysNet, which is supervised training in the source domain.

TABLE V

ONE-SHOT EVALUATIONS BETWEEN UBFC AND PURE. THE RESULTS SHOWN IN LIGHT GRAY BACKGROUND REPRESENT THE CROSS-DOMAIN TESTING BETWEEN UBFC AND PURE, WHERE OTHERS ARE INTRADOMAIN TESTING. BEST RESULTS ARE MARKED IN BOLD

Method	Train data		Test data					
			UBFC			PURE		
	Source	Target	MAE	RMSE	R	MAE	RMSE	R
PhysNet [20]	UBFC	-	0.75	1.47	0.99	3.81	-	0.87
	PURE	-	7.02	-	0.60	0.99	5.22	0.93
SHOT [30]	UBFC	PURE	3.83	11.06	0.57	0.33	1.15	0.99
	PURE	UBFC	1.08	1.87	0.98	0.56	2.00	0.98
CoWA [38]	UBFC	PURE	0.75	1.47	0.99	1.33	3.06	0.97
	PURE	UBFC	0.92	1.78	0.98	0.50	1.73	0.99
SFDA-rPPG (Ours)	UBFC	PURE	0.67	1.41	0.99	**0.28**	**1.11**	**0.99**
	PURE	UBFC	**0.58**	**1.22**	**0.99**	0.44	1.70	0.99

The analysis demonstrates that our method can achieve commendable cross-domain performance, even when employing only one video from the target domain for training, thus indicating its superior learning and generalization capability.

G. Ablation Study

1) Impact of Branchings: To show the effectiveness of our overall framework, we compare the results produced by our methodology with the detachments of the branch architecture, and also with the direct cross-domain assessment of our pretrained models. As shown in Table VI, our approach outperforms other approaches in both 1-shot and ALL settings, establishing its adaptability and effectiveness across diverse conditions. Our findings also indicate that the transition

TABLE VI

ABLATION RESULTS OF DIFFERENT BRANCHING SETTINGS. THE RESULTS SHOWN IN LIGHT GRAY BACKGROUND REPRESENT THE CROSS-DOMAIN TESTING BETWEEN UBFC AND PURE, WHERE OTHERS ARE INTRADOMAIN TESTING. BEST RESULTS ARE MARKED IN BOLD

Method	Setting	Train data		Test data					
				UBFC			PURE		
		Source	Target	MAE	RMSE	R	MAE	RMSE	R
PhysNet	0-shot	UBFC	-	0.58	1.22	0.99	1.06	2.93	0.97
		PURE	-	0.83	1.73	0.99	0.22	1.10	0.99
Only Temporal	1-shot	UBFC	PURE	3.83	11.06	0.57	0.33	1.15	0.99
		PURE	UBFC	1.08	1.87	0.98	0.56	2.00	0.98
	ALL	UBFC	PURE	5.83	15.38	0.17	0.28	1.11	0.99
		PURE	UBFC	0.67	1.41	0.99	0.78	2.40	0.98
Only Spatial	1-shot	UBFC	PURE	1.50	3.11	0.96	0.50	1.97	0.98
		PURE	UBFC	0.83	2.00	0.99	2.39	9.69	0.70
	ALL	UBFC	PURE	5.83	15.38	0.17	0.33	1.15	0.99
		PURE	UBFC	0.58	1.22	0.99	0.83	2.60	0.98
SFDA-rPPG (Ours)	1-shot	UBFC	PURE	0.67	1.41	0.99	**0.28**	**1.11**	**0.99**
		PURE	UBFC	0.58	1.22	0.99	0.44	1.70	0.99
	ALL	UBFC	PURE	10.92	22.30	0.29	0.33	1.15	0.99
		PURE	UBFC	**0.42**	**1.08**	**0.99**	0.56	1.76	0.99

from UBFC to PURE is relatively easy; consequently, the 1-shot learning outcomes surpass the ALL learning outcomes in this scenario. This also explains why the ALL learning outcomes from the sole temporal branch marginally exceed the performance of our approach. Nevertheless, in aggregate, our complete framework is better than the single-branch framework.

2) Impact of Loss Functions: As shown in Table I, we conduct a comparative assessment of various loss functions within PhysNet [20], employing the UBFC and PURE datasets for supervised training. To prove the effectiveness of our FWD loss, we compare the losses commonly used in the rPPG method, including Frequency loss [18] and KL loss in [42]. Experiments show that our FWD loss achieves the best performance not only within each dataset but also across datasets. This indicates that our FWD loss method is fully effective, and it is expected to be widely used in the rPPG field in the future.

3) Impact of Backbones: To verify the generalization of our framework, in addition to PhyNet [20] as the backbone, we also consider two other backbones, namely PhysFormer [42] and EfficientPhys [46]. Specifically, we train the backbones to obtain pretrained models, and split the backbones of PhysFormer and EfficientPhys into encoder and decoder parts. Then, we perform SFDA based on our proposed three-branch framework. As shown in Table VII, the method incorporating SFDA has significantly improved the cross-domain performance of the original method. This demonstrates that our spatiotemporal consistency framework possesses strong generalization capabilities and can be effectively applied to various backbone architectures.

4) Impact of Resampling Factor r: To analyze the sensitivity of resampling factor r, we conduct additional experiments examining different resampling factor ranges from UBFC-rPPG to the PURE dataset. As shown in Fig. 8, we evaluated three different ranges: [0.57, 0.66], [0.66, 0.8], and [0.8, 1.0],

TABLE VII

ABLATION OF DIFFERENT BACKBONES. "BASELINE" REFERS TO THE VERSION WITH THE BARE BACKBONE, WHILE "SFDA" REFERS TO THE VERSION WHERE THE BACKBONE IS INTEGRATED WITH OUR PROPOSED SFDA-RPPG FRAMEWORK

UBFC → PURE

Method	Baseline			SFDA		
	MAE	RMSE	R	MAE	RMSE	R
PhysNet [20]	3.81	-	0.87	**0.33**	**1.15**	**0.99**
PhysFormer [42]	5.44	14.17	0.43	**1.33**	**3.65**	**0.95**
EfficientPhys [46]	4.16	15.82	0.60	**2.38**	**9.98**	**0.81**

TABLE VIII

COMPARISON OF PARAMETERS AND COMPUTATIONAL EFFICIENCY. THE BEST RESULTS ARE MARKED IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Method	Param.(M)	MACs(G)
PhysNet [20]	**0.77**	**56.1**
TS-CAN [47]	7.5	<u>96</u>
DeepPhys [48]	4.16	<u>96</u>
Contrast-phys+ [37]	<u>0.86</u>	109.5
SFDA-rPPG(Ours) (training stage)	2.53	328.6
SFDA-rPPG(Ours) (testing stage)	<u>0.86</u>	109.5

which increased heart rate frequency by [1.5, 1.75], [1.25, 1.5], and [1.0, 1.25], respectively. The results clearly demonstrate that the middle range [0.66, 0.8] yields the best performance with the lowest MAE (0.33 bpm) and RMSE (1.15 bpm) values, confirming that our selected range is indeed optimal for the rPPG estimation task.

H. Discussion and Visualization

1) Further Discussion on FWD: Here, we use PhysNet as the basic framework and compare intra- and cross-dataset results among the proposed FWD loss and other three recent WD-related time/frequency-domain losses, that is, time-domain WD [28], mutual information loss [55], and frequency-domain MCC [41]. The results in the time and frequency domains are shown in Fig. 6. The experimental results show that the models trained with FWD loss perform obviously better than those with time-domain WD. For example, in the case of cross-datasets from UBFC to PURE, models with FWD loss can reach the result of MAE 1.06, while only 5.83 bpm from the models with time-domain WD. We can also find that our FWD performs more stably compared with mutual information loss and frequency-domain MCC in most intra- and cross-domain testing.

2) Parameters and Computational Efficiency: In Table VIII, we compare the number of parameters and multiply–accumulate operations (MACs) with other models. TSTC-Net demonstrates balanced efficiency with 2.53M parameters. While training requires 328.6G MACs due to its three-branch architecture, the inference phase uses only the

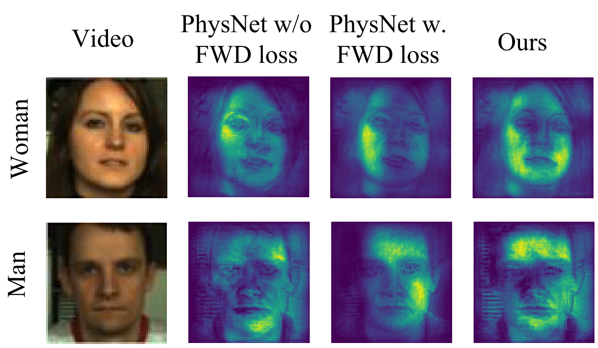

Fig. 9. Saliency maps of representative samples on the PURE dataset for the method with or without FWD loss. Brighter areas (yellowish-green) indicate regions where the model focuses more attention on extracting physiological signals.

main branch, reducing computational cost to 109.5G MACs with input size of $128 \times 128 \times 128$ ($T \times H \times W$). This design supports effective deployment on resource-limited devices.

3) Frequency-Domain Stability: To validate the stability advantage of frequency-domain analysis, we conducted a qualitative experiment applying 90° phase shifts to rPPG signals, as shown in Fig. 2. While the time-domain signals exhibit significant visual differences (left panel), the frequency-domain PSD distributions remain nearly identical with preserved peak locations around 1.8 Hz (right panel), demonstrating the inherent robustness of frequency-domain representations to phase variations commonly encountered in real-world scenarios.

4) Saliency Maps: We follow [29] to generate a comparative saliency map of our method and PhysNet [20], both with and without FWD loss. Specifically, we employ a gradient-based method to calculate the negative Pearson correlation gradient of input video results using fixed model weights, and then get the corresponding saliency map. To evaluate its cross-domain performance from UBFC to PURE, we train PhysNet weights with and without FWD loss on UBFC. Additionally, our SFDA-rPPG is pretrained on UBFC and then adapts to PURE. We select videos of two subjects from the PURE dataset to conduct a significant graph comparison. As illustrated in Fig. 9, experimental results demonstrate saliency map of our method activates the majority of skin regions in rPPG data, particularly concentrating on the cheek and forehead with rich rPPG signal. The results of PhysNet with FWD loss are more concentrated than normal PhysNet, again demonstrating the effectiveness of our FWD loss.

5) Visualization of Predicted and PSD: In the visualization tests of predicted BVP and PSD, we evaluate cross-dataset testing from PURE to UBFC, comparing our approach with PhysNet [20]. The ground truth and predicted BVP and PSD are compared for a test sample of subject 42 from the UBFC dataset. The frames from 100 to 550 are adopted for evaluation. The BVP value is normalized from 0 to 1. By converting the BVP into PSD, the frequency distribution is depicted alongside the BVP figure. As can be seen from Fig. 10, our approach aligns better with the ground truth in both BVP and PSD compared to PhysNet. This indicates that the model not only learns well on the heart rate prediction but also learns the trend of the BVP signal.

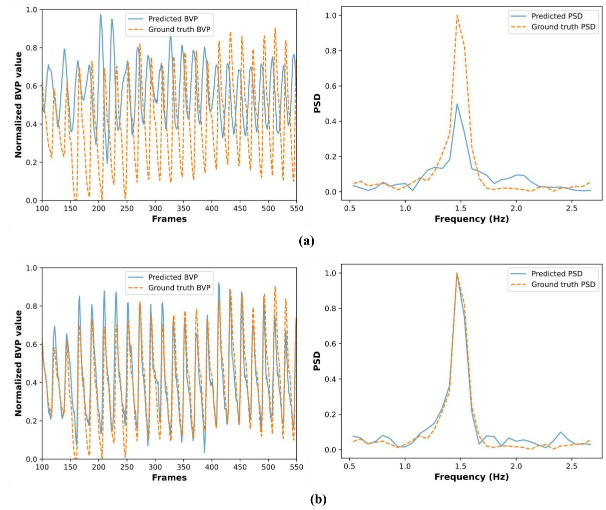

Fig. 10. Comparison of the predicted and ground-truth BVP and PSD visualization results for PhysNet and SFDA-rPPG in cross-dataset prediction from PURE to UBFC. (a) BVP and PSD visualization on cross-dataset test with PhysNet. (b) BVP and PSD visualization on cross-dataset test with SFDA-rPPG.

V. Conclusion

In this article, we introduced SFDA-rPPG, a novel framework for SFDA in rPPG measurement. Our framework addresses domain shift challenges by incorporating a TSTC-Net, which effectively reduces the domain gap by learning consistent features across domains. Additionally, we proposed a FWD loss to align the spectrum distributions of rPPG signals, significantly enhancing adaptation performance. SFDA-rPPG presents a breakthrough for cross-domain and privacy-preserving applications in rPPG technology. In future work, we plan to explore pretraining on multiple source domains to further improve SFDA in rPPG measurement.

References

[1] Z. Yue et al., "Deep super-resolution network for rPPG information recovery and noncontact heart rate estimation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[2] M. Das, M. K. Bhuyan, and L. N. Sharma, "Time–frequency learning framework for rPPG signal estimation using scalogram-based feature map of facial video data," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.

[3] M. Hu, F. Qian, D. Guo, X. Wang, L. He, and F. Ren, "ETA-rPPGNet: Effective time-domain attention network for remote heart rate measurement," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[4] B.-F. Wu, B.-J. Wu, B.-R. Tsai, and C.-P. Hsu, "A facial-image-based blood pressure measurement system without calibration," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[5] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Opt. Exp.*, vol. 16, no. 26, pp. 21434–21445, 2008.

[6] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3430–3437.

[7] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Exp.*, vol. 18, no. 10, pp. 10762–10774, 2010.

[8] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2396–2404.

[9] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.
[10] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017.
[11] S. B. Park, G. Kim, H. J. Baek, J. H. Han, and J. H. Kim, "Remote pulse rate measurement from near-infrared videos," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1271–1275, Aug. 2018.
[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[13] Y. Tian et al., "3D tooth instance segmentation learning objectness and affinity in point cloud," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 4, pp. 1–16, Nov. 2022.
[14] Y. Tian et al., "A revised approach to orthodontic treatment monitoring from oralscan video," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 12, pp. 5827–5836, Dec. 2023.
[15] Y. Tian et al., "RGB oralscan video-based orthodontic treatment monitoring," *Sci. China Inf. Sci.*, vol. 67, no. 1, Jan. 2024, Art. no. 112107.
[16] B. Dong, Y. Liu, K. Yang, and J. Cao, "Realistic pulse waveforms estimation via contrastive learning in remote photoplethysmography," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, 2024.
[17] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "The benefit of distraction: Denoising camera-based physiological measurements using inverse attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4935–4944.
[18] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "AutoHR: A strong end-to-end baseline for remote heart rate measurement with neural searching," *IEEE Signal Process. Lett.*, vol. 27, pp. 1245–1249, 2020.
[19] Z. Yu, X. Li, P. Wang, and G. Zhao, "TransRPPG: Remote photoplethysmography transformer for 3D mask face presentation attack detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 1290–1294, 2021.
[20] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–12.
[21] R. Spetlík, V. Franc, J. Cech, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2018, pp. 3–6.
[22] H. Lu, Z. Yu, X. Niu, and Y. Chen, "Neuron structure modeling for generalizable remote physiological measurement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18589–18599.
[23] J. Du, S.-Q. Liu, B. Zhang, and P. C. Yuen, "Dual-bridging with adversarial noise generation for domain adaptive rPPG estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10355–10364.
[24] GaninYaroslav et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.
[25] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, "Maximum density divergence for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3918–3930, Nov. 2021.
[26] J. Wang, H. Lu, A. Wang, Y. Chen, and D. He, "Hierarchical style-aware domain generalization for remote physiological measurement," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 3, pp. 1635–1643, Mar. 2024.
[27] W.-H. Chung, C.-J. Hsieh, S.-H. Liu, and C.-T. Hsu, "Domain generalized rppg network: Disentangled feature learning with domain permutation and domain augmentation," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 807–823.
[28] Z. Sun, J. Junttila, M. Tulppo, T. Seppänen, and X. Li, "Non-contact atrial fibrillation detection from face videos by learning systolic peaks," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 9, pp. 4587–4598, Sep. 2022.
[29] Z. Sun and X. Li, "Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 492–510.
[30] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6028–6039.
[31] J. N. Kundu, N. Venkat, and R. V. Babu, "Universal source-free domain adaptation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4544–4553.
[32] K. Li, D. Patel, E. Kruus, and M. R. Min, "Source-free video domain adaptation with spatial–temporal-historical consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14643–14652.
[33] H. Xia, H. Zhao, and Z. Ding, "Adaptive adversarial network for source-free domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9010–9019.
[34] Y. Li, J. Huang, J. Zhao, D. Wu, and M. Zheng, "TS-CAN+: A improved TS-CAN architecture for non-contact heart rate measurement," *IEEE Trans. Consum. Electron.*, vol. 71, no. 1, pp. 1393–1401, Feb. 2025.
[35] H. Lu, H. Han, and S. K. Zhou, "Dual-GAN: Joint BVP and noise modeling for remote physiological measurement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12404–12413.
[36] R.-N. Yin, R.-S. Jia, Z. Cui, and H.-M. Sun, "PulseNet: A multitask learning network for remote heart rate estimation," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 108048.
[37] Z. Sun and X. Li, "Contrast-Phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5835–5851, Aug. 2024.
[38] J. Lee, D. Jung, J. Yim, and S. Yoon, "Confidence score for source-free unsupervised domain adaptation," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, 2022, pp. 12365–12377.
[39] Z. Yue, M. Shi, and S. Ding, "Facial video-based remote physiological measurement via self-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 1–16, 2023.
[40] H. Shao, L. Luo, J. Qian, S. Chen, C. Hu, and J. Yang, "TranPhys: Spatiotemporal masked transformer steered remote photoplethysmography estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 3030–3042, Apr. 2024.
[41] J. Gideon and S. Stent, "The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 3995–4004.
[42] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4186–4196.
[43] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.
[44] A. Ramdas, N. Trillos, and M. Cuturi, "On Wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, Jan. 2017.
[45] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[46] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "EfficientPhys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4997–5006.
[47] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proc. NIPS*, 2020, pp. 19400–19411.
[48] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 349–365.
[49] K. I. Park, M. Park, and James, *Fundamentals of Probability and Stochastic Processes With Applications to Communications*. Cham, Switzerland: Springer, 2018.
[50] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. 23rd IEEE Int. Symp. Robot Human Interact. Commun.*, Aug. 2014, pp. 1056–1062.
[51] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.*, vol. 124, pp. 82–90, Jun. 2019.
[52] G. Heusch, A. Anjos, and S. Marcel, "A reproducible study on remote heart rate measurement," 2017, *arXiv:1709.00962*.
[53] J. Tang et al., "MMPD: Multi-domain mobile video physiology dataset," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–5.
[54] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 151–160.
[55] S. Zhao, Y. Wang, Y. Zheng, and D. Cai, "Region mutual information loss for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.