Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement

Hao Lu[1,2], Hu Han[1,3,*], S. Kevin Zhou[4,1]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Peng Cheng Laboratory, Shenzhen, China
[4] Medical Imaging, Robotics, and Analytic Computing Laboratory and Engineering (MIRACLE),
School of Biomedical Engineering & Suzhou Institute for Advance Research,
University of Science and Technology of China, Suzhou 215123, China
hao.lu@miracle.ict.ac.cn, hanhu@ict.ac.cn, s.kevin.zhou@gmail.com

Abstract

Remote photoplethysmography (rPPG) based physiological measurement has great application values in health monitoring, emotion analysis, etc. Existing methods mainly focus on how to enhance or extract the very weak blood volume pulse (BVP) signals from face videos, but seldom explicitly model the noises that dominate face video content. Thus, they may suffer from poor generalization ability in unseen scenarios. This paper proposes a novel adversarial learning approach for rPPG based physiological measurement by using Dual Generative Adversarial Networks (Dual-GAN) to model the BVP predictor and noise distribution jointly. The BVP-GAN aims to learn a noise-resistant mapping from input to ground-truth BVP, and the Noise-GAN aims to learn the noise distribution. The two GANs can promote each other's capability, leading to improved feature disentanglement between BVP and noises. Besides, a plug-and-play block named ROI alignment and fusion (ROI-AF) block is proposed to alleviate the inconsistencies between different ROIs and exploit informative features from a wider receptive field in terms of ROIs. In comparison to state-of-the-art methods, our approach achieves better performance in heart rate, heart rate variability, and respiration frequency estimation from face videos.

1. Introduction

Physiological signals such as heart rate (HR), respiration frequency (RF), and heart rate variability (HRV) are important indicators of human health status. Traditionally, these physiological signals are measured using electrocardiography (ECG) and photoplethysmography (PPG); both are skin-contact based approach, which is intrusive and may cause may cause discomfort for human. Recently, no-contact physiological measurement approaches based on remote photoplethysmography (rPPG) have attracted increasing attention, and most of the approaches use face videos recorded by commodity cameras to perform rPPG based physiological measurement [18, 19, 25]. The principle behind rPPG based physiological measurement is that the optical absorption by skin changes periodically along with the periodic blood volume change due to heartbeat. Thus, if we can capture the periodic skin color changes, we can obtain the heart rate. However, such skin color changes are very weak and can be easily affected by various noises such as illumination and head movement.

Early approaches for rPPG based physiological measurement usually use PCA or ICA to decompose the raw temporal signal [1, 20, 28] or perform color space transformations like CHROM [8] and POS [37] to extract the BVP signals. These hand-crafted algorithms usually make certain assumptions about the background noises, e.g., the motion has the same influence to the intensity variations of different color channels, based on which the CHROM method can remove the motion influence by computing the ratio between three channels [8]. However, since the human face is not an ideal Lambertian object, such assumptions do not always hold for every facial region. As a result, while the hand-crafted methods may not require training and have relatively good generalization ability, there is still big room to improve the physiological measurement accuracy. In addition, hand-crafted methods may not effectively leverage big training dataset to learn informative features even when a

*Corresponding author.

large dataset is available.

With the great success of deep learning (DL) in various computer vision tasks [14, 23, 24], DL methods have also been studied for rPPG based physiological measurement [25, 27]. Considering the low PSNR of BVP signals in face videos, DL methods usually first compute a spatial-temporal map (STMap) [25] or the difference of frames (DOF) [27] before using convolutional neural networks (CNNs) to learn informative features for physiological signals. However, most DL methods only focus on how to extract BVP signals from videos but ignore the modeling of background noises that dominate the video's content.

Some studies demonstrated that using synthetic physiological signals with artificial noises can benefit the training of DL-based physiological measurement methods [21, 27]. However, these synthetic noises are generally compiled by mixing Gaussian noise with trigonometric functions (e.g., sine or cosine) or step pulses. Apparently, such a synthetic physiological signal generation approach is not able to replicate the real noise distribution in practical scenarios. As a result, their effectiveness for improving the model robustness is also limited.

To address these issues, we propose a novel adversarial learning approach for robust rPPG based remote physiological measurement by using dual generative adversarial networks (Dual-GAN) to simultaneously model the BVP predictor and noise distribution. As illustrated in Fig. 1, we first compute the spatial-temporal map (STMap) from the input video to obtain a preliminary representation of the BVP signal. Then, we use one GAN model to learn the mapping from STMap to BVP, in which the generator (a.k.a. BVP estimator) aims to generate a BVP signal as similar as the ground-truth BVP, while the discriminator aims to distinguish between the generated BVP from the ground-truth BVP. The other GAN is used to model the noise distribution w.r.t. the BVP, in which the generator (including a noise-free and a noise STMap generator) aims to generate a synthetic STMap as similar as the STMap computed from video, while the discriminator (shared) aims to distinguish between the synthetic STMap and the real STMap. The Dual-GAN can enhance the capability of each other. The former improves the noise distribution learning ability of the latter, and the latter, like online data augmentation, can in turn improve the robustness of the former BVP predictor against unseen noises.

In addition, existing methods [25] treat the temporal signals of different ROIs (individual rows of STMap) indiscriminately during convolution; however face is not an ideal Lambertian object, the temporal signals of different ROIs should have different BVP and noise distributions. Therefore, we also propose a plug-and-play module named ROI alignment and fusion (ROI-AF) block, which can perform ROI-wise convolution to handle such noise distribution inconsistency of individual ROIs, and by fusing BVP features from a wider receptive field.

The contributions of this work are as follows:

1) We propose a novel Dual-GAN architecture for rPPG based physiological measurement, which can not only model BVP predictor but also explicitly model noise distribution via adversarial learning, and thus can obtain more robustness BVP representation against unseen noises.

2) We propose a plug-and-play ROI-AF block, which can be used after conventional convolution layers to address noise and BVP distribution inconsistency among different ROIs.

3) The proposed approach outperforms the state-of-the-art methods in HR, HRV and RF estimation under both intra-dataset and cross-dataset testings, showing its robustness again complicated scenarios.

2. Related Work

2.1. Remote Physiological Measurement

Remote physiological measurement aims to achieve HR, HRV, and RF estimation from videos recorded by commodity cameras. Traditional methods are mainly based on certain skin reflection model to perform signal decomposition to obtain the BVP signal. De Haan $et\ al.$ proposed a chrominance based color space projection (CHROM) to eliminate the influence of head movement for HR estimation [8]. Later, pixel-wise CHROM was proposed to further improve the HR estimation [38]. Wang $et\ al.$ proposed a spatial subspace rotation (2SR) based on the correlation of the three color channels to improve the HR estimation robustness. De Haan $et\ al.$ studied the signature of rPPG signals at different wavelengths and then proposed a blood-volume pulse vector method to extract the pulse signal [9]. Independent component analysis (ICA) was employed from the perspective of blind signal separation to separate the BVP signals from three color channels for HR estimation [28]. These methods are designed manually under certain assumptions, and may not work very well when the image acquisition conditions change.

Deep learning (DL) model has powerful nonlinear fitting capabilities and has been successfully used in various computer vision tasks. There are also attempts of studying DL-based remote HR estimation [6, 15, 21, 25, 32, 39, 42]. DeepPhys computed the difference of frames and used a deep convolutional network to extract physiological signals. An attention mechanism was also proposed to reduce the effects of motion [6]. Wang $et\ al.$ proposed a novel two-step CNN and adopted a low-rank constraint loss to derive reliable features [39]. A 3D convolutional network (named Spatial-Temporal Net) was used to estimate HR directly from the video by fusing both spatial and temporal information [32]. The face video contains lots of irrelevant in-

formation besides the BVP signal, and thus some studies tried to design effective hand-crafted representations for the physiological signal [15, 25]. Time-frequency representation was proposed as the input of CNN, which directly accumulated the frequency component of the physiology signals extracted from video [15, 29]. A spatial-temporal representation was designed as the input of CNN, which was a combination of the temporal physiology signals extracted from different ROIs of the face [25]. There are also a number of studies focusing on how to suppress the noises mixed together with the physiological signals [18, 26]. Niu $et\ al.$ attempted to remove the noise from the MSTMap via cross-verified feature disentangling [26]. Lee $et\ al.$ tried to use meta-learning to cope with the noise distribution changes during model deployment [18]. Song et al. proposed to learn the BVP signal distribution via GAN to improve the BVP waveform quality [31]. Most of the existing methods focus on how to enhance and extract the BVP signal from the video, but ignore explicit modeling of the background noises that is also important for improving the BVP signal extraction. Different from the existing methods, we propose to simultaneously model BVP predictor and noise distribution via adversarial learning. As a result, we can better disentangle BVP features and noise features to achieve more robust physiological measurement.

2.2. Deep Generative Noise Modeling

Generative models aim at learning the true data distribution from a limited number of data. Recent advances in parameterizing these models using deep neural networks have led to successful applications in various computer tasks like image conversion [44], face editing [7], image super-resolution [10] and denoising [4]. One of the most commonly used and efficient deep generative models is Generative Adversarial Networks (GAN) [12]. GAN is effective for learning unknown data distributions using unsupervised learning and has achieved tremendous success in many applications.

For example, GAN has been widely utilized to generate noise-free images [5, 16, 43]. For remote physiological measurement task, some methods tried to use synthetic physiological signals with noises for data augmentation and to improve the model robustness [21, 27]. These synthetic physiological signals were generated by mixing clean signals, such as trigonometric functions (e.g., sine or cosine) and step pulses, with Gaussian noises, which may not replicate real-world noises very well. Considering the effectiveness of GAN for noise modeling, some studies investigated GAN-based physiological signal generation, and reported improved performance than previous physiological measurement methods [11, 30].

We propose a Dual-GAN for modeling the noise distribution and physiological estimator jointly for remote physiological measurement. Our approach falls under the GAN-based approach, but with several significant differences compared with existing methods: (i) Compared with the noise synthesis in [21, 27], our model can learn the distribution of real physiological signals by GAN and therefore can generate more realistic physiological signals that can replicate the ones in real applications. (ii) Different from the image denoising methods [5, 16, 43], our method does not require paired data (e.g., noise and noise-free) to learn the noise distribution. (iii) Different from the synthesis methods in [11, 30], we simultaneously learn a BVP estimator and a generative noise model, which can enhance the capability of each other, leading to better feature disentanglement between BVP and noise signals.

3. Proposed Method

We denote the i-th input video as v^i, and the corresponding ground-truth BVP signal as s^i_{gt}. The goal of remote physiological measurement is to learn a mapping: $F : v^i \rightarrow s^i_{gt}$. Since the BVP signal is very weak compared with the background face content (characterizing identity and attributes) and the noises in the video, existing methods usually build a composite function for F by leveraging either hand-crafted transformations [8, 28, 37] or deep learning [6, 18, 42] to extract the BVP signal. However, these methods do not *explicitly* model the noise distribution. As a result, the disentanglement between the BVP signal and the background noises can be sub-optimum, leading to poor generalization ability in new scenarios. We propose an adversarial learning framework to jointly model both BVP predictor and noise distribution using a Dual-GAN network. As shown in Fig. 1, besides learning the mapping from STMap m^i to BVP s^i_{gt} via a GAN consisting of a generator (named BVP estimator F_b) and a discriminator D, we also learn the noise distribution with a peer GAN consisting of two sub-generators G_{phy} and G_{noise}, and a discriminator D, in which D is shared by both GANs. The details of our approach are described as follows.

3.1. Spatial-Temporal Map

As discussed in [6, 25], direct applying CNNs to the face video may not effectively exploit the information of the physiological signal. Therefore, we choose to use STMap as the input of CNN like [22, 25], which establishes a preliminary representation of the physiological signal by discarding most of the irrelevant background content. Let m^i denote the STMap computed from v^i. The dimension of m^i is $n \times l \times c$, in which n denotes the number of ROIs, l denote the number of frames of a clip, and $c = 3$ denotes the three channels of R, G and B. Then our goal is to establish a mapping $F_b : m^i \rightarrow s^i_{gt}$.

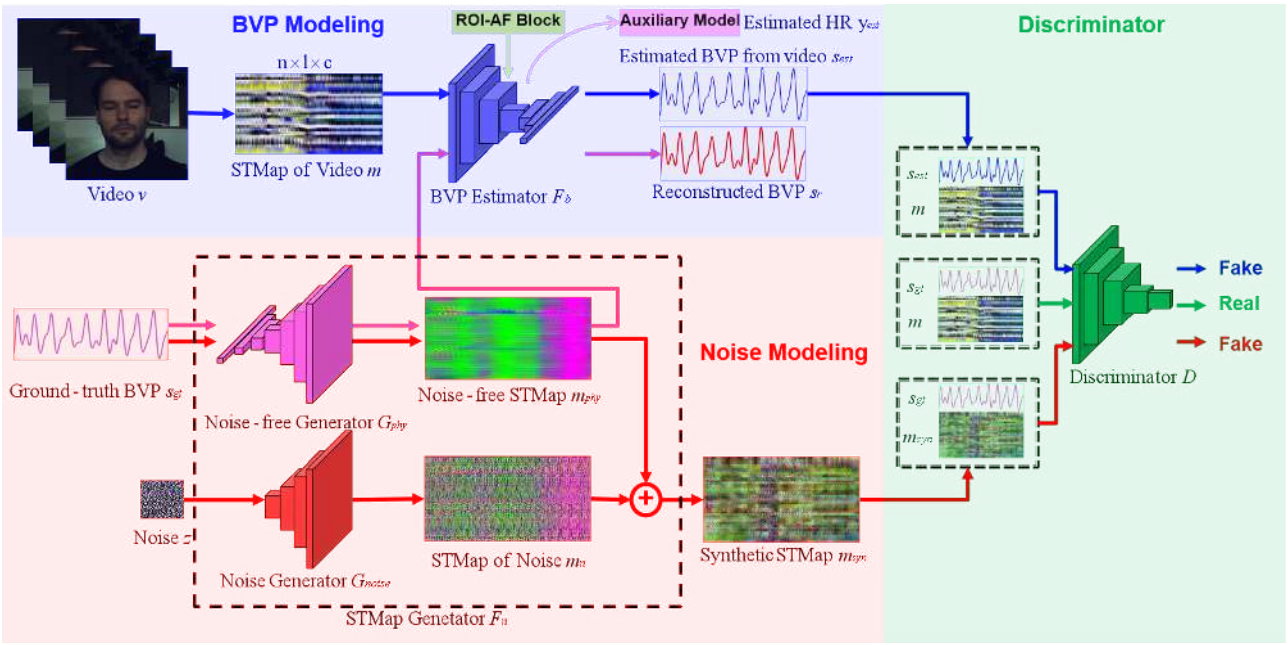

Figure 1. The architecture of our Dual-GAN for jointly modeling BVP predictor and noise distribution. The GAN for BVP modeling consists of a generator F_b for BVP signal predictor from STMap, and a discriminator D for distinguishing between the predicted BVP signal s_{est} and the ground-truth BVP signal s_{gt}. The GAN for noise modeling consists of a two-path generator for generating a synthetic STMap from ground-truth BVP and random noise variable, and uses a shared discriminator D to distinguishing between synthetic STMap m_s and the real STMap m. The Dual-GAN structure allows us to perform indirect supervision w.r.t. the noises so that we can better model the noise distribution and achieve better feature disentanglement for the BVP signal.

3.2. ROI Alignment and Fusion Block

Each row of STMap m^i represents the raw temporal signal for one ROI on the face. Since the face is not an ideal Lambertian object, the BVP and noise distributions of different ROIs are different. However, existing methods usually perform the same convolutions for the temporal signals of different ROIs [25, 21]. Such a manner can be suboptimum when filtering out the noises in different ROIs to obtain the BVP signals. A more reasonable approach is to perform different convolutions for different ROIs. In another aspect, as indicated in [17, 26], the BVP signals from different ROIs of the facial region should be nearly synchronized, and a large receptive field w.r.t. ROIs can be helpful for extracting the periodic BVP signals with higher PSNR. However, given the limited receptive filed of convolution kernels, current CNNs may only leverage a few adjacent ROIs (*i.e.*, a few adjacent rows in the STMap and feature map) at each convolution operation.

To solve these problems, we propose a simple yet efficient block, named ROI alignment and fusion block (ROI-AF), to perform feature alignment and fusion across ROIs. As shown in Fig. 2, a feature map of CNN or the input STMap is split by rows, and per-row 1D-Conv. is performed to alleviate BVP and noise distribution differences and achieve across-ROI feature alignment purpose. Then, the aligned features are concatenated in terms of channels and 1-D conv. with a channel attention model, *i.e.*, a global average pooling (GAP) followed by two linear layers (FC) is applied to fuse the aligned features from individual ROIs. Finally, the fused feature map is reshaped to its original dimensions. The proposed ROI-AF block is actually plug-and-play and can be inserted into different convolution blocks of a CNN network.

3.3. Dual-GAN

Dual-GAN jointly models the BVP predictor and noise distribution via an adversarial learning. We detail BVP-GAN and Noise-GAN below.

BVP-GAN. BVP-GAN consists of a generator F_b which learns a mapping from STMap m to the ground-truth BVP signal s_{gt}, *i.e.*, $s_{est} = F_b(m)$, and a discriminator D which aims to distinguish between the estimated BVP signal s_{est} and the ground-truth BVP signal s_{gt}. The generator F_b consists of four "ConvBlock" and four "UpBlock" as shown in Fig. 3 (a), which takes the STMap as input, and outputs an estimated BVP signal. Four ROI-AF blocks are inserted ahead of the four "ConvBlock" to reduce BVP and noise distribution inconsistencies across ROIs. Besides BVP signal prediction, we also introduce an auxiliary task, *i.e.*, performing HR regression from STMap, which is expected to improve the feature learning robustness of generator F_b via multi-task learning.

To measure the quality of the estimated BVP signal s_{est},

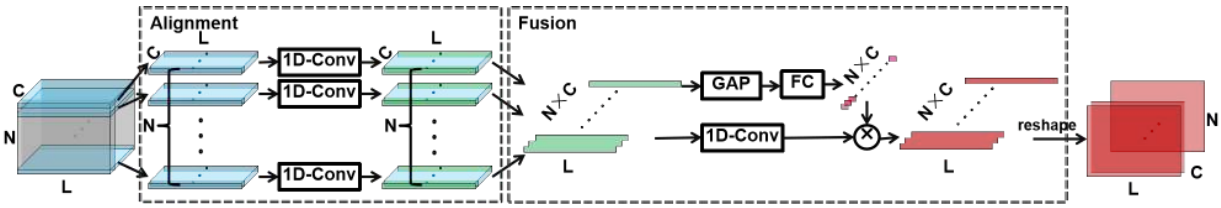

Figure 2. ROI Alignment and Fusion Block (ROI-AF): A feature map or STMap is split by rows, and per-row 1D-Conv. is performed to handle BVP and noise distribution differences to achieve feature alignment. Then, the aligned features are concatenated in terms of channels and 1-D conv. with a channel attention model (*i.e.*, a global average pooling (GAP) followed by two linear layers (FC)) is performed to obtain a fused feature map. Finally, the fused feature map is reshaped to its original dimensions.

we only need to focus on its periodicity instead of the amplitude changes. Therefore, we choose to use a negative Pearson correlation loss calculated between the estimated BVP signals s_{est} and the ground-truth BVP signals s_{gt}:

$$\mathcal{L}_p = 1 - PCor(s_{est}, s_{gt}) \tag{1}$$

where $PCor(\cdot)$ is the Pearson correlation [42]. We also use a frequency domain loss [26], which is defined as a cross-entropy loss between the spectral distribution of the estimated BVP signal and the one-hot code of ground-truth HR frequency o_{gt}:

$$\mathcal{L}_{fre} = CE(PSD(s_{est}), o_{gt}) \tag{2}$$

where $PSD(\cdot)$ denotes the power spectral density of s_{est}, and $CE(\cdot)$ denotes the cross-entropy loss. The ground-truth HR y_{gt} can also be represented by a one-hot vector $o_{gt} = [0, ..., 0, 1, 0, ...]$, and '1' denotes the index corresponding to y_{gt}. $PSD(s_{est})$ can be regarded as one-hot vector of HR, e.g., $p = [0.1, ..., 0.1, 0.5, 0.1, ...]$. For the auxiliary HR regression task, we calculate the L_1 distance between the estimated HR y_{est} and the ground-truth HR y_{gt} for supervision. In summary, the overall loss function for generator F_b can be written as:

$$\mathcal{L}_{phy} = \lambda_1 ||y_{est} - y_{gt}||_{L_1} + \lambda_2 \mathcal{L}_p + \lambda_3 \mathcal{L}_{fre} \tag{3}$$

where λ_1, λ_2 and λ_3 are balancing parameters. In our experiments, we set $\lambda_1 = 0.2$, $\lambda_2 = 1$, and $\lambda_3 = 0.1$ empirically according to individual loss scales.

The discriminator D of BVP-GAN consists of five "ConvBlock", a "GAP" and an "FC", which takes the combination of STMap and BVP signals as input, and outputs fake or real as shown in Fig. 3 (d). The loss function of D is defined as:

$$\begin{aligned} \max_{D} \min_{F_b, G_{noise}} \mathcal{L}_{joint} = & log(D(s_{gt}, m)) - log(D(F_b(m), m)) \\ & - log(D(s_{gt}, F_n(g_{gt}, z))) \end{aligned} \tag{4}$$

where $F_n(\cdot)$ denotes the generator of the Noise-GAN that aims to generate a STMap from ground-truth BVP and random noise variable, which will be discussed in the next subsection. Through the adversarial learning between generators F_b and discriminator D, we expect that F_b can predict BVP signals as close as possible to the ground-truth BVP signals.

Noise-GAN. One key challenge that prohibits existing methods to explicitly model the noise distribution during BVP estimation is that there is no paired data (*i.e.*, data with and without noises) for supervision. We address this challenge by using an indirect supervision manner, *i.e.*, supervising the sum of a known BVP distribution and an unknown noise distribution.

Specifically, Noise-GAN consists of a STMap generator F_n for generating STMap from ground-truth BVP signal and random noise variables, and a discriminator D, which is shared with BVP-GAN. The generator F_n uses a two-path structure, consisting of two sub-generators (G_{phy} and G_{noise}) and a sum operation. While G_{phy} aims to generate a noise-free STMap from the ground-truth BVP signal, *i.e.*, $m_{phy} = G_{phy}(s_{gt})$, G_{noise} aims to generate a noise STMap from random noise variable z sampled from a Gaussian distribution, *i.e.*, $m_n = G_{noise}(z)$. We assume an additive model between noises and BVP signal, and sum m_{phy} and m_n together to obtain a synthetic STMap m_{syn}. Thus, F_n aims to generate a synthetic STMap m_{syn} that is as close as possible to the real STMap m used in BVP-GAN.

As shown in Figs. 3 (b) and (c), G_{phy} consists of four "ConvBlock", and four "UpBlock", and G_{noise} consists of four "UpBlock". The adversarial loss for F_n is also Eq. (4).

Our Dual-GAN structure also allows us to introduce extra supervision to ensure individual components can function as we expect. Specifically, we expect the generator F_b in BVP-GAN can only filter out the noises while retain the useful BVP information. Then, based on F_b and G_{phy}, the reconstructed BVP signal s_r is expected to be the same as s_{gt}: $s_{gt} = s_r = F_b(G_{phy}(s_{gt}))$. Such a constraint can be achieved by minimizing a negative Pearson correlation coefficient L_r

$$\mathcal{L}_r = 1 - PCor(F_b(G_{phy}(s_{gt})), s_{gt}) \tag{5}$$

Training Strategy. We train our Dual-GAN using an alternative training strategy. In each batch, we sample both real training data (m, s_{gt} and y_{gt}) and synthetic data (m_{syn} and s_{est}) generated by our models[1]. These data are then

[1]For the first iteration, the synthetic data is generated by network with random weight.

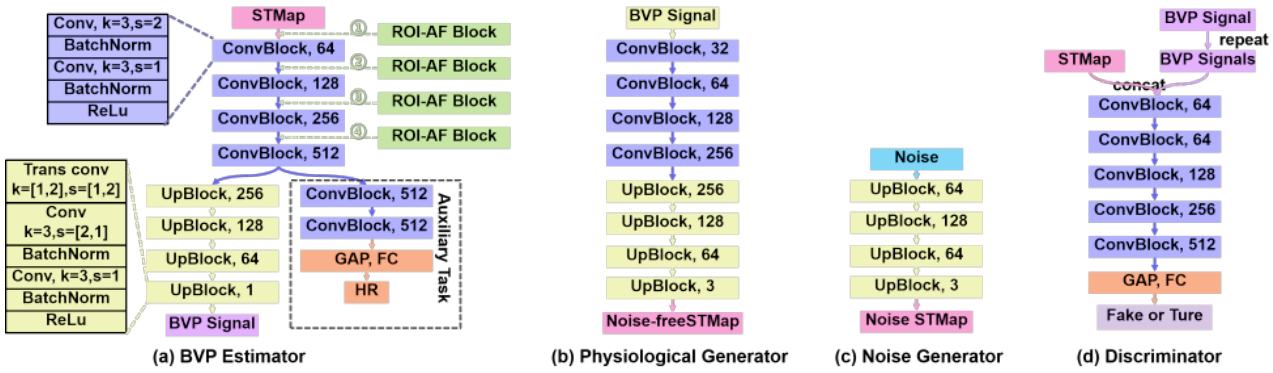

Figure 3. The architectures of the (a) BVP estimator F_b, (b) noise-free STMap generator G_{phy}, (c) noise STMap generator G_{noise} and (d) shared discriminator D. "ConvBlock" denotes two convolution layers, "UpBlock" denotes transposed convolution and two convolution layers, "GAP" denotes global averaging pooling, "FC" denotes two linear layers.

used as follows: (i) training F_b using training data (m, s_{gt}, and y_{gt}), and synthetic data m_{syn} with corresponding labels s_{gt} and y_{gt} using loss function $\mathcal{L}_{phy}$ in Eq. (3); (ii) training G_{phy} using loss function $\mathcal{L}_r$ in Eq. (5) with F_b fixed; (iii) training F_b and G_{noise} by minimizing $\mathcal{L}_{joint}$ in Eq. (4); and (iv) training the shared discriminator D by maximizing $\mathcal{L}_{joint}$ in Eq. (4) with all generators fixed.

Given the proposed Dual-GAN structure and the alternative training strategy, both $F_b(G_{phy}(s_{gt}))$ and $F_b(G_{phy}(s_gt) + G_{noise}(z))$ are expected to obtain the ground-truth BVP signal s_{gt}. Thus, the improved noise modeling capability of G_{noise} will in turn improve the robustness of F_b against noises that are not present in the original training data. This can improve the generalization ability of the proposed approach for physiologic measurement under unseen scenarios.

Some methods are also based on Dual-GAN such as [13, 40, 41], but our work is very different from theirs: (i) Dual-GAN is used in [13] for novel retina and segmentation image synthesis, which requires paired data (mask and retina image) for supervision; however the noise modeling in our method does not require paired noise-free data for supervision. (ii) [41] used Dual-GAN to perform style-level image translation, but does not retain pixel-level correspondence. Our BVP feature disentanglement from STMap with noises require precise temporal alignment. (iii) Dual-GAN was used in [40] to reuse multiple pre-trained networks for multi-label classification via knowledge amalgamation, which is a completely different task than our feature disentanglement.

4. Experimental Results

We perform rPPG based physiological measurement for three types of physiological signals, *i.e.*, heart rate (HR), heart rate variability (HRV), and respiration frequency (RF), using three public-domain datasets (UBFC-rPPG [2], VIPL-HR [22], and PURE [33]).

4.1. Databases and Experimental Settings

UBFC-rPPG [2] contains 42 RGB videos containing sunlight and indoor illumination. The videos were recorded with a Logitech C920 HD Pro webcam in a resolution of 640 × 480 and 30 fps. The ground-truth BVP signals and HR values were collected by CMS50E.

VIPL-HR [22] is a challenging large-scale database for remote physiology measurement under less-constrained scenarios. It contains 2,378 RGB videos of 107 subjects captured with different head movements, lighting conditions and acquisition devices. In addition, the frame rate of the videos varies due to different recording scenarios and devices. We normalize the videos and the corresponding BVP signals to 30 fps by cubic spline interpolation.

PURE [33] contains 60 RGB videos from 10 subjects with 6 different activities (sitting still, talking, four variations of rotating and moving head), which were recorded using an eco274CVGE camera at 30 fps and a resolution of 640 × 480. The BVP signal was collected using CMS50E at 60 fps. The BVP signals are reduced to 30 fps with cubic spline interpolation to align with the videos.

Tasks and Evaluation Metrics: We perform HR, HRV, and RF estimation on UBFC-rPPG, HR estimation on VIPL-HR and PURE, and cross-database HR estimation on UBFC-rPPG with training on PURE. For HR estimation, we follow [18, 22] and report the standard deviation of the error (Std), mean absolute error (MAE), root mean square error (RMSE), mean error rate percentage (MER), and Pearson's correlation coefficient (r). For HRV and RF estimation, we follow existing methods [26, 42] and report low frequency (LF), high frequency (HF), and LF/HF ratio in terms of Std, RMSE, and r.

Parameters: Our algorithm is implemented in the PyTorch and trained on NVIDIA TITAN Xp. All CNNs are trained for 10 epochs, using random initialization, via adam optimizer with default beats of 0.9 and 0.999, weight decay of 0, learning rate of 0.0001, batch size of 32 and without any decay strategy for learning rate. For all the experiments,

Table 1. RF and HRV estimation results by our method and several state-of-the-art methods on the UBFC-rPPG database.

Method	LF-(u.n)			HF-(u.n)			LF/HF			RF-(Hz)		
	Std↓	RMSE↓	r↑	Std↓	RMSE↓	r↑	Std↓	RMSE↓	r↑	Std↓	RMSE↓	r↑
POS [37]	0.171	0.169	0.479	0.171	0.169	0.479	0.405	0.399	0.518	0.109	0.107	0.087
CHROM [8]	0.243	0.240	0.159	0.243	0.240	0.159	0.655	0.645	0.226	0.086	0.089	0.102
Green [36]	0.186	0.186	0.280	0.186	0.186	0.280	0.361	0.365	0.492	0.087	0.086	0.111
CVD [26]	0.053	0.065	0.740	0.053	0.065	0.740	0.169	0.168	0.812	0.017	0.018	0.252
Dual-GAN (Ours)	**0.034**	**0.035**	**0.891**	**0.034**	**0.035**	**0.891**	**0.131**	**0.136**	**0.881**	**0.010**	**0.010**	**0.395**

Table 2. HR estimation results of our method and several state-of-the-art methods on the UBFC-rPPG database.

Method	MAE↓	RMSE↓	MER↓	r↑
POS [37]	8.35	10.00	9.85%	0.24
CHROM [8]	8.20	9.92	9.17%	0.27
Green [36]	6.01	7.87	6.48%	0.29
SynRhythm [21]	5.59	6.82	5.5%	0.72
PulseGAN [31]	1.19	2.10	1.24%	0.98
Dual-GAN (Ours)	**0.44**	**0.67**	**0.42%**	**0.99**

Table 3. HR estimation results by our method and several state-of-the-art methods on the VIPL-HR database.

Method	Std↓	MAE↓	RMSE↓	r↑
SAMC [35]	18.0	15.9	21.0	0.11
POS [37]	15.3	11.5	17.2	0.30
CHROM [8]	15.1	11.4	16.9	0.28
I3D [3]	15.9	12.0	15.9	0.07
DeepPhy [6]	13.6	11.0	13.8	0.11
RhythmNet [25]	8.11	5.30	8.14	0.76
CVD [26]	7.92	5.02	7.97	0.79
Dual-GAN (Ours)	**7.63**	**4.93**	**7.68**	**0.81**

the length of each video clip is set to 256 frames, and the step between clips is 10 frames. We use the data balance and random horizontal flipping in [25] for data augmentation when computing STMap from face videos. In all experiments, we compute HR, HRV, and RF based on the average duration between two adjacent BVP signal peaks [26, 42].

4.2. Results

HR, HRV and RF estimation on UBFC-rPPG: Following the protocol in [31], the videos of the first 30 subjects are used for training, and the videos of the remaining 12 subjects are used for testing. For HRV and RF estimation, we compare our approach with POS [37], CHROM [8] and Green [36] as shown in Table 1, which are performed in iPhys [2]. CVD [26], a DL method, is also used for comparison [3]. For HR estimation, besides our approach and the above three methods, we also provide the results of SynRhythm [21] and PulseGAN [31]. We implement SynRhythm by ourselves and use the results of PulseGAN from the original paper since we use the same protocol as [31].

The HR estimation results by individual methods are shown in Table 2. We can see that DL methods like SynRhythm and PulseGAN perform much better than the hand-crafted methods such as POS, CHROM, and Green. This suggests that DL methods can learn more informative features than hand-crafted methods for BVP signal prediction and HR estimation. Compared with the best of the baseline methods, *i.e.*, PulseGAN, the proposed Dual-GAN can further reduce the errors (MAE, RMSE, and MER) significantly, and improve the Pearson correlation coefficient between the estimation HR and the ground-truth HR. The reasons why our Dual-GAN outperforms PulseGAN are threefold: (i) While PulseGAN can also generate new physiological signals for data augmentation by learning the physiological signal distribution output by CHROM, which does not explicitly model the noises mixed with the BVP signal. By contrast, our jointly modeling of BVP predictor and noise distribution enables us to obtain better disentangled features for BVP signals; (ii) joint training of Dual-GAN in our approach can make them promote each other's capability; (iii) the ROI-AF block can improve the representation learning capability of the network (see our ablation study).

The HRV and RF estimation results are shown in Table 1. LF, HF, and LF/HF are three measures for HRV estimation, each reported with Std, RMSE, and r, respectively. We can see that the proposed approach outperforms all the baseline methods for HRV and RF estimation under all measures. This is reasonable because HRV and RF computation rely on the BVP estimation accuracy.

HR estimation on VIPL-HR and PURE: We further evaluate the effectiveness of the proposed approach by performing HR estimation on two more challenging VIPL-HR and PURE datasets. We follow [25, 26] and use a subject-exclusive 5-fold cross-validation protocol on VIPL-HR. We compared our method with seven baseline methods (SAMC [35], POS [37], I3D [3], DeepPhy [6], RhythmNet [25], and CVD [26]) on VIPL-HR, in which the performance of these methods are directly from [25, 26]. As shown in Table 4, the proposed Dual-GAN outperforms all the baseline methods under all measures. For PURE dataset, we follow the same testing protocol in [32] and compare with 2SR [9], CHROM [8], and HR-CNN [32], for which their performance are available in [32]. From Table 4, we can see that the proposed approach again outperforms all the baseline methods. These results show that the proposed Dual-GAN is still very effective for HR estimation under less-constrained scenarios.

[2] https://github.com/danmcduff/iphys-toolbox

[3] https://github.com/nxsEdson/CVD-Physiological-Measurement

Table 4. HR estimation results of our method and several state-of-the-art methods on the PURE database.

Method	MAE↓	RMSE↓	r↑
2SR [9]	2.44	3.06	0.98
CHROM [8]	2.07	9.92	0.99
HR-CNN [32]	1.84	2.37	0.98
Dual-GAN (Ours)	**0.82**	**1.31**	**0.99**

Table 5. Cross-database HR estimation (training on PURE and testing on UBFC-rPPG) by our Dual-GAN and baseline methods.

Method	MAE↓	RMSE↓	MER↓	r↑
GREEN [36]	8.29	15.82	7.81%	0.68
ICA [28]	4.39	11.60	4.30%	0.82
POS [37]	3.52	8.38	3.36%	0.90
CHROM [8]	3.10	6.84	3.83%	0.93
PulseGAN [31]	2.09	4.42	2.23%	0.97
Siamese-rPPG [34]	1.29	8.73	\	\
Dual-GAN (Ours)	**0.74**	**1.02**	**0.73%**	**0.997**

Table 6. Ablation study of our Dual-GAN in terms of Noise-GAN and ROI-AF modules for HR estimation on UBFC-rPPG.

Noise-GAN	ROI-AF	MAE↓	RMSE↓	MER↓
w/o	**w/o**	1.59	3.14	1.53%
w/o	**w.**	0.84	1.81	1.26%
w.	**w/o**	0.66	1.52	0.82%
w.	**w.**	**0.44**	**0.67**	**0.42%**

Cross-database HR estimation: Generalization ability is very important for remote physiological measurement model. We perform cross-database evaluations to verify the generalization ability of our approach. We follow [31, 34] to train our model on PURE and test it on UBFC-rPPG. The cross-database HR estimation results of our method and the baseline methods are given in Table 5, in which the results of GREEN [36], ICA [28], POS [37], CHROM [8] are from [31], and the results of PulseGAN and Siamese-rPPG are form [31] and [34], respectively. From Table 5, we can see that our model still achieves the best results under all evaluation metrics compared with these state-of-the-art methods. These results indicate that the proposed method generalizes well into new scenarios with unknown noises.

4.3. Ablation Study

We provide ablation study of our approach in terms of noise modeling and ROI-AF by performing HR estimation on UBFC-rPPG. We cover the following ablation studies: (I) without Noise-GAN and ROI-AF; (II) without Noise-GAN but with ROI-AF; (III) without ROI-AF but with Noise-GAN; and (IV) the whole Dual-GAN method. The results are shown in Table 6.

Comparing the results by (I) and (II), we can see that using ROI-AF can reduce the MAE and RMSE errors by 0.75 and 1.31, respectively. This suggests that reducing the noise and BVP diversities across different ROIs via ROI-AF is helpful for improving BVP estimation accuracy. In addition, we found that $\mathcal{L}_p$ converges to 0.0002 with ROI-AF blocks but converges to 0.0023 without ROI-AF.

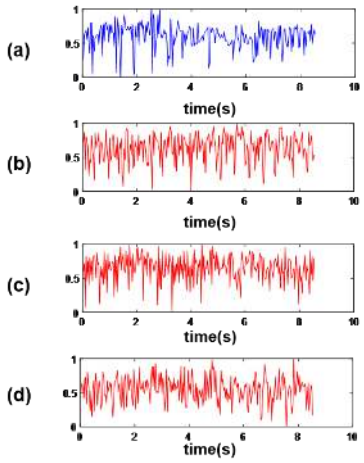

Figure 4. (a) The BVP signal with real noises computed from video (one row from STMap m). (b,c,d) Synthetic BVP signals with synthetic noises (one row from synthetic STMap m_{syn}) by using different noise variable z.

Similarly, using Noise-GAN for noise distribution modeling can greatly improve HR estimation accuracy. These results indicate that the Noise-GAN does learn a good distribution about the noises mixed with the BVP signals, which leads to better BVP feature disentanglement.

We also visualize the synthetic STMap by our Noise-GAN by changing the random noise variable z. From Fig. 4, we can notice that the same row from three synthetic STMaps shown in Figs. 4 (b, d, d) are very similar to the same row from the STMap computed from video. This also explains why Noise-GAN is able to work as online data augmentation to improve the BVP estimator's robustness.

5. Conclusion

Remote physiological measurement based on rPPG is challenging because of the very weak BVP signal and the strong noises. We propose to jointly model the BVP predictor and noise distribution using Dual-GAN, in which a BVP-GAN learns a mapping from input STMap to BVP signal, and a Noise-GAN learns a mapping from random noise variable and ground-truth BVP to STMap, both in an adversarial learning manner. We also propose an alternative training strategy for optimizing the generators and discriminators in Dual-GAN end-to-end. The dual GANs can promote each other's capability, leading to improved noise distribution estimation accuracy and enhanced feature disentanglement for BVP signals, which finally improves the physiological measurement accuracy.

6. Acknowledgment

This research was supported in part by the National Key R&D Program of China (grant 2018AAA0102501), and Youth Innovation Promotion Association CAS (grant 2018135).

References

[1] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proc. IEEE CVPR*, pages 3430–3437, 2013.

[2] 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.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE CVPR*, pages 6299–6308, 2017.

[4] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proc. IEEE CVPR*, pages 3155–3164, 2018.

[5] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proc. IEEE CVPR*, pages 3155–3164, 2018.

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

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE CVPR*, pages 8789–8797, 2018.

[8] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Trans. Biomed. Eng.*, 60(10):2878–2886, 2013.

[9] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiol. Meas.*, 35(9):1913, 2014.

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proc. ECCV*, pages 184–199, 2014.

[11] Tomer Golany and Kira Radinsky. Pgans: Personalized generative adversarial networks for ecg synthesis to improve patient-specific deep ecg classification. In *Proc. AAAI*, volume 33, pages 557–564, 2019.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS*, pages 2672–2680, 2014.

[13] John T. Guibas, Tejpal S. Virdi, and Peter S. Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017.

[14] Hu Han, Jie Li, Anil K Jain, Shiguang Shan, and Xilin Chen. Tattoo image search at scale: Joint detection and compact representation learning. *IEEE Trans. Pattern Anal. Mach. Intell*, 41(10):2333–2348, 2019.

[15] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In *Proc. IJCB*, pages 383–389, 2017.

[16] Dong-Wook Kim, Jae Ryun Chung, and Seung-Won Jung. Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In *Proc. IEEE CVPR*, pages 0–0, 2019.

[17] Antony Lam and Yoshinori Kuno. Robust heart rate measurement from video using select random patches. In *Proc. ICCV*, 2016.

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

[19] Yu-Chen Lin, Yu-Jen Wang, Jason Chia-Hsien Cheng, and Yuan-Hsiang Lin. Contactless monitoring of pulse rate and eye movement for uveal melanoma patients undergoing radiation therapy. *IEEE Trans. Instrum. Meas.*, 68(2):474–482, 2018.

[20] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Trans. Biomed. Eng.*, 61(10):2593–2601, 2014.

[21] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *Proc. IEEE ICPR*, pages 3580–3585, 2018.

[22] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Proc. ACCV*, pages 562–576, 2018.

[23] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proc. CVPR*, 2019.

[24] Xuesong Niu, Hu Han, Jiabei Zeng, Xuran Sun, Shiguang Shan, Yan Huang, Songfan Yang, and Xilin Chen. Automatic engagement prediction with gap feature. In *Proc. ICMI*, pages 599–603, 2018.

[25] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Trans. on Image Process.*, 29:2409–2423, 2020.

[26] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *Proc. ECCV*, 2020.

[27] Olga Perepelkina, Mikhail Artemyev, Marina Churikova, and Mikhail Grinenko. Hearttrack: Convolutional neural network for remote video-based heart rate monitoring. In *Proc. IEEE CVPR Workshops*, pages 288–289, 2020.

[28] 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.

[29] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[30] P. Singh and G. Pradhan. A new ecg denoising framework using generative adversarial network. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, pages 1–1, 2020.

[31] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse

waveforms in remote photoplethysmography. *IEEE J-BHI*, pages 1–1, 2021.

[32] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proc. BMVC*, pages 3–6, 2018.

[33] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *Proc. IEEE ISRHIC*, pages 1056–1062, 2014.

[34] Yun-Yun Tsou, Yi-An Lee, Chiou-Ting Hsu, and Shang-Hung Chang. Siamese-rppg network: remote photoplethysmography signal estimation from face videos. In *Proc. ACM SAC*, pages 2066–2073, 2020.

[35] 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 *Proc. IEEE CVPR*, pages 2396–2404, 2016.

[36] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, 2008.

[37] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Trans. Biomed. Eng.*, 64(7):1479–1491, 2017.

[38] Wenjin Wang, Sander Stuijk, and Gerard De Haan. Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE Trans. Biomed. Eng.*, 62(2):415–425, 2015.

[39] Zhi-Kuan Wang, Ying Kao, and Chiou-Ting Hsu. Vision-based heart rate estimation via a two-stream cnn. In *Proc. IEEE ICIP*, pages 3327–3331, 2019.

[40] J. Ye, Y. Ji, X. Wang, X. Gao, and M. Song. Data-free knowledge amalgamation via group-stack dual-gan. In *Proc. IEEE CVPR*, pages 12513–12522, 2020.

[41] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proc. IEEE ICCV*, pages 2868–2876, 2017.

[42] 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 *Proc. IEEE ICCV*, pages 151–160, 2019.

[43] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *Proc. IEEE ECCV*, 2020.

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE ICCV*, pages 2223–2232, 2017.