PhysFlow: Skin tone transfer for remote heart rate estimation through conditional normalizing flows

Joaquim Comas [1]
joaquim.comas@upf.edu
Antonia Alomar[1]
antonia.alomar@upf.edu
Adrià Ruiz[2]
adriaruiz@seedtag.com
Federico Sukno[1]
federico.sukno@upf.edu

[1] Department of Information and Communication Technologies
Pompeu Fabra University
Barcelona, Spain

[2] Seedtag
Madrid, Spain

Abstract

In recent years, deep learning methods have shown impressive results for camera-based remote physiological signal estimation, clearly surpassing traditional methods. However, the performance and generalization ability of Deep Neural Networks heavily depends on rich training data truly representing different factors of variation encountered in real applications. Unfortunately, many current remote photoplethysmography (rPPG) datasets lack diversity, particularly in darker skin tones, leading to biased performance of existing rPPG approaches. To mitigate this bias, we introduce PhysFlow, a novel method for augmenting skin tone diversity in remote heart rate estimation using conditional normalizing flows. PhysFlow adopts end-to-end training optimization, enabling simultaneous training of supervised rPPG approaches on both original and generated data. Additionally, we condition our model using CIELAB color space skin features directly extracted from the facial videos without the need for skin-tone labels. We validate PhysFlow on publicly available datasets, UCLA-rPPG and MMPD, demonstrating reduced heart rate error, particularly in dark skin tones. Furthermore, we demonstrate its versatility and adaptability across different data-driven rPPG methods.

1 Introduction

In recent years, the interest on camera-based measurement of physiological signals has undergone substantial growth caused by its potential applicability in clinical [21] and human-computer interaction applications [4, 37]. The recent advancements in deep learning have further propelled this field. However, data-driven methods often require a large amount of training data to achieve good generalization, making their performance highly dependent on the distribution and scale of the dataset chosen for training. Additionally, many rPPG

datasets suffer from biases, particularly regarding demographic diversity, leading to unfair performance depending on skin tone, especially for underrepresented ethnicities.

Previous studies [16, 38, 59] have investigated the effects of demographic biases in existing rPPG datasets, revealing a significant decline in remote heart rate estimation performance for subjects with the darkest skin tones. While these studies highlighted the consequences of skin tone bias, none of them proposed concrete solutions to address the problem. Although the creation of datasets that accurately represent all skin types may seem like a simple solution to overcome this problem, it is often impractical and resource-intensive, especially during the recruiting of the participants. Consequently, each dataset typically over-represents the population of the country in which it was collected. For example, datasets like VIPL-HR [35] or PURE [49] predominantly consist of Asian or Caucasian populations, respectively.

Only a few recent studies have addressed skin tone bias. Two of them [33, 61] used synthetic avatars to tackle data imbalance. However, these approaches are often not photo-realistic and can be time-consuming in data generation. Another work [57] employed radar hardware to improve rPPG signal recovery for dark skin tones. Nevertheless, this method is not always feasible and limits rPPG versatility with conventional cameras. Finally, Ba et al. [3] proposed a two-stage optimization using external pre-training [65] to transfer skin tone content, which is an interesting possibility to balance skin-tone distributions as long as external data with the underrepresented skin tones are available for pre-training.

In this context, we propose a novel data augmentation approach leveraging normalizing flows [13, 14] to disentangle the skin tone information from the rest of the facial video content. Additionally, they allow for attribute control by concatenating parameters to embeddings, facilitating the conditioning of specific properties such as skin tone. Furthermore, in this paper, we approach the representation of skin tone for rPPG estimation differently than before. Previous works typically used the Fitzpatrick scale [17] to evaluate or annotate skin tone, categorizing it into six levels from I (lightest) to VI (darkest). In contrast, we propose using a bi-dimensional representation [53] to quantify the apparent skin color, which extracts the perceptual light and hue angle from facial videos, offering three key advantages over the Fitzpatrick scale. Firstly, it enables automated skin tone measurement in each facial video, eliminating the need for manual annotations and allowing to use it into unlabeled data, common in the majority of rPPG datasets. Secondly, it simplifies the collection and annotation for new rPPG datasets, reducing human error inherent in manual Fitzpatrick scale annotations due to subjective perceptions. Lastly, unlike the Fitzpatrick scale, which considers only lightness, this bi-dimensional representation accounts for variations in hue, distinguishing between red and yellow hues. This broader perspective better accommodates diverse skin tones, particularly given that common experiments in existing rPPG datasets involve varying lighting conditions using different types of light sources.

The main contributions of the paper are three-fold:

- We introduce PhysFlow, a novel skin tone data augmentation approach for rPPG estimation, which adopts conditional normalizing flows to disentangle skin tone information from other appearance and temporal facial video features.
- We propose a novel training that leverages skin tone CIELAB color space representation without requiring external labels.
- We train the PhysFlow model end-to-end to optimize any supervised rPPG approach using the original and generated data within the same data generation process promoting a fast adaptation.

2 Related work

Camera-based PPG measurement. Since Takano et al. [51] and Verkruysse et al. [56] demonstrated remote HR measurement feasibility from facial videos, diverse methods for physiological data recovery have emerged. Some focus on regions of interest, employing techniques like Blind Source Separation [25, 43, 44], Normalized Least Mean Squares [26], or self-adaptive matrix completion [55]. Others utilize the skin optical reflection model, projecting RGB skin pixel channels into an optimized subspace to mitigate motion artifacts [12, 59]. Deep learning-based methods [24, 31, 39, 42, 48, 63] have surpassed conventional techniques, achieving state-of-the-art performance in estimating vital signs from facial videos. Some combine traditional methods with Convolutional Neural Networks (CNNs) to leverage advanced features [34, 36, 47]. Other recent works [24, 29] explore unsupervised approaches using meta-learning, enhancing generalization in out-of-distribution cases. Unlike previous approaches, some propose end-to-end models [6, 42, 62] to directly extract the rPPG signal from facial videos. Transformer-based models like Physformer [64] and RADIANT [19] have gained attention for leveraging long-range spatiotemporal features, although they currently lack optimization for mobile deployment. While promising, they may not yet demonstrate a significant performance advantage over CNN-based models [30]. Finally, works like [30] and [7] propose lightweight rPPG frameworks with competitive HR results while controlling computational cost.

Skin tone bias in rPPG measurement. Earlier works [2, 16, 59] noted the lowest signal-to-noise ratio and highest error rates for darker skin tones in assessments on private datasets. Prior to the work of Nowara et al. [38], the influence of skin tone on rPPG extraction had only been examined in limited datasets without population diversity and data-driven approaches. In their meta-analysis, Nowara et al. investigated the impact of skin type on rPPG signal recovery for each Fitzpatrick skin tone using three datasets. They found significantly poorer rPPG signal recovery for subjects with category VI compared to other skin types. In [15], a new color space for rPPG estimation was proposed, showing consistent results across all skin tones, albeit with a performance drop for darker skin tones (V and VI), similar to [38]. Subsequently, Dasari et al. [11] conducted an extensive study analyzing estimation biases of camera PPG methods across diverse demographics. They observed similar biases as with contact-based devices and environmental conditions.

rPPG data-augmented approaches. While various works have identified skin tone bias in rPPG estimation, only a few solutions have directly addressed its mitigation. Recent studies have proposed different data augmentation approaches considering factors like motion [40] or data scarcity [20, 54], but without targeting skin tone bias. Two recent works aimed to balance demographic groups present in current rPPG datasets. Firstly, Vilesov et al. [57] proposed a method to mitigate bias across skin tones using multi-modal fusion, combining information from an RGB camera and 77 GHz radar. However, the need for a radar device limits the applicability of camera-based measurements in all scenarios. Alternatively, Ba et al. [3] introduced a two-step joint optimization framework to translate light-skinned to dark-skinned facial videos while maintaining their pulsatile signals, resulting in a considerable reduction of HR estimation error in their private dataset. However, this approach has two main limitations. Firstly, the use of a predefined model [65] to translate skin tone is limited to only four labels (African, Asian, Caucasian, and Indian) and may not discern different variations for each skin label. Secondly, the two-stage joint optimization can lead to unrealistic skin tone transformations, with certain facial regions appearing more illuminated than others. In contrast, our novel approach is independent of external pre-training and can trans-

late skin tone from the same data or a reduced set of data. Our end-to-end approach using conditional normalizing flows produces more homogeneous, realistic, and diverse skin tone transfers. Also some studies have tackled the limited availability of data in existing benchmarks by generating synthetic data. For example, McDuff et al. [33] used Blender to create synthetic avatars aligned with cardiac and respiratory signals across various scenarios, while Wang et al. [61] employed a 3D morphable face model to generate synthetic facial videos. Although both methods show promise, they have limitations. Primarily, the generation of synthetic data can be excessively time-consuming, especially in the case of [33], and they may suffer from domain shift, where models trained on synthetic data fail to perform well on real datasets due to the domain gap between the datasets, requiring a domain adaptation stage [58]. Additionally, in both instances, the facial appearance lacks photo-realism, and certain physiological factors like Pulse Transit Time (PTT) are not considered. In contrast, our data augmentation approach proposes direct training that simultaneously trains the rPPG model while generating new data, preserving the photo-realism of the original data and the characteristics of pulsatile data.

3 Methodology

3.1 Preliminaries: Continous Normalizing Flows

Normalizing flows (NFs) [46, 50] are a class of generative neural networks that estimate an unknown complex data distribution, $p(z)$, from a known and simpler base probability distribution, $p(u)$, such as a normal distribution $\mathcal{N}(\mu,\sigma^2)$. The mapping between both distributions is achieved by applying a sequence of bijective and differentiable transformations $f_1,...,f_K : \mathbb{R}^d \to \mathbb{R}^d$. Additionally, the mapping between both distributions can be constrained to some set of conditions or attributes c, these types of NFs are commonly known as conditional normalizing flows (c-NFs). Hence, the relationship between $z|c$ and $u|c$ can be defined as:

$$u = f_\theta(z;c), z = f_\theta^{-1}(u;c) \tag{1}$$

where $u \sim p(u|c)$, $z \sim p(z|c)$, and f_θ is a bijective neural network parameterized by θ. Mathematically, c-NFs can express the log-likelihood by applying the change of variables rule:

$$\log p(z;c,\theta) = \log p(f_\theta(z;c)) - \sum_{i=1}^{k-1} \log|det(J_{\theta_i}(z;c)| \tag{2}$$

where $J_{\theta_i}(z;c) = \frac{\partial f_{\theta_i}(z;c)}{\partial z}$ is the Jacobian, and $|det(J_{\theta_i}(z;c)|$ represents the volume change caused by the transformations of f_θ.

The normalizing flow can be generalized from a discrete sequence to a continuous transformation [5, 18], parameterizing the dynamics of data transformation over time $\frac{\partial u_\tau}{\partial \tau} = g_\theta(\tau,c,u_\tau)$, where g_θ is a neural network, u_τ is the state at time τ, $u \sim \mathcal{N}(\mu,\sigma^2)$ and c is the conditioning attributes. The Continuous Normalizing Flow (CNF) computes z from u by integrating g_θ over time, express as:

$$z = u_{\tau_0} + \int_{\tau_0}^{\tau_1} g_\theta(\tau,c,u_\tau)d\tau \tag{3}$$

Following the Instantaneous Change of Variables theorem [5] where the change in log probability of a continuous random variable is equal to the trace of the Jacobin matrix, the

total change in log density can be written as:

$$\log p(u(\tau_1)) = \log p(u(\tau_0)) - \int_{\tau_0}^{\tau_1} Tr\left(\frac{\partial g_\theta}{\partial u_\tau}\right) d\tau \tag{4}$$

To compute the gradients with respect to the parameters a black-box ordinary differential equation (ODE) solver is applied using an adjoint sensitivity method [45]. In this work, we adopt CNFs because they are demonstrated to have better expressiveness and versatility than discrete NFs [1, 18].

3.2 PhysFlow

The proposed PhysFlow approach, outlined in Fig.1, aims to mitigate the effect of skin tone bias in rPPG data-driven models by generating a diverse range of skin tone-augmented facial videos through the usage of CIELAB skin tone features. Our model comprises three main components: a 3D auto-encoder (AE), a conditional CNF network, and the rPPG model.

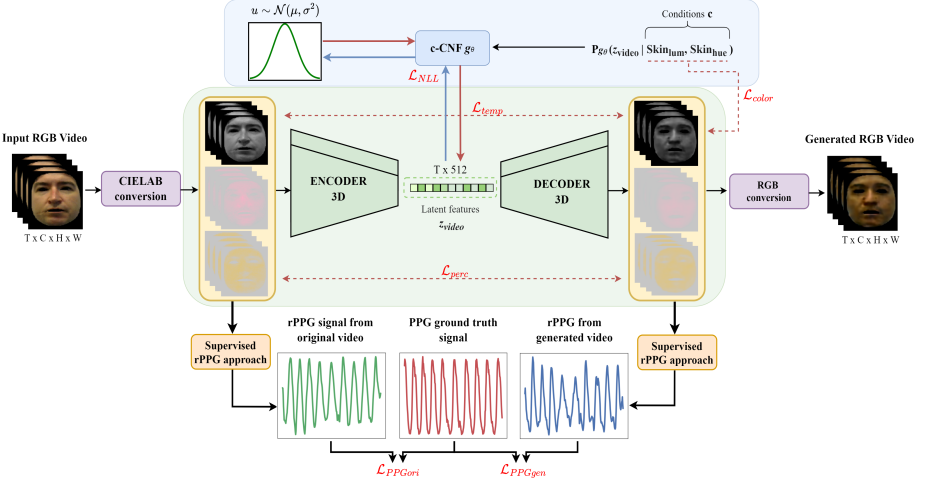

Figure 1: PhysFlow pipeline: A 3D-CNN AE encodes entangled video facial content into a latent embedding. This embedding is then processed by c-CNFs to disentangle the skin tone content. Simultaneously, the rPPG model is iteratively trained using both original and skin tone-augmented data.

3.2.1 3D video reconstruction

Inspired by several flow-based generative models [27, 28] in different computer vision tasks, we employ a conventional auto-encoder model, which is trained to encode and decode the high-dimensional information into low-dimensional feature embedding, independently of any disentanglement. Unlike previous works, we aim to preserve the underlying physiological signals that depend on subtle spatio-temporal variations. Therefore, the preservation of inter-frame content is crucial to recover the rPPG signal, thus we choose a 3D-CNN AE to model the spatio-temporal information of the input facial videos (the architecture details are explained in supplementary material). During the 3D-CNN AE training, the features are

only spatially downsampled and upsampled in order to not degrade the temporal content. To train the proposed 3D-CNN AE, we utilize the following combined loss:

$$\mathcal{L}_{\text{AE}} = \alpha_1 \mathcal{L}_{\text{rec}} + \alpha_2 \mathcal{L}_{\text{perc}} + \alpha_3 \mathcal{L}_{\text{rPPG}},$$
$$\mathcal{L}_{\text{rec}} = \left\|\hat{X}_t - X_t\right\|_1, \quad \mathcal{L}_{\text{perc}} = \frac{1}{C_j H_j W_j}\left\|\phi_j(\hat{X}_t)) - \phi_j(X_t))\right\|_1, \quad \mathcal{L}_{\text{rPPG}} = \left\|rPPG_{rec} - rPPG_{ori}\right\|_2 \tag{5}$$

where α_*, controls the importance between different loss components and $\mathcal{L}_{rec}$ is the reconstructed loss between the reconstructed $\hat{X}_t$ and original X_t facial videos computed by the L_1-loss. The $\mathcal{L}_{perc}$ is a VGG perceptual loss [22] adopted to retain high-frequency details between the reconstructed and original video, where ϕ_j represents the perceptual function which outputs the activation function of the jth layer in the VGG network and C_j, H_j, W_j are the dimensions of the tensor feature map. Finally, $\mathcal{L}_{rPPG}$ is a L_2-loss between the extracted rPPG in the original and reconstructed facial video to ensure the preservation of the physiological data.

3.2.2 Skin tone representation and conditioning

Our data augmentation approach relies on two primary assumptions. Firstly, for effective augmentation, the unbalanced dataset should feature a minimum level of skin tone diversity, enabling the transfer of skin tone without relying on an external dataset. Secondly, consistent luminance conditions within the same video are necessary for successful skin tone transfer. Hence, in this work, we select the UCLA-rPPG dataset [61]. Despite its significant imbalance in skin tone representation, the dataset offers sufficient diversity and samples, while maintaining constant illumination across each facial video. Based on these assumptions, we model our skin tone transfer employing a bi-dimensional representation [53] expressed in CIELAB color space through two components: luminance and hue. The luminance component determines the lightness or darkness of the skin, while the hue component, derived by dividing the alpha and beta channels, spans from red to yellow. Fig. 2 illustrates the comparison between our bi-dimensional skin tone label representation and the annotated Fitzpatrick labels from the UCLA-rPPG dataset. We notice that the luminance and hue values are not well captured by the subjective Fitzpatrick annotations, i.e. some subjects annotated in a specific skin type may belong to another one, particularly between skin types I to V, as seen in Fig. 2. These subjective annotations may stem from the challenge of labelling a diverse range of skin tones within only six main classes, where considerable diversity exists within each class. Therefore, the use of CIELAB skin tone labels not only facilitates the skin tone transfer without relying on manual Fitzpatrick labels but also ensures the ability to map large variations of skin tones and overcome illumination changes between different trials on the same subject, which could affect the perception of skin type.

To apply skin tone transfer, once the 3D auto-encoder has converged, we freeze its parameters and proceed to train the whole system. During the end-to-end training, the c-CNF module is responsible for mapping the entangled latent representation to a disentangled form. The invertibility of the flow-based model ensures that no information is lost during the transition from the entangled to the disentangled latent space. To achieve skin tone disentanglement, our CNF model is conditioned on luminance and hue values computed for a sequence of frames.

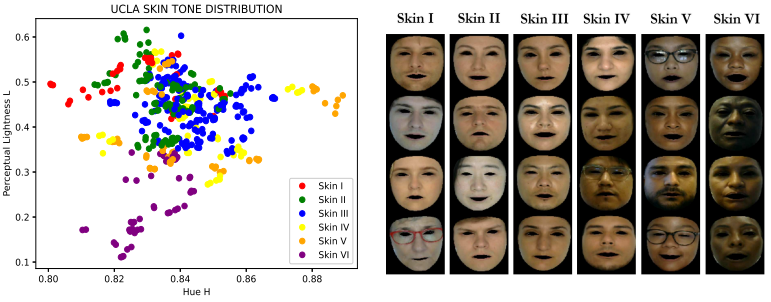

Figure 2: Skin tone representation in UCLA-rPPG. Left: Distribution representation of skin tone in terms of CIELAB luminance and hue compared to annotated Fitzpatrick scale labels from the dataset. Right: Visual examples of different skin tones with Fitzpatrick labels.

3.2.3 Model training and optimization

For training the PhysFlow model, we utilize end-to-end training, optimizing both the c-CNF module and the rPPG approach simultaneously. This involves minimizing the continuous form of the negative log-likelihood, as previously introduced in Equation 4:

$$\mathcal{L}_{\text{NLL}} = -\log p(u(\tau_0)) + \int_{\tau_0}^{\tau_1} Tr\left(\frac{\partial g_\theta}{\partial u_\tau}\right) d\tau \tag{6}$$

During the PhysFlow training, we also control the synthesis of the skin tone transfer by conditioning the c-CNF with specific transferable skin tone values and assessing the generated output. For that, we introduce three regularization losses. Firstly, we compute a consistency color loss $\mathcal{L}_{\text{color}}$ between the conditioned skin tone values st_{cond} and the resulting skin tone measured from the generated video $\hat{st}_{pred}$. Secondly, to enforce the preservation of the temporal content we consider a temporal consistency loss $\mathcal{L}_{\text{temp}}$, which is computed by minimizing the error of the first-order derivative of the original z_{video} and the conditioned $\hat{z}_{video}$ latent space:

$$\mathcal{L}_{\text{color}} = \left\|\hat{st}_{pred} - st_{cond}\right\|_1, \qquad \mathcal{L}_{\text{temp}} = \left\|\frac{\partial \hat{z}_{video}}{\partial t} - \frac{\partial z_{video}}{\partial t}\right\|_1 \tag{7}$$

Lastly, we use the same perceptual loss $\mathcal{L}_{\text{perc}}$ from Eq. 5 to keep the overall appearance of the original face. At the same time, to optimize the supervised rPPG approach we incorporate a physiological loss which combines the minimization of the rPPG signal concerning the ground truth PPG from the original and also from the generated video, which can be written as:

$$\mathcal{L}_{\text{phys}} = \mathcal{L}_{\text{PPGgen}} + \mathcal{L}_{\text{PPGori}} \tag{8}$$

To compute $\mathcal{L}_{\text{phys}}$, in our experiments, we use the Negative Pearson correlation [62] and TALOS [7] losses depending on the selected rPPG approach. Finally, we can express the overall training objective for PhysFlow as the minimization of the sum of all the losses:

$$\mathcal{L}_{\text{overall}} = \lambda_1\mathcal{L}_{\text{NLL}} + \lambda_2\mathcal{L}_{\text{perc}} + \lambda_3\mathcal{L}_{\text{color}} + \lambda_4\mathcal{L}_{\text{temp}} + \lambda_5\mathcal{L}_{\text{phys}} \tag{9}$$

where λ_*, controls the importance between different loss components.

4 Experiments

4.1 Experimental Setup

Datasets and models: We consider two recent datasets, UCLA-rPPG [61] and MMPD [52]. Both benchmarks consist of facial videos and corresponding gold-standard PPG signals, while also containing Fitzpatrick labels (see supplementary materials for more details). To evaluate PhysFlow performance, we use three deep-learning models: PhysNet [62], EfficientPhys [30], and DPMN+TDM [9]. For our experiments, we use the UCLA-rPPG dataset for training and MMPD for cross-dataset evaluation. Also, we include some traditional methods such as POS [60] and CHROM [12] as baseline results.

Metrics and evaluation protocol: To evaluate the HR estimation performance of the proposed model, we adopt the same metrics used in the literature, such as the mean absolute HR error (MAE), the root mean squared HR error (RMSE), the mean absolute percentage error (MAPE) and Pearson's correlation coefficients R [26, 29, 30]. Our experiments are performed using subject evaluation with a 10-sec sequences average with no overlap to compute HR estimation for all reported metrics. For our cross-dataset experiment, we utilize the MMPD dataset considering only the skin tone. For this reason, we followed the same protocol in [52] to exclude from our analysis other factors such as exercise or lighting conditions.

Implementation details: In all our experiments, we utilize the Mediapipe Face Mesh model [23, 32] to focus our analysis only on facial skin pixels. After masking the facial video, each frame is resized to 96×96 pixels. The PPG ground truth is pre-processed following [10] to denoise the raw PPG signal, which facilitates a better model convergence during the training. We use PyTorch 2.1.2 [41] and train on a single NVIDIA RTX3060 using sequences of 300 frames without overlap and AdamW optimizer with a learning rate of 0.0001. As explained in section 3.2.2, we first pre-trained the 3D AE for 310 epochs and then froze the weights. Following [28], during the 3D AE training, we add a small Gaussian noise of 0.05 to the latent space to prevent the latent distribution from collapsing to single delta peaks. Furthermore, for fast convergence, once the 3D AE is trained we also pre-trained the c-CNF module for 50 epochs and then trained end-to-end jointly with the supervised rPPG model. The weights for different losses in 3D AE training are set as $\alpha = 500$, $\alpha = 10$ and $\alpha = 5$, while in PhysFlow training are set as $\lambda_1 = 1$, $\lambda_2 = 100$, $\lambda_3 = 1000$, $\lambda_4 = 0.0001$ and $\lambda_5 = 10$. Finally, the predicted rPPG signal is filtered using a band-pass filter with cut-offs of 0.66-3Hz while the heart rate is calculated using Chirp-Z Transform (CZT) [8].

4.2 Experimental Results

Table 1 summarizes the assessment of remote HR performance across different Fitzpatrick scale splits in MMPD. These baseline evaluations are conducted using both traditional and data-driven approaches. From these results, several conclusions can be drawn. Firstly, across all methods, there is a noticeable decline in heart rate performance with darker skin tones, consistent with findings in previous studies [2, 15, 38]. Despite the similar representation of skin types IV, V, and VI between the datasets, a significant performance gap is evident, potentially influenced by their proximity to the more prevalent types such as skin tones II and III (as depicted in supplementary materials). Finally, deep learning models demonstrate substantial superiority over traditional approaches for lighter skin types (III and IV). However, for the most challenging skin types (V and VI), certain traditional methods perform comparably or even outperform some deep learning models, indicating skin tone-biased performance

	Hand-crafted methods						Deep learning methods								
Method	POS [60]			CHROM [12]			PhysNet [62]			EfficientPhys [30]			DPMN+TDM [9]		
Skin Tone	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑	MAE↓	RMSE↓	R↑
Lighter skin types															
III	5.76	9.67	0.48	6.57	10.46	0.33	3.90	7.98	0.56	4.18	9.40	0.60	3.09	5.92	0.78
IV	9.06	13.51	0.23	10.57	13.80	0.15	7.69	10.93	0.56	7.42	14.35	0.41	4.98	8.78	0.69
Darker skin types															
V	12.78	16.69	-0.03	14.65	18.91	-0.12	10.65	13.73	0.16	11.92	19.84	0.04	7.77	11.90	0.46
VI	11.17	15.34	0.26	12.53	16.47	0.06	13.74	18.19	-0.04	16.25	23.84	0.05	11.08	15.22	0.25

Table 1: Baseline cross-database evaluation on MMPD dataset for each skin tone (beats per minute).

attributed to dataset imbalance.

To showcase the advantages of integrating the PhysFlow approach into rPPG data-driven models for addressing biased skin tone challenges, we assess our method on skin types V and VI, which represent the most challenging skin types, as mentioned previously. To mitigate the skin tone unbalancing effect, during training PhysFlow generates an augmented range of darker skin by conditioning each sequence with a luminance value between 0.15 and 0.40, which allows the model to generate and learn from a wider diversity of darker skin tones. Fig. 3 illustrates a qualitative demonstration of the PhysFlow augmentation for dark skin tones, while ensuring the preservation of pulsatile information extracted from the facial region.

Table 2 depicts the cross-dataset HR performance comparison among the three evaluated data-driven models for skin types V and VI in the MMPD dataset, with and without the integration of PhysFlow. The findings reveal a significant reduction in HR error, with the introduction of augmented dark skin tone videos resulting in a decrease in MAE ranging from 1 to 5 BPM, depending on the method. Skin type VI demonstrates a more substantial improvement compared to skin type V, attributed to its distinctiveness from the predominant skin type in the training dataset. While skin type V also experiences improvement, it is slightly less pronounced due to its proximity to the predominant skin type III in rPPG datasets. Additionally, there are variations in improvement among the adopted methods. For instance, the HR error for the EfficientPhys model is reduced by approximately 36 % for skin types V and VI, whereas the DPMN+TDM model shows reductions ranging from 13% to 22%. This can be explained since EfficientPhys was designed as a lightweight model for rPPG deployment which due to its limited characteristics is more sensitive to skin tone bias than PhysNet or DPMN+TDM. Nevertheless, after the PhysFlow data augmentation, the HR error reduction is observed across all tested models, and their performance aligns more closely with the tendency observed for skin type III in Table 1.

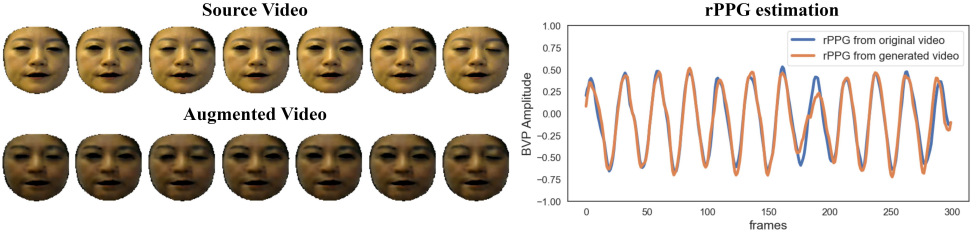

Figure 3: Visual example of dark skin tone data augmentation. PhysFlow transfers skin tone while preserving the pulsatile wave from the source to the augmented video.

Skin Tone		V				VI			
Method	Strategy	MAE↓	RMSE↓	MAPE↓	R↑	MAE↓	RMSE↓	MAPE↓	R↑
PhysNet [62]	Baseline	10.65	13.73	15.66	0.16	13.74	18.19	16.18	-0.04
	PhysFlow	**7.35**	**11.71**	**11.73**	**0.47**	**9.47**	**13.72**	**11.35**	**0.40**
	Difference	**-3.30**	**-2.02**	**-3.93**	**+0.31**	**-4.27**	**-4.47**	**-4.83**	**+0.43**
EfficientPhys [30]	Baseline	11.92	19.84	18.64	0.04	16.25	23.84	21.22	0.05
	PhysFlow	**7.71**	**13.36**	**11.51**	**0.49**	**10.53**	**16.22**	**13.69**	**0.36**
	Difference	**-4.21**	**-6.48**	**-7.13**	**+0.45**	**-5.72**	**-7.62**	**-7.53**	**+0.31**
DPMN+TDM [9]	Baseline	7.77	11.90	11.64	0.46	11.08	15.22	13.48	0.25
	PhysFlow	**6.76**	**10.33**	**10.27**	**0.63**	**8.72**	**12.32**	**10.63**	**0.47**
	Difference	**-1.01**	**-1.57**	**-1.37**	**+0.17**	**-2.36**	**-2.90**	**-2.85**	**+0.22**

Table 2: PhysFlow cross-evaluation on darkness skin types of MMPD dataset (beats per minute).

5 Conclusions and future work

In this work, we introduce PhysFlow, a novel method for augmenting skin tone diversity in remote heart rate estimation. PhysFlow utilizes conditional normalizing flows to disentangle skin tone information from other facial video content. Physflow is trained end-to-end, allowing simultaneous training of any supervised rPPG approach on original and augmented data during data generation. To train PhysFlow we use a novel bi-dimensional skin tone representation using CIELAB color space, offering adaptability to unbalanced rPPG datasets and skin tone variations. Our cross-dataset experiments on the MMPD dataset using three different data-driven models demonstrate the capability of PhysFlow for skin tone diversity augmenting in any supervised rPPG, showing how our approach significantly reduces heart rate estimation error, particularly in underrepresented skin tone categories, favoring equitable performance across different skin tones. Future work will focus on extending the data augmentation process to include the generation of new appearance data and modification of pulsatile information through a multi-modal conditional normalizing flow approach.

Acknowledgments

This work is partly supported by the eSCANFace project (PID2020-114083GB-I00) funded by the Spanish Ministry of Science and Innovation.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

[2] Paul S Addison, Dominique Jacquel, David MH Foo, and Ulf R Borg. Video-based heart rate monitoring across a range of skin pigmentations during an acute hypoxic challenge. *Journal of clinical monitoring and computing*, 32:871–880, 2018.

[3] Yunhao Ba, Zhen Wang, Kerim Doruk Karinca, Oyku Deniz Bozkurt, and Achuta Kadambi. Style transfer with bio-realistic appearance manipulation for skin-tone inclusive rppg. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2022.

[4] Yannick Benezeth, Peixi Li, Richard Macwan, Keisuke Nakamura, Randy Gomez, and Fan Yang. Remote heart rate variability for emotional state monitoring. In *EMBS Int. Conf. Biomed. Health Inform. BHI*, pages 153–156. IEEE, 2018.

[5] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

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

[7] Joaquim Comas, Adria Ruiz, and Federico Sukno. Efficient remote photoplethysmography with temporal derivative modules and time-shift invariant loss. In *CVPR*, pages 2182–2191, 2022.

[8] Joaquim Comas, Adria Ruiz, and Federico Sukno. Deep adaptative spectral zoom for improved remote heart rate estimation. *arXiv preprint arXiv:2403.06902*, 2024.

[9] Joaquim Comas, Adria Ruiz, and Federico Sukno. Deep pulse-signal magnification for remote heart rate estimation in compressed videos. *arXiv preprint arXiv:2405.02652*, 2024.

[10] Lorenzo Dall'Olio, Nico Curti, Daniel Remondini, Yosef Safi Harb, Folkert W Asselbergs, Gastone Castellani, and Hae-Won Uh. Prediction of vascular aging based on smartphone acquired ppg signals. *Scientific reports*, 10:1–10, 2020.

[11] Ananyananda Dasari, Sakthi Kumar Arul Prakash, László A Jeni, and Conrad S Tucker. Evaluation of biases in remote photoplethysmography methods. *NPJ digital medicine*, 4(1):91, 2021.

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

[13] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *ICLR*, 2015.

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017.

[15] Hannes Ernst, Matthieu Scherpf, Hagen Malberg, and Martin Schmidt. Optimal color channel combination across skin tones for remote heart rate measurement in camera-based photoplethysmography. *Biomedical Signal Processing and Control*, 68:102644, 2021.

[16] Bennett A Fallow, Takashi Tarumi, and Hirofumi Tanaka. Influence of skin type and wavelength on light wave reflectance. *Journal of clinical monitoring and computing*, 27:313–317, 2013.

[17] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[18] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJxgknCcK7.

[19] Anup Kumar Gupta, Rupesh Kumar, Lokendra Birla, and Puneet Gupta. Radiant: Better rppg estimation using signal embeddings and transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4976–4986, 2023.

[20] Cheng-Ju Hsieh, Wei-Hao Chung, and Chiou-Ting Hsu. Augmentation of rppg benchmark datasets: Learning to remove and embed rppg signals via double cycle consistent learning from unpaired facial videos. In *European Conference on Computer Vision*, pages 372–387. Springer, 2022.

[21] Bin Huang, Weihai Chen, Chun-Liang Lin, Chia-Feng Juang, Yuanping Xing, Yanting Wang, and Jianhua Wang. A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks. *Engineering Applications of Artificial Intelligence*, 106:104447, 2021.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[23] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.

[24] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *ECCV*, pages 392–409. Springer, 2020.

[25] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *FedCSIS*, pages 405–410. IEEE, 2011.

[26] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *CVPR*, pages 4264–4271, 2014.

[27] Yuan-kui Li, Yun-Hsuan Lien, and Yu-Shuen Wang. Style-structure disentangled features and normalizing flows for diverse icon colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2022.

[28] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[29] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. Metaphys: few-shot adaptation for non-contact physiological measurement. In *CHIL*, pages 154–163, 2021.

[30] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 5008–5017, 2023.

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

[32] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[33] Daniel McDuff, Miah Wander, Xin Liu, Brian Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. Scamps: Synthetics for camera measurement of physiological signals. *Advances in Neural Information Processing Systems*, 35:3744–3757, 2022.

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

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

[36] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE TIP*, 29:2409–2423, 2019.

[37] Ewa M Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE transactions on intelligent transportation systems*, 23(4):3589–3600, 2020.

[38] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 284–285, 2020.

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

[40] Akshay Paruchuri, Xin Liu, Yulu Pan, Shwetak Patel, Daniel McDuff, and Soumyadip Sengupta. Motion matters: Neural motion transfer for better camera physiological measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5933–5942, 2024.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*, 32, 2019.

[42] Olga Perepelkina, Mikhail Artemyev, Marina Churikova, and Mikhail Grinenko. Heartrack: Convolutional neural network for remote video-based heart rate monitoring. In *CVPRW*, pages 288–289, 2020.

[43] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in non-contact, multiparameter physiological measurements using a webcam. *IEEE Trans. Biomed. Eng.*, 58(1):7–11, 2010.

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

[45] Lev Semenovich Pontryagin. *Mathematical theory of optimal processes*. Routledge, 2018.

[46] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[47] 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.Biomed.Health Inform.*, 25(5):1373–1384, 2021.

[48] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *BMVC*, 2018.

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

[50] Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[51] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007.

[52] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak N. Patel, Daniel J. McDuff, and Xin Liu. Mmpd: Multi-domain mobile video physiology dataset. *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5, 2023. URL https://api.semanticscholar.org/CorpusID:256662570.

[53] William Thong, Przemyslaw Joniak, and Alice Xiang. Beyond skin tone: A multidimensional measure of apparent skin color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4903–4913, 2023.

[54] Yun-Yun Tsou, Yi-An Lee, and Chiou-Ting Hsu. Multi-task learning for simultaneous video generation and remote photoplethysmography estimation. In *ACCV*, 2020.

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

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

[57] Alexander Vilesov, Pradyumna Chari, Adnan Armouti, Anirudh Bindiganavale Harish, Kimaya Kulkarni, Ananya Deoghare, Laleh Jalilian, and Achuta Kadambi. Blending camera and 77 ghz radar sensing for equitable, robust plethysmography. *ACM Trans. Graph.*, 41(4):36–1, 2022.

[58] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[59] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Trans. Biomed. Eng.*, 63(9): 1974–1984, 2015.

[60] 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, 2016.

[61] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *CVPR*, pages 20587–20596, 2022.

[62] Z. Yu, Xiao-Bai Li, and G. Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *BMVC*, 2019.

[63] 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 *ICCV*, pages 151–160, 2019.

[64] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *IJCV*, 131(6):1307–1330, 2023.

[65] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.