

SkinMap: Weighted Full-Body Skin Segmentation for Robust Remote Photoplethysmography

Zahra Maleki* Amirhossein Akbari* Amirhossein Binesh Babak Khalaj

{zahra.maleki, amirhoseinakbari}@ee.sharif.edu, ah.binesh@ce.sharif.edu, khalaj@sharif.edu

Sharif University of Technology

Abstract

Remote photoplethysmography (rPPG) is an innovative method for monitoring heart rate and vital signs by using a simple camera to record a person, as long as any part of their skin is visible. This low-cost, contactless approach helps in remote patient monitoring, emotion analysis, smart vehicle utilization, and more. Over the years, various techniques have been proposed to improve the accuracy of this technology, especially given its sensitivity to lighting and movement. In the unsupervised pipeline, it is necessary to first select skin regions from the video to extract the rPPG signal from the skin color changes. We introduce a novel skin segmentation technique that prioritizes skin regions to enhance the quality of the extracted signal. It can detect areas of skin all over the body, making it more resistant to movement, while removing areas such as the mouth, eyes, and hair that may cause interference. Our model is evaluated on publicly available datasets, and we also present a new dataset, called SYNC-rPPG, to better represent real-world conditions. The results indicate that our model demonstrates a prior ability to capture heartbeats in challenging conditions, such as talking and head rotation, and maintain the mean absolute error (MAE) between predicted and actual heart rates, while other methods fail to do so. In addition, we demonstrate high accuracy in detecting a diverse range of skin tones, making this technique a promising option for real-world applications.

1. Introduction

Remote photoplethysmography (rPPG) is an advanced non-contact technique that enables the measurement of vital physiological signals [60], such as heart rate (HR), respiratory frequency (RF), and heart rate variability (HRV), by analyzing video captured from any part of the skin surface. The light reaching the camera sensor has an AC component that reflects variations in light absorption caused by changes

in arterial blood volume [17, 22, 43]. Unlike traditional contact-based sensors such as PPG or ECG, which require specialized equipment that can be costly and inaccessible [9], rPPG offers a scalable and non-invasive solution for user monitoring. This technology holds significant promise for applications in remote healthcare, emotion analysis, and facial security [54], as it can capture data from any exposed area of the skin without requiring physical proximity.

The extraction of the rPPG signal generally follows unsupervised methods that rely on a structured pipeline [33], where regions of interest (ROIs) on the skin are isolated using computer vision techniques [8, 24, 29, 45, 46, 57, 58, 62, 67]. Then conventional algorithms are applied to convert the RGB signal into the rPPG signal and estimate the heart rate [8, 15, 18, 36, 42, 44, 48, 56, 59]. However, over the past decade, deep learning-based approaches have taken over many parts of processing. These deep learning methods combine conventional techniques with deep learning models or provide end-to-end solutions [12, 16, 26, 32, 34, 39, 41, 49, 52, 64–66]. In the case of end-to-end deep learning methods, the raw video input is processed through various network architectures to directly output the physiological signal.

Unsupervised methods for recovering physiological signals typically begin by selecting the area of the skin that is most likely to produce high-quality signals, with factors such as skin thickness, movement, and lighting playing a significant role [24]. Previous studies have shown that regions such as the cheeks and forehead are often reliable sources of extraction of strong rPPG signals, while areas around the mouth and eyes tend to produce noisy data [5, 23]. In addition, some research has focused on regions such as the hands [2, 51, 53] or neck [3, 6, 25], which can provide high-quality signals due to the abundance of capillaries and arteries in these areas. However, there is a lack of research on dynamic approaches that utilize multiple skin regions across the face and body, enabling more robust signal extraction. This would reduce reliance on specific areas that can be blocked or compromised due to factors such as facial expressions, occlusions, or challenging lighting con-

*Contributed equally to this work

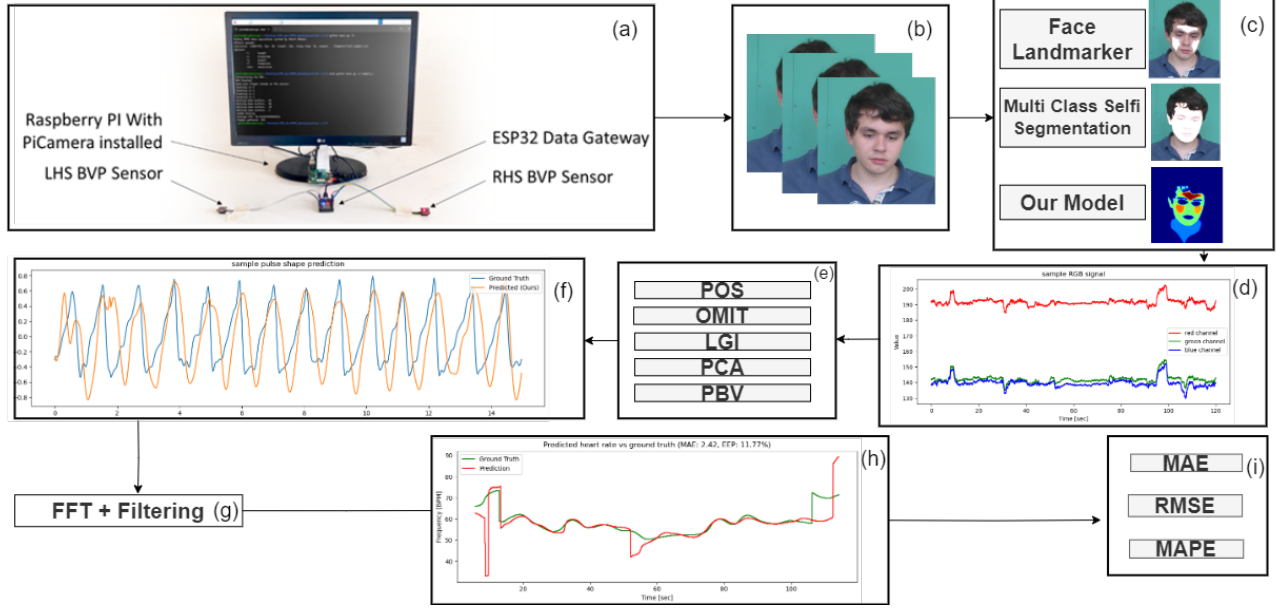


Figure 1. Unsupervised pipeline for heart rate estimation from video. (a) Data acquisition. (b) Video dataset collection synchronized with PPG signals. (c) Skin segmentation or ROIs selection process. (d) RGB signal extraction by averaging skin pixels. (e) rPPG signal extraction methods are applied to the RGB signal. (f) Comparison of the extracted rPPG signal with the reference PPG pulse. (g) Heart rate estimation. (h) Heart rate analysis over time. (i) Evaluation of our estimation using statistical metrics.

ditions [7, 40], ultimately offering a more versatile and reliable approach for extracting rPPG signals.

A key requirement for validating rPPG methods is testing them on realistic datasets. Although existing datasets provide video recordings with ground-truth physiological signals [4, 37, 42, 47, 50], the two are not synchronized. Videos are recorded at a fixed frame rate (FPS), whereas reference signals are sampled at different frequencies, resulting in misaligned timestamps. The interpolation or re-sampling needed to align them can distort the data. Moreover, existing datasets lack real-world complexity, as they do not include head movements, facial expressions, and variations in lighting. Some are restricted to controlled scenarios with plain, high-contrast backgrounds and depend on high-quality cameras, making them not representative of real-world settings.

We introduce a unified pre-processing model for extracting the rPPG signal. Our model generates a dense mask that identifies skin pixels throughout the body, enabling the segmentation of skin even when insufficient facial regions are visible. Furthermore, it generates a mask that provides a pixel-wise weighting to indicate its contribution to signal extraction. In this way, the model suppresses noisy areas and highlights regions with better conditions for signal extraction. Both segmentation and weighting are integrated into a single model, simplifying the pipeline and enabling real-time performance. Overall, the technique provides robustness under real-world conditions and, through its unsu-

pervised pipeline, guarantees strong generalization across datasets. In addition, we introduce a new dataset with diverse real-life scenarios and synchronized sampling rates, which offers a realistic benchmark for evaluating rPPG methods.

Contributions

The contribution of this paper falls into two categories:

- We propose SkinMap, a DeepLabV3-based model that segments facial and body skin, producing a weighted mask for each frame that prioritizes regions with stronger signal quality, without requiring any additional face or landmark detection.
- We present SYNC-rPPG, a new dataset that captures data in four real-world scenarios across 80 samples. Data collection was done using an affordable camera and sensor with the same sampling rates.

2. Related Work

In recent years, multiple approaches have been proposed for extracting heart rates from video cameras, ranging from unsupervised to fully end-to-end supervised models. Unsupervised methods typically involve segmentation of the skin region followed by applying conventional techniques, such as LGI [42], POS [59], CHROM [15], PBV [18], PCA [28], OMIT [8], GREEN [36, 56], to extract the rPPG signal and apply denoising.

2.1. Unsupervised Methods

In facial video-based rPPG, several works have focused on selecting ROIs to achieve the most reliable rPPG signals. [24] conducted experiments on 39 anatomically divided facial regions to identify the most accurate regions for rPPG extraction, highlighting that the cheeks and forehead are more reliable based on their anatomical characteristics. Similarly, in the work [19], the same three main rPPG signal sources (cheeks and forehead) are selected. Many similar studies proposing new unsupervised algorithms use face detection in combination with spatial averaging over the entire skin area as the ROI. In particular, the rPPG-toolbox [33], a comprehensive toolbox for rPPG signal processing, utilizes spatial averaging for each frame in an unsupervised pipeline. Face2PPG [8] has been introduced to stabilize movement and expressions using rigid mesh normalization to extract consistent RGB signals, and combines dynamic multi-region selection with OMIT techniques for accurate heart rate estimation. However, it depends on skin detection and facial segmentation and is limited to facial regions.

As suggested by [27], excluding active areas such as the eyelids and lips helps mitigate motion artifacts, while glasses and hair can contaminate the signal. Narrowing the face area or dividing it into smaller sections without proper skin segmentation increases sensitivity to noise. These issues can be addressed by accurately segmenting the largest possible skin area to improve signal reliability. [45] provides three methods for skin segmentation: two classical color thresholding approaches (Cheref [14], Levelset [55]) and a model (DeepLabV3+ [11]) to create a valid face skin mask for the extraction of the rPPG signal. Another interesting study [2] proposes the use of rPPG signals to prevent spoofing in palm images by converting RGB frames to YCbCr, with skin pixels identified in the Cb-Cr plane. In [3], both the neck and the face are used as ROIs to extract the rPPG signal.

Many rPPG signal extraction algorithms rely on a well-defined, dynamic, weighted skin mask to improve rPPG signal reliability and robustness, and spatially-based techniques often prove to be effective [58]. [38] reviewed the past decade of skin segmentation techniques, including deep and non-deep learning approaches. Many studies use MediaPipe’s 3D face mesh for ROI extraction [21]. The MediaPipe multi-class selfie segmentation model detects face and body skin in real-time [35]. However, it does not differentiate between non-skin areas, such as the eyes, mouth, or glasses, and there is limited published work on this model.

2.2. Supervised Methods

Deep Neural Networks, particularly Convolutional Neural Networks, have gained significant attention in computer vision and signal processing, including healthcare appli-

cations. An end-to-end model directly maps raw video frames to physiological signals, requiring dataset-specific training with ground-truth rPPG signals so that the network learns the entire extraction process without a chain of pre-processing steps such as face detection, skin segmentation, color space transformation, or signal filtering.

DeepPhys [12] is an end-to-end convolutional attention network that estimates heart rate and breathing rate directly from video. The method introduces a motion representation based on normalized frame differences. It uses an appearance-guided attention mechanism that learns soft masks to highlight informative skin regions. EfficientPhys [34] introduces a convolution-based network with a custom normalization module (difference + batchnorm), tensor-shifted convolutions, and self-attention for efficient spatiotemporal modeling. PhysFormer [66] is also an end-to-end video transformer. It introduces temporal difference transformer blocks that combine temporal-difference multi-head self-attention (TD-MHSA) and spatio-temporal feed-forward (ST-FF) modules.

Extraction of rPPG signal relies on extremely subtle, quasi-periodic pixel changes caused by blood volume fluctuations in the skin, which are easily overshadowed by much larger variations from factors such as lighting conditions and motion. Unsupervised approaches tend to offer better generalization in different applications [31]. Although attention mechanisms are powerful in computer vision tasks with rich spatial semantics, in rPPG, they often amplify dataset-specific textures, lighting artifacts, or camera noise [13]. This not only results in a lack of understanding of the underlying physiological mechanisms, but also introduces substantial computational overhead [67]. In contrast, pre-processing models can attenuate illumination and motion artifacts without requiring the use of the signal itself. The pipeline does not require heavy attention modules to learn which pixels to trust based on the signal. If the pre-processing model enforces physiology-driven priors, it makes the extracted features less biased toward dataset-specific appearances.

3. Methodology

As shown in Fig. 1, the unsupervised pipeline for extracting rPPG signals typically involves the following steps:

1. **Dataset Collection:** This step involves collecting video data synchronized with a reference signal and providing the necessary information to read and manage the available or collected dataset.
2. **Video Processing:** A skin segmentation or ROIs selection technique is applied, followed by average or weighted averaging of the pixel values within the skin region to obtain the RGB signal throughout the video.
3. **RGB to rPPG conversion:** Transforming skin color variations into physiological signals using algorithms that

combine the RGB channels, band-pass filtering, and denoising to extract the rPPG signal.

4. Heart Rate Estimation: Heart rate is estimated by performing a frequency analysis on the rPPG signal.
5. Evaluation of results: The estimated Heart rate is evaluated based on various metrics to assess its accuracy, robustness, and reliability.

3.1. Preliminary

In this work, our objective is to develop a robust skin segmentation framework with a pixel-wise weighting mask to improve the extraction of PPG signals from video. For the skin segmentation task, we adopt a variant of the well-established DeepLabV3 architecture with a ResNet-50 backbone, chosen because it represents a state-of-the-art solution for semantic segmentation while remaining computationally efficient. DeepLabV3 incorporates dilated convolutions and an Atrous Spatial Pyramid Pooling (ASPP) module, which together balance fine spatial detail, which is essential for generating accurate pixel-level masks, with contextual understanding that helps separate skin from background under challenging conditions. Since rPPG depends on pixel averaging rather than edge precision, we only require reliable skin separation with pixel-wise weighting. DeepLabV3 provides this balance efficiently, while heavier models add unnecessary complexity without clear benefit.

In the process of training and evaluating our model, we use two state-of-the-art MediaPipe skin segmentation methods. The first is Face Landmark Detection of MediaPipe [21]. A real-time model that predicts 468 3D facial landmarks, employing BlazeFace face detection followed by 3D landmark regression using a MobileNetV2 backbone optimized through transfer learning and Euclidean loss minimization. It can be used to determine the pulse of the cheeks and forehead, which are widely used as ROIs in rPPG extraction pipelines [24]. The second one is Multi-Class Selfie Segmentation of MediaPipe (MCSS). A Vision Transformer-based model designed for real-time segmentation of human subjects. It outputs segmentation masks at $256 \times 256 \times 6$ and $512 \times 512 \times 6$ resolutions, including background, hair, body skin, face skin, clothing, and accessories classifications. However, it does not explicitly differentiate non-skin facial areas, such as the eyes, mouth, or glasses [35]. It is worth noting that all three models are strong performers and capable of real-time processing. A comparison of the results from MediaPipe Landmarker, Multi-Class Selfie Segmentation, and our trained model is illustrated in Fig. 2. As shown, the Landmarker failed to detect the face at harsh angles and when it was not fully visible.

3.2. Proposed Architecture

Traditional algorithms average pixel values from selected areas of facial skin. However, an intelligent system is

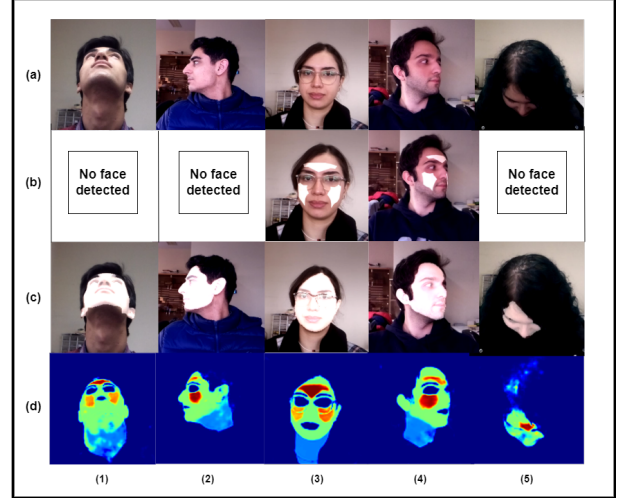


Figure 2. Illustration of our dataset and segmentation results. (a) Frame samples from the rotation task of our dataset. (b) Segmentation results using Face Landmark Detection, where white areas indicate detected ROIs. In some frames, the Landmarker failed to detect a face. (c) Segmentation results using the Multi-Class Selfie Segmentation, where white areas represent detected skin regions. (d) Heat-map visualization of the output of our segmentation model.

needed to automatically segment skin regions, compute their average pixel values, and feed them into the pulse extraction algorithm. As mentioned, we are using the DeepLabV3-ResNet50 architecture [10]. We replace the final layer of the default and auxiliary classifier with a single-channel convolutional layer. A sigmoid activation function is attached to the final layer to confine the output values between 0 and 1. After fine-tuning the model on a large and suitable dataset, we expect it to effectively segment all available skin areas by generating a mask, assigning weights based on the subject’s position and lighting conditions in each frame of the video dataset.

3.3. Photo Dataset Creation

Training our segmentation model requires a diverse dataset of human images under various lighting and environmental conditions. Although there are some public skin segmentation datasets [1, 20, 63], they contain a very limited number of precise samples and are not suitable for our training [38]. To address this, we curated a custom dataset by extracting human images from the COCO dataset [30], which offers a rich variety of real-world scenes with diverse backgrounds.

In the first step of generating our training dataset, we need to extract images containing humans with fully visible faces. To filter the dataset, we utilize the MediaPipe Face Landmark Detector to identify and select relevant images, which feature individuals and groups in various age ranges. Secondly, we need to generate a reference skin mask for

each of these extracted images to train our model later.

As discussed, various regions of the face influence signal quality unequally. The lips and mouth have a different color from the rest of the skin and the amplitude of the heart pulse extracted from these areas is negligible. The eye region is prone to excessive movement, which introduces noise into the signal [24, 27]. Previous studies have shown that the cheeks and forehead exhibit the highest amplitude of the pulse signal [24]. Therefore, prioritizing the segmentation of these regions over other skin areas is essential for accurate signal extraction. For the synthesized photo dataset, we need to systematically assign importance to different facial regions. We classify them into three priority levels:

- Priority 1: Forehead and cheeks, as they provide the highest-quality pulse signals.
- Priority 2: Other facial skin regions, excluding areas around the eyes, eyebrows, and lips.
- Priority 3: Other skin surfaces on the body.

Regions with higher priority should have greater weight in the final skin mask. To achieve this, we introduce a weighting mechanism that considers both the angular orientation of the skin relative to the camera and the assigned priority level. For Priority 1 regions, the weight varies between 4 and 2 depending on the angle, while Priority 2 and 3 regions are assigned fixed weights of 2 and 1, respectively. For example, suppose that the subject is looking directly at the camera; in this case, the specular reflection of the forehead is maximized [61]. We use facial landmarks to estimate the orientation of each region. The weighting function for Priority 1 regions is defined in equation Eq. (1). To explain Eq. (1), we assign a weight P_i to each region based on the angle between the normal vector of the surface and the direction of the camera θ_i . We adjusted the ROIs weighting so that smaller angles receive a higher weight. This results in a weight curve ranging from 2 to 4. The cosine function, used for the effective area, smooths the curve and minimizes noise in challenging poses.

$$P_i = 2(\cos(\theta_i) + 1), \quad \theta_i \leq \frac{\pi}{2} \quad (1)$$

To construct the skin mask for our photo dataset, we first use the MediaPipe Selfie Segmentation model to extract the face and body skin (assigned as a priority 2 and 3 region). Next, we utilize the MediaPipe Face Landmarker to exclude the eyes, eyebrows, and mouth, thereby minimizing noise and also defining priority-1 regions. We then combine these outputs and assign weights based on Eq. (1) and the priorities. After these steps, the mask is normalized to a value between 0 and 1 to maintain consistency with the network output scale. This process is applied to selected human images from the COCO dataset, producing an image-skin mask dataset of 8,000 samples with reliable ground-truth masks for skin segmentation and weighting. This dataset is



Figure 3. Model output on a random sample from the COCO [30] dataset, showcasing its reliability in real-world applications.

subsequently used to train our DeepLabV3 model for accurate and weighted skin segmentation.

We should emphasize that the MediaPipe models are used only for creating the photo dataset. Furthermore, our experiments demonstrate that the final trained model surpasses all of these baseline models in performance while remaining computationally as efficient as they are. Fig. 3 presents sample output of the trained model on randomly selected images from the COCO dataset [30]. These results demonstrate the robustness of the model to skin tone variations.

3.4. Model Training and Heart Rate Estimation

We trained our DeepLabV3-ResNet50 model for 30 epochs, using 90 percent of the data for training and 10 percent for validation. The training was conducted on an RTX 4090 GPU with 20GB of VRAM usage, supported by 198GB DDR5 RAM and an Intel i7-14700K CPU. The training took approximately 4 hours. During training, the training loss steadily decreased and converged, and the validation loss, despite initial fluctuations, trended downward. We stopped at epoch 30, ensuring effective learning and generalization. For each video in the rPPG dataset, a weighted average of the pixels is computed for every frame based on the model’s output. From this, the corresponding RGB signal is generated. Based on the implementation of [33], the RGB signal of each video is processed using commonly used rPPG algorithms. The extracted rPPG signal is then used to estimate heart rate (HR). Heart rate is determined using the Fourier transform (FFT), and band-pass filtering, which extracts frequency components within the physiological heart rate range. The strongest frequency in this range is identified as the heart rate in beats per minute (BPM).

4. Experiments

We evaluated the proposed segmentation model in terms of both accuracy and robustness of rPPG signal extraction by comparing it to state-of-the-art unsupervised and supervised settings. We incorporate several pre-processing techniques into our comparison, including the rPPG-Toolbox pre-processing based on spatial averaging [33], Mediapipe Landmark Detection to isolate the cheek and forehead areas with equal weighting, and Mediapipe Multi-Class Selfie Seg-

mentation (MCSS) to detect facial skin regions. In addition, we include the non-weighted version of our model, which performs full-body skin segmentation, to highlight the benefits of intelligent weighting and the inclusion of body skin. We refer to this model as the full-body model in this section. We mainly employ the POS algorithm [59] for unsupervised pipelines due to its proven superior performance compared to other methods [19]. In the supervised pipeline, we report the results of the pre-trained models of DeepPhys [12], EfficientPhys [34], and PhysFormer [66].

The comparison is conducted on our dataset as well as the UBFC-PHYS dataset [37]. These datasets are selected because they provide sufficient subject diversity and include several real-world scenarios. Since our goal is to evaluate the reliability of these techniques, we focus on identifying the best model that can generalize across scenarios and subjects. To ensure fairness, we utilized the pre-trained versions of the supervised models, which were originally trained on the UBFC-rPPG dataset [4]. This dataset represents a relatively simple setting with stationary subjects and ideal lighting conditions. In contrast, UBFC-PHYS and SYNC-rPPG are significantly more complex.

4.1. Experimental Setup

In this paper, we present a new rPPG dataset. Most available datasets are captured under ideal lighting and environmental conditions, with minimal subject movement, which does not accurately represent real-life applications. Based on these circumstances, we consider it essential to design and implement a dedicated sampling device that ensures the precise, simultaneous acquisition of image and pulse data. An overview of our setup is provided in Fig. 1 (a). We select a Raspberry Pi 4B development board featuring a 64-bit processor clocked at 1.5 GHz and 8 GB of RAM. The system runs the Raspberry Pi OS and utilizes Python for rapid development and seamless integration. For video capture, we employ the Raspberry Pi Camera V2 module, which provides imaging at a resolution of 1280×720 pixels and a frame rate of 30 frames per second (fps). For the data collection, we integrate a laboratory-grade MAX30102 sensor to capture heart pulse data. Furthermore, to improve measurement reliability and reduce errors in pulse signal capture, an additional MAX30102 sensor is integrated to simultaneously acquire pulse data from both hands.

The sensors and camera are precisely synchronized at 30 FPS, which is the maximum achievable rate limited by the intrinsic capabilities of the camera module, thus ensuring a stable high-rate data stream for accurate rPPG analysis. As demonstrated in [8], the PPG signals from fingertip contact-based sensors in publicly available datasets exhibit fluctuations due to finger movement or disconnections, resulting in errors in heart rate estimation. Since we collected the data ourselves, we know the challenges of working with asso-

Attribute	UBFC-rPPG	UBFC-PHYS	SYNC-rPPG
Sample count	50	168	80
Scenarios	rest	rest, talk, exercise	rest, talk, rotation, exercise
Video (FPS)	30	35	30
Sensor (Hz)	60	64	30
Resolution	640×480	1024×1024	1280×720
Heart rate Range (bpm)	~60–80	~60–100	~60–140
Lighting	perfect	perfect	day-light + artificial
Sensor count	1	1	2

Table 1. Comparison of rPPG Datasets

ciated devices. SYNC-rPPG incorporates two sensors, and we use the mean value of their signals.

4.2. Datasets

Our dataset, named SYNC-rPPG, was collected from 20 individuals, with each video lasting 30 seconds. All subjects gave their informed consent for their data to be made publicly available. Each participant was recorded in four different scenarios. In the first scenario, the subject remained calm with no head movement and minimal facial expressions. In the second scenario, the subject was asked to read an emotional passage or discuss an important personal memory. In the third scenario, the subject performed rapid head rotations. In the fourth scenario, the recording took place after exercise, under conditions similar to the first scenario. A comparison between SYNC-rPPG and other datasets used in this study is presented in Tab. 1.

The UBFC-RPPG database [4] utilizes a Logitech C920 HD Pro webcam at 30 frames per second (fps) and 640x480 resolution in uncompressed 8-bit RGB format. A CMS50E pulse oximeter is used to capture PPG data. The database includes 50 videos, each approximately 1 minute long and featuring minimal movement. The UBFC-PHYS dataset [37] includes data from 56 subjects, participating in three tasks: rest, speech, and arithmetic. Participants are filmed and wear a wristband that records BVP and EDA signals.

4.3. Experimental Results

To evaluate the extracted heart rate, we employ five metrics: mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) as introduced in [27], Pearson correlation coefficient (PCC), and signal-to-noise ratio (SNR). The runtime performance of our model on SYNC-rPPG achieved an average processing speed of 211.85 FPS with an average latency of 6.65 ms on an NVIDIA RTX 3060 GPU, demonstrating that the model is capable of real-time operation.

This work focuses on improving the rPPG signal extraction, rather than introducing the most powerful segmentation model; however, we evaluate segmentation accuracy to ensure reliability. We analyze the accuracy and diversity of our model, with and without weights, across different skin tones using the annotations (light, dark, unsure, and nan) provided in [68] for the COCO human image dataset. For

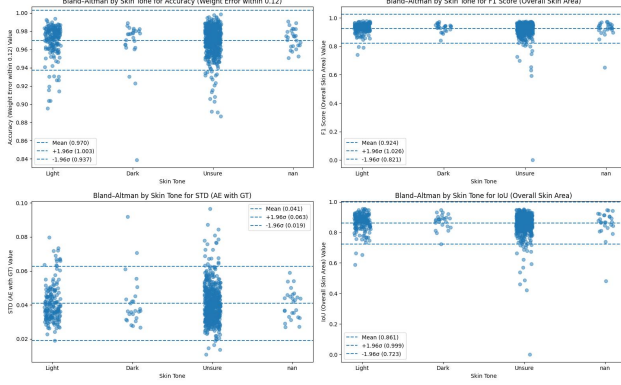


Figure 4. Evaluating skin segmentation by skin tone. Top left: accuracy (Weight Error within 0.12). Top right: F1 score (Overall Skin Area). Bottom left: standard deviation (AE with GT). Bottom right: IoU (Overall Skin Area)

validation, we used 10 percent of our synthesized dataset. The weighted mask achieves a mean accuracy of 0.97 and a mean F1 score of 0.924 for overall skin detection, as shown in Fig. 4. These results indicate that the model is capable of highly accurate and consistent skin segmentation across diverse skin tones, demonstrating strong generalization to population diversity.

One way to evaluate models is by measuring the average number of frames in which they fail to adjust a mask and therefore cannot contribute to the final RGB signal. This problem is more common in ROIs-based models during motion, as shown in Fig. 2, where they do not detect the correct region or the face detector could not locate the face. In our dataset, Face Landmark Detection misses an average of 0.75 frames per video in talking tasks and 118 frames per video in rotation tasks, whereas SkinMap and Multi-Class Selfie Segmentation perform flawlessly. Fig. 5 are extracted signals from one of the samples of UBFC-Phys dataset in the talking scenario. It is evident that our model can reconstruct the true shape and peaks of the signal much more accurately.

As illustrated in Fig. 6, plots (a) to (c) and (g) present the pre-processing results of unsupervised pipelines, while panels (d) to (f) correspond to the supervised settings on the UBFC-Phys dataset. The results indicate that SkinMap achieves the tightest clustering along the diagonal, reflecting more consistent and accurate predictions. Furthermore, SkinMap demonstrates superior and more generalized performance compared to pre-trained supervised models. This finding suggests that supervised models trained on simplified settings fail to outperform an unsupervised pipeline equipped with SkinMap, highlighting their limited reliability in healthcare applications. Fig. 7 presents the results on the SYNC-rPPG dataset; SkinMap pipeline achieves the lowest variance around the diagonal. In addition, we observed that rPPG extraction algorithms, such as POS, struggle

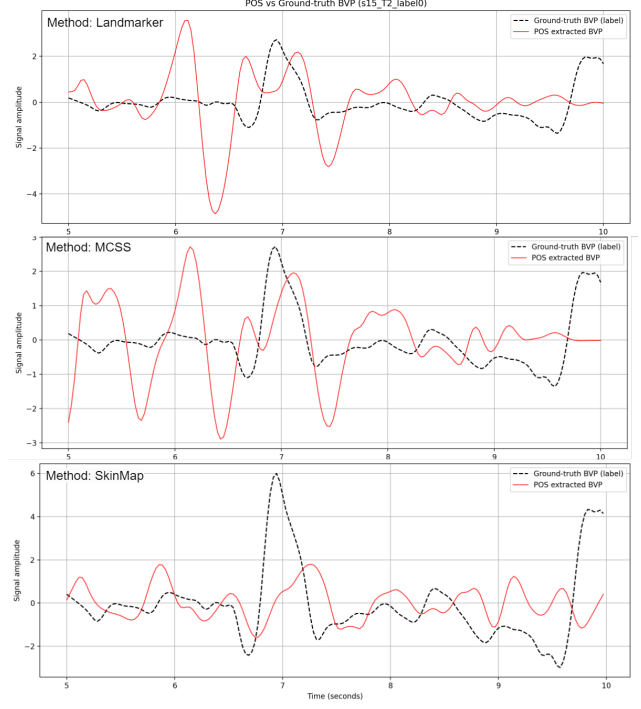


Figure 5. Extracted signals using models, up: Face Landmark Detection, middle: Multi-Class Selfie Segmentation, down: SkinMap.

to reconstruct signals at higher heart rates, while supervised approaches handle them better; however, their performance is highly dependent on the training data.

Tab. 2 gives a detailed comparison of models on the SYNC-rPPG and UBFC-Phys datasets. SkinMap outperforms other models in the head rotation task in SYNC-rPPG and in the talking and arithmetic tasks in UBFC-Phys. It stays competitive in stable scenarios and maintains its accuracy during challenging tasks, showing robustness. Pre-trained supervised models, especially EfficientPhys, excel in the rest scenario, where the samples closely match the UBFC-rPPG dataset on which they were trained. However, they fail in the talking scenario, with MAE nearly twice that of SkinMap. SkinMap’s superior performance comes from its ability to adjust weights during a video to use neck regions when the face is partly hidden. In contrast, Spatial Averaging, Landmarker, and MCSS perform well in stationary tasks but struggle in challenging real-world scenarios, particularly Landmarker. The full-body model, a non-weighted version of SkinMap, performs robustly but still under-performs compared to SkinMap, highlighting the value of the weighting mechanism. This evaluation suggests that while simple ROIs selection or supervised approaches may suffice for static conditions, they are insufficient for real-life applications. For practical use, models must leverage all available sources of information. Skin-

Dataset	Scenario	Metric	Models							
			Spatial Average [33]	Landmarker	MCSS	Full-body	DaapPhys [12]	EfficientPhys [34]	PhysFormer [66]	SkinMap (ours)
SYNC-rPPG	Rest	MAE ↓	11.34±2.48	10.02±2.81	11.51±3.16	10.55±2.15	5.89±2.97	1.85±0.61	11.87±2.48*	6.86±1.62
		RMSE ↓	15.86±9.63	16.08±11.37	18.24±11.91	14.26±8.60	14.54±12.31	3.31±2.27	16.24±9.54	9.97±5.54
		MAPE ↓	15.41±3.59	13.92±4.51	16.41±4.93	14.72±3.44	6.98±3.33	2.39±0.78	16.49±3.89	9.09±2.15
		PCC ↑	-0.04±0.24	0.11±0.23	0.20±0.23	0.039±0.236	0.233±0.229	0.933±0.085	-0.042±0.235	0.497±0.205
		SNR (dB) ↑	-5.49±0.45	-4.82±0.45	-4.75±0.43	-4.74±0.52	-1.66±0.74	-1.44±0.64	-5.21±0.43	-4.20±0.51
	Talking	MAE ↓	13.45±2.39	13.54±2.59	12.83±2.40	12.30±2.04	29.53±4.44*	22.85±4.28	12.39±2.55	12.66±2.17
		RMSE ↓	17.16±9.50	17.81±9.92	16.74±9.31	15.31±8.63	35.59±16.91	29.81±15.64	16.85±9.28	15.95±9.33
		MAPE ↓	15.44±2.72	14.86±2.53	14.35±2.52	13.75±2.11	31.41±4.38	24.65±4.26	13.96±2.97	14.42±2.31
		PCC ↑	0.243±0.229	0.31±0.22	0.32±0.22	0.439±0.212	-0.281±0.226	-0.262±0.227	0.128±0.234	0.242±0.229
		SNR (dB) ↑	-6.57±0.51	-6.09±0.56	-6.34±0.64	-6.21±0.59	-8.35±0.77	-7.24±0.70	-5.95±0.34	-6.15±0.67
	Head Rotation	MAE ↓	14.85±2.10	24.17±3.51	13.80±1.92	13.45±2.47	27.25±2.36*	21.45±3.18	15.21±2.71	11.95±2.13
		RMSE ↓	17.58±8.32	28.82±14.04	16.25±7.91	17.41±9.88	29.22±11.35	25.74±12.80	19.44±10.63	15.29±8.17
		MAPE ↓	19.18±2.93	31.30±4.39	17.74±2.57	17.65±3.40	34.05±2.51	27.80±4.44	20.50±4.24	14.99±2.59
		PCC ↑	-0.028±0.236	0.50±0.20	0.03±0.24	0.170±0.232	-0.072±0.235	-0.335±0.222	0.107±0.234	0.343±0.221
		SNR (dB) ↑	-6.25±0.38	unstable	-6.90±0.47	-5.62±0.44	-9.35±0.60	-7.77±0.48	-6.07±0.41	-5.82±0.29
	After Exercise	MAE ↓	36.47±4.86	29.53±5.33	32.70±5.27	33.05±4.89	45.18±9.34*	37.88±7.69	28.56±5.67	32.96±4.64
		RMSE ↓	42.46±17.89	37.96±18.98	40.31±19.11	39.62±18.05	61.53±31.25	51.16±25.13	38.21±20.06	38.94±17.70
		MAPE ↓	29.05±3.38	22.95±3.80	25.77±3.54	25.82±3.29	34.82±6.73	28.44±5.49	21.98±3.95	26.11±3.07
		PCC ↑	0.241±0.229	0.00±0.24	-0.47±0.21	0.033±0.236	-0.317±0.224	-0.450±0.210	-0.038±0.236	0.312±0.224
		SNR (dB) ↑	-10.77±0.95	-9.64±1.03	-10.18±1.05	-10.02±0.83	-8.65±1.08	-7.34±0.92	-8.92±1.03	-9.49±0.88
UBFC-Phys	Rest	MAE ↓	4.91±1.23	5.13±1.55	5.28±1.52	4.65±1.10	5.57±1.43	3.75±0.98	6.25±1.46*	5.18±1.36
		RMSE ↓	10.13±6.50	12.29±8.53	12.19±8.52	9.19±5.72	11.20±6.57	7.63±5.17	11.87±7.12	10.86±6.95
		MAPE ↓	6.88±1.89	6.83±2.43	7.03±2.41	5.98±1.64	7.47±2.02	5.34±1.52	8.97±2.27	7.27±2.07
		PCC ↑	0.751±0.093	0.577±0.116	0.597±0.113	0.770±0.090	0.718±0.105	0.834±0.083	0.678±0.108	0.717±0.102
		SNR (dB) ↑	0.69±0.71	2.82±0.90	3.06±0.87	2.04±0.91	0.322±0.771	0.71±0.75	-0.75±0.84	0.37±0.80
	Talking	MAE ↓	12.75±1.80	25.00±2.72*	24.85±2.80	16.09±2.01	19.45±2.37	16.91±2.11	18.19±1.95	12.04±1.73
		RMSE ↓	18.20±9.03	31.77±14.53	32.03±13.29	21.64±10.36	25.10±10.69	22.46±10.25	21.98±9.48	17.35±8.90
		MAPE ↓	18.38±3.07	35.87±4.80	35.38±4.44	22.24±2.90	24.31±2.94	23.51±3.42	25.33±3.21	16.82±2.88
		PCC ↑	0.143±0.140	-0.262±0.136	-0.073±0.141	0.193±0.139	-0.062±0.152	-0.126±0.145	0.214±0.158	0.124±0.140
		SNR (dB) ↑	-5.14±0.41	-7.42±0.54	-6.18±0.57	-6.30±0.54	-6.14±0.43	-5.53±0.43	-6.24±0.38	-5.19±0.40
	Arithmetic	MAE ↓	10.31±1.62	22.13±2.51*	20.51±2.27	19.89±2.08	13.18±1.87	12.19±1.89	16.44±2.12	10.12±1.61
		RMSE ↓	15.68±7.74	28.72±12.60	26.34±11.57	24.99±10.63	18.68±9.31	17.99±8.39	21.94±10.02	15.46±7.91
		MAPE ↓	15.06±2.70	35.88±4.84	33.45±4.56	31.37±4.21	16.86±2.26	17.56±3.00	23.29±3.24	14.72±2.70
		PCC ↑	0.325±0.132	-0.166±0.138	0.152±0.138	-0.044±0.140	0.436±0.130	0.248±0.141	-0.024±0.149	0.394±0.129
		SNR (dB) ↑	-4.57±0.36	-6.57±0.60	-6.76±0.59	-6.13±0.54	-4.83±0.44	-4.00±0.47	-5.38±0.34	-4.17±0.36

Table 2. Evaluation results. Unsupervised models: Spatial Average, Landmarker, MCSS, Full-body, and SkinMap. Supervised models: DeepPhys, EfficientPhys, and PhysFormer. The best MAE are highlighted in bold and (*) indicates the worst values

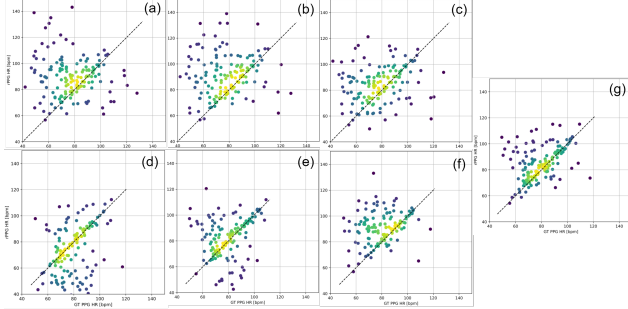


Figure 6. Predicted vs. ground-truth heart rate (BPM) on the UBFC-Phys: (a) Landmarker, (b) MCSS, (c) Full-body, (d) DeepPhys, (e) EfficientPhys, (f) PhysFormer, (g) SkinMap. The dashed line shows perfect prediction.

Map, which does not require additional face detection or extensive pre-processing, offers a robust solution by prioritizing skin pixels, therefore filtering out sudden environmental noise from body movements and lighting variations, thereby enhancing signal quality.

5. Conclusions and Future Works

This study presents SkinMap, a full-body skin segmentation model that utilizes all available skin regions and as-

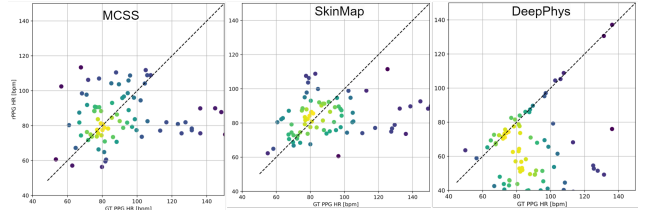


Figure 7. Predicted vs. ground-truth heart rate (BPM) scatter plot on the SYNC-rPPG dataset. Left: MCSS, middle: SkinMap, and right: DeepPhys.

signs optimized pixel-wise weights for unsupervised rPPG signal extraction pipelines. A new video-PPG dataset was collected at a uniform sampling rate across four real-world scenarios. SkinMap was trained using a synthesized dataset of image-mask pairs. Experimental results indicate that SkinMap accurately detects skin regions and generalizes effectively across diverse skin tones, while distinguishing non-skin areas such as accessories and hair. Compared with existing skin segmentation, ROIs selection approaches, and state-of-the-art supervised methods, SkinMap demonstrates superior performance and robustness in complex, dynamic scenarios. Future work will focus on reducing the model size for deployment on mobile devices.

References

- [1] Abdallah S. Abdallah, Mohamad Abou El-Nasr, and Amos Lynn Abbott. A new color image database for benchmarking of automatic face detection and human skin segmentation techniques. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 1:3769–3773, 2007. 4
- [2] Byeongseon An, Hyeji Lim, and Eui Lee. Fake biometric detection based on photoplethysmography extracted from short hand videos. *Electronics*, 12:3605, 2023. 1, 3
- [3] Byeong An, Hyeji Lim, Hyeon Seong, and Eui Chul Lee. Facial and neck region analysis for deepfake detection using remote photoplethysmography signal similarity. *IET Biometrics*, 2024, 2024. 1, 3
- [4] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 92:35–42, 2017. 2, 6
- [5] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro D’Amelio, Giuliano Grossi, and Raffaella Lanzarotti. Enhancing rppg pulse-signal recovery by facial sampling and psd clustering. *Biomedical Signal Processing and Control*, 101:107158, 2025. 1
- [6] Meiyun Cao, Gennadi Saiko, Timothy Burton, and Alexandre Douplik. Remote physiological monitoring of neck blood vessels. page 26, 2023. 1
- [7] Mingyue Cao, Xu Cheng, Xingyu Liu, Yan Jiang, Hao Yu, and Jingang Shi. St-phys: Unsupervised spatio-temporal contrastive remote physiological measurement. *IEEE Journal of Biomedical and Health Informatics*, 28(8):4613–4624, 2024. 2
- [8] Constantino Álvarez Casado and Miguel Bordallo López. Face2ppg: An unsupervised pipeline for blood volume pulse extraction from faces. *IEEE Journal of Biomedical and Health Informatics*, 27(11):5530–5541, 2023. 1, 2, 3, 6
- [9] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, 4(4):195–202, 2018. 1
- [10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 4
- [11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018. 3
- [12] Weixuan ‘Vincent’ Chen and Daniel J. McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. *ArXiv*, abs/1805.07888, 2018. 1, 3, 6, 8
- [13] Chun-Hong Cheng, Kwan-Long Wong, Jing-Wei Chin, Tsz-Tai Chan, and Richard H. Y. So. Deep learning methods for remote heart rate measurement: A review and future research agenda. *Sensors*, 21(18), 2021. 3
- [14] Djamila Dahmani, Mehdi Cheref, and Slimane Larabi. Zero-sum game theory model for segmenting skin regions. *Image and Vision Computing*, 99:103925, 2020. 3
- [15] Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2
- [16] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3975–3984, 2021. 1
- [17] Amogh Gudi, Marian Bittner, and Jan van Gemert. Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation. *Applied Sciences*, 10(23), 2020. 1
- [18] G. Haan, de and A.J. Leest, van. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological Measurement*, 35(9):1913–1926, 2014. 1, 2
- [19] Fridolin Haugg, Mohamed Elgendi, and Carlo Menon. Effectiveness of remote ppg construction methods: A preliminary analysis. *Bioengineering*, 9:485, 2022. 3, 6
- [20] Lei Huang, Tian Xia, Yongdong Zhang, and Shouxun Lin. Human skin detection in images by msr analysis. *2011 18th IEEE International Conference on Image Processing*, pages 1257–1260, 2011. 4
- [21] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *ArXiv*, abs/1907.06724, 2019. 3, 4
- [22] Fatema-Tuz-Zohra Khanam, Ali Abdulelah Al-Naji, and Javan Chahl. Remote monitoring of vital signs in diverse non- clinical and clinical scenarios using computer vision systems: A review. *Applied Sciences*, 9:4474, 2019. 1
- [23] Adam Kiddle, Helen Barham, Simon Wegerif, and Connie Petronzio. Dynamic region of interest selection in remote photoplethysmography: Proof-of-concept study. *JMIR Formative Research*, 7, 2023. 1
- [24] Dae-Yeol Kim, Kwangkee Lee, and Chae-Bong Sohn. Assessment of roi selection for facial video-based rppg. *Sensors*, 21(23), 2021. 1, 3, 4, 5
- [25] Benjamin Kossack, Eric Wisotzky, Anna Hilsmann, and Peter Eisert. Local remote photoplethysmography signal analysis for application in presentation attack detection. 2019. 1
- [26] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision*, 2020. 1
- [27] Kunyoung Lee, Jaemu Oh, Hojoon You, and Eui Chul Lee. Improving remote photoplethysmography performance through deep-learning-based real-time skin segmentation network. *Electronics*, 12:3729, 2023. 3, 5, 6
- [28] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam – a non-contact method for evaluating cardiac activity. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 405–410, 2011. 2
- [29] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam – a non-contact method for evaluating cardiac activity.

- In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 405–410, 2011. 1
- [30] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 4, 5
- [31] Tianqi Liu, Hanguang Xiao, Yisha Sun, Yulin Li, Shiyi Zhao, Zhenyu Yi, and Aohui Zhao. Style-rppg: Exploration and analysis of style transfer in unsupervised remote physiological measurement. *Expert Systems with Applications*, 269: 126310, 2025. 3
- [32] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement, 2020. 1
- [33] Xin Liu, Xiaoyu Zhang, Girish Narayanswamy, Yuzhe Zhang, Yuntao Wang, Shwetak Patel, and Daniel McDuff. Deep remote ppg toolbox. *arXiv preprint arXiv:2210.00716*, 2022. 1, 3, 5, 8
- [34] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4997–5006, 2023. 1, 3, 6, 8
- [35] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubaweja, Michael Hays, Fan Zhang, Chuoling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. 3, 4
- [36] Luis Francisco Corral Martínez, Gonzalo Páez, and Marija Strojnik. Optimal wavelength selection for noncontact reflection photoplethysmography. In *The International Commission for Optics*, 2011. 1, 2
- [37] Rita Meziati, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, and Fan Yang. Ubf-cphys, 2021. 2, 6
- [38] Loris Nanni, Andrea Loreggia, Alessandra Lumini, and Alberto Dorizza. A standardized approach for skin detection: Analysis of the literature and case studies. *Journal of Imaging*, 9(2), 2023. 3, 4
- [39] Girish Narayanswamy, Yujia Liu, Yuzhe Yang, Chengqian Ma, Xin Liu, Daniel McDuff, and Shwetak N. Patel. Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 7899–7909. IEEE, 2024. 1
- [40] Nhi Nguyen, Le Nguyen, Honghan Li, Miguel Bordallo López, and Constantino Álvarez Casado. Evaluation of video-based rppg in challenging environments: Artifact mitigation and network resilience. *Computers in Biology and Medicine*, 179:108873, 2024. 2
- [41] Xuesong Niu, S. Shan, Hu Han, and Xilin Chen. Rhythmet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 1
- [42] Christian S. Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1335–13358, 2018. 1, 2
- [43] Ming-Zher Poh, Daniel McDuff, and Rosalind Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58:7–11, 2010. 1
- [44] Ming-Zher Poh, Daniel McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18 10:10762–74, 2010. 1
- [45] Matthieu Scherpf, Hannes Ernst, Leo Misera, Hagen Malberg, and Martin Schmidt. Skin segmentation for imaging photoplethysmography using a specialized deep learning approach. In *2021 Computing in Cardiology (CinC)*, pages 1–4, 2021. 1, 3
- [46] Yi Sheng, Wu Zeng, Qiuyu Hu, Weihua Ou, Yuxuan Xie, and Jie Li. An improved approach to the performance of remote photoplethysmography. *Computers, Materials and Continua*, 73(2):2773–2783, 2022. 1
- [47] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. 2
- [48] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021. 1
- [49] Radim Spetlik, Jan Cech, Vojtěch Franc, and Jiri Matas. Visual heart rate estimation with convolutional neural network. 2018. 1
- [50] R. Stricker, S. Müller, and H.-M. Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man 2014)*, pages 1056–1062, Edinburgh, Scotland, UK, 2014. IEEE. 2
- [51] Lu Sun, Liting Wang, Wentao Shen, Changsong Liu, and Fengshan Bai. Robust rppg method based on reference signal envelope to improve wave morphology. *Electronics*, 12(13), 2023. 1
- [52] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. page 492–510, Berlin, Heidelberg, 2022. Springer-Verlag. 1
- [53] Xin Tian, Chau-Wai Wong, Sushant Ranadive, and Min Wu. A multi-channel ratio-of-ratios method for noncontact hand video based spo2 monitoring using smartphone cameras, 2021. 1
- [54] Akito Tohma, Maho Nishikawa, Takuya Hashimoto, Yoichi Yamazaki, and Guanghao Sun. Evaluation of remote photoplethysmography measurement conditions toward telemedicine applications. *Sensors*, 21(24), 2021. 1
- [55] Alexander Trumpp, Stefan Rasche, Daniel Wedekind, Martin Schmidt, Thomas Waldow, Frederik Gaetjen, Katrin Plötze,

- Hagen Malberg, Klaus Matschke, and Sebastian Zaunseder. Skin detection and tracking for camera-based photoplethysmography using a bayesian classifier and level set segmentation. In *Bildverarbeitung für die Medizin*, 2017. 3
- [56] Wim Verkruysse, Lars Svaasand, and John Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16:21434–21445, 2008. 1, 2
- [57] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, 2008. 1
- [58] W. Wang, Sander Stuijk, and Gerard Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on bio-medical engineering*, 0: 1, 2015. 1, 3
- [59] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 1, 2, 6
- [60] Wenjin Wang, Albertus C. den Brinker, and Gerard de Haan. Single-element remote-ppg. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2019. 1
- [61] Kwan Long Wong, Jing Wei Chin, Tsz Tai Chan, Ismoil Odinaev, Kristian Suhartono, Kang Tianqu, and Richard Hau Yue So. Optimising rppg signal extraction by exploiting facial surface orientation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 2164–2170. IEEE, 2022. 5
- [62] Yuting Yang, Chenbin Liu, Hui Yu, Dangdang Shao, Francis Tsow, and Nongjian Tao. Motion robust remote photoplethysmography in cielab color space. *Journal of Biomedical Optics*, 21(11):117001, 2016. 1
- [63] Hojoon You, Kunyoung Lee, Jaemu Oh, and Eui Chul Lee. Efficient and low color information dependency skin segmentation model. *Mathematics*, 11(9), 2023. 4
- [64] 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. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 151–160, 2019. 1
- [65] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, PP:1–1, 2020.
- [66] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H. S. Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4176–4186, 2021. 1, 3, 6, 8
- [67] Qi Zhan, Wenjin Wang, and Gerard Haan. Analysis of cnn-based remote-ppg to understand limitations and sensitivities. *Biomedical Optics Express*, 11, 2020. 1, 3
- [68] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. pages 14810–14820. Institute of Electrical and Electronics Engineers Inc., 2021. Publisher Copyright: © 2021 IEEE.; 18th IEEE/CVF International Conference on Computer Vision, ICCV 2021 ; Conference date: 11-10-2021 Through 17-10-2021. 6