engineering proceedings

MDPI

Proceeding Paper

FusionPhys: A Multi-Source rPPG Method for Estimating Heart Rate from Facial Videos †

Tolotra Frédéric Rabarijaona [1] and Ghazal Rouhafzay [1,2,*]

[1] Department of Computer Science, Université de Moncton, Moncton, NB E1A 3E9, Canada; etr7975@umoncton.ca
[2] Perception, Robotique et Intelligence Machine (PRIME) Research Laboratory, Université de Moncton, Moncton, NB E1A 3E9, Canada
* Correspondence: ghazal.rouhafzay@umoncton.ca
† Presented at the 13th International Electronic Conference on Sensors and Applications, 18-20 November 2026; Available online: https://sciforum.net/event/ECSA-13.

Abstract

Remote photoplethysmography (rPPG) estimates cardiovascular pulse information from facial videos without the use of contact sensors. Considerable research effort has been devoted over the past decade to improve the accuracy of this technology using both classical color-based estimators and learned models. However, most of them remain vulnerable in real-world conditions, including subject movements, lighting variations, or low image quality. To overcome these limitations, we propose FusionPhys, an architecture that exploits three complementary sources of information extracted from facial video: the raw appearance of the face, motion variations between successive frames, and pulse indices provided by classical estimators. The model is trained to predict the standardized first derivative of PPG, from which heart rate is estimated by the dominant spectral peak. Two subject-independent evaluations were performed on UBFC-rPPG. On Dataset 1, leave-one-subject-out evaluation gave heart-rate MAEs of 1.02 BPM for POS, 1.38 BPM for FusionPhys, 2.60 BPM for DeepPhys, and 6.60 BPM for Green. On the larger Dataset 2, using 30/6/6 train/validation/test subjects, MAEs were 1.93, 2.37, 14.17, and 18.45 BPM, respectively. FusionPhys therefore substantially improved upon DeepPhys while remaining close to the strongest classical baseline, and it provided the lowest waveform-derivative RMSE among the evaluated methods.

Keywords: remote photoplethysmography; rPPG; DeepPhys; POS; Green method; multi-source fusion; heart-rate estimation; physiological sensing

1. Introduction

Remote photoplethysmography (rPPG) estimates cardiovascular pulse information from subtle, periodic changes in skin reflectance captured by conventional cameras. The technique has found its application in contactless physiological monitoring for telemedicine, driver monitoring, and human–computer interaction, as well as in ambient health sensing. Early work demonstrated that pulse information can be recovered from ordinary RGB recordings under ambient illumination [1]. However, an rPPG signal has a small amplitude relative to changes caused by head motion, facial expression, shadows, camera auto-exposure, and illumination variation, all of which can make the extraction of the signal challenging.

Classical approaches rely on explicit colorimetric assumptions. The Green method [1] uses the average green-channel intensity, relying on the fact that hemoglobin absorption

Check for updates

Academic Editor: Jean-Marc Laheurte

Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

produces a strong pulsatile component in this spectral band. More advanced methods such as plane-orthogonal-to-skin (POS) [2] project normalized RGB traces onto a chrominance plane designed to suppress intensity variation. The POS [2] method assumes that the skin-tone direction remains approximately stable over short temporal windows, and that such projection can help emphasize pulsatile chrominance changes while reducing common-mode intensity variations. Classial methods are computationally light and interpretable but can degrade when their assumptions are violated.

Deep learning provides an alternative by learning spatial and temporal representations directly from video. DeepPhys [3] introduced a two-stream convolutional attention network in which normalized frame differences encode motion while appearance features generate spatial attention masks [3]. Building on this direction, PhysNet [4] moved toward end-to-end spatiotemporal modeling of pulse dynamics from video sequences, while TS-CAN [5] combined temporal-shift operations with attention mechanisms to better capture temporally informative changes in facial regions.

Deep learning methods can learn rich spatiotemporal representations, but their performance depends strongly on the diversity and size of the training data, and they may not explicitly exploit well-established colorimetric priors that classical rPPG methods encode by design. This motivates FusionPhys, which combines these complementary strengths rather than requiring a neural network to rediscover all colorimetric priors from limited training data. In the proposed architecture, Green [1] and POS [2] methods are treated as independent learned branches, each processed separately before feature-level concatenation with the DeepPhys [3] representation. The standardized derivative of the synchronized pulse-oximeter PPG serves as the supervision target, while all model inputs are derived exclusively from the facial video.

This study therefore investigates whether combining classical color-based rPPG priors with learned representations can improve pulse estimation across different acquisition conditions. FusionPhys is evaluated using subject-independent protocols on UBFC-rPPG [6] Dataset 1 and Dataset 2 and compared with DeepPhys [3], Green [1], and POS [2] on identical held-out subjects. Following the emphasis on standardized and reproducible evaluation promoted by frameworks such as rPPG-Toolbox (commit b7500b8, 15 September 2025) [7], the remainder of the paper describes the proposed architecture and experimental methodology, presents the comparative results, and discusses the main findings and limitations.

2. Framework and Methods

Figure 1 summarizes the overall architecture of FusionPhys. Further discussions about the design are provided in this section.

2.1. Classical rPPG Estimators

We explore both Green [1] and POS [2] as complementary classical estimators because they represent two distinct strategies for extracting rPPG information from facial color variations. Incorporating both, therefore, exposes the fusion model to complementary physiological and colorimetric cues rather than to two variants of the same estimator.

Let the spatially averaged facial-region color trace be $\mathbf{C}(t) = [R(t), G(t), B(t)]^\top$, where the RGB values are averaged over the square facial ROI obtained by expanding the detected face bounding box by a factor of 1.35 and resizing it to 36 × 36 pixels. The Green baseline uses $G(t)$ as the raw pulse carrier, followed by pulse-band filtering.

https://doi.org/10.3390/engproc2026163063

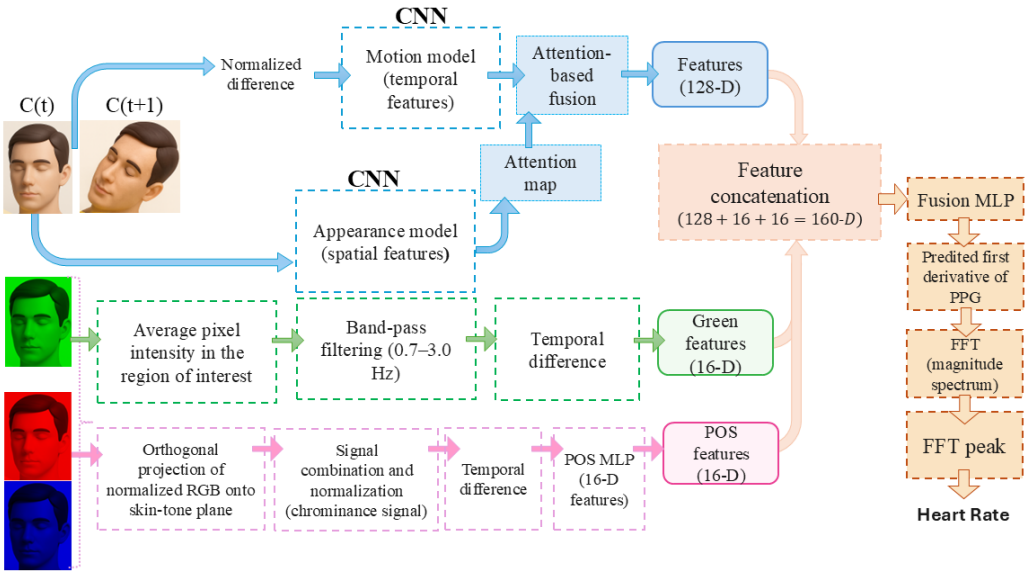

Figure 1. Overview of the proposed FusionPhys architecture. The blue branch represents DeepPhys-based motion and appearance processing, while the green and pink branches correspond to the Green and POS rPPG estimators, respectively. The orange blocks represent feature fusion and heart-rate estimation. Arrows indicate the direction of data flow between processing stages.

POS [2] takes a different approach from the Green [1] method by exploiting relationships among all three color channels. After normalizing the RGB traces by their temporal means over a short sliding window, the signal is projected onto a plane orthogonal to the skin-tone direction. Following Wang et al. [2], this projection is performed using

$$\mathbf{P}_{\mathrm{POS}} = \begin{bmatrix} 0 & 1 & -1 \\ -2 & 1 & 1 \end{bmatrix}. \tag{1}$$

The resulting projection produces two chrominance components:

$$\begin{bmatrix} S_1(t) \\ S_2(t) \end{bmatrix} = \mathbf{P}_{\mathrm{POS}}\, \mathbf{C}_n(t), \tag{2}$$

where $S_1(t)$ and $S_2(t)$ correspond to the linear combinations $G - B$ and $-2R + G + B$, respectively. These components are intended to attenuate common intensity variations while retaining pulse-related color modulation. The POS algorithm then adaptively combines them according to their temporal variability:

$$h(t) = S_1(t) + \alpha S_2(t), \qquad \alpha = \frac{\sigma(S_1)}{\sigma(S_2)}, \tag{3}$$

where $\sigma(\cdot)$ denotes the temporal standard deviation computed over the analysis window [2]. In our implementation, the normalization, projection, and combination are performed over an approximately 1.6 s sliding window.

As illustrated in Figure 1, the computed Green [1] and POS [2] traces are each treated as one-dimensional physiological signals. Both traces are band-pass-filtered between 0.7 and 3.0 Hz (42–180 BPM), standardized, converted to first temporal differences, and then temporally aligned with the DeepPhys [3] motion samples.

The Green [1] and POS [2] streams are then passed through two separate multilayer perceptrons (MLPs), with no parameter sharing. Each branch maps a scalar input through two 16-unit fully connected layers with hyperbolic-tangent activations and dropout. This produces two 16-dimensional feature vectors that remain independent until the final fusion stage.

2.2. DeepPhys Branch

Following DeepPhys [3], consecutive RGB face crops are used to construct a normalized motion representation

$$\mathbf{D}_t = \frac{\mathbf{I}_t - \mathbf{I}_{t-1}}{\mathbf{I}_t + \mathbf{I}_{t-1} + \varepsilon} \tag{4}$$

where $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$ represent consecutive image frames, $\mathbf{D}_t$ denotes the normalized frame-difference tensor, and $\varepsilon > 0$ is a small numerical stabilizer added to the denominator to prevent division by zero. All arithmetic operations are performed element-wise. The resulting motion values are clipped to within three standard deviations of their mean and subsequently standardized per recording. The appearance input consists of the current RGB face crop, standardized to zero mean and unit variance.

The DeepPhys [3] branch contains parallel motion and appearance convolutional streams. Each uses 3 × 3 convolutions with 32 filters in the first block and 64 filters in the second block. The appearance stream generates sigmoid attention maps that are L_1 normalized and multiplicatively applied to the corresponding motion features. Average pooling and dropout are used between blocks. The final DeepPhys [3] representation is projected to a 128-dimensional learned feature vector.

2.3. Feature-Level Fusion

The three learned representations are concatenated only after independent processing:

$$\mathbf{z}_{\text{fusion}} = \left[\mathbf{z}_{\text{DeepPhys}} \parallel \mathbf{z}_{\text{Green}} \parallel \mathbf{z}_{\text{POS}}\right] \in \mathbb{R}^{160}. \tag{5}$$

A fusion head maps the 160-dimensional concatenated representation to 128 hidden units, followed by a single linear output, with a hyperbolic-tangent activation and dropout applied in the hidden layer.

The target is the synchronized reference PPG measured by the pulse oximeter. The PPG is interpolated to the video timestamps, and the standardized first temporal difference $y_t = \mathcal{Z}(p(t) - p(t-1))$ is used for supervision, consistent with the derivative-based formulation of DeepPhys [3]. Here, $\mathcal{Z}(\cdot)$ denotes standardization. The loss is the mean squared error (MSE) between the predicted and reference PPG derivatives.

2.4. Video Preprocessing

For each video, facial regions of interest (ROIs) are localized using the MediaPipe face detection framework, based on the BlazeFace architecture [8], with a minimum detection confidence of 0.5. Face detection is refreshed every 30 frames, corresponding to approximately 1 s at the 30 fps acquisition rate, and the most recent valid bounding box is reused between detections. The detected bounding box is expanded by a factor of 1.35 to form a square facial ROI, which is then resized to 36 × 36 RGB pixels. Channel-specific computations are performed after explicit BGR-to-RGB conversion.

For each recording, the mean RGB value within the facial ROI is extracted at every frame and used to compute the Green and POS traces. The appearance input is standardized independently for each frame, while the motion input is constructed using Equation (4), clipped to ±3 standard deviations around the recording mean, and then standardized. The Green and POS traces are band-pass-filtered between 0.7 and 3.0 Hz using a third-order Butterworth filter applied with zero-phase filtering. The reference PPG is interpolated to the video frame timestamps, and its first temporal difference is computed. Consequently, the DeepPhys motion samples, the differentiated Green and POS traces, and the reference PPG derivative all correspond to the same consecutive-frame interval, allowing sample-wise temporal alignment.

https://doi.org/10.3390/engproc2026163063

2.5. UBFC-rPPG Datasets [6]

Experiments were conducted using the publicly available UBFC-rPPG benchmark [6]. The dataset comprises two subsets acquired under different experimental conditions. Facial videos were recorded using a Logitech C920 HD Pro webcam (Logitech International S.A., Lausanne, Switzerland) at approximately 30 fps and a spatial resolution of 640 × 480 pixels, in an uncompressed 8-bit RGB format. Reference PPG signals were acquired using a CMS50E finger-clip pulse oximeter (Contec Medical Systems Co., Ltd., Qinhuangdao, China). Participants were seated approximately 1 m from the camera. In Dataset 1, participants were instructed to remain relatively still, whereas Dataset 2 involved a time-constrained mathematical task intended to induce heart-rate variations under simulated human-computer interaction conditions.

Dataset 1 contains eight recordings (approximately 16,500 frames). Subjects were primarily asked to remain seated, although several sequences contain appreciable movement. Because a fixed 70/20/10 split would leave only one test subject and provide an unstable estimate, Dataset 1 is evaluated by leave-one-subject-out (LOSO) cross-validation. Each of the eight folds uses six subjects for training, one for validation, and one for testing; every subject serves as the test subject exactly once.

Dataset 2 contains 42 recordings acquired while participants performed a time-sensitive mathematical task designed to induce heart-rate variation during a human–computer interaction scenario. For this experiment, a reproducible subject-level random split assigned 30 subjects to training, six to validation, and six to testing. The held-out test subjects are S3, S8, S10, S16, S25, and S26. The same split is used for DeepPhys [3], FusionPhys, Green [1], and POS [2].

2.6. Training and Model Selection

Both DeepPhys and FusionPhys are optimized via Adam using a learning rate of 10^{-3} and MSE loss. Training is run for 29 epochs. For each experiment, the checkpoint with the lowest validation MSE is retained for final testing. The classical Green [1] and POS [2] baselines require no learned fitting and are evaluated directly on the corresponding held-out subjects.

The original 36 × 36 DeepPhys [3] convolutional configuration is retained: 3 × 3 kernels, 32 then 64 filters, 2 × 2 average pooling, and dropout rates of 0.25 and 0.50.

2.7. Evaluation Metrics

The principal outcome is heart-rate error. For each held-out recording, heart rate is estimated from the dominant spectral peak of the standardized PPG derivative within 0.7–3.0 Hz (42–180 BPM). Differentiation preserves the fundamental frequency, allowing heart rate to be obtained directly from the derivative signal. The same spectral procedure is applied to the reference PPG derivative to obtain the reference heart rate, with the peak frequency converted to beats per minute via multiplication by 60. The heart-rate trace provided in the UBFC-rPPG ground-truth file is not used for evaluation.

We report

$$\mathrm{MAE}_{\mathrm{HR}} = \frac{1}{N}\sum_{i=1}^{N} |HR_i - \widehat{HR}_i| \tag{6}$$

and

$$\mathrm{RMSE}_{\mathrm{HR}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (HR_i - \widehat{HR}_i)^2}. \tag{7}$$

Because FusionPhys is also intended to estimate a continuous pulse representation, we additionally report the RMSE and Pearson correlation r between the predicted and reference standardized PPG derivatives. These metrics can help determine if a method

can identify the correct dominant frequency while still possessing a poor waveform phase or morphology.

3. Results and Discussion

3.1. Dataset 1: LOSO Evaluation

Table 1 reports the LOSO results on the eight-recording Dataset 1. POS [2] achieves the lowest HR MAE at 1.02 BPM, while FusionPhys follows at 1.38 BPM. Importantly, FusionPhys improves substantially upon DeepPhys [3] (2.60 BPM) and Green [1] (6.60 BPM). Relative to DeepPhys [3], the FusionPhys HR MAE is reduced by approximately 47%. FusionPhys also provides the lowest waveform-derivative RMSE (0.974) and the highest mean waveform correlation ($r = 0.237$), whereas the POS [2] waveform correlation is negative despite its excellent frequency-based HR estimate.

We further evaluate the method on the larger Dataset 2 to mitigate the risk of overfitting, since each Dataset 1 training fold contains only six subjects.

Table 1. HR MAE reported as the mean ± SD across subject-wise absolute errors.

Method	HR MAE (BPM)	HR RMSE (BPM)	Standardized ΔPPG RMSE	ΔPPG Pearson r
POS	**1.02 ± 1.61**	**1.90**	1.583	−0.256
FusionPhys	1.38 ± 1.77	2.24	**0.974**	**0.237**
DeepPhys	2.60 ± 2.57	3.65	1.027	0.082
Green	6.60 ± 7.07	9.67	1.485	−0.103

Note: Bold values indicate the lowest error or highest Pearson correlation for each metric.

3.2. Dataset 2: Subject-Independent Evaluation

Table 2 summarizes performance on the six held-out subjects from the 42-subject Dataset 2 experiment. POS [2] again achieves the lowest HR MAE (1.93 BPM), with FusionPhys close behind (2.37 BPM). In contrast, DeepPhys [3] and Green [1] show much larger mean errors of 14.17 and 18.45 BPM, respectively. FusionPhys reduces the HR MAE by approximately 83% relative to DeepPhys [3].

The waveform-level results provide complementary information. FusionPhys achieves the lowest waveform-derivative RMSE (0.976) and the highest mean waveform Pearson correlation ($r = 0.217$). DeepPhys reaches $r = 0.129$, whereas Green [1] and POS [2] have negative average correlations. Thus, although POS [2] slightly outperforms FusionPhys on the dominant-frequency HR MAE, FusionPhys provides a continuous output that is more aligned with the reference derivative waveform.

Figure 2 shows the subject-wise absolute HR errors for the four methods on Dataset 2.

Across both protocols, FusionPhys consistently improves upon the DeepPhys-only network, with a substantially larger relative improvement on Dataset 2. This pattern suggests that video-derived colorimetric priors may help stabilize a learned rPPG model when the purely learned branch encounters difficult subjects.

At the same time, POS [2] remains a strong baseline and achieves the best HR MAE on both datasets. The present results therefore do not support a claim that FusionPhys universally outperforms classical methods. However, they show that the proposed fusion closes most of the gap between a learned DeepPhys [3] baseline and a robust analytical estimator, while producing a waveform derivative that is more strongly correlated with the synchronized PPG derivative than the Green [1] and POS [2] traces in these experiments.

Table 2. Dataset 2 performance on the common six-subject held-out test set. The HR MAE is reported as the mean ± SD across subject-wise absolute errors.

Method	HR MAE (BPM)	HR RMSE (BPM)	Standardized ΔPPG RMSE	ΔPPG Pearson r
POS	**1.93 ± 3.17**	**3.71**	1.530	−0.175
FusionPhys	2.37 ± 3.29	4.06	**0.976**	**0.217**
DeepPhys	14.17 ± 28.07	31.44	0.992	0.129
Green	18.45 ± 22.93	29.43	1.470	−0.083

Note: Bold values indicate the lowest error or highest Pearson correlation for each metric.

The subject-level results also illustrate the limitations of relying on mean error alone, as isolated large failures substantially affect the aggregate performance of DeepPhys [3] and Green [1]. In contrast, FusionPhys and POS [2] show more consistent errors across the six held-out subjects.

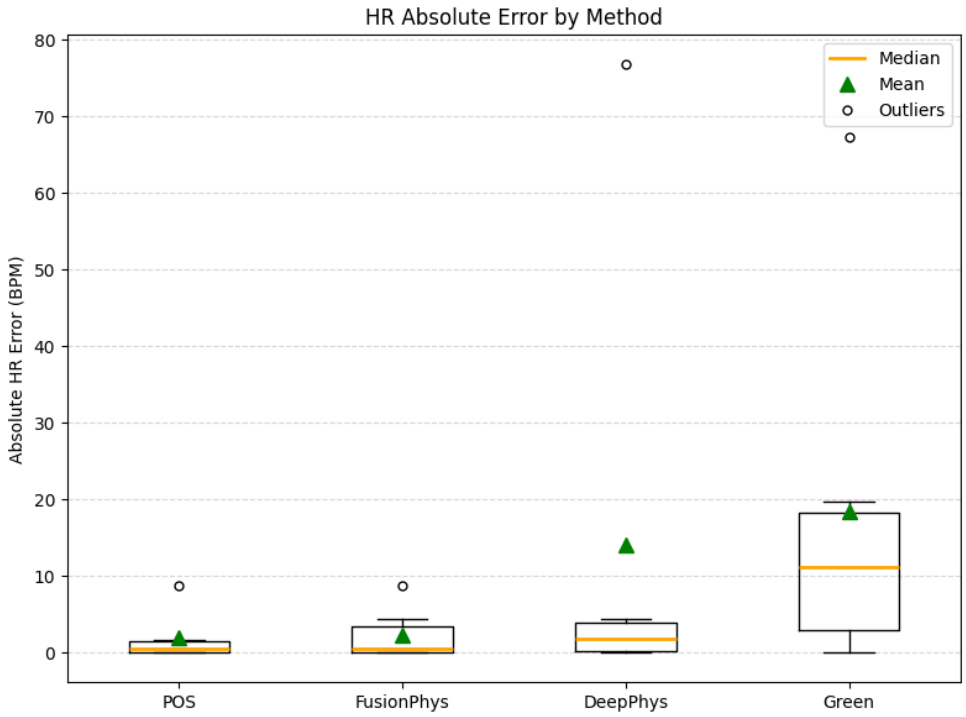

Figure 2. Absolute heart-rate estimation errors for the six held-out Dataset 2 subjects, comparing POS [2], FusionPhys, DeepPhys [3], and Green [1].

4. Conclusions

This paper presents FusionPhys, which enhances the well-known DeepPhys [3] architecture for rPPG-based heart-rate estimation by incorporating classical Green [1] and POS priors.

The standardized derivative of the synchronized PPG is used as the supervision target. On UBFC-rPPG Dataset 1, FusionPhys achieves a 1.38 BPM LOSO MAE, improving on DeepPhys [3] (2.60 BPM) and Green [1] (6.60 BPM), while remaining close to POS [2] (1.02 BPM). On the larger 42-subject Dataset 2 experiment, FusionPhys obtains a 2.37 BPM MAE compared with 14.17 BPM for DeepPhys [3] and 18.45 BPM for Green [1]; POS [2] remains slightly better at 1.93 BPM. FusionPhys nevertheless achieves the best waveform-derivative RMSE and Pearson correlation in both evaluations.

https://doi.org/10.3390/engproc2026163063

Author Contributions: Conceptualization, T.F.R. and G.R.; methodology, T.F.R. and G.R.; software, T.F.R.; validation, T.F.R.; formal analysis, T.F.R. and G.R.; writing—original draft preparation, T.F.R.; writing—review and editing, G.R.; supervision, G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ResearchNB through the Talent Recruitment program awarded to Ghazal Rouhafzay. No article processing charge (APC) was required for this publication.

Institutional Review Board Statement: Not applicable. This study reanalyzed a publicly available benchmark dataset and did not recruit participants or collect new human-subject data.

Informed Consent Statement: Not applicable. Participant consent for the original UBFC-rPPG data collection was handled by the dataset creators; no new participant data were collected in this study.

Data Availability Statement: The UBFC-rPPG dataset analyzed in this study is publicly available from the dataset authors at https://sites.google.com/view/ybenezeth/ubfcrppg (accessed on 3 July 2026). The subject-level result tables used to generate the figures are included with the manuscript source files.

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI ChatGPT (GPT-5.6 Sol, August 2026) for language editing and LaTeX template formatting. The authors reviewed and edited all outputs.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Verkruysse, W.; Svaasand, L.O.; Nelson, J.S. Remote plethysmographic imaging using ambient light. *Opt. Express* **2008**, *16*, 21434–21445. [CrossRef]
2. Wang, W.; den Brinker, A.C.; Stuijk, S.; de Haan, G. Algorithmic principles of remote PPG. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1479–1491. [CrossRef]
3. Chen, W.; McDuff, D. DeepPhys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 349–365. [CrossRef]
4. Yu, Z.; Li, X.; Zhao, G. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019.
5. Liu, X.; Fromm, J.; Patel, S.; McDuff, D. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19400–19411.
6. Bobbia, S.; Macwan, R.; Benezeth, Y.; Mansouri, A.; Dubois, J. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognit. Lett.* **2019**, *124*, 82–90. [CrossRef]
7. Liu, X.; Narayanswamy, G.; Paruchuri, A.; Zhang, X.; Tang, J.; Zhang, Y.; Wang, Y.; Sengupta, S.; Patel, S.; McDuff, D. rPPG-Toolbox: Deep remote PPG toolbox. *arXiv* **2022**, arXiv:2210.00716.
8. Bazarevsky, V.; Kartynnik, Y.; Vakunov, A.; Raveendran, K.; Grundmann, M. BlazeFace: Sub-millisecond neural face detection on mobile GPUs. *arXiv* **2019**, arXiv:1907.05047.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.