
VitalLens 2.0: High-Fidelity rPPG for Heart Rate Variability Estimation from Face Video

Philipp V. Rouast
Rouast Labs
philipp@rouast.com

Abstract

This report introduces VitalLens 2.0, a new deep learning model for estimating physiological signals from face video. This new model demonstrates a significant leap in accuracy for remote photoplethysmography (rPPG), enabling the robust estimation of not only heart rate (HR) and respiratory rate (RR) but also Heart Rate Variability (HRV) metrics. This advance is achieved through a combination of a new model architecture and a substantial increase in the size and diversity of our training data, now totaling 1,413 unique individuals. We evaluate VitalLens 2.0 on a new, combined test set of 422 unique individuals from four public and private datasets. When averaging results by individual, VitalLens 2.0 achieves a Mean Absolute Error (MAE) of 1.57 bpm for HR, 1.08 bpm for RR, 10.18 ms for HRV-SDNN, and 16.45 ms for HRV-RMSSD. These results represent a new state-of-the-art, significantly outperforming previous methods. This model is now available to developers via the VitalLens API at <https://rouast.com/api>.

1 Introduction

Remote Photoplethysmography (rPPG) harnesses signals from standard video to estimate physiological information, offering immense potential for non-invasive health monitoring [11]. Our initial release, VitalLens 1.0, provided real-time estimation of heart rate (HR) and respiratory rate (RR) [7].

This report introduces VitalLens 2.0, the next generation of our rPPG model. The primary goal is to move beyond simple rate estimation to achieve high-fidelity physiological waveform reconstruction. This moves the challenge from simple rate estimation (i.e., finding a dominant frequency) to high-fidelity waveform reconstruction, which requires the precise, sub-second temporal accuracy of Inter-Beat Intervals (IBIs) to be robust.

This leap in performance was achieved through these two key developments:

1. **An expanded and meticulously curated training dataset.** We combined our in-house data with the Vital Videos public dataset, then manually curated all samples to ensure high-quality video and labels. The final set totals 1,413 unique individuals.
2. **A new model architecture and training methodology,** designed specifically to capture the subtle inter-beat variations necessary for accurate HRV analysis.

To validate these improvements, we establish a new, large-scale combined test set. This set is comprised of 1,081 video samples from 422 unique individuals, combining our in-house test set (PROSIT) with several publicly available datasets (Vital Videos [9, 10], UBFC-Phys [6], UBFC-rPPG [1]). All video chunks are processed to be 20-60 seconds in duration, making them suitable for time-domain HRV analysis. For our model development purposes we created our own strict, participant-disjoint training, validation, and test sets to ensure a robust, unbiased evaluation of generalization.

Our key contributions are:

- **VitalLens 2.0.** We introduce a new rPPG model capable of accurately estimating HRV metrics from face video, in addition to HR and RR.
- **Comprehensive evaluation.** We benchmark VitalLens 2.0 on a large, diverse test set of 422 individuals, demonstrating its superior accuracy.
- **State-of-the-art performance.** We show that VitalLens 2.0 significantly outperforms handcrafted algorithms (e.g., POS [12]), other learning-based methods (e.g., MTTs-CAN [4]), and our previous VitalLens 1.0 architecture.

This new model is the engine behind the VitalLens API, enabling developers to integrate robust, real-time HRV analysis into their applications. For more information, visit <https://rouast.com/api>.

2 Architecture

The VitalLens 2.0 model is an end-to-end deep convolutional neural network, building upon an EfficientNet-based backbone [8]. The architecture incorporates novel temporal-attentive mechanisms optimized for extracting high-fidelity physiological waveforms. This design is specifically focused on minimizing signal noise and preserving the precise temporal location of systolic peaks, which is the critical prerequisite for reliable IBI extraction and HRV analysis.

The model takes a sequence of video frames of a person’s face as input and outputs two continuous time-series signals: the pulse (PPG) waveform and the respiration (RESP) waveform. From these waveforms, downstream metrics such as Heart Rate (HR), Respiratory Rate (RR), and HRV metrics (e.g., SDNN, RMSSD, LF/HF) are derived using industry-standard signal processing techniques.

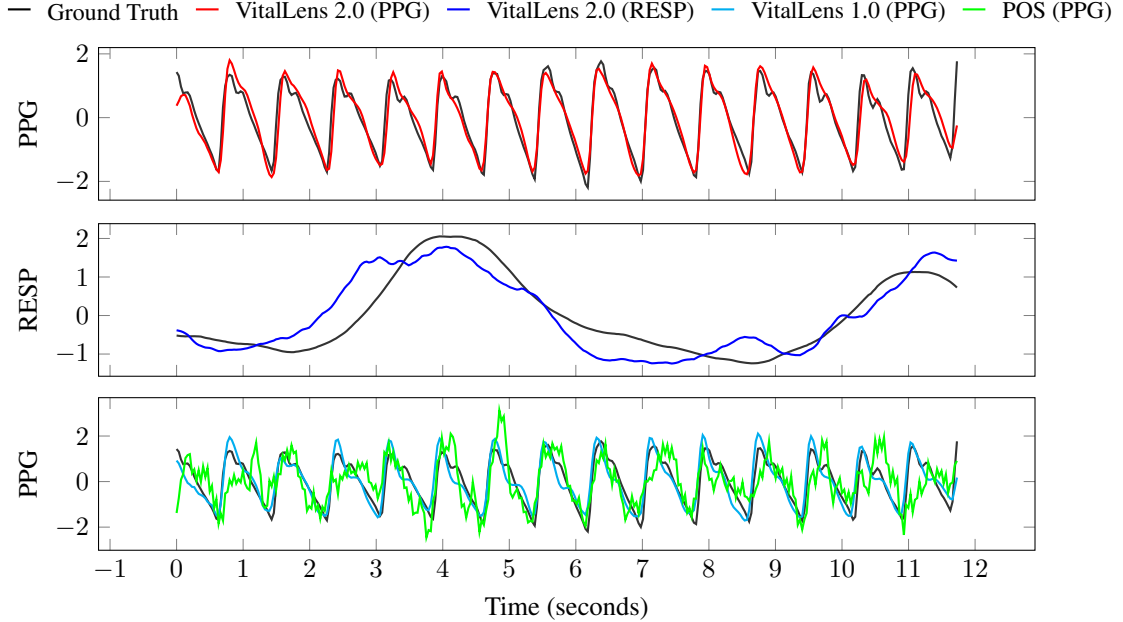


Figure 1: Visual comparison of estimated waveforms from a sample handheld video segment.¹ Top: VitalLens 2.0 PPG vs. Ground Truth. Middle: VitalLens 2.0 Respiration vs. Ground Truth. Bottom: VitalLens 1.0 and POS PPG vs. Ground Truth. VitalLens 2.0 achieves higher fidelity, accurately reconstructing the precise timing of the systolic peaks.

Figure 1 provides a visual comparison of the high-fidelity waveforms generated by VitalLens 2.0 against the ground truth and other models for a sample handheld video segment.

¹The source video and ground truth vitals are publicly available as `sample_video_1` at: <https://github.com/Rouast-Labs/vitalens-python/tree/main/examples>

3 Datasets

The development of VitalLens 2.0 utilized a large-scale, multi-source dataset, combining our in-house PROSIT dataset with several publicly available datasets, including the Vital Videos (VV) collection [9, 10], UBFC-Phys [6], and UBFC-rPPG [1].

Training Dataset. The training dataset combines the training splits of PROSIT and the Vital Videos datasets we have access to. This results in a total training set of 1,413 unique individuals, a significant increase in size and diversity over the data used for VitalLens 1.0 [7]. Beyond demographics, this diversity includes a wide variety of real-world filming locations, diverse lighting conditions, cameras, and varied backgrounds, and unscripted participant behavior, including significant camera motion from handheld devices. The composition of this training set is detailed in Table 1.

Table 1: VitalLens 2.0 Training Dataset Size

Source	# Participants	# Chunks	Time (hours)
PROSIT (In-house)	128	5,700	16.8
Vital Videos (EU)	589	2,238	9.3
Vital Videos (Africa)	383	2,869	7.7
Vital Videos (South Asia)	313	3,887	10.8
Total	1,413	14,694	44.6

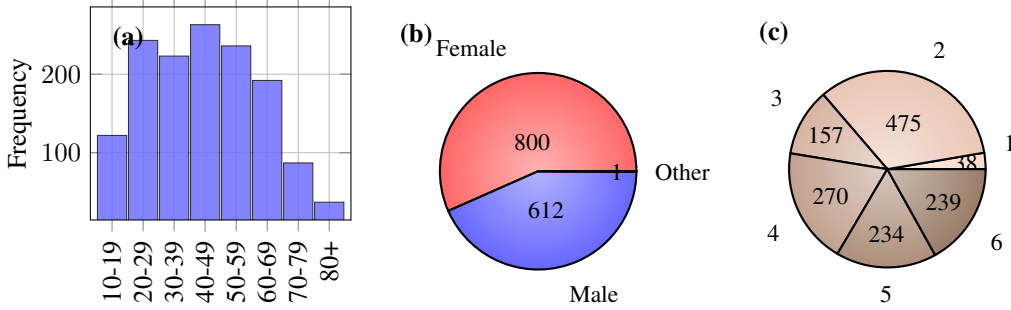


Figure 2: Participant demographics in the training dataset. (a) Age, (b) Gender, (c) Skin type.

Figures 2 and 3 detail the demographic and physiological composition of this new training set. Data is reported per-individual. It is well-balanced in gender and includes a comprehensive representation across all six Fitzpatrick skin types. The vitals distributions cover a wide range of physiological states, including a broad spectrum of heart rate variability.

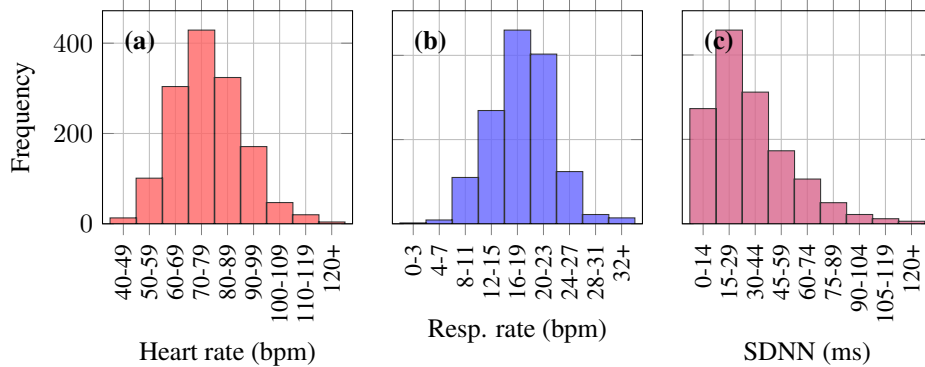


Figure 3: Distributions of per-individual average vitals in the combined training dataset. (a) Heart Rate, (b) Respiratory Rate, (c) HRV-SDNN.

Test Dataset. To validate VitalLens 2.0, we constructed a new, combined test set. All individuals in this test set are disjoint from the training and validation sets. Chunks were re-processed to have durations between 20 and 60 seconds, enabling the calculation of time- and frequency-domain HRV metrics. The composition of this test set is detailed in Table 2. While all test samples have a ground truth signal for PPG, some (VV in part and both UBFC-rPPG and UBFC-Phys) do not have a ground truth signal for RESP. In addition, samples with known label-quality or synchronization issues (e.g., in parts of UBFC-Phys) were excluded to ensure a fair and reliable benchmark.

Table 2: VitalLens 2.0 Combined Test Set Composition

Source Dataset	# Participants	# Chunks
PROSIT (In-house)	22	251
Vital Videos (EU)	146	282
Vital Videos (Africa)	93	260
Vital Videos (South Asia)	78	168
UBFC-Phys	34	67
UBFC-rPPG	49	53
Total	422	1,081

4 Methodology

Model Training and Validation. We follow a strict participant-disjoint methodology for training, validation, and testing. The 1,413 individuals in the training set were used to optimize model parameters. A separate validation set, also participant-disjoint from the training set, was used to monitor for overfitting and to perform hyperparameter tuning and model selection. The final VitalLens 2.0 model chosen for evaluation is the one that demonstrated the best performance on this validation set.

Models Compared. We benchmark VitalLens 2.0 against a comprehensive set of methods. These include traditional handcrafted algorithms (G [11], CHROM [3], and POS [12]) and several prominent learning-based methods (DeepPhys [2], MTTS-CAN [4], and EfficientPhys [5]). To ensure a fair and direct comparison, all learning-based baselines were re-trained from scratch on the exact same 1,413-participant training dataset used for VitalLens 2.0.

We also include two internal baselines:

- **VitalLens 1.0*:** The original model as presented in [7], trained on the smaller, original dataset.
- **VitalLens 1.1:** The original VitalLens 1.0 architecture re-trained on our new, larger 1,413-participant dataset.

This comparison allows us to isolate performance gains from architectural improvements (VitalLens 2.0 vs. 1.1) versus data improvements (VitalLens 1.1 vs. other baselines and 1.0*).

Evaluation Metrics. We evaluate performance at both the waveform and vital sign levels. For waveforms, we report Pearson Correlation (r) and Signal-to-Noise Ratio (SNR). For vital signs, we report Mean Absolute Error (MAE) for HR, RR, and the HRV metrics SDNN, RMSSD, and LF/HF.

HRV Calculation Pipeline. The estimation of HRV metrics from the predicted PPG waveform follows a multi-stage signal processing pipeline. First, cardiac cycles are identified by detecting valid peaks in the waveform, considering signal prominence, width, and periodicity relative to a rolling frequency estimate. These detections are filtered to retain only high-confidence peaks. From the resulting peak train, Inter-Beat Intervals (IBIs) are calculated. This IBI time-series is then cleaned by interpolating outlier intervals that fall outside a physiologically plausible range (e.g., due to missed or false detections). Finally, standard time-domain (SDNN, RMSSD) and frequency-domain (LF/HF) metrics are computed from the cleaned IBI sequence, contingent on meeting minimum duration (e.g., ≥ 20 s for SDNN) and beat count thresholds.

Reporting Strategy. All results in Section 5 are reported by first averaging metrics for all chunks belonging to a single individual, and then averaging these per-individual scores. This approach prevents individuals with more video chunks from disproportionately influencing the aggregate results. HRV metrics (SDNN, RMSSD) are calculated for all chunks ≥ 20 s. LF/HF, which requires a longer window, is calculated only for chunks ≥ 55 s (approx. 1/3 of the test set).

5 Results

5.1 Overall Performance

Table 3 summarizes the performance of all models on our 422-participant combined test set. VitalLens 2.0 significantly outperforms all other methods across almost all metrics, establishing a new state-of-the-art for learned rPPG models.

Table 3: Vitals estimation results on the combined test set (N=422 individuals). Results are averaged per-individual. Best performance is in **bold**.

Method	Pulse			Respiration ^a			Heart Rate Variability (HRV)		
	HR	PPG		RR	RESP		SDNN	RMSSD	LF/HF ^b
	MAE ↓	r ↑	SNR ↑	MAE ↓	r ↑	SNR ↑	MAE ↓	MAE ↓	MAE ↓
G	12.33	0.34	−8.20	—	—	—	58.12	81.29	1.31
CHROM	3.26	0.56	1.87	—	—	—	30.23	48.87	1.12
POS	4.03	0.61	2.68	—	—	—	50.56	80.09	1.35
DeepPhys	5.17	0.65	3.43	—	—	—	26.57	40.65	1.13
MTTS-CAN	2.17	0.76	7.47	—	—	—	19.58	31.73	0.96
EfficientPhys	2.96	0.75	7.16	—	—	—	19.34	31.38	0.98
VitalLens 1.0*	2.13	0.81	9.88	1.91	0.75	7.15	20.55	33.17	1.03
VitalLens 1.1	1.64	0.85	12.89	1.09	0.81	10.09	12.70	20.00	0.85
VitalLens 2.0	1.57	0.86	13.52	1.08	0.82	10.25	10.18	16.45	0.83

^a Calculated on subset containing a ground truth RESP signal (approx. 40% of test set).

^b Calculated on subset of chunks ≥ 55 s in duration (approx. $\frac{1}{3}$ of test set).

* VitalLens 1.0 was trained on a smaller dataset.

Notably, the performance gap is most pronounced in HRV estimation. While retraining the original model architecture on new data (VitalLens 1.1) yields significant improvements over both the original VitalLens 1.0* and other baselines, the new architecture in VitalLens 2.0 provides a further substantial leap in accuracy. For example, VitalLens 2.0 achieves an SDNN MAE of 10.18 ms.

This demonstrates that the architectural and training methodology improvements in VitalLens 2.0 were critical for achieving the high-fidelity waveform estimation necessary for robust HRV analysis. This result is the basis for enabling HRV metrics only for users of the VitalLens 2.0 model in our API.

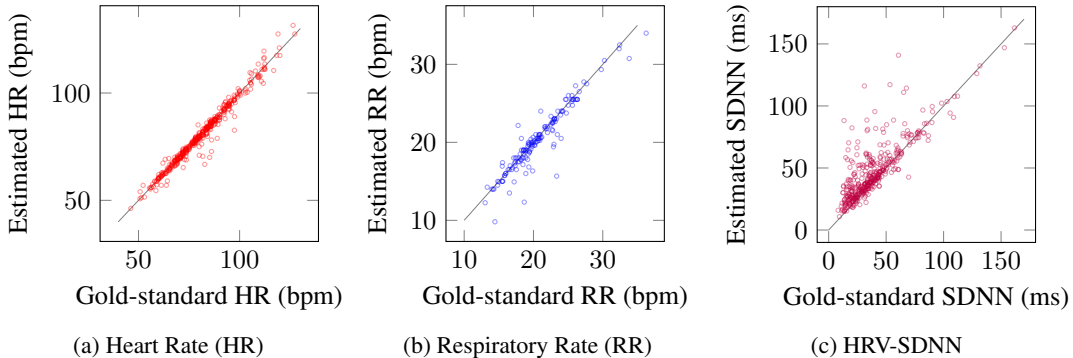


Figure 4: VitalLens 2.0 estimated vitals vs. gold-standard true vitals on the combined test set.

Figure 4 provides scatter plots for HR, RR, and SDNN, visually demonstrating the high correlation between VitalLens 2.0 estimates and the gold-standard ground truth labels.

5.2 Results by Dataset

To demonstrate the generalization of VitalLens 2.0, Table 4 breaks down performance across all constituent test sets. The results confirm the model’s strong performance, and the variance between datasets is explained by their specific, known characteristics. The in-house PROSIT set, which uniquely features unscripted participant and camera motion, serves as our most challenging real-world benchmark. As expected, it shows the highest MAE for HR, RR, and SDNN. Similarly, Vital Videos (Africa) shows a high SDNN MAE of 24.77 ms, comparable to PROSIT’s. This is also anticipated, as this dataset’s primary challenge is its high concentration of participants with Fitzpatrick skin types 5 and 6. In contrast, the Vital Videos (EU) and (South Asia) subsets, which feature stationary subjects and a wider mix of skin tones, demonstrate excellent performance, with SDNN MAE as low as 6.00 ms.

Table 4: VitalLens 2.0 estimation performance by dataset subset. Results are averaged per-individual.

Source Dataset	HR MAE ↓	RR MAE ^a ↓	SDNN MAE ↓
PROSIT (In-house)	3.22	3.04	27.45
Vital Videos (EU)	1.30	1.15	7.69
Vital Videos (Africa)	1.37	0.80	24.77
Vital Videos (South Asia)	0.90	0.69	6.00
UBFC-Phys	2.33	–	2.81
UBFC-rPPG	2.53	–	3.76

^a RR metrics only given where ground truth labels available.

A notable finding comes from the UBFC-Phys and UBFC-rPPG datasets, which show mediocre HR MAE (2.33 and 2.53 bpm, respectively) but state-of-the-art SDNN MAE (2.81 and 3.76 ms). This apparent discrepancy is likely attributable to the composition of these datasets. The long duration of the UBFC samples (typically 60 seconds) provides a highly stable window for time-domain HRV analysis, which is less sensitive to the challenges that affect frequency-domain HR estimation. Furthermore, the higher average heart rates in these datasets (approx. 90 bpm) may present a greater challenge for frequency-based HR algorithms. Finally, the RR MAE results reinforce these themes, with the high-motion PROSIT set showing significantly higher error (3.04 bpm) than the stationary Vital Videos subsets. This confirms that the core challenges identified, namely participant motion and dataset composition, are consistent across all estimated vital signs [7].

5.3 Robustness Analysis

The dataset-level analysis in Section 5.2 demonstrates that performance is dictated by specific, known challenges. The high error rates in the PROSIT and Vital Videos (Africa) subsets, for example, directly point to participant movement and darker skin tones as the two most critical factors for robust, real-world HRV estimation. To isolate and quantify the impact of these factors, we conducted this more granular analysis, comparing VitalLens 2.0 against the previous VitalLens 1.0* model. The results in Figure 5 not only confirm this hypothesis but also highlight the practical value of the new architecture.

Figure 5(a) illustrates the model’s performance against participant movement, a critical factor for any non-clinical or handheld application. We binned the test set into terciles (Low, Medium, and High) based on a computed motion metric. The results clearly show that while motion artifacts remain a challenge for both models, with error rates increasing in line with motion, the new architecture in VitalLens 2.0 provides a substantial, systematic improvement. The key finding is a clear downward shift in absolute error across all three bins. For instance, we observe that the SDNN MAE for VitalLens 2.0 in the ‘High’ movement category is lower than the error for VitalLens 1.0* in the ‘Low’ movement category.

Similarly, Figure 5(b) analyzes performance across all six Fitzpatrick skin types. The data reveals two key findings. First, VitalLens 2.0 delivers a new level of performance, achieving a low and highly

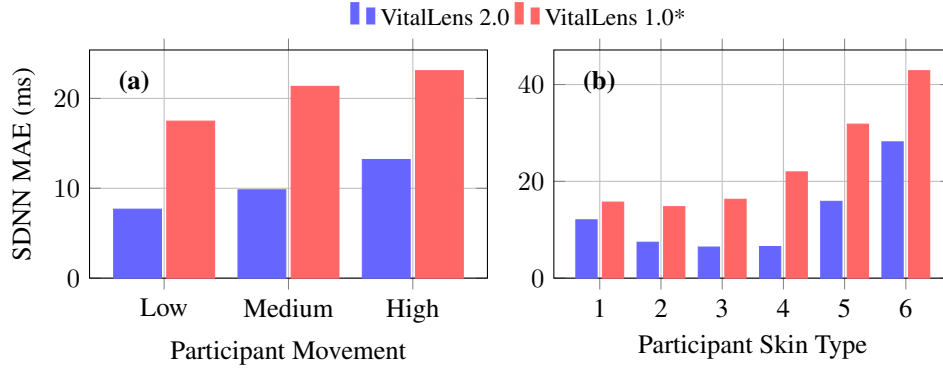


Figure 5: Comparing the robustness in HRV-SDNN estimation between VitalLens 2.0 and VitalLens 1.0* under increasing participant movement and different participant skin types.

stable error rate across skin types 1 through 4, in contrast to the rising error of VL 1.0*. Second, for skin types 4, 5, and 6, where the rPPG signal is most attenuated, VitalLens 2.0 provides a strong reduction in error. This confirms the effectiveness of our expanded and diversified training dataset, which included a significant number of individuals with darker skin tones. However, the results also transparently show that a performance gap remains; error rates for types 5 and 6 are still notably higher than for types 1-4. While VL 2.0 makes HRV estimation more equitable and reliable, further research is required. Much as our previous work focused on stabilizing HR estimation across all skin tones, achieving this same equity for high-fidelity HRV metrics represents the new frontier, and it remains a key focus for our ongoing development.

6 Conclusion

VitalLens 2.0 represents a significant advancement in remote physiological monitoring. By combining a large-scale, diverse dataset with a novel, optimized architecture, it achieves state-of-the-art accuracy not only for heart and respiration rates but also for challenging Heart Rate Variability metrics. The demonstrated high-fidelity performance, particularly for HRV, opens new possibilities for accessible, non-invasive health and wellness tracking.

Acknowledgments and Disclosure of Funding

We are thankful to Pieter-Jan Toye for his helpful suggestions and access to the Vital Videos dataset. This work was funded by Rouast Labs Pty Ltd.

References

- [1] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.
- [2] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *ECCV*, 2018.
- [3] Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.
- [4] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In *NeurIPS*, 2020.
- [5] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5008–5017, 2023.

- [6] UBFC-Phys: A multimodal database for psychophysiological studies of social stress. Sabour, rita meziati and benezeth, yannick and de oliveira, pierre and chappe, julien and yang, fan. *IEEE Transactions on Affective Computing*, 14(1):622–636, 2021.
- [7] Philipp V Rouast. VitalLens: Take A Vital Selfie. *arXiv preprint arXiv:2312.06892*, 2023.
- [8] Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. In *ICML*, 2021.
- [9] Pieter-Jan Toye. Vital videos: A public dataset of videos with ppg and bp ground truths. *arXiv preprint arXiv:2306.11891*, 2023.
- [10] Pieter-Jan Toye. A large and diverse rppg dataset with rich ground truths. In *Proc. ICCV*, 2025.
- [11] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008.
- [12] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Trans. Biomedical Eng.*, 64(7):1479–1491, 2017.