CNN-Transformer Framework with Hybrid Loss for Robust FMCW Radar-based Heartbeat Sensing

Ying Wang[1], Zhaodong Sun[1,*], Xu Cheng[1] and Zuxian He[1]

[1]School of Computer Science, Nanjing University of Information Science and Technology, 219 Ningliu Road, Nanjing, Jiangsu, 210004, China

Abstract

Remote physiological sensing using Frequency Modulated Continuous Wave (FMCW) radar has emerged as a promising alternative to contact-based methods due to its non-intrusive nature and privacy preservation. However, existing signal-processing and CNN-based approaches suffer from phase wrapping ambiguities, noise sensitivity, and limited ability to capture long-range dependencies in heartbeat dynamics. In this work, we propose a novel CNN-Transformer framework for supervised radar-based heartbeat measurement. The CNN component extracts local temporal features, while the transformer encoder models long-range dependencies critical for periodic cardiac motion. To further enhance performance, we design a hybrid loss function that integrates Negative Pearson Loss, Signal-to-Noise Ratio (SNR) Loss, and Sparsity Loss, effectively balancing temporal fidelity, noise robustness, and physiologically meaningful frequency representation. We additionally introduce RadHR, a new FMCW radar dataset with recordings from 50 participants, providing a high-quality benchmark for non-contact heartbeat estimation. Extensive experiments on both the public EquiPleth dataset and RadHR demonstrate that our method consistently outperforms existing baselines, achieving state-of-the-art accuracy and robustness under realistic conditions.

Keywords

FMCW Radar, Remote Heart Rate Sensing, Deep Learning, Vital Signs

1. Introduction

Radar-based heartbeat sensing has attracted increasing attention as a promising non-contact physiological monitoring technique. Compared with traditional contact-based methods such as electrocardiography (ECG) and photoplethysmography (PPG), radar sensing provides unique advantages in privacy preservation, environmental robustness, and suitability for continuous long-term monitoring. The underlying principle is to detect subtle chest wall vibrations (typically 0.1–0.5 mm) induced by cardiac activities [1]. Frequency Modulated Continuous Wave (FMCW) radars are widely employed due to their ability to precisely track motion by measuring relative phase variations in received chirp signals.

In recent years, radar-based physiological monitoring has developed into a highly active research field, with applications ranging from healthcare to safety and smart environments. Several review studies have systematically summarized these advances. For instance, comprehensive surveys on Doppler radar technology highlight its potential for continuous healthcare monitoring without requiring physical contact, enabling early diagnosis and chronic disease management in clinical and home-care settings [2]. Similarly, microwave radar sensing systems have been investigated in the context of search-and-rescue operations, where non-contact monitoring of vital signs in complex and cluttered environments can support timely detection of survivors [3]. In parallel, biomedical MIMO radar systems have been extensively studied for their ability to achieve both vital sign detection and fine-grained human localization, offering a promising avenue for multi-person monitoring scenarios [4]. Collectively, these reviews establish radar sensing as a versatile modality capable of addressing challenges that are difficult to solve with traditional sensors.

The 4th Vision-based Remote Physiological Signal Sensing (RePSS) Challenge & Workshop, August 29, 2025, Guangzhou, China
*Corresponding author.
ying.wang@nuist.edu.cn (Y. Wang); zhaodong.sun@nuist.edu.cn (Z. Sun); xcheng@nuist.edu.cn (X. Cheng); 202412200693@nuist.edu.cn (Z. He)
0009-0001-2524-5579 (Y. Wang); 0000-0002-0597-0765 (Z. Sun); 0000-0003-2355-9010 (X. Cheng)

Beyond reviews, numerous radar architectures and signal processing pipelines have been explored for real-world vital sign detection. Impulse-radio ultra-wideband (IR-UWB) radar, owing to its fine temporal resolution, has been widely applied to monitor heartbeat and respiration. Early systems demonstrated non-contact heart rate monitoring using IR-UWB signals under controlled conditions [5], while subsequent preclinical studies validated simultaneous monitoring of respiration and carotid pulsation, paving the way for clinical applicability [6]. IR-UWB radar has also been deployed in challenging enclosed environments, such as inside vehicles, to detect and localize passengers while extracting vital signs through non-line-of-sight measurements [7].

In addition to IR-UWB methods, self-calibrating radar systems have been proposed to improve stability and adaptability across different users and scenarios. These approaches automatically adjust system parameters to mitigate the effects of channel variation and hardware non-idealities, thereby enhancing robustness for long-term monitoring [8]. Meanwhile, mm-wave FMCW radars have been validated for remote monitoring of human vital signs, showing strong resilience to environmental interference and enabling compact, low-power implementations suitable for pervasive healthcare systems [9].

Another emerging application domain is radar-based vital sign monitoring in automotive environments. Detecting passenger presence and health conditions inside vehicles is particularly challenging due to vibrations and motion artifacts. Recent studies have conducted both theoretical investigations [10] and practical experiments [11], demonstrating the feasibility of extracting respiration and heartbeat information even in the presence of strong vehicle vibrations. These findings extend the applicability of radar sensing from controlled laboratory settings to highly dynamic real-world conditions.

Despite these advancements, conventional radar-based heartbeat sensing approaches typically rely on extracting and unwrapping signal phases to recover heartbeat dynamics [9, 12, 13]. However, such methods remain highly susceptible to motion artifacts, multipath interference, and low signal-to-noise ratio (SNR) conditions. Phase wrapping ambiguities and noise sensitivity often lead to significant performance degradation, particularly in realistic environments where subjects are not perfectly stationary.

To overcome these limitations, recent advances in supervised deep learning have demonstrated the ability to learn complex spatiotemporal representations directly from radar signals [14, 15, 16]. By bypassing explicit phase unwrapping, these methods achieve greater robustness under noise and motion. Nevertheless, most existing deep learning frameworks are dominated by convolutional neural networks (CNNs), which are effective at capturing local temporal features but struggle to model long-range dependencies that are critical for representing periodic heartbeat dynamics.

In this work, we propose a supervised FMCW radar-based heartbeat measurement framework that combines CNNs with Transformers. Our main contributions are summarized as follows:

- We design a novel framework for radar heartbeat sensing, which integrates 1D CNN layers for local feature extraction with Transformer encoders for modeling long-range temporal dependencies, addressing the limitations of CNN-only baselines.
- We collected a new radar dataset (RadHR) containing 50 individuals for heartbeat sensing benchmark, and the dataset will be made public upon request.
- The proposed model demonstrates superior resilience against motion artifacts, multipath interference, and low-SNR conditions, which commonly degrade the performance of traditional signal-processing and CNN-based methods.
- We conduct extensive experiments on FMCW radar data, showing that our approach consistently outperforms state-of-the-art CNN-based baselines in heartbeat sensing accuracy and robustness.

2. Methodology

2.1. Preliminaries

A range matrix is obtained from FMCW radar raw data to facilitate subsequent analysis and processing. Specifically, the procedure of constructing a range matrix is as follows. In each chirp loop, the FMCW

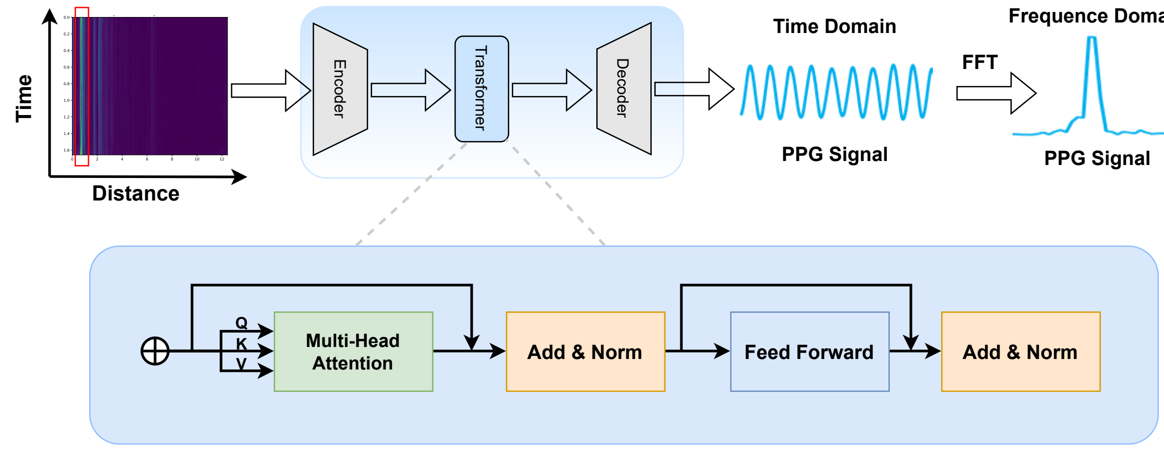

Figure 1: The framework of our CNN-Transformer method.

radar transmits a chirp signal $s(t)$ and simultaneously receives the corresponding reflected chirp signal $u(t)$. Both $s(t)$ and $u(t)$ are linear frequency modulation signals, commonly referred to as chirp signals. In particular, the received signal $u(t)$ is mixed with the in-phase and quadrature (IQ) components of the transmitted signal, denoted as $s_I(t)$ and $s_Q(t)$, to produce the complex intermediate frequency (IF) signal $p(t) \in \mathbb{R}^D$, which can be expressed as:

$$p(t) \propto \mathrm{LPF}[s_I(t) \cdot u(t)] + j, \mathrm{LPF}[s_Q(t) \cdot u(t)] \;\propto\; \exp(j(2\pi f t + \varphi)), \quad f = 2kd/c, \; \varphi = 4\pi d/\lambda \tag{1}$$

where LPF denotes the low-pass filter, k is the frequency slope of the FMCW signal, d is the distance, c is the speed of light, and λ is the wavelength associated with the FMCW starting frequency. The frequency f of the IF signal $p(t)$ corresponds to the frequency difference between the transmitted signal $s(t)$ and the received signal $u(t)$. Consequently, f is directly proportional to the signal round-trip time and the distance d between the radar and the target. Likewise, the phase φ is also proportional to the distance, but it is inherently wrapped within the interval $[-\pi, \pi]$.

To capture heartbeat dynamics continuously, the radar sequentially transmits N chirps $[s_1(t), s_2(t), \ldots, s_N(t)]$ and receives the corresponding reflected signals $[u_1(t), u_2(t), \ldots, u_N(t)]$. Accordingly, N IF signals $[p_1(t), p_2(t), \ldots, p_N(t)]$ are obtained, where each $p_n(t) \in \mathbb{R}^D$. Since the frequency of each IF signal $p_n(t)$ is a function of the target distance, a fast Fourier transform (FFT) is applied to each $p_n(t)$ to generate the corresponding range profile $P_n[f]$. Finally, by concatenating all N range profiles, the range matrix is constructed as:

$$P = [P_1[f], P_2[f], \ldots, P_N[f]] \in \mathbb{R}^{N \times D}, \tag{2}$$

where N represents the number of chirps and D denotes the number of range bins. This range matrix serves as the foundation for subsequent feature extraction and heartbeat estimation in our framework.

2.2. CNN-Transformer Module

As described in Section 2.1, the raw FMCW radar signals are converted into a range matrix $P \in \mathbb{R}^{N \times D}$, where N is the number of chirps and D is the number of range bins. This range matrix captures both the distance-related amplitude and phase variations, serving as the input for subsequent heartbeat estimation. Before feeding into the neural network, we take a window of the range matrix $P_w \in \mathbb{R}^{N \times D}$ around the central range bin d to get the windowed heartbeat matrix $P_w(\cdot, d \pm \Delta d) \in \mathbb{R}^{N \times (2\Delta d + 1)}$ as the input following the previous work [14].

The first stage of our model is a 1D CNN-based feature extractor, which operates along the temporal dimension of the range matrix. Specifically, for each range bin $d_w \in [1, d \pm \Delta d]$, the corresponding temporal sequence $P_w[:, d_w]$ is processed by stacked convolutional layers with small kernel sizes. These layers aim to capture local temporal patterns corresponding to heartbeat-induced chest movements. Formally, the CNN feature extraction can be expressed as:

$$F_{CNN} = CNN(P_w), F_{CNN} \in \mathbb{R}^{N \times C} \tag{3}$$

where C is the number of feature channels output by the CNN. We adopt ReLU activations, batch normalization, and dropout to improve training stability and prevent overfitting.

While CNNs effectively capture local patterns, heartbeat signals exhibit long-range temporal dependencies that CNNs alone may fail to model. To address this, we incorporate a Transformer encoder after the CNN stage. The Transformer employs self-attention mechanisms to model interactions between distant time steps, allowing the network to capture the periodicity and subtle dynamics of cardiac motion. Given the CNN features $F_{\text{CNN}} \in \mathbb{R}^{N \times C}$, the Transformer computes:

$$F_{Trans} = Transformer(F_{CNN}), F_{Trans} \in \mathbb{R}^{N \times C} \tag{4}$$

where F_{Trans} encodes both local and global temporal information. We adopt multi-head attention to allow the model to focus on multiple temporal patterns simultaneously, followed by a feed-forward network with residual connections and layer normalization.

3. Losses

In this work, we design a composite loss function that combines Negative Pearson Loss, Signal-to-Noise Ratio (SNR) Loss[14], and a Sparsity Loss[17] to optimize the supervised heartbeat measurement task using FMCW radar. This hybrid design allows the model to achieve accurate estimation, suppress noise, and encourage physiologically meaningful spectral representations.

3.1. Negative Pearson Loss

The Negative Pearson Loss evaluates the linear correlation between the predicted signal and the ground truth. The Pearson correlation coefficient ranges from -1 to 1, with higher values indicating stronger correlations. To maximize similarity, we minimize the negative Pearson coefficient:

$$L_{Pearson} = -Pearson(\hat{y}, y) \tag{5}$$

where $\hat{y}$ and y denote the predicted and reference heartbeat signals, respectively. This loss function encourages the model to preserve temporal waveform consistency with the ground truth.

3.2. Signal-to-Noise Ratio(SNR) Loss

To enhance robustness against noise and motion artifacts, we employ an SNR-based loss that emphasizes spectral energy concentration around the true heartbeat frequency. Specifically, the loss is defined as:

$$L_{SNR}(\mathrm{y}, \hat{\mathrm{y}}) = \frac{\int_{f_0-w}^{f_0+w} |\hat{\mathrm{Y}}(f)|^2 df}{\int_{-\infty}^{f_0-w} |\hat{\mathrm{Y}}(f)|^2 df + \int_{f_0+w}^{\infty} |\hat{\mathrm{Y}}(f)|^2 df}, f_0 = argmax Y(f) \tag{6}$$

where $Y(f)$ and $\hat{Y}(f)$ are the respective Fourier transforms of y and $\hat{y}$ and w is the chosen window size.

3.3. Sparsity Loss

We integrate Sparsity Loss with Negative Pearson Loss and SNR Loss motivated by the fact that in FMCW radar-based heartbeat sensing, the heartbeat frequency typically manifests as the dominant spectral peak within a physiologically plausible range (e.g., 45–250 bpm). The Sparsity Loss penalizes predictions that fail to concentrate energy near the main peak within this frequency band:

$$L_s = \frac{1}{\sum_{i=a}^{b} Y_i} \left[\sum_{i=a}^{Y^*-\Delta Y} Y_i + \sum_{i=Y^*+\Delta Y}^{b} Y_i \right] \tag{7}$$

where $Y_* = \text{argmax(Y)}$ and ΔY are the frequencies of the spectral peak and padding around the peak, respectively. For all experiments $\Delta Y = 6$ beats per minute[18].

3.4. Overall Loss

The final loss function is a weighted combination of the three terms:

$$L_{total} = \lambda_1 L_{Pearson} + L_{SNR} + \lambda_2 L_{Sparsity} \tag{8}$$

where λ_1, λ_2 are hyperparameters balancing the contribution of each term.

4. Experiments

4.1. Datasets and Experimental Setup

4.1.1. Equipleth Dataset

The Equipleth radar dataset [14] comprises 550 paired facial video and FMCW radar recordings collected from 91 participants. Skin tones are classified using the Fitzpatrick scale [19], with 28, 49, and 14 subjects representing light, medium, and dark skin tones, respectively, for fairness evaluation. Each participant contributed six 30-second recordings. Additional details are provided in the supplementary materials.

4.1.2. RadHR

Our self-collected radar heartbeat dataset (RadHR) consists of recordings from 50 participants, each measured in a stationary condition to minimize motion artifacts. For every subject, FMCW radar signals were continuously collected for 30 seconds at a sampling rate of 120 frames per second (fps), and subsequently converted into range matrices following the standard FMCW signal processing pipeline. This dataset provides high-temporal-resolution radar measurements of subtle chest wall movements, serving as a reliable benchmark for supervised heartbeat estimation research.

4.1.3. Experimental Setup

Following prior work [14], we use 10-second windows for training and heart rate evaluation. For the Equipleth and RadHR dataset, we use the same training protocol as [14]. The model is optimized using the AdamW algorithm with a learning rate of 1×10^{-4} for 200 epochs, and the best-performing checkpoint is selected based on validation set performance. For evaluation, we follow prior work and report mean absolute error (MAE), root mean squared error (RMSE), and the Pearson correlation coefficient (r) as the primary metrics for heart rate estimation.

4.2. Comparison with State-of-the-Art method

Table 1 summarizes the intra-dataset heart rate estimation results using radar modality on both the EquipIeth dataset and our proposed RadHR dataset. We compare our method with three baselines: FFT-based Radar, EquipIeth Radar, and VitaNet. The evaluation metrics include mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (r).

On the EquipIeth dataset, our approach achieves an MAE of 1.82, RMSE of 5.39, and correlation $r = 0.89$, surpassing previous methods and demonstrating robust performance. Similarly, on the RadHR

Table 1
Intra-dataset heart rate results of radar modality on Equipleth dataset and our RadHR dataset. The best results are in bold.

Method	Equipleth			RadHR		
	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑
FFT-based Radar[9]	13.51	21.07	0.24	12.23	18.33	0.21
Equipleth Radar[14]	2.18	6.12	0.89	3.15	7.13	0.84
VitaNet[15]	3.14	7.70	0.77	5.26	9.17	0.63
ours	**1.82**	**5.39**	**0.89**	**2.11**	**2.73**	**0.92**

Table 2
Ablation study of the overall loss on EquiPleth dataset. The best results are in bold.

Pearson Loss	SNR Loss	Sparsity Loss	MAE↓	RMSE↓	r↑
✓			8.52	14.12	0.33
✓	✓		1.92	5.33	0.89
✓	✓	✓	**1.82**	**5.39**	**0.89**

dataset, our method achieves an MAE of 2.11, RMSE of 2.73, and correlation $r = 0.92$, outperforming all baselines by a clear margin. Notably, compared with the FFT-based Radar method, our approach reduces the RMSE by more than 85% on RadHR, highlighting the effectiveness of combining CNN and Transformer architectures with our tailored loss design.

4.3. Ablation Study

To investigate the contribution of each component in the overall loss function, we conduct an ablation study on the EquiPleth dataset. The results are summarized in Table 2.

When only the Pearson Loss is used, the model achieves a relatively high MAE (8.52) and RMSE (14.12), with a poor correlation coefficient (r = 0.33). This indicates that although Pearson Loss enforces correlation, it alone is insufficient for stable reconstruction.

Introducing the SNR Loss significantly improves performance, reducing the MAE to 1.92 and RMSE to 5.33, while the correlation r increases to 0.89. This suggests that the SNR constraint effectively enhances the signal fidelity by improving the signal-to-noise ratio.

Finally, when all three losses (Pearson Loss, SNR Loss, and Sparsity Loss) are combined, the model achieves the best performance, with the lowest MAE (1.82), competitive RMSE (5.39), and the highest correlation (r = 0.89). The improvement demonstrates that the Sparsity Loss further regularizes the prediction, helping the model suppress redundant information and capture more discriminative features.

5. Conclusion

In this paper, we presented a CNN-Transformer framework for FMCW radar-based heartbeat estimation, coupled with a novel hybrid loss design. By leveraging CNNs for local feature extraction and Transformers for long-range dependency modeling, our approach effectively captures both fine-grained and global temporal patterns of cardiac dynamics. The integration of Pearson Loss, SNR Loss, and Sparsity Loss further enhances robustness by encouraging waveform fidelity, noise suppression, and physiologically consistent spectral concentration. To support the community, we introduced RadHR, a new radar heartbeat dataset comprising recordings from 50 subjects under stationary conditions. Experimental results on both RadHR and the EquiPleth dataset demonstrated that our method outperforms conventional signal-processing and deep learning baselines in terms of MAE, RMSE, and correlation coefficient.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62572249), the Natural Science Foundation of Jiangsu Province (Grant No. BK20250742), the Startup Foundation for Introducing Talent of NUIST (Grant No. 1083142501006), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. SJCX25_0518).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase, and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. D. Droitcour, Non-contact measurement of heart and respiration rates with a single-chip microwave doppler radar, Stanford University, 2006.

[2] C. Li, V. M. Lubecke, O. Boric-Lubecke, J. Lin, A review on recent advances in doppler radar sensors for noncontact healthcare monitoring, IEEE Transactions on microwave theory and techniques 61 (2013) 2046–2060.

[3] N. Van Thi Phuoc, L. Tang, V. Demir, S. Hasan, N. Duc Minh, S. Mukhopadhyay, Review-microwave radar sensing systems for search and rescue purposes, Sensors 19 (2019) 2879.

[4] E. Cardillo, A. Caddemi, A review on biomedical mimo radars for vital sign detection and human localization, Electronics 9 (2020) 1497.

[5] Y. Lee, J.-Y. Park, Y.-W. Choi, H.-K. Park, S.-H. Cho, S. H. Cho, Y.-H. Lim, A novel non-contact heart rate monitor using impulse-radio ultra-wideband (ir-uwb) radar technology, Scientific reports 8 (2018) 13053.

[6] J.-Y. Park, Y. Lee, Y.-W. Choi, R. Heo, H.-K. Park, S.-H. Cho, S. H. Cho, Y.-H. Lim, Preclinical evaluation of a noncontact simultaneous monitoring method for respiration and carotid pulsation using impulse-radio ultra-wideband radar, Scientific reports 9 (2019) 11892.

[7] S. Lim, S. Lee, J. Jung, S.-C. Kim, Detection and localization of people inside vehicle using impulse radio ultra-wideband radar sensor, IEEE Sensors Journal 20 (2019) 3892–3901.

[8] M.-C. Huang, J. J. Liu, W. Xu, C. Gu, C. Li, M. Sarrafzadeh, A self-calibrating radar sensor system for measuring vital signs, IEEE transactions on biomedical circuits and systems 10 (2015) 352–363.

[9] M. Alizadeh, G. Shaker, J. C. M. D. Almeida, P. P. Morita, S. Safavi-Naeini, Remote monitoring of human vital signs using mm-wave fmcw radar, IEEE Access 7 (2019) 54958–54968. doi:10.1109/ACCESS.2019.2912956.

[10] S. D. Da Cruz, H.-P. Beise, U. Schröder, U. Karahasanovic, A theoretical investigation of the detection of vital signs in presence of car vibrations and radar-based passenger classification, IEEE Transactions on Vehicular Technology 68 (2019) 3374–3385.

[11] S. D. Da Cruz, H.-P. Beise, U. Schröder, U. Karahasanovic, Detection of vital signs in presence of car vibrations and radar-based passenger classification, in: 2018 19th International Radar Symposium (IRS), IEEE, 2018, pp. 1–10.

[12] J. Tu, T. Hwang, J. Lin, Respiration rate measurement under 1-d body motion using single continuous-wave doppler radar vital sign detection system, IEEE Transactions on Microwave Theory and Techniques 64 (2016) 1937–1946. doi:10.1109/TMTT.2016.2560159.

[13] M. Mercuri, I. R. Lorato, Y.-H. Liu, F. Wieringa, C. V. Hoof, T. Torfs, Vital-sign monitoring and spatial tracking of multiple people using a contactless radar-based sensor, Nature Electronics 2 (2019) 252–262.

[14] A. Vilesov, P. Chari, A. Armouti, A. B. Harish, K. Kulkarni, A. Deoghare, L. Jalilian, A. Kadambi,

Blending camera and 77 ghz radar sensing for equitable, robust plethysmography., ACM Trans. Graph. 41 (2022) 36–1.
[15] Q. Hu, Q. Zhang, H. Lu, S. Wu, Y. Zhou, Q. Huang, H. Chen, Y.-C. Chen, N. Zhao, Contactless arterial blood pressure waveform monitoring with mmwave radar, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8 (2024) 1–29.
[16] Z. Wu, Y. Xie, B. Zhao, J. He, F. Luo, N. Deng, Z. Yu, Cardiacmamba: A multimodal rgb-rf fusion framework with state space models for remote physiological measurement, IEEE Transactions on Instrumentation and Measurement (2025).
[17] J. Speth, N. Vance, P. Flynn, A. Czajka, Non-contrastive unsupervised learning of physiological signals from video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14464–14474.
[18] E. M. Nowara, D. McDuff, A. Veeraraghavan, Systematic analysis of video-based pulse measurement from compressed videos, Biomedical Optics Express 12 (2020) 494–508.
[19] S. Sachdeva, Fitzpatrick skin typing: Applications in dermatology, Indian journal of dermatology, venereology and leprology 75 (2009) 93.