

Extraction of Blood Volume Pulse Morphology from Facial Videos Using an LSTM-Based Temporal Encoder-Decoder Model

Jonathan Tyler

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

A. Lynn Abbott, Chair

Abhijit Sarkar

Creed Jones

February 12, 2025

Blacksburg, Virginia

Keywords: Temporal Encoder-Decoder, Machine Learning, Remote Photoplethysmography,
iPPG, BVP, Instantaneous Heart Rate

Copyright 2025, Jonathan Tyler

Extraction of Blood Volume Pulse Morphology from Facial Videos Using an LSTM-Based Temporal Encoder-Decoder Model

Jonathan Tyler

ABSTRACT

This thesis introduces a method for extracting blood volume pulse (BVP) signals from facial videos, moving beyond basic heart rate estimation to capture full pulse waveforms. Our approach adapts techniques from audio signal separation and applies them to video, using a machine learning model capable of processing complex time-based data. By incorporating both regular RGB (red, green, blue) and infrared (850nm, 940nm) video, we enhance the quality of the extracted signals, making signal extraction more reliable under different lighting conditions. This method not only improves accuracy in measuring real-time heart rate but also captures unique heart patterns that could support biometric identification. We evaluate our approach on the VT Tricam iPPG dataset, demonstrating improvements over prior methods. Our model achieves up to a 9.8% increase in shape metric correlation and up to a 26% reduction in heart rate prediction error compared to our baseline system. Our findings show that this approach effectively recovers detailed BVP shapes from video, paving the way for advancements in health monitoring and identity verification technologies.

Extraction of Blood Volume Pulse Morphology from Facial Videos Using an LSTM-Based Temporal Encoder-Decoder Model

Jonathan Tyler

GENERAL AUDIENCE ABSTRACT

This thesis focuses on how to measure heart signals from facial videos in a way that captures more detail than just average heart rate. We use a machine learning model designed for an audio separation task, adapting it to separate blood flow signals from noise in signals extracted from video of the face. By adding infrared video data along with regular color channels, our method becomes more accurate, especially in low-light situations. This allows us to not only calculate a person's heart rate more precisely but also to create unique patterns from their heartbeat, which could help in personal identification. Through testing, we show that our method can successfully extract clear heart signals from video, opening up new uses for health monitoring and security.

Dedicated to Virginia Tech.

Acknowledgments

I would like to extend my deepest gratitude to everyone who has supported me along my M.S. journey. In particular, I am thankful to my thesis advisors, Dr. Abbott and Dr. Sarkar, for their expert guidance and consistent encouragement, which have been pivotal in shaping this research.

I also appreciate the commitment of the faculty and staff at Virginia Tech for providing an environment rich in resources and academic support. My sincere thanks go to the Commonwealth Cyber Initiative (CCI) and the National Science Foundation (NSF) for their financial backing that made this work possible.

Lastly, I am grateful to my family and friends, whose steady encouragement and support have been indispensable throughout this process.

Contents

List of Figures	viii
1 Introduction	1
2 Review of Literature	5
2.1 PPG Background	5
2.1.1 iPPG	6
2.1.2 Recognition of Bias	7
2.2 Previous work on iPPG	8
2.2.1 Hybrid Spectrogram-Waveform Approach	10
2.2.2 Hybrid Transformer-Based Architecture	11
2.2.3 NIR iPPG	12
3 Method	14
3.1 Data Conversion	14
3.2 Data Preprocessing	17
3.2.1 Preprocessing the output/ground truth	17
3.3 Neural Network Model	18
3.3.1 LSTM Temporal Encoder-Decoder Model	18

3.3.2	Loss Function	25
3.4	Postprocessing	27
3.4.1	Peak Detection Method using Pan-Tompkins and Interpolation	27
3.4.2	Smoothing of Instantaneous Heart Rate Using Kalman and Savitzky-Golay Filters	32
3.5	Evaluation Metrics	37
3.5.1	Morphology Metrics	37
3.5.2	Heart Rate Metrics	39
4	Results	42
4.1	Datasets and Experimentation	42
4.2	PPG Shape Recovery with LSTM	45
4.3	Heart Rate Prediction Performance	49
4.4	Impact of NIR on Shape Recovery and Heart Rate prediction	51
4.4.1	Impact of NIR on diversity robustness	55
4.5	Comparison with existing methods	57
4.6	Dataset Limitations	58
5	Conclusions	59
5.0.1	Future Work	60
	Bibliography	62

List of Figures

1.1	Representation of this paper’s iPPG workflow. In the first step, a face video is recorded by synchronized cameras. Skin color variation data extracted from the video is used to estimate the BVP signal with the iPPG system.	2
2.1	(a) Example of PPG morphology, including systolic and diastolic peaks, with time steps shown on the horizontal axis. The lower half zooms into a single pulse, highlighting key features such as the diastolic peaks, systolic peaks, and dicrotic notch (Figure credit: [1]). (b) Representation of the skin reflection model including representations of the specular reflection and diffuse reflection. Only diffuse reflections contain pulsatile information (Figure credit: [2]).	6
2.2	Structure of the Demucs system for source separation [3]. The input is a stereo audio sequence, and the output consists of separated waveforms representing distinct instruments.	9
3.1	(Left) Example of MediaPipe Facemesh[4] landmark detection. Each mesh forms a mask of vertices that can be used to identify locations on the face. Regions of Interest are outlined here in green. (Right) Complementary example of face mesh in 850nm infrared video.	16
3.2	Comparison of before/after average amplitude normalization	19

3.3 Diagram representing the structure of our model. This design is inspired by the LSTM Demucs variant [3]. (a) Overall architecture that includes a U-Net structure and bridges between encoder and decoder blocks with skip connections. LSTM layers are placed in between the encoder and decoder modules, and a linear layer is used to bring the output size back to the output size of the encoder module. The input to the system is a temporal sequence of 3 to 5 video channels (red, green, blue, or NIR 850nm, 940nm) across 5 regions of interest. The output of the system is a temporal PPG waveform of the similar length to the input sequence. (b) Encoder blocks are composed of a 1D convolutional layer with ReLU activation, followed by another 1D convolutional layer utilizing gated linear unit (GLU) activation. GLU activation modulates the output, introducing enhanced non-linearity to the model. The decoder blocks mirror the encoder blocks in structure, but in reverse. Decoder blocks include a 1D convolutional layer with GLU activation, followed by a transposed 1D convolutional layer with ReLU activation. . . . 24

3.4	Network structure of prior model by Li et al. [1], inspired by the LSTM Demucs variant [3]. (a) Overall architecture that includes a U-Net structure and bridges between encoder and decoder blocks with skip connections. LSTM layers are placed in between the encoder and decoder modules, and a linear layer is used to bring the output size back to the output size of the encoder module. The input to the system is a temporal sequence of 3 to 5 video channels (red, green, blue, or NIR 850nm, 940nm) across 5 regions of interest. (b) Encoder blocks are composed of a 1D convolutional layer with ReLU activation, followed by another 1D convolutional layer utilizing gated linear unit (GLU) activation. GLU activation modulates the output, introducing enhanced non-linearity to the model. The decoder blocks mirror the encoder blocks in structure, but in reverse. Decoder blocks include a 1D convolutional layer with GLU activation, followed by a transposed 1D convolutional layer with ReLU activation.	26
3.5	Instantaneous Heart Rate prediction error by heart rate due to quantization error at several different sampling frequencies	29
4.1	Spectral response for VT Tricam iPPG video channels (combining filters and sensor sensitivity information from equipment specifications[5, 6, 7, 8])	43

4.2	Pulse shape recovery for subjects P315 (left) and P320 (right) from the VT Tricam iPPG dataset. Each column represents a different subject, with results shown for various video segment (e.g. Seg8, Seg12, Seg16). The top row corresponds to the ground truth (GT) PPG signal, the middle row shows the estimated (Est) signal from Li et al.’s model trained on the new dataset, and the bottom row presents the estimated signal from our proposed model. All results are derived from RGB channel inputs.	47
4.3	(a) Shape recovery results of subject ‘P315’ from VT Tricam iPPG dataset (amplitudes normalized for comparison). Two segments from the testing set with session 1 lighting and from task 2 are displayed. The top left subplot shows the detected pulse shape of the measured PPG signal, with solid lines representing the average waveform and shaded areas showing ± 1 standard deviation. The remaining subplots show the results of models with different combinations of video channels among standard Red, Green, and Blue channels (RGB), and 850nm, 940 Near Infrared(NIR) channels. All models were trained on the same training set using data from the VT Tricam iPPG dataset. (Continued on Next Page)	53
4.3	(Continued from Previous Page.) (b) Shape recovery results of subject ‘P320’ from VT Tricam iPPG dataset (amplitudes normalized for comparison). Again, two segments from the testing set with session 1 lighting and from task 2 are displayed.	54

Chapter 1

Introduction

In an era where non-invasive health monitoring is becoming increasingly essential, imaging photoplethysmography (iPPG) stands out as a transformative technology. Physiological signals play a crucial role in evaluating vital health metrics, with most current monitoring systems relying on wearable devices designed for activity tracking, fitness monitoring, and health assessment [9]. However, there is an increasing push toward developing non-contact techniques for monitoring vital signs and physiological functions [10, 11]. With advancements in deep learning, it is now possible to extract indicators such as heart rate from camera-based systems, though current methods have limitations in shape precision and reliability [12]. Compared to wearable devices, which may be cumbersome or invasive, camera systems present a less intrusive and more accessible method for passive, continuous monitoring [13]. Such systems have great potential in remote healthcare services and various surveillance scenarios, where monitoring can occur without physical interaction with subjects.

Research into methods for deriving physiological signals from video data has advanced rapidly in recent years [12, 14, 15, 16]. One established technique is photoplethysmography (PPG), which monitors blood volume changes using light measurements. This technique is widely used in consumer-grade devices, such as fitness trackers and smartwatches, to capture vital metrics like heart rate [17]. Our study focuses on a variation of this method, known as iPPG or remote PPG, which uses video recordings of subjects to infer physiological data rather than relying on direct skin contact via a sensor. The iPPG workflow that is used in this

work is briefly summarized in Figure 1.1. To address the limitations of current methods, this research specifically targets the detection of blood volume changes, aiming to capture blood volume pulses (BVP) that coincide with each heartbeat.

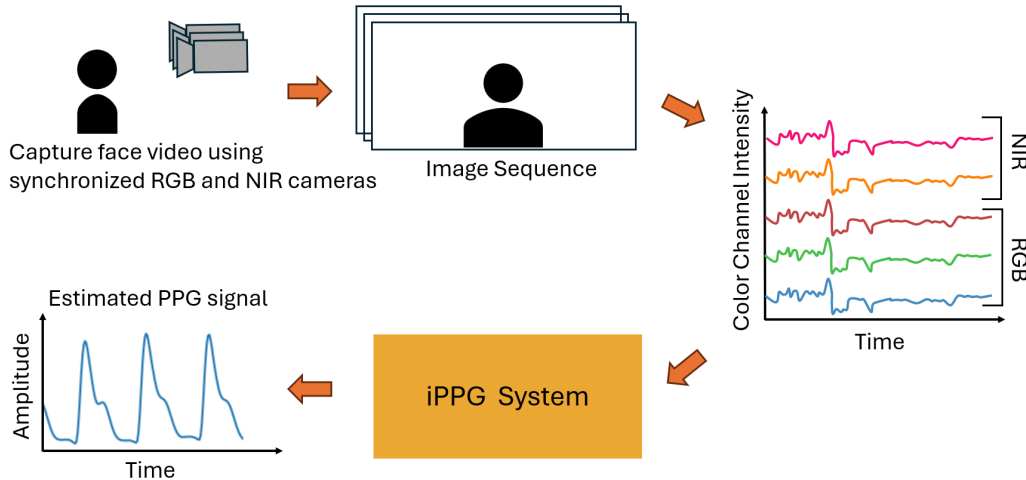


Figure 1.1: Representation of this paper’s iPPG workflow. In the first step, a face video is recorded by synchronized cameras. Skin color variation data extracted from the video is used to estimate the BVP signal with the iPPG system.

To overcome these challenges, we introduce a novel approach to remote PPG signal extraction by utilizing Long Short-Term Memory (LSTM) layers within a 1D encoder-decoder neural network model. LSTM layers, a type of recurrent neural network well-suited for handling time-series data, enhance the temporal resolution of the extracted signals while effectively reducing noise. This marks the first application of this architecture to the problem of extracting detailed BVP morphology from video data, providing a more granular and accurate assessment of cardiovascular activity.

Additionally, we explore the integration of infrared (IR) channels into the signal extraction process. Infrared light offers deeper skin penetration and different reflection characteristics compared to visible light, providing a clearer signal in some cases. Our experiments demonstrate that adding IR channels significantly enhances the quality of the recovered

BVP waveform, particularly in terms of noise reduction and improved shape preservation.

Recognizing the demographic variations that can influence the effectiveness of IR channels, we evaluate the robustness of our approach across diverse skin tones. This focus on inclusivity ensures that our system is effective for a wide range of users, contributing to a more equitable understanding of non-contact physiological monitoring methods.

Several researchers have used iPPG to estimate average heart rate methods, often using significant post processing [12, 18, 19, 20, 21]. A few more recent studies have attempted to directly estimate the BVP waveforms using iPPG [22, 23]. This work builds on the work of Li et al.[23] in using an encoder-decoder model to extract the PPG signal from time series color data, and provides results that improve upon it by most metrics.

Key Contributions:

- **Technical Advances:** We apply LSTM layers within a 1D encoder-decoder model to remote PPG signal extraction, providing improved temporal resolution and noise reduction.
- **Enhancement with IR Channels:** We demonstrate that incorporating near-infrared(NIR) channels enhances BVP signal recovery over RGB-only variants in some experiments, with shape metrics such as cross correlation improving by up to 4.5% and heart rate prediction error metrics showing up to a 51% reduction.
- **Demographic Inclusivity:** We evaluate the robustness of IR-based signal extraction across skin tones, underscoring the inclusivity of our approach.
- **Instantaneous HR Applications:** We address the challenge of instantaneous heart rate calculation, going beyond the traditional analysis of average heart rate to allow granular and short-exposure cardiovascular evaluations. This is not a usual approach

for iPPG systems, but advancements in this area may lead to potential for real-time operation.

By addressing these key challenges, this research aims to push the boundaries of what non-contact monitoring can achieve. The organization of this thesis is as follows: Chapter 2 discusses the prior literature related to this thesis. Chapter 3 describes the methods used by our system and discusses each step in the overall workflow. Chapter 4 details the configurations and comparisons we performed and provides an analysis of the findings. Chapter 5 provides a summary of our work, and the conclusions drawn from it.

Chapter 2

Review of Literature

2.1 PPG Background

In recent years, photoplethysmography (PPG) signals have gained significant attention due to their non-invasive nature and the valuable insights they provide into cardiovascular function. PPG signals, commonly used in health-monitoring applications, have also been explored as a method for biometric authentication due to the distinct waveform patterns generated by the heart's systolic and diastolic phases [24].

The blood volume pulse (BVP) recovered from PPG has a shape that is determined by an individual's physiology. Because of the relation to physiological condition, the shape can provide insight into a person's health, emotional state, and other factors present in their physiology. These waveforms, as shown in Figure 2.1a, can form unique enough biometric templates that they can allow comparisons for authentication purposes [25, 26, 27].

Traditional PPG systems, however, often require physical contact with a person's skin. These methods include fingertip sensors and wearables that have sensors directly against a subject's skin.

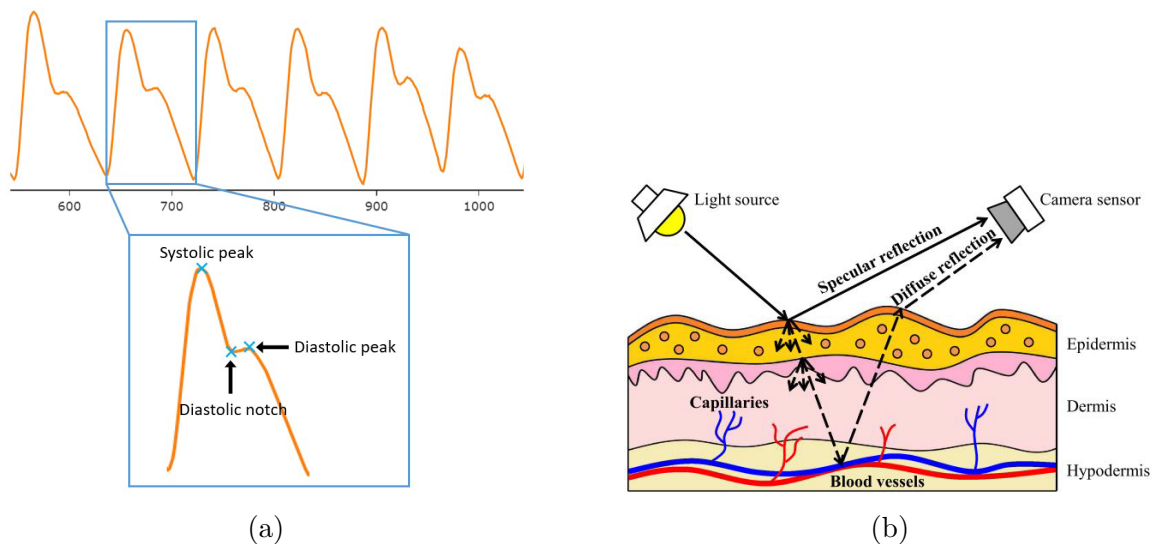


Figure 2.1: (a) Example of PPG morphology, including systolic and diastolic peaks, with time steps shown on the horizontal axis. The lower half zooms into a single pulse, highlighting key features such as the diastolic peaks, systolic peaks, and diastolic notch (Figure credit: [1]). (b) Representation of the skin reflection model including representations of the specular reflection and diffuse reflection. Only diffuse reflections contain pulsatile information (Figure credit: [2]).

2.1.1 iPPG

A key advantage of camera-based PPG biometrics is that the signals are not visible to the human eye, making them more secure than traditional biometric modalities such as fingerprint or facial recognition systems [28]. Additionally, the ability to extract PPG signals remotely using video cameras enhances their appeal for short-exposure, non-contact monitoring [12]. However, extracting clean PPG signals from video remains challenging due to external noise sources such as lighting variations, camera noise, and motion artifacts. These challenges may disproportionately affect diverse populations or environments where lighting conditions and access to high-quality video equipment vary significantly. For instance, under-resourced settings with inconsistent lighting or older imaging technologies may experience degraded signal quality, potentially impacting the inclusivity of PPG-based solutions in global health

applications [29].

Remote PPG signals extracted from facial videos are further complicated by physiological variations, such as differences in facial vascular distribution and autonomic nervous system control [30, 31]. Temporal variations in cardiovascular activity across different regions of the face can also introduce inconsistencies [20, 21]. Many approaches average signals across the entire face, treating it as a single region of interest (ROI). However, this simplification may reduce the accuracy of the extracted pulse shapes due to the heterogeneous nature of facial vasculature [20]. To address this limitation, some studies have adopted a multi-ROI approach, selecting smaller regions across the face to improve signal localization and accuracy [1].

2.1.2 Recognition of Bias

The challenges associated with these physiological and methodological variations highlight the need for solutions that are adaptable to diverse populations. For example, differences in skin tone, age, or gender may influence the ability to recover heart rate metrics accurately using remote PPG. Individuals with darker skin tones, which absorb more light, may experience reduced signal quality under certain lighting conditions. Similarly, age-related changes in skin structure and vascularity, as well as gender-based physiological differences, could further impact signal extraction. Addressing these factors is essential to ensure that PPG technologies are inclusive and perform consistently across diverse groups.

Bias due to differences in aspects such as skin tone is a well-known challenge in computer vision, commonly in face recognition [32]. Further, this issue is present in any computer vision application that processes human subjects [33, 34].

A discussion paper published by the FDA addresses concerns regarding the accuracy of pulse

oximeters in individuals with different skin tones, highlighting potential biases in PPG-based oxygen saturation estimates [35]. Studies indicate that darker skin pigmentation can lead to overestimation of oxygen saturation, raising concerns about device performance in diverse populations. The FDA proposes improving evaluation methods by incorporating objective pigmentation assessments, such as the Monk Skin Tone (MST) Scale and Individual Typology Angle (ITA) measurement, alongside statistical modeling to mitigate bias [35]. These findings align with broader challenges in iPPG, where variations in skin reflectance impact signal quality and physiological measurement accuracy.

2.2 Previous work on iPPG

Extracting clean PPG signals from video data has driven researchers to explore signal separation techniques such as Independent Component Analysis (ICA) [18, 19] and Non-negative Matrix Factorization (NMF) [36, 37]. While these methods have shown success in other domains, their performance in video-based PPG extraction has been inconsistent due to external influences like lighting changes, respiratory signals, and head movements [2, 38]. Image-based PPG commonly uses the skin reflection model depicted in Figure 2.1b [2], where the sensor captures specular and diffuse reflections. For the purposes of iPPG only the diffuse reflection is used as it contains the pulsatile data.

Recently, deep learning has emerged as a powerful tool to overcome these challenges. One promising model is Demucs, originally developed for music source separation, which employs a fully convolutional architecture with dilated convolutions and Long Short-Term Memory (LSTM) layers [3]. The LSTM layers enable the model to capture long-term dependencies in time-series data, which is crucial for separating PPG signals from video noise. The structure of the LSTM variation Demucs model is depicted in Figure 2.2. Researchers have

adapted Demucs’ architecture to PPG extraction, using it to effectively filter out noise while preserving the morphology of the PPG signal [1].

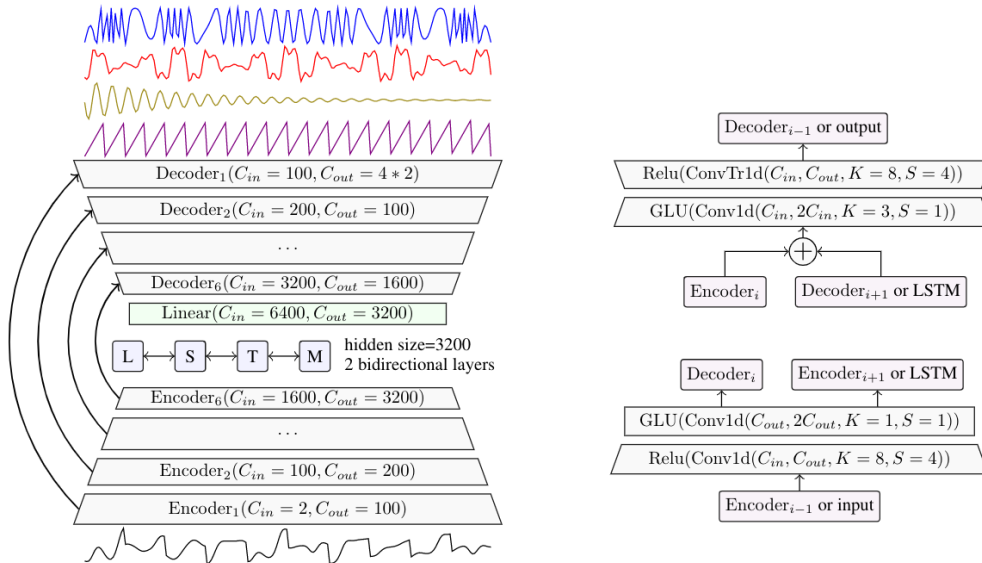


Figure 2.2: Structure of the Demucs system for source separation [3]. The input is a stereo audio sequence, and the output consists of separated waveforms representing distinct instruments.

Additional variations of the Demucs model have been developed to address the challenges of more complex signal separation tasks. One notable variation is the hybrid spectrogram-waveform model, which integrates both time-domain and frequency-domain representations for source separation [39]. This model harnesses the complementary benefits of these two representations: the spectrogram provides detailed frequency-domain insights, making it useful for isolating rhythmic components such as PPG signals, while the waveform maintains high temporal resolution, preserving the signal’s fine structure. The dual-input architecture enhances the separation of intricate signals that may overlap in frequency and time.

2.2.1 Hybrid Spectrogram-Waveform Approach

One additional variation is the Hybrid Spectrogram-Waveform Architecture, introduced in [39], which combines spectral and temporal processing pathways to leverage the strengths of both frequency and time-domain representations. In this model, the input waveform is simultaneously passed through a temporal encoder and a spectral encoder. The latter first transforms the waveform into a spectrogram using the Short-time Fourier Transform (STFT), capturing frequency-specific information. After encoding, the two representations are summed when their dimensions align, allowing the model to benefit from the complementary nature of these domains. The decoder follows a similar structure, with the spectral output converted back to the time domain via the Inverse Short-time Fourier Transform (ISTFT) before merging with the temporal output to produce the final signal.

This architecture holds potential relevance to iPPG signal extraction tasks, where balancing frequency-domain insights with precise temporal dynamics is critical. The spectrogram processing could enable the model to better distinguish PPG signals from noise based on their frequency characteristics, such as separating physiological signals from ambient lighting variations. Meanwhile, the temporal branch maintains the resolution necessary for recovering subtle shape features, including the dicrotic notch, which is essential for detailed morphological analysis.

While the Hybrid Spectrogram-Waveform Architecture demonstrates a compelling theoretical framework for signal separation tasks, its practical application to iPPG remains an open question. Future research could explore strategies such as data augmentation or hybrid pre-training approaches to overcome these challenges and further assess its potential for robust PPG signal isolation under noisy conditions.

2.2.2 Hybrid Transformer-Based Architecture

A third variation is the the Hybrid Transformer Demucs architecture, as introduced in [40]. This model builds on the original Hybrid Demucs by integrating a cross-domain Transformer Encoder between the encoder and decoder layers, while retaining the dual-branch U-Net structure for processing temporal and spectral features. Self-attention mechanisms, employed within the Transformer Encoder, allow the model to capture long-range dependencies and interactions across domains. The hybrid design merges temporal and spectral features at a shared layer, enabling a unified representation before decoding.

Transformers, known for their self-attention mechanism, excel at capturing long-range dependencies in sequential data without the need for recurrent structures. This capability is particularly beneficial for tasks involving prolonged sequences where global context is necessary. In the context of PPG extraction, such models could improve the detection of subtle and complex temporal patterns that might be less effectively modeled by traditional LSTM layers. However, these transformer-based variations introduce higher computational demands and require larger datasets to mitigate overfitting, posing challenges for specialized PPG datasets with limited sample sizes.

In the context of iPPG, this architecture’s cross-domain approach and self-attention mechanisms are theoretically promising. The ability to model long-range dependencies could aid in isolating subtle temporal features of the PPG signal while mitigating noise components like quantization noise and positioning noise. Additionally, the integration of temporal and spectral domains could provide richer representations of the signal, potentially improving the separation of PPG from overlapping noise.

2.2.3 NIR iPPG

In addition to architectural advancements, the integration of infrared (IR) channels has been explored to enhance signal quality in video-based photoplethysmography (PPG) extraction. Infrared light, particularly in the near-infrared (NIR) spectrum, has unique characteristics that make it highly suitable for biomedical sensing applications. Unlike visible light, which is subject to greater scattering and absorption by melanin and other skin components, NIR light penetrates deeper into tissue structures, allowing for more reliable signal acquisition even under suboptimal lighting conditions [41].

This deeper tissue penetration is particularly beneficial for addressing challenges associated with variations in skin tone, age, and other demographic factors that influence visible light absorption and reflection. Studies in other domains, such as agricultural and remote sensing applications, have demonstrated that NIR wavelengths (e.g., 800–900 nm and beyond) provide more accurate measurements of biological material by minimizing interference from external lighting conditions [?]. This aligns with the objectives of PPG research, where consistent and accurate signal extraction is critical for robust physiological monitoring.

Furthermore, NIR technology has been employed in hyperspectral imaging to enhance signal detection through improved spectral resolution, reducing the impact of noise and environmental variability [?]. By incorporating IR channels, video-based PPG methods can potentially overcome biases that would otherwise affect individuals with diverse skin tones, ensuring more inclusive and equitable biometric monitoring solutions. The ability of NIR wavelengths to function effectively in varying illumination environments further strengthens its applicability in real-world, non-contact PPG measurements.

This contextual detail reinforces the motivation for integrating infrared wavelengths into PPG research, positioning it as a key technological enhancement for improving accuracy,

inclusivity, and robustness in biometric sensing.

NIR iPPG has shown promise in improving signal reliability under challenging conditions like low-light environments, motion artifacts, and ambient light fluctuations. Studies have identified optimal NIR wavelengths, such as 799 nm and 861 nm, for reducing noise and enhancing accuracy[42]. Research on applications like driver monitoring has demonstrated the benefits of narrow-band filtering (e.g., 940 nm) to minimize ambient light interference [43]. Additionally, the use of models like the TURNIP model, which combines time-series U-Net and recurrence, has improved robustness in uncontrolled lighting conditions [44]. These advancements underscore NIR's potential to enhance non-contact cardiovascular monitoring by reducing artifacts and improving signal extraction.

The development of advanced deep learning architectures, signal separation techniques, and the integration of multiple ROIs and infrared channels represent important strides in improving the accuracy and reliability of remote PPG extraction. These advancements lay a strong foundation for future research, particularly in the fields of biometric authentication and short-exposure cardiovascular health monitoring. Addressing challenges related to diversity, such as skin tone, age, and gender, remains crucial to ensuring that PPG technologies are both inclusive and effective in all populations.

Chapter 3

Method

This section outlines the key steps involved in our iPPG signal extraction approach, starting with data conversion, followed by preprocessing, neural network modeling, and post-processing. We begin by transforming 4D video frames into a 2D time-series of ROI-based color values, enabling more focused analysis. Preprocessing is then applied to both the input and ground-truth signals to remove noise and ensure consistency. We employ a neural network model, incorporating a U-Net[45] architecture with LSTM layers to capture long-term temporal dependencies. Finally, post-processing steps, including morphology-based metrics and peak detection, are used to refine the extracted signals and evaluate their accuracy. This systematic approach allows us to isolate the PPG signal from noisy input data and capture subtle facial color variations caused by cardiovascular activity.

3.1 Data Conversion

Data Conversion is the first phase of our process, and involves extracting time-series data from video frames, which is essential for subsequent signal processing. We begin by converting a 4D image sequence with dimensions (L, W, H, C) —where the number of frames is represented by L , the width and height of each frame by W and H , and the number of color channels by C —into a 2D time-series with dimensions $(L, N \cdot C)$. Here, N represents the number of sampled Regions of Interest (ROIs). This step reduces variability by focusing

the analysis on specific ROIs, minimizing the influence of factors such as movement or hair color.

ROI Selection

A primary challenge in video-based iPPG is the subtlety of facial color changes caused by cardiovascular activity. At each camera pixel, this weak signal is further degraded by noise and typically only sampled with around six quantization levels, resulting in a noisy, low-quality signal at each pixel. To mitigate this, a spatial average of pixels is commonly used to reduce quantization errors from the camera, as highlighted in previous research [46]. Traditionally, many approaches average color over large ROIs, but this can introduce problems due to temporal differences in cardiovascular activity across various regions of the face [20, 21].

To address this issue, we chose to average color across multiple small ROIs, which allows for a more localized and temporally precise measurement of facial color changes, following a similar decision by previous work[1, 23]. Averaging was chosen for aggregation to be robust to small outliers while presenting a smooth, continuous estimate. We also chose similar ROIs to the foundational work of Li et al.[1, 23], based on blood flow and common occlusions of the face.

MediaPipe Face Mesh

MediaPipe Face Mesh model[4] is a model developed by Google for 3D face landmark estimation. Instead of a simple bounding box face detection model, this approach directly plots each facial landmark. A visualization of the face mesh and some extracted regions of interest are shown in Figure 3.1.

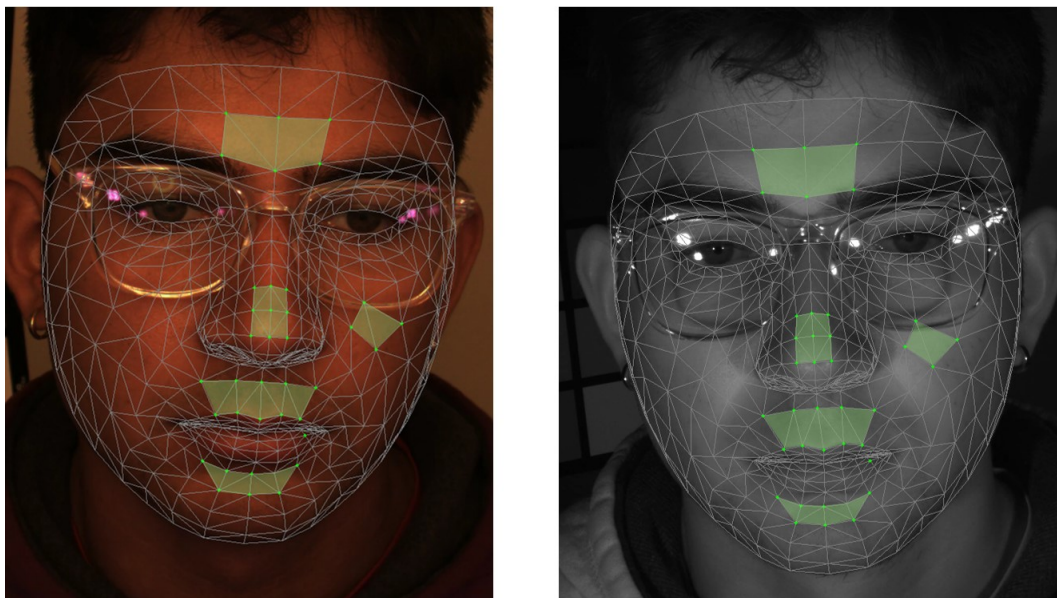


Figure 3.1: (Left) Example of MediaPipe Facemesh[4] landmark detection. Each mesh forms a mask of vertices that can be used to identify locations on the face. Regions of Interest are outlined here in green. (Right) Complementary example of face mesh in 850nm infrared video.

This model fits a high-accuracy 3D triangular mesh to the face in the image. The mesh consists of 468 points which, when mapped by the MediaPipe Face Mesh model, each correspond to a pixel location and depth. Using the pixel locations output by the model, we can identify the locations of our regions of interest in a given camera frame.

MediaPipe Face Mesh has been widely adopted in research and industry due to its efficiency, accuracy, and real-time performance. The model is trained on a large and diverse dataset, ensuring robustness across various facial structures, lighting conditions, and poses. Based on our observations, glasses and other obstructions have had minimal effect on the mesh estimation.

We use the face mesh in our data conversion approach to convert video data into time-series.

The dataset we use is a recently developed database known as VT Tricam iPPG. This dataset was developed in response to challenges encountered with existing data sets such

as DEAP[47]. The main challenges that the dataset intends to address are the inclusion of age and skin tone diversity, minimal facial occlusion, and controlled lighting with a color calibration card. The dataset also provides higher resolution video and provides 850nm and 940nm infrared video in addition to the traditional RGB video. The dataset consists of 18 persons with a total of 126 5-minute videos.

In our work, we extract the average value for each triangle in the mesh, later selecting the triangles that make up an ROI and re-constructing it. This approach allowed for experimentation with ROIs without the need for frequent video processing.

3.2 Data Preprocessing

Both the ground-truth BVP (Blood Volume Pulse) signal extracted from a fingertip (contact) sensor and the input temporal ROI color data require preprocessing. We preprocess the ground-truth signal due to the significant noise present in the BVP sensor data, particularly from low-frequency components. To mitigate this, we apply bandpass filtering to remove unwanted noise. Additionally, we normalize each pulse to ensure consistent starting, ending, and average amplitudes, which improves the efficiency of model training.

Similarly, the input temporal data undergoes bandpass filtering to filter out constants that are unrelated to the BVP signal.

3.2.1 Preprocessing the output/ground truth

Owing to physical limitations of the sensors, raw PPG data often contains substantial low-frequency noise. This noise can interfere with reconstructing the shape of individual pulses, particularly when using the Mean Squared Error (MSE) loss function during model training.

To address this, we preprocess the raw sensor data by applying a bandpass filter using FFT (Fast Fourier Transform) tools. Following the approach of prior research [1], we retain the first 45 terms of the FFT. By applying a bandpass filter, the shape and phase of each individual pulse is preserved, but the low-frequency noise is removed.

The variation in pulse amplitudes present another challenge within the ground-truth PPG data, which is primarily due to changes in sensor contact conditions rather than the physiological signal itself. To correct for this, we normalize the average amplitude of each pulse in the signal. Using peak detection tools, we first isolate individual pulses and then interpolate them to ensure uniform starting, ending, and average amplitudes. After normalization, we reapply bandpass filtering to eliminate any low-frequency noise introduced during the normalization process. Figure 3.2 compares the bandpass-filtered and normalized PPG data, showing that each pulse now has a consistent average amplitude, while maintaining the shapes and relative amplitudes within the pulse.

3.3 Neural Network Model

The core of our system is the neural network model. The model will estimate the PPG signal using our preprocessed temporal ROI color data as input.

3.3.1 LSTM Temporal Encoder-Decoder Model

The goal of this stage is to extract the PPG signal from a mixture of noise and components that have a linear relationship with the PPG signal using the temporal encoder-decoder neural network. The model we use is based on a modified version of Demucs [3], originally used for audio source separation. The similarities between audio separation and our signal

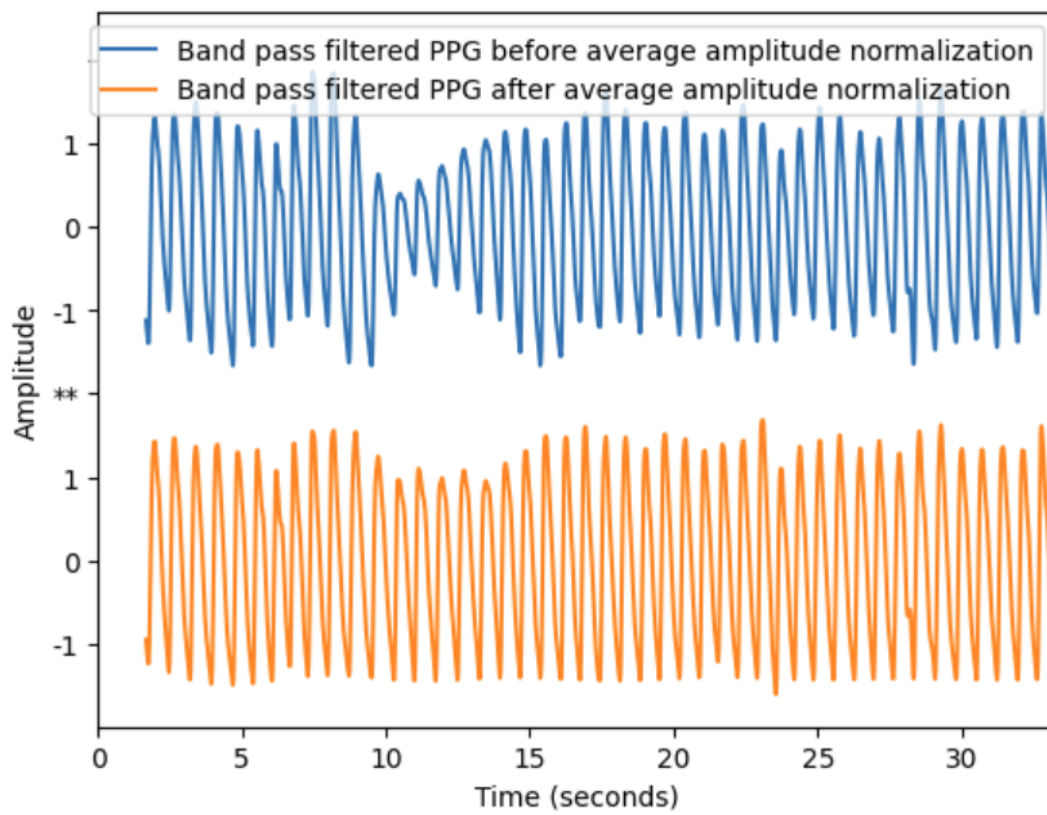


Figure 3.2: Comparison of before/after average amplitude normalization

extraction task, where both require isolating a target signal from noisy input, prompted the adoption of Demucs as the base model.

The neural network takes a data array input of dimensions $(L) \times (N \cdot C)$, where the number of Regions of Interest (ROIs) are represented by N , the number of video channels are represented by C , and sequence length is L . This array is constructed by stacking preprocessed data sequences from all ROIs across their color channels. The model outputs a vector of length L , representing the estimated PPG signal.

Our model employs a U-Net structure with skip connections between encoder and decoder blocks, as shown in Figure 2.2. Encoder blocks starts with a 1D ReLU activation convolution layer, primarily for down-sampling. The core processing happens in the second layer, a 1D GLU activation convolution[48], which has input based activation modulation, adding non-linearity that is particularly useful for temporal sequence processing. The decoder block mirrors this structure, using a 1D GLU activation convolution layer and up-sampling via transposed 1D ReLU activation convolution.

Introduction of LSTM

The key enhancement in our model, compared to prior work in iPPG using similar architectures, is the reincorporation of a LSTM layer. Although Demucs [3] originally included LSTM to handle temporal dynamics in audio source separation, prior iPPG research using Demucs as a base model did not include LSTM [1]. The rationale for introducing LSTM in our model is to improve temporal modeling, particularly to distinguish periodic PPG signals from noise components that exhibit structured temporal patterns.

LSTM, a type of recurrent neural network (RNN), is designed to capture long-term dependencies in sequential data, making it well-suited for time-series tasks. In our model, many

terms can be approximated under assumptions of stationary subjects and fixed illumination, but quantization noise and positioning noise remain significant. These noise components fluctuate over time and can overlap with the PPG signal, necessitating a mechanism to model long-range dependencies.

The inclusion of an LSTM layer between the encoder and decoder provides several advantages over a pure encoder-decoder approach:

- **Enhanced Temporal Consistency:** While an encoder-decoder alone compresses and reconstructs features at each timestep independently, LSTM maintains continuity across time steps, helping to distinguish the periodic nature of the PPG signal from temporally structured noise.
- **Improved Signal Reconstruction:** By retaining information across frames, LSTM helps recover details lost in the encoding process, particularly in preserving PPG waveform features such as systolic peaks and the dicrotic notch.
- **Noise Mitigation:** Positioning noise and other artifacts often exhibit temporal correlations that a standard encoder-decoder may not explicitly capture. LSTM provides a mechanism to learn these dependencies, improving the robustness against structured distortions.

Despite these advantages, previous iPPG studies [1] omitted LSTM when adapting Demucs-based architectures. One likely reason for this omission is the increased computational cost associated with training LSTM layers. Compared to a pure encoder-decoder, adding LSTM significantly increases the number of trainable parameters, leading to longer convergence times and higher memory requirements. During model training, it was observed that the LSTM variant required considerably more iterations to reach stable performance, which may

have deterred prior implementations, especially when computational efficiency was prioritized.

As shown in Figure 3.3, the LSTM layer is positioned between the encoder and decoder to introduce temporal modeling capabilities without altering the fundamental structure of the network. While prior iPPG studies opted for a purely convolutional encoder-decoder structure to expedite training, our results indicate that structured noise patterns in iPPG data benefit from explicit temporal modeling.

In summary, integrating LSTM into the encoder-decoder framework strengthens the model’s ability to extract clean PPG signals by preserving temporal coherence, mitigating structured noise, and improving waveform reconstruction. While the increased computational cost is a valid concern, the trade-off is justified by the improved robustness in signal extraction, aligning with the broader goal of enhancing iPPG signal processing for real-world applications.

Rationale for Chosen Model In the course of developing our temporal encoder-decoder model, we explored several existing variations of Demucs to assess their suitability for the PPG signal extraction task. These variations included hybrid spectrogram-waveform approaches and transformer-based architectures, each offering unique advantages in handling complex signal components and enhancing temporal context.

Given the limitations encountered with more complex architectures, we selected a streamlined version of Demucs that incorporates LSTM layers. This choice strikes a balance between model complexity and the dataset’s available size. Using LSTM, the model gains an enhanced ability to capture temporal dynamics over extended sequences, which is essential to maintain continuity in PPG signal reconstruction. This variation offered sufficient performance improvements without the overfitting risks seen in more data-intensive models.

Ultimately, while hybrid approaches and transformers presented promising theoretical benefits, our findings underscored the importance of tailoring model complexity to data availability. The LSTM-based Demucs model provided the most reliable results within the context of our chosen dataset.

Architecture of Chosen Model The proposed model (Figure 3.3) enhances a U-Net framework by introducing LSTM layers between the encoder and decoder modules, enabling it to better capture temporal dependencies within the input data. The skip connections inherent in the U-Net structure ensure that low-level temporal details extracted by the encoder are directly accessible to the decoder, preserving critical signal features during the reconstruction process. The LSTMs provide the added capability to learn long-term temporal dependencies, which is essential for accurately capturing subtle, sustained dynamics in PPG waveforms. A linear layer bridges the LSTM outputs to the decoder input size, ensuring seamless integration within the network architecture.

The encoder blocks utilize a combination of 1D convolution layers with ReLU and GLU activations. While the ReLU activation supports effective feature extraction, the GLU activation introduces adaptive gating, adding nonlinearity and enhancing the model’s representational power. The decoder blocks invert the encoder’s operations, incorporating transposed convolutions to reconstruct the PPG waveform while retaining the features extracted during encoding.

This work seeks to evaluate the inclusion of LSTMs, a feature absent from the previous model by Li et al. [1] (Figure 3.4). While the prior work’s U-Net structure similarly utilizes skip connections and nonlinear activations for effective signal reconstruction, the lack of LSTMs suggests a reliance on convolutional layers alone for temporal feature modeling. The rationale behind omitting LSTMs in the prior model is unclear, but their inclusion

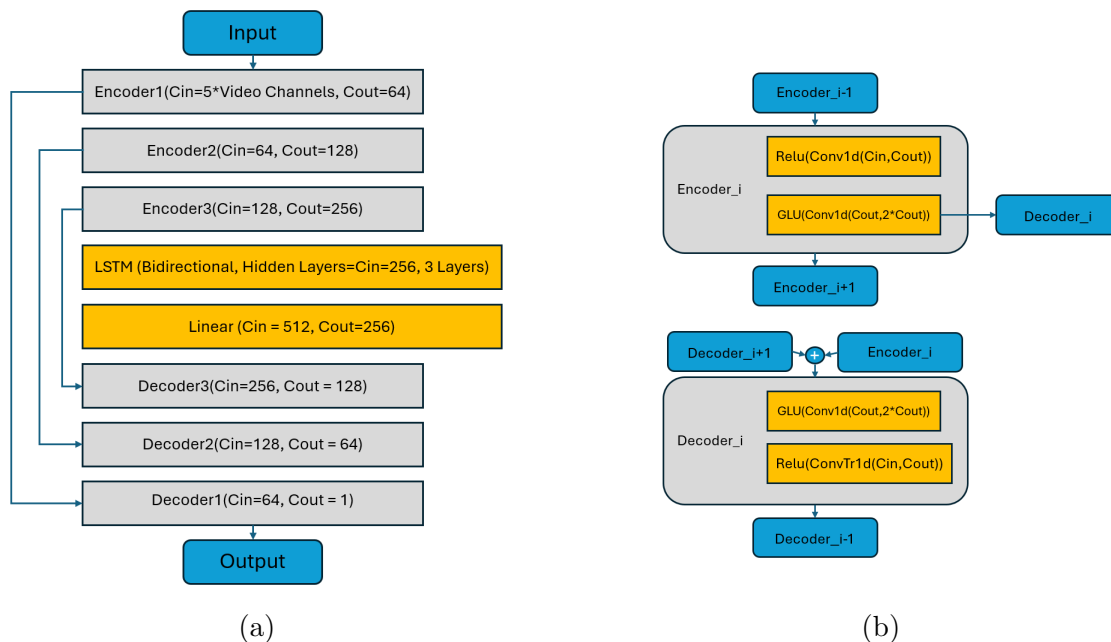


Figure 3.3: Diagram representing the structure of our model. This design is inspired by the LSTM Demucs variant [3]. (a) Overall architecture that includes a U-Net structure and bridges between encoder and decoder blocks with skip connections. LSTM layers are placed in between the encoder and decoder modules, and a linear layer is used to bring the output size back to the output size of the encoder module. The input to the system is a temporal sequence of 3 to 5 video channels (red, green, blue, or NIR 850nm, 940nm) across 5 regions of interest. The output of the system is a temporal PPG waveform of the similar length to the input sequence. (b) Encoder blocks are composed of a 1D convolutional layer with ReLU activation, followed by another 1D convolutional layer utilizing gated linear unit (GLU) activation. GLU activation modulates the output, introducing enhanced non-linearity to the model. The decoder blocks mirror the encoder blocks in structure, but in reverse. Decoder blocks include a 1D convolutional layer with GLU activation, followed by a transposed 1D convolutional layer with ReLU activation.

in this work allows a direct comparison of how explicit temporal modeling impacts PPG reconstruction performance. By introducing LSTMs, the proposed model is designed to address potential limitations of convolution-only architectures, particularly in capturing long-term dependencies crucial for accurate physiological signal reconstruction.

3.3.2 Loss Function

Our loss function is defined as follows:

$$L(y, \hat{y}) = (1 - \lambda) \|y - \hat{y}\| + \lambda \left(\|Re(Y) - Re(\hat{Y})\| + \|Im(Y) - Im(\hat{Y})\| \right) \quad (3.1)$$

This loss function combines mean squared error (MSE) in the time domain with a similar loss in the frequency domain based on the Fast Fourier Transform (FFT). In this loss function representation, y is the predicted output and \hat{y} is the ground truth, and similarly, Y is the Fourier transform of the output and \hat{Y} is the Fourier transform of the ground truth. The operator $Re(\cdot)$ denotes the real components and operator $Im(\cdot)$ denotes the imaginary components extracted from transformed sequences.

For the calculation of the frequency component, we compute the Fourier transforms and the MSE is applied separately to the real and imaginary parts of the output and ground truth transforms. This process is implemented using the PyTorch’s auto-gradient module. The loss function incorporates the frequency domain because some features of the PPG signal, namely the diastolic peak, appear more distinguishable in the frequency domain than in the time domain. As such, including this frequency-domain loss term can help the model capture finer details of the PPG waveform morphology.

The inclusion of a frequency-domain component in the loss function is particularly pertinent

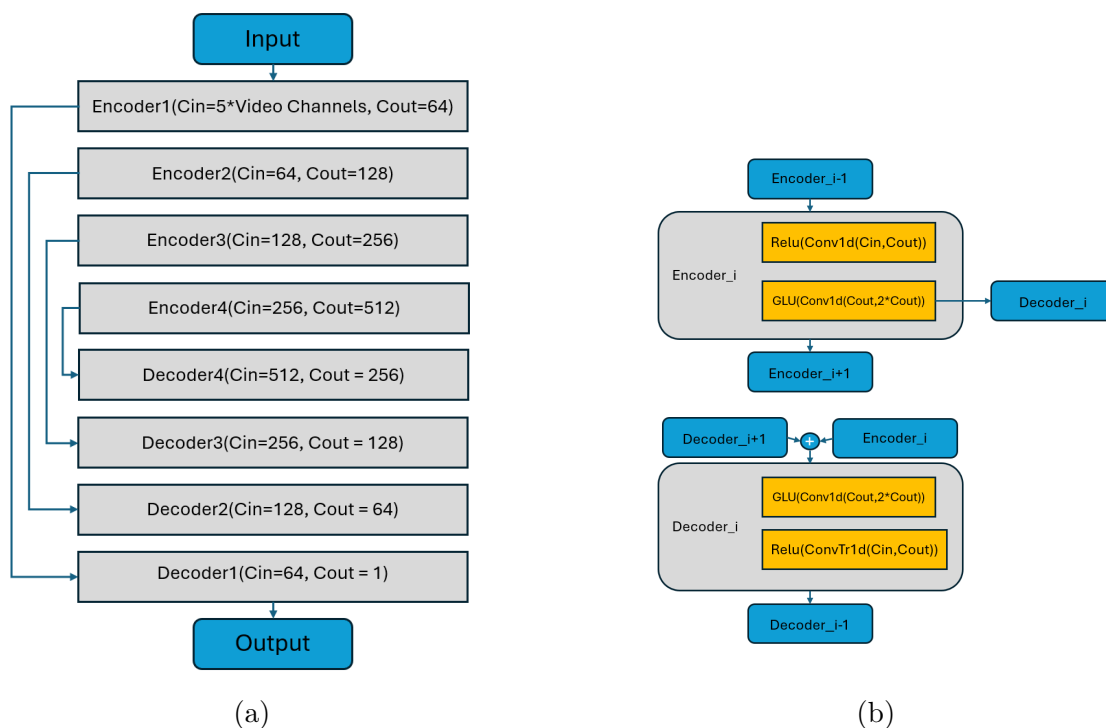


Figure 3.4: Network structure of prior model by Li et al. [1], inspired by the LSTM Demucs variant [3]. (a) Overall architecture that includes a U-Net structure and bridges between encoder and decoder blocks with skip connections. LSTM layers are placed in between the encoder and decoder modules, and a linear layer is used to bring the output size back to the output size of the encoder module. The input to the system is a temporal sequence of 3 to 5 video channels (red, green, blue, or NIR 850nm, 940nm) across 5 regions of interest. (b) Encoder blocks are composed of a 1D convolutional layer with ReLU activation, followed by another 1D convolutional layer utilizing gated linear unit (GLU) activation. GLU activation modulates the output, introducing enhanced non-linearity to the model. The decoder blocks mirror the encoder blocks in structure, but in reverse. Decoder blocks include a 1D convolutional layer with GLU activation, followed by a transposed 1D convolutional layer with ReLU activation.

to shape recovery and signal extraction for several reasons. First, the PPG waveform contains subtle features, such as the diastolic peak and inflection points, that are closely linked to physiological parameters. These features often manifest as distinct harmonics or patterns in the frequency domain, making them more readily identifiable when compared to the time domain, where noise and overlapping signal components can obscure them. By incorporating the FFT loss, the model is encouraged to align not only the overall amplitude and trends of the signal but also the spectral characteristics that define its fine structure.

Moreover, the dual-domain approach effectively balances the trade-offs between time-domain and frequency-domain accuracy. Time-domain metrics like mean squared error ensure that the predicted signal closely follows the temporal progression of the ground truth. However, they may not penalize discrepancies in periodicity or subtle waveform oscillations effectively. The frequency-domain component compensates for this limitation by directly targeting spectral mismatches, thus helping the model capture periodicities and amplitude variations that are critical for applications such as heart rate variability analysis. Ultimately, this hybrid loss function should enhance the model’s ability to recover physiologically meaningful waveforms, contributing to more robust and interpretable signal extraction outcomes.

3.4 Postprocessing

3.4.1 Peak Detection Method using Pan-Tompkins and Interpolation

Our approach to detecting peaks in the PPG signal combines the Pan-Tompkins[49] algorithm with cubic spline interpolation to address the limitations of fixed sampling rates and improve the accuracy of instantaneous heart rate (IHR) calculations. This method is espe-

cially important given that small errors in peak detection can lead to significant deviations in heart rate estimation.

The peak detection process begins with the application of the Pan-Tompkins algorithm, which is widely used for detecting sharp transitions in physiological signals like ECG. The raw PPG signal is first bandpass-filtered to remove unwanted noise, followed by differentiation to highlight sharp transitions corresponding to heartbeats. The signal is then squared to amplify these transitions, and a moving window integration smooths the signal, making the peaks more prominent and easier to detect.

Once the signal is preprocessed, we detect peaks using a threshold and a minimum peak distance, both of which are derived from the sampling frequency (f_s). Peaks are initially identified in the smoothed, integrated signal using the `find_peaks` function of the SciPy signal package. However, these initial peaks are aligned with the sampling grid of the integrated signal, which can introduce quantization errors when mapping back to the original PPG signal.

Quantization errors arise because the fixed sampling rate forces peak detection to “snap” to the nearest sample, rather than accurately capturing the true peak location. Given the typical sampling rates for PPG signals (e.g., 30–60 Hz), this snapping can lead to temporal errors of 16 to 33 milliseconds. These small timing errors, when propagated over successive heartbeats, can cause significant inaccuracies in the calculation of IHR, especially at higher heart rates where the interval between beats is shorter. For example, at a heart rate of 120 BPM, a 33-millisecond error can lead to a prediction error of several BPM, as illustrated in Figure 3.5.

To mitigate these errors, we employ cubic spline interpolation to refine the detected peaks. After identifying a peak in the integrated signal, a small window around the peak is extracted,

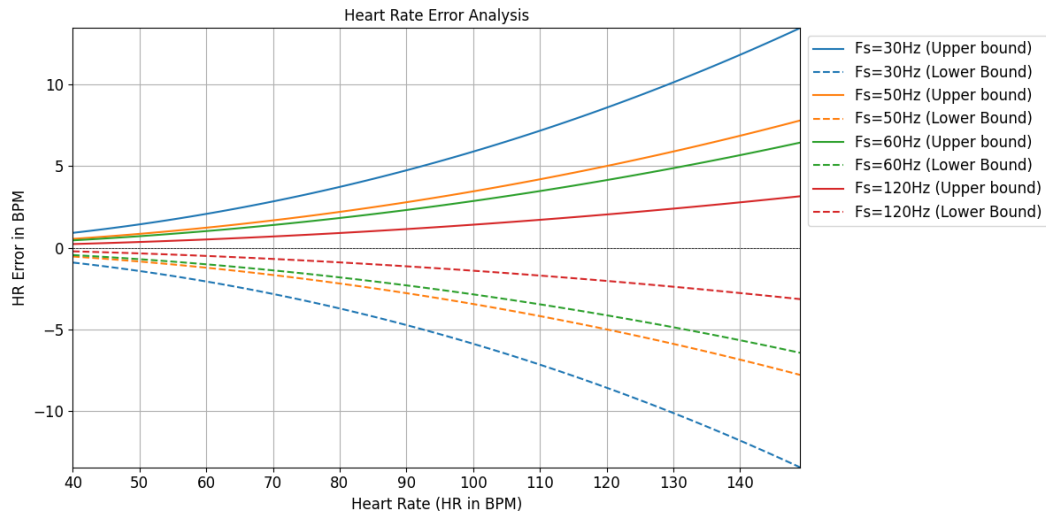


Figure 3.5: Instantaneous Heart Rate prediction error by heart rate due to quantization error at several different sampling frequencies

and cubic spline interpolation is applied. This allows us to create a finer, continuous approximation of the signal within the window, enabling us to localize the peak position with sub-sample precision. By using this interpolation technique, we avoid the snapping effect caused by quantization and significantly improve the temporal resolution of the detected peaks. By refining the peak locations, the cubic interpolation step ensures that the detected peaks better align with the actual physiological events, reducing errors in heart rate estimation that arise from the fixed sampling grid.

Figure 3.5 illustrates how the quantization caused by a fixed sampling rate can affect IHR predictions at different heart rate ranges. These variations in predictions demonstrate that without interpolation, the impact of peak detection error increases with heart rate, leading to larger deviations in BPM. In contrast, applying cubic spline interpolation reduces these errors by ensuring that peaks are detected more precisely, regardless of the heart rate range. Accounting for such sources of error is critical in applications that require high accuracy in short-exposure heart rate monitoring, such as fitness tracking and health diagnostics.

Quantization errors arise because the fixed sampling rate forces peak detection to “snap” to the nearest sample, rather than accurately capturing the true peak location. This snapping effect results from discretizing a continuous-time signal onto a fixed temporal grid. For a signal sampled at a frequency f_s Hz, the maximum temporal quantization error Δt_{\max} is given by:

$$\Delta t_{\max} = \frac{1}{f_s}.$$

In our setup, fingertip PPG signals are sampled at $f_s = 125$ Hz, which corresponds to a maximum temporal error of 8 ms. However, in the case of iPPG signals extracted from video, the sampling rate is often constrained by video frame rates, typically $f_s = 30\text{--}60$ Hz. This results in larger quantization errors of 16.7–33.3 ms, limiting the temporal resolution of peak detection and IHR predictions.

Quantization errors become particularly significant for instantaneous heart rate (IHR) estimation, as the error in detecting inter-beat intervals (IBI) is inversely proportional to the heart rate and sampling frequency. The expected IBI for a reference heart rate H_{ref} (in beats per minute) is given by:

$$T_{\text{ref}} = \frac{60}{H_{\text{ref}}}.$$

Due to the effect of quantization, we determine that the relation with the predicted heart rate, H_{mes} is given by:

$$T_{\text{ref}} \pm \Delta t_{\max} = \frac{60}{H_{\text{mes}}}$$

Solving for both the measured and reference heart rate, we get the following two equations

$$H_{\text{mes}} = \frac{60}{T_{\text{ref}} \pm \Delta t_{\text{max}}}, H_{\text{ref}} = \frac{60}{T_{\text{ref}}}.$$

We can measure the error in the IHR prediction by comparing the predicted heart rate to the expected reference heart rate:

$$\Delta H = H_{\text{ref}} - H_{\text{mes}} = H_{\text{ref}} - \frac{60}{\frac{60}{H_{\text{ref}}} \pm \frac{1}{f_s}}$$

After simplifying, we determine the relation:

$$\Delta H = \frac{H_{\text{ref}}^2}{60f_s \pm H_{\text{ref}}}$$

For example, at $H_{\text{ref}} = 120$ BPM:

- For PPG at $f_s = 125$ Hz:

$$\Delta H_{\text{max}} = \frac{120^2}{60(125) - 120} \approx 1.95 \text{ BPM.}$$

- For iPPG at $f_s = 30$ Hz:

$$\Delta H = \frac{33.3 \cdot 120^2}{60} \approx 8.57 \text{ BPM.}$$

These calculations highlight how lower sampling rates result in significantly higher errors, especially at higher heart rates.

To address this limitation, we employ cubic spline interpolation to refine peak localization.

After detecting a peak at position t_p in the integrated signal, a small window $[t_p - \Delta, t_p + \Delta]$

is extracted. Cubic spline interpolation is applied within this window to create a continuous representation of the signal:

$$s_{\text{interp}}(t) = \sum_{i=0}^3 c_i (t - t_p)^i,$$

where c_i are the coefficients of the cubic polynomial. The refined peak position t_{refined} is identified as the local maximum of $s_{\text{interp}}(t)$ within the window.

This process allows for sub-sample precision in peak detection, mitigating the snapping effect caused by quantization errors. Figure 3.5 illustrates the impact of sampling rate on IHR prediction errors across various heart rates. Without interpolation, quantization errors lead to significant deviations in BPM, particularly at higher heart rates and lower sampling rates. By applying cubic spline interpolation, we significantly reduce these errors, as shown in the figure.

This step is especially critical for iPPG, where the video frame rate limits the sampling frequency and exacerbates quantization effects. Addressing these errors is essential for accurate heart rate estimation in short-exposure applications such as remote health monitoring and fitness tracking.

3.4.2 Smoothing of Instantaneous Heart Rate Using Kalman and Savitzky-Golay Filters

Once the instantaneous heart rate (IHR) predictions are derived from detected peaks, the data can be subject to fluctuations caused by noise or small errors in peak timing. To enhance the consistency of these IHR predictions, post-processing is applied using Kalman and Savitzky-Golay filters. These filtering techniques aim to smooth the heart rate signal,

reducing spurious variations while preserving the physiological dynamics of the IHR.

Kalman Filtering

The Kalman filter[50] is a powerful recursive algorithm that operates by predicting the next value of the heart rate signal based on prior observations and adjusting this prediction with the current measurement. This approach smooths the signal by balancing process noise (which accounts for inherent variability in heart rate) and measurement noise (which reflects uncertainties in detected peaks). As each new heart rate measurement is received, the Kalman filter updates its estimate, producing a smoothed signal that is less susceptible to sudden, non-physiological spikes.

The Kalman filter operates using two primary equations: the ****prediction step**** and the ****update step****. For a signal x_t at time t :

1. ****Prediction Step:****

$$\hat{x}_t^- = F\hat{x}_{t-1} + Bu_t$$

$$P_t^- = FP_{t-1}F^\top + Q$$

Here, \hat{x}_t^- is the predicted state, F is the state transition matrix, B is the control matrix, u_t is the control input, P_t^- is the predicted error covariance, and Q represents process noise.

2. ****Update Step:****

$$K_t = P_t^- H^\top (HP_t^- H^\top + R)^{-1}$$

$$\hat{x}_t = \hat{x}_t^- + K_t(z_t - H\hat{x}_t^-)$$

$$P_t = (I - K_t H)P_t^-$$

Here, K_t is the Kalman gain, H is the measurement matrix, z_t is the measurement, R is

the measurement noise covariance, and I is the identity matrix. The Kalman filter balances process and measurement noise through Q and R , providing a smoothed estimate of the heart rate. [50]

By carefully tuning the process and measurement variance parameters, the Kalman filter can be adapted to the specific characteristics of the heart rate data. This ensures that genuine changes in heart rate are tracked while high-frequency noise and errors due to peak timing inaccuracies are minimized. The result is a more stable heart rate estimate, particularly useful in continuous monitoring settings where large, rapid variations may otherwise obscure meaningful trends.

Savitzky-Golay Filtering

After Kalman filtering, further smoothing is achieved with the Savitzky-Golay (SG) filter, a polynomial-based smoothing technique that operates over localized windows of the data [51]. The SG filter fits a polynomial to the data points within each window and uses this to produce a smoothed approximation of the signal. This method excels at reducing random noise while preserving key signal features such as the shape and slope of heart rate peaks.

The SG filter can be expressed as:

$$y_i = \sum_{j=-k}^k c_j x_{i+j}$$

where y_i is the smoothed value, x_{i+j} are the data points in the window, and c_j are the coefficients derived from fitting the polynomial to the data. The size of the window ($2k + 1$) and the order of the polynomial are key parameters that determine the balance between smoothing and feature preservation. A carefully chosen window size provides sufficient smoothing to reduce noise while maintaining sensitivity to physiological variations in heart rate. The

polynomial order is typically kept low to avoid overfitting the data while still capturing the underlying trend. This combination of smoothing and trend preservation makes the SG filter an ideal complement to the Kalman filter for post-processing IHR predictions. [51]

Impact on Instantaneous Heart Rate Predictions

Table 3.1 highlights the improvements in heart rate estimation achieved through post-processing with Kalman and Savitzky-Golay filters. The primary goal of these filtering techniques is to reduce noise-induced fluctuations while preserving meaningful physiological dynamics. This is particularly critical for instantaneous heart rate predictions, which are more susceptible to noise due to small timing inaccuracies in detected peaks.

Table 3.1: Heart Rate performance comparison (using our RGB model testing results) with/without filtering using instantaneous predictions and predictions over 15 second clips.

Metrics	No Filtering	with Kalman	with Savitsky-Golay	Both Combined
Instant HR MAE (BPM) ↓	12.13	10.39	11.06	10.24
Instant HR RMSE (BPM) ↓	32.476	21.645	24.429	20.430
Clip HR MAE (BPM) ↓	5.77	6.45	5.74	5.74
Clip HR RMSE (BPM) ↓	9.825	12.193	9.810	9.810

The Association for the Advancement of Medical Instrumentation (AAMI) sets a maximum allowable error of ± 5 BPM for clinical heart rate monitoring [52], though iPPG methods have yet to consistently meet this standard. Reported errors vary depending on factors such as motion, illumination, and HR estimation techniques, typically falling within a broader range [16]. While some approaches achieve lower errors under controlled conditions, they often rely on long averaging windows or filtering techniques not directly comparable to peak-based instantaneous HR estimation.

The Kalman filter provides significant benefits by dynamically adjusting its estimates based on prior measurements and current observations. This recursive smoothing process effectively reduces spurious variations in the signal, leading to lower instantaneous heart rate errors.

The observed reduction in MAE underscores the filter’s ability to balance noise suppression and responsiveness to true physiological changes. The Kalman filter is particularly well-suited for short exposure applications where accurate and stable instantaneous predictions are required.

The Savitzky-Golay filter further enhances the signal by applying localized polynomial smoothing, which excels at preserving the underlying shape of the heart rate waveform while reducing high-frequency noise. When combined with the Kalman filter, this approach yields the most stable and accurate instantaneous heart rate estimates, as evidenced by the reduced errors. The complementary nature of the two filters ensures that both random noise and peak detection inaccuracies are mitigated, resulting in a more reliable signal.

As instantaneous predictions of heart rate are more volatile, they are not common in iPPG applications due to an inherent decrease in the quality of signal compared to fingertip PPG or other contact-based monitoring. To mitigate this effect, some sort of smoothing is often used. In our case, we average the instantaneous heart rate predictions across each 15-second clip to better represent the overall performance.

For clip-wise heart rate calculations, only the Savitzky-Golay filter is applied. The Kalman filter operates on individual instantaneous predictions, smoothing variations at a finer temporal scale. However, when calculating averages over longer intervals, the noise is inherently reduced through the averaging process, making the Kalman filter redundant in this context. By applying only the Savitzky-Golay filter, we ensure adequate smoothing of residual noise while avoiding unnecessary steps that do not provide additional benefits at the clip level.

Overall, the use of both filters for instantaneous predictions capitalizes on their distinct strengths: the Kalman filter’s adaptability to noise and variability, and the Savitzky-Golay filter’s ability to refine local signal features. For clip-level averages, the targeted applica-

tion of the Savitzky-Golay filter maintains data integrity while further reducing residual noise. These results highlight the importance of tailoring filtering strategies to the specific requirements of each metric to optimize the accuracy and reliability of heart rate estimates.

3.5 Evaluation Metrics

3.5.1 Morphology Metrics

For morphology-based metrics, we calculate the mean normalized cross-correlation between the model’s output signals and the ground truth BVP signals for each subject in the dataset. These metrics are computed across the time, frequency, and power domains, providing a comprehensive evaluation of the model’s ability to reproduce the correct signal shape morphology, as explored in prior work [22].

The normalized cross-correlation (NCC) is defined as:

$$ncc(x_{gt}(n), x_{op}(n)) = \frac{\sum_{i=1}^N x_{gt}(n_i)x_{op}(n_i)}{\sqrt{\sum_{i=1}^N x_{gt}^2(n_i)}\sqrt{\sum_{i=1}^N x_{op}^2(n_i)}} \quad (3.2)$$

where $x_{gt}(n)$ is the ground truth signal, $x_{op}(n)$ is the model’s output signal, and N is the total number of samples.

In the time domain, the signals are denoted as $x_{gt}(t)$ for the ground truth and $x_{op}(t)$ for the model output. For the frequency domain, the corresponding signals are computed using the magnitude of the Fast Fourier Transform (FFT):

$$x_{gt}(f) = FFT_{mag}(x_{gt}(t)) \quad (3.3)$$

$$x_{op}(f) = FFT_{mag}(x_{op}(t)) \quad (3.4)$$

where FFT_{mag} represents the magnitude of the FFT applied to the time-domain signals.

Similarly, in the power domain, the signals are represented using power spectral density (PSD) as follows:

$$psd(x(n), f_s) = \lim_{x \rightarrow \infty} \frac{1}{T} \left| \sum_N^{n=1} x_n e^{-i2\pi f n} \right|^2 \quad (3.5)$$

$$x_{gt}(p) = psd(x_{gt}(t), f_s) \quad (3.6)$$

$$x_{op}(p) = psd(x_{op}(t), f_s) \quad (3.7)$$

where psd is the power spectral density, f_s is the sampling frequency, and N is the number of signal samples.

Using these representations, we compute the shape morphology metrics in the time, frequency, and power domains, denoted as smm_t , smm_f , and smm_p , respectively. These metrics are calculated as follows:

$$smm_t = \frac{1}{C} \sum_{i=1}^C ncc(x_i(t)_{gt}, x_i(t)_{op}) \quad (3.8)$$

$$smm_f = \frac{1}{C} \sum_{i=1}^C ncc(x_i(f)_{gt}, x_i(f)_{op}) \quad (3.9)$$

$$smm_p = \frac{1}{C} \sum_{i=1}^C ncc(x_i(p)_{gt}, x_i(p)_{op}) \quad (3.10)$$

where C is the number of candidates in the dataset.

3.5.2 Heart Rate Metrics

This section describes the heart rate (HR) metrics used in our analysis to evaluate the performance of our models. These metrics assess the accuracy and consistency of instantaneous HR predictions as well as average HR predictions over 15-second clips. Additionally, the signal-to-noise ratio (SNR) is used to evaluate the quality of the recovered HR signal.

Mean Absolute Error (MAE) of Instantaneous HR Predictions

The mean absolute error (MAE) measures the average absolute difference between the predicted instantaneous heart rate (\widehat{HR}_i) and the ground truth heart rate (HR_i) for N samples, where N is the total number of instant heart rate predictions across the entire training set:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N \left| \widehat{HR}_i - HR_i \right|$$

Root Mean Square Error (RMSE) of Instantaneous HR Predictions

The root mean square error (RMSE) quantifies the square root of the mean squared difference between the predicted instantaneous heart rate and the ground truth:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\widehat{HR}_i - HR_i \right)^2}$$

MAE of Average HR Over 15-Second Clips

For 15-second clips, the mean absolute error is computed over the averaged instantaneous HR predictions (\widehat{HR}_j) and the corresponding ground truth averages (\overline{HR}_j) across each of the heart rate samples for M clips, where M is the total number of clips in the dataset, each containing several instant heart rate predictions:

$$\text{MAE}_{\text{clip}} = \frac{1}{M} \sum_{j=1}^M \left| \widehat{HR}_j - \overline{HR}_j \right|$$

RMSE of Average HR Over 15-Second Clips

Similarly, the RMSE for 15-second clip averages is calculated as:

$$\text{RMSE}_{\text{clip}} = \sqrt{\frac{1}{M} \sum_{j=1}^M \left(\widehat{HR}_j - \overline{HR}_j \right)^2}$$

Signal-to-Noise Ratio (SNR)

The signal-to-noise ratio (SNR) evaluates the quality of the recovered HR signal by comparing the power of the signal (P_{signal}) to the power of noise (P_{noise}):

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \text{ dB}$$

The signal power (P_{signal}) is computed as the variance of the HR signal, while the noise power (P_{noise}) is the variance of the error between the predicted HR and the ground truth:

$$P_{\text{signal}} = \text{Var}(HR), \quad P_{\text{noise}} = \text{Var}(\widehat{HR} - HR)$$

Interpretation

These metrics collectively provide a comprehensive assessment of model performance. MAE and RMSE measure the accuracy of HR predictions, both instantaneously and over time intervals, while SNR reflects the signal quality, highlighting the model's ability to suppress noise in HR estimation.

Chapter 4

Results

4.1 Datasets and Experimentation

In this study, we employed the VT Tricam iPPG dataset for training and evaluation. This dataset was chosen for its noteworthy characteristics:

- High-resolution videos recorded at a frame rate of 60 frames per second (fps).
- Synchronized RGB and near-infrared (NIR) video streams, enabling a comprehensive dual-modality analysis.
- Minimal facial occlusions and controlled lighting conditions, which help maintain signal integrity.

The dataset includes recordings from 18 participants, with each participant contributing seven five-minute video sessions involving various tasks such as mental math, reading, and object-focused activities. Two of these recordings were captured under different lighting conditions to test model performance under varied scenarios.

The recorded video for the dataset was aggregated from three cameras. The first is an RGB camera, providing red, green, and blue video channels. The remaining two video cameras are monochrome cameras with narrow NIR filters installed, one at 850nm, one at 940nm. The spectral response of each video channel is portrayed in Fig. 4.1.

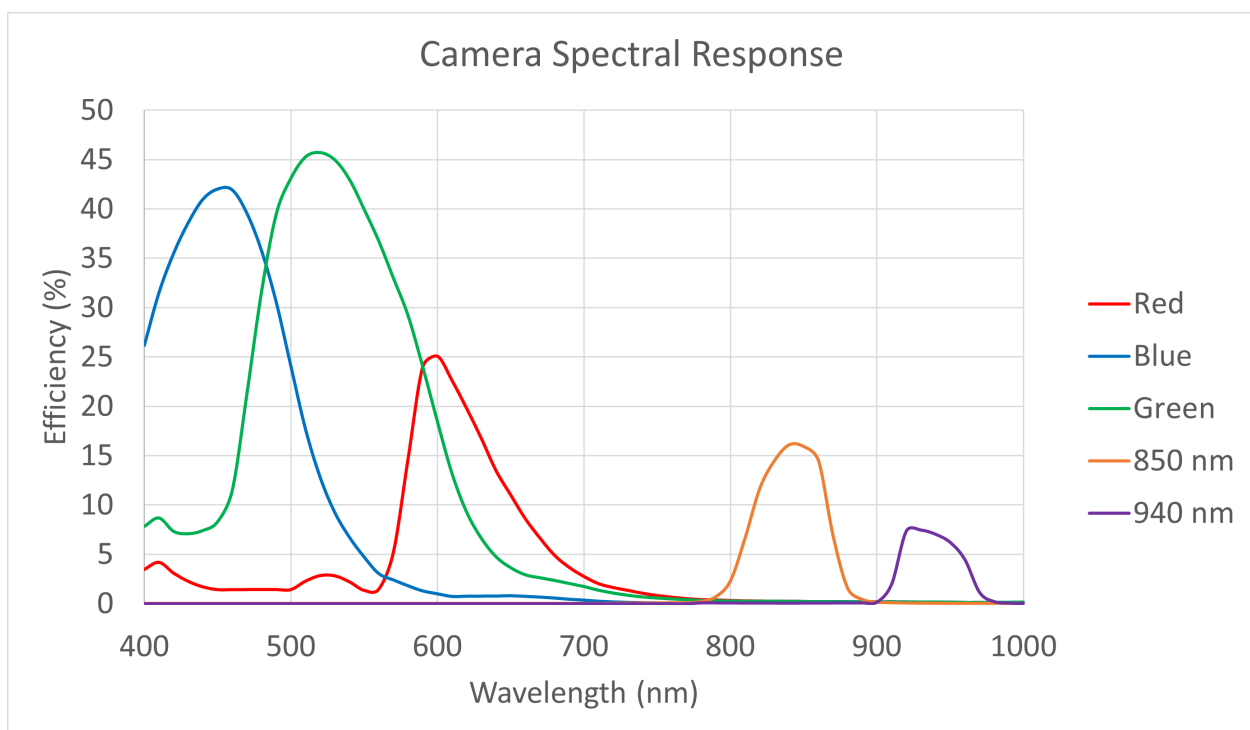


Figure 4.1: Spectral response for VT Tricam iPPG video channels (combining filters and sensor sensitivity information from equipment specifications[5, 6, 7, 8])

Due to limited access, this study used data from only 9 of the intended 18 participants, totaling 63 video sessions. Due to one task having the participant look sideways, resulting in significant occlusion of the face, we selected the six remaining tasks as usable video. The selection resulted in a total of 54 high-quality video recordings for analysis. These were segmented into distinct 15-second clips, yielding 16 segments for training and 4 for testing for each participant. The clips did not overlap, such that no clips used for training contained any content of the testing clips.

Ground-truth PPG signals underwent preprocessing, including bandpass filtering and pulse height normalization as described in section 3.2. We applied a high-pass filter with a cut-off frequency of 0.75 Hz to retain frequencies associated with heart rates above 45 beats per minute (BPM). Filtered signals were manually checked for pulse shape integrity and resampled to match the video’s 60 Hz frame rate.

To augment training data and enhance temporal diversity, we implemented random temporal clipping (a process to generate clips of varying lengths or start times within existing training data), introducing variability across data sequences. The Demucs architecture was adapted to handle three input channels (expanded to five with NIR inclusion), capturing regions of interest (ROIs) and outputting a single-channel PPG signal. The network consisted of three encoder-decoder layer pairs, with each convolutional (Conv1D/ConvTr1D) layer using a kernel size of 10 and a stride of 2, effectively covering the full range of a PPG pulse. For modeling temporal dependencies, three LSTM layers were integrated within the source separation pipeline. The Demucs architecture was adapted for three input channels (using combinations of RGB and NIR video channels), with LSTM layers incorporated for temporal handling. The RGB/NIR channel combinations used in experimentation were standard Red/Green/Blue (RGB), Red/Green/850nm (RG850), Red/Green/940nm (RG940), Red/850nm/940nm (RIR), Green/850nm/940nm (GIR), Blue/850nm/940nm (BIR), and a

5-channel variation Red/Green/Blue/850nm/940nm (RGBIR).

The model was trained using the Adam optimizer[53] with a learning rate of 0.0003. Our loss function combined mean squared error (MSE) and Fourier transform (FFT) loss to improve pulse shape accuracy. Training on the VT Tricam iPPG dataset was conducted over 5000 epochs to ensure convergence.

We explored Demucs variants incorporating hybrid spectrogram-waveform approaches and transformer/self-attention mechanisms for enhanced temporal context handling. However, due to the limited number of high-quality video samples, these more complex models could not be trained effectively, demonstrating insufficient convergence or overfitting. Consequently, the simpler Demucs model with LSTM-based temporal handling was selected as it offered a balance between effective training on the available data and reasonable shape recovery performance.

4.2 PPG Shape Recovery with LSTM

Traditional iPPG methodologies often depend on extensive post-processing techniques, such as filtering and frequency analysis, to extract cardiovascular metrics like heart rate. Our proposed method departs from this norm by emphasizing direct PPG waveform recovery with minimal post-processing. This approach facilitates immediate shape-based analysis and enhances temporal coherence, offering a more streamlined and potentially more accurate means of extracting physiological information.

Building on the foundational work of Li et al. [1], our model integrates LSTM layers and NIR channels to significantly enhance waveform fidelity. The inclusion of LSTM layers plays a critical role in maintaining temporal dependencies, essential for accurately capturing the

dynamic nature of the PPG waveforms. Similar to Li et al.’s approach, our model aims to reduce reliance on extensive signal conditioning, thereby simplifying the processing pipeline. However, by leveraging LSTM layers, our method further improves the robustness and stability of the recovered waveforms, highlighting the added benefits of temporal modeling in enhancing PPG signal recovery.

Figure 4.2 illustrates the comparative performance of our LSTM-based model against Li et al.’s non-LSTM approach using the VT Tricam iPPG dataset. The left and right columns show the recovered pulse shapes for two participants, with shaded areas representing the range of ± 1 standard deviation. Our model consistently exhibits narrower standard deviations than non-LSTM method, indicating reduced noise and improved waveform stability. In the case of participant P315 (left), we can also note that the prediction from our model shows a clearer second notch (Diastolic Peak). This enhanced stability underscores the effectiveness of LSTM layers in capturing temporal patterns, leading to more accurate recovery of key waveform features, such as amplitude and morphology.

The performance metrics in Table 4.1 further substantiate the improvements achieved by our approach. Specifically, our model with RGB alone achieves the higher frequency correlation (0.954) and a notable improvements in both power correlation (0.689) and time correlation (0.878) compared to Li et al.’s architecture. Frequency correlation, which measures the alignment of the predicted signal’s dominant frequency content with the ground truth, is crucial for capturing heart rate variations accurately. Power correlation, which assesses how well the predicted signal’s energy distribution matches the ground truth, reflects the model’s capacity to preserve signal strength and overall waveform dynamics. Time correlation assesses the synchronization of the features.

These findings highlight the superior performance of our LSTM-based approach in recovering both the timing and amplitude characteristics of the PPG signal. By enhancing waveform

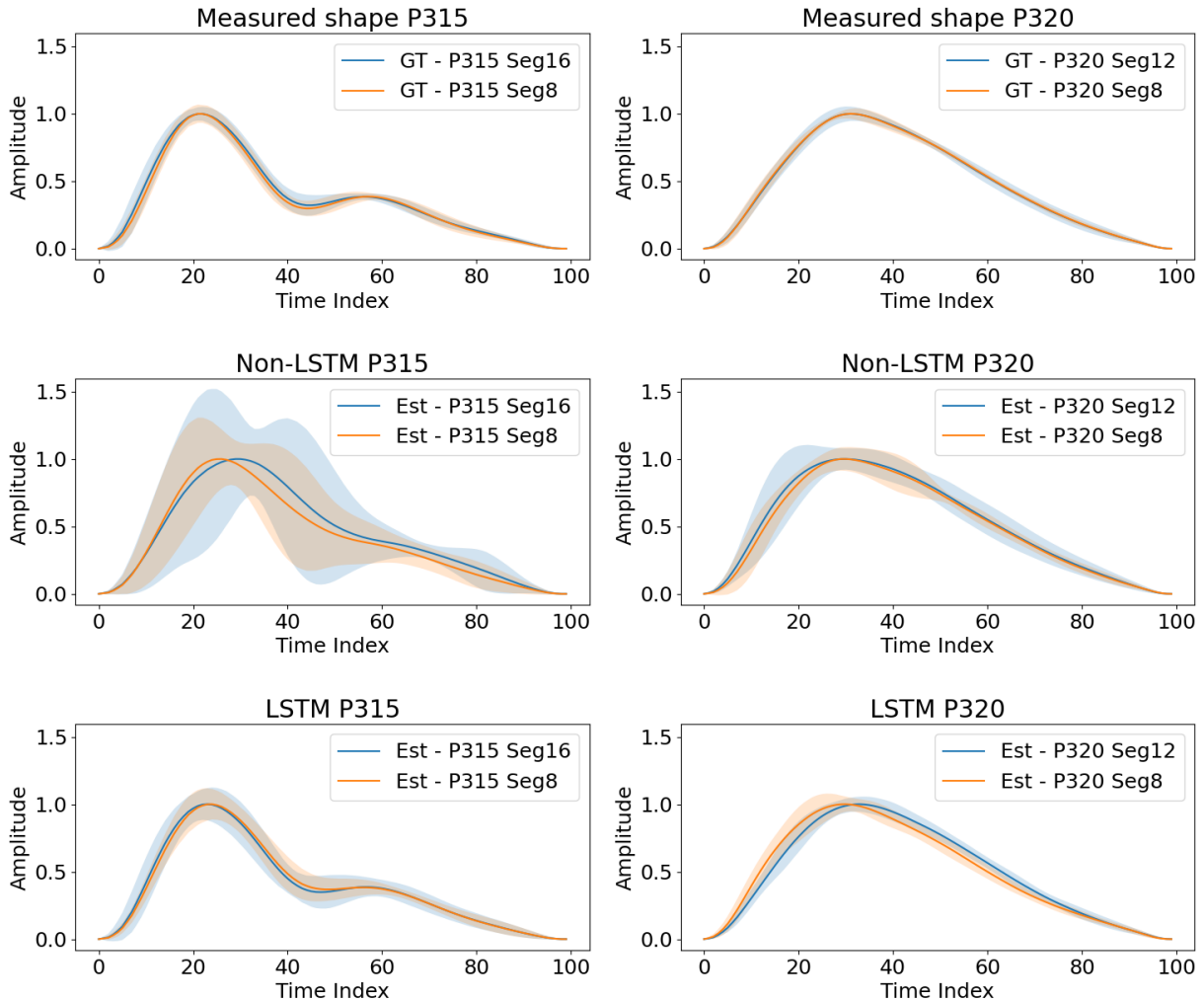


Figure 4.2: Pulse shape recovery for subjects P315 (left) and P320 (right) from the VT Tricam iPPG dataset. Each column represents a different subject, with results shown for various video segment (e.g. Seg8, Seg12, Seg16). The top row corresponds to the ground truth (GT) PPG signal, the middle row shows the estimated (Est) signal from Li et al.'s model trained on the new dataset, and the bottom row presents the estimated signal from our proposed model. All results are derived from RGB channel inputs.

fidelity and stability, our method demonstrates the critical role of LSTM layers in advancing the accuracy and reliability of remote PPG signal extraction.

Table 4.1: Shape recovery performance comparison on the VT Tricam iPPG test set.

Training Dataset	VT Tricam iPPG		DEAP (cross-testing)		
	LSTM		No LSTM	LSTM	No LSTM
Model	RGB	RGBIR	RGB	RGB	RGB
Time Corr. \uparrow	0.878	0.870	0.872	0.010	0.011
Frequency Corr. \uparrow	0.954	0.957	0.949	0.263	0.252
Power Corr. \uparrow	0.689	0.676	0.646	0.282	0.242

Variations in time-domain, frequency-domain, and power spectrum correlations between the VT Tricam iPPG and DEAP datasets highlight the challenges of cross-dataset generalization in iPPG. One contributing factor is the difference in sampling rates (60 Hz in VT Tricam iPPG vs. 50 Hz in DEAP), which can introduce temporal misalignment and affect the consistency of waveform reconstruction. Additionally, the DEAP dataset differs in key environmental conditions such as lighting, camera positioning, environment, and has different participants, all of which can impact signal quality and model predictions. The inherent complexity of iPPG, which relies on subtle color variations influenced by skin tone, motion artifacts, and ambient illumination, makes it particularly sensitive to dataset-specific characteristics. Furthermore, the training dataset remains limited in diversity, restricting the model’s exposure to a broad range of physiological and environmental variations. These factors collectively emphasize the need for training on larger, more representative datasets to ensure reliable performance across different acquisition settings.

4.3 Heart Rate Prediction Performance

Although the primary goal of this research is to recover PPG waveform shapes and evaluate the impact of incorporating IR channels, we also assessed the model’s ability to predict heart rate (HR). This analysis focused on instantaneous HR detection based on the time between consecutive peaks, as opposed to traditional methods that rely on evaluating HR predictions over a time window. By calculating inter-beat intervals, our method provides an immediate estimation of HR on a beat-by-beat basis.

The HR prediction process uses a peak detection algorithm adapted from a modified Pan-Tompkins method, enhanced with cubic spline interpolation for improved peak accuracy. This pipeline also includes smoothing techniques—Kalman filtering and Savitzky-Golay smoothing—to reduce timing noise. This approach is detailed in Section 3.4. We evaluated the model using mean absolute error (MAE), root mean square error (RMSE), and signal-to-noise ratio (SNR) to assess both instantaneous and clip-averaged HR predictions. Instantaneous predictions are directly compared to the detected heart rate of the ground truth, while the clip-wise predictions compare the average prediction over the 15-second clip to the average detected heart rate over the corresponding interval in the ground truth. These metrics are covered in more detail in Section 3.5.1.

The results in Table 4.2 show that LSTM layers provide improvements in HR prediction performance, especially on the VT Tricam iPPG dataset. The LSTM model with RGB achieves an instantaneous HR MAE of 10.24 bpm and a clip HR MAE of 6.45 bpm, compared to 10.49 bpm and 7.14 bpm for the Non LSTM model, indicating a slight improvement in accuracy. Similarly, the RGBIR model achieves an even lower instantaneous HR MAE of 7.06 bpm, showing a more significant improvement over the Non LSTM model.

The model’s performance on the DEAP dataset, as indicated by higher MAE and RMSE

Table 4.2: Heart rate performance comparison on the VT Tricam iPPG test set.

Training dataset	VT Tricam iPPG		DEAP (cross-testing)		
	LSTM		No LSTM	LSTM	No LSTM
Model	RGB	RGBIR	RGB	RGB	RGB
MAE_inst ↓	10.24	7.06	10.49	18.77	17.59
RMSE_inst ↓	20.43	10.93	19.76	30.93	23.17
MAE_clip ↓	6.45	4.18	7.14	15.00	12.70
RMSE_clip ↓	12.21	6.46	10.07	19.21	16.12
SNR (dB) ↑	2.10	5.18	0.01	-12.94	-14.62

values, suggests challenges in adapting to a different data distribution. Beyond the sampling rate mismatch, which can affect temporal alignment, the dataset differences in illumination, participant characteristics, and recording setup likely introduce variations in signal quality that impact reconstruction accuracy. The limited diversity in this incomplete stage of the training dataset further restricts the model’s ability to generalize to unseen conditions. These findings underscore the necessity of a more comprehensive dataset that captures a wider range of real-world variability to improve generalization across different test domains.

These findings suggest that the LSTM layers enable the model to capture the temporal dynamics essential for both shape recovery and HR prediction. The improvements in MAE demonstrate the benefits of LSTM layers for HR estimation. While cross-dataset performance variations highlight the need for further model adjustments, the LSTM model still outperforms the No LSTM approach in both the VT Tricam iPPG and DEAP datasets in terms of MAE.

4.4 Impact of NIR on Shape Recovery and Heart Rate prediction

The incorporation of NIR channels into RGB-based models was evaluated across various metrics for waveform recovery and heart rate estimation. Table 4.3 presents the morphology results, including correlations for timing, frequency, and power between the recovered and ground truth signals, while Table 4.4 summarizes heart rate metrics such as Instantaneous HR MAE, RMSE, and signal-to-noise ratio (SNR). These metrics allow for a detailed comparison of the performance between RGB-only, RGBIR, and different combinations of RGB and NIR inputs. A more in-depth description of the morphology metrics can be found in 3.5.1, and heart rate metrics can be found in Section 3.5.2. A morphology metric of 1.0 would perfectly correlate with the original signal in that domain.

Table 4.3: Morphology results on VT Tricam iPPG for RGB/NIR channel variations.

Metrics	RGB	RGBIR	RG850	RG940	RIR	GIR	BIR
Time Corr. \uparrow	0.878	0.870	0.872	0.880	0.879	0.876	0.874
Frequency Corr. \uparrow	0.954	0.957	0.955	0.955	0.954	0.950	0.946
Power Corr. \uparrow	0.689	0.676	0.679	0.720	0.706	0.685	0.650

Table 4.4: Heart Rate performance comparison on VT Tricam iPPG for RGB/NIR channel variations.

Metrics	RGB	RGBIR	RG850	RG940	RIR	GIR	BIR
Instant HR MAE (bpm) \downarrow	10.24	7.06	7.48	9.06	9.99	9.82	9.90
Instant HR RMSE (bpm) \downarrow	20.43	10.93	12.09	17.80	21.03	17.10	15.27
Clip HR MAE (bpm) \downarrow	6.45	4.18	4.27	5.12	6.23	6.26	6.42
Clip HR RMSE (bpm) \downarrow	12.21	6.46	7.18	8.62	10.52	9.78	9.56
SNR (dB) \uparrow	2.10	5.18	4.61	4.21	3.34	2.59	0.96

Figure 4.3 illustrates the variability in waveform recovery across different channel configu-

rations, highlighting the enhanced stability provided by the inclusion of NIR channels. The RGBIR configuration demonstrates reduced variability in waveform recovery compared to RGB alone. This configuration achieves a slight improvement in frequency correlation (0.957 vs. 0.954) as shown in Table 4.3. In heart rate estimation, the RGBIR setup outperforms the RGB-only model with lower Instantaneous HR RMSE (10.93 bpm vs. 20.43 bpm) and 15-second clip HR MAE (4.18 bpm vs. 6.45 bpm) as shown in Table 4.4. However, the 5-channel variant slightly underperforms in time and power correlation metrics compared to RGB alone (0.870 vs. 0.878 and 0.676 vs. 0.689, respectively). These results suggest that while NIR channels enhance certain aspects of signal recovery, the benefits are context-dependent and do not uniformly improve all metrics, thus calling for an investigation into various combinations of RGB and IR channels to determine their impact.

Among the three-channel configurations, distinct patterns emerge based on the specific channel combinations. The RG850 setup achieves strong shape recovery performance, with a frequency correlation of 0.955 and a power correlation of 0.679 (Table 4.3). Similarly, the RG940 configuration demonstrates a frequency correlation of 0.955 and a higher power correlation of 0.720, indicating that the 940 channel contributes more effectively to amplitude-related features. These results highlight the ability of specific NIR wavelengths to enhance certain aspects of shape recovery when paired with red and green channels.

Configurations using one RGB channel with both NIR channels (850 nm and 940 nm) provide insights into the contributions of individual RGB channels and the supporting role of NIR. For example, the RIR configuration achieves the highest power correlation among single RGB plus dual IR setups (0.706), reflecting better amplitude preservation compared to GIR (0.685) and BIR (0.650) (Table 4.3). Similarly, the GIR setup shows competitive time (0.876) and power correlation (0.685) metrics, outperforming the BIR configuration across most shape metrics (Table 4.3).

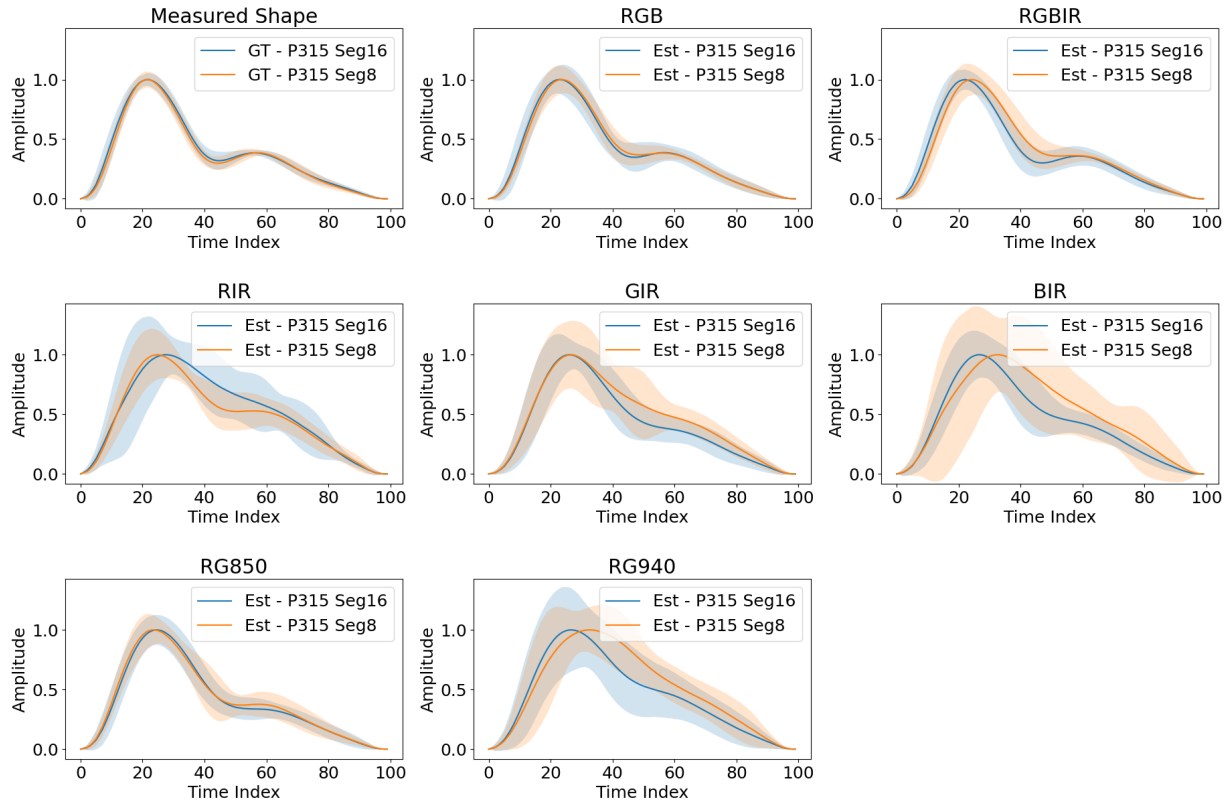


Figure 4.3: (a) Shape recovery results of subject ‘P315’ from VT Tricam iPPG dataset (amplitudes normalized for comparison). Two segments from the testing set with session 1 lighting and from task 2 are displayed. The top left subplot shows the detected pulse shape of the measured PPG signal, with solid lines representing the average waveform and shaded areas showing ± 1 standard deviation. The remaining subplots show the results of models with different combinations of video channels among standard Red, Green, and Blue channels (RGB), and 850nm, 940 Near Infrared(NIR) channels. All models were trained on the same training set using data from the VT Tricam iPPG dataset. (Continued on Next Page)

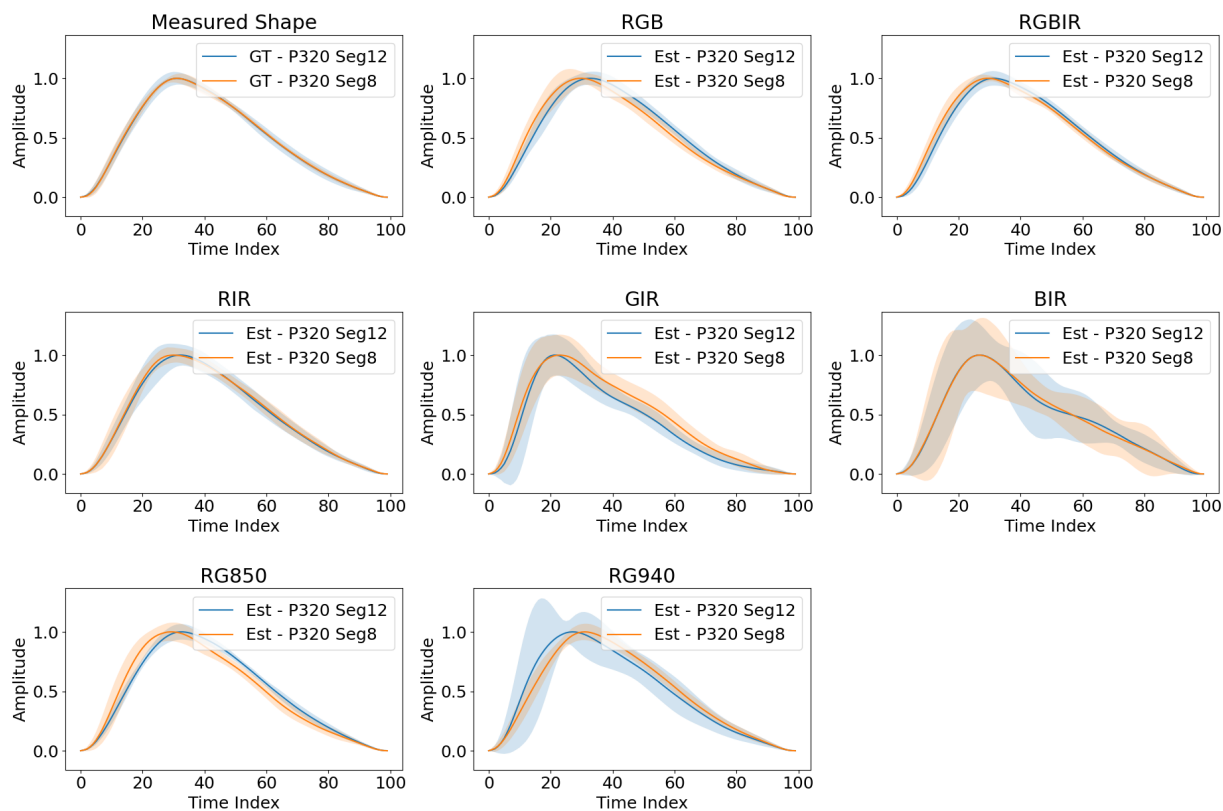


Figure 4.3: (Continued from Previous Page.) (b) Shape recovery results of subject ‘P320’ from VT Tricam iPPG dataset (amplitudes normalized for comparison). Again, two segments from the testing set with session 1 lighting and from task 2 are displayed.

While single RGB/dual NIR combinations generally do not consistently surpass the RGB-only configuration, they offer valuable insights into the channel-specific contributions to waveform recovery. The BIR setup, for instance, lags behind in all shape metrics, including the lowest power correlation (0.650) and higher variability in waveform recovery (Figure 4.3), suggesting the blue channel’s limited effectiveness in shape recovery. In contrast, the red and green combinations with NIR show better performance in specific metrics, such as RIR excelling in time and power correlation and green + NIR providing competitive heart rate predictions (Table 4.4).

For heart rate estimation, the five-channel RGBIR configuration delivers the best overall performance, with the lowest Instantaneous HR RMSE (10.93 bpm) and 15-second clip HR MAE (4.18 bpm) (Table 4.4). Among the three-channel setups, the RG850 configuration stands out with an IHR MAE of 7.48 bpm, significantly outperforming the RGB-only IHR MAE of 10.24 bpm.

In shape recovery, while the five-channel RGBIR configuration achieves the highest frequency correlation (0.957), the RG940 setup excels in time and power metrics while maintaining a competitive frequency correlation (0.955) (Table 4.3).

In summary, the inclusion of NIR channels in various configurations, especially the five-channel RGB + NIR and RG(NIR) channel setups, enhances shape recovery and heart rate prediction performance. The RG940 configuration, in particular, outperforms the RGB-only setup across all metrics, demonstrating the potential of NIR channels to improve waveform fidelity and HR estimation. These findings emphasize the significance of NIR channels in refining signal recovery processes and advancing remote photoplethysmography applications.

4.4.1 Impact of NIR on diversity robustness

When measuring the impact of different skin tones, we base our scale off of the Fitzpatrick skin phototypes[54]. These types represent different skin tones by virtue of their photosensitivity, and delineate levels I through VI. Our current participants consist of levels II, III, and IV, which will be referred to as skin tones 2, 3, and 4.

Table 4.5 presents the VT Tricam iPPG performance across different channel configurations, focusing on Skin Tones 2, 3, and 4. The data highlight the significant role of NIR channels in enhancing algorithm robustness across diverse skin tones.

The top-performing configurations consistently include RG850 and RG940, integrating NIR

Table 4.5: Comparison of VT Tricam iPPG performance for RGB/NIR channel variations (skin Tones 2, 3, and 4).

Metrics	LSTM Models						No-LSTM Models		
	RG850	RG940	RIR	GIR	BIR	RGB	RGBIR	RGB	RGBIR
Skin Tone 2 (lighter)									
Time Corr. \uparrow	0.871	0.879	0.873	0.873	0.871	0.872	0.872	0.869	0.879
Frequency Corr. \uparrow	0.954	0.949	0.952	0.945	0.947	0.953	0.956	0.949	0.945
Power Corr. \uparrow	0.746	0.759	0.759	0.727	0.727	0.744	0.743	0.701	0.725
Instant HR MAE (bpm) \downarrow	5.849	5.664	7.468	8.329	7.410	7.368	6.069	8.527	8.336
SNR (dB) \uparrow	7.266	5.439	5.575	4.362	3.908	4.854	7.304	2.885	3.949
Skin Tone 3									
Time Corr. \uparrow	0.859	0.866	0.867	0.865	0.856	0.871	0.851	0.857	0.867
Frequency Corr. \uparrow	0.953	0.955	0.955	0.948	0.946	0.953	0.956	0.951	0.948
Power Corr. \uparrow	0.903	0.776	0.761	0.736	0.740	0.703	0.701	0.470	0.446
Instant HR MAE (bpm) \downarrow	8.297	11.145	12.303	9.565	11.881	12.595	7.023	11.838	10.285
SNR (dB) \uparrow	1.599	5.876	4.459	2.753	-0.532	2.951	5.769	0.519	2.884
Skin Tone 4 (darker)									
Time Corr. \uparrow	0.895	0.903	0.907	0.900	0.898	0.901	0.893	0.900	0.898
Frequency Corr. \uparrow	0.960	0.965	0.958	0.958	0.957	0.960	0.961	0.958	0.957
Power Corr. \uparrow	0.513	0.508	0.518	0.575	0.541	0.473	0.470	0.470	0.446
Instant HR MAE (bpm) \downarrow	9.425	10.965	11.686	12.671	11.894	12.498	8.598	12.418	14.053
SNR (dB) \uparrow	-1.586	-0.756	-2.719	-1.196	-2.719	-4.676	-0.656	-6.522	-7.419

channels at 850 nm and 940 nm, respectively. These setups exhibit superior performance in waveform fidelity and heart rate estimation across all skin tones. Notably, RG940 achieves the highest correlations and the lowest heart rate errors, demonstrating its effectiveness in capturing PPG signals even with varying skin pigmentation.

On the other hand, the lower-performing configurations, such as RGB and RGBIR (non-LSTM), struggle with darker skin tones. The RGB configuration, lacking NIR integration, shows higher errors and lower correlations across these tones. Similarly, the RGBIR model without LSTM layers fails to improve performance, indicating that merely adding more channels without effective temporal modeling does not enhance the algorithm’s robustness.

These findings underscore the critical importance of NIR channels, especially when paired with LSTM layers, in enhancing the fairness and inclusivity of iPPG algorithms. By reducing

the impact of skin tone variations, these configurations ensure more equitable performance across diverse skin tones.

4.5 Comparison with existing methods

Unsupervised methods for iPPG signal recovery, as summarized in the rPPG-Toolbox by Liu et al. [16], are primarily designed for frequency-based heart rate estimation and are not optimized for waveform shape recovery. Therefore, our evaluation metrics, which emphasize pulse-specific shape preservation through Time, Frequency, and Power Correlation, do not fully capture their strengths.

To address this limitation, we instead evaluated their performance using FFT-based metrics, which are more appropriate for assessing methods that prioritize frequency domain analysis. FFT-based metrics measure the dominant frequency components of the signal and provide a reliable estimate of heart rate, regardless of the preservation of waveform shape. This aligns with the strengths of unsupervised methods, which excel at isolating periodic heart rate information without requiring high fidelity in waveform reconstruction. The results of this comparison are provided in Table 4.6.

Using the results from Table 4.6, we can determine that our model performs well relative to these unsupervised methods on the VT-rPPG and DEAP datasets. On the VT Tricam iPPG dataset, our approach achieves the best metrics across the board with relatively low variability, and achieves promising results on the DEAP dataset, with best or second best metrics outside of the Pearson correlation.

Table 4.6: Performance of state-of-the-art unsupervised methods compared to our method using FFT-based metrics. (Calculated using rPPG-Toolbox [16].)

Dataset	Method	MAE ↓	RMSE ↓	MAPE ↓	Pearson ↑	SNR ↑
VT Tricam iPPG	GREEN [55]	7.47 ± 0.65	12.13 ± 4.75	9.80 ± 0.88	0.54 ± 0.06	-7.97 ± 0.47
	ICA [56]	8.94 ± 0.83	15.12 ± 5.94	12.61 ± 1.31	0.41 ± 0.06	-7.19 ± 0.53
	LGI [57]	13.23 ± 0.74	17.09 ± 5.29	18.49 ± 1.17	0.04 ± 0.07	-12.26 ± 0.31
	PBV [58]	9.77 ± 0.74	14.64 ± 5.58	13.03 ± 1.04	0.44 ± 0.06	-10.01 ± 0.39
	POS [2]	15.25 ± 0.90	20.23 ± 6.15	22.29 ± 1.47	0.04 ± 0.07	-11.50 ± 0.36
	LSTM (ours)	5.19 ± 0.53	9.31 ± 4.48	6.83 ± 0.69	0.66 ± 0.05	0.36 ± 0.69
	No LSTM ([23])	6.33 ± 0.50	9.75 ± 3.65	8.58 ± 0.77	0.59 ± 0.06	-3.72 ± 0.66
DEAP	GREEN [55]	9.60 ± 0.67	13.08 ± 5.1	13.58 ± 1.01	0.31 ± 0.07	-8.62 ± 0.30
	ICA [56]	11.92 ± 0.86	16.51 ± 6.08	17.82 ± 1.49	0.04 ± 0.08	-8.95 ± 0.29
	LGI [57]	8.68 ± 0.70	12.73 ± 5.07	12.13 ± 1.05	0.41 ± 0.07	-7.71 ± 0.30
	PBV [58]	12.26 ± 0.76	15.82 ± 5.38	17.64 ± 1.18	0.18 ± 0.07	-10.06 ± 0.26
	POS [2]	13.49 ± 0.99	18.76 ± 6.73	20.30 ± 1.69	-0.06 ± 0.08	-8.73 ± 0.29
	LSTM (ours)	8.93 ± 0.51	11.2 ± 3.51	13.03 ± 0.85	0.07 ± 0.08	-6.28 ± 0.66
	No LSTM ([23])	8.93 ± 0.47	10.88 ± 3.44	13.19 ± 0.82	0.09 ± 0.08	-6.91 ± 0.63

4.6 Dataset Limitations

While the VT Tricam iPPG dataset provides a valuable baseline, its 60 fps sampling rate limits cross-dataset validation, as models may perform differently under alternative frame rates, impacting temporal signal fidelity. Furthermore, the dataset currently includes only 9 participants, providing approximately half the intended sample size of 18. The strict requirements for consistent lighting and minimal movement limited the usable data to 54 five-minute video sessions. This scarcity of training data complicates the use of deep neural network (DNN) models, potentially affecting the generalizability and robustness of the performance evaluations.

Chapter 5

Conclusions

This study investigated a modified approach to extracting photoplethysmography (PPG) waveforms from facial video, focusing on pulse shape recovery and leveraging a 1D encoder-decoder architecture with Long Short-Term Memory (LSTM) layers. While our model builds on a previous framework, the reintroduction of LSTM adds temporal sensitivity to the signal extraction process, allowing the model to capture sequential dependencies in the PPG signal more effectively. Although this adjustment does not represent a novel architecture, it demonstrates some promising improvements in waveform fidelity.

The primary contribution of this work lies in achieving more accurate shape recovery for PPG signals, which could be of value in scenarios requiring detailed pulse morphology. Results indicate that the model's ability to capture PPG shape details could provide a basis for more precise signal recovery in contexts with minimal post-processing. While our work remains an incremental step, the improvement in waveform representation suggests that models incorporating temporal sensitivity could offer advantages in the field of image based PPG (iPPG) extraction.

Additionally, the inclusion of infrared (IR) channels alongside RGB input enhanced the reliability of the recovered PPG signal, particularly in challenging lighting conditions. The dual-modality input provides additional signal stability, indicating that IR data can serve as a useful complement to RGB channels, especially in low-light settings. This finding supports the potential benefits of multi-modality approaches in iPPG tasks.

However, this study is constrained by limitations of the VT Tricam iPPG dataset, which, while controlled, captures a limited range of variability in lighting and participant characteristics. Future work with larger and more diverse datasets would further assess the model’s robustness and adaptability under varying conditions.

In summary, this work contributes to the ongoing exploration of PPG signal extraction from video by adapting an architecture for improved shape recovery and by illustrating the potential of IR channels in enhancing signal quality. These results highlight possible directions for further investigation in iPPG applications requiring accuracy across diverse environments.

5.0.1 Future Work

To build upon the findings of this study, future work could explore advancements in model architecture, application domains, and performance under realistic conditions:

- **Enhanced Model Architectures for BVP Recovery:** Exploring variations in temporal processing layers, such as gated recurrent units (GRUs), attention mechanisms, or transformer-based architectures, could refine BVP waveform extraction. These alternatives may capture temporal dynamics more efficiently, supporting better recovery of waveform morphology from complex and noisy video data.
- **Instantaneous Heart Rate Monitoring in Dynamic Settings:** Given the initial success of instantaneous heart rate recovery, future work could adapt the model for continuous heart rate monitoring across settings with natural head movements or dynamic lighting. This extension would be especially valuable for applications in fitness or clinical monitoring, where precise, low-exposure HR estimates are critical.

- **Potential for Biometric Authentication:** The clear recovery of individual pulse shapes suggests applications in biometric identification. Future studies could focus on optimizing the model's precision and testing reliability in authenticating individuals based on unique pulse waveforms, particularly when paired with advanced temporal models.
- **Exploiting Infrared (IR) Modalities for Broader Applicability:** While this study suggests IR benefits in signal extraction, further exploration into using IR under challenging conditions, like low-light settings, could improve reliability for nighttime monitoring or low-visibility scenarios.
- **Multi-Task Learning for Comprehensive Vital Sign Monitoring:** Future models could incorporate multi-task learning to predict multiple vital signs, such as respiratory rate or blood oxygen levels, alongside BVP. This extension would increase the practical applications of iPPG technology in health monitoring.

These directions would not only advance model performance but also increase the feasibility of iPPG applications in varied real-world environments, enabling more reliable and versatile health monitoring tools.

Bibliography

- [1] F. Li, *A temporal encoder-decoder approach to extracting blood volume pulse signal morphology from face videos*. PhD thesis, Virginia Tech, 2023.
- [2] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan, “Algorithmic principles of remote PPG,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.
- [3] A. Défossez, N. Usunier, L. Bottou, and F. Bach, “Demucs: Deep extractor for music sources with extra unlabeled data remixed,” *arXiv preprint arXiv:1909.01174*, 2019.
- [4] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, “Mediapipe: A framework for building perception pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.
- [5] Allied Vision Technologies GmbH, “Alvium 1800 U-511.” <https://www.alliedvision.com/en/products/alvium-configurator/alvium-1800-u/511/>. Accessed: 2025-03-04.
- [6] HOYA, “Color Compensating Filters (C).” <https://hoyaoptics.com/colored-glass-filters/color-compensating-filters-c/>. Accessed: 2025-03-04.
- [7] Midwest Optical Systems, Inc., “BN850 Narrow Near-IR Bandpass Filter.” <https://midopt.com/filters/bn850/>. Accessed: 2025-03-04.
- [8] Midwest Optical Systems, Inc., “BN940 Narrow Near-IR Bandpass Filter.” <https://midopt.com/filters/bn940/>. Accessed: 2025-03-04.

- [9] Z. Lv and Y. Li, “Wearable sensors for vital signs measurement: a survey,” *Journal of Sensor and Actuator Networks*, vol. 11, no. 1, p. 19, 2022.
- [10] M. Kumar, A. Veeraraghavan, and A. Sabharwal, “Distanceppg: Robust non-contact vital signs monitoring using a camera,” *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, 2015.
- [11] M. Villarroel, A. Guazzi, J. Jorge, S. Davis, P. Watkinson, G. Green, A. Shenvi, K. McCormick, and L. Tarassenko, “Continuous non-contact vital sign monitoring in neonatal intensive care unit,” *Healthcare Technology Letters*, vol. 1, no. 3, pp. 87–91, 2014.
- [12] D. McDuff, “Camera measurement of physiological vital signs,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–40, 2023.
- [13] R. van Esch, K. Ebrahimkheil, I. Cramer, W. Wang, T. Kaandorp, F. Sammali, A. Dierick, C. Kloeze, C. Verstappen, M. van’t Veer, *et al.*, “Remote PPG for heart rate monitoring: lighting conditions and camera shutter time,” in *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2021*, 2021.
- [14] F. Haugg, M. Elgendi, and C. Menon, “GRGB rPPG: An efficient low-complexity remote photoplethysmography-based algorithm for heart rate estimation,” *Bioengineering*, vol. 10, no. 2, p. 243, 2023.
- [15] W. Wang, S. Leonhardt, L. Tarassenko, C. Shan, and D. McDuff, “Guest editorial: Camera-based monitoring for pervasive healthcare informatics,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1358–1360, 2021.
- [16] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, R. Sengupta,

- S. Patel, Y. Wang, and D. McDuff, “rppg-toolbox: Deep remote PPG toolbox,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [17] D. Yang, J. Zhu, and P. Zhu, “Spo2 and Heart Rate measurement with wearable watch based on PPG,” in *2015 IET International Conference on Biomedical Image and Signal Processing (ICBISP 2015)*, pp. 1–5, 2015.
- [18] C. J. James and C. W. Hesse, “Independent Component Analysis for biomedical signals,” *Physiological Measurement*, vol. 26, p. R15–R39, Dec. 2004.
- [19] J. V. Stone, “Independent Component Analysis: an introduction,” *Trends in Cognitive Sciences*, vol. 6, no. 2, pp. 59–64, 2002.
- [20] J. Liu, H. Luo, P. P. Zheng, S. J. Wu, and K. Lee, “Transdermal optical imaging revealed different spatiotemporal patterns of facial cardiovascular activities,” *Scientific Reports*, vol. 8, no. 1, 2018.
- [21] J. Wei, H. Luo, S. J. Wu, P. P. Zheng, G. Fu, and K. Lee, “Transdermal optical imaging reveal basic stress via heart rate variability analysis: A novel methodology comparable to electrocardiography,” *Frontiers in Psychology*, vol. 9, 2018.
- [22] Y. Deshpande, S. Thapa, A. Sarkar, and A. L. Abbott, “Camera-based recovery of cardiovascular signals from unconstrained face videos using an attention network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5974–5983, June 2023.
- [23] F. Li, S. Thapa, S. Bhat, A. Sarkar, and A. L. Abbott, “A temporal encoder-decoder approach to extracting blood volume pulse signal morphology from face videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5965–5974, June 2023.

- [24] R. D. Labati, V. Piuri, F. Rundo, and F. Scotti, “Photoplethysmographic biometrics: A comprehensive survey,” *Pattern Recognition Letters*, vol. 156, pp. 119–125, 2022.
- [25] J. Sancho, Á. Alesanco, and J. García, “Biometric authentication using the PPG: A long-term feasibility study,” *Sensors*, vol. 18, no. 5, p. 1525, 2018.
- [26] A. Sarkar, *Cardiac signals: remote measurement and applications*. Blacksburg, VA: Virginia Tech, Aug. 2017.
- [27] A. Sarkar, A. L. Abbott, and Z. Doerzaph, “Biometric authentication using photoplethysmography signals,” in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–7, 2016.
- [28] L. Li, C. Chen, L. Pan, L. Y. Zhang, Z. Wang, J. Zhang, and Y. Xiang, “A survey of PPG’s application in authentication,” *Computers amp; Security*, vol. 135, Dec. 2023.
- [29] H. Lu, H. Han, and S. K. Zhou, “Dual-GAN: Joint BVP and noise modeling for remote physiological measurement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12404–12413, 2021.
- [30] H. Kashima, T. Ikemura, and N. Hayashi, “Regional differences in facial skin blood flow responses to the cold pressor and static handgrip tests,” *European Journal of Applied Physiology*, vol. 113, pp. 1035–1041, 2013.
- [31] T. P. Whetzel and S. J. Mathes, “Arterial anatomy of the face: an analysis of vascular territories and perforating cutaneous vessels,” *Plastic and Reconstructive Surgery*, vol. 89, no. 4, pp. 591–603, 1992.
- [32] S. Yucer, F. Tektas, N. Al Moubayed, and T. Breckon, “Racial bias within face recognition: A survey,” *ACM Computing Surveys*, vol. 57, Dec. 2024.

- [33] M. Benčević, M. Habijan, I. Galić, D. Babin, and A. Pižurica, “Understanding skin color bias in deep learning-based skin lesion segmentation,” *Computer Methods and Programs in Biomedicine*, vol. 245, 2024.
- [34] X. Li, Z. Chen, J. Zhang, F. Sarro, Y. Zhang, and X. Liu, “Bias behind the wheel: Fairness testing of autonomous driving systems,” *ACM Transactions on Software Engineering and Methodology (TOSEM)*, Oct. 2024.
- [35] “Approach for improving the performance evaluation of pulse oximeter devices taking into consideration skin pigmentation, race and ethnicity: Discussion paper and request for feedback,” tech. rep., U.S. Food and Drug Administration (FDA), 2024.
- [36] D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Advances in Neural Information Processing Systems*, vol. 13, 2000.
- [37] Y.-X. Wang and Y.-J. Zhang, “Nonnegative matrix factorization: A comprehensive review,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.
- [38] G. Boccignone, D. Conte, V. Cuculo, A. d’Amelio, G. Grossi, and R. Lanzarotti, “An open framework for remote-PPG methods and their assessment,” *IEEE Access*, vol. 8, pp. 216083–216103, 2020.
- [39] A. Défossez, “Hybrid spectrogram and waveform source separation,” 2022. arXiv preprint arXiv:2111.03600.
- [40] S. Rouard, F. Massa, and A. Défossez, “Hybrid Transformers for Music Source Separation,” 2022. arXiv preprint arXiv:2211.08553.
- [41] M. Xu, G. Zeng, Y. Song, Y. Cao, Z. Liu, and X. He, “Ivrr-PPG: An illumination

- variation robust remote-PPG algorithm for monitoring heart rate of drivers,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.
- [42] F. Würtenberger, T. Haist, C. Reichert, A. Faulhaber, T. Boettcher, and A. Herkommer, “Optimum wavelengths in the near infrared for imaging photoplethysmography,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2855–2860, 2019.
- [43] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, “Near-infrared imaging photoplethysmography during driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3589–3600, 2022.
- [44] A. Comas, T. K. Marks, H. Mansour, S. Lohit, Y. Ma, and X. Liu, “Turnip: Time-series u-net with recurrence for nir imaging ppg,” in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 309–313, 2021.
- [45] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.
- [46] W. Chen and D. McDuff, “DeepPhys: Video-based physiological measurement using convolutional attention networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–365, 2018.
- [47] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis ;using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.
- [48] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *Proceedings of the 34th International Conference on Machine*

- Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 933–941, PMLR, 06–11 Aug 2017.
- [49] J. Pan and W. J. Tompkins, “A real-time qrs detection algorithm,” *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.
- [50] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, vol. 82, pp. 35–45, 03 1960.
- [51] A. Savitzky and M. J. Golay, “Smoothing and differentiation of data by simplified least squares procedures.,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [52] A. for the Advancement of Medical Instrumentation *et al.*, “Cardiac monitors, heart rate meters, and alarms,” *American National Standard (ANSI/AAMI EC13: 2002)* Arlington, VA, pp. 1–87, 2002.
- [53] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, (San Diego, CA, USA), 2015.
- [54] S. Eilers, D. Q. Bach, R. Gaber, H. Blatt, Y. Guevara, K. Nitsche, R. V. Kundu, and J. K. Robinson, “Accuracy of self-report in assessing Fitzpatrick skin phototypes I through VI,” *JAMA Dermatology*, vol. 149, pp. 1289–1294, Nov. 2013.
- [55] W. Verkruyse, L. O. Svaasand, and J. S. Nelson, “Remote plethysmographic imaging using ambient light.,” *Optics Express*, vol. 16, no. 26, pp. 21434–21445, 2008.
- [56] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in noncontact, multiparameter physiological measurements using a webcam,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2010.

- [57] C. S. Pilz, S. Zaunseder, J. Krajewski, and V. Blazek, “Local group invariance for heart rate estimation from face videos in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pp. 1254–1262, 2018.
- [58] G. De Haan and A. Van Leest, “Improved motion robustness of remote-ppg by using the blood volume pulse signature,” *Physiological Measurement*, vol. 35, no. 9, p. 1913, 2014.