Remote Photoplethysmography: Spatiotemporal Architecture

Tianyue Sun

CMU-CS-25-158

December 2025

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Dr. Artur W. Dubrawski, Chair
Dr. László A. Jeni

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Keywords: Remote Photoplethysmography, Foundation Models

For my family.

Abstract

Remote photoplethysmography (rPPG) enables contactless measurement of physiological signals such as heart rate and respiratory rate from videos, offering a practical alternative to traditional contact-based sensor measurements. Recent deep learning methods have achieved strong rPPG accuracy, but these approaches often depend on controlled settings and struggle to generalize to real-world environments with motion and varying lighting. These limitations are in part due to the reliance on techniques such as manual parameter tuning and the need for large labelled datasets that are often captured under clean conditions.

This research thesis presents an exploration of the end-to-end rPPG pipeline. The primary contribution is a novel spatiotemporal architecture for rPPG that combines DINOv2, a vision transformer, and Chronos, a time series model. This represents the first multimodal rPPG framework that leverages a combination of spatial and temporal representations from foundation models for physiological measurement. The two foundation models are kept frozen, and a lightweight prediction head is trained.

The proposed model achieves strong performance on the synthetic SCAMPS dataset for heart rate estimation, establishing benchmarks for future rPPG research. On real-world datasets, including PURE and UBFC-rPPG, the model demonstrates effective learning of blood volume pulse (BVP) waveforms and heart rate estimation, despite the increased errors reflecting the difficulty of more challenging conditions. Extensions of the model to respiratory rate illustrate the generalizability of the architecture across different physiological measurement tasks. Overall, the results show that foundation models can improve rPPG robustness and generalization, offering a promising path towards practical rPPG systems with applications in inpatient monitoring, telehealth, and emergency response.

In addition to model development, this thesis analyzes components of the full rPPG pipeline, including signal processing and ground truth extraction. It is shown that common signal processing methods applied to the same BVP signal can lead to discrepancies in the estimation of the scalar heart rate value. Moreover, the method of obtaining the ground truth from data affects the reported performances. These insights motivate the need to further discuss reliable signal processing and evaluation procedures to ensure reliable comparisons and interpretations of rPPG methods.

Acknowledgments

I would like to sincerely thank my thesis committee members, Dr. Artur W. Dubrawski and Dr. László A. Jeni, for their support of the thesis presentation and document. I am incredibly grateful for all those at The Auton Lab whom I have had a chance to work with; The Auton Lab is full of exceptional people who push the boundaries of research. From The Auton Lab, I would like to additionally thank Dr. Dubrawski, Dr. Cecilia Morales, and Dr. Chi-En Teh; I have learned so much under your mentorship. I would also like to thank everyone on the MSCS Program team. Finally, I would like to thank my friends and family for their continued support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Human health parameters, such as heart rate, are traditionally measured with direct-contact sensors that touch a person's body. A popular and inexpensive method for monitoring health-related information is photoplethysmography (PPG), an optical measurement method where the technology uses light and a photodetector at the skin surface to measure the volumetric variations of blood circulation [3].

PPG is often used for heart rate monitoring purposes, and wearable PPG sensors are commonly placed on body locations such as the finger, earlobe, forehead, wrist, torso, and ankle [3]. Examples of devices include pulse oximeters, chest straps, smartwatches, and fitness trackers. The traditional alternative to PPG for cardiac monitoring is electrocardiogram (ECG), which records the electric activity of the heart, showing voltage over time [3]. This has been taken as the gold standard for heart rate measurement [12]. Effective ECG, however, requires the placement of multiple bioelectrodes at certain body locations, limiting movement [3]. Another heart rate measurement method is ambulatory blood pressure (ABP), an oscillometric technique which monitors a person's systolic and diastolic blood pressure and uses a cuff around the arm and a machine. With PPG, ECG, and ABP, heart rate monitoring involves an electronic measurement tool.

Over the past decade, remote photoplethysmography (rPPG) has emerged as a non-invasive alternative method to heart rate measurement [13]. rPPG is a contactless technology that uses ordinary videos or frames of human faces to extract health parameters. From a video recording of a person, contactless systems estimate health metrics, such as heart rate, respiratory rate, blood oxygen saturation, and blood pressure. In the case of heart rate, this involves transforming red/green/blue (RGB) colour signals into blood volume pulse signals by analyzing subtle changes in skin colour caused by blood flow.

The rPPG pipeline typically involves a video or sequence of frames as input, and then some form of image processing such as face detection, cropping, and centralizing. Then, this is fed into an algorithm that uses the RGB variations over time to predict a continuous blood volume pulse (BVP) signal. A BVP signal consists of consecutive pulse waves, consisting of a systolic phase, dicrotic notch, and diastolic phase; the time between heartbeats (systolic time intervals)

is called the interbeat interval (IBI). From the BVP signal, a signal processing method is applied to determine the frequency of pulse signals. This frequency corresponds to the heart rate; the relationship between IBI (milliseconds) and heart rate (beats per minute, BPM) is BPM = $\frac{60000}{\text{IBI}}$. The signal processing method used to obtain the heart rate value is typically based on Fourier Transform or peak detection.

Although the colour variations that rPPG captures appear to be visually imperceptible, they originate from hemodynamics, the dynamics of blood flow. Each cardiac cycle involves changes in blood volume, leading to varying arterial translucency. Hemoglobin, a protein in red blood cells, absorbs green light, and thus pulses are reflected in the RGB intensities. Cameras, particularly those with high resolution, have enough sensitivity to capture such small variations, allowing for pulse extraction.

The rPPG approach to physiological signal monitoring has been promising, particularly in situations where direct contact with a person may be undesirable. In populations with damaged or highly sensitive skin, such as burned patients, the elderly, and newborns, the application of contact-based monitoring may even be impossible [18]. Telemedicine services have seen advantages in using rPPG techniques for monitoring the heart rate variability (HRV), which can be used as indicators for patient monitoring, sleep monitoring, and neonate monitoring [30]. This solution is seen to be beneficial in telehealth due to affordability, portability, and the lack of requirement for dedicated devices. In addition, in physical locations where a risk of infection is present or resources may be limited, patient and clinical safety is improved.

In addition to use in clinical settings, through deployment on autonomous systems such as ground robots and drones, rPPG technology has applications in emergency and disaster response situations where the environment is hazardous and human access is limited. These systems are able to utilize cameras to assess patients from a safe distance [28], allowing a remote evaluation of health parameters. In such high-risk or low-resource situations, scalability and safety are improved.

Although rPPG has many applications, there exist challenges to accurate prediction. First, most rPPG research involves controlled settings where the camera is pointed directly at participants with limited or controlled motion and uniformly consistent backgrounds and lighting. Moreover, factors such as skin tone, facial hair, accessories (mask, glasses, hat), and makeup can lead to different performances of rPPG algorithms. In real-world scenarios, there may be uncontrolled lighting and motion from subjects, the camera, and in the background. In such cases, existing rPPG solutions often perform considerably worse.

Research on rPPG has involved both unsupervised and supervised methods. Deep learning with convolutional neural network and transformer-based architectures has advanced rPPG significantly. End-to-end training on videos of subjects has led to strong performance and accurate heart rate estimations. These systems, however, exhibit deteriorating performance when evaluated outside of what is seen in their training data. Most publicly available rPPG datasets, such as UBFC-rPPG [2] and PURE [25], were collected under controlled laboratory conditions with stable lighting and controlled motion. In real-world scenarios without light source control, the results of the models are mixed; performance depends on the database, sample diversity, and environment variety [24].

To address challenges with generalization, the research of this thesis involves developing an architecture and evaluating the feasibility of rPPG using foundation models. Foundation models

are large AI models trained on vast datasets, suitable for a wide range of tasks. Their large-scale training has the possibility to allow for understanding of subtle signals that correspond to the desired health metrics, despite varied conditions.

This thesis proposes a spatiotemporal foundation model architecture that combines DINOv2 [20] and Chronos [1] in parallel for rPPG. To the best of my knowledge, this architecture is the first multimodal approach that utilizes vision and time series. The architecture leverages DINOv2, a self-supervised Vision Transformer foundation model, for spatial information, and Chronos, a time series forecasting model, to capture the temporal information. A lightweight prediction head is trained on the embeddings of the two models. The method is evaluated on both simulation and real-world datasets.

1.2 DARPA Triage Challenge

Triage is the assessment of patients to determine the urgency of need for medical care. The Defense Advanced Research Projects Agency (DARPA) Triage Challenge (DTC) [8] aims to improve mass casualty triage through the advancement of scalable, timely, and accurate tools. It involves a series of challenge events to drive innovations in identifying injuries, assessing casualties, and transmitting critical data. The challenge consists of three separate competitions Systems, Data, and Virtual.

The Systems competition of the DTC requires teams to operate autonomously to geolocate and assess casualties. The evaluation criteria includes rubric items on heart rate in beats per minute (BPM) and respiratory rate in breaths per minute (BrPM). Teams are awarded points for correct reporting of the heart rate and/or respiratory rate, where correctness is evaluated by whether the reported value is within a threshold (5 BPM for heart rate, 3 BrPM for respiratory rate) of the ground truth value.

The research in this thesis is motivated by the DTC. The development of algorithms that can be deployed on autonomous systems has potential for improvement of medical care in mass casualty incidents.

1.3 Research Contributions

This research thesis provides the following contributions:

- A novel spatiotemporal architecture and evaluations
- Results and benchmarks on synthetic dataset (HR and BR); results on real-world datasets
- Analysis of signal processing for PPG waveforms
- Analysis of ground truth discrepancies

Chapter 2

Related Work

Techniques for accurate rPPG have evolved considerably over the last decade, progressing from traditional unsupervised signal extraction techniques to deep learning methods, and more recently, to foundation models.

2.1 Unsupervised rPPG Methods

Early rPPG methods applied unsupervised signal extraction techniques, which use subtle colour changes in facial videos caused by blood flow. Independent Component Analysis (ICA) [23] was among the first widely adopted methods, involving the decomposition of RGB signals into independent sources and identifying the components that correlated most with the pulse signal. Further methods introduced techniques that were more physiologically motivated, such as CHROM [7] and PBV [6], enabled chrominance formulations. POS [32] further improved robustness by projecting colour signals in a normalized plane. LGI [22] uses spatial relationships in video data to estimate physiological signals.

These methods are effective computationally and do not require labelled data, making them practical for deployment in controlled settings. They are, however, highly sensitive to motion artifacts, lighting variations, camera characteristics, and skin tone differences, requiring careful parameter tuning. As such, their performance degrades in real-world environments.

Summary of notable methods:

- ICA [23]
- CHROM [7]
- LGI [22]
- PBV [6]
- POS [32]

2.2 Supervised rPPG Methods

To address the limitations of careful tuning of traditional signal extraction techniques, supervised deep learning models have been proposed for rPPG applications. DeepPhys [4] introduced an end-to-end convolutional architecture that leverages appearance and motion using a reflection model. PhysNet [34] adopted a lightweight 3D CNN to capture temporal dynamics and reduce overfitting.

More recent work has explored transformer-based architectures to model long-range temporal dependencies. TS-CAN [9] incorporated a CNN architecture to extract spatial and temporal features, with attention mechanisms for learning where the strongest pulse signals are contained. PhysFormer [35] applied temporal difference multi-head self-attention. EfficientPhys-C [11] focuses on simple, fast, and accurate measurements.

While supervised models achieve strong performance under matched training and testing conditions, they depend on large quantities of labelled PPG data. This limits scalability and generalization.

Summary of notable methods:

- TS-CAN [9]
- PhysNet [34]
- PhysFormer [35]
- DeepPhys [4]
- EfficientPhys-C [11]

2.3 Self-Supervised and Unsupervised Learning for rPPG

To reduce the dependence on labelled physiological data, recent work has explored self-supervised and semi-supervised strategies. Contrast-Phys [27] introduces a contrastive learning framework that is tailored to rPPG, and encourages representations that preserve learning without needing explicit PPG labels.

Although promising, self-supervised rPPG methods remain relatively underexplored, and their robustness in unconstrained real-world settings is not well established.

Summary of notable methods:

- Contrast-Phys [27]

2.4 Foundation Model rPPG Methods

Foundation models pretrained on large-scale datasets have begun to be adopted in rPPG research more recently. PhysLLM [33] uses large language models in rPPG estimation, introducing a novel Text Prototype Guidance (TPG) strategy to establish cross-modal alignment by projecting

hemodynamic features with linguistic tokens. It relies heavily on LLM components, limiting practical real-world deployability.

As well, general purpose foundation models for physiological signals have emerged. PaPaGei [21] is pretrained on more than 57,000 hours of 20 million unlabelled PPG signals, demonstrating transferability for downstream physiological tasks, although not designed specifically for vision-based rPPG.

Summary of notable methods:

- PhysLLM [33]
- PaPaGei [21]

2.5 Signal Processing for rPPG

Signal processing is a critical component of the rPPG pipeline, in both unsupervised and supervised systems. Methods for improving frequency resolution and heart rate estimation accuracy include deep adaptive spectral zoom with the Chirp-Z transform [5], enabling a refinement of the spectrum to the narrow-band range of interest for heart rate. Other approaches have explored improved resolutions of the power spectral density due to the occurrence of harmonics [15].

Summary of notable methods:

- Chirp-Z transform [5]
- Improved resolution of PSD [15]

Chapter 3

Methodology

The spatiotemporal architecture involves two pretrained foundation models operating in parallel. The same input frames are fed into the two models, allowing spatial and temporal features to be obtained. Both models are frozen during training, and a lightweight prediction head is trained on top of the embeddings.

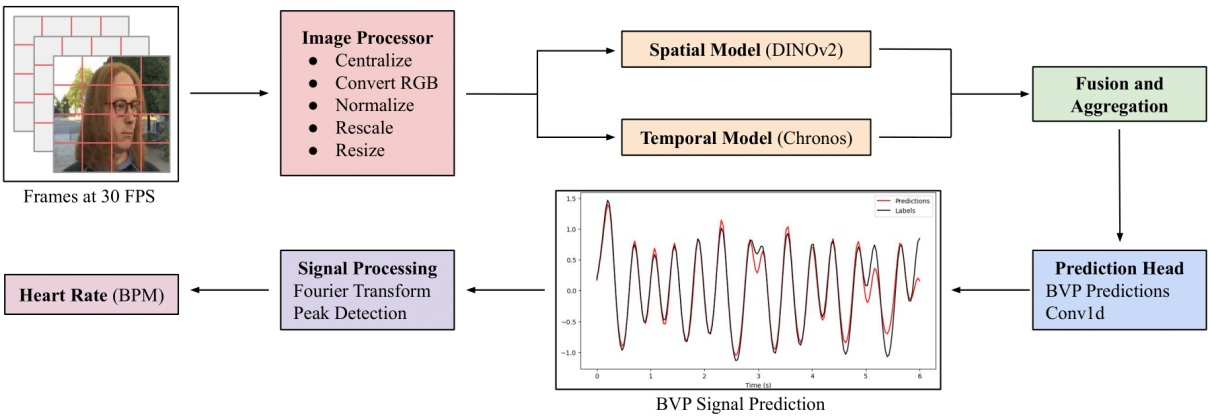

Figure 3.1: Spatiotemporal rPPG model architecture.

This chapter includes descriptions of the models used, prediction head, training pipeline, dimensions, and signal processing.

3.1 Spatial Model

The architecture employs DINOv2 as the spatial model for this architecture. DINOv2 [20], a family of self-supervised Vision Transformer foundation models, produces features that are suitable for image-level visual tasks, such as image classification, instance retrieval, and video understanding, as well as pixel-level visual tasks, such as depth estimation and semantic segmentation.

Pretrained DINOv2 foundation models are available on the DINOv2 GitHub repository. Pretrained backbones via PyTorch Hub are provided. For both with and without registers, there are four DINOv2 model sizes: `small` with 21M parameters, `base` with 86M parameters, `large` with 300M parameters, and `giant` with 1,100M parameters.

3.2 Temporal Model

The temporal model used is Chronos. Chronos is a family of pretrained time series forecasting models based on language model architectures [1]. It is notable that Chronos achieves zero-shot forecasting performance out of the box.

Pretrained Chronos foundation models are available on the Chronos GitHub repository. Five sizes are available from the original Chronos family: `tiny` with 8M parameters, `mini` with 20M parameters, `small` with 46M parameters, `base` with 200M parameters, and `large` with 710M parameters.

3.3 Prediction Head

The embeddings from the two models are concatenated into a unified representation. This feature map is aggregated across spatial patches through mean pooling, resulting in a compact descriptor for the input sequence. This is passed into a prediction head as input, which takes the spatiotemporal representation as input and outputs a predicted BVP waveform. The prediction head consists of three layers: a 1D convolutional layer, Rectified Linear Unit (ReLU), and another 1D convolutional layer. The output signal represents the predicted blood volume changes over time, and is used to extract the heart rate.

3.4 Training

Training Pipeline. The training pipeline is:

- Process frames: in rPPG-Toolbox
  - The input video is sliced into a sequence of frames at 30 frames per second
  - The frames containing human faces are detected and cropped at dimensions of 72 $\times$ 72 pixels by the rPPG-Toolbox preprocessor
- Feed processed frames into foundation models in parallel
  - DINOv2: vision foundation model
    - Pass through the DINOv2 image processor, which transforms images into dimensions 224 $\times$ 224; these dimensions are for passing into the DINOv2 model
      - DINOv2 uses a patch size of 14 $\times$ 14
      - The number of patches is
        (224 / 14) $\times$ (224 / 14) = 16 $\times$ 16 = 256
    - Extract only the normalized patch tokens
    - Pass through DINOv2 to obtain embeddings
  - Chronos: time series forecasting foundation model
    - For each of the RGB channels, split the image into square subpatches such that the patch size matches the patch size used for DINOv2
    - Reshape to combine spatial dimensions (2D representation of image into 1D representation)
    - Average the colour signal values over each patch, for each RGB channel
    - Pass through Chronos to obtain embeddings for each of the 14 $\times$ 14 patches
    - Take the mean of the embeddings over all of the 256 patches
- Fusion and aggregation
  - The feature map outputs of both models are fused into one feature map by concatenation
  - Patch aggregate to reduce by a factor of the number of patches through mean pooling
- Prediction head
  - Conv1D (1D convolution) prediction head
  - The prediction head includes three layers:
    Conv1d $\rightarrow$ ReLU $\rightarrow$ Conv1d
  - Pass through the prediction head to obtain final BVP prediction
- Signal processing: use a method such as the discrete Fourier transform or peak detection to process the signal and obtain the heart rate value

3.5 Dimensions

Engineering the spatiotemporal architecture involves appropriately calculating dimensions throughout the pipeline. The dimensions flow is presented.

Define the following variables:

- `B` = batch size
- `T` = context length
- `C` = number of channels per frame
- `H` = height of frame (pixels)
- `W` = width of frame (pixels)
- `D1` = number of features output by DINOv2
- `D2` = number of features output by Chronos

The chunk size is 6 seconds, and frames are at 30 FPS. Each chunk has 180 frames.

Architecture

Raw frames

{raw frames;
`[B, T, C, H, W]` = `[1, 180, 3, 72, 72]`}

⋆ 1 video, 180 frame/video, 3 channels (RGB)/frame, 72 pixels by 72 pixels/frame

Preprocessed frames

{raw frames;
`[B, T, C, H, W]` = `[1, 180, 3, 72, 72]`}
→ preprocess
→ {preprocessed frames;
`[B, T, C, H, W]` = `[1, 180, 3, 224, 224]`}

⋆ 1 video, 180 frame/video, 3 channels (RGB)/frame, 224 pixels by 224 pixels/frame

DINOv2

{preprocessed frames; `[B, T, C, H, W]` = `[1, 180, 3, 224, 224]`}
→ cut 224 × 224 into 14 × 14 patches and model has features to describe each patch
→ {patches with features; `[B, T, P, D1]` = `[1, 180, 256, 768]`}

⋆ 1 video, 180 frame/video, 16 × 16 = 256 patch/frame, 768 channel (num features)/patch

Chronos

{preprocessed frames;
`[B, T, C, H, W]` = `[1, 180, 3, 224, 224]`}
→ cut 224 × 224 into 14 × 14 patches and take mean of pixels; there are 16 × 16 patches
→ {patches with mean;
`[B, T, C, P, mean]` = `[1, 180, 3, 256, 1]`}
→ reshape
→ {batches and number of frames per time series (180 values per time series);
`[B * C * P, T]` = `[1 * 3 * 256, 180]`}
→ {each patch has 768 features;
`[B * C * P, T, D2]` = `[1 * 3 * 256, 180, 768]`}
→ reshape to be same as DINOv2
→ {`[B, T, P, C * D2]` = `[1, 180, 256, 3 * 768]`}

⋆ 1 video, 180 frame/video, 3 channels (RGB)/frame, 256 patch/frame, 1 mean of pixels in one patch

Combined output result of DINOv2 and Chronos

{feature map;
`[B, T, P, D1 + C * D2]` = `[1, 180, 256, 768 + 3 * 768]`}
→ feature reduction/aggregation (used average across patches)
→ {reduced feature map;
`[B, T, D1 + C * D2]` = `[1, 180, 768 + 3 * 768]`}

Prediction head

{reduced feature map;
`[B, T, D1 + C * D2]` = `[1, 180, 768 + 3 * 768]`}
→ prediction head Conv1d
→ {prediction, for each frame in each video, get a value;
`[B, T, 1]` = `[B, T]` = `[1, 180]`}

3.6 Signal Processing

In the context of rPPG, signal processing involves converting the signal from the time domain to the frequency domain, such as extracting the heart rate value or respiratory rate value from the PPG waveform. Typically, the methods used in literature include discrete Fourier transform and peak detection.

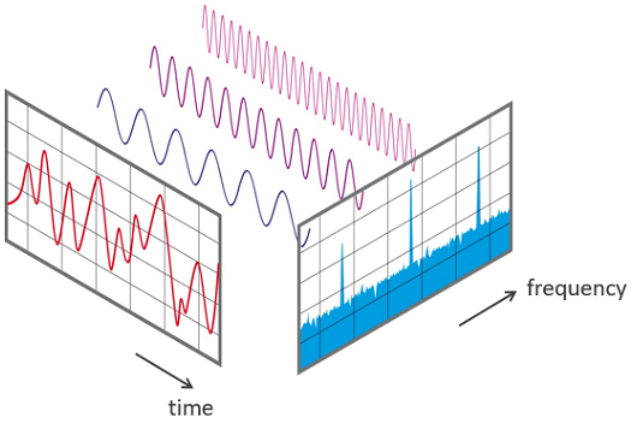

Figure 3.2: View of a signal in the time and frequency domain. [19]

3.6.1 Fourier Transform

The Fourier method is a tool widely used in mathematics to find underlying frequency components in signals, often ones that are more complicated. Fourier analysis transforms a signal from the time domain into the frequency domain, by decomposing the signal into frequency distributions over a continuous range. Specifically, the Fourier Transform breaks down a complex signal into basic functions, in this case sine/cosine waves, at different frequencies. The weights correspond to how strong each frequency component is. The Fourier Transform often handles continuous, infinite signals.

The Discrete Fourier Transform (DFT) processes finite, sampled data, taking samples of the continuous transform. It works with finite sequences, assuming periodicity. The Fast Fourier Transform (FFT) is an efficient algorithm that computes the DFT.

3.6.2 Chirp-Z Transform

The Chirp-Z Transform (CZT) is a signal processing algorithm that generalizes the DFT. CZT samples the Z plane along spiral arcs, in contrast to DFT, which samples the Z plane uniformly along the unit circle. CZT is high-resolution and allows the ability to specify the bandwidth.

3.6.3 Signal Processing Methods for rPPG Pipeline

For the same PPG waveform, the predicted heart rate value can vary considerably depending on the signal processing method utilized. As such, this work also explores the impact of different signal processing methods. The following signal processing methods are explored:

- **rPPG-Toolbox FFT**
  The rPPG-Toolbox FFT implementation takes as input the PPG signal, sampling rate (Hz, default value = 60 Hz), a low pass (Hz, default value = 0.6 Hz = 36 BPM), and a high pass (Hz, default value = 3.3 Hz = 198 BPM). The function takes the next power of 2 for FFT length and estimates the power spectral density using a periodogram with `scipy.signal.periodogram`. The frequencies and powers within the low/high pass band are kept, and the frequency value corresponding to the maximum power is extracted. The frequency is multiplied by 60 to turn from Hz into BPM.
- **rPPG-Toolbox Peak Detection**
  The rPPG-Toolbox Peak Detection implementation takes as input the PPG signal and the frame rate. It calls with default values `scipy.signal.find_peaks`, which finds the local maxima by comparison with neighbour values. Then, the mean difference between peaks is divided by the frame rate and converted to BPM.
- **SciPy Signal Periodogram**
  The SciPy Signal Periodogram takes as input the PPG signal, frame rate, and maximum resolution (Hz, default value = 0.01 Hz). The function calculates the number of points as $2^{\lceil \frac{\texttt{frame\_rate}}{\texttt{max\_resolution}} \rceil - 1}$ and estimates the power spectral density using a periodogram with `scipy.signal.periodogram`, using a `linear` detrend and a `boxcar` window. The frequency value corresponding to the maximum power is extracted. The frequency is multiplied by 60 to turn from Hz into BPM.
- **SciPy Signal Chirp-Z Transform (CZT)**
  The SciPy Signal Chirp-Z Transform takes as input the PPG signal, frame rate, maximum resolution (Hz, default value = 0.01 Hz), beginning frequency (Hz, default value = 0.5 Hz), and ending frequency (Hz, default value = 3.75 Hz). The signal is processed by subtracting the mean. The function calculates the number of points as $2^{\lceil \frac{60 \cdot \texttt{frame\_rate}}{\texttt{max\_resolution}} \rceil - 1}$, and the `boxcar` window is applied to the signal. The `bandwidth` is defined as `frequency_end - frequency_begin`. The `scipy.signal.czt` function computes the frequency response around a spiral in the Z plane. It takes the windowed signal, `m` = the number of output points, `w` = the ratio between points in each step, and `a` = the starting point in the complex plane. The CZT function is applied with `w` = $\exp(-i \cdot 2\pi \cdot \frac{\texttt{bandwidth}}{n_{\text{points}} \cdot \texttt{frame\_rate}})$, and `a` = $\exp(i \cdot 2\pi \cdot \frac{\texttt{bandwidth}}{\texttt{frame\_rate}})$. This represents uniform steps around the unit circle, starting at 0. From the frequency response, the frequency value corresponding to the maximum power is extracted. The frequency is multiplied by 60 to turn from Hz into BPM.

Chapter 4

Experiments

The experimental procedure and results are detailed in this chapter. The proposed spatiotemporal architecture is evaluated with multiple (Train Set, Test Set) pairs to assess its accuracy, robustness, and generalization. This chapter details the experimental procedure, including the datasets, evaluation metrics, model and training configuration, quantitative results, and ablation studies.

The experiments are designed to answer these key research questions:

- How well does the proposed spatiotemporal architecture estimate the blood volume pulse and heart rate, compared to existing rPPG methods?
- How robust is the proposed spatiotemporal architecture under challenging conditions, and what might be the failure cases?
- How well does the proposed spatiotemporal architecture generalize across different rPPG datasets?

4.1 Datasets

The spatiotemporal model is trained and evaluated on datasets containing videos/frames with time-aligned PPG waveforms.

SCAMPS [14]

SCAMPS (Synthetics for Camera Measurement of Physiological Signals) is a dataset of synthetics that contains 2800 videos with aligned cardiac and respiratory signals and facial action intensities. Each file contains ground truth heart rate, breathing rate, interbeat intervals, and heart rate variability.

The synthetic videos are created using a graphics pipeline, involving realistically modelling facial blood flow; this is simulated by adjusting properties of physically-based shading material.

Prior work has primarily relied on SCAMPS simulation data for training. The authors of SCAMPS trained on SCAMPS and validated and tested on real-world datasets such as PURE, and the results showed that the synthetic data are sufficient to train a reasonable supervised model, implying generalization is possible.

In this thesis, model performance is also explored when training and testing on this dataset, using the officially provided Train, Validation, and Test splits with 2000, 400, and 400 files respectively. This provides a benchmark for confirming model effectiveness.

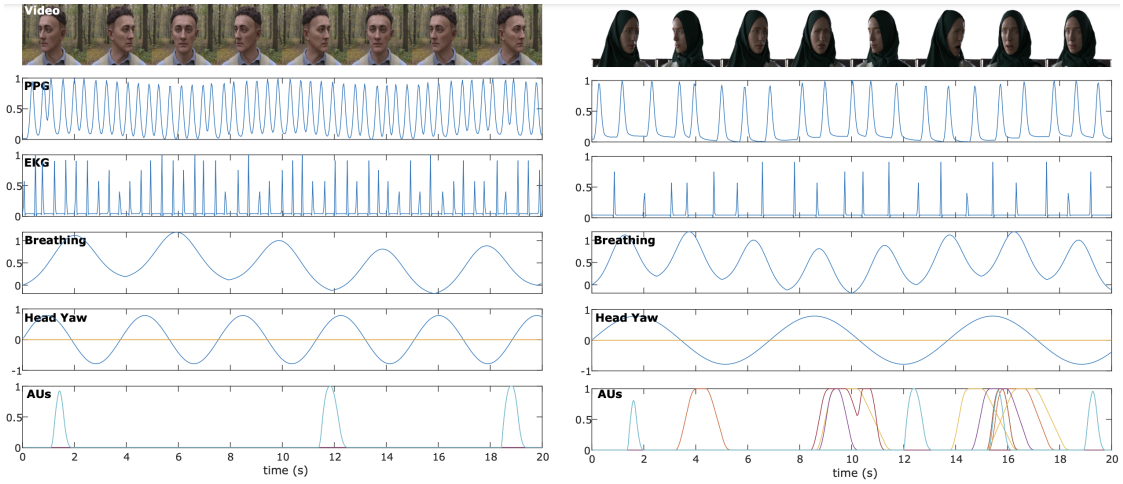

Figure 4.1: SCAMPS Dataset Structure [14]

PURE [25]

PURE - Pulse Rate Detection Dataset consists of 10 persons (8 male, 2 female) performing different, controlled head motions; each person was recorded in 6 different setups for a total of 60 sequences of around 1 minute each. The videos were captured with an eco274CVGE camera by SVS-Vistek GmbH at a frame rate of 30 Hz with a cropped resolution of 640×480 pixels and a 4.8mm lens. Reference data was captured in parallel using a finger clip pulse oximeter (pulox CMS50E), which delivers pulse rate wave and SpO_2 readings with a sampling rate of 60 Hz.

Test subjects were in front of the camera with an average distance of 1.1 meters, and lighting condition was daylight. The six setups were steady, talking, slow translation, fast translation, small rotation, and medium rotation. Translations involved head movements parallel to the camera plane, and rotations involved looking at targets in a predefined sequence for head rotation angles of 20°and 35°.

UBFC-rPPG [2]

UBFC-rPPG (Univ. Bourgogne Franche-Comté Remote PhotoPlethysmoGraphy) is a database comprising two datasets for rPPG analysis. The videos were taken with a Logitech C920 HD Pro webcam at 30 fps with a 640×480 resolution in uncompressed 8-bit RGB format. The ground truth PPG data, comprising the PPG waveform and PPG heart rates, was collected using a CMS50E transmissive pulse oximeter. Subjects were seated about 1m away from the camera with face visible, and recordings were conducted indoors with a varying amount of sunlight and indoor illumination.

Dataset 1, intended to be simple, includes 8 videos where participants were asked to sit still. Dataset 2, intended to be realistic, includes 42 available videos where participants play a time-sensitive mathematical game. For all experiments, only videos in Dataset 2 were used.

UBFC-Phys [16]

The UBFC-Phys dataset is a multimodal dataset consisting of 56 participants who underwent an experiment with a protocol inspired by the Trier Social Stress Test (TSST), conducted in three stages: a rest, a speech, and arithmetic tasks.

During the experiment, participants were filmed and wore a wristband that measured their blood volume pulse (BVP) and electrodermal activity (EDA) signals. The experience took place in a laboratory room, and filmed with an EO-23121C RGB digital camera by Edmund Optics, with a Motion JPEG compression and a 35 frame per second rate. The frame resolution was of 1024×1024 pixels. An artificial light source was used for uniform lighting conditions. Participants were seated around 1m away from the camera and the light source. The contact measurements were realized using the Empatica E4 wristband, which records BVP, skin temperature, and EDA responses. The E4 bracelet has an accelerometer and computes the interbeat intervals from the BVP signal.

All participants were aged between 19 and 38 with a mean age of 21.8, with 46 females and 10 males. For each subject, three videos are available, corresponding to the three tasks.

MMPD [29]

The MMPD: MMPD: Multi-Domain Mobile Video Physiology Dataset collected by Tsinghua University comprises 11 hours (1152K) frames of recordings from mobile phones of 33 subjects, for a total of 660 one-minute videos. The dataset was designed to capture videos with greater representation across skin tone, body motion, and lighting conditions. The skin types ranged from 3 to 6 on the Fitzpatrick skin type classification, with four different lighting conditions (LED-high, LED-low, incandescent, natural), and four different activities (stationary, head rotation, talking, walking). The videos were filmed on a Samsung Galaxy S22 Ultra at 30 frames per second with a resolution of 1280×720 pixels, compressed to 320×240 pixels. The PPG signals were recorded using an HKG-07C+ oximeter.

NATURE

The NATURE rPPG dataset is a dataset collected by the Auton Lab at Carnegie Mellon University that consists of 17 adult participants with 11 different outdoor setups. Participant ages ranged from 19 to 31 years, with 12 males, 4 females, 1 X.

The dataset includes a diverse set of environmental conditions in 5 categories. The baseline condition is outdoor background, no accessories, no sunlight, no head movement, and 1m distance from the camera. The variables changed one at a time compared to the baseline: background (outdoor background, white background), accessories (glasses, face mask), lighting (partial shadow, direct sunlight), head movement rotation at 45°left and right (slow 12 s/cycle, medium 8 s/cycle, fast 4 s/cycle), and distance from the camera (2m, 3m). Participants also reported their skin tones on the Fitzpatrick skin tone chart.

The videos were recorded with an Intel RealSense D435 camera, with 1920×1080 RGB streams at 30 Hz for around 60 s per clip. Ground truths were obtained using a ZOLL X Series monitor that reported PPG and ECG as well as a Garmin HRM-Pro Plus chest strap.

4.2 Evaluation Metrics

The goal of evaluation metrics is to quantitatively assess how accurately a model predicts heart rate. Since the full rPPG pipeline involves predicting a scalar heart rate or respiratory rate value from a sequence of frames, the ideal indicator of prediction quality is how close the predicted rate is to the ground truth rate. To this extent, metrics that capture different aspects of prediction quality are employed, including absolute error, outlier robustness, and tolerance for accuracy. These metrics provide a comprehensive view of model performance.

Initial experiments demonstrated that there is a noticeable difference in the results depending on the signal processing method used for extracting the rate from the waveform. This motivated the exploration of multiple signal processing and evaluation methods. Two sets of metrics are reported: rPPG-Toolbox waveform window signal processing evaluation metrics and ground truth comparison metrics. These two types of ground truth extraction are further assessed in the Discussion chapter.

For all metrics, R_p represents the predicted signal rate, R_g represents the ground truth signal rate, and N represents the number of instances. Each instance corresponds to one non-overlapping 6-second window. For the sensor end-of-window ground truth comparison metrics, $\mathbb{1}(\cdot)$ is the indicator function that outputs 1 if the condition is true and 0 if the condition is false.

4.2.1 [rPPG-Toolbox] Waveform Window Signal Processing Metrics

In literature, many comparisons are done through the metrics proposed by rPPG-Toolbox, which involves signal processing on waveform windows of the waveform. The rPPG-Toolbox computes and displays the following metrics.

- **Mean Absolute Error (MAE)**
  MAE = $\frac{1}{N}\sum_{n=1}^{N}|R_g - R_p|$
- **Root Mean Squared Error (RMSE)**
  RMSE = $\sqrt{\frac{1}{N}\sum_{n=1}^{N}(R_g - R_p)^2}$
- **Mean Absolute Percentage Error (MAPE)**
  MAPE = $\frac{1}{N}\sum_{n=1}^{N}|\frac{R_g - R_p}{R_g}|$
- **Pearson Correlation Coefficient (Pearson)**
- **Signal-to-Noise Ratio (SNR)**
- **Maximum Amplitude of Cross Correlation (MACC)**

In reporting results, the primary focus is on MAE and MAPE, and at times RMSE.

rPPG-Toolbox evaluates the heart rate prediction accuracy (MAE, RMSE, MAPE) by retrieving the contact ground truth PPG signal and the predicted PPG signal, and then applying a signal processing method on the 6-second window for both to obtain a scalar value corresponding to the heart rate ground truth or prediction. The ground truth values obtained through this method are dependent on the signal processing method used. Thus, an alternative ground truth metric is suggested.

4.2.2 Sensor End-of-Window Ground Truth Comparison Metrics

Since the datasets provide heart rate values at timesteps provided by the sensor, another evaluation metric is to compare directly to the sensor heart rate ground truth at a timestep that represents the window. As heart rate does not fluctuate dramatically in a 6-second window, the ground truth is taken at the end of the 6-second window. That is, for each non-overlapping 6-second window of video, the heart rate value at the timestamp corresponding to the end of the window is recorded. For instance, for the window from 12 seconds to 18 seconds, the sensor heart rate value at 18 seconds is taken.

With the sensor end-of-window ground truth comparison metrics, the following metrics are computed.

- **Mean Absolute Error (MAE)**
  MAE = $\frac{1}{N}\sum_{n=1}^{N}|R_g - R_p|$
- **Median Absolute Error (MedAE)**
  MedAE = median $(|R_g - R_p|)$
- **Root Mean Squared Error (RMSE)**
  RMSE = $\sqrt{\frac{1}{N}\sum_{n=1}^{N}(R_g - R_p)^2}$
- **Tolerance Threshold within 5 BPM (Heart Rate)**
  $\text{Acc}_{\text{HR 5 BPM}} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}(|R_g - R_p| \leq 5)$
  The percentage of 6-second windows in the test set for which the heart rate prediction is within 5 BPM of the ground truth value.
- **Tolerance Threshold within 3 BrPM (Respiratory Rate)**
  $\text{Acc}_{\text{RR 3 BrPM}} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}(|R_g - R_p| \leq 3)$
  The percentage of 6-second windows in the test set for which the respiratory rate prediction is within 3 BrPM of the ground truth value.

The tolerance threshold metrics align with evaluation criteria from the DARPA Triage Challenge [8].

For the UBFC-rPPG dataset, some heart rate values were missing. The missing values at certain timesteps were calculated by using the nearest available ground truths and linearly interpolating.

4.3 Model Configuration

For experimentation, different-sized models and finetuning were explored. Strong results were found with DINOv2 size small and Chronos size base, with a frozen architecture. Although finetuning demonstrated potential to improve the predicted BVP signal, the diastolic peak would cause signal processing methods to overestimate the heart rate. Results are reported for different configurations, but the primary focus will be on the DINOv2 small, Chronos base, frozen configuration.

In all experiments, non-overlapping 6-second windows are utilized. This means signal processing is applied every 6 seconds to obtain a heart rate estimation for that window.

4.4 Training Configuration

The number of epochs used for testing was based on the learning loss and validation results. For the supervised neural methods, rPPG-Toolbox found that stable results with convergence were achieved at a maximum learning rate of 0.009 and 30 epochs. The learning rates were selected naively through experimentation, starting at a larger value and finding values at which the losses do not diverge. When training, the learning rate scheduler provided by rPPG-Toolbox was used.

		Training Parameter	
Spatiotemporal Model	**Train Set**	Learning Rate$^{\downarrow}$	Epochs$^{\downarrow}$
DINOv2-Small + Chronos-Tiny	SCAMPS	9e-4	10
	UBFC-rPPG	3e-3	4
	PURE	3e-3	3
DINOv2-Small + Chronos-Base	SCAMPS	9e-4	10
	UBFC-rPPG	1e-4	30
	PURE	1e-3	30

Table 4.1: Training Configuration for Spatiotemporal DINOv2 + Chronos rPPG Model.

For all datasets, the raw frame data type is utilized. The face detection backend is Haar Cascade, with a large face box. The training batch size is 1 for all configurations. The rPPG-Toolbox configurations relevant to training were generally set as the following for all datasets:

- `DATA_TYPE: ['Raw']`
- `DATA_AUG: ['None']`
- `LABEL_TYPE: 'Raw'`
- `DO_CHUNK: True`
- `CHUNK_LENGTH: 180`
- `DO_CROP_FACE: True`
- `BACKEND: 'HC'`
- `USE_LARGE_FACE_BOX: True`
- `LARGE_BOX_COEF: 1.5`
- `DO_DYNAMIC_DETECTION: False`

4.5 Experimental Results

4.5.1 Heart Rate

Synthetic Dataset

To evaluate the feasibility of the proposed architecture, it is first tested with the SCAMPS synthetic dataset. Training and testing is done in accordance with the provided Train, Validation, and Test splits provided by SCAMPS.

Performance for heart rate of the Spatiotemporal DINOv2 + Chronos model.
Train Set: SCAMPS (Train Split); Test Set: SCAMPS (Test Split)

Evaluation: rPPG-Toolbox Waveform Window Signal Processing Metrics

MAE	RMSE	MAPE
3.65	12.71	6.56

Table 4.2: DINOv2-Small + Chronos-Tiny; rPPG-Toolbox Metrics; SCAMPS/SCAMPS.

MAE	RMSE	MAPE
2.46	10.43	4.55

Table 4.3: DINOv2-Small + Chronos-Base; rPPG-Toolbox Metrics; SCAMPS/SCAMPS.

Evaluation: Sensor End-of-Window Ground Truth Comparison

Signal Processing	**MAE**	**MedAE**	**RMSE**	**5 BPM**
rPPG-Toolbox FFT	6.57	2.04	15.18	87.75%
rPPG-Toolbox Peak	28.95	11.86	44.12	43.00%
SciPy Periodogram	3.63	0.69	13.54	94.33%
SciPy CZT	3.29	0.67	12.80	94.83%

Table 4.4: DINOv2-Small + Chronos-Tiny; Sensor End-of-Window Comparison; SCAMPS/SCAMPS.

Signal Processing	**MAE**	**MedAE**	**RMSE**	**5 BPM**
rPPG-Toolbox FFT	5.47	2.00	13.30	90.50%
rPPG-Toolbox Peak	20.77	2.32	33.56	56.00%
SciPy Periodogram	1.64	0.63	7.15	98.08%
SciPy CZT	1.39	0.62	5.87	98.50%

Table 4.5: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; SCAMPS/SCAMPS.

Real-World Datasets

Benchmark Results. The rPPG-Toolbox includes benchmarks for performances across different algorithms and datasets, evaluated with rPPG-Toolbox Metrics. To compare the performance of the Spatiotemporal DINOv2 + Chronos model with other models, the table is replicated and two columns are added with the results of DINOv2-Small + Chronos-Tiny and DINOv2-Small + Chronos-Base. Train Sets: UBFC-rPPG, PURE, SCAMPS. Test Sets: PURE, UBFC-rPPG.

Table 4.6: **Benchmark Results.** (Table from rPPG-Toolbox [10]). Performance on the UBFC-rPPG [2], PURE [25], UBFC-Phys [16], and MMPD [29] datasets generated using the rPPG-Toolbox. Supervised methods show cross-dataset training results using the UBFC-rPPG, PURE, and SCAMPS datasets. Added results with spatiotemporal model.
SpaTe = SpatioTemporal DINOv2 + Chronos Model
SmTi = DINOv2-Small + Chronos-Tiny
SmBa = DINOv2-Small + Chronos-Base

			Test Set							
			PURE [25]		UBFC-rPPG [2]		UBFC-Phys [16]		MMPD [29]	
	Method	**Train Set**	MAE$^{\downarrow}$	MAPE$^{\downarrow}$	MAE$^{\downarrow}$	MAPE$^{\downarrow}$	MAE$^{\downarrow}$	MAPE$^{\downarrow}$	MAE$^{\downarrow}$	MAPE$^{\downarrow}$
UNSUPERVISED	GREEN [31]	N/A	10.09	10.28	19.81	18.78	13.55	16.01	21.68	24.39
	ICA [23]	N/A	4.77	4.47	14.70	14.34	10.03	11.85	18.60	20.88
	CHROM [7]	N/A	5.77	11.52	3.98	3.78	4.49	6.00	13.66	15.99
	LGI [22]	N/A	4.61	4.96	15.80	14.70	6.27	7.83	17.08	18.98
	PBV [6]	N/A	3.91	4.82	15.90	15.17	12.34	14.63	17.95	20.18
	POS [32]	N/A	3.67	7.25	4.00	3.86	4.51	6.12	12.36	14.43
SUPERVISED	TS-CAN [9]	UBFC-RPPG	3.69	3.38	N/A	N/A	5.13	6.53	14.00	15.47
		PURE	N/A	N/A	1.29	1.50	5.72	7.34	13.93	15.14
		SCAMPS	4.66	5.83	3.62	3.53	5.55	6.91	19.05	21.77
	PHYSNET [34]	UBFC-RPPG	8.06	13.67	N/A	N/A	5.79	7.69	9.47	11.11
		PURE	N/A	N/A	0.98	1.12	4.78	6.15	13.93	15.61
		SCAMPS	13.30	20.01	5.40	5.43	8.53	11.22	20.78	24.43
	PHYSFORMER [35]	UBFC-RPPG	12.92	23.92	N/A	N/A	6.63	8.91	12.1	15.41
		PURE	N/A	N/A	1.44	1.66	6.04	7.67	14.57	16.73
		SCAMPS	26.58	42.79	4.56	5.18	11.91	15.57	22.69	27.06
	DEEPPHYS [4]	UBFC-RPPG	5.54	5.32	N/A	N/A	6.62	8.21	17.49	19.26
		PURE	N/A	N/A	1.21	1.42	8.42	10.18	16.92	18.54
		SCAMPS	3.95	4.25	3.10	3.08	4.75	5.89	15.22	16.56
	EFF.PHYS-C [11]	UBFC-RPPG	5.47	5.39	N/A	N/A	4.93	6.25	13.78	15.15
		PURE	N/A	N/A	2.07	2.10	5.31	6.61	14.03	15.31
		SCAMPS	10.24	11.70	12.64	11.26	6.97	8.47	20.41	23.52
	SPATE-SMTI	UBFC-RPPG	9.88	17.11	N/A	N/A	N/A	N/A	N/A	N/A
		PURE	N/A	N/A	10.91	11.08	N/A	N/A	N/A	N/A
		SCAMPS	17.07	27.46	16.79	24.53	N/A	N/A	N/A	N/A
	SPATE-SMBA	UBFC-RPPG	12.45	24.10	N/A	N/A	15.17	21.28	14.80	18.62
		PURE	N/A	N/A	5.51	5.22	10.79	14.55	17.64	21.16
		SCAMPS	18.21	32.21	4.50	4.74	18.17	25.10	22.35	27.71

MAE = Mean Absolute Error in HR estimation (Beats/Min), MAPE = Mean Percentage Error (%).

Performance for heart rate of the Spatiotemporal DINOv2 + Chronos model.
Train Set: PURE; Test Set: UBFC-rPPG

Evaluation: Sensor End-of-Window Ground Truth Comparison

Signal Processing	MAE	MedAE	RMSE	5 BPM
rPPG-Toolbox FFT	12.25	4.53	20.38	52.50%
rPPG-Toolbox Peak	32.99	28.11	41.03	12.50%
SciPy Periodogram	22.34	5.75	36.23	48.41%
SciPy CZT	18.73	4.70	31.37	50.68%

Table 4.7: DINOv2-Small + Chronos-Tiny; Sensor End-of-Window Comparison; PURE/UBFC-rPPG.

Signal Processing	MAE	MedAE	RMSE	5 BPM
rPPG-Toolbox FFT	7.59	3.41	13.92	64.77%
rPPG-Toolbox Peak	37.25	34.05	44.52	7.50%
SciPy Periodogram	11.04	2.23	25.36	66.59%
SciPy CZT	9.32	2.19	20.37	67.73%

Table 4.8: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; PURE/UBFC-rPPG.

Performance for heart rate of the Spatiotemporal DINOv2 + Chronos model.
Train Set: SCAMPS (Train Split); Test Set: UBFC-rPPG

Evaluation: Sensor End-of-Window Ground Truth Comparison

Signal Processing	MAE	MedAE	RMSE	5 BPM
rPPG-Toolbox FFT	7.21	3.41	13.04	67.05%
rPPG-Toolbox Peak	39.96	37.17	47.70	9.55%
SciPy Periodogram	12.84	2.32	27.70	63.64%
SciPy CZT	10.59	2.28	22.42	65.45%

Table 4.9: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; SCAMPS/UBFC-rPPG.

Performance for heart rate of the Spatiotemporal DINOv2 + Chronos model.
Train Set: UBFC-rPPG; Test Set: PURE

Evaluation: Sensor End-of-Window Ground Truth Comparison

Signal Processing	MAE	MedAE	RMSE	5 BPM
rPPG-Toolbox FFT	10.89	4.47	18.86	53.32%
rPPG-Toolbox Peak	65.60	71.12	77.11	8.81%
SciPy Periodogram	17.18	4.46	28.72	52.09%
SciPy CZT	11.50	3.48	21.46	57.34%

Table 4.10: DINOv2-Small + Chronos-Tiny; Sensor End-of-Window Comparison; UBFC-rPPG/PURE.

Signal Processing	MAE	MedAE	RMSE	5 BPM
rPPG-Toolbox FFT	13.67	3.59	23.87	61.82%
rPPG-Toolbox Peak	54.43	54.57	64.45	8.50%
SciPy Periodogram	16.34	2.61	31.18	63.06%
SciPy CZT	13.37	2.26	28.00	67.39%

Table 4.11: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; UBFC-rPPG/PURE.

Performance for heart rate of the Spatiotemporal DINOv2 + Chronos model.
Train Set: SCAMPS (Train Split); Test Set: PURE

Evaluation: Sensor End-of-Window Ground Truth Comparison

Signal Processing	MAE	MedAE	RMSE	5 BPM
rPPG-Toolbox FFT	19.08	5.59	30.07	48.38%
rPPG-Toolbox Peak	67.59	70.45	73.18	3.09%
SciPy Periodogram	20.78	3.73	36.64	55.02%
SciPy CZT	19.51	3.44	35.04	55.80%

Table 4.12: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; SCAMPS/PURE.

Performance for heart rate of the Spatiotemporal DINOv2 + Chronos model.
Train Set: SCAMPS (Train Split); Test Set: NATURE

Evaluation: Sensor End-of-Window Ground Truth Comparison

Signal Processing	MAE	MedAE	RMSE	5 BPM
rPPG-Toolbox FFT	13.36	7.49	20.63	36.69%
rPPG-Toolbox Peak	70.80	72.18	76.18	0.68%
SciPy Periodogram	26.57	10.49	41.15	32.62%
SciPy CZT	22.78	9.34	36.49	26.61%

Table 4.13: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; SCAMPS/NATURE.

Ground Truth vs Predicted Heart Rate Values

The ground truths are obtained using sensor end-of-window comparisons for the figures.

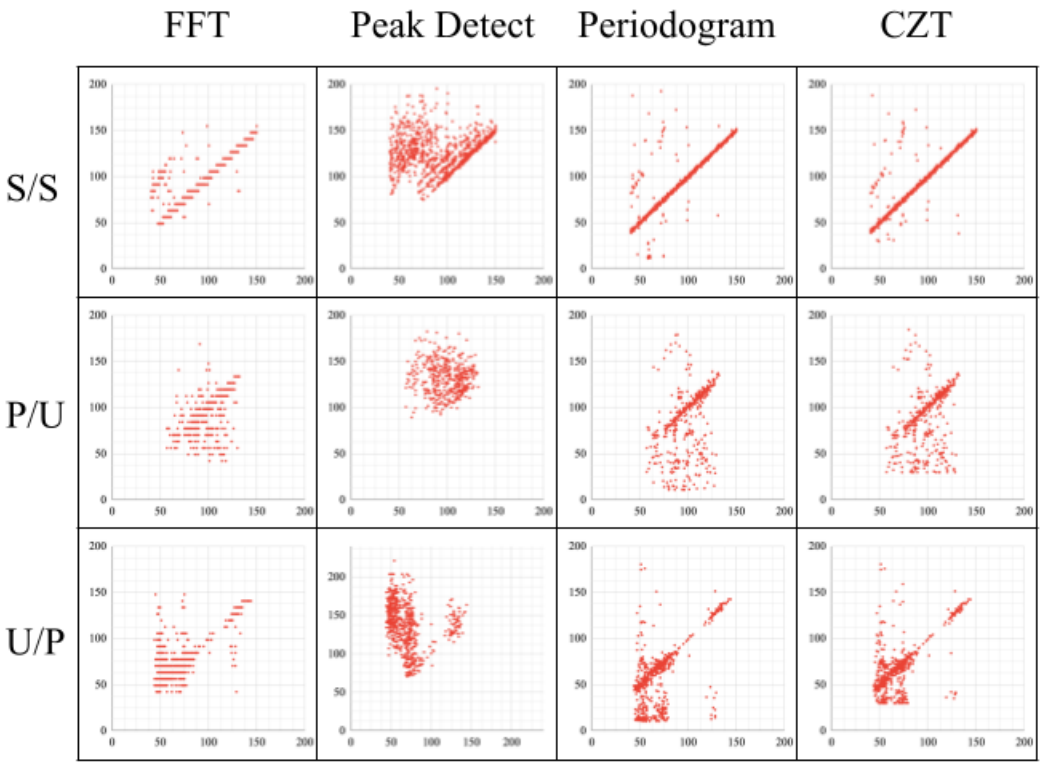

Figure 4.2: DINOv2-Small + Chronos-Tiny; Sensor End-of-Window Comparison.
Comparison of ground truth (x-axis) and predicted (y-axis) heart rate using the spatiotemporal model across multiple datasets. S, P, and U denote SCAMPS, PURE, and UBFC-rPPG, respectively.

Train Set: SCAMPS (Train Split); Test Set: SCAMPS (Test Split)
Model: DINOv2-Small + Chronos-Base

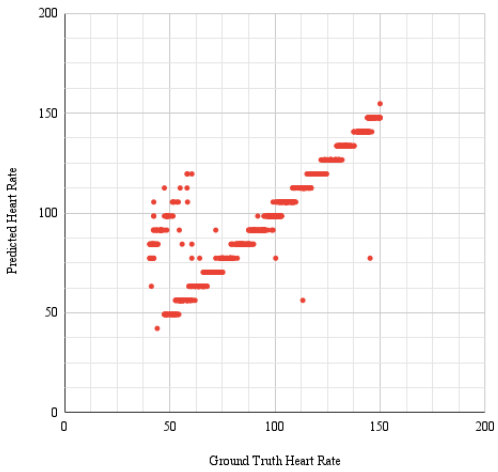

Figure 4.3: Ground Truth vs Predicted Heart Rate Values, SCAMPS Dataset (rPPG-Toolbox FFT)

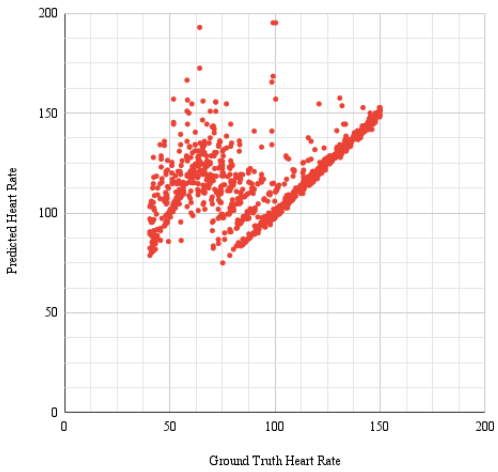

Figure 4.4: Ground Truth vs Predicted Heart Rate Values, SCAMPS Dataset (rPPG-Toolbox Peak Detection)

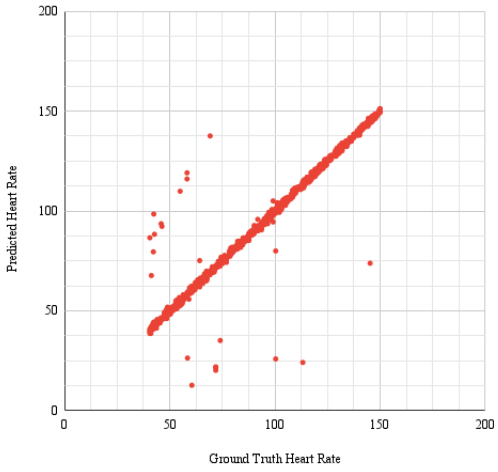

Figure 4.5: Ground Truth vs Predicted Heart Rate Values, SCAMPS Dataset (Periodogram)

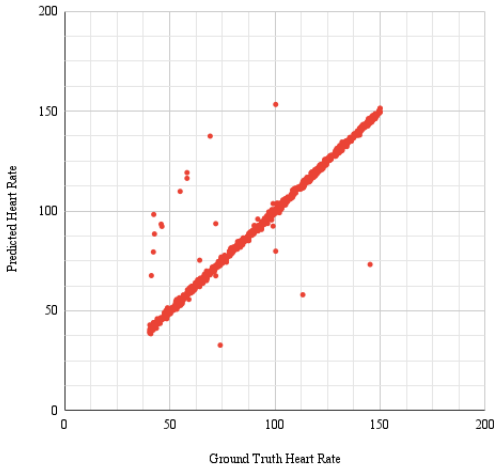

Figure 4.6: Ground Truth vs Predicted Heart Rate Values, SCAMPS Dataset (CZT)

Train Set: UBFC-rPPG; Test Set: PURE
Model: DINOv2-Small + Chronos-Base

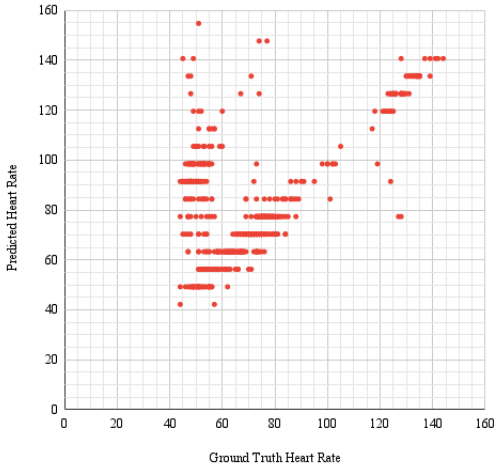

Figure 4.7: Ground Truth vs Predicted Heart Rate Values, PURE Dataset (rPPG-Toolbox FFT)

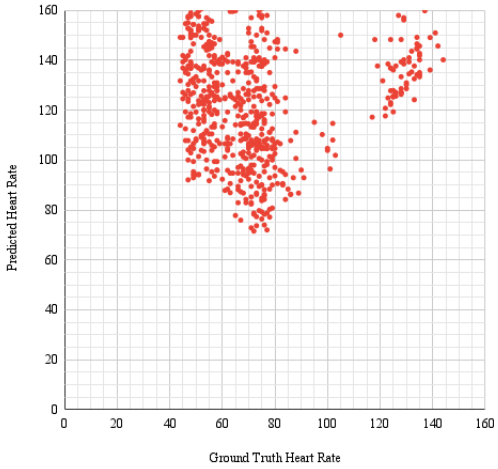

Figure 4.8: Ground Truth vs Predicted Heart Rate Values, PURE Dataset (rPPG-Toolbox Peak Detection)

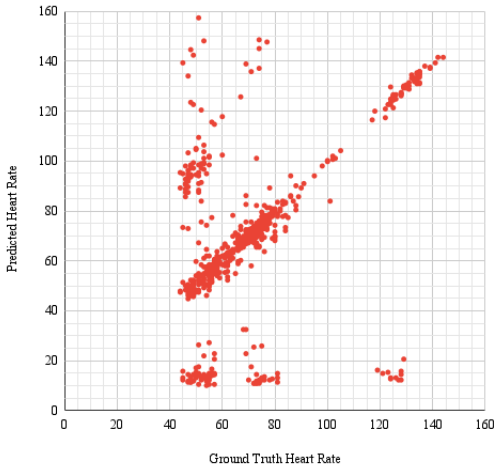

Figure 4.9: Ground Truth vs Predicted Heart Rate Values, PURE Dataset (Periodogram)

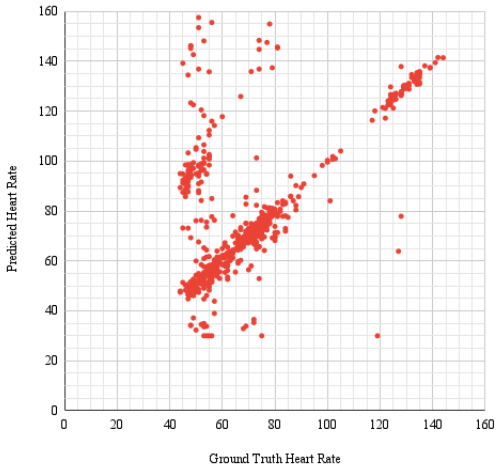

Figure 4.10: Ground Truth vs Predicted Heart Rate Values, PURE Dataset (CZT)

Train Set: PURE; Test Set: UBFC-rPPG
Model: DINOv2-Small + Chronos-Base

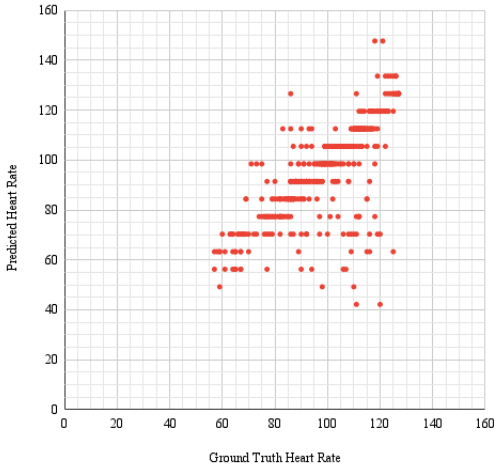

Figure 4.11: Ground Truth vs Predicted Heart Rate Values, UBFC-rPPG Dataset (rPPG-Toolbox FFT)

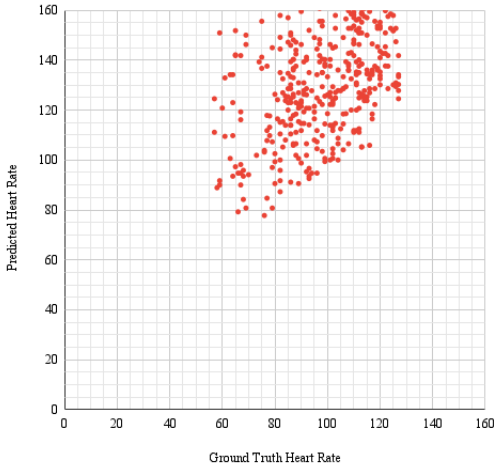

Figure 4.12: Ground Truth vs Predicted Heart Rate Values, UBFC-rPPG Dataset (rPPG-Toolbox Peak Detection)

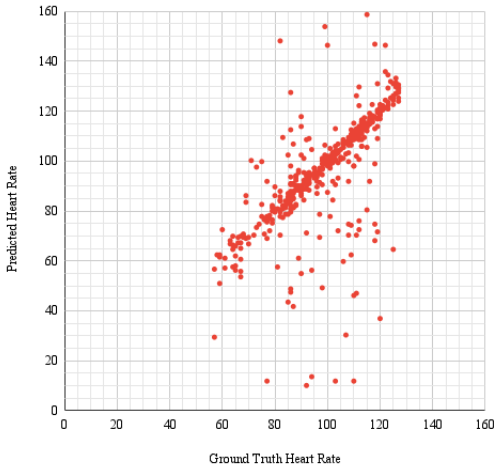

Figure 4.13: Ground Truth vs Predicted Heart Rate Values, UBFC-rPPG Dataset (Periodogram)

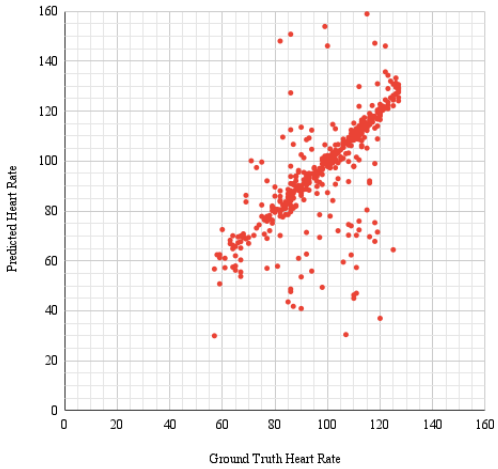

Figure 4.14: Ground Truth vs Predicted Heart Rate Values, UBFC-rPPG Dataset (CZT)

4.5.2 Respiratory Rate

Synthetic Dataset

In addition to heart rate, the SCAMPS synthetic dataset provides aligned respiratory signals and ground truth breathing rate. The breathing rate pipeline is the same as the heart rate pipeline, except the spatiotemporal architecture is trained on the respiratory signals. Evaluations are conducted similarly to heart rate, and using signal processing methods periodogram and CZT with a broad bandwidth of 0.0 Hz to 10.0 Hz.

Performance for respiratory rate of the Spatiotemporal DINOv2 + Chronos model. Train Set: SCAMPS (Train Split); Test Set: SCAMPS (Test Split)

Evaluation: Sensor End-of-Window Ground Truth Comparison

Signal Processing	MAE	MedAE	RMSE	3 BrPM
SciPy Periodogram	4.42	2.52	6.81	56.17%
SciPy CZT	3.00	1.52	5.00	71.58%

Table 4.14: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; SCAMPS/SCAMPS.

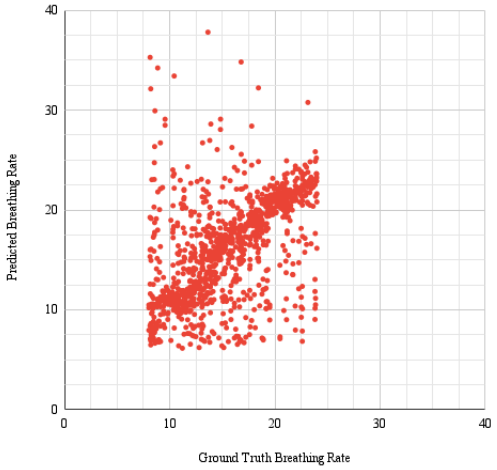

Figure 4.15: Ground Truth vs Predicted Breathing Rate Values, SCAMPS Dataset (CZT)

4.6 Ablation Studies

To understand the individual contributions of the visual and temporal foundation models in the rPPG pipeline, ablation studies were conducted with DINOv2 and Chronos in isolation. In the ablation studies, DINOv2-Small and Chronos-Tiny were evaluated with the 5 BPM threshold metric on the SCAMPS dataset for heart rate. When DINOv2-Small and Chronos-Tiny were combined for the full spatiotemporal architecture on this dataset, the correctness within 5 BPM for heart rate is 94.83%.

When the model used solely DINOv2 features, it consistently failed to generate a BVP-like signal. The predictions were noisy and did not resemble a plausible physiological waveform. The visual model had 5.92% of predictions within 5 BPM of the ground truth heart rate on the SCAMPS dataset. This indicates that spatial features alone, without temporal modelling, cannot capture the temporal dynamics despite providing rich semantic and structural information.

Ablations studies with Chronos alone allowed the model to learn some temporal structure from raw frame sequences. The temporal Chronos configuration achieved 63.42% of predictions within 5 BPM of the ground truth heart rate on the SCAMPS dataset. This demonstrates the ability of the temporal-only architecture to extract some amount of the physiological signal, but it is still considerably below the full system, which achieves 94.83%.

Thus, these results demonstrate that the accurate rPPG estimations emerge from the integration of both spatial and temporal modelling. DINOv2 captures spatial representations of subtle variations in the skin appearance, while Chronos captures the temporal patterns that correspond to physiological signals. The joint use of two foundation models for a spatiotemporal architecture allows for more accurate BVP prediction.

Chapter 5

Analysis

The analysis chapter provides a discussion of the experimental results, as well as a systematic investigation of components within the rPPG pipeline. The goal is to assess the spatiotemporal architecture as a whole, while also identifying possibilities for improvement and potential flaws present within data and evaluation.

5.1 Results with Datasets

The results of the spatiotemporal architecture for rPPG are discussed for various datasets, and the datasets themselves are assessed.

5.1.1 Synthetic Dataset

With the official Train/Validation/Test splits provided by the SCAMPS dataset, the spatiotemporal architecture performs well during evaluation.

The strongest result is obtained with the DINOv2-Small + Chronos-Base configuration, with an MAE of 1.39 BPM, MedAE of 0.62 BPM, and within 5 BPM percentage of 98.50%. This indicates that the spatiotemporal architecture can recognize the PPG signal from the simulation data, and prompts further experiments with real-world data.

It is notable that the RMSE, relative to the other metrics, is somewhat high. This can be attributed to the outliers with a high error that cause the squared error to be large. The MedAE of 0.62 BPM with CZT shows that most of the errors when tested on the SCAMPS dataset are very small. The 90th percentile absolute error for the SCAMPS evaluation is 2.70 BPM, meaning 90% of the results have an absolute error of $\leq$ 2.70 BPM. Some large estimation errors occur also in part due to other components of the rPPG pipeline, such as signal processing.

To the author's knowledge, there has not been much prior research conducted on testing SCAMPS using the provided Train/Validation/Test splits. Although SCAMPS is not designed as a test set for evaluating the clinical efficacy of a model and success with SCAMPS does not necessarily indicate generalizability to real people, testing on SCAMPS is an appropriate, helpful step to validate the feasibility of a model. The results obtained on SCAMPS by the spatiotemporal

architecture can serve as a benchmark for such purposes, for heart rate and possibly respiratory rate.

5.1.2 Real World Datasets

The results on the real-world datasets reveal that the spatiotemporal architecture demonstrates an ability to learn the PPG waveforms on real people, albeit with more errors compared to the synthetic dataset.

When compared to the benchmark results 4.6 provided by rPPG-Toolbox, the spatiotemporal model is comparable to some of the values achieved by other methods.

Failure Cases

Due to the lower performance on real-world datasets, it can be helpful to diagnose what the failure cases are.

The 10 people in the PURE dataset were each set up in 6 different settings: steady, talking, slow translation, fast translation, small rotation, and medium rotation. These configurations are evaluated in terms of signal end-of-window MAE, grouped by the 6 different settings, with UBFC-rPPG as the train set, PURE as the test set, and CZT for signal processing.

PURE Dataset Configurations

Index	**Configuration**	**Model Results**			
		DINOv2-Small + Chronos-Tiny		DINOv2-Small + Chronos-Base	
		MAE	5 BPM	MAE	5 BPM
	Overall	11.50	57.34%	13.37	67.39%
01	Steady	4.42	78.64%	9.54	75.73%
02	Talking	16.16	54.64%	17.28	59.79%
03	Slow Translation	9.02	53.72%	15.91	64.46%
04	Fast Translation	22.59	30.08%	20.33	30.08%
05	Small Rotation	7.39	62.38%	7.62	80.20%
06	Medium Rotation	7.84	56.86%	7.83	79.41%

Table 5.1: Evaluations grouped by configurations of the PURE dataset

From the failure cases, it is evident that the spatiotemporal architecture overall performs the best when the subject is steady. The worst model performance was obtained during fast translation, which involved head movements parallel to the camera plane. Talking also considerably lowered model performance compared to steady. Comparatively, the model was robust to the head rotation settings.

5.1.3 Dataset Review

Train Dataset

In the evaluation of rPPG methods, supervised methods are shown with cross-dataset training results, as presented by the Benchmark Results table 4.6. The train set in literature is typically

one of UBFC-rPPG [2], PURE [25], and SCAMPS [14].

It is noteworthy, although consistent with expectations, that the train set influences the evaluation results considerably. That is, when trained on different datasets and tested on the same dataset, the results differ. For instance, when tested on UBFC-Phys [16], the MAEs when the spatiotemporal model was trained on UBFC-rPPG [2], PURE [25], and SCAMPS [14] differed by up to around 8 BPM. This demonstrates that the train set can influence the prediction accuracy of the spatiotemporal rPPG architecture significantly. The results did not show there to be a universal best train set; a train set could yield relatively better results on one dataset and worse on another. This may be caused by different environmental conditions, such as variations in lighting, subject motion, camera characteristics, that result in biases that are specific to a dataset. It is also possible for models to overfit to certain properties, leading to poorer generalization to unseen settings. This highlights the importance of diverse training data, and the possible benefits of increasing the scope of the train set, for instance, by combining multiple datasets in training. The use of foundation models attempts to mitigate such generalization issues by having seen large amounts of data, such that no rPPG-specific adjustment of the architecture is necessary, requiring only training the lightweight prediction head. Regardless, the spatiotemporal model may be improved further through combining multiple datasets for the training of the prediction head.

Dataset Limitations

Accurate reporting of results is also dependent on correct and complete data. Two dataset limitations can affect the accuracy of the results: incomplete PPG waveforms and assumptions regarding frame rate and time.

Incomplete PPG waveforms can occur due to sensor malfunctions of data loss during collection, which lead to flatlines or missing segments in the signal. This can affect the training and evaluation of rPPG models, since the data does not represent the physiological signals accurately. Erroneous data annotations add noise to the labels, misleading the model. This error is observed in datasets such as UBFC-rPPG and MMPD, where on occasion, the heart rate values or signal were dropped. UBFC-rPPG had some timestamps with a corresponding heart rate ground truth value of 1–4 BPM, which were incorrect and fixed for sensor end-of-window metrics through linear interpolation. Unless excluded, errors in PPG segments would cause a lower reported performance of the model.

Another possible limitation involves assumptions with data collection. The frame rate is assumed to be fixed, in most cases at 30 FPS. Time windows, signal processing, and the rPPG pipeline depend on this sampling rate for window lengths and frequency resolutions. In practice, real-world video data may exhibit variations in the frame rate, in part due to the hardware which may cause dropped or duplicated frames. Inconsistent frame rates can cause issues in both training and deployment. There can be misalignments between visual inputs and physiological signals, causing inaccurate frequency estimation. Additionally, time alignment errors between the frames and corresponding PPG signals may cause distortions in the learned relationship between visual cues and the physiological signal. This may arise due to sensor latency, delays, or data collection post-processing procedures. Temporal shifts can add additional noise during training and evaluation. Challenges with the frame rate and time alignment were experienced

first-hand during data collection with the NATURE dataset, where the frame rate of the camera was not perfectly stable and post-processing was necessary during data cleaning.

5.2 Signal Processing

The goal of signal processing in the context of rPPG is to accurately extract the heart rate given a BVP waveform. This is a post-processing step that does not concern the rPPG model itself, but can have a substantial impact on the overall heart rate value prediction accuracy. The effects are shown by the experimental results: with the same BVP waveform, the signal processing method can lead to vastly different results for the scalar heart rate value. The dependency of the rPPG pipeline on a signal processing method adds an additional variable to the development and analysis of rPPG systems.

Accurate signal extraction from rPPG waveform predictions can be challenging in part because PPG waves are not pure sinusoidal waves, which are what DFT works best on. Peak detection can also be difficult to adjust due to this reason. One interbeat interval typically consists of a systolic phase and a diastolic phase, with four main points: foot/onset, systolic peak, dicrotic notch, and diastolic peak.

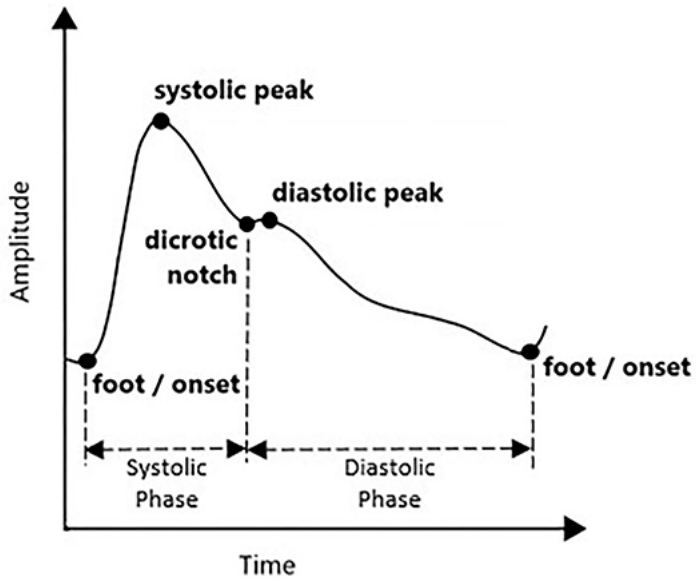

Figure 5.1: Typical PPG waveform. [26]

Depending on the sensor and subject, the diastolic peak may be more or less prominent compared to the systolic peak. In cases where the diastolic peak is more visible, simple signal processing methods will mistake the diastolic peak belonging to a separate interbeat interval compared to the systolic peak, which could cause incorrect heart rate values.

Out of the four signal processing methods utilized (rPPG-Toolbox FFT, rPPG-Toolbox Peak Detection, SciPy Periodogram, SciPy CZT), SciPy CZT generally has the best performance, achieving a lower MAE and higher percentage of test files within 5 BPM. This may be due to

the finer frequency bins as well as the ability to provide frequency bounds that correspond to physiological signals. Still, it encounters issues with choosing the correct frequency.

5.2.1 Fourier Transform: Harmonics

The Discrete Fourier Transform transforms a signal from the time domain into the frequency domain, allowing the computation of the power spectral density (PSD), or power spectrum, which shows how the power of a signal is distributed across frequencies. Higher powers or peaks in the PSD indicate that the corresponding frequency is more present in the original signal.

A prevalent issue among the tests occurs with the step of selecting the correct frequency or heart rate during signal processing, once the PSD is obtained. That is, given the PSD, find the correct frequency corresponding to the heart rate. This can be challenging, not necessarily only because of the spatiotemporal model, but the existence of harmonics in the PSD; even with a perfect PPG waveform prediction, the PSD after signal processing may not be simple to process.

Harmonics are integer multiples of a fundamental frequency. For instance, for a fundamental frequency of 60 Hz, harmonics include 120 Hz, 180 Hz, 240 Hz, and so on. Additionally, it is also possible that a selected frequency is $\frac{1}{N}$ of the true frequency for some integer N, which indicates that the true frequency was interpreted as a harmonic.

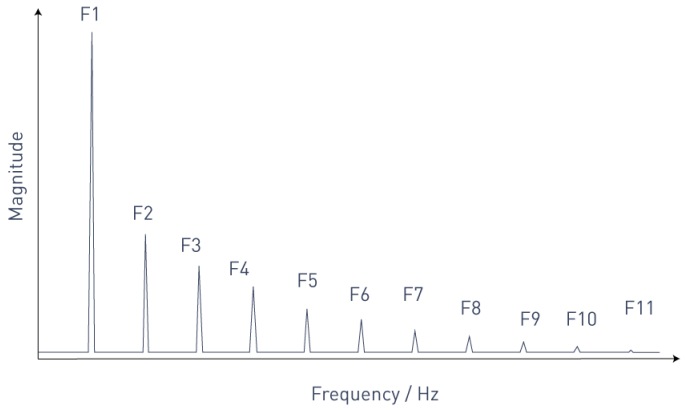

Figure 5.2: FFT-spectrum of a distorted sinusoidal signal with multiple harmonics. [17]

In heart rate predictions, harmonics are observed due to the PPG signal not being a pure unmodulated sine wave, as multiple peaks occur in a single interbeat interval. Other noise in predicted PPG signals can also cause inaccuracies. In the ground truth vs predicted heart rate value plots, harmonics would appear as $y = mx$ patterns for $\ldots, \frac{1}{3}, \frac{1}{2}, 2, 3, \ldots$, in addition to the correct $y = x$. The $y = 2x$ pattern is the most prevalent harmonic in the estimations of the spatiotemporal model.

5.2.2 Peak Detection Algorithm

An alternative to DFT-based PPG signal processing methods is to use peak detection on the PPG signal. The challenge with peak detection is finding an optimal configuration for what

classifies as a peak in the context of rPPG. An ideal method could try to find the local maxima corresponding to the systolic peaks, and attempt to filter out the diastolic peaks.

Peak detection can be implemented with the function `scipy.signal.find_peaks`, which provides many customization options, including optional parameters for `height`, `threshold`, `distance`, `prominence`, `width`, `wlen`, `rel_height`, and `plateau_size`. This function returns the indices of the peaks in the signal that satisfy all conditions along with properties of the returned peaks. Optimally, the distance between adjacent peaks represents the interbeat interval.

5.2.3 CZT Augmentations

To address the challenges due to harmonics of the fundamental frequency occurring in the power spectrum, signal processing methods are explored to more accurately predict the fundamental frequency, which corresponds to the heart rate value. The methods are built on top of the CZT signal processing method, operating directly on the power spectrum produced by CZT. That is, CZT is frozen, and augmentations are built purely on top of the power spectra that CZT returns. The inputs to CZT are the predicted PPG waveforms of the spatiotemporal architecture.

The base SciPy CZT implementation used throughout the experiments finds the tallest peak in the power spectrum and returns it as the predicted heart rate. In an ideal situation, the most prominent peak in the power spectrum is the desired fundamental frequency, however, there are cases when it is not.

To demonstrate the effect of CZT augmentations, identical predicted BVP signals are processed with four configurations: CZT Filtered, CZT Dot Product Halving, CZT Dot Product Decrease 0.25, and CZT Dot Product Decrease 0.3. The unaltered CZT is taken as the benchmark to evaluate the effectiveness of the augmentations on CZT.

CZT Filtered

Given the power spectrum returned by CZT, the CZT Filtered augmentation takes the lowest peak from the power spectrum that satisfies certain criteria.

It takes three additional parameters `n_tallest`, `ratio_to_tallest`, `distance_peaks`, and uses `scipy.signal.find_peaks` on the power spectrum to find all peaks that are at least `distance_peaks` apart. From these peaks, it takes the top `n_tallest` peaks. From these `n_tallest` peaks, it finds the peak with the minimum index that is at least as tall as `ratio_to_tallest * tallest_peak`, where `tallest_peak` is the height or power of the tallest peak in the power spectrum. The frequency corresponding to this peak is selected.

In the implementations, the parameter values for CZT Filtered were `n_tallest` = 5, `ratio_to_tallest` = 0.4, `distance_peaks` = 40000.

CZT Dot Product

Given the power spectrum returned by CZT, the CZT Dot Product augmentation takes a weighted dot product of powers corresponding to the harmonics of a frequency value, returning a new power spectrum. That is, for each frequency f, and $a_1 = 1$,

$$\begin{aligned}\text{power}_{\text{dot\_product}} &= \sum_{h=1}^{k} a_h \text{power}(hf) \\ &= \text{power}(f) + a_2\text{power}(2f) + a_3\text{power}(3f) + \ldots + a_k\text{power}(kf)\end{aligned}$$

Here, power represents the original power spectrum that CZT finds. As a function, power(f) for any frequency value f (Hz) returns the power corresponding to f in the CZT power spectrum. The frequency corresponding to the tallest peak in the new dot product power spectrum, $\text{power}_{\text{dot\_product}}$, is selected.

Three configurations of CZT Dot Product are explored in the implementations. In all cases, $k = 5$ = maximum harmonic appended.

- CZT Dot Product Halving: $a_h = \frac{1}{2^{h-1}}$
- CZT Dot Product Decrease 0.25: $a_h = 1 - 0.25 * (h - 1)$
- CZT Dot Product Decrease 0.30: $a_h = 1 - 0.30 * (h - 1)$

Results

To evaluate the CZT augmentation methods, the results focus on the PURE dataset, which especially demonstrated notable diastolic peaks within the ground truth PPG waveforms. Comparisons are made to the CZT baseline of selecting the tallest peak in the power spectrum without further processing.

CZT Augmentation	MAE	MedAE	RMSE	5 BPM
CZT Base	13.37	2.26	28.00	67.39%
CZT Filtered	6.68	2.01	14.75	74.65%
CZT Dot Product Halving	10.48	2.23	22.99	69.09%
CZT Dot Product Decrease 0.25	8.67	2.39	17.89	68.32%
CZT Dot Product Decrease 0.30	9.46	2.23	20.18	69.40%

Table 5.2: DINOv2-Small + Chronos-Base;
CZT augmentation methods on the PURE dataset, trained on UBFC-rPPG.

The results of CZT augmentation methods on the PURE dataset demonstrate improvements across evaluation metrics compared to the CZT base. All of these methods partially eliminate the second harmonic, represented by the $y = 2x$ line in the ground truth vs predicted heart rate values plots.

Ground Truth vs Predicted Heart Rate Values

CZT Augmentation Comparisons. The ground truths are obtained using sensor end-of-window comparisons.

Train Set: UBFC-rPPG; Test Set: PURE

Model: DINOv2-Small + Chronos-Base

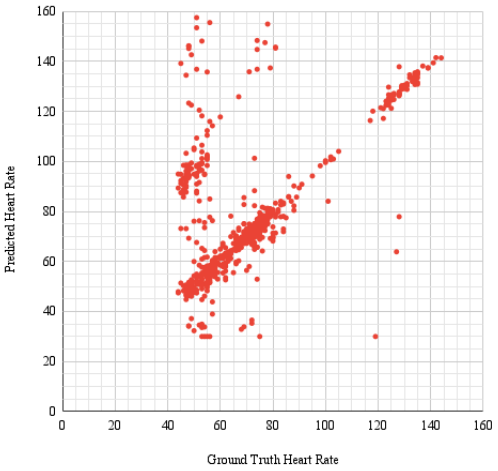

Figure 5.3: Ground Truth vs Predicted Heart Rate Values, PURE Dataset
(CZT Base)

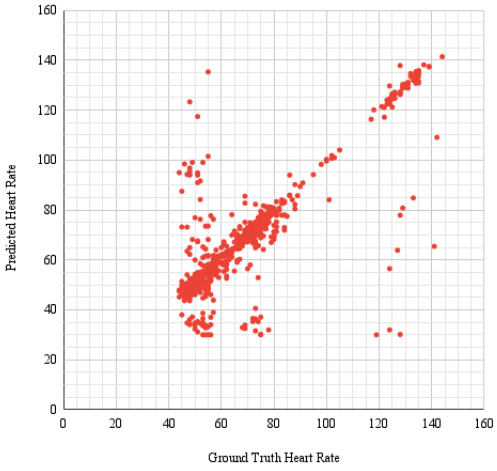

Figure 5.4: Ground Truth vs Predicted Heart Rate Values, PURE Dataset
(CZT Filtered)

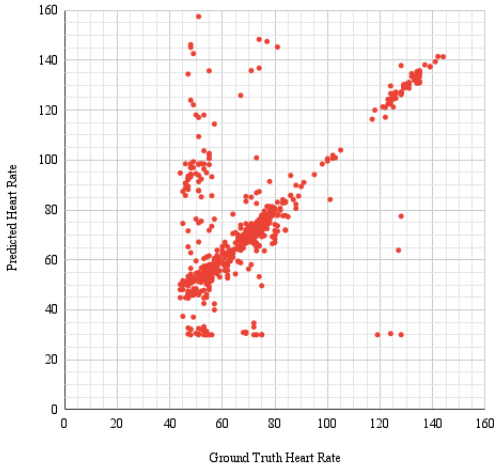

Figure 5.5: Ground Truth vs Predicted Heart Rate Values, PURE Dataset
(CZT Dot Product Halving)

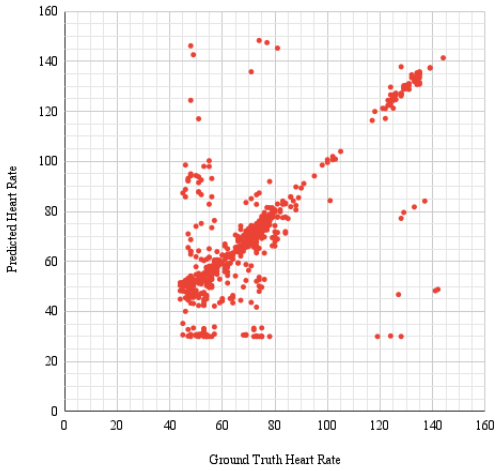

Figure 5.6: Ground Truth vs Predicted Heart Rate Values, PURE Dataset
(CZT Dot Product Decrease 0.25)

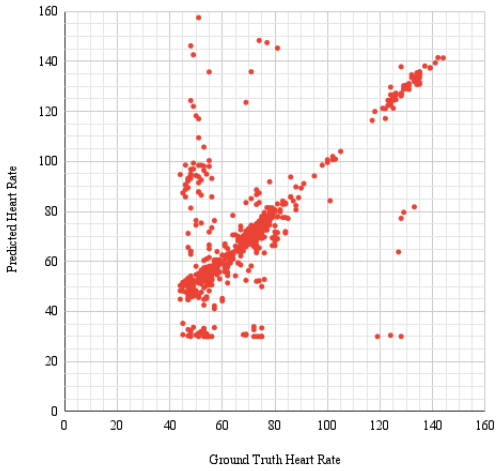

Figure 5.7: Ground Truth vs Predicted Heart Rate Values, PURE Dataset
(CZT Dot Product Decrease 0.30)

5.3 Ground Truth Discrepancies

In this work, the spatiotemporal rPPG architecture is evaluated with two classes of metrics: the waveform window signal processing metrics adopted from the rPPG-Toolbox, and the new proposed sensor end-of-window ground truth comparison metrics. These metric groups differ in how the ground truth heart rate values are defined, leading to differences quantitatively and qualitatively. This section examines these distinctions, focusing specifically on how ground truth heart rate values are extracted from data, and how different methods of extracting the ground truth can lead to varied results.

Despite identical train and test configurations, the scalar heart rate ground truth values reported by the two evaluation metrics yield different quantitative results. These discrepancies arise from differences in how ground truth heart rate values are extracted from the reference sensor data within each evaluation metric set. Thus, the metric choice impacts the evaluation and interpretation of the rPPG pipeline, and it can thus be helpful to assess this difference. Through this analysis, it can be shown that the proposed sensor end-of-window ground truth comparison provides a more physiologically consistent evaluation of rPPG predictions.

In literature, most rPPG models are evaluated using the rPPG-Toolbox waveform window signal processing metrics. The ground truth heart rate value is obtained by using signal processing on each segment of the PPG data. A criticism of this method is that it depends on the signal processing implementation, since each signal processing implementation can yield a different heart rate value. In a naive, hypothetical situation, the signal processing implementation returns 70 BPM regardless of the signal. Then, since the rPPG prediction pipeline also uses signal processing, both the ground truth and predicted heart rates would simply become 70 BPM, suggesting an exact, perfect prediction. While this is not the case in reality, it illustrates the influence of signal processing on the ground truth, and how it can be problematic for justified

evaluation of an rPPG model.

As a result of these observations, it is desired that the ground truth values are fully isolated from any form of external implementations, and extracted directly from the dataset. Many rPPG datasets, such as PURE and UBFC-rPPG, provide ground truth readings that include the sensor-reported heart rate value at each timestep. The sensor end-of-window evaluation metrics extract the scalar heart rate values from this without further processing.

Qualitatively, the sensor end-of-window evaluation metrics are beneficial over the rPPG-Toolbox waveform window signal processing metrics in stability and realistic fluctuations. Realistically, it is normal and expected for the heart rate to fluctuate subtly without large jumps, and not remain exactly identical over a long period of time. These fluctuations usually are not extremely large and result from reasons such as stress and physical activity.

The first observation is that there were instances of the FFT causing unrealistic heart rate jumps between 6-second windows, where the participant does not seemingly exhibit this behaviour.

The second observation is that the rPPG-Toolbox implementation of FFT has large frequency bins, which led to there being cases where the heart rate was reported as the exact same for extended periods of time, for instance, 48 seconds ($8 \times$ 6-second windows). While this could be improved with finer FFT bins, the takeaway is that generic signal processing methods such as Fourier transform and peak detection are not tuned specifically to fit physiologically realistic fluctuations.

Examples of ground truth heart rate values obtained using both evaluation metric sets are shown across the duration of files from UBFC-rPPG and PURE. These demonstrate cases of large differences between evaluation metrics with poor stability from waveform window signal processing metrics from rPPG-Toolbox. The figures and tables show Window Number (WN), the heart rate obtained by sensor end-of-window ground truth (R_{EoW}), and the heart rate obtained by rPPG-Toolbox waveform window signal processing (R_{WSP}).

UBFC-rPPG, Subject 36

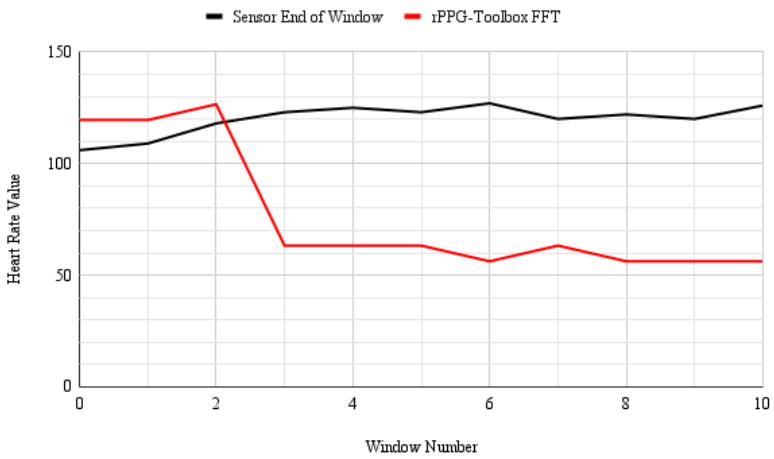

Figure 5.8: Ground Truth Discrepancy Example, Window Number vs Heart Rate Value Obtained by Sensor End-of-Window and [rPPG-Toolbox] Waveform Window Signal Processing (rPPG-Toolbox FFT).
Data: UBFC-rPPG, Subject 36

WN	R_{EoW}	R_{WSP}
0	106	119.53125
1	109	119.53125
2	118	126.5625
3	123	63.28125
4	125	63.28125
5	123	63.28125
6	127	56.25
7	120	63.28125
8	122	56.25
9	120	56.25
10	126	56.25

Table 5.3: Ground Truth Discrepancy Example, Window Number vs Heart Rate Values: Sensor End-of-Window (R_{EoW}) and [rPPG-Toolbox] Waveform Window Signal Processing (R_{WSP}) (rPPG-Toolbox FFT).
Data: UBFC-rPPG, Subject 36

UBFC-rPPG, Subject 39

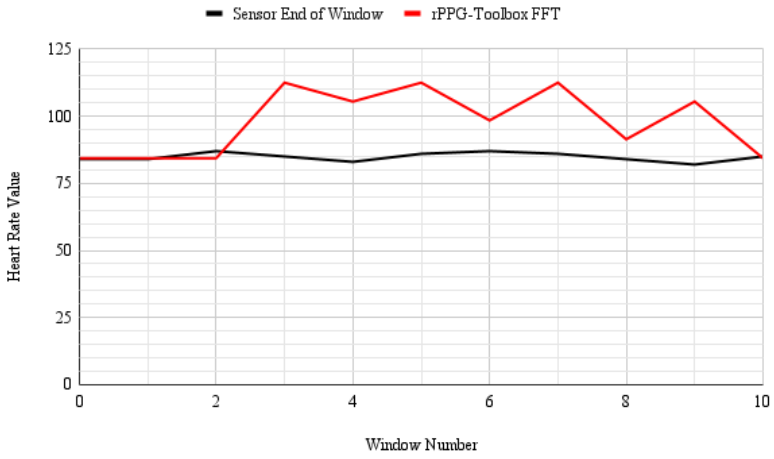

Figure 5.9: Ground Truth Discrepancy Example, Window Number vs Heart Rate Value Obtained by Sensor End-of-Window and [rPPG-Toolbox] Waveform Window Signal Processing (rPPG-Toolbox FFT).
Data: UBFC-rPPG, Subject 39

WN	R_{EoW}	R_{WSP}
0	84	84.375
1	84	84.375
2	87	84.375
3	85	112.5
4	83	105.46875
5	86	112.5
6	87	98.4375
7	86	112.5
8	84	91.40625
9	82	105.46875
10	85	84.375

Table 5.4: Ground Truth Discrepancy Example, Window Number vs Heart Rate Values: Sensor End-of-Window (R_{EoW}) and [rPPG-Toolbox] Waveform Window Signal Processing (R_{WSP}) (rPPG-Toolbox FFT).
Data: UBFC-rPPG, Subject 39

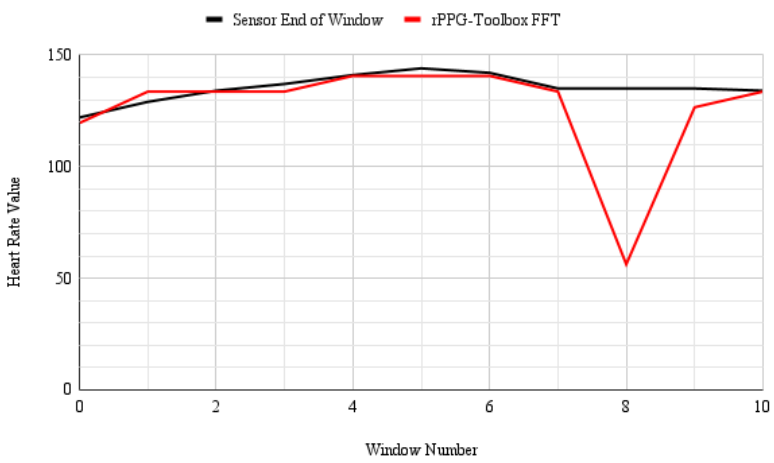

Figure 5.10: Ground Truth Discrepancy Example, Window Number vs Heart Rate Value Obtained by Sensor End-of-Window and [rPPG-Toolbox] Waveform Window Signal Processing (rPPG-Toolbox FFT).
Data: PURE, 07-02

WN	R_{EoW}	R_{WSP}
0	122	119.53125
1	129	133.59375
2	134	133.59375
3	137	133.59375
4	141	140.625
5	144	140.625
6	142	140.625
7	135	133.59375
8	135	56.25
9	135	126.5625
10	134	133.59375

Table 5.5: Ground Truth Discrepancy Example, Window Number vs Heart Rate Values: Sensor End-of-Window (R_{EoW}) and [rPPG-Toolbox] Waveform Window Signal Processing (R_{WSP}) (rPPG-Toolbox FFT).
Data: PURE, 07-02

To quantitatively assess the overall difference between the two sets of evaluation metrics, the mean absolute difference (MAD), the median absolute difference (MedAD), and the root mean squared difference (RMSD) are computed. These are calculated as follows, where R_{WSP} represents the heart rate obtained by rPPG-Toolbox waveform window signal processing, R_{EoW} represents the heart rate obtained by sensor end-of-window ground truth, and N represents the number of instances.

- **Mean Absolute Difference (MAD)**
  MAD = $\frac{1}{N}\sum_{n=1}^{N}|R_{WSP} - R_{EoW}|$
- **Median Absolute Difference (MedAD)**
  MedAD = median $(|R_{WSP} - R_{EoW}|)$
- **Root Mean Squared Difference (RMSD)**
  RMSD = $\sqrt{\frac{1}{N}\sum_{n=1}^{N}(R_{WSP} - R_{EoW})^2}$

For the rate obtained through rPPG-Toolbox waveform window signal processing R_{WSP}, four different signal processing methods are included in the difference computations, which are those used throughout the experimental results: rPPG-Toolbox FFT, rPPG-Toolbox Peak, SciPy Periodogram, and SciPy CZT. rPPG-Toolbox uses rPPG-Toolbox FFT for the metrics calculations, including the Benchmark Results 4.6.

Dataset	Signal Processing	MAD	MedAD	RMSD
PURE	rPPG-Toolbox FFT	3.10	2.25	5.21
	rPPG-Toolbox Peak	14.99	4.52	24.74
	SciPy Periodogram	2.59	1.42	6.89
	SciPy CZT	2.55	1.41	6.71
UBFC-rPPG	rPPG-Toolbox FFT	4.27	2.50	7.11
	rPPG-Toolbox Peak	4.64	1.74	8.31
	SciPy Periodogram	3.73	1.38	6.98
	SciPy CZT	3.74	1.37	7.00

Table 5.6: Discrepancy between rPPG-Toolbox Waveform Window Signal Processing Metrics and Sensor End-of-Window Ground Truth Comparison Metrics.

Ground Truth Discrepancies for Heart Rate

Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing

PURE Dataset

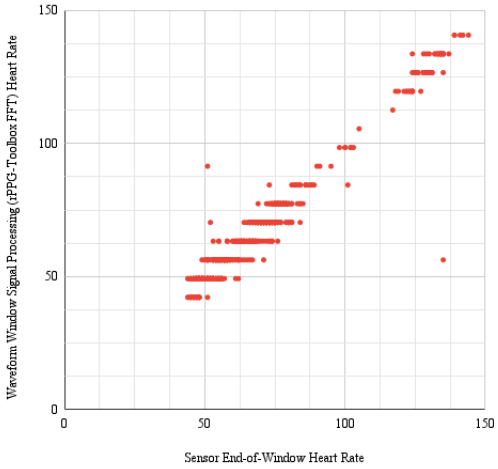

Figure 5.11: Ground Truth Discrepancies for Heart Rate, PURE Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (rPPG-Toolbox FFT)

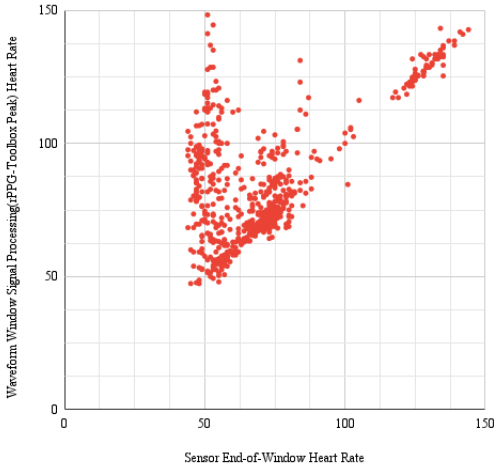

Figure 5.12: Ground Truth Discrepancies for Heart Rate, PURE Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (rPPG-Toolbox Peak)

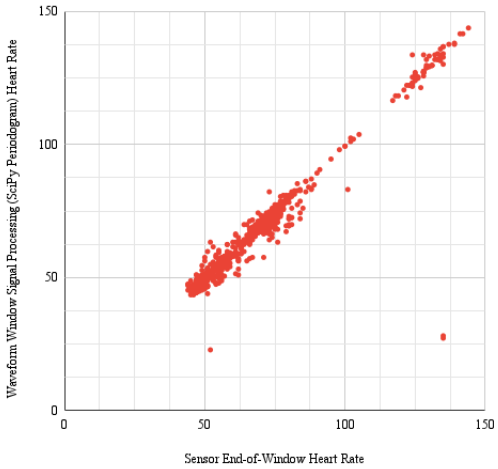

Figure 5.13: Ground Truth Discrepancies for Heart Rate, PURE Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (SciPy Periodogram)

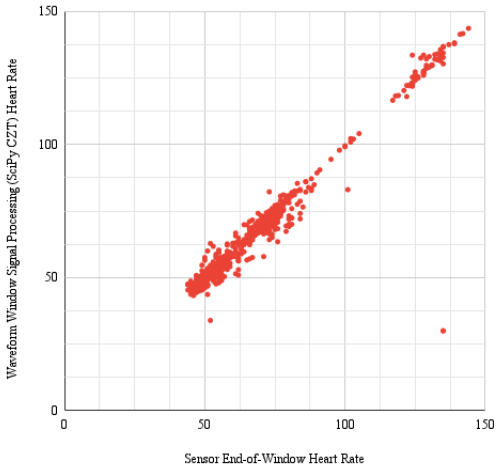

Figure 5.14: Ground Truth Discrepancies for Heart Rate, PURE Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (SciPy CZT)

UBFC-rPPG Dataset

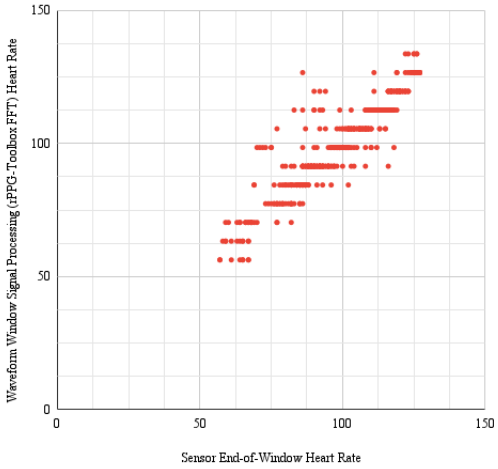

Figure 5.15: Ground Truth Discrepancies for Heart Rate, UBFC-rPPG Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (rPPG-Toolbox FFT)

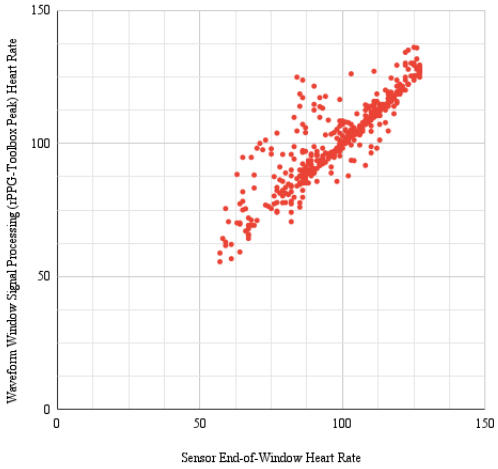

Figure 5.16: Ground Truth Discrepancies for Heart Rate, UBFC-rPPG Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (rPPG-Toolbox Peak)

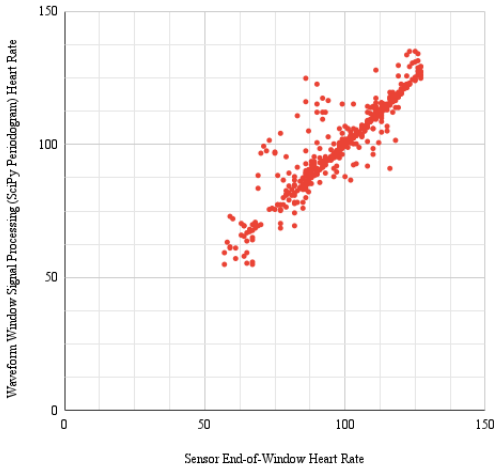

Figure 5.17: Ground Truth Discrepancies for Heart Rate, UBFC-rPPG Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (SciPy Periodogram)

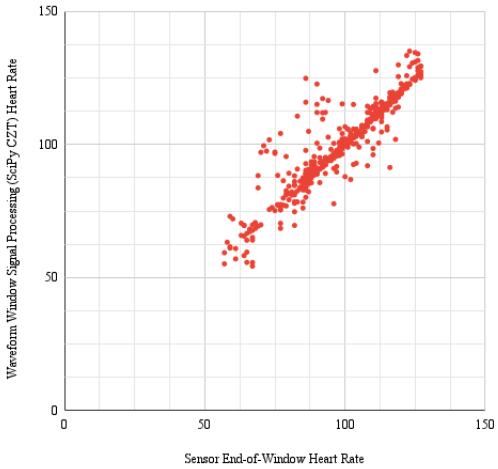

Figure 5.18: Ground Truth Discrepancies for Heart Rate, UBFC-rPPG Dataset: Sensor End-of-Window vs [rPPG-Toolbox] Waveform Window Signal Processing (SciPy CZT)

These results indicate that the quantitative results differ when computed using the two sets of evaluation metrics. While this inconsistency is not extreme, it demonstrates that the reported performance is dependent on the choice of the evaluation framework.

An assumption of the Signal End-of-Window is that there is not a large amount of fluctuation occurs throughout each window. To address this, absolute difference metrics are computed for the windows: the mean absolute difference (MAD), the median absolute difference (MedAD), and the p^{th} percentile absolute difference ($\%_p$AD) at $p \in \{25, 75\}$ are computed. These are calculated as follows, where $R_{EoW\text{-}s}$ represents the heart rate obtained by sensor start-of-window ground truth, $R_{EoW\text{-}e}$ represents the heart rate obtained by sensor end-of-window ground truth, and N represents the number of windows.

- **Mean Absolute Difference (MAD)**
  MAD = $\frac{1}{N}\sum_{n=1}^{N} |R_{EoW\text{-}e} - R_{EoW\text{-}s}|$
- **Median Absolute Difference (MedAD)**
  MedAD = median $(|R_{EoW\text{-}e} - R_{EoW\text{-}s}|)$
- **Root Mean Squared Difference (RMSD)**
  $\%_p$AD = percentile$_p$ $(|R_{EoW\text{-}e} - R_{EoW\text{-}s}|)$

Dataset	MAD	$\%_{25}$AD	MedAD	$\%_{75}$AD
PURE	1.47	0	1	2
UBFC-rPPG	4.60	1	2	5

Table 5.7: Absolute Difference Metrics for Start and End of Window Heart Rate Values.

The data demonstrates that, within a 6-second window, the heart rate typically does not fluctuate a large amount. Notably, UBFC-rPPG involves participants playing a time-sensitive mathematical game, so relatively larger fluctuations are expected.

Ultimately, an accurate ground truth value allows a stronger, more justified evaluation of whether a predicted value is reasonable.

5.4 Model Size Comparison

The proposed spatiotemporal architecture is modular and generalizable, which allows the ability to change model size and the models themselves easily. It is beneficial to change separate components and factors to assess what affects the performance. One such element is the model size.

In this thesis, two variations of the spatiotemporal rPPG architecture are evaluated: DINOv2-Small + Chronos-Tiny and DINOv2-Small + Chronos-Base. It is observed that changing the size of pretrained Chronos from a smaller (Tiny) model to a larger model (Base) results in improvements in PPG and heart rate prediction accuracy. In particular, both DINOv2-Small + Chronos-Tiny and DINOv2-Small + Chronos-Base perform well on the synthetic SCAMPS dataset, but DINOv2-Small + Chronos-Base outperforms DINOv2-Small + Chronos-Tiny on the real-world datasets.

On the SCAMPS dataset, the qualitative observation is that the DINOv2-Small + Chronos-Tiny PPG waveform predictions were not as precise to the label as the DINOv2-Small + Chronos-Base PPG waveform predictions. The smaller model, however, was able to pick up a pattern that, after signal processing, still resulted in accurate heart rate predictions. By changing the Chronos size from Tiny to Base, the MAE with rPPG-Toolbox metrics improved from 3.65 BPM to 2.46 BPM; with Sensor End-of-Window comparison using SciPy CZT Signal Processing, the MAE improved from 3.29 BPM to 1.39 BPM and the 5 BPM threshold percentage improved from 94.83% to 98.50%.

Model Size Comparison: PPG Predictions, SCAMPS P000443

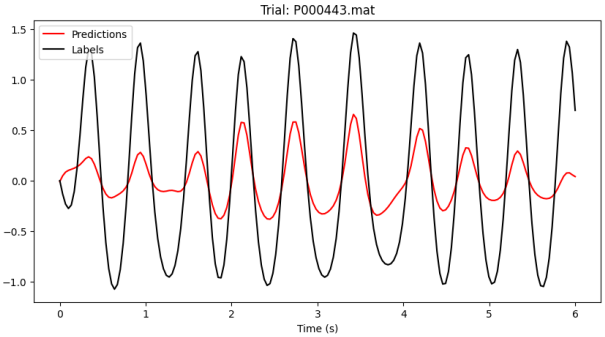

Figure 5.19: Spatiotemporal Model DINOv2-Small + Chronos-Tiny, PPG Prediction, SCAMPS P000443

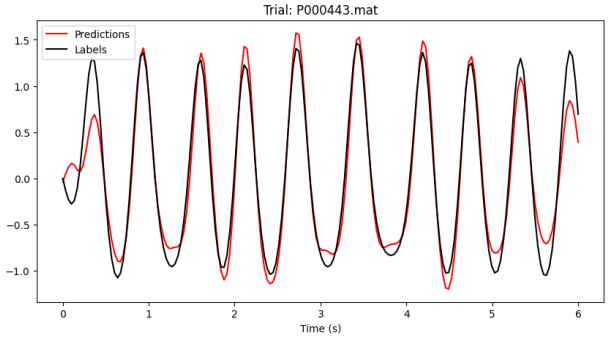

Figure 5.20: Spatiotemporal Model DINOv2-Small + Chronos-Base, PPG Prediction, SCAMPS P000443

On real-world datasets, the improvement is more significant. For instance, when trained on UBFC-rPPG and tested on PURE, changing the Chronos size from Tiny to Base, and evaluated with Sensor End-of-Window comparison using SciPy CZT Signal Processing, resulted in an improvement in 5 BPM threshold percentage from 57.34% to 67.39%. Most PPG waveform predictions show that the model with a larger Chronos size learned physiological signals better.

Model Size Comparison: PPG Predictions, PURE 10-05

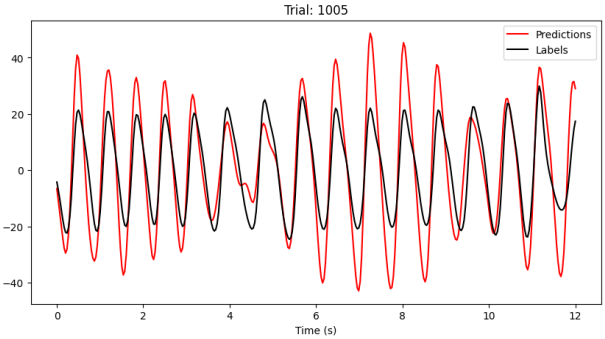

Figure 5.21: Spatiotemporal Model DINOv2-Small + Chronos-Tiny, PPG Prediction, PURE 10-05

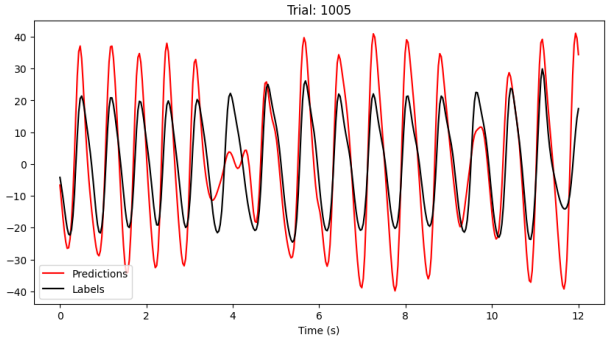

Figure 5.22: Spatiotemporal Model DINOv2-Small + Chronos-Base, PPG Prediction, PURE 10-05

Model Size Comparison: PPG Predictions, PURE 02-04

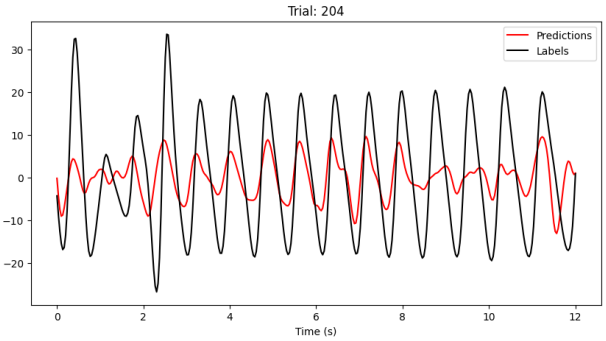

Figure 5.23: Spatiotemporal Model DINOv2-Small + Chronos-Tiny, PPG Prediction, PURE 02-04

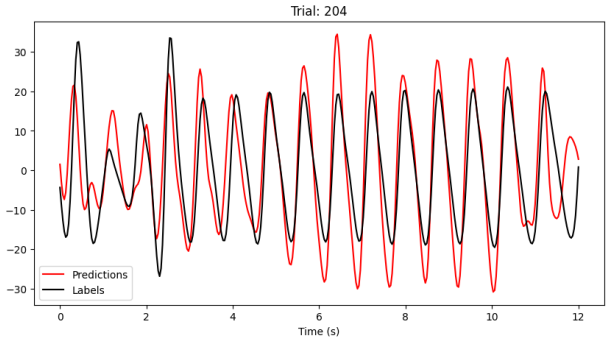

Figure 5.24: Spatiotemporal Model DINOv2-Small + Chronos-Base, PPG Prediction, PURE 02-04

Model Size Comparison: PPG Predictions, PURE 03-02

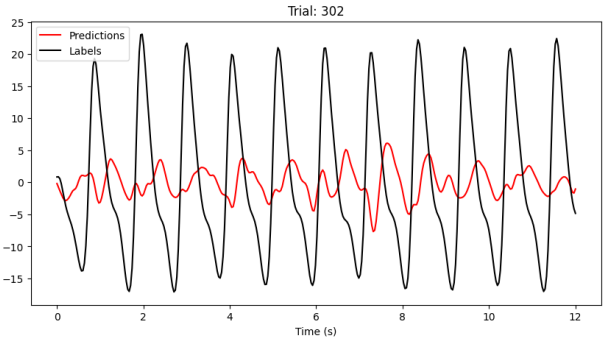

Figure 5.25: Spatiotemporal Model DINOv2-Small + Chronos-Tiny, PPG Prediction, PURE 03-02

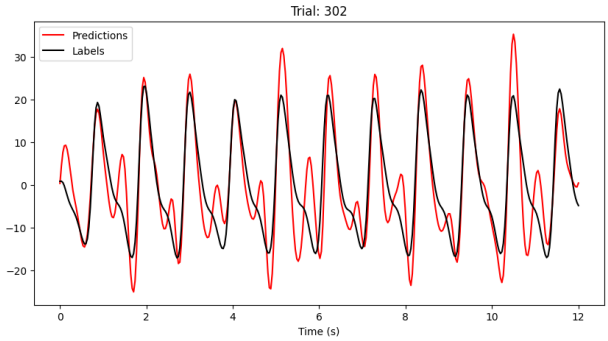

Figure 5.26: Spatiotemporal Model DINOv2-Small + Chronos-Base, PPG Prediction, PURE 03-02

Model Size Comparison: PPG Predictions, PURE 08-03

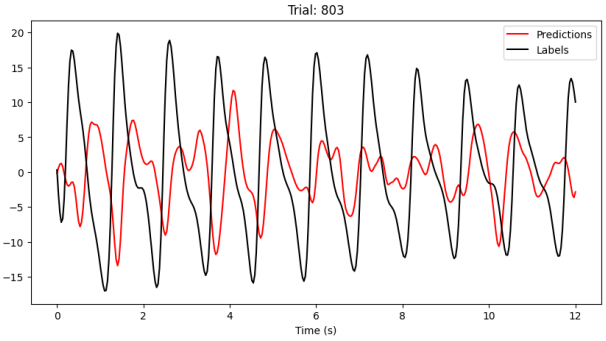

Figure 5.27: Spatiotemporal Model DINOv2-Small + Chronos-Tiny, PPG Prediction, PURE 08-03

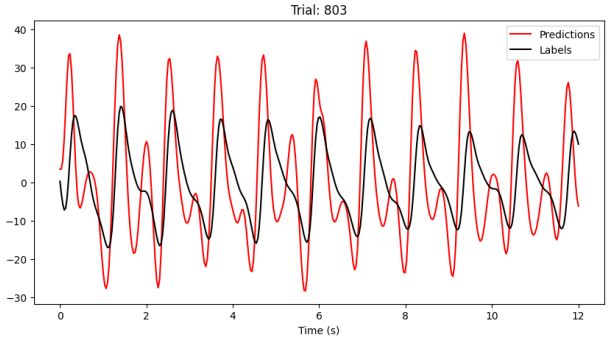

Figure 5.28: Spatiotemporal Model DINOv2-Small + Chronos-Base, PPG Prediction, PURE 08-03

5.5 Finetuning

One possibility to explore is finetuning DINOv2 and/or Chronos to further optimize for rPPG applications. Initial finetuning experiments showed that the finetuned DINOv2-Small + Chronos-Base spatiotemporal model learned PPG waveforms well, but sometimes predicted the diastolic peaks to be more prominent with reference to the systolic peaks compared to the frozen DINOv2-Small + Chronos-Base spatiotemporal model. Due to signal processing and harmonics, this caused the reported accuracy to be worse; with further filtering, however, there is space for improvement.

Signal Processing	MAE	MedAE	RMSE	5 BPM
Frozen	13.37	2.26	28.00	67.39%
Finetune	19.97	3.97	33.76	53.48%

Table 5.8: DINOv2-Small + Chronos-Base; Sensor End-of-Window Comparison; UBFC-rPPG/PURE. Frozen/Finetune Comparison. Evaluated with CZT.

Train Set: UBFC-rPPG; Test Set: PURE
Model: DINOv2-Small + Chronos-Base

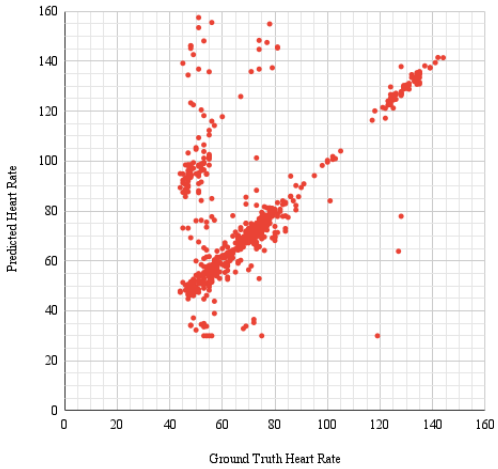

Figure 5.29: Ground Truth vs Predicted Heart Rate Values, PURE Dataset, Frozen (CZT)

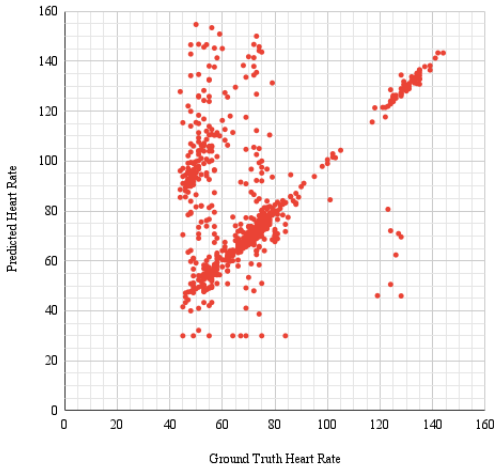

Figure 5.30: Ground Truth vs Predicted Heart Rate Values, PURE Dataset, Finetune (CZT)

It can be observed that there are more errors along the $y = 2x$ line in the results of the finetuned model, possibly corresponding to harmonics.

Frozen vs Finetune Comparison: PPG Predictions, PURE 03-02

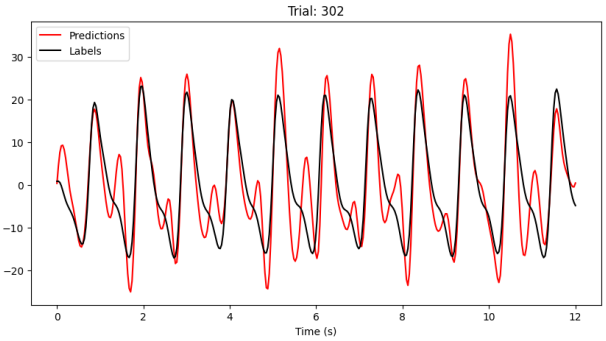

Figure 5.31: Spatiotemporal Model DINOv2-Small + Chronos-Base, Frozen, PPG Prediction, PURE 03-02

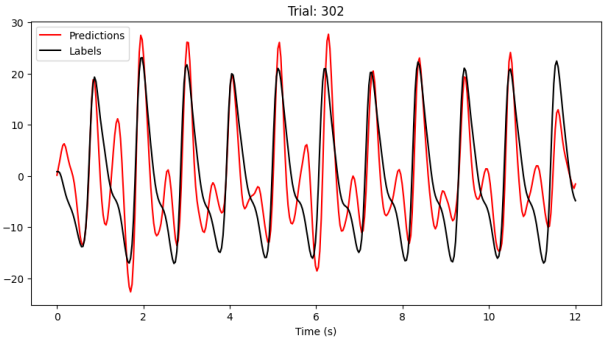

Figure 5.32: Spatiotemporal Model DINOv2-Small + Chronos-Base, Finetune, PPG Prediction, PURE 03-02

Frozen vs Finetune Comparison: PPG Predictions, PURE 08-03

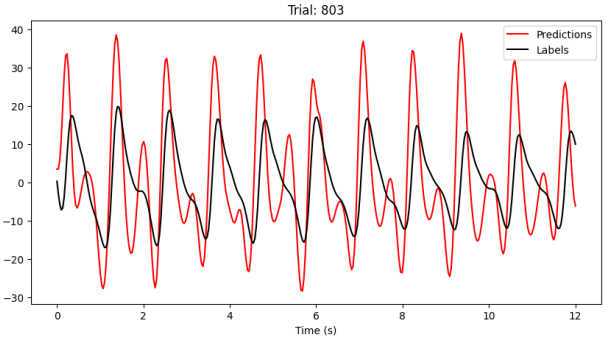

Figure 5.33: Spatiotemporal Model DINOv2-Small + Chronos-Base, Frozen, PPG Prediction, PURE 08-03

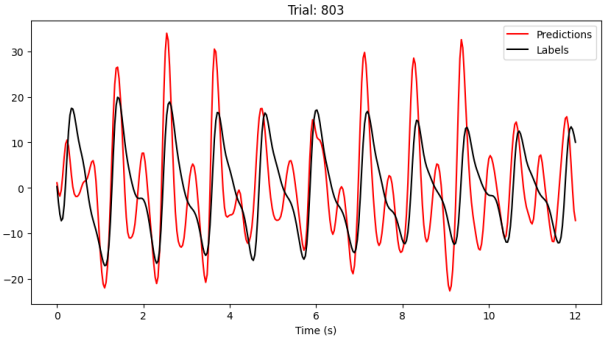

Figure 5.34: Spatiotemporal Model DINOv2-Small + Chronos-Base, Finetune, PPG Prediction, PURE 08-03

5.6 Continuous Monitoring and Priors

In practice, heart rate values do not typically fluctuate very rapidly. For instance, it is physiologically unrealistic for a person's heart rate to jump between 80 and 160 in a matter of seconds. With this knowledge, it is possible to improve rPPG techniques in continuous monitoring situations, where a subject's health parameters are taken persistently. This uses prior knowledge.

Current rPPG pipelines that rely on signal processing may suffer from falsely predicting sudden large jumps, as seen in examples from Ground Truth Discrepancies, where signal processing on the label signal can be unstable.

A simple method to prevent such leaps could be a bound that limits the difference of heart rate between adjacent windows by a physiologically reasonable limit. Two other methods are explored: frequency tracking and the sequential probability ratio test. Although these ideas were not fully incorporated into evaluation, they act as plausible paths to explore for future work and real-life applications.

Chapter 6

Conclusion

A major takeaway from the spatiotemporal architecture research and evaluations is that integrating two off-the-shelf foundation models, without elaborate task-specific conditioning, generates practically feasible performance for rPPG applications. This finding suggests that representations required for physiological inference are already embedded within visual and temporal foundation models. Engineering and design of architectures plays a key role in leveraging representations to rPPG applications.

In addition to model design, the evaluations involving individual components within the entire rPPG pipeline motivate further discussion about considerations for the end-to-end rPPG system. Factors such as signal processing have a large impact on the predicted heart rate values. As well, reliable and accurate ground truth values are necessary for fair and robust evaluation of rPPG systems, since inaccuracies can lead to less meaningful performance comparisons. Progress in rPPG research includes many directions, not only model architecture, but also robust data collection and organization, fair evaluation metrics, and holistic design of the end-to-end pipeline.

6.1 Limitations

Model Size

The integration of multiple foundation models results in a relatively large overall architecture, which increases memory required and may limit scalability. This might cause challenges on real-world deployment, especially in low-resource environments, where there are constraints on the size of the model and energy usage.

Training and Inference Speed

The large size of the models causes the training time to be long and inference time to be slow. Training with large datasets such as SCAMPS could take days. Inference was relatively longer as well. This can pose challenges for real-time applications and large-scale evaluation, and potentially with longer videos. At present, the architecture may not be suited for applications where fast evaluation is necessary.

Fixed 6-second Windows

The use of fixed 6-second windows can restrict the amount of video available to the model. Optimally, the length of a video would be a multiple of 6. Excluding some portion of the video, in this case at the end, may lead to evaluations that do not cover the dataset fully.

6.2 Future Work

The research presented in this work motivates many directons for future research.

Direct Heart Rate Prediction

Current rPPG systems involve the pipeline of explicitly predicting and constructing a PPG waveform. As shown through experiments, this can cause issues with signal processing. It may be valuable to research the possibility of a model that does not need the waveform prediction step, and whether this step is necessary for performance or primarily as an intermediate constraint. In the case that the scalar heart rate value prediction is the goal of the model, eliminating the need for the waveform step and signal processing, and directly predicting heart rate, could lead to simpler and more efficient models.

Environmental Conditions Decision Mechanisms

The incorporation of a decision mechanism, such as a decsion tree or classifier informed by environmental conditions, could allow the pipeline to adapt the processing strategy dynamically. This can improve robustness under varying conditions.

Preprocessing and Segmentation

All of the experiments with the spatiotemporal architecture were done with simple, raw conditions, without additional processing such as segmentation or dynamic detection and tracking. This already generates feasible performance, however, incorporating region-based facial segmentation, such as focusing on the forehead and cheeks, could improve the signal quality further by emphasizing the regions that show stronger changes associated with BVP. Such processing steps could reduce noise from less informative regions. Attention mechanisms can also allow the model to focus on spatial regions or temporal segments that are more informative for rPPG estimation.

Motion Artifacts

Explicitly modeling motion through optical flow or other representations could distinguish physiological signals from noise due to motion. This is particularly beneficial with significant movement. Enabling dynamic tracking, involving real-time monitoring and adapting to changes, can lead to stronger relationships between patches and pixels over time.

Training with Multiple Datasets

Throughout training and evaluation, the focus has been training on one dataset and testing on another. Joint training across multiple datasets may improve generalization, as the model is

exposed to a wider range of recording conditions, subjects, and environments. This can help reduce dataset-specific biases and improve robustness.

Respiratory Rate for Real-World Datasets

Initial experiments with the spatiotemporal architecture on the SCAMPS dataset suggested an ability to learn respiratory rate signals in addition to heart rate signals. Further validating this on real-world breathing rate datasets would strengthen the claim, and overall broaden the applicability of the spatiotemporal model for multiple parameters. In the real world, respiratory signals may be weaker and more susceptible to noise.

Signal Processing Specific to BVP

As models depend on accurate heart rate extraction from a PPG waveform, developing and improving signal processing methods to be specific to BVP waveforms can result in increased accuracy in heart rate predictions. Such research directions could include filtering, frequency estimation, or possibly a supervised signal processing method trained on BVP waveforms and heart rate values.

Modular Changes

The modular design of the architecture allows ease of replacement of visual and temporal foundation models, as well as prediction head changes. Exploring alternative models, including ones that may be more specialized, can yield further performance gains or efficiency improvements.

PPG Ground Truth Validation and Frame Rate Normalization

Improved validation of ground truth rPPG signals are beneficial to reducing label noise and allow for better training of supervised models as well as more reliable evaluation. Frame rate normalization could further mitigate errors that arise from variable or inconsistent video sampling rates.

6.3 Ethics Statement

This work relies on datasets that contains videos of participant faces and associated physiological signals. Datasets were publicly available, collected under protocols that obtained consent from subjects and adhere to guidelines.

Bibliography

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR. 1.1, 3.2

[2] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 2017. (document), 1.1, 4.1, 4.6, 5.1.3

[3] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int. J. Biosens. Bioelectron.*, 4(4):195–202, aug 2018. 1.1

[4] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks, 2018. URL https://arxiv.org/abs/1805.07888. 2.2, 4.6

[5] Joaquim Comas, Adria Ruiz, and Federico Sukno. Deep adaptative spectral zoom for improved remote heart rate estimation, 2024. URL https://arxiv.org/abs/2403.06902. 2.5

[6] G de Haan and A van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological Measurement*, 35(9):1913, aug 2014. doi: 10.1088/0967-3334/35/9/1913. URL https://doi.org/10.1088/0967-3334/35/9/1913. 2.1, 4.6

[7] Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. doi: 10.1109/TBME.2013.2266196. 2.1, 4.6

[8] Defense Advanced Research Projects Agency (DARPA). Darpa triage challenge. https://www.darpa.mil/research/challenges/darpa-triage-challenge, 2025. 1.2, 4.2.2

[9] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement, 2021. URL https://arxiv.org/abs/2006.03790. 2.2, 4.6

[10] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe

Zhang, Yuntao Wang, Soumyadip Sengupta, Shwetak Patel, and Daniel McDuff. rPPG-Toolbox: Deep Remote PPG Toolbox. *arXiv preprint arXiv:2210.00716*, 2022. (document), 4.6

[11] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4997–5006, 2023. doi: 10.1109/WACV56688.2023.00498. 2.2, 4.6

[12] Huiwen Loh, Shuting Xu, Oliver Faust, Chui Ooi, Prabal Datta Barua, Subrata Chakraborty, Ru San Tan, Filippo Molinari, and U Acharya. Application of photoplethysmography signals for healthcare systems: An in-depth review. *Computer Methods and Programs in Biomedicine*, 216:106677, 02 2022. doi: 10.1016/j.cmpb.2022.106677. 1.1

[13] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9):176:1–176:40, January 2023. doi: 10.1145/3558518. 1.1

[14] Daniel McDuff, Miah Wander, Xin Liu, Brian L Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. SCAMPS: Synthetics for Camera Measurement of Physiological Signals. *arXiv preprint arXiv:2206.04197*, 2022. (document), 4.1, 4.1, 5.1.3

[15] James Mcnames, Cristina Crespo, Mateo Aboy, J. Bassale, L. Jenkins, and Brahm Goldstein. Harmonic spectrogram for the analysis of semi-periodic physiologic signals. volume 1, pages 143 – 144 vol.1, 02 2002. ISBN 0-7803-7612-9. doi: 10.1109/IEMBS.2002.1134427. 2.5

[16] Rita Meziati, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, and Fan Yang. Ubfc-phys, 2021. URL `https://dx.doi.org/10.21227/5da0-7344`. (document), 4.1, 4.6, 5.1.3

[17] Monolithic Power Systems, Inc. Harmonics in ac power systems. `https://www.monolithicpower.com/en/about-mps.html`. (document), 5.2

[18] Andreia Vieira Moço, Sander Stuijk, and Gerard de Haan. Ballistocardiographic artifacts in ppg imaging. *IEEE Transactions on Biomedical Engineering*, 63(9):1804–1811, 2016. doi: 10.1109/TBME.2015.2502398. 1.1

[19] NTi Audio. Fast fourier transformation (fft) — basics. `https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft`. (document), 3.2

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL `https://arxiv.org/abs/2304.07193`. 1.1, 3.1

[21] Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open Foundation Models for Optical Physiological Signals. In *The Thir-*

teenth International Conference on Learning Representations, ICLR 2025, Singapore, April 2025. URL https://arxiv.org/abs/2410.20542. Accepted. arXiv preprint arXiv:2410.20542. 2.4

[22] Christian S. Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1335–13358, 2018. doi: 10.1109/CVPRW.2018.00172. 2.1, 4.6

[23] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011. doi: 10.1109/TBME.2010.2086456. 2.1, 4.6

[24] Ali S. Salim and Abdul Sattar M. Khidhir. A comprehensive review of rppg methods for heart rate estimation. *Open Access Library Journal*, 11(11):1–20, 2024. doi: 10.4236/oalib.1112482. 1.1

[25] R. Stricker, S. Müller, and H.-M. Gross. Non-contact Video-based Pulse Rate Measurement on a Mobile Service Robot. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1056–1062, Edinburgh, Scotland, UK, 2014. IEEE. (document), 1.1, 4.1, 4.6, 5.1.3

[26] Mohd Zubir Suboh, Rosmina Jaafar, Nazrul Anuar Nayan, Noor Hasmiza Harun, and Mohd Shawal Faizal Mohamad. Analysis on four derivative waveforms of photoplethysmogram (ppg) for fiducial point detection. *Frontiers in Public Health*, Volume 10 - 2022, 2022. ISSN 2296-2565. doi: 10.3389/fpubh.2022.920946. URL https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.920946. (document), 5.1

[27] Zhaodong Sun and Xiaobai Li. *Contrast-Phys: Unsupervised Video-Based Remote Physiological Measurement via Spatiotemporal Contrast*, page 492–510. Springer Nature Switzerland, 2022. ISBN 9783031197758. doi: 10.1007/978-3-031-19775-8_29. URL http://dx.doi.org/10.1007/978-3-031-19775-8_29. 2.3

[28] Amir Tahernejad, Ahmad Sahebi, Amirhossein Soleimani Shokouh Abadi, and Mohammad Safari. Application of artificial intelligence in triage in emergencies and disasters: a systematic review. *BMC Public Health*, 24(1):3203, 2024. doi: 10.1186/s12889-024-20447-3. 1.1

[29] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: Multi-domain mobile video physiology dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5, 2023. doi: 10.1109/EMBC40787.2023.10340857. (document), 4.1, 4.6

[30] Atsushi Tohma, Makoto Nishikawa, Takayuki Hashimoto, Yuuki Yamazaki, and Guoxiang Sun. Evaluation of Remote Photoplethysmography Measurement Conditions toward Telemedicine Applications. *Sensors (Basel)*, 21(24):8357, Dec 2021. ISSN 1424-8220. doi: 10.3390/s21248357. 1.1

[31] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, Dec 2008. doi:

10.1364/OE.16.021434. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-16-26-21434. 4.6

[32] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. doi: 10.1109/TBME.2016.2609282. 2.1, 4.6

[33] Yiping Xie, Bo Zhao, Mingtong Dai, Jian-Ping Zhou, Yue Sun, Tao Tan, Weicheng Xie, Linlin Shen, and Zitong Yu. PhysLLM: Harnessing Large Language Models for Cross-Modal Remote Physiological Sensing, 2025. URL https://arxiv.org/abs/2505.03621. 2.4

[34] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 2.2, 4.6

[35] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer, 2022. URL https://arxiv.org/abs/2111.12082. 2.2, 4.6