ST2S-rPPG: A Spatiotemporal Two-Stage Learning Approach for Pulse Estimation Using Video

Eirini Kateri E.Kateri@soton.ac.uk
Katayoun Farrahi K.Farrahi@soton.ac.uk
Electronics and Computer Science, University of Southampton, Southampton, UK

Abstract

Remote physiological monitoring presents an opportunity to enhance patient care, particularly in scenarios where traditional monitoring methods are impractical or unavailable. Heart rate, being a principal indicator of health, has been a focal point of video-based monitoring systems. Despite significant advancements in remote photoplethysmography technology, several challenges persist, including motion artifacts, data homogeneity and availability, which impact the accuracy and reliability of such solutions. In this work, we introduce a novel framework aimed at addressing these challenges, ST2S-rPPG. Our methodology involves a stabilization method to mitigate motion artifacts. We propose a spatiotemporal representation of video data, which captures predictive available information in the video and assists in transforming the input video. We present a unique approach to ground truth representation for capturing more informative features. Finally, we incorporate a two-stage learning component into our framework to optimize estimation accuracy. Through evaluations on benchmark datasets, we demonstrate the effectiveness of our contributions and their practical relevance in healthcare applications.

Keywords: two-stage learning, remote photoplethysmography, rPPG, vital sign estimation, spatiotemporal methods

Data and Code Availability We evaluate the performance of our proposed framework on two benchmark datasets, MMSE-HR Zhang et al. (2016) and UBFC-rPPG Sabour et al. (2021). Both datasets can be made available after contacting the authors cited. Our code is available at: https://github.com/eirkateri/ST2S-RPPG.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

In an era characterised by the integration of technology into our daily lives, the intersection of computer vision and remote physiological monitoring, has captured research's interest by promising to revolutionize the way we view human health in both clinical and non-clinical settings. Remote physiological monitoring can have a positive effect on patient care, especially in scenarios where traditional monitoring methods are impractical or unavailable, thus can lead to delays in diagnosis and sub-optimal management of health conditions. The increased risk of infection for healthcare workers and patients, the fragile skin of newborn infants or the elderly, the importance of continuous surveillance for chronic disease patients, individuals in inaccessible locations, lack of mobility, staff shortage, or financial constraints are some of healthcare's current barriers that could benefit from remote monitoring solutions. Such approaches not only enhance patient comfort and convenience, but also have the potential to improve clinical outcomes by allowing a continuous monitoring of vital signs, facilitating early detection of abnormalities. Heart rate in particular, being one of the principal indicators of health problems, has been the main focus of video-based monitoring systems since 2008 with Verkruysse et al. (2008)'s work, using either signal processing [Poh et al. (2010a,b)] or deep learning approaches [Chen and McDuff (2018); Liu et al. (2020)].

While remote photoplethysmography (rPPG) technology has made significant advancements in recent decades, there are still several factors that limit its effectiveness. Motion artifacts, illumination changes and the lack of data diversity are among the key barriers that have an effect on accuracy and reliability of such solutions. Researchers have attempted to mitigate the effects of these factors in heart rate estimation [Li et al. (2014); Feng et al. (2014); Wang et al. (2014)], however this still remains an open problem.

Such limitations significantly affect the adoption of these technologies in clinical settings and raise concerns about their impact on patient care.

Addressing these challenges requires innovative approaches that leverage the capabilities of computer vision and machine learning. Researchers also need to consider the significance and availability of existing datasets, specifically their lack of accurate representation of real life scenarios and the natural diversity in humans.

The proposed paper introduces several novel contributions that address some of these problems and hold significance for machine learning applications in healthcare. By leveraging a spatiotemporal representation of data, we offer a unique approach to transforming the input video in a spatiotemporal manner. This methodology could be particularly valuable in healthcare settings where data availability and resources are often limited, allowing for more robust training of machine learning algorithms. The proposed stabilization method addresses the challenge of motion artifacts in videos, a common issue in healthcare related video data, improving the accuracy and reliability of machine learning models applied to such data. Lastly, the incorporation of a two-stage learning component into the framework aims to optimize estimation accuracy by selecting the best-performing images for model training, a technique with broader applicability beyond pulse estimation, particularly in scenarios where selecting informative data samples is crucial for model performance. The two-stage learning approach can lead to recommendations that are more accurate and calibrated to each individual, ultimately improving patient outcomes. Subsequent sections of the paper will provide evidence of the effectiveness of these contributions through evaluations on benchmark datasets, demonstrating their practical relevance and impact.

Our contributions can be summarized as follows:

- We develop a **spatiotemporal representation** to transform input video into spatiotemporal images;
- We propose a **stabilization method** using the Persistent Independent Particles algorithm to combat motion artifacts;
- We design a **two stage learning framework** to optimise estimation accuracy by selecting the most informative spatiotemporal images

2. Related Work

Early work establishes the sufficiency of information in video recordings of a person's face under ambient light for pulse measurements [Verkruysse et al. (2008)]. This is the first step in modeling how light interacts with skin, using Blind Source Separation (BSS) techniques like Independent Component Analysis (ICA) [Poh et al. (2010b,a)] and CHROM [De Haan and Jeanne (2013), Wang et al. (2016)] to extract pulse measurements. In 2013, research begins to investigate the Newtonian reaction of the face to the movement of blood. Balakrishnan et al. (2013) suggests that displacement on the skin due to the cardiac cycle could be employed for pulse calculation. Other researchers propose that the choice of region of interest (ROI) greatly influences rPPG measurements, as the density distribution of blood vessels varies in different facial regions [Lempe et al. (2013); Li et al. (2014); Kumar et al. (2015)]. Some experiments include attempts at noise reduction with adaptive bandpass filters [Li et al. (2014); Feng et al. (2014); Wang et al. (2014)], other color frequencies (e.g., orange, cyan) [McDuff et al. (2014b,a)], and different BSS methods [Lewandowska et al. (2011); Kwon et al. (2012)]. These earlier studies help identify two main issues with rPPG: motion artifacts and signal strength variation caused by illumination.

Despite the progress made by previously presented methods, deep learning approaches are becoming increasingly popular, making significant advancements in the field of remote monitoring. DeepPhys [Chen and McDuff (2018)] is the first system using a deep attention CNN to calculate pulse and respiration rates. It is based on the skin reflection model and has an attention mechanism for robust estimation under lighting changes and motion artifacts. In 2020, Liu et al. (2020) base their work on DeepPhys, creating a multitask temporal-shift convolutional attention network (MTTS-CAN) for real-time heart and respiratory rate measurements. They employ temporal shift modules to remove noise, an attention mechanism to improve signal source separation, and a multitask mechanism to estimate pulse and respiration rates jointly. Even though such solutions moderately help with the motion and light artifact issues, they also introduce a number of constraints. Such deep complex models lack interpretability, making it challenging to understand how the models make decisions. This lack of transparency can be problematic, especially in healthcare applications where understanding the reasoning behind predictions is crucial for clinical decision-making. Other issues include the need for large labeled datasets, constraints in real-time processing capabilities and lack of accounting for inter-subject variability.

In contrast, hybrid methods have been developed that first extract handcrafted features followed by a deep learning network for pulse estimation. These methods take advantage of spatiotemporal maps which are subsequently fed through a machine or deep learning framework to extract the pulse rate. Qiu et al. (2018) employs the Eulerian Video Magnification approach to extract spatiotemporal features and Niu et al. (2018, 2019b,a) generates spatiotemporal maps by aggregating signal from multiple ROIs. Similarly Jaiswal and Meenpal (2022) generates spatiotemporal maps of particular ROIs and creates a compressed 2D representation of a video by decreasing spatial data redundancy while maintaining temporal dynamics of the original video. Finally, Shao et al. (2023) designs a spatiotemporal transformer module to extract physiological cues from facial regions and aggregate them. The use of spatiotemporal features as a means to estimate heart rate has many benefits compared to the traditional use of videos as highlighted by the research above. Such approaches can result in higher temporal resolution compared to video frames, providing more detailed information about the skin changes over time and can help mitigate the impact of motion and light artifacts by integrating information over time, leading to more robust pulse estimation. These works have demonstrated promising results, however there are still components that have not been addressed. All above works rely on pre-defined ROIs, which could neglect regions with sufficient signal to assist in the increased accuracy of pulse estimation. Additionally, averaging information from multiple frames in a single image result in significant signal variations being suppressed.

3. Methodology

The proposed ST2S-rPPG framework is divided into four steps, face identification and stabilization, spatiotemporal image generation, pulse estimation using a regression CNN and a second learning component to improve estimation. In the following sections, these steps are described in detail.

3.1. Face Identification and Tracking

In the context of estimating pulse signals from facial video data, the identification of ROIs is a fundamental step. However, the accurate tracking of these regions becomes challenging when confronted with video instability. To address this concern, we propose the application of the Persistent Independent Particles (PIPs) algorithm for video stabilization.

The primary objective of this proposed approach is to facilitate the tracking of the facial region within video sequences. Our inspiration is drawn from the work of [Sand and Teller (2008)] and the follow up work of [Harley et al. (2022)], who introduced a novel motion representation paradigm referred to as "particle video". The fundamental concept underlying particle video stems from particle filters and is the representation of a video as an ensemble of particles, each traversing multiple frames. The key advantage lies in the utilization of long-range temporal priors during the tracking of these particles.

Even within the same video, different segments may exhibit unique physiological patterns, hence we segment the original video V into discrete 10-second segments. Equation 1 represents each of the 10-second segments V_c. This spatiotemporal transformation enriches the existing dataset and reduces the computational complexity in later stages.

$$V_c = \lfloor \frac{V}{10} \rfloor \tag{1}$$

We isolate the first frame of each video and apply the Viola-Jones algorithm [Viola and Jones (2001)] to extract the precise facial location within the frame. We assume the dimensions of the Viola Jones bounding box are (h, w), where h is the box's height and w is its width in pixels. We then identify the box's central point (x_0, y_0):

$$(x_0, y_0) = (l + \frac{h}{2}, c + \frac{w}{2}) \tag{2}$$

where (l, c) denote the pixel coordinates of the top-right corner of the bounding box.

We employ the PIPs algorithm to track the trajectory represented by (x_0, y_0) coordinates in conjunction with the video input. Once the central point's trajectory is obtained for every frame, we proceed to extract a sub-frame from each original video frame using the following formula:

$$Ic_n(x_n, y_n, z) = Io_n(x_n - \frac{w}{2} : x_n + \frac{w}{2}, y_n - \frac{h}{2} : y_n + \frac{h}{2}, z) \tag{3}$$

where $Ic_n(x_n, y_n, z)$ represents the cropped frame at index n, Io_n represents the original frame, $(x_n,\ y_n)$ denote the x and y coordinates of the central point in frame n and (w, h, z) represent the width, height and color channel (RGB) dimensions of the desired crop in pixels. The resulting stabilized video consists of the concatenated frames with background removed.

3.2. Spatiotemporal Image Generation

Utilizing spatiotemporal images offers several advantages over analyzing a single continuous video stream. Firstly, it increases the dataset size, as each spatiotemporal image encapsulates a temporal sequence of a single facial region, since each facial region consists of slightly distinct features. This spatiotemporal transformation facilitates more robust training of machine learning models, enhancing their ability to distinguish subtle changes in pulse signals over time. Additionally, the process of stabilizing the images ensures consistent tracking of specific facial areas across the temporal sequence. By maintaining alignment between consecutive frames, the analysis remains focused on the same regions, enabling more precise examination of physiological variations.

In order to generate spatiotemporal images we employ a technique that involves the division of the stabilized videos into six equal vertical segments. Subsequently, the first and last segments are discarded to exclude any residual background or non-essential facial regions that may not have been adequately eliminated by the Viola-Jones algorithm. Then, we segment each remaining frame into L vertical segments of three pixels:

$$L = \lfloor \frac{w}{3} \rfloor \tag{4}$$

where w represents the width of each frame in pixels.

Heuristically we find that three pixel wide slices provide a balance between spatial resolution and computational efficiency. At the same time, facial features relevant to pulse estimation may exhibit variations on the order of a few pixels. By segmenting the frames into three-pixel-wide segments, we aim to capture these subtle variations more effectively.

We can represent each frame as:

$$Frame^{(k)} = \begin{bmatrix} S_{0,0}^{(k)} & S_{0,1}^{(k)} & \dots & S_{0,L-1}^{(k)} \\ S_{1,0}^{(k)} & S_{1,1}^{(k)} & \dots & S_{1,L-1}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ S_{i,0}^{(k)} & S_{i,1}^{(k)} & \dots & S_{i,L-1}^{(k)} \end{bmatrix} \tag{5}$$

for,

$$S_{i,n}^{(k)} = (P_{i,3j}^{(k)}, P_{i,3j+1}^{(k)}, P_{i,3j+2}^{(k)}) \tag{6}$$

where k is the k^{th} frame, (i,j) are the height and width of the frame in pixels respectively, P represents the pixel values and S each 3 pixel slice values. For clarity, j refers to the width of the frame in terms of pixel groups or slices and n is the time dimension or different frames in the sequence.

In order to construct the spatiotemporal images, we arrange the corresponding vertical segments from each frame sequentially, frame by frame. The m^{th} spatiotemporal image, generated by the m^{th} vertical slice is represented by:

$$ST_m = \begin{bmatrix} S_{0,m}^{(0)} & S_{0,m}^{(1)} & \dots & S_{0,m}^{(k)} \\ S_{1,m}^{(0)} & S_{1,m}^{(1)} & \dots & S_{1,m}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ S_{i,m}^{(0)} & S_{i,m}^{(1)} & \dots & S_{i,m}^{(k)} \end{bmatrix} \tag{7}$$

These resulting images provide a comprehensive representation of the video's content, with each image capturing a distinct set of three-pixel-wide segments spanning the entire duration of the video. It must be noted that the number of images per subject can vary, depending on their approximate location to the camera or their facial size.

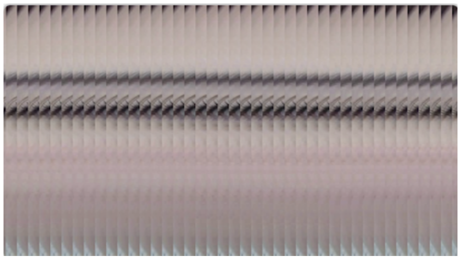

Figure 1: Example of a spatiotemporal image

3.3. Convolutional Neural Network

Let ST_m represent the input spatiotemporal image, with dimensions $w \times h \times z$, where w is the width, h

is the height, and z is the number of channels. The CNN is designed to process the input spatiotemporal image ST_m and predict the number of beats $\hat{y}$. The architecture consists of three convolutional layers, a flattening step, and fully connected layers, as summarized in Table 1. During the forward pass, the input ST_m is reshaped and passed through each layer of the CNN in sequence, with ReLU activation functions applied after each convolutional and fully connected layer. The optimizer used is the Adam optimizer with a learning rate of 0.001, and the loss function is the L1 loss (Mean Absolute Error) between the predicted number of beats $\hat{y}$ and the ground truth number of beats. A visual representation of the CNN can be found in Figure 7 in the Appendix.

Table 1: Parameters of the CNN Architecture

Layer	**Input**	**Parameters/Output**
Conv1	ST_m	Kernel: $K_1 = 3 \times 3$, Stride: $S_1 = 1$, Activation: ReLU, Output: O_1
Conv2	O_1	Kernel: $K_2 = 3 \times 3$, Stride: $S_2 = 3$, Activation: ReLU, Output: O_2
Conv3	O_2	Kernel: $K_3 = 3 \times 3$, Stride: $S_3 = 3$, Activation: ReLU, Output: O_3
Flatten	O_3	Flattened Output: F
Fully Connected	F	Units: $H_1 = 128$, Output: H_1
Output Layer	H_1	Predicted Beats: $\hat{y}$
Optimizer	Adam, Learning Rate: 0.001	
Loss Function	L1 Loss (Mean Absolute Error)	

3.4. Two-stage learning

It is apparent that not all spatiotemporal images exhibit similar performance, and certain regions within them may contain significant noise. Rather than making the assumption on which areas the CNN finds the most informative based on convention, we have implemented a second learning stage. Following the CNN's pulse prediction on individual images, we construct a new binary dataset. This dataset is formed based on the Mean Absolute Error (MAE) between the CNN's predictions and the ground truth on number of beats. Utilizing a predefined threshold, t=0.5, corresponding to a MAE of 3 beats per minute (bpm), we categorize the images into two classes according to whether their MAE surpasses or remains below the threshold. The 3 bpm criterion for categorizing images automatically distinguishes "good" images from "bad" ones. This threshold was chosen as it represents an acceptable margin of error for pulse estimation. A Multi-Layer Perceptron (MLP) is trained to classify the images in the custom binary dataset, ensuring that only the most informative "good" images are utilized for further analysis. This automated selection process eliminates the need for subjective assumptions about image quality, enhancing the robustness of the pipeline. A Multi-Layer Perceptron (MLP) comprising of 5 layers with 200 neurons per layer, is trained to classify the spatiotemporal images, using the custom "good" and "bad" image dataset as described above. A 10-fold cross-validation experiment is conducted, selecting images that the classifier categorizes as "good" 70% of the time. The evaluation metrics presented in the subsequent section are estimated based on the MLP's predictions for the "good image" class.

4. Results

4.1. Datasets and Ground Truth

We evaluate the performance of ST2S-rPPG on two benchmark datasets, the Multimodal Spontaneous Emotion database for heart rate estimation (MMSE-HR) and the Université Bourgogne Franche-Comté dataset for rPPG (UBFC-rPPG). These are datasets widely used in the field of rPPG and allow us to compare our methodology with state-of-the-art approaches. The two datasets have different baselines due to their different characteristics (frame rate, resolution, collection protocol). Our work uses subsets of the original datasets, specifically formatted for rPPG. Despite our continuous efforts to expand our study with additional datasets such as PURE and VIPL-HR — both of which are also commonly used in rPPG research — we were unable to secure access to these datasets. Despite this, the chosen MMSE-HR and UBFC-rPPG datasets provide sufficient variability to validate the generalizability of our approach.

MMSE-HR: [Zhang et al. (2016)] Comprising of 98 RGB videos and corresponding heart-rate data obtained from 40 participants, each video is recorded at a resolution of 1040x1392 pixels and a frame rate of 25 frames per second (fps). Video lengths are varying from 30 seconds to 1 minute. We extract 48,415 spatiotemporal images through the spatiotemporal image generation process.

UBFC-rPPG: [Sabour et al. (2021)] Comprising of 40 RGB videos and corresponding heart-rate data obtained from 40 participants. Each video provided is recorded at a resolution of 640x480 pixels and a frame rate of 30fps. Video duration varies from 46 seconds to 1 minute 8 seconds. Each video is synchronized with a pulse oximeter finger clip sensor to collect the

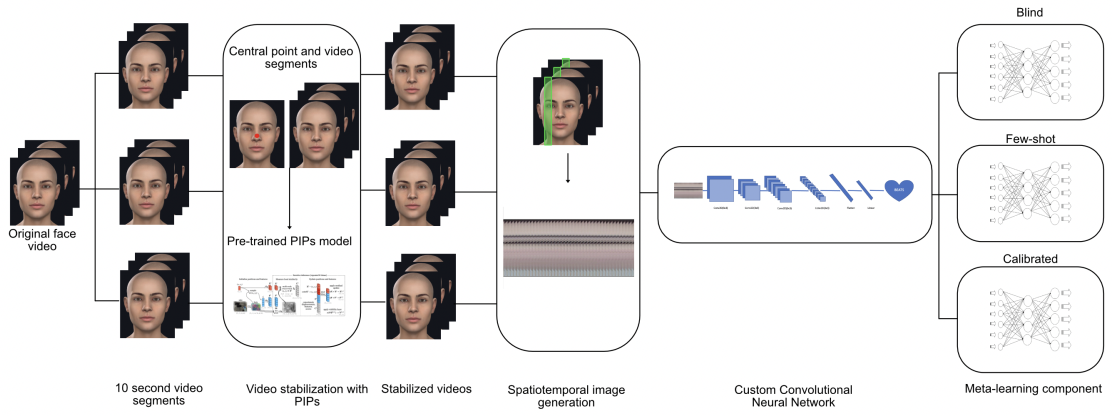

Figure 2: Overview of ST2S-rPPG for rPPG based remote HR measurement via spatiotemporal images. Key steps include video stabilization using PIPs to reduce motion artifacts, spatiotemporal image generation for capturing combined spatial and temporal features, and the second-stage learning process for automated selection of high-quality, informative images. Each stage is designed to optimize pulse estimation accuracy while minimizing computational complexity

ground truth. We extract 9,659 spatiotemporal images from the UBFC-rPPG database through the spatiotemporal image generation process. Ground truth measurements are provided for both datasets. We convert the measurements provided to beats per 10 second segment using the procedure below:

For the MMSE-HR dataset, ground truth heart rate data is acquired through a contact sensor operating at a sample rate of 1 KHz, providing pulse measurements per frame. In MMSE-HR, the definition of the heart rate ground truth data is that the measurement changes every time there is a heartbeat. We define each 10 second time segment as $[t_{start}, t_{end}]$, where t_{start} is the starting time of the segment and t_{end} the ending time of the segment. To identify the location of these segments we multiply the start and end time with the sampling rate. Within each segment we count the number of changes in the provided ground truth files, each change is a heartbeat.

For the UBFC-rPPG dataset, we use the raw signal data and the scipy library *find_peaks* to identify the beats. With the same process as above we count the number of peaks per segment. This process provides granular information regarding the pulse variability within each 10 second segment, which is not necessarily visible by using the average measurements for the whole video. The idea behind this choice is that the pulse estimation CNN will be able to distinguish beats easier than extrapolated bpm in each 10 second segment. After we compile our results we multiply the predicted value by 6 to extract the BPM measurement and compare performance with existing methodologies.

4.2. Evaluation Metrics

We evaluate the performance of ST2S-rPPG using five metric indicators commonly utilized to assess rPPG regression approaches, namely Mean Absolute Error (MAE), Mean Error (ME), Standard Deviation (SD), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) as defined in Section A.1 in the Appendix. It is worth noting that some entries in the comparison tables include missing values due to the absence of reported metrics in the referenced literature. Our work provides a comprehensive evaluation across all relevant metrics, ensuring a complete and consistent comparison.

4.3. Implementation Details

We implement ST2S-rPPG on the PyTorch framework and one NVIDIA GeForce GTX 1080 GPU. For both datasets, during the CNN prediction, we implement the Leave One Subject Out (LOSO) method. This is due to the fact that individual variability is significant and LOSO ensures that the model is trained on a wide variety of subjects and tested on a completely independent individual while preventing overfitting to individual characteristics. For the second-stage learning component, we conduct three distinct experiments: **Blind Scenario:** The classifier is withheld samples of the individual it is predicting on (i.e. LOSO), ensuring no data leakage. **Few-Shot Scenario:** The classifier is provided with data from all participants and only 6 random samples from the individual it is predicting on (3 per class). **Calibrated Scenario:** The classifier is trained using data from one individual. We ensure a separate training/test set within the individual's data to prevent data leakage; the training set is comprised of 80% of the balanced image dataset and the test set 20%.

4.4. Evaluation on MMSE-HR

We compare our proposed method with several state-of-the-art methods, ranging from approaches mitigating motion artifacts [Li et al. (2014), Tulyakov et al. (2016)] to other spatiotemporal approaches [Niu et al. (2019a), Jaiswal and Meenpal (2022)]. To ensure the validity of the comparison, we report on work that has been evaluated on the same dataset. All related results are presented in Table 2 and supporting visualisations can be found in section A.2. of the Appendix.

We decide to keep the calibrated results separate from the evaluation as our classifier is trained with personalised data and comparison would not be fair. Our MAE excluding the second-stage learning highlights the challenges of rPPG without the selection of informative data regions. Despite this, the standard deviation of the first-stage learning is favorable compared to literature, which indicates that our approach produces more consistent predictions with lower variability across individuals. We did not conduct experiments without the stabilisation, as the facial regions would not be consistently tracked, which would make it impossible to generate the images required for our method. We can observe that ST2S-rPPG achieves promising results on most commonly used metrics. Specifically, both blind and few-shot two-stage learning approaches achieve the best results in $HR_{MAE}, HR_{ME}, HR_{SD}$. The few shot two-stage learning also achieves second best performance in HR_{RMSE}, HR_{MAPE}. We demonstrate the most significant improvement compared to Li et al. (2014) and Tulyakov et al. (2016). These methods do not use spatiotemporal representations, further proving their efficiency. They also use adaptive band-pass filters for noise reduction, proving our stabilization method's capability. Our advantage over Niu et al. (2019a,b) is that instead of using an aggregate signal from multiple ROIs, we take advantage of all regions of the face, do not aggregate the spatiotemporal signal and do not choose the optimal images (ROIs) empirically. Finally, compared to all spatiotemporal approaches in Table 2, ST2S-rPPG does not perform any RGB transformations, since the lighting in the MMSE-HR database is not heterogeneous.

It's important to mention that while HR_{ME} is provided by some papers, its informativeness can be misleading, as the absolute value of error in not used. In our analysis, we prioritize HR_{MAE} as the most informative metric, as it directly measures the average error without biasing negative or positive deviations. Compared to Jaiswal and Meenpal (2022), the HR_{MAE} error was reduced by 13.64%. At the same time we have achieved the lowest standard deviation, suggesting more consistent predictions across individuals. Our calibrated two-stage learning experiment achieves the best performance across all metrics, keeping in mind that the classifier is trained with personalised data.

4.5. Evaluation on UBFC-rPPG

We compare ST2S-rPPG to several state-of-the-art approaches that have been evaluated on the UBFC-rPPG database. Supporting visualisations can be found in section A.2. of the Appendix.

In Table 3, we observe a similar trend with Table 2 regarding our results without the second-stage learning. Our ST2S-rPPG blind and few-shot method achieves the best results most reported metrics ($HR_{ME}, HR_{SD}, HR_{RMSE}, HR_{MAPE}$) and the second best results in HR_{MAE}, demonstrating its efficiency in accurately estimating heart rate even with limited training data. Additionally, ST2S-rPPG exhibits improvements in HR_{SD}, indicating more precise predictions and reduced variability in heart rate estimations. Similarly to the previous database, we demonstrate the most significant improvement of performance against non spatiotemporal traditional approaches [Poh et al. (2010a); Wang et al. (2016)] and 3D CNN approaches [Bousefsaf et al. (2019)]. TransPhys [Shao et al. (2023)] seems to be performing best in the HR_{MAE} metric, suggesting that spatiotemporal transformers show promising results, but can be computationally expensive. Finally, Meta-

Table 2: A summary of average HR estimation per video for ST2S-rPPG on the MMSE-HR dataset. Bold numbers indicate best performance and underlined numbers indicate second best performance.

Method	HR_{MAE}	HR_{ME}	HR_{SD}	HR_{RMSE}	HR_{MAPE}
Li et al. (2014)	-	11.56	20.02	19.95	14.64%
SAMC [Tulyakov et al. (2016)]	-	7.61	12.24	11.37	10.84%
RythmNet [Niu et al. (2019a)]	-	-0.85	4.99	**5.03**	**3.67%**
Niu et al. (2019b)	-	-3.10	9.66	10.10	6.61%
Jaiswal and Meenpal (2022)	6.4	-	6.63	6.82	-
ST2S-rPPG - No second-stage learning	10.21	1.59	5.58	11.75	14.94%
ST2S-rPPG - Blind (ours)	5.94	0.65	4.78	7.67	7.66%
ST2S-rPPG - Few-shot (ours)	**5.13**	**-0.39**	**4.11**	6.57	6.16%
ST2S-rPPG - Calibrated (ours)	*2.06*	*-0.23*	*2.35*	*3.11*	*2.88%*

Table 3: A summary of average HR estimation per video for ST2S-rPPG on the UBFC-rPPG dataset. Bold numbers indicate best performance and underlined numbers indicate second best performance

Method	HR_{MAE}	HR_{ME}	HR_{SD}	HR_{RMSE}	HR_{MAPE}
ICA[Poh et al. (2010a)]	8.43	-	18.6	18.8	-
CHROM [Wang et al. (2016)]	10.6	6.78	19.1	20.3	-
3D CNN [Bousefsaf et al. (2019)]	5.45	-1.31	8.55	8.64	-
Meta-rPPG [Lee et al. (2020)]	5.97	-	7.12	7.42	-
TransPhys [Shao et al. (2023)]	**4.66**	-	7.22	7.36	-
ST2S-rPPG - No second-stage learning	8.51	-1.93	4.75	9.84	8.25%
ST2S-rPPG - Blind (ours)	5.62	0.04	4.76	7.24	5.6%
ST2S-rPPG - Few-shot (ours)	5.24	**-0.02**	**3.81**	**6.36**	**5.21%**
ST2S-rPPG - Calibrated (ours)	*3.05*	*-1.04*	*2.82*	*3.98*	*2.95%*

rPPG [Lee et al. (2020)], showcases slightly lower estimation accuracy, potentially indicating that a second stage learning component, trained on predictions captures more valuable information for estimation.

5. Discussion

Telehealth is a field continuously growing in popularity, especially in the wake of the Covid-19 pandemic. Despite the increased adoption and improved accuracy of telehealth systems, there is a preference for complex deep learning methods, which might distance themselves from intuitive human understanding. This research seeks to address this issue by demonstrating the efficiency of spatiotemporal methods, which offer a more intuitive and comprehensive approach compared to the increasingly complex end-to-end deep learning methods. These spatiotemporal approaches aim to make the process more accessible and understandable, as they can be visualized (as in Figure 1) and further analyzed for signal vs. noise (a topic for future work). However, current spatiotemporal approaches often rely on pre-defined ROIs, which could neglect regions containing sufficient signal that could enhance pulse estimation accuracy. Additionally, aggregating multiple frames into a single image can suppress significant signal variations.

We addresses these limitations through several key contributions. By stabilizing original videos using the PIPs algorithm [Harley et al. (2022)], we mitigate the effects of natural motion artifacts, enabling the generation of more consistent spatiotemporal images. These images allow us to capture both spatial and temporal information within video data, enriching datasets (especially those with limited samples). Compared to complex deep learning models requiring continuous video processing, our methodology achieves significant computational efficiency by using spatiotemporal images and a simpler CNN architecture while maintaining high accuracy. This balance makes it particularly suitable for real-time healthcare applications. Finally, the integration of a second-stage learning framework marks a significant advancement by automating the selection of ROIs rather than relying on predefined assumptions. This approach enhances overall system efficiency by focusing on the most informative data, a common practice in real-world applications, particularly in clinical settings where minimizing noise is essential for accurate and reliable results. The calibrated two-stage learning method further enhances performance in both research and real-world settings. For instance, in medical consultations, collecting initial calibration data during in-person visits can improve model accuracy by accounting for individual variability in physiology and behavior. This personalized approach leads to more tailored recommendations, ultimately improving patient outcomes.

The findings from this study contribute to the broader telehealth landscape by demonstrating how spatiotemporal approaches can balance efficiency and accuracy. Compared to traditional end-to-end deep learning methods, this framework offers a more interpretable solution, which is particularly advantageous in clinical settings where understanding the rationale behind predictions is crucial. Integrating clinical feedback to validate and refine these methods further would strengthen their translational potential. Our approach is particularly relevant in scenarios where traditional monitoring tools, such as wearable devices, are impractical or inaccessible. This includes settings like resource-limited healthcare environments, monitoring fragile populations (e.g., newborns or elderly patients), and scenarios requiring non-intrusive continuous surveillance. The intuitive and computationally efficient design of our framework facilitates its potential integration into telehealth applications, reducing the dependency on contact-based systems and improving patient comfort and safety.

Limitations and Future Work

While the contributions of this work are significant, we acknowledge several limitations that highlight opportunities for future research. A substantial gap persists between research advancements and practical implementation in real-world scenarios.

One of the most pressing challenges is the significant lack of diverse and publicly available datasets. Current datasets fail to represent the wide spectrum of human characteristics, including variations in skin tone, ethnicity, gender, age, general appearance, and natural fluctuations in heart rate and medical conditions. Moreover, they lack environmental diversity, such as variations in motion and lighting, making current approaches impractical for real-world applications. Addressing these gaps requires collaborative efforts to access underrepresented populations and leveraging synthetic data generation to simulate more diverse scenarios.

While our proposed framework demonstrates robust performance across datasets with varying resolutions, its utility in scenarios with low-quality or non-standardized video data remains uncertain. We recognize that no rPPG method can reliably extract pulse signals from extremely low-quality data. We will explore techniques to mitigate this limitation, such as adaptive preprocessing methods or enhanced spatiotemporal transformations that emphasize robustness over fine detail resolution.

Another important consideration is the ethical implications of video-based monitoring, particularly regarding privacy and consent. Future research should prioritize strategies to address these challenges, including anonymization techniques and clear protocols for obtaining informed consent, ensuring compliance with ethical standards in healthcare applications.

Future work will focus on understanding the “bad” classifications in greater depth. Preliminary observations suggest that these include frames with poor alignment, occlusions, or regions like the eyes, which contribute less to accurate pulse estimation. An in-depth analysis of these representations could inform the design of adaptive and dynamic ROI selection mechanisms or enhanced preprocessing strategies.

Despite these obstacles, incremental progress continues to bring research closer to real-world applications. By addressing the challenges outlined above, future work can help bridge the gap between research findings and practical implementation, ultimately contributing to the advancement of telehealth technologies.

References

Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3430–3437, 2013.

Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9(20):4364, 2019.

Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018.

Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013.

Litong Feng, Lai-Man Po, Xuyuan Xu, Yuming Li, and Ruiyi Ma. Motion-resistant remote imaging photoplethysmography based on the optical properties of skin. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):879–891, 2014.

Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022.

Kokila Bharti Jaiswal and Toshanlal Meenpal. Heart rate estimation network from facial videos using spatiotemporal feature image. *Computers in Biology and Medicine*, 151:106307, 2022.

Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015.

Sungjun Kwon, Hyunseok Kim, and Kwang Suk Park. Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone. In *2012 annual international conference of the IEEE engineering in medicine and biology society*, pages 2174–2177. IEEE, 2012.

Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 392–409. Springer, 2020.

Georg Lempe, Sebastian Zaunseder, Tom Wirthgen, Stephan Zipser, and Hagen Malberg. Roi selection for remote photoplethysmography. In *Bildverarbeitung für die Medizin 2013: Algorithmen-Systeme-Anwendungen. Proceedings des Workshops vom 3. bis 5. März 2013 in Heidelberg*, pages 99–103. Springer, 2013.

Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011.

Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014.

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering*, 61 (10):2593–2601, 2014a.

Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera. *IEEE Transactions on Biomedical Engineering*, 61 (12):2948–2954, 2014b.

Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *2018 24th international conference on pattern recognition (ICPR)*, pages 3580–3585. IEEE, 2018.

Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019a.

Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019b.

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, 2010a.

Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010b.

Ying Qiu, Yang Liu, Juan Arteaga-Falconi, Haiwei Dong, and Abdulmotaleb El Saddik. Evm-cnn: Real-time contactless heart rate estimation from facial video. *IEEE transactions on multimedia*, 21 (7):1778–1787, 2018.

Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14(1):622–636, 2021.

Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *International journal of computer vision*, 80:72–91, 2008.

Hang Shao, Lei Luo, Jianjun Qian, Shuo Chen, Chuanfei Hu, and Jian Yang. Tranphys: Spatiotemporal masked transformer steered remote photoplethysmography estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016.

Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

Wenjin Wang, Sander Stuijk, and Gerard De Haan. Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE transactions on Biomedical Engineering*, 62(2):415–425, 2014.

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, 2016.

Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016.

Appendix A. Appendix

A.1. Definition of evaluation metrics

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{8}$$

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i) \tag{9}$$

$$\text{SD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left((y_i - \hat{y}_i) - \text{ME}\right)^2} \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{11}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{12}$$

A.2. Plots between ground truth beats and predicted beats

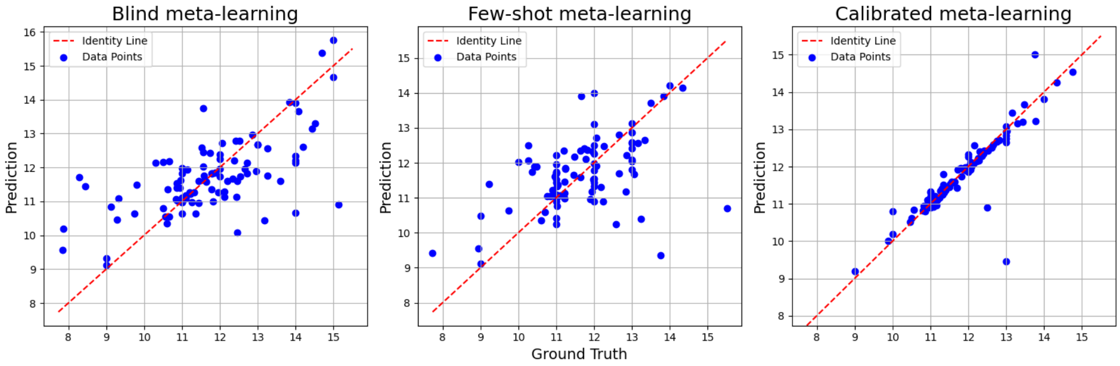

Figure 3: Scatter plot between ground truth HR and predicted HR for the MMSE-HR Dataset

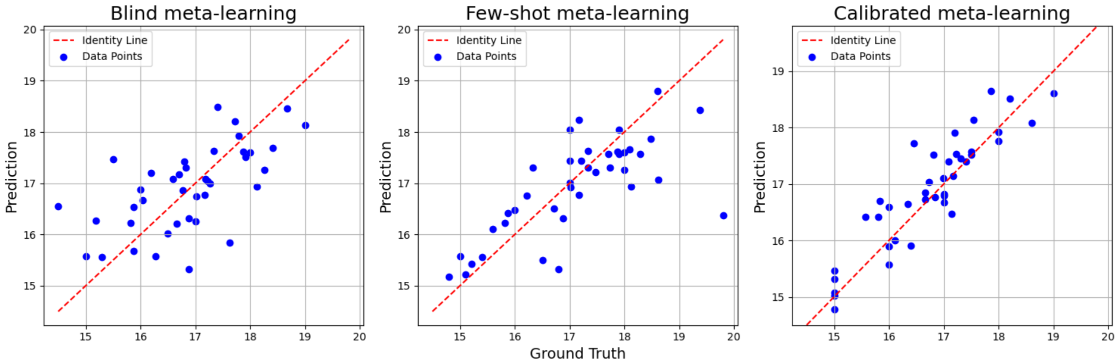

Figure 4: Scatter plot between ground truth HR and predicted HR for the UBFC-rPPG Dataset

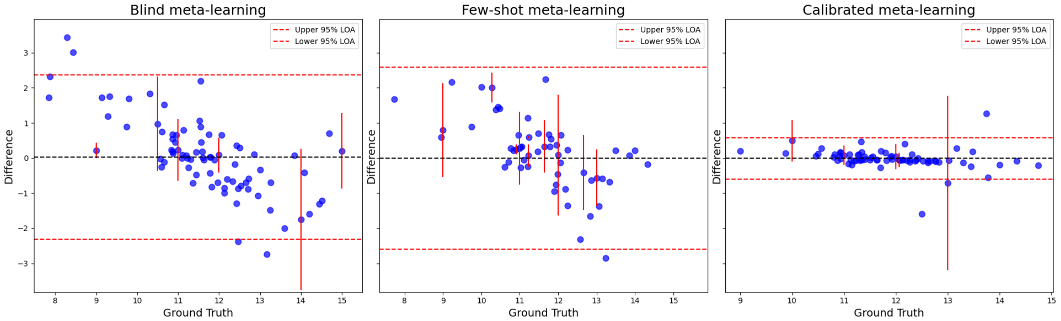

Figure 5: Bland-Altman plot with adjustments for ST2S-rPPG on the MMSE-HR Dataset, the black line represents the mean and the red lines the 95% limits of agreement

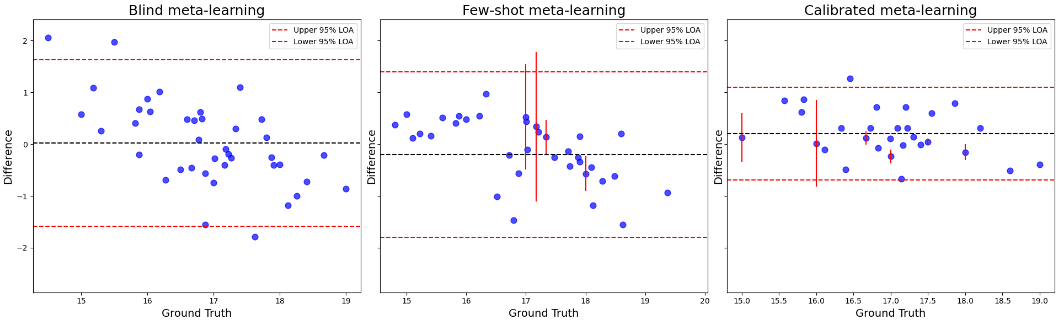

Figure 6: Bland-Altman plot with adjustments for ST2S-rPPG on the UBFC-rPPG Dataset, the black line represents the mean and the red lines the 95% limits of agreement

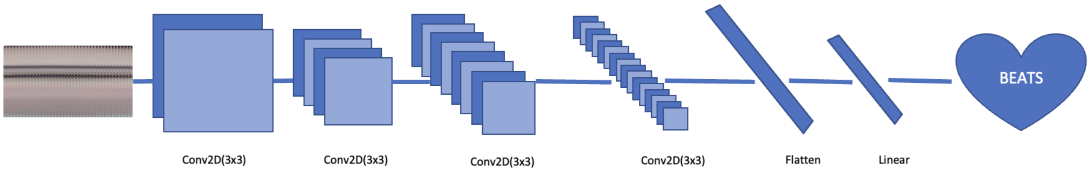

Figure 7: Architecture Diagram for the CNN