Smart Health 23 (2022) 100240

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

CamSense: A camera-based contact-less heart activity monitoring

Zahid Hasan *, Sreenivasan Ramasamy Ramamurthy, Nirmalya Roy

Mobile, Pervasive, and Sensor Computing Laboratory, Department of Information Systems, UMBC, 1000 Hilltop Cir, Baltimore, MD 21250, USA

ARTICLE INFO

Keywords:
Remote PPG
Deep learning
Multitask network
rPPG reconstruction

ABSTRACT

Remote Photoplythysmograpy (rPPG) systems enable contactless heart activities (heart rate, heart rate variability) monitoring by estimating Photoplethysmogram (PPG) signal, blood's volumetric variation in the skin tissues, leveraging the occurred diffused reflection from the exposed skin in skin video. They can primarily monitor heart activities using off-the-shelf video sensors while ensuring the safety of concerned individuals during contagious diseases. However, developing the rPPG systems is challenging due to the marginal presence of PPG signal in the video stream, data variations, limited and noisy rPPG data. In this regard, we propose an end-to-end deep learning-based approach for camera-based contactless sensing *CamSense* for recovering PPG signals from consecutive raw video frames. Firstly, we design and validate a personalized model to bypass data variation and noise across data collection with a modified objective function under realistic settings. Secondly, we explore and design multi-task learning (MTL) network to address the rPPG data variabilities for learning generalized rPPG representations. We also propose a transfer learning approach that integrates an efficient weight initialization to scale the rPPG systems under different domains and settings for fast and generalized training and inferences. Finally, we offer a new dataset, MPSC-rPPG dataset, containing multiple RGB videos and corresponding PPG ground truth for end–end rPPG network training. We evaluate *CamSense* on two public datasets and our MPSC-rPPG dataset across multiple subjects and heterogeneous camera sensors such as DSLR and near-infrared sensors with different ground truth provider PPG sensors (wrist, finger) to showcase its' generalizability. We further validate our components' design choices by performing ablation studies using different settings. Our developed model approximates accurate PPG signals with an average root mean square error (RMSE) of 0.08, 0.10, and 0.06 for personalized models, MTL model, and transfer learnings on the held-out test videos.

1. Introduction

At this crucial time of a contagious pandemic of COVID-19, contactless physiological signal sensing is of utmost importance. One of the most critical bio-signals that could indicate an underlying medical condition is cardiovascular activity. Heart rate (HR) is the crucial parameter that measures the heart's function and reflects the balance of sympathetic and parasympathetic activity (Arnold, Fitchett, Howlett, Lonn, & Tardif, 2008). Studies have shown the relation between high natural resting HR, heart rate variability (HRV) with cardiovascular diseases, coronary artery diseases, and risk of stroke (Draghici & Taylor, 2016; Zhang et al., 2016). High HR can indicate infections or any cause of fever, asthma or breathing problems, and anemia (Publishing, 2020).

* Corresponding author.
E-mail address: zhasan3@umbc.edu (Z. Hasan).
URL: https://mxahan.github.io/ (Z. Hasan).

https://doi.org/10.1016/j.smhl.2021.100240
Received 25 October 2021; Accepted 1 November 2021
Available online 24 November 2021
2352-6483/Published by Elsevier Inc.

Remote Photoplethysmography (rPPG) measures cardiovascular activity in terms of PPG (Hasan & Haque, 2015) without making skin contact by deploying the video sensors to leverage the variation in the reflection of human skin from a video sequence (Deng, Hung, Ho, & Lin, 2017; Verkruysse, Svaasand, & Nelson, 2008). PPG signal measures the volumetric variation in the uncovered skin tissues from the video streams. This blood variation in the skin is the direct result of rhythmic heart activities. HR, HRV parameters can be conveniently inferred by analyzing the approximated PPG signal (Alqaraawi, Alwosheel, & Alasaad, 2016; Hasan & Haque, 2015). Additionally, the estimated raw PPG signals' quality helps interpret the rPPG system performance during inference. The approximated signal provides relevant feedback to adjust the design choice of the rPPG system during its development phase. Due to the interpretability and design-related feedback, we are motivated to investigate the rPPG systems that estimate the encoded PPG signal from skin video instead of measuring the HR, HRV parameters directly.

Developing rPPG is an active area of research due to its potential for instant remote physiological monitoring while ensuring better safety for the healthcare workers, especially during a pandemic of contagious diseases. The contactless nature also provides a quick screening of cardiovascular activity with requiring only a video camera sensors. Traditionally, extracting PPG from a facial video sequence consists of two distinguished parts; detect, localize the bare skin, and apply systematic signal processing techniques on the localized skin to extract the PPG. Such algorithms heavily depend on the skin/face detection aspect and skin signal quality, strength, and variation. Besides, these frameworks also fail to generate a consistent PPG signal from video data due to different heuristic dependencies of the signal processing techniques (Osman et al., 2015; Verkruysse et al., 2008). Alternatively, deep learning (DL) frameworks can approximate the function for PPG extraction from skin videos. They have shown promise in extracting PPG information from video sequences. However, recent DL methodologies depend on video post-processing of network output (Zhan et al., 2020). Other works focus on detecting ECG signal (Mirvis & Goldberger, 2001) peaks in the video sequences (Huynh, Balan, Ko, & Lee, 2019), which leaves a void in developing a scalable end-to-end network design to acquire PPG signal from raw video with embedded PPG information (Glasmachers, 2017).

Our study's overarching objective is to develop a scalable end-to-end DL-based PPG signal estimator from video sensors to continuously extract PPG, eventually HR and HRV, without depending on the handcrafted features (Draghici & Taylor, 2016). The continuous PPG estimation allows monitoring the system performance during inference. Moreover, during real-world data collections for PPG, there is a variance introduced in the PPG signal specific to a person, ambiance temperature, type of underlying conditions, motion artifacts, and so on. In this study, we plan to investigate two particular research questions. Firstly, we explore the prospect of recovering the PPG signal from the video sequence data using an end-to-end DL-based architecture and investigate their scalability to mimic real-world scenarios in contactless bio-signal sensing applications. Secondly, we explore the knowledge transfer of the learned model for faster convergence of the model while ensuring the same accuracy with reduced training data instances for the new scenario. In our research works, we postulate the following contributions:

- ***CamSense:* DL-based rPPG estimator.** We design and validate a contact-less end-to-end personalized camera-based rPPG estimator framework that helps to derive the heart activity across different modalities of video sensors (NIR, RGB, and RGB demosaiced). Our proposed framework eliminates the need for any additional device contact sensors to measure heart activity.
- **Generalized Deep Framework using Multi-task Learning.** We postulate a multi-head multi-task learning model to account for target variations specific in the rPPG context. We posit a shared network to learn the features representation of multiple users towards generalization *CamSense*.
- **Transfer Learning enabled *CamSense*.** To showcase the scalability of *CamSense*, we design a transfer learning-enabled *CamSense* model that is capable of preserving its performance over various personal and sensing heterogeneities.
- **Evaluation and Validation of *CamSense*.** We offer an MPSC-rPPG dataset to enable end-to-end rPPG system development and validation. We validate *CamSense* on three video modalities using two public datasets and our MPSC-rPPG dataset. Further, we perform extensive ablation studies to validate the effectiveness of the network architecture and the learning objective choices.

We organize the paper as follows. Section 2 discusses the related work while Section 3 articulates the problem statement and architecture of our proposed DL-based rPPG estimator. Section 4 describes the datasets, model pipelines employed to conduct the experiments. The results are presented in Section 5 while Section 6 depicts the challenges *CamSense* tackles. Section 7 presents an ablation study, Section 8 posits limitations with future research directions, and Section 9 concludes the work.

2. Related works

Previous work on rPPG can be broadly categorized as (i) Signal Processing and (ii) DL approaches. The former approach involves identifying and tracking the bare skin region from consecutive video frames, which results in skin-intensity time-series. Various approaches such as filtering the green channel with a 3rd-order Butterworth bandpass filter (Verkruysse et al., 2008), frequency filtering (Gudi, Bittner, Lochmans, & van Gemert, 2019), chrominance signal analysis method CHROM (De Haan & Jeanne, 2013), plane orthogonal to the skin tone (POS) (Wang et al., 2016), multi-channel spectral matrix decomposition with Kalman filter for spectral peak tracking in motion corrupted videos (Wu et al., 2018), source separation (Macwan, Benezeth, & Mansouri, 2019), feature extraction with SVM method (Osman et al., 2015), showed their success on extracting PPG in videos. Additionally, several source separation methods include Independent Component Analysis (ICA) (Poh et al., 2010), Singular Spectrum Analysis (SSA) on compressed videos (Zhao et al., 2018), nonlinear mode decomposition (Demirezen & Erdem, 2018) have shown their success in PPG decoding from the video signal. The authors of (McDuff, Gontarek, & Picard, 2014) have localized systolic and diastolic peak from facial videos using Independent components analysis and Source Separation (SS) techniques. As an alternative to RGB camera modality, Near-infrared (NIR) videos contain PPG information and have been used to extract PPG information via signal de-noising with spectral estimation (Magdalena Nowara et al., 2018), Empirical mode decomposition (EMD) (Zhang et al., 2018), discriminative signature-based extraction (Wang et al., 2019a). Moreover, infrared (IR) videos have also been used as an input source to monitor infant HR (Abbas, Heimann, Jergus, Orlikowsky, & Leonhardt, 2011). In another study, (Hu et al., 2019) leverage illumination variance from the IR videos was leveraged to monitor PPG signals. The authors of (Adib, Mao, Kabelac, Katabi, & Miller, 2015) investigated the wireless reflection and (Tarassenko, Villarroel, Guazzi, Jorge, Clifton, & Pugh, 2014) proposed autoregressive and pole cancellation methods to extract breathing rate and HR.

In recent times, DL-based approaches have shown promise to acquire PPG from facial or bare skin videos. The authors of *VitaMon* (Huynh et al., 2019) have demonstrated the ability of convolutional neural network (CNN) based architecture to detect PPG peaks (acquired at a higher sampling rate) from raw video frames. The authors used CNN and Fully connected feed-forward (FCFF) network in the second step neural network to further localize the ECG signal peak position in the finer scale from the video frames with regression loss in the frame prediction. The authors proposed a network that uses phase information to predict the ECG peak occurrence time and HRV in higher resolution. Our model, *Camsense*, is closely comparable to the *VitaMon* (Huynh et al., 2019), however, our focus greatly differs. We focus on reconstructing the entire PPG signal rather than the ECG peaks (R peak in Huynh et al., 2019; not the entire ECG reconstruction). Our proposed novel loss function trains the network to approximate the underlying PPG signal from the raw video. Moreover, we offer a Multi-Task Learning solution to adapt the variations caused due to user's individuality and other relevant causes to scale our proposed model. Authors of the (Chen & McDuff, 2018) used Convolutional Attention Network (CAN) to learn facial features to extract PPG in both RGB and IR videos. Various state-of-the-art video activity detection architectures such as two-stream-CNN (Wang et al., 2019b) and 3D-CNN (Bousefsaf et al., 2019) have approximated HR from video input.

3. *CamSense:* DL-based rPPG estimator

The input to CamSense is the video with sufficient PPG information. Videos are susceptible to noise such as offset position of the face, different color gamut, uneven resolution, and so on. To minimize such noise or variations, we employ a pre-processing stage to crop, reshape, and normalize the image as elaborated in Section 4.2. To train *CamSense*, we require simultaneous PPG data as the target for the input videos. Fig. 1(a) depicts the overview of the *CamSense* model that comprises of Multi-head multi-task learning and transfer learning components. A trained *CamSense* model estimates PPG from the facial video input only. In this section, we describe the details of problem statement, *CamSense* model, and its various components.

3.1. Problem statement

The facial video with sufficient PPG information undergoes two functions to recover the underlying PPG. Firstly, independent of the target, the model should localize the exposed skin pixels, and extract the facial PPG from the spatial–temporal information of the consecutive frames. Secondly, depending on sensors, the extracted facial PPG signals require transformations to match the target finger/wrist PPG signal. The transformation function differs based on the individual, sensor characteristics, placement, and alignment between video and the PPG signal and hinders the development of generalized model by training with the combined dataset. We hypothesize that a neural network with adequate search space and appropriate objective would implicitly discover the two functions to learn and mimic the ground truth PPG signal from video data only (Cybenko, 1989). Mathematically we can formulate these as follow:

$$task_{final} = f_{head}(f_{shared}(Input)) \tag{1}$$

Where f_{shared} and f_{head} represents functions performing two discussed transformation. To accommodate a large variance, compatibility, alignment mismatch in the target PPG signal, We postulate that a multi-head multi-task learning (MTL) model as a potential solution. MTL can generalize the PPG extraction model by leveraging the shared task (Ruder, 2017) of learning the general spatio-temporal features from the video frames using a shared body network. The shared body allows us to build a robust feature extraction network. Mathematically, we express the scenario for total J subjects as follow,

$$task_{Sub_j} = f_{head_j;\theta_1}(f_{shared;\theta_s}(video_j)); \forall j \in J \tag{2}$$

3.2. Network architecture

PPG information is encrypted in both the spatial region containing skin and temporal domain across the channels of video (Wu et al., 2018) capturing the skin reflection variation. A 2-D Convolutional Neural Network (CNN) would learn the necessary spatial and temporal features from the video to retrieve the PPG signal from exposed skin videos (Huynh et al., 2019). We use four CNN layers with an l2 regularization. Each of the convolutional layer is followed by Batch Normalization (BN), Rectified Linear Unit (ReLU) activation function, and max-pooling layers. The Inception layers with multi-scale parallel convolutional filters have shown to be better feature selection in the case of facial images (Schroff et al., 2015) and the global average pooling (GAP) helps to localize the object (Girshick et al., 2014). Subsequently, we put two naive inception layers (Szegedy et al., 2015) each followed by a GAP layer. FCFF layers help learn to transform the CNN extracted features to the PPG signal. We flatten the output of the final GAP layer. The GAP layer output passes through two FCFF layers with ReLU activation. Motivated from a Generalized Adversarial Networks

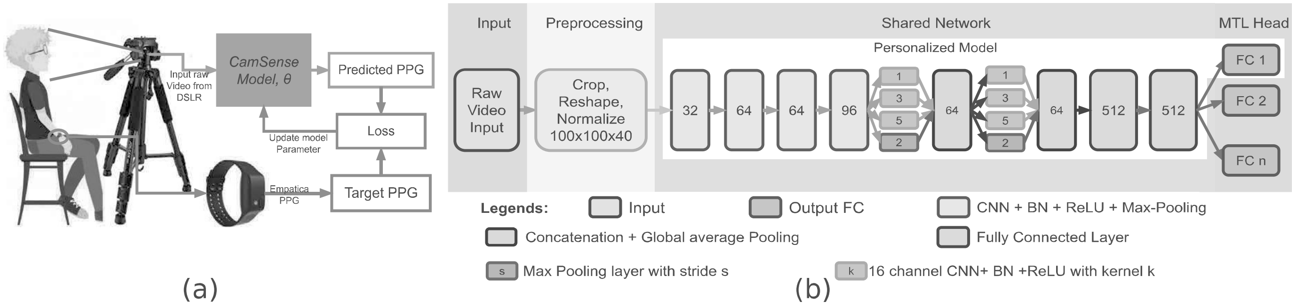

Fig. 1. (a) Overall setting for the rPPG system. The blue line sections exist during the training phase only. (b) *CamSense* Architecture for the personalized, MTL and transfer setup.

(GAN), we propose to use hyperbolic tangent *tanh* activation function to ensure smoothness of the regressed output layer. For the experiments related to Multi-task learning (MTL) (Long et al., 2015) networks, the output layer also acts as the head network (Ruder, 2017). Our overall model with a single head consists of 351,349 parameters of which 350,517 are trainable parameters with an input size of $100 \times 100 \times 40$. The shared-body network and the downstream regression head network consist of $\sim 69K$ and $\sim 11K$ parameters. We open-source our code, model, and data to reproduce the exact results reported in this study and further development in our[1] GitHub repository. We depict the overall architecture for both personalized (single FCFF layer) and MTL models in Fig. 1(b).

3.3. Loss objective

To effectively improve the learning phase, we proposed to optimize a novel loss objective in rPPG extraction to grasp the magnitude and phase difference of target PPG. The loss function consists of weighted sum of two loss components: root mean square error (RMSE) loss, and Sign agreement loss. The RMSE loss penalizes the network output for the point-wise approximation error in the magnitude. The sign loss component prevents the network from learning a fixed sequence of PPG over the training instance by penalizing the phase difference for the network predicted PPG, guiding the model to learn the appropriate relationship between the PPG phase and input video sequences. Let there be m training instances; $(X_1, Y_1), (X_2, Y_2), \ldots, (X_M, Y_M)$, where X_m be the mth input video data and Y_m be the corresponding output PPG signal vector consists of components $[y_{gt,1,m}, y_{gt,2,m}, \ldots, y_{gt,N,m}]$. The network predicts $Y_{p,m}$ consists of components $[y_{p,1,m}, y_{p,2,m}, \ldots, y_{p,N,m}]$. The RMSE and sign loss for mth sample output prediction $\mathcal{L}_{RMSE,m}$ and $\mathcal{L}_{sign,i,m}$ are as follow,

$$\mathcal{L}_{RMSE,m} = \sqrt{\frac{\sum_{i=1}^{N}(y_{gt,i,m} - y_{p,i,m})^2}{N}} \tag{3}$$

$$\mathcal{L}_{sign,m} = \frac{\sum_{i=1}^{N} \mathcal{L}_{sign,i,m}(y_{gt,i,m}, y_{p,i,m})}{N} \tag{4}$$

$$\mathcal{L}_{sign,i,m}(y_{gt,i,m}, y_{p,i,m}) = \begin{cases} 1 & y_{gt,i,m} \text{ and } y_{p,i,m} \text{ signs mismatch} \\ 0 & y_{gt,i,m} \text{ and } y_{p,i,m} \text{ signs match} \end{cases} \tag{5}$$

Thus, we propose our final objective function as below to better optimize by both penalizing the magnitude and phase differences.

$$\mathcal{L}_{Total} = \frac{\sum_{m=1}^{M}(w_{RMSE}\mathcal{L}_{RMSE,m} + w_{sign}\mathcal{L}_{sign,m})}{M} \tag{6}$$

4. Experiments

Here, we describe the data collection procedure, data processing, and the pipeline of *CamSense* model. All the DL models were implemented using Python and Tensorflow 2.1. The codes were run on a Linux server housing an Intel Core i76850K CPU, 4x NVIDIA GeForce GTX 1080Ti GPUs, and 64GB RAM. For benchmark comparison, we recreated the *VitaMon* work using the PPG signal.

[1] https://github.com/mxahan/project_rppg.

4.1. Datasets

We have experimented using two of the largest available public rPPG datasets and our collected MPSC-rPPG dataset, described in the below sections.

MERL-Rice NIR Pulse Indoor (MR-NIRP Indoor) dataset (Nowara et al., 2018). This MR-NIRP dataset consists of 8 persons with both gender and various facial expressions. The dataset provides three video modalities: near-infrared (NIR), RGB raw, and RGB demosaiced for each subject. Each video is captured at 30 frame per second (fps) for about 3 minutes with simultaneous finger PPG at 60 Hz as the ground truth.

UBFC-rPPG dataset (Bobbia, Macwan, Benezeth, Mansouri, & Dubois, 2019). The UBFC-rPPG dataset contains 8 simple and 42 realistic RGB videos fps varies from 28 to 30 with PPG records sampled from 30 to 62.5 Hz. These parameter variances caused an incompatibility with both MR-NIRP and MPSC-rPPG datasets. Hence, we picked 6 subjects with 8 different trials from UBFC-rPPG in our experiments that met the criteria of consistent sampling frequencies.

MPSC-rPPG dataset. In this work, we collect and open-source simultaneous PPG and video sensor data to develop and validate rPPG methods. We captured HD video using RGB stationery (in a stand) DSLR (Canon D3500) at 30 fps in a laboratory setup under artificial light for 5 min while each volunteer was sitting still on a chair three feet away in front of the camera. Concurrently, Empatica E4 Wristwatch tracks the wrist PPG sampled at 64 Hz from volunteers' dominant hand. To align the video and PPG data to the same time-stamp, we leveraged the event marker feature in the Empatica E4. During that, the volunteers count to ten for 10 s. To process the data, we identified the exact position of the event marker press of Empatica E4 by locating event marker light at particular video frames. The Empatica E4 locates the exact PPG position for the event markers pressed. This allows us to align PPG and video with an error margin of 1/30 s. The dataset consists of two females, six males volunteers. Our dataset has introduced person heterogeneity such as sex, facial hair, fitness level, and spectacles usage. Two male volunteers provided data with and without a beard and glass. All subjects are healthy and have no known medical condition having HR ranges from 50 BPM to 95 BPM. In RGB video times, our dataset suppressed both public datasets, and more suitable for developing generalized DL models. Moreover, the MPSC-rPPG dataset introduces heterogeneity by placing the PPG sensor at the wrist. Interestingly, the hypotheses remain the same as now the network aims either wrist or finger PPG. This would showcase the generalizability of *CamSense*.

The three datasets are well collected for rPPG. Nevertheless, the public datasets sometimes suffer a random phase shift between video and the PPG signal. Moreover, each PPG depends on the sensor location and personal characteristics, and few PPG signals contain random trends, missing values, and magnitude variances across the data collected and lacks anomalies in HR.

4.2. Data preparation

To prepare the datasets for training, consecutive 40 frames (1.33 s) green channel of facial videos from various starting times were stacked as an input instance. We reshaped each frame to 100×100 resolution and input data instance becomes $100 \times 100 \times 40$. The trained network uses this as input. To create a training instance, we take the simultaneous PPG signal of the corresponding input data as a target. For the MR-NIRP dataset, we take 40 consecutive frames corresponding 80 consecutive PPG samples. For the UBFC-rPPG and MPSC-rPPG dataset, 85 points of PPG data matched the 40 frames of video from the same starting time. We normalized the input pixel values between 0 to 1, and the target output PPG values between -1 to 1. The output range is suitable to implement the sign agreement loss and *tanh* activation function.

4.3. Model description

Firstly, we employ a personalized PPG extraction model for each individual to show the learning capabilities of the *CamSense* model across heterogeneous sensor locations (finger and wrist), personal traits, and sensor modalities (NIR, RGB-raw, and RGB-demosaiced). Next, we generalize the model across multiple users using an MTL framework. Further, we transfer learned knowledge across subjects and modalities to scale *CamSense*. We fine-tune the model with a small amount of target domain (a new individual/sensor) data while initializing weights from the source domain (Yosinski, Clune, Bengio, & Lipson, 2014) pre-trained personalized or MTL model. Here, we seek a fast and robust fine-tuning using minimal training instances from the target domain.

Personalized model. To develop the personalized *CamSense* model, we utilize the 166.67 s video and corresponding PPG from a single subject to prepare the training and validation set. We take random face cropping as augmentation. We leave the rest of the data to evaluate the test performance. We create about 4500 training instances and 500 validation instances. We trained the model for 150 epochs with a batch size of 16 without early stopping criteria starting with random weights initialization (Glorot & Bengio, 2010). We used a constant learning rate (LR) of 0.0008 with Adam optimizer.

MTL model. We address the alignment and compatibility issue between the video and the PPG across datasets towards generalizing the PPG extraction model using the MTL model. To develop these, we attach two separate MTL head FCFF network on top of the shared network output in parallel 1. To prepare the training data for MTL *CamSenes* model, we considered 83 seconds of data from each person, out of which 2250 were used as training instances and 250 as validation instances for each head network (to keep data amount same for personalized and MTL model). The MTL model can conveniently be extended to incorporate more subjects by allowing more MTL heads. In contrast to the personalized model, each person's data is associated with training different head networks. We train each head with one persons' data and alternate for the next head using other persons' data in the next batch to complete a training step and so on so forth. We iterative train the head networks for 150 epochs. While training one head, we

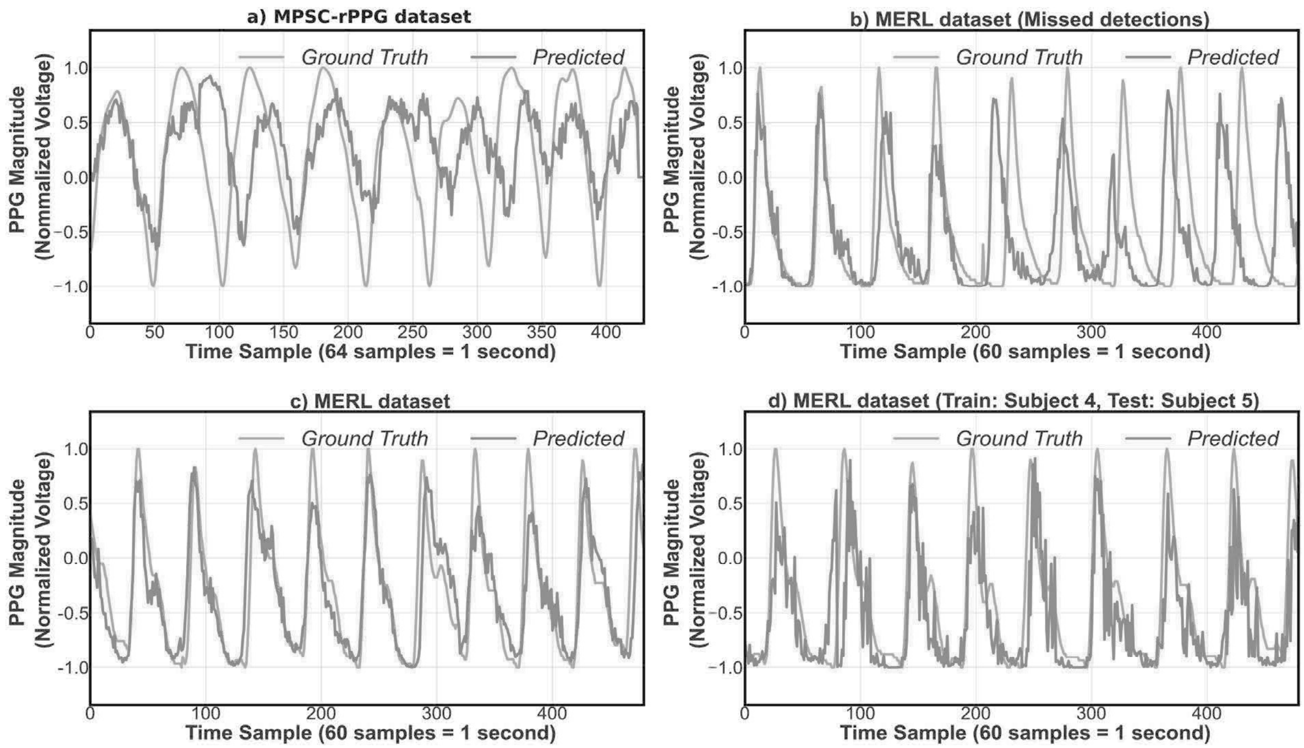

Fig. 2. PPG estimation on individual's test data of (a) Personalized model for Subject 4 of the MPSC-rPPG data. (b) A missed peak in the ground truth around 200 samples. The Model predicted the peaks right and exhibit phase difference for the rest of the time frames. (c) Personalized model for Subject 3 of MR-NIRP dataset. (d) Prediction on new individual by a different personalized model.

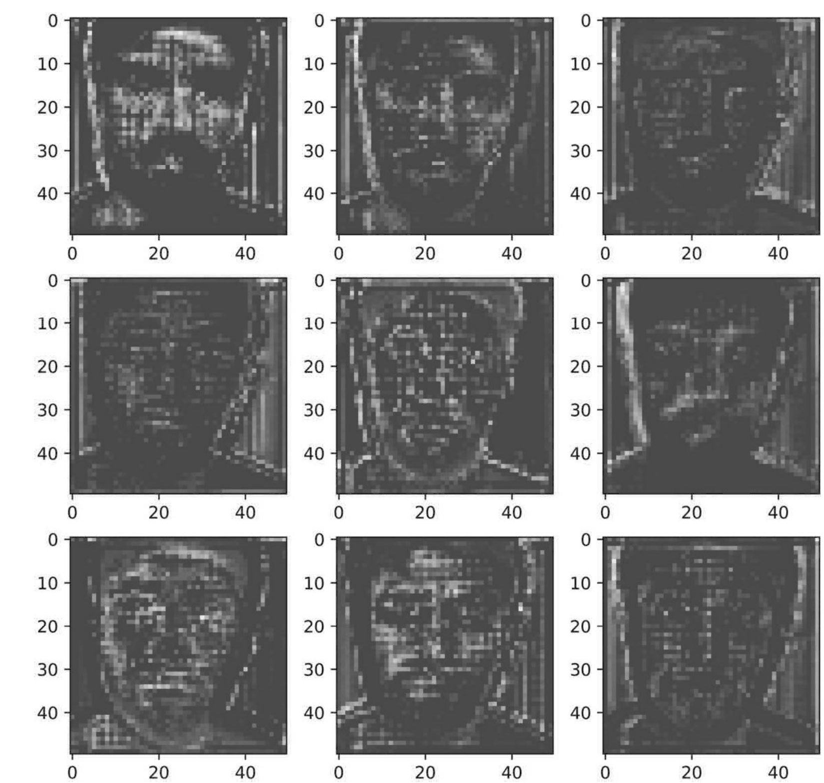

Fig. 3. Activation map of few filters of inception layer shows the models attention to the face region.

Table 1
Test results of HR errors and peak detection error on the MR-NIRP dataset for raw RGB, RGB demosaiced and NIR green channel video data.

Subj	Test len (sec)	Raw RGB			RGB demosaiced			NIR		
		TP /Act. peak	FP	FN	TP /Act. peak	FP	FN	TP /Act. peak	FP	FN
Sub 1	12 s	16/18	1	2	17/18	1	1	16/18	2	2
Sub 2	43 s	40/47	7	7	40/47	7	7	40/47	7	7
Sub 3	74 s	81/83	3	2	81/83	4	2	81/83	4	2
Sub 4	13 s	0/14	14	14	0/14	14	14	0/14	14	14
Sub 5	13 s	15/15	0	0	15/15	0	0	15/15	1	1
Sub 6	27 s	32/32	0	0	32/32	0	0	32/32	0	0
Sub 7	13 s	10/14	5	4	10/14	4	4	10/14	4	4
Sub 8	13 s	10/10	1	0	9/10	1	1	9/10	1	1

freeze the other head network's weights. In MTL the Shared network gets updated for each batch, the head network updates only for the associate persons' data. We used Adam optimizer with an LR of 0.0005 and 0.0008 for the shared body and the individual head network. We experimented and empirically selected loss components weights to be 1 and 0.5 for w_{RMSE} and w_{sign}, respectively, for both personalized and MTL models.

Transfer learning model. We transfer our learned knowledge from both personalized and more generalized MTL *CamSense* model to a new domain (individuals, sensors, and modalities). For personalized model transfer, we initialize the whole network with the weights of the learned model. To transfer from MTL, initialize only the shared network and initialize with random weights for the MTL head. To fine-tune the model in the target domain (sensors or persons), we use 33 seconds of target domain video and corresponding PPG signal and train for 100 epochs. We have selected 1 as the weights for both RMSE and sign agreement loss components. We apply Adam optimizer with an LR of 0.0001 for fine-tuning.

During the learning phase, we closely monitor and find the learning curves' small gap between the validation set consistently [Fig. 4 b]. We further monitor the intermediate CNN filter activation map to interpret and understand it better. After the training, the higher layers CNN filters learned to focus on the face. The trained filter in different trained models shows similar characteristics 3.

5. Results

We report the results for the leave-out test video for all the datasets. We train our models from different initialization and reach 0.11 in test RMSE on average for all the datasets. Besides, we identified sensitivity (True Positive, TP), False Positive, FP), and missed detection (False Negative, FN) for evaluation. Next, we compared the estimated PPG shape and the number with the target PPG. To qualify as a correct peak, we overlap the predicted and target PPG to identify the peak location of both signals. If the positions of prediction and ground truth are within less than 0.5 s, they are labeled as TP; otherwise, they are labeled as FP or FN based on the presence of target or predicted peaks. As we aim for PPG extraction, we mostly focus on comparing the estimated rPPG with ground truth PPG instead of evaluating the PPG features (Castaneda et al., 2018).

5.1. Personalized model

We report the results for three modalities of the MR-NIRP dataset and RGB modality of the MPSC-rPPG dataset for all the subjects. We ignored UBFC dataset for this experiment as it has a small video per subject. The personalized *CamSense* consistently tracked correct PPG in the test regions as shown in Fig. 2c for MR-NIRP dataset. The model underperforms in general near the extreme values since the *tanh* activation needs a higher absolute value to map the two extremes. Table 1 shows the overall performance of the personalized model all the subjects of the MR-NIRP dataset for three different modalities (RGB, RGB demosaiced, and NIR) for peak detection. We observe that the RGB demosaiced model performs very close to the RGB model. However, the NIR modality performed a little inferior comparatively. The reduction in performance can be attributed to using a single channel for the analysis of an already limited information modality (NIR). The overall results comparison is depicted in Fig. 6. The alignment of target and input can explain the results of the subject 4 and 7 of the MR-NIRP dataset in Table 1 and Fig. 2b. The ground truth PPG signal has an irregular repetitive pattern due to unknown reasons. This irregularity may be caused due to a sensor or an abnormality in the heart activity of the subject. We located the location where the model missed the shift in the data temporally and found it just before the test sequence began 2b. This missed-detection needs to be investigated on a case-by-case scenario with a physician. Overall the models suffered 0.09 RMSE between predicted and test PPG.

Similarly, for the MPSC-rPPG dataset, the personalized *CamSense* show its ability to estimate wrist PPG from the video sequence (MSE of 0.10 on held-out data) as shown in Table 2. We observed that a 30% drop in sensitivity with a beard compared to a no-beard. A similar drop (28%) was observed with glasses when compared to no glasses. As the beard or glasses reduce exposed skin and introduce complexity, the model performance drops. In general, for personalized *CamSense* model performed consistently on the RGB and RGB demosaiced modality using a single green channel. The *CamSense* underperforms in NIR compared to the previous two modalities. We tested *CamSense* on the same persons' held-out dataset. There was no leave-one-person-out analysis because of

Table 2
Test result on the MPSC-rPPG dataset for the personalized, MTL model and transfer learning experiments.

Subject (RGB)	Test Length (Second)	Personalized Model			Transfer from subject 1			MTL model sub 2 and 4 fine tuned rest		
		TP /Actual Peaks	FP	FN	TP /Actual Peaks	FP	FN	TP /Actual Peaks	FP	FN
Sub 1 (No Beard)	10 s	12/12	1	0	–	–	–	12/12	0	0
Sub 1 (Beard)	10 s	8/12	4	4	13/12	2	1	11/12	2	1
Sub 2 (Female)	10 s	10/11	2	1	12/11	2	1	9/11	2	2
Sub 3 (Glass)	10 s	8/11	2	3	12/11	2	3	9/11	1	2
Sub 3 (No glass)	10 s	11/11	0	0	11/11	0	0	11/11	0	0
Sub 3 (Different day)	10 s	11/11	0	0	11/11	0	0	11/11	0	0
Sub 4	10 s	12/12	0	0	10/12	3	2	10/12	1	2
Sub 5	10 s	15/14	3	2	14/14	0	0	14/14	2	2
Sub 6	10 s	9/9	0	0	9/9	0	0	9/9	0	0
Sub 7	10 s	12/13	1	2	13/13	0	0	12/13	1	2
Sub 8	10 s	12/12	0	0	12/12	0	0	12/12	0	0

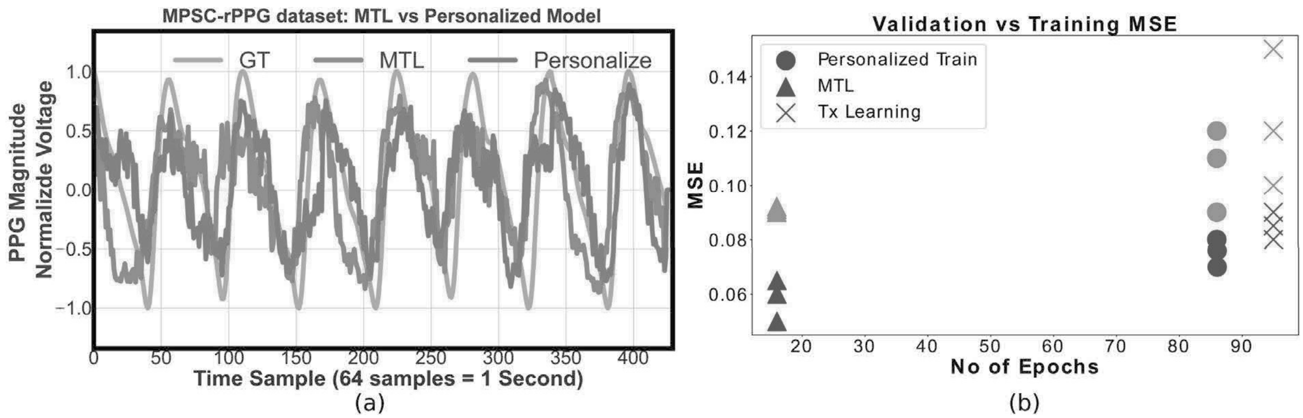

Fig. 4. (a) Comparison of PPG estimation among Personalized and MTL *CamSense* on same subjects' held-out data. (b) Final Training (blue) and validation (red) RMSE loss values for the proposed approaches from different initialization.

limited data and PPG starting position uncertainty. Moreover, some ground truth is not zero mean, biased to either positive or negative. This forced the sign loss to skew the output towards the sign of the mean. As a result, some peaks never crossed zero in its' cycle.

5.2. Multi-task learning model

We overcome small maladjustment between input and target using MTL setup. MTL architecture reduces the computational cost and increases generalization. The Fig. 4 compares the result of personalized model and jointly trained MTL *CamSense* models for common individual. The MTL framework can be extended to include more persons conveniently. We report our MTL *CamSense* model performance over all the MR-NIRP data subjects 3 and the MPSC-rPPG dataset 2. The results of Table 2 shows that generalized *CamSense* model quickly adapts to target domain and predicts PPG signal peak accurately.

In comparison with personalized models, MTL *CamSense* underperforms on the training subjects (MSE of 0.12) in general. It can be explained by the fact that the MTL model shared network needs to extract features from multiple subjects, whereas the personalized model learns to focus single person as depicted in Fig. 1. Despite this drawback, the MTL model provides a way to scale our methods for multiple subjects in presence of input–output misalignment with better representation learning on the shared network.

5.3. Transfer learning

The personalized model can track HR for different persons with various HR bpm profiles in case of consistent sensors without any fine-tuning 2d. A similar conclusion always holds for transferring from RGB to RGB-demosaiced and vice versa as evident in Table 3. However, in the case of input data distribution heterogeneity introduced by sensors variability, the *CamSense* needs fine-tuning using the target domain. We facilitate a transfer learning approach to save memory and computational cost. We restrict our analysis for transferring from a particular subject due to the combinatorial option available for transfer across persons and modalities.

Table 3
Test results of HR errors and peak detection error on the MR-NIRP dataset in transfer learning setting.

Subject	RGB transfer From sub 3 (Fine Tune)			RBG to NIR Same subject (Fine tuned)			Raw RGB to RGB demosaiced (No fine tune)			RGB MTL Trained on 1 & 2 (Fine tuning rest)			NIR MTL Trained on 1 & 2 (Fine tuning rest)		
	TP /Actual Peaks	FP	FN	TP /Actual Peaks	FP	FN	TP /Actual Peaks	FP	FN	TP /Actual Peaks	FP	FN	TP /Actual Peaks	FP	FN
Sub 1	9/12	4	3	10/12	3	2	9/12	4	3	11/12	2	1	12/12	0	0
Sub 2	12/12	0	0	12/12	0	0	11/12	1	0	12/12	0	0	12/12	0	0
Sub 3	–	–	–	13/13	0	0	13/12	2	1	13/12	2	1	11/12	1	2
Sub 4	13/13	0	0	13/13	0	0	13/13	0	0	13/13	0	0	12/13	1	2
Sub 5	14/14	0	0	14/14	0	0	14/14	0	0	14/14	0	0	14/14	0	0
Sub 6	13/13	0	0	13/13	0	0	13/13	0	0	13/13	0	0	13/13	0	0
Sub 7	11/13	1	2	11/13	1	2	11/13	2	2	11/13	1	2	11/13	1	2
Sub 8	13/13	0	0	13/13	0	0	13/13	0	0	11/13	3	2	13/13	0	0

Table 4
Transfer Result on UBFC-rPPG dataset.

Subject	Test Period Second	Transfer from personalized			Transfer from MTL		
		TP /Actual Peaks	FP	FN	TP /Actual Peaks	FP	FN
Sub 1	10 s	8/15	6	7	12/15	4	3
Sub 2	10 s	11/11	0	0	11/11	0	0
Sub 3 (F)	10 s	11/11	0	0	11/11	0	0
Sub 4 (F)	10 s	10/12	2	2	11/12	2	1
Sub 5	10 s	12/12	0	0	12/12	0	0
Sub 6	10 s	11/11	0	0	11/11	0	0

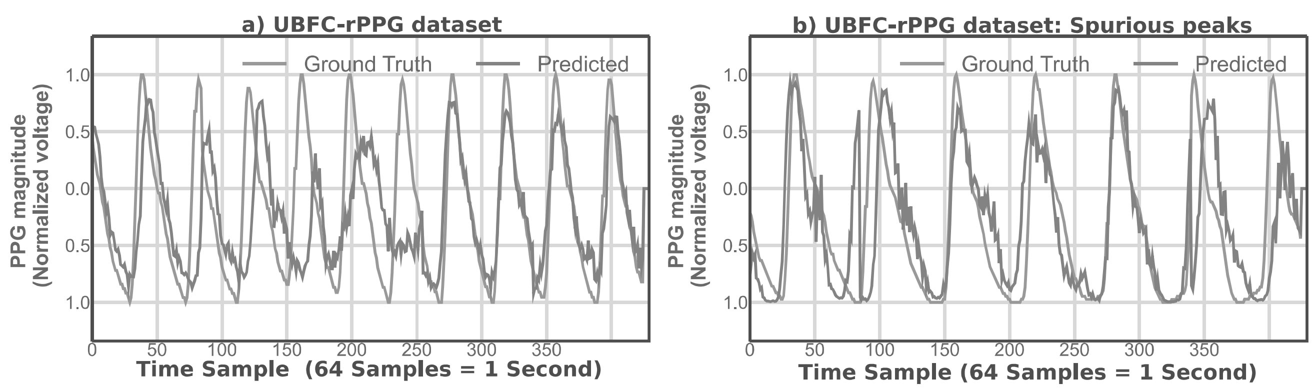

Fig. 5. (a) Recovered peaks in UBFC data after estimation failure. (b) Consistent tracking with some spurious peaks.

For the MR-NIRP dataset, all the transfer results are shown in Table 3 exporting from both the personalized and MTL model. For MTL transfer, we trained only using the 1st two subjects (both male) for MR-NIRP and two subjects (female and male) and fine-tuned for the rest using 30 s (900 frames) of target domain data. We find improvement in the results for MR-NIRP 1 and MPSC-rPPG data 2 in general. Interestingly, the model improved after exporting the pre-trained model developed using no beard data. For the UBFC-rPPG dataset, we find excellent performance for fine-tuned *CamSense* model on UBFC-rPPG dataset 5 transferring the model knowledge from the MPSC-rPPG data. These results are crucial as we quickly overcome the device heterogeneity (both camera and PPG sensor) via fine-tuning as shown in Table 4. The table results show improvement upon using the MTL transfer for the subjects in general as evident in Fig. 6a and 6b. In modality adaptation, The RGB to NIR modality transfer needed fine-tuning since the pixel properties vary significantly among them. After fine-tuning the personalized and MTL model, the RGB model learned to extract PPG information from NIR as shown in Table 3. Interestingly, in the transfer setting, MTL models transferred and learned better and faster consistently across the dataset. We kept the source person constant for this experiment (subject 4 and 7). The overall results for MR-NIRP are summarized in Fig. 6d, e. The plots show that for MTL transfer, the percentage of the FP and FN peaks have decreased.

In transfer learning, we gain two-fold benefits. Primarily we start the model from a good initialization for the target domain, leading to faster and better convergence. Secondly, we achieve high accuracy performance by fine-tuning the network with a small dataset. This allows us to scale and develop *CamSense* models for a new domain with a smaller dataset. The model reaches a similar

Table 5
Comparison between State-of-the-art (*VitaMon*) and our work, *CamSense* using our MPSC-rPPG dataset.

	VitaMon	*CamSense*
Goal	ECG peak Detection (Classify)	PPG reconstruct (regression)
Model approach	Two step Global and Personalized	End-to-end Personalized, MTL, Transfer learning
Pros	Mobile implement, Low fps video, ECG peak detection	Low-cost PPG, Scalable, accurate continuous PPG reconstruct, Interpretation
Cons	Two step, ECG, less interpretable, training target alignment issue	high fps HD dslr video, offline processing
MPSC-rPPG RMSE	N/A	Personalized (0.1), MTL (0.11), Transfer (0.08)
% TP (recreated for PPG)	Personalized (70), Global (40)	Personalized (87), MTL (83), Transfer (89)
Evaluation	Vitamon's MPSC-rPPG dataset	Two Public dataset, MPSC-rPPG dataset

Table 6
Average results of models with different loss function and final layer choices.

Metrics in %	Model design choice during training		
	Tanh+RMSE+Sign Loss	Relu+RMSE+Sign Loss	tanh RMSE
TP	*93*	72	60
FP	*10*	19	35

Table 7
Illustrate importance of the sign loss and the *tanh* activation function. * Fixed PPG memorized by the network. ** Peak counts may contain peak selection bias. ***base model is the model of MTL with fine tuning.

Subject	RGB*** (*tanh* + RMSE + Sign loss)			RGB** (ReLU + RMSE + Sign loss)			RGB (*tanh* + RMSE)		
	TP/Actual peaks	FP	FN	TP/Actual peaks	FP	FN	TP/Actual peaks	FP	FN
Sub 1	11/12	2	1	6/12	1	6	8/12	1	4
Sub 2	12/12	0	0	8/12	3	4	8/12	3	4
Sub 3	11/12	2	1	7/12	2	5	9/12	5	3
Sub 4	13/13	0	0	9/13	2	4	4/13	8*	9
Sub 5	14/14	0	0	8/14	4	6	3/14	9*	11
Sub 6	13/13	0	0	11/13	2	4	13/13	0	0
Sub 7	11/13	1	2	9/13	2	4	9/13	4*	4
Sub 8	11/13	3	2	7/13	3	6	7/13	3	6

training and validation error after fine-tuning for a smaller number of epochs and data instances. Our empirical result demonstrates that while transferring the higher weights for sign loss leads to better results. This resonates with the importance of our proposed sign loss components with RMSE loss as reported in Table 6.

6. Discussion

Apart from the architectural similarity, our works are fundamentally different from baseline *VitaMon* both in an objective and overall setting 5. *VitaMon* aims at the ECG peak (requires complex wearable) detection (Mirvis & Goldberger, 2001) on a higher scale using a two-step approach from low frame-rate videos, whereas, we reconstruct continuous PPG using a single network in an end-to-end setting. Instead of using supervised information from ECG peaks, we train the low-cost easily achievable wearable sensors' complete PPG information. We recreated the *VitaMon* global and personalized architecture and experimented with the MR-NIRP dataset and our MPSC-rPPG dataset using PPG signals instead of ECG. The global model fails to learn because of the temporal mismatch of alignment between the PPG and video as discussed earlier. The personalized *VitaMon* model detected the PPG peaks with appropriate window size. However, the networks faced instability during both the training and testing phases while predicting the single frame with the peak. This may be attributed to the primary difference between ECG and PPG signal, such as the presence of QRS complex in ECG, PPG motion artifact due to wearable sensor modality. Further, we faced the problem where the video window contains multiple peaks [high HR] or no peak [low HR] as *VitaMon* model returns one peak point for each window. Regardless, we understand the huge scope of *VitaMon* for a low sampling rate data in the availability of the simultaneous ECG.

It might be counter-intuitive that *CamSense* might learn the target PPG bias and inherent randomness. We argue that with re-scaling PPG and large training instances, the *CamSense* learns the underlying PPG to minimize the errors in expectation. The *CamSense* model mostly fails in the presence of severe head movement. The network mostly produces flat lines during this time. It tracks the PPG again after the head retains the rest position. In the presence of head movement, and Signal to Noise Ratio

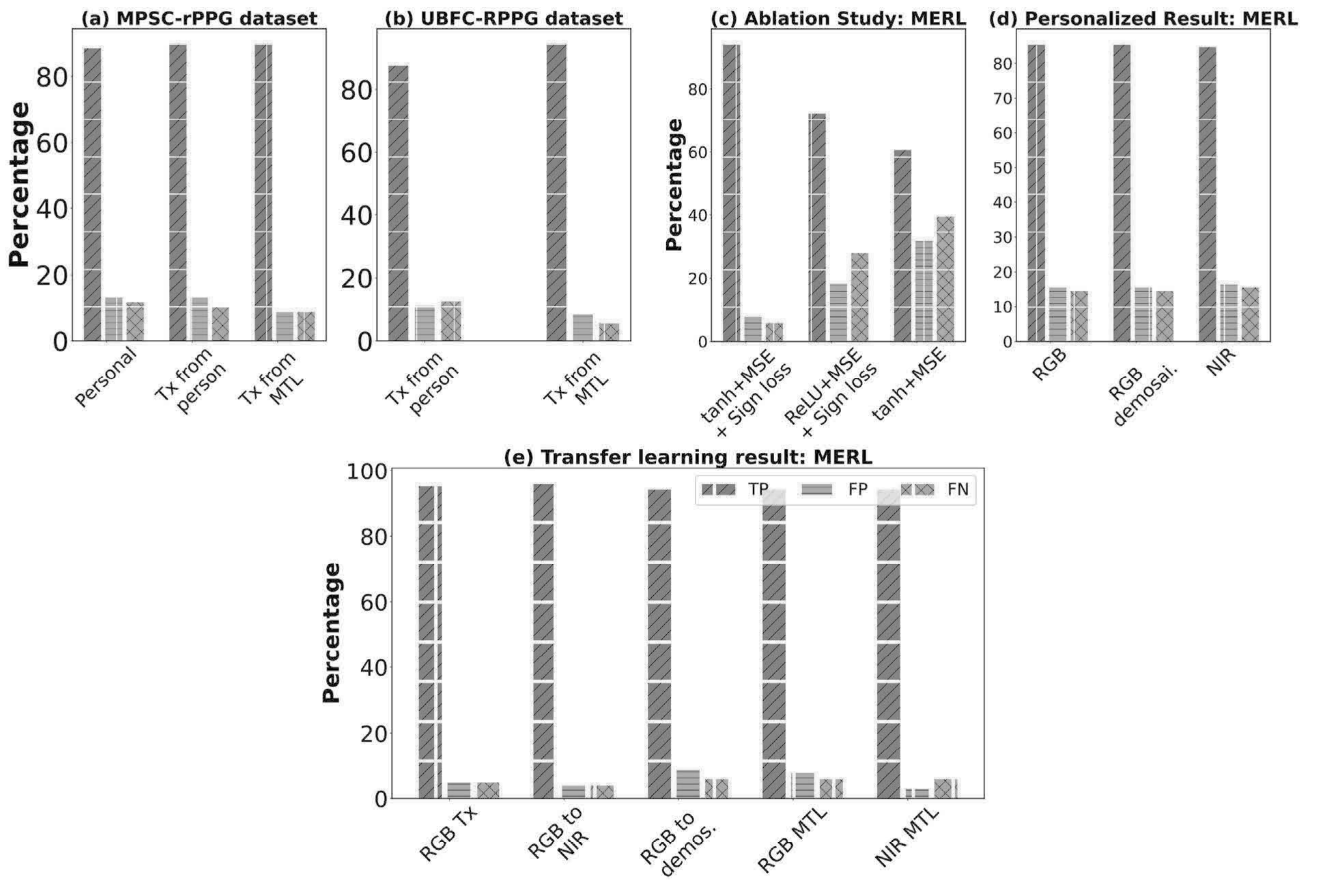

Fig. 6. Result summary of the models in terms of the peak detection. The performance increases with transfer from MTL model as minimal FP with a high TP for (a) MPSC-rPPG data (b) UBFC-rPPG data. (c) Ablation study shows the best performance with our design choices. (d) Models performance for different modalities of MERL dataset shows the generalization. (e) Transfer learning result across modalities demonstrate high accuracy in PPG estimation.

(SNR) becomes low. The models are also robust of the head orientation to some extends. Another concern may emerge about the scalability of the MTL model. But, we argue that MTL is an intelligent solution in the case of PPG data variability while training. The personalized *CamSense* with a single head can incorporate multiple people if the video and PPG are well aligned for compatible PPG sensors. We trained multiple persons' data for our MPSC-rPPG dataset and model learned and performed consistently with reduced accuracy. MTL model outperforms a single generalized model; it has extra parameters to learn simultaneously in the presence of multiple person data. This reverberated our hypothesis that one end–end network can extract continuous PPG signal.

7. Ablation study

Firstly, we study the importance of the *tanh* as activation function for the last layer. Since this is a regression problem, sigmoid activation in output makes little sense. The threshold of ReLU activation disrupted the output smoothness. The output also needs scaling between −1 to 1 to implement the sign loss function. The results of Table 7 shows the improvement in output smoothness via *tanh*. Moreover, The learning curves with *tanh* compared to ReLU converges faster to lower loss value. Fig. 6 depicts that replacing *tanh* by ReLU activation while keeping the rest of the set fixed causes increasing FP and FN.

Secondly, we ablate the impact of the objective loss function in learning to reconstruct PPG from the videos. Our experiments with the RMSE loss only model shows inconsistent convergences. In some cases, the system finds a local expectation in the PPG outputs to minimize the training loss. In these scenarios, the network generates the same PPG disregarding the phase relation between the input video and PPG. Besides, The sign loss alone with RMSE provides better feedback for both the personalized and MTL *CamSense* models. Table 7 shows the impact of sign loss in personalized model. The sign agreement loss provided a high gradient near the zero crossings. This gradient forced the network to learn the correct sign of the network output. The impacts of the components are depicted in Fig. 6. The sign agreement feedback reduced both the FP and FN in expectation in the test case. Moreover, we monitored individual losses and found the magnitude of both RMSE and sign agreement loss values are comparable. This helped us to experiments with the relative weights for the loss functions and check their contributions towards final loss function and gradient calculation.

Finally, we empirically demonstrate the better generalization of the MTL model than the personalized models by transfer learning weights from both models and comparing their performances. In our experiments, MTL models consistently matched the performance of a personalized model using smaller target domain data. These generalization can be attributed to the regularization ability of parameter sharing MTL model. This helped to the faster adaptation to the target domain concerning personalized model weight transfer. This can be observed from Table 3. The NIR results have also improved for the MTL model transfer. This demonstrates that the pretraining with the MTL model learns better generalization with the same number of training instances.

8. Limitation and future direction

The critical limitation of this work is the requirement of low variability, alignment problems in the clean dataset to develop a robust model. In some instances, the scaling of the ground truth PPG signal between −1 to 1 leaves a skewed signal. In this case, the sign loss biases the network to the dominant sign. The sudden shift or missing in PPG makes the whole learning setup error-prone. Moreover, the network is susceptible to a person's sudden movement. In some UBFC-rPPG sample, the network failed to track the correct PPG signal in the presence of the heavy head movement.

Despite some advantages, the sign loss provides a gradient only at a point near zero. Alternatively, we tested absolute difference loss if signs do not agree instead of the sign loss for a smooth gradient for the optimization. This loss underperforms concerning sign loss but better than RMSE loss only. In the future, we also plan to incorporate and experiment with the contrastive loss (Hadsell, Chopra, & LeCun, 2006) and minimize the source domain dependency (Ganin & Lempitsky, 2015). For the MR-NIRP dataset, we test our model for the same person on the same video. We somewhat tackle this issue by considering each subject on two different trials/videos with the same background in the MPSC-rPPG data and UBFC-rPPG dataset. In the future, we seek to improve consistent performance by training in multiple environments for the same individual. Another drawback of the *CamSense* models is that after fine-tuning for a new target domain, their performance on the source domain data drops.

In the future, we intend to look at the possibility of incorporating a respiratory rate measurement, another important physiological parameter. We seek to extend our *CamSense* to develop a unified, robust physiological information extraction system using remote and available video sensor data. Besides application extension, we also plan to explore algorithms and architecture to extract these parameters with a focus on better and meaningful representations. We hope to study the feasibility analysis for extracting different physiological measurements simultaneously. We are currently aiming to implement our system in memory-constraint computing devices towards developing a complete system. We are also interested to look into the prospect of federated learning to our application problem as we start with a base and fine-tune it for different sensors. We might aggregate the evolved model in a privacy-preserving setting towards better generalization (McMahan, Moore, Ramage, Hampson, & y Arcas, 2017).

9. Conclusion

In this research work, we propose, demonstrate, and validate a novel approach to extract continuous PPG signals from the raw video signals of different modalities (RGB, near-infrared, and RGB demosaiced). We propose a MTL approach to train a robust model in presence of variance of target wearable PPG sensors. Moreover, we depict two transfer learning schemes to train the network faster and ensure better convergences with limited training data in the target domain. We evaluate our end-to-end rPPG reconstruction system with our MPSC-rPPG and two public datasets. Our results show that the *CamSense* architecture learns to consistently extract PPG from facial skin videos. We believe our proposed contact-less heart activity sensing and measuring framework can be useful to minimize the interactions between caregivers and patients.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by NSF CAREER grant 1750936, U.S. Army grant W911NF2120076, ONR, United States grant N00014-18-1-2462, and Alzheimer's Association, United States grant AARG-17-533039.

References

Abbas, A. K., Heimann, K., Jergus, K., Orlikowsky, T., & Leonhardt, S. (2011). Neonatal non-contact respiratory monitoring based on real-time infrared thermography. *Biomedical Engineering Online*, *10*(1), 93.

Adib, F., Mao, H., Kabelac, Z., Katabi, D., & Miller, R. C. (2015). Smart homes that monitor breathing and heart rate. In Proceedings of the 33rd annual acm conference on human factors in computing systems (pp. 837–846).

Alqaraawi, A., Alwosheel, A., & Alasaad, A. (2016). Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach. *Healthcare Technology Letters*, *3*(2), 136–142.

Arnold, J. M., Fitchett, D. H., Howlett, J. G., Lonn, E. M., & Tardif, J.-C. (2008). Resting heart rate: a modifiable prognostic indicator of cardiovascular risk and outcomes? *Canadian Journal of Cardiology*, *24*.

Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., & Dubois, J. (2019). Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, *124*, 82–90.

Bousefsaf, F., et al. (2019). 3D convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, *9*(20).

Castaneda, D., et al. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, *4*(4), 195.

Chen, W., & McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the European conference on computer vision (pp. 349–365).

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314.

De Haan, G., & Jeanne, V. (2013). Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, *60*(10), 2878–2886.

Demirezen, H., & Erdem, C. E. (2018). Remote photoplethysmography using nonlinear mode decomposition. In *ICASSP* (pp. 1060–1064). IEEE.

Deng, G.-F., Hung, Y.-S., Ho, W., & Lin, H.-H. (2017). Remote measurement of infant emotion via heart rate variability.

Draghici, A. E., & Taylor, J. A. (2016). The physiological basis and measurement of heart rate variability in humans. *Journal of Physiological Anthropology*.

Ganin, Y., & Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In International conference on machine learning (pp. 1180–1189).

Girshick, R., et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR* (pp. 580–587).

Glasmachers, T. (2017). Limits of end-to-end learning. arXiv preprint arXiv:1704.08305.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings (pp. 249–256).

Gudi, A., Bittner, M., Lochmans, R., & van Gemert, J. C. (2019). Efficient real-time camera based estimation of heart rate and its variability. In 2019 IEEE/CVF international conference on computer vision workshop. (pp. 1570–1579).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *2*, In *CVPR* (pp. 1735–1742). IEEE.

Hasan, Z., & Haque, M. A. (2015). Robust heart rate estimate from PPG signal utilizing the harmonic analysis of motion signal. In *TENCON 2015 - 2015 IEEE region 10 conference* (pp. 1–4). http://dx.doi.org/10.1109/TENCON.2015.7372992.

Hu, J., et al. (2019). Illumination robust heart-rate extraction from single-wavelength infrared camera using spatial-channel expansion. In *EMBC* (pp. 3896–3899). IEEE.

Huynh, S., Balan, R. K., Ko, J., & Lee, Y. (2019). VitaMon: measuring heart rate variability using smartphone front camera, In Proceedings of the 17th conference on embedded networked sensor systems (pp. 1–14).

Long, M., et al. (2015). Learning multiple tasks with multilinear relationship networks. arXiv preprint arXiv:1506.02117.

Macwan, R., Benezeth, Y., & Mansouri, A. (2019). Heart rate estimation using remote photoplethysmography with multi-objective optimization. *Biomedical Signal Processing and Control*, *49*, 24–33.

Magdalena Nowara, E., et al. (2018). Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *CVPR* (pp. 1272–1281).

McDuff, D., Gontarek, S., & Picard, R. W. (2014). Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera. *IEEE Transactions on Biomedical Engineering*, *61*(12), 2948–2954.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273–1282). PMLR.

Mirvis, D. M., & Goldberger, A. L. (2001). Electrocardiography. *Heart Disease*, *1*, 82–128.

Nowara, E. M., et al. (2018). SparsePPG: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *2018 IEEE/CVF CVPRW* (pp. 1353–135309). IEEE.

Osman, A., et al. (2015). Supervised learning approach to remote heart rate estimation from facial videos. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), Vol. 1* (pp. 1–6). IEEE.

Poh, M.-Z., et al. (2010). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, *18*(10), 10762–10774.

Publishing, H. H. (2020). *How's your heart rate and why it matters?*. Harvard Health, URL https://www.health.harvard.edu/heart-health/hows-your-heart-rate-and-why-it-matters.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Schroff, F., et al. (2015). Facenet: A unified embedding for face recognition and clustering. In *CVPR* (pp. 815–823).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *2015 IEEE conference on computer vision and pattern recognition* (pp. 1–9). http://dx.doi.org/10.1109/CVPR.2015.7298594.

Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., Clifton, D., & Pugh, C. (2014). Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement*, *35*(5), 807.

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics Express*, *16*(26), 21434–21445.

Wang, W., et al. (2016). Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, *64*(7), 1479–1491.

Wang, W., et al. (2019a). Discriminative signatures for remote-PPG. *IEEE Transactions on Biomedical Engineering*, *67*(5), 1462–1473.

Wang, Z.-K., et al. (2019b). Vision-based heart rate estimation via a two-stream CNN. In *2019 IEEE international conference on image processing* (pp. 3327–3331). IEEE.

Wu, B.-F., et al. (2018). Motion resistant image-photoplethysmography based on spectral peak tracking algorithm. *IEEE Access*, *6*, 21621–21634.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? arXiv preprint arXiv:1411.1792.

Zhan, Q., et al. (2020). Analysis of CNN-based remote-PPG to understand limitations and sensitivities. *Biomedical Optics Express*.

Zhang, D., et al. (2016). Association between resting heart rate and coronary artery disease, stroke, sudden death and noncardiovascular diseases: a meta-analysis. *Cmaj*, *188*(15), E384–E392.

Zhang, Q., et al. (2018). Heart rate extraction based on near-infrared camera: Towards driver state monitoring. *IEEE Access*, *6*, 33076–33087.

Zhao, C., et al. (2018). A novel framework for remote photoplethysmography pulse extraction on compressed videos. In *CVPR* (pp. 1299–1308).