# Remote Pulse Estimation in the Presence of Face Masks

Jeremy Speth, Nathan Vance, Patrick Flynn, Kevin Bowyer, Adam Czajka
University of Notre Dame
{jspeth, nvance1, flynn, kwb, aczajka}@nd.edu

## Abstract

*Remote photoplethysmography (rPPG), a family of techniques for monitoring blood volume changes, may be especially useful for contactless health monitoring via face videos from consumer-grade cameras. The COVID-19 pandemic caused widespread use of protective face masks, which results in a domain shift from the typical region of interest. In this paper we show that augmenting unmasked face videos by adding patterned synthetic face masks forces the deep learning-based rPPG model to attend to the periocular and forehead regions, improving performance and closing the gap between masked and unmasked pulse estimation. This paper offers several novel contributions: (a) deep learning-based method designed for remote photoplethysmography in a presence of face masks, (b) new dataset acquired from 54 masked subjects with recordings of their face and ground-truth pulse waveforms, (c) data augmentation method to add a synthetic mask to a face video, and (d) evaluations of handcrafted algorithms and two 3D convolutional neural network-based architectures trained on videos of unmasked faces and with masks synthetically added.*

## 1. Introduction

Remote pulse estimation is especially useful in settings where health diagnostics are desired, but using contact sensors is expensive, presents some risk, or professional sensors (*e.g.* pulse oximeters) are not available. The COVID-19 pandemic is one such scenario where extracting cardiac diagnostics without surface contact mitigates the risk of viral transmission, and can potentially allow for ubiquitous health monitoring at a critical time. Contactless kiosks could be used to screen entrants to public spaces to minimize breakout events. Furthermore, camera sensors are becoming increasingly common in personal devices, and health monitoring could be performed easily from mobile phones or laptops for telehealth applications.

The widespread adoption of face mask usage caused significant problems for existing technologies that assume an unobstructed view of the face [21]. People from certain religious denominations also occlude the face with coverings. Nearly all recent rPPG algorithms extract the signal from the face [5, 7, 11, 16, 19, 29, 36, 44, 47], sometimes even limiting the analyzed region to the cheeks [17, 40], which is a region generally occluded by a mask. While early contactless pulse estimation algorithms used hand-crafted features in both the temporal and spatial domains, more recent works have shown that convolutional neural networks (CNN) fed with spatiotemporal representations may outperform handcrafted approaches [5, 23, 28, 36, 39, 47]. We thus select two 3DCNN-based architectures for the models evaluated in this work, as shown in Fig. 1.

To accommodate the research community's need for large-scale realistic physiological datasets, we present a **new Masked Physiological Monitoring (MPM) dataset** of face recordings with masked subjects. High resolution videos at 90 frames per second were simultaneously recorded with oximeter pulse waveforms from 54 subjects. In this paper, we use the MPM dataset to analyze the effects of realistic face occlusions on rPPG algorithms, and we answer the following three **research questions**:

(Q1) Is the accurate pulse rate estimation possible on subjects wearing masks?

(Q2) If the answer to (Q1) is affirmative, does inclusion of face videos with synthetic masks result in better performance on videos of subjects wearing actual masks?

(Q3) What adaptations to the existing rPPG methods are useful to fine-tune them to COVID-19 and future health crises?

To the authors' knowledge, this is the first paper to explore the effects of face occlusions on the accuracy of remote pulse estimation algorithms. Along with the MPM dataset, we also offer source codes of the method adding synthetic face masks to existing (unmasked) face videos.

## 2. Background and Related Work

Remote photoplethysmography is the process of estimating the blood volume pulse from changes in reflected light

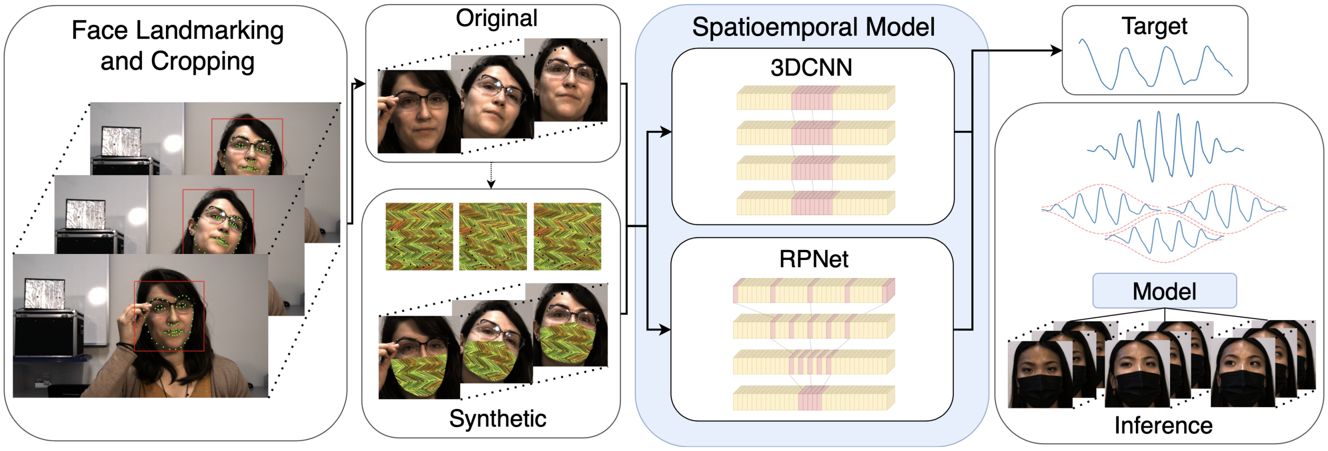


Figure 1. Training and inference pipeline for the spatiotemporal modeling task of remote pulse estimation. Raw RGB frames are landmarked and cropped, and then either fed directly into the model or a synthetic mask is added. Multiple frame sequences are overlap-added [7] to produce the full pulse waveform.

from the skin. Microvasculature beneath the skin's surface fills with blood, which changes reflected color due to the optical absorption of hemoglobin. Estimation from visible light is especially desirable due to low-cost sensors, but thermal [10,27,33,38] and near-infrared [25,26,41] sensors are also effective. In practice, the color changes in visible-light are subtle and may be obscured by noise factors such as illumination changes and body movements. The problem is further compounded by face coverings, which decreases surface area available to detect the pulse, and consequently, the signal-to-noise ratio.

Early studies began with stationary subjects and manually selected skin regions [42, 46]. Poh *et al.* [29, 30] applied blind source separation through independent component analysis (ICA) to the color channels. Several advancements combined color channels to locate the pulse signal [7, 8, 43–45]. The first approach considered the chrominance signal (CHROM), which was agnostic to illumination and robust to movement [7]. Later improvements relaxed assumptions on the distortion signals from movement [8], examined rotation of the skin pixels' subspace [45], and dynamically selected projections based on signal strength [43]. Wang *et al.* [44] introduced plane-orthogonal-to-skin (POS), which defines a projection plane for separating the specular and pulse components.

Until Li *et al.* [17] designed an effective pulse detector on the MAHNOB-HCI database [35], many approaches had been designed and tested on relatively small private datasets. After using the public MAHNOB-HCI dataset many groups were able to compare their estimators [5, 17, 40, 47], and it spurred the creation of more public datasets such as AFRL [9], MMSE-HR [40], VIPL-HR [22], and UBFC-RPPG [2]. The increased size of datasets made it possible to train deep neural networks. The first deep learning approach [15] trained a regression model on ICA and chrominance features.

Later, deep learning models for rPPG were trained on the spatial [5, 14] and spatiotemporal [23, 28, 36, 39, 47] dimensions of the video rather than extracted temporal features alone. Hsu *et al.* [14] trained VGG-15 on the frequency representation of averaged color signals. Chen *et al.* [5] used two-stream networks [34] frame differences and raw frames to a two-stream architecture, predicting the waveform derivative. A recent approach fed spatial-temporal maps from a grid of facial regions into ResNet-18 followed by a gated recurrent unit (GRU) to predict heart rate [23].

Yu *et al.* [47] constructed a 3DCNN that takes video clips as input and minimizes the negative Pearson correlation between waveforms. An advantage is the network's capability of producing a waveform, rather than a single heart rate value. Speth *et al.* [36] created the RPNet architecture by adding temporal dilations based on empirical results from frame rate experiments. We use both the 3DCNN and RPNet in our experiments. We improve the 3DCNN architecture by increasing the width of temporal kernels, such that longer-range time dependencies can be captured.

A later extension added enhancement and attention networks to help with compressed video [48]. Lee *et al.* [16] presented a transductive learner to adapt quickly to new samples. Disentangled representations were used to separate non-physiological signals from the pulse signal [24]. Liu *et al.* [19] further improved a DeepPhys-like architecture to form their MTTS-CAN model. Gideon *et al.* [11] recently performed unsupervised learning for rPPG.

While rPPG has been used for many applications such as presentation attack detection [13, 18] to distinguish between no pulse detected (presentation attack) and pulse detection (live), our goal is to determine how accurately the pulse rate can be estimated from a known live face wearing a mask. **Face occlusions have never been explored in rPPG**.

## 3. Datasets

We utilized three remote physiological monitoring datasets. The first is a large-scale publicly available dataset [37] recorded with unmasked subjects. The second is an augmentated version of the DDPM dataset with synthetic face masks. The third is a newly-collected dataset to assess remote pulse estimation algorithms in the presence of face masks.

**DDPM Dataset [37]** We used the publicly available *Deception Detection and Physiological Monitoring (DDPM)* dataset, which consists of 86 simultaneous video and pulse recordings of nearly 11 minutes per subject. During the recording, a paid actress conducted an interview consisting of 24 questions. Each subject was instructed beforehand to answer particular questions truthfully or deceptively. Subjects were free to complete the interview without constraints on motion, facial expressions, and talking, which accurately represents scenarios for pulse estimation in the wild. The interview setting also introduced variability in the pulse rate, as shown in Fig. 2. Such variability is rarely observed in rPPG datasets, and thus overall, the DDPM dataset's size and setting make it useful for our analyses.

**DDPM-Mask Dataset.** We augment the DDPM dataset with synthetic face masks by occluding the lower face region to create the *DDPM-Mask* corpus. We use the same landmarks selected in [21] to define a wide, medium coverage mask. We use two landmarkers: the OpenFace (OF) toolkit [1] and Bulat *et al*.'s 2D landmarker [4]. Along with black masks, we also added patterned masks by randomly selecting images from the Describable Textures Dataset (DSD) [6] and overlaying the image onto the 2D mask. The pattern was transformed with head rotation and translation to cover the same portions of the masked region. We first resized the pattern to 64 × 64 pixels. Then we randomly translated the pattern image such that the face landmarks for the first frame of the sequence were still within the pattern. Using these landmark points as anchors on the pattern image, we estimated the similarity transformation (rotation, translation, and scaling) from the anchor landmarks to the face landmarks in every following frame of the sequence, then applied the transformations on the pattern image before adding the masked region to the face frames. The second column of Fig. 1 illustrates a patterned synthetic mask added to the DDPM dataset over a sequence of frames.

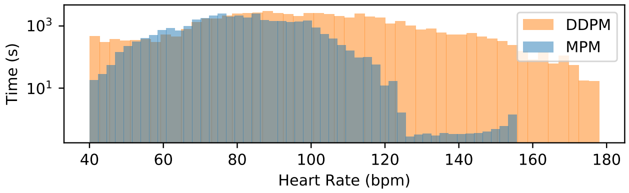


Figure 2. Distribution of heart rates over the MPM and DDPM [37] datasets. The deception interview setting for the DDPM results in a higher variance in heart rates. Note that the y-axis is log-scale.

**MPM Dataset.** We collected a new *Masked Physiological Monitoring (MPM)* dataset for remote physiological monitoring of *masked* subjects. A plexiglass screen was placed between acquisition personnel and the subject to reduce COVID-19 transmission risk. Subjects were asked to bring 3 different face masks to increase variability in color, texture, and shape. Subjects sat approximately 1 to 2 meters from the RGB camera. The ground truth heart rate, blood oxygenation, and blood volume pulse waveforms were collected by a Contec CMS50EA finger oximeter recording at 60 Hz. RGB videos were recorded with 1920 × 1080 pixels at 90 frames per second (fps) by TheImagingSource DFK 33UX290 camera. Videos were losslessly compressed with H.264 encoding using a constant rate factor of 0 to avoid damaging the optical pulse signal and allow for future compression studies. The MPM dataset was collected from consenting subjects under a human subjects research protocol approved by the authors' Human Subjects Institutional Review Board.

We captured videos for 54 subjects over 3 different sessions, where the participant wore a different mask in each recording. We divided each session into three different tasks: (a) natural conversation with free head movement, (b) directed head movement, and (c) frontal view without head movement. The natural conversational task consisted of sustained interaction with an acquisition worker for 2 minutes. The directed head movement task aimed to stress the pulse estimation algorithms by adding nonfrontal gaze and head motion. Subjects were directed to look at a total of 6 different targets for approximately 5 seconds each, resulting in a 30 second interval. The final task consisted of the subject maintaining frontal gaze and avoiding movement or talking for 30 seconds. Three subjects were only recorded for 2 sessions, resulting in a total of 159 recordings, over 3 minutes in length per video, giving us around 8 hours of recorded data. The reliability of the ground truth physiological signals was improved by using two Contec CMS50EA oximeters placed on both index fingers. A copy of the MPM dataset can be requested at `https://cvrl.nd.edu/projects/data/`.

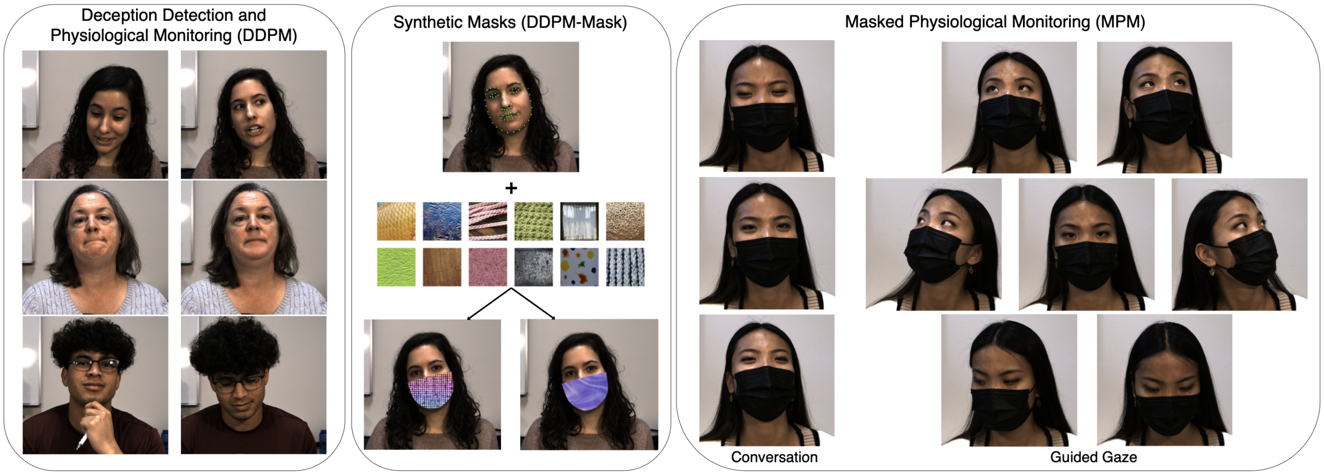

Deception Detection and Physiological Monitoring (DDPM)
Synthetic Masks (DDPM-Mask)
+
Masked Physiological Monitoring (MPM)
Conversation
Guided Gaze

Figure 3. Frames from all pulse databases used throughout this paper are shown for the unmasked, synthetically masked, and masked videos. Patterns for the synthetic masks are randomly sampled from the Describable Textures Dataset [6].

# 4. Approach

We model the pulse prediction task as a regression problem with the blood volume waveform from the oximeter as the target. This task generates a real value for every image in a sequence. We deploy two 3DCNN architectures on cropped frame sequences from the original video to contain the face only, as shown by Fig. 1. The following sections describe the model architectures and pipeline used to prepare the videos and target waveforms.

## 4.1. Spatiotemporal Architectures

We select two 3DCNN architectures to learn the pulse waveform from frame sequences. The 3DCNN was selected for three reasons. Firstly, it is capable of producing a high-resolution pulse waveform, not only selected statistics such as heart rate. Second, the 3DCNN is capable of learning from the raw image sequences. Lastly, the rPPG task benefits from joint learning of spatiotemporal features.

The first architecture is similar to the PhysNet-3DCNN [47], but with modified temporal dimensions of the kernels from a width of 3 to a width of 9 to capture longer time dependencies and help filter out high-frequency noise. The second is the RPNet architecture [36], which increases the temporal receptive field with dilated convolutions. We selected RPNet due to its good performance compared to the aforementioned 3DCNN with a kernel width of only 5, resulting in fewer parameters.

## 4.2. Preprocessing

To make the rPPG task easier for the model, we cropped the face region from all frames. We used both the OpenFace (OF) toolkit [1] and the 2D landmarker of Bulat *et al.* (AB) [4]) to detect 68 facial landmarks as a basis for defining the bounding box. Using two landmarkers allows us to investigate reliance of the results on the landmarker. Defining a bounding box from landmarks results in less jitter over time than simpler face detection methods. Additionally, the face landmarks gave us keypoints to approximate the shape and location of a synthetic mask.

From the minimum and maximum $(x, y)$ landmark locations we extended the sides and bottom by 5% and the top by 30% to include the forehead. We then extended the shorter of the two axes to the length of the other to form a square. The cropped region was then downsized to $64 \times 64$ pixels with bicubic interpolation. The model is given clips of the video consisting of 135 frames (1.5 seconds). We selected this as the minimum length of time an entire heartbeat would occur, considering 40 beats per minute (bpm) as the lowest frequency for healthy subjects.

For the DDPM dataset, we used the upsampled oximeter waveforms which had been phase shifted to match CHROM's waveforms [36]. For MPM, the oximeters recorded ground truth waveform values at 60 Hz, which differed from the native 90 fps of the videos. We upsampled the ground truth waveforms with cubic interpolation to the video timestamps. During training, we apply min-max normalization to the waveform labels for each clip to keep the values in [0,1].

## 4.3. Video Augmentation

We augment the input data by horizontal flipping with 50% probability, adding random illumination changes with mean of zero and standard deviation of 10 when operating on 8-bit grayscale images, and adding pixel-wise Gaussian noise $\sim \mathcal{N}(0, 4)$. The image values are subsequently scaled to floating point values between 0 and 1. We augment every frame within each video clip in the same manner.

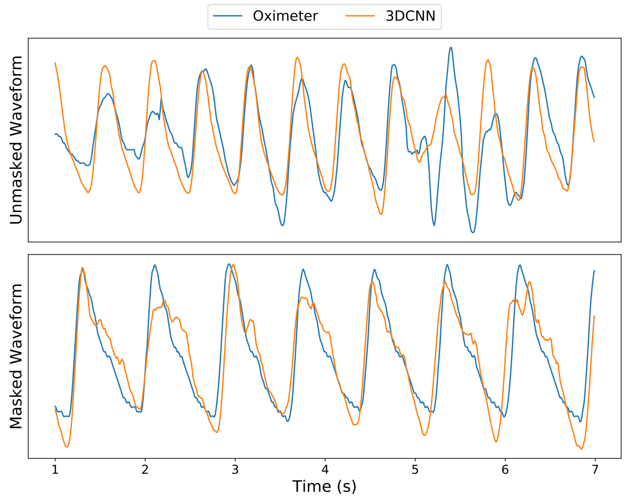


Figure 4. Ground truth and predicted waveforms for a short time segment on the unmasked (top) and masked (bottom) datasets using the same 3DCNN trained on subjects without face occlusions.

### 4.4. Optimization and Training

We optimize the 3DCNN for the temporal regression problem by minimizing the negative Pearson correlation between waveforms [36, 39, 47], each of the length of 135 frames. We apply the Adam optimizer without weight decay, with a learning rate of $\alpha = 0.0001$, and parameter values of $\beta_1 = 0.99$ and $\beta_2 = 0.999$. We apply dropout during training with 75% probability, since 3DCNNs are prone to overfitting. Example ground-truth and predicted waveforms are shown in Fig. 4. Visual inspection suggests that the trained 3DCNN models perform well on both masked and unmasked faces.

### 4.5. Overlap Adding

The model is given short video clips and predicts a waveform value for every frame. For videos longer than the clip length, it is necessary to perform predictions in sliding window fashion over the full video. Similar to [7], we use a stride of half the clip length to slide across the full video. For validation and testing we use a clip length of 136 frames to accommodate the half overlaps. The windowed outputs are first standardized, then a Hann function is applied to mitigate edge effects. Finally all overlapped outputs are summed to give a final waveform, as shown in Fig. 1 (right).

## 5. Experiments

Our experiments attempt to understand how face masks adversely affect remote pulse estimation performance, and whether adding synthetically-generated masks to face videos during training helps improve performance in the presence of real face masks. To give a complete evaluation, we evaluate all models on both the masked (MPM) and unmasked (DDPM) datasets, and with two different face landmarkers, in the following **four scenarios**:

(s1) training / tuning all methods on **unmasked** face videos (train/validation partition of DDPM), and testing also on **unmasked** face videos (test partition of DDPM),

(s2) training / tuning all methods on face videos with **synthetically added masks** (DDPM-Mask dataset), and testing on **unmasked** subject-disjoint face videos (test partition of DDPM),

(s3) training / tuning all methods on **unmasked** face videos (train/validation partition of DDPM), and testing on **masked** face videos (MPM dataset),

(s4) training / tuning all methods on face videos with **synthetically added masks** (DDPM-Mask dataset), and testing on **masked** face videos (MPM dataset).

### 5.1. Dataset Partitions

We used the provided training, validation, and testing partitions provided with the unmasked DDPM dataset. In total, there were 64 subjects used for training, another 11 subjects for validation, and the remaining 11 for testing. Splits were crafted with stratified random sampling across race, gender, and age, in order of importance in the cases that equal splits were not possible. By setting a portion of the unmasked data (DDPM) aside for testing, we can effectively examine the change in performance when evaluating on the entire masked (MPM) dataset of 54 subjects.

### 5.2. Compared Methods

We selected several previous state-of-the-art algorithms to evaluate the efficacy of our approach. All hand-crafted methods evaluated in the paper, including chrominance-based (CHROM) [7] and plane-orthogonal-to-skin (POS) [44], were reimplemented by us with minor help from components of Heusch *et al*. [12].

Two algorithms employing blind-source separation of the color channels through independent component analysis (ICA) [29, 30] were also tested, due to their initial popularity in the field. For simplicity, we refer to the ICA approach presented in [30] as POH10, and refer to the improved ICA approach with detrending [29] as POH11. Both of the ICA approaches perform spatial averaging on the cropped facial region after applying a face detector. We apply OpenFace and use the landmarks to define the region of interest in the same protocol presented in section 4.2.

We use the previously described 3DCNN architectures as examplars for deep learning approaches. Given the output waveforms from the 4 handcrafted approaches, in addition to the 3DCNN and RPNet trained on DDPM with OpenFace

(OF) and Bulat (AB) landmarking approaches (denoted in the results as 3DCNN OF and 3DCNN AB), DDPM-Mask with black synthetic masks from both landmarkers (3DCNN OF+B and 3DCNN AB+B), and DDPM-Mask with patterned synthetic masks from both landmarkers (3DCNN OF+P and 3DCNN AB+P), we evaluate each method in the same manner for a fair comparison. The efforts described above related to acquisition, re-implementation of various rPPG methods, and their evaluation may be regarded as a strong comparison for modern rPPG methods.

### 5.3. Evaluation Metrics

We evaluated the performance in the frequency domain (associated with the pulse rate). The errors are calculated between the oximeter and predicted heart rates, defined by the dominant frequency in the waveform for short time periods. The window size for estimating the dominant frequency can significantly affect the evaluation [20]. Since the time window used to predict heart rate within the oximeter is unknown, we calculate the ground truth heart rate frequencies from the oximeter's waveforms. We use a 30 second sliding window and apply a Hamming window prior to converting the signal to the frequency domain with the Fast Fourier Transform (FFT). The frequency with the maximum spectral peak between $0.\overline{66}$ Hz and 3 Hz (40 bpm to 180 bpm) is selected as the heart rate. A five-second moving average filter is applied to the resultant heart rate to smooth noisy regions containing finger movement. To compare the heart rate estimates, we used standard metrics from the rPPG literature, such as mean error (ME), mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient ($r_f$).

Since two oximeters are present in the masked dataset, we perform the same procedure over both waveforms and average the heart rate value at each time step. In a very small number of cases, the noise from hand movement gave different heart rate values from the oximeters. To remove these portions, if the heart rates differed by more than 10 beats per minute, the calculated heart rate closest to the average of the original heart rate estimates from the oximeters was selected. The resultant signals were smoothed with a three-second moving average filter to avoid spurious jumps in the heart rate. A subset of the samples were manually verified by calculating peak-to-peak distances.

Table 1. Pulse rate estimation comparison when the methods are tested on videos without face masks (scenarios s1 and s2). "OF" and "AB" denote two different face landmarkes: OpenFace and Bulat *et al.*'s, respectively. "+B" and "+P" denote black mask and patterned synthetic masks added to the training data, respectively.

| Method | ME (bpm) | MAE (bpm) | RMSE (bpm) | $r_f$ (bpm) |
|---|---|---|---|---|
| CHROM [7] | -0.26 | 3.48 | 10.37 | 0.93 |
| POS [44] | **0.11** | 3.16 | 11.19 | 0.92 |
| POH10 [30] | 18.54 | 20.56 | 33.10 | 0.56 |
| POH11 [29] | 10.47 | 14.30 | 28.86 | 0.54 |
| 3DCNN OF [47] | -1.18 | **1.96** | **6.99** | **0.97** |
| 3DCNN AB [47] | -1.25 | **1.96** | 7.17 | **0.97** |
| RPNet OF [36] | -1.18 | 2.09 | 7.30 | **0.97** |
| RPNet AB [36] | -1.21 | 2.05 | 7.22 | **0.97** |
| 3DCNN OF+B | -1.18 | 2.06 | 7.29 | **0.97** |
| 3DCNN OF+P | -1.27 | 2.00 | 7.29 | **0.97** |
| 3DCNN AB+B | -1.16 | 2.03 | 7.28 | **0.97** |
| 3DCNN AB+P | -0.81 | 2.30 | 7.76 | 0.96 |
| RPNet OF+B | -0.93 | 2.23 | 7.61 | 0.96 |
| RPNet OF+P | -1.16 | 2.07 | 7.31 | **0.97** |
| RPNet AB+B | -1.06 | 2.17 | 7.48 | **0.97** |
| RPNet AB+P | -1.07 | 2.20 | 7.35 | **0.97** |

Table 2. Same as in Tab. 1 except that the methods are tested on videos with face masks (scenarios s3 and s4).

| Method | ME (bpm) | MAE (bpm) | RMSE (bpm) | $r_f$ (bpm) |
|---|---|---|---|---|
| CHROM [7] | 3.52 | 12.59 | 16.34 | 0.03 |
| POS [44] | 16.24 | 19.27 | 26.79 | 0.14 |
| POH10 [30] | 25.83 | 27.26 | 33.08 | -0.01 |
| POH11 [29] | 38.74 | 38.76 | 41.16 | -0.05 |
| 3DCNN OF [47] | -1.59 | 3.73 | 9.65 | 0.77 |
| 3DCNN AB [47] | -2.40 | 3.99 | 10.60 | 0.75 |
| RPNet OF [36] | -2.00 | 4.60 | 10.85 | 0.73 |
| RPNet AB [36] | -2.98 | 4.74 | 11.76 | 0.71 |
| 3DCNN OF+B | -1.80 | 3.94 | 9.90 | 0.77 |
| 3DCNN OF+P | -2.00 | 3.91 | 9.69 | 0.78 |
| 3DCNN AB+B | -2.00 | 3.84 | 9.82 | 0.77 |
| 3DCNN AB+P | **-0.72** | **3.55** | **8.75** | **0.80** |
| RPNet OF+B | -1.80 | 3.93 | 9.55 | 0.78 |
| RPNet OF+P | -1.83 | 3.91 | 9.54 | 0.78 |
| RPNet AB+B | -2.48 | 4.14 | 10.84 | 0.74 |
| RPNet AB+P | -2.13 | 3.89 | 10.08 | 0.76 |

## 6. Results

**Scenario s1 (baseline): training and testing on *unmasked* face videos.** Performance for unmasked participants (DDPM) is shown in the **top portion of Table 1**. The two chrominance-based methods achieve lower mean error rates, showing that they are well calibrated for predicting heart rate and don't exhibit bias. Both ICA-based methods give worse performance than the chrominance and neural network approaches on every metric. The 3DCNN and RPNet models contain slightly higher mean error rates from bias than the chrominance models, but perform remarkably well in terms of MAE and RMSE. The choice of landmarker does not appear to significantly affect performance.

**Scenario s2: training on face videos *with synthetic masks*, testing on *unmasked* face videos.** The results for models trained on synthetically masked participants

(DDPM-Mask) are shown in the **bottom portion of Table 1**. Error discrepancies between the black and patterned masks are negligible, considering they perform better on different metrics. We find that models trained with synthetic masks perform slightly worse than the models trained to use the entire facial region, but they still give very strong positive correlations with ground truth heart rates. We find that synthetic masks slightly decrease performance on unmasked subjects, since the models do not learn to use the cheeks and lower face during training.

**Scenario s3: training on *unmasked* face videos, testing on videos of faces wearing *real masks*.** Performance of the handcrafted methods, 3DCNN model, and RPNet model trained on DDPM and evaluated on masked subjects (MPM) are shown in the **upper portion Table 2**. As expected, performance degrades compared to maskless subjects, since the number of available skin pixels is decreased. The best MAE among the handcrafted methods is given by CHROM, with over 12 bpm – more than 3 times worse than on DDPM. Both RPNet and the 3DCNN model give significantly better performance than the chrominance and ICA approaches. Fortunately, correlation between heart rate predictions and ground truth remains strong with positive correlations $r_f > 0.7$ for all spatiotemporal models. For general purposes, the increase in error is likely not large enough to change an assessment of one's state of health, but improving performance to the unmasked baseline is desirable. The performance decrease indicates that face occlusions cause difficulties for all analysed approaches. Strangely, the spatiotemporal models exhibit a bias towards predicting a higher pulse rate, as shown by the mean error.

**Scenario s4: training on face videos *with synthetic masks*, testing on videos of faces wearing *real masks*.** The **lower portion of Table 2** shows the performance of the models trained on black (3DCNN OF+B, 3DCNN AB+B, RPNet OF+B, and RPNet AB+B) and patterned (3DCNN OF+P, 3DCNN AB+P, RPNet OF+P, and RPNet AB+P) synthetic masks. For the 3DCNN we don't find a consistent improvement with black synthetic masks, however, we see the best-performing approach is trained with patterned synthetic masks and Bulat *et al.*'s landmarker, giving the lowest ME, MAE, and RMSE, along with the highest correlation between heart rates. Interestingly, the patterned synthetic masks with the OpenFace landmarker do not help the model, showing that the choice of landmarker is important.

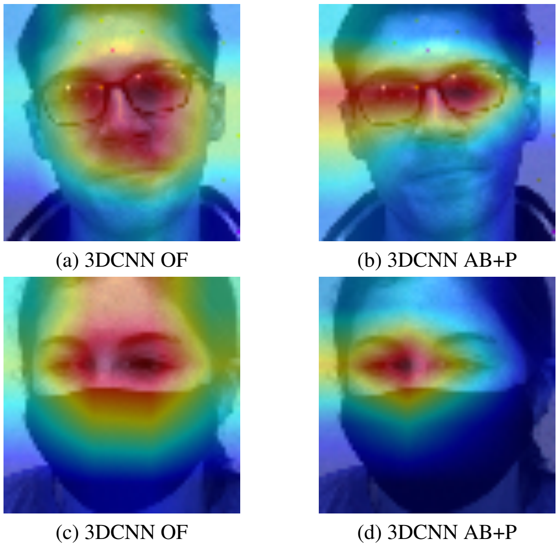
For the RPNet model we find that for both landmarkers the performance is improved when training with synthetic face masks. We find that the patterned masks give the greates improvement during training and reduce the RMSE by more than 1 bpm for both landmarkers. For both landmarkers the correlation is also increased by 0.05. Adding textured patterns to the masks seems to act as a useful augmentation strategy for the model, and helps refine the predictions to the periocular and forehead regions. Note that the 3DCNN has nearly twice as many parameters as RPNet, due to the wider temporal kernel width, but performance is comparable between the RPNet and 3DCNN models.

## 7. Discussion

**Visual and Anatomical Explanations.** We apply Grad-CAM [32] to visually explain the performance differences between the 3DCNNs. Since Grad-CAM is traditionally used for single images, we collected the pixel-wise sum over all images in a clip followed by normalization for image viewing. We then overlay the heatmap over the middle frame of the sequence. Figure 5 shows the attended spatial regions in the eighth convolutional layer for 3DCNN OF and 3DCNN AB+P, the best performing models for unmasked and masked subjects, respectively. The heatmaps clearly show that 3DCNN OF attends to the center of the face region, even when partially occluded by a face mask, while 3DCNN AB+P has learned to attend to the periocular region and forehead, since the synthetic masks during training occluded the lower face.

The anatomical reason for attention on the periocular region is likely the ophthalmic arteries, present in the canthi. Remote blood flow monitoring literature showed that this was visible in the infrared spectrum [33], but Fig. 5d indicates that it emits further across the spectrum. Such findings could be especially promising for ocular pulse detection from devices that are traditionally examining the iris, gaze, or pupil. Further, this is also promising for estimating the pulse from subjects whose religious denominations prescribe face coverings. Similarly, for the forehead, the likely culprits are the supraorbital vessels, which have also been documented in thermal imagery [49]. While we present a promising approach for restricting model attention to the periocular and forehead regions by synthetic occlusion during training, there is still room for improvement in guiding learning-based models to particular regions of the skin.

**Importance of Landmarker.** Our experimental results show the importance of selecting an accurate landmarker when generating the synthetically masked videos. Figure 6 shows erroneous face landmarks produced by the OF method when the subject gazed away from the cameras, which occurs frequently in the DDPM dataset due to the interview scenario. Interestingly, training on synthetic masks defined by the AB landmarker improves performance for the 3DCNN, but generally gives worse performance for RPNet. While the performance differences are minor, it could still indicate that learning-based rPPG estimators could benefit from jointly learning to predict the region of interest, as is done in a region proposal network [31].


(a) 3DCNN OF

(b) 3DCNN AB+P

(c) 3DCNN OF

(d) 3DCNN AB+P

Figure 5. Grad-CAM heatmaps for the top performing networks on DDPM (3DCNN OF) and MPM (3DCNN AB). The top and bottom rows show samples from the DDPM and MPM datasets, respectively. The attended regions for the model trained on DDPM covers a larger portion of the face than the the model trained on DDPM-Mask, which was guided to focus on the periocular and forehead regions. Images are scaled for viewing purposes.

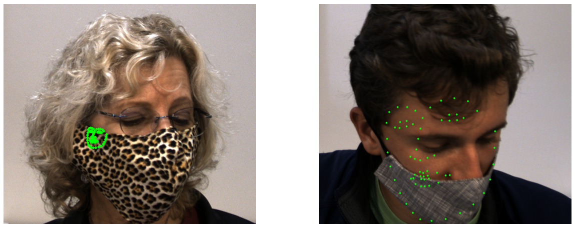


Figure 6. Face masks present difficulties to face detection and landmarking algorithms, as shown by the errors when using Open-Face on heavily occluded faces.

**Future Directions.** Training the proposed models with images of faces partially covered by synthetically-generated masks improved the performance of these models. Such data augmentation approach "removes" information anticipated to be missing in future face videos, and thus discourages the model from using these areas in inference. We hypothesise that instead of handcrafted "removal" of input information, an approach that guides the model towards salient regions by appropriate loss function formulation may be a promising future research direction. The salient regions can be sourced from automatic face image segmentation, or from human annotations. Independently of the source, this extra knowledge "where to look" can be used to penalize model's spatial attention to regions not associated with the task at hand [3].

## 8. Conclusions

In this paper, we present a new physiological monitoring dataset of high resolution RGB videos and oximeter recordings of subjects wearing masks to evaluate remote pulse estimators on masked people. In **answering the research questions** posed in the introduction, we find: (**re: Q1**) accurate pulse estimation is possible when subjects are wearing face masks, but the performance is slightly worse, (**re: Q2**) training with synthetically generated mask videos improves performance *when using robust face landmarkers*, and (**re: Q3**) face landmarkers and skin detectors robust to heavy face occlusion should be deployed in the early phases of the pulse detection algorithms to define reliable regions of interest. Several previous state-of-the-art pulse estimators built for unoccluded face video are found to perform worse on masked subjects, while the 3DCNN and RPNet spatiotemporal models exhibit only a moderate drop in performance. We find training the model with patterned synthetic masks created with accurate face landmarkers is sufficient to increase the robustness of pulse detection in the presence of masks and close the gap in performance.

## Acknowledgements

This research was sponsored by the Securiport Global Innovation Cell, a division of Securiport LLC. Commercial equipment is identified in this work in order to adequately specify or describe the subject matter. In no case does such identification imply recommendation or endorsement by Securiport LLC, nor does it imply that the equipment identified is necessarily the best available for this purpose. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of our sponsors.

## References

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. 3, 4

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 2

[3] Aidan Boyd, Patrick J. Tinsley, Kevin W. Bowyer, and Adam Czajka. CYBORG: Blending Human Saliency Into the Loss Improves Deep Learning. *CoRR*, abs/2112.00686, 2021. 8

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*, 2017. 3, 4

[5] Weixuan Chen and Daniel McDuff. DeepPhys: Video-Based Physiological Measurement Using Convolutional At-

tention Networks. *European Conference on Computer Vision (ECCV)*, pages 356–373, 2018. 1, 2

[6] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3, 4

[7] G. de Haan and V. Jeanne. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2, 5, 6

[8] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological Measurement*, 35(9):1913–1926, 2014. 2

[9] Justin R. Estepp, Ethan B. Blackford, and Christopher M. Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1462–1469, 2014. 2

[10] Marc Garbey, Nanfei Sun, Arcangelo Merla, and Ioannis Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Biomedical Engineering*, 54(8):1418–1426, 2007. 2

[11] John Gideon and Simon Stent. The Way to my Heart is through Contrastive Learning : Remote Photoplethysmography from Unlabelled Video. *IEEE International Conference on Computer Vision (ICCV)*, pages 3995–4004, 2021. 1, 2

[12] Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *CoRR*, abs/1709.00962, 2017. 5

[13] Guillaume Heusch and Sebastien Marcel. Pulse-based features for face presentation attack detection. *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018. 2

[14] G. Hsu, A. Ambikapathi, and M. Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 383–389, 2017. 2

[15] Yungchien Hsu, Yen Liang Lin, and Winston Hsu. Learning-based heart rate detection from remote photoplethysmography features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4433–4437, 2014. 2

[16] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rPPG: Remote Heart Rate Estimation Using a Transductive Meta-Learner. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[17] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4264–4271, 06 2014. 1, 2

[18] Si-Qi Liu, Xiangyuan Lan, and Pong Yuen. Remote Photoplethysmography Correspondence Feature for 3D Mask Face Presentation Attack Detection. In *European Conference on Computer Vision (ECCV)*, pages 577–594, 2018. 2

[19] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19400–19411. Curran Associates, Inc., 2020. 1, 2

[20] Yuriy Mironenko, Konstantin Kalinin, Mikhail Kopeliovich, and Mikhail Petrushan. Remote photoplethysmography: Rarely considered factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 6

[21] Mei Ngan, Patrick Grother, and Kayee Hanaoka. Ongoing Face Recognition Vendor Test (FRVT)Part 6A: Face recognition accuracy with masks using pre-COVID-19 algorithms. Technical Report NISTIR 8311, National Institute of Standards and Technology, July 2020. 1, 3

[22] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. VIPL-HR: A Multi-modal Database for Pulse Estimation from Less-constrained Face Video. In *Asian Conference on Computer Vision (ACCV)*, 2018. 2

[23] 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, 2020. 1, 2

[24] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[25] Ewa Magdalena Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan. SparsePPG: Towards Driver Monitoring Using Camera-Based Vital Signs Estimation in Near-Infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1353–135309, 2018. 2

[26] Sang Bae Park, Gyehyun Kim, Hyun Jae Baek, Jong Hee Han, and Joon Ho Kim. Remote pulse rate measurement from near-infrared videos. *IEEE Signal Processing Letters*, 25(8):1271–1275, 2018. 2

[27] I. Pavlidis, J. Dowdall, N. Sun, C. Puri, J. Fei, and M. Garbey. Interacting with human physiology. *Computer Vision and Image Understanding*, 108(1):150–170, 2007. Special Issue on Vision for Human-Computer Interaction. 2

[28] Olga Perepelkina, Mikhail Artemyev, Marina Churikova, and Mikhail Grinenko. Hearttrack: Convolutional neural network for remote video-based heart rate monitoring. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1163–1171, 2020. 1, 2

[29] M. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011. 1, 2, 5, 6

[30] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, May 2010. 2, 5, 6

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M.

Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. Curran Associates, Inc., 2015. 7

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7

[33] Dvijesh Shastri, Panagiotis Tsiamyrtzis, and Ioannis Pavlidis. Periorbital thermal signal extraction and applications. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 102–105, 2008. 2, 7

[34] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press. 2

[35] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. 2

[36] Jeremy Speth, Nathan Vance, Adam Czajka, Kevin Bowyer, and Patrick Flynn. Unifying frame rate and temporal dilations for improved remote pulse detection. *Computer Vision and Image Understanding (CVIU)*, pages 1056–1062, 2021. 1, 2, 4, 5, 6

[37] Jeremy Speth, Nathan Vance, Adam Czajka, Kevin Bowyer, Diane Wright, and Patrick Flynn. Deception detection and remote physiological monitoring: A dataset and baseline experimental results. In *International Joint Conference on Biometrics (IJCB)*, pages 4264–4271, 2021. 3

[38] Nanfei Sun, Marc Garbey, Arcangelo Merla, and Ioannis Pavlidis. Imaging the cardiovascular pulse. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005. 2

[39] Yun-Yun Tsou, Yi-An Lee, Chiou-Ting Hsu, and Shang-Hung Chang. *Siamese-rPPG Network: Remote Photoplethysmography Signal Estimation from Face Videos*, page 2066–2073. Association for Computing Machinery, New York, NY, USA, 2020. 1, 2, 5

[40] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2396–2404, 2016. 1, 2

[41] Mark van Gastel, Sander Stuijk, and Gerard de Haan. Motion Robust Remote-PPG in Infrared. *IEEE Transactions on Biomedical Engineering*, 62(5):1425–1433, 2015. 2

[42] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, Dec 2008. 2

[43] Wenjin Wang, Albertus C. Den Brinker, and Gerard De Haan. Single-Element Remote-PPG. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2019. 2

[44] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 1, 2, 5, 6

[45] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation. *IEEE Transactions on Biomedical Engineering*, 63(9):1974–1984, 2016. 2

[46] F. P. Wieringa, F. Mastik, and A. F.W. Van Der Steen. Contactless multiple wavelength photoplethysmographic imaging: A first step toward ”SpO2 camera” technology. *Annals of Biomedical Engineering*, 33(8):1034–1041, 2005. 2

[47] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 1, 2, 4, 5, 6

[48] Zitong Yu*, Wei Peng*, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[49] Zhen Zhu, Panagiotis Tsiamyrtzis, and Ioannis Pavlidis. The segmentation of the supraorbital vessels in thermal imagery. In *International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 237–244, 2008. 7