MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Unrolled IPPG: Video Heart Rate Esitmation via Unrolling Proximal Gradient Descent

Shenoy, Vineet; Marks, Tim K.; Mansour, Hassan; Lohit, Suhas

TR2023-116 September 13, 2023

Abstract

Imaging photoplethysmography (iPPG) is the process of estimating a person's heart rate from video. In this work, we propose Un- rolled iPPG, in which we integrate iterative optimization updates with deep learning-based signal priors to estimate the pulse wave- form and heart rate from facial videos. We model the signal extracted from video as the sum of an underlying pulse signal and noise, but instead of explicitly imposing a handcrafted prior (e.g., sparsity in the frequency domain) on the signal, we learn priors on the signal and noise using neural networks. We solve for the underlying pulse sig- nal by unrolling proximal gradient descent; the algorithm alternates between gradient descent steps and application of learned denoisers, which replace handcrafted priors and their proximal operators. Using this method, we achieve state-of-the-art heart rate estimation on the challenging MMSE-HR dataset.

IEEE International Conference on Image Processing (ICIP) 2023

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

UNROLLED IPPG: VIDEO HEART RATE ESTIMATION VIA UNROLLING PROXIMAL GRADIENT DESCENT

Vineet R. Shenoy[*2] *Tim K. Marks*[1] *Hassan Mansour*[1] *Suhas Lohit*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Johns Hopkins University, MD, USA

ABSTRACT

Imaging photoplethysmography (iPPG) is the process of estimating a person's heart rate from video. In this work, we propose Unrolled iPPG, in which we integrate iterative optimization updates with deep learning-based signal priors to estimate the pulse waveform and heart rate from facial videos. We model the signal extracted from video as the sum of an underlying pulse signal and noise, but instead of explicitly imposing a handcrafted prior (e.g., sparsity in the frequency domain) on the signal, we learn priors on the signal and noise using neural networks. We solve for the underlying pulse signal by unrolling proximal gradient descent; the algorithm alternates between gradient descent steps and application of learned denoisers, which replace handcrafted priors and their proximal operators. Using this method, we achieve state-of-the-art heart rate estimation on the challenging MMSE-HR dataset.

Index Terms— Heart rate estimation, imaging photoplethysmography, remote photoplethysmography, unrolling algorithms.

1. INTRODUCTION

Recent years have witnessed increasing interest in non-contact monitoring of vital signs, particularly for telemedicine [1], including estimation of heart rate [2, 3, 4, 5], breathing rate [6, 7], and blood pressure [8] from video of the face. In addition to healthcare, remote monitoring can be used in safety-critical applications such as driving [9, 10] or heavy equipment operation. In this work, we estimate heart rate from facial video using the technique of Imaging Photoplethysmography (iPPG).

Prior work has shown that cameras recording facial videos capture the subtle changes in skin color corresponding to the blood volume pulse [4, 11, 12]. However, the blood volume pulse signal is a small fraction of the pixel intensity and can be easily masked by illumination changes and motion. We consider heart rate estimation systems in which estimation is predicated on extracting the weak and noisy pulse signal from the face, followed by denoising the resulting pulse wave.

Of particular interest are works such as SparsePPG [9] and AutoSparsePPG [10], which aim to model the pulse wave as a sparse signal in the Fourier domain and use variations of the Iterative Shrinkage Thresholding Algorithm (ISTA) to estimate this sparse signal from a noisy time series obtained from various facial regions. These methods model the signal as sparse in the frequency domain and noise as sparse in the spatio-temporal domain, based on two assumptions: (1) that the true pulse signal can be modeled using a sparse set of frequencies that are shared across face regions; and (2) that the noise primarily affects a small number of regions. Solving for the signal and noise components can be done via alternating gradient updates and soft-thresholding projection.

More recent methods are based on deep learning and use training data consisting of videos and their ground-truth waveforms. They are trained to map from either the face videos [13, 14, 15] or time series extracted from the videos [16] to the pulse wave.

In this work, we propose *Unrolled iPPG*, which combines the advantages of model-based iterative algorithms with the expressive power of deep neural networks. In particular, rather than explicitly enforcing signal sparsity in Fourier domain, we use deep neural networks to learn priors on the pulse signal and noise component. The proximal gradient descent algorithm is unrolled, and the traditional gradient updates are followed by two learned denoisers, one for the pulsatile signal and the other for the noise component. We achieve state-of-the-art results on the task of heart-rate estimation on the challenging MMSE-HR dataset [17], outperforming both purely handcrafted methods like AutoSparsePPG [10] and purely data-driven methods like InverseCAN [15].

2. RELATED WORK

2.1. Unrolling iterative algorithms to solve inverse problems

Unrolling algorithms integrate learnable parameters into traditional iterative algorithms such as FISTA [18], harnessing the power of learning while exploiting known structure and retaining interpretability. These algorithms repeatedly apply two steps: first, they ensure that the intended result is consistent with measurements by minimizing a data fidelity term using a learned or fixed forward operator, and second, they apply a signal denoiser – parametrized as a learnable function – to fit the solution to an implicit prior. Unrolling has been applied to a diverse set of inverse problems including super-resolution [19], blind image denoising [20], and multi-spectral image fusion [21]. Early work, called Learnable Iterative Shrinkage Thresholding Algorithm (LISTA) [22], aimed to improve sparse coding by learning the sparsifying basis upon which the sparse code is generated. Research in compressive sensing [23] specifically learned the proximal operators with neural networks using pixel losses and GAN-based perceptual losses to replace conventional sparsity terms. Similarly, our work learns a denoising operation to capture the structure of the pulse signal in the Fourier domain.

2.2. Heart rate estimation from videos

Remote estimation of the pulse wave and heart rate from videos, known as imaging photoplethysmography (iPPG) or remote photoplethysmography (rPPG), can be categorized into blind source separation methods, model-based methods, and data-driven methods.

[*]VS worked on this project when he was an intern at MERL.

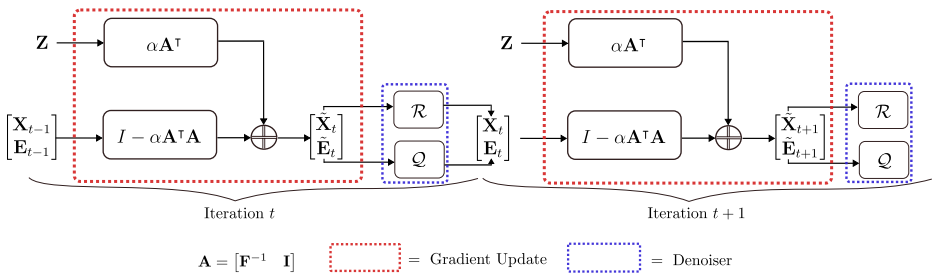


Fig. 1: Illustration of the proposed Unrolled iPPG algorithm. We show two unrolled iterations of the algorithm, which integrates gradient update steps with respect to a fixed forward model and learned neural network denoisers $\mathcal{R}$ and $\mathcal{Q}$.

Blind source separation [3, 4] assumes that the extracted time-series signal is composed of both the noise and the underlying pulse signal, and that these signals are either statisically independent or uncorrelated. Techniques such as Independent Component Analysis (ICA) and Principal Components Analysis (PCA) separate the signal from the noise. Several model-based algorithms [24, 25, 5] explicitly model light absorption and reflection in the skin; [25, 5] both aim to reduce the dependence of the extracted signal on the average skin reflection color. The model based methods SparsePPG and AutoSparsePPG [9, 10] model the underlying pulse waveform as the sum of a sparse set of periodic signals, and they aim to solve for the underlying pulse waveform. Data-driven methods [13, 14, 15, 16] learn directly from the training data and are typically instantiated as neural networks, whose parameters are learned by minimizing a loss between the output of the network and the ground-truth waveform. Methods such as [13, 15] take frames as input directly, learn which regions of each frame contain the signal using attention mechanisms, extract the signal, and reconstruct the underlying pulse waveform.

Our work incorporates data-driven methods into a model-based algorithm, achieving state-of-the-art performance while maintaining interpretability. Instead of imposing handcrafted sparsity constraints on the underlying pulse signal, we use deep neural networks to learn implicit priors from data.

3. METHOD

3.1. Signal model

For each input time window containing S frames, we first extract a time series containing the average intensity of each of K face regions in every frame. Stacking these signals into a matrix $\mathbf{Z} \in \mathbb{R}^{S\times K}$, we assume that these region-specific signals share a quasi-periodic pulse signal that admits a structured representation in the Fourier domain. Therefore, we can model the observation of the heartbeat signal as

$$\mathbf{Z} = \mathbf{Y} + \mathbf{E} = \mathbf{F}^{-1}\mathbf{X} + \mathbf{E}, \tag{1}$$

where $\mathbf{F}^{-1} \in \mathbb{C}^{S\times N}$ is the oversampled inverse Fourier Transform matrix, $\mathbf{Y} \in \mathbb{R}^{S\times K}$ and $\mathbf{X} \in \mathbb{C}^{N\times K}$ represent the pulsatile signal in all K regions in the time domain and the frequency domain, respectively, and $\mathbf{E} \in \mathbb{R}^{S\times K}$ is a noise component that captures the non-pulse-related fluctuations in the iPPG signals $\mathbf{Z}$.

AutoSparsePPG [10] formulated the recovery of $\mathbf{X}$ and $\mathbf{E}$ as a joint-sparse recovery problem:

$$\min_{\mathbf{X},\mathbf{E}} \frac{1}{2}\underbrace{\left\|\mathbf{Z}-\mathbf{F}^{-1}\mathbf{X}-\mathbf{E}\right\|_F^2}_{\text{Data Fidelity}=D} + \lambda\cdot\underbrace{(\|\mathbf{X}\|_{2,1}+\left\|\mathbf{E}^\intercal\right\|_{2,1})}_{\text{Regularization}}. \tag{2}$$

Here, the $\ell_{2,1}$ norm $\|\mathbf{X}\|_{2,1} = \sum_t \sqrt{\Sigma_j \mathbf{X}(t,j)^2}$ encourages the same small set of Fourier coefficients to be nonzero across multiple regions, while $\|\mathbf{E}^\intercal\|_{2,1}$ encourages the noise to be limited to a sparse set of regions. The data fidelity term, D, encourages the solution to be close to the observed data. We can rewrite it as:

$$D = \frac{1}{2}\left\|\mathbf{Z}-\mathbf{A}\begin{bmatrix}\mathbf{X}\\ \mathbf{E}\end{bmatrix}\right\|_F^2, \quad \text{where} \quad \mathbf{A}=[\mathbf{F}^{-1}\quad\mathbf{I}]. \tag{3}$$

3.2. Unrolled iPPG

While AutoSparsePPG [10] has shown good performance, the $\ell_{2,1}$ penalty function employed by AutoSparsePPG assumes that the nonzero frequency components are independent of each other. However, this assumption does not hold in general since the shape of the heart beat signal necessitates that a group of frequency bins should be nonzero together to describe the heart beat waveform. To that end, we train a deep denoiser to discover the appropriate structure in the Fourier domain. Instead of using explicit priors defined by $\ell_{2,1}$ regularization in Eq. 2, our contribution is to encode the signal and noise priors as an implicit penalty function $\rho(\cdot,\cdot)$ and employ its learnable scores $\mathcal{R}$ and $\mathcal{Q}$ as deep denoisers for $\mathbf{X}$ and $\mathbf{E}$, respectively. Using the same data fidelity formulation as in Eq. (3), we aim to solve:

$$\min_{\mathbf{X},\mathbf{E}} \frac{1}{2}\left\|\mathbf{Z}-\mathbf{A}\begin{bmatrix}\mathbf{X}\\ \mathbf{E}\end{bmatrix}\right\|_F^2 + \lambda\cdot\rho(\mathbf{X},\mathbf{E}), \tag{4}$$

where $\mathbf{A} = [\mathbf{F}^{-1}\quad\mathbf{I}]$. Unlike AutoSparsePPG [10], which solves Eq. (2) through alternating gradient update and soft-thresholding projection steps, we unroll proximal gradient descent in Eq. (4) for T iterations. In each iteration, we perform gradient updates on $\mathbf{X}$ and $\mathbf{E}$ followed by forward propagation through $\mathcal{R}$ and $\mathcal{Q}$. Given a step size α, the updates on $\mathbf{X}$ and $\mathbf{E}$ are given by

$$\begin{aligned}\begin{bmatrix}\tilde{\mathbf{X}}_{t+1}\\ \tilde{\mathbf{E}}_{t+1}\end{bmatrix} &= \begin{bmatrix}\mathbf{X}_t\\ \mathbf{E}_t\end{bmatrix} - \alpha\nabla_{\left[\begin{smallmatrix}\mathbf{X}\\ \mathbf{E}\end{smallmatrix}\right]} D\left(\begin{bmatrix}\mathbf{X}_t\\ \mathbf{E}_t\end{bmatrix}\right),\\ &= \alpha\mathbf{A}^\intercal\mathbf{Z} + (\mathbf{I}-\alpha\mathbf{A}^\intercal\mathbf{A})\begin{bmatrix}\mathbf{X}_t\\ \mathbf{E}_t\end{bmatrix},\\ \mathbf{X}_{t+1} &= R(\tilde{\mathbf{X}}_{t+1}),\\ \mathbf{E}_{t+1} &= Q(\tilde{\mathbf{E}}_{t+1}).\end{aligned} \tag{5}$$

Here, $\nabla_{\left[\begin{smallmatrix}\mathbf{X}\\ \mathbf{E}\end{smallmatrix}\right]} D\left(\begin{bmatrix}\mathbf{X}_t\\ \mathbf{E}_t\end{bmatrix}\right)$ is the gradient of D with respect to $\begin{bmatrix}\mathbf{X}\\ \mathbf{E}\end{bmatrix}$ evaluated at the desired point $\begin{bmatrix}\mathbf{X}_t\\ \mathbf{E}_t\end{bmatrix}$. The full algorithm

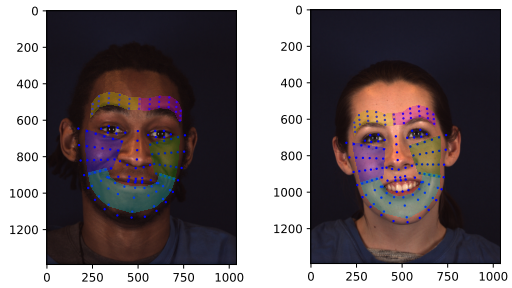


Fig. 2: Facial landmark detection and construction of regions. We segment the face into 5 regions (left cheek, right cheek, left forehead, right forehead, and chin) and spatially average the pixels to obtain the noisy 5-dimensional time series $\mathbf{Z}$.

is shown in Fig. 1. During training, we update the parameters of the neural networks $\mathcal{R}$ and $\mathcal{Q}$ to minimize the mean squared error loss between the output signal after T unrolled iterations, $\mathbf{Y}_T = \mathbf{F}^{-1}\mathbf{X}_T$, and the ground-truth waveform $\mathbf{Z}_{\text{gt}}$.

To find the heart rate, we sum the power in every frequency bin across all K regions, then select the frequency with the most power.

4. IMPLEMENTATION AND RESULTS

4.1. MMSE-HR dataset

The MMSE-HR dataset [17], with 23 female subjects and 17 male subjects, contains 102 videos capturing the face and simultaneous blood pressure wave from a finger sensor as various emotions are elicited. This results in substantial motion in some videos, to which our algorithm is robust. Videos were captured at a resolution of 1040×1392 at 25 frames per second, while the blood pressure wave was measured at 1000 samples per second. The ground-truth data are downsampled to match the frame rate of the videos. We train and evaluate on this dataset using leave-one-subject-out cross validation. For each of the 40 subjects, we test using a model that was trained on the other 39 subjects.

4.2. Time series extraction

We extract the time-series $\mathbf{Z}$ by first detecting the face in each RGB video frame using FaceBoxes [26]. Next, we use LUVLi [27] landmark localization and interpolate/extrapolate its 68-landmark output to 145 landmarks. These landmarks are grouped into 48 small spatial areas, in each of which we compute the mean pixel intensity of the Red and Green channels. Instead of using multiple color channels, we take the ratio of the Red and Green Channels for further processing [5]. We then group these small areas into $K = 5$ facial regions as indicated in Figure 2, taking the median intensity value of the areas within each facial region. This yields a 5-dimensional time series for each video. We then apply a Butterworth filter with cutoff frequencies [0.7, 2.5] Hz as in [15] to capture frequencies in a typical range of heart rates.

4.3. Neural network denoiser architecture and training details

In Unrolled iPPG, the learned neural network denoisers $\mathcal{R}$ and $\mathcal{Q}$ are modeled using an encoder-decoder architecture. Since the network $\mathcal{R}$ operates on Fourier coefficients $\mathbf{X}_t$, its network weights are complex-valued, while the network weights for $\mathcal{Q}$ (which operates on $\mathbf{E}_t$) are real-valued. The networks consist of two downsampling convolutional blocks with a stride of 2, in which the number of channels is increased from 5 to 32 and then from 32 to 64. The two upsampling blocks are implemented using transposed convolution with a stride of 2, decreasing the number of channels from 64 to 32 and from 32 to 5. Each convolutional layer has a kernel size of 16 and is followed by a ReLU nonlinearity and then a batch normalization layer. We initialize $\mathcal{Q}$ to output $\mathbf{0}$, and we initialize $\mathcal{R}$ to the identity transform by initializing the convolution layers to output $\mathbf{0}$ and adding a single skip connection at the highest convolution layer. The variable $\mathbf{X}$, which is input to $\mathcal{R}$, is initialized as the Fourier transform of $\mathbf{Z}$. The noise $\mathbf{E}$, which is input to $\mathcal{Q}$, is initialized as the $\mathbf{0}$ matrix. We calculate the mean squared error between each of the five output channels and the ground-truth, and update the network using the Adam [28] optimizer with a learning rate of 3×10^{-4} for 8 epochs. To be able to estimate heart rates at the lower and higher end that are not well represented in the dataset, we augment the training data using augmentations called "SpeedUp" and "SlowDown". For the "SlowDown" augmentation, an input window of length S is cropped by a random percentage between 20% and 40%, and interpolated back to the original window size S using linear interpolation. For the "SpeedUp" augmentation, given our window length S, we randomly chose an input window length that is 20% to 40% larger than our target time windows (e.g. $1.2 \times S$), and linearly interpolate it back to length S.

During training, we partition each empirical and ground-truth waveform into 10-second windows, then shift the window by 2.4 seconds to get the next partially overlapping window for training. The windows are loaded randomly during training with a batch size of 100. At test time, we reconstruct 10-second segments in a non-overlapping fashion.

4.4. Evaluation protocol

Following the protocol in [15], we report the mean absolute error (MAE) and root mean squared error (RMSE) of the ground-truth and predicted heart rate computed for 30-second time windows for the test videos. However, as our algorithm uses 10-second windows, we concatenate our output signal for three adjacent 10-second windows in order to evaluation on 30-second windows as in [15]. The MAE and RMSE metrics, which are averaged over all B windows for all test videos and over all test set partitions, are given by

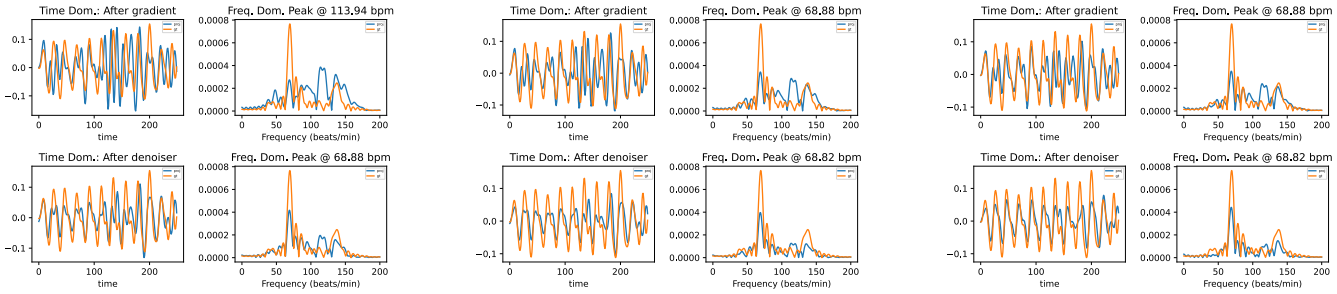
$$\text{MAE} = \frac{1}{B}\sum_{i=1}^{B}|R_i - \hat{R}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(R_i - \hat{R}_i)^2} \quad (6)$$

where $\hat{R}_i$ is the predicted heart rate and R_i is the ground-truth heart rate for the time window. We also report a metric that we call PTE6, the percent of time the heart rate error is less than 6 beats per minute (bpm), which is a way to measure how often the estimated heart rate is correct:

$$\text{PTE6} = \frac{1}{B}\sum_{i=1}^{B}P_i, \text{ where } P_i = \begin{cases} 1 & \text{if } |R_i - \hat{R}_i| < 6 \text{ bpm}, \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

4.5. Results

We compare our results to previous methods in Table 1. As shown, we signficantly outperform model-based methods such as AutoSparsePPG [10] by reducing the MAE error from 4.55 to 1.11, and the RMSE from 14.42 to 2.97. We also increase the PTE6 from



(a) Iteration 1 (b) Iteration 2 (c) Iteration 3

Fig. 3: Three iterations of the unrolling algorithm. The signals in orange are the ground-truth, with a peak frequency at 69.42 bpm, while the signals in blue are the (intermediate) outputs of our algorithm for one face region. For each interation t, the top row shows the estimate before $\mathcal{R}(\cdot)$, $\tilde{\mathbf{X}}_t$, and the bottom row shows the estimate after $\mathcal{R}(\cdot)$, $\mathbf{X}_t$. Each iteration shows the time-domain signals on the left and power spectrum on the right. The power spectrum plot titles show the peak frequency of each estimated signal. Viewing left to right, the blue signal progressively denoises from an incorrect estimation (113.94 bpm) to a reasonable estimation (68.82 bpm) for the ground-truth signal with peak frequency 69.42 bpm.

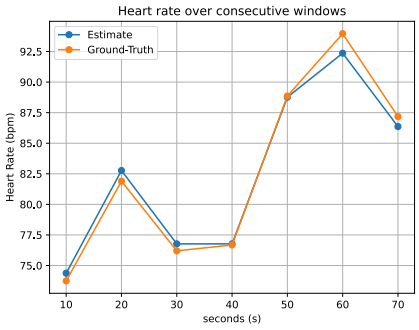


Fig. 4: A window-by-window heart-rate estimate (shown on 10-second windows for clarity). Our results show that we can accurately predict the heart rate from facial videos.

Table 1: Heart rate estimation results on the MMSE-HR dataset

Method	MAE (bpm) ↓	RMSE (bpm) ↓	PTE6 (%) ↑
ICA [29]	5.44	12.00	-
CHROM [5]	3.74	8.11	-
POS [24]	3.90	9.61	-
CAN [13]	4.06	9.51	-
InverseCAN [15]	2.27	4.90	-
AutoSparsePPG [10]	4.55	14.42	88.10
Unrolled iPPG (Ours)	**1.11** $\pm$ **0.01**	**2.97** $\pm$ **0.17**	**93.53** $\pm$ **0.73**

88.10 to 93.53. Given that we incorporate learned components, we compare our results to the data-driven methods CAN [13] and InverseCAN [15], outperforming both. Compared to InverseCAN, our Unrolled iPPG method reduces the MAE from 2.27 to 1.11 and the RMSE from 4.90 to 2.97. In Fig. 3, we show an example of how our algorithm iteratively estimates the underlying pulse signal and spectrum for one test video. An example of our performance on consecutive 10-second time windows is shown in Fig. 4, where we see that our Unrolled iPPG algorithm correctly predicts the ground-truth heart rate over a wide range of heart rates and across the duration of an entire video.

We also study the effect of not explicitly modeling the noise component $\mathbf{E}$ and not using $\mathcal{Q}$, i.e., solving the optimization problem

$$\min_{\mathbf{X}} \frac{1}{2} \left\| \mathbf{Z} - \mathbf{F}^{-1}\mathbf{X} \right\|_F^2 + \lambda \cdot \tilde{\rho}(\mathbf{X}). \quad (8)$$

Table 2: Heart rate estimation with and without modeling the noise component $\mathbf{E}$. Using $\mathcal{Q}$ to denoise $\tilde{\mathbf{E}}$ improves performance.

Method	MAE (bpm) ↓	RMSE (bpm) ↓	PTE6 (%) ↑
Without $\mathcal{Q}$ (Eq. (8))	1.44	3.87	92.25
With $\mathcal{Q}$ (Eq. (4))	**1.11**	**2.82**	**93.02**

Table 3: Heart rate estimation performance based on the number T of unrolling iterations

Number of iter.	MAE (bpm) ↓	RMSE (bpm) ↓	PTE6 (%) ↑
1	1.33	**3.12**	92.25
3	**1.11**	3.22	**93.02**
5	2.14	6.72	90.7
10	2.59	8.89	90.7

As shown in Table 2, modeling the noise explicitly using $\mathcal{Q}$ (instead of modeling the Fourier coefficients and noise simultaneously in $\mathcal{R}$) significantly improves the results. The network $\mathcal{Q}$ learns the structure of the noise, which can be subtracted from the signal $\mathbf{Z}$ in Eq. (1) to obtain the underlying pulse signal.

Finally, in Table 3 we analyze the effect of the number of iterations T over which we unroll, corresponding to the number of successive gradient and projection steps. The results show that $T = 3$ unrolled iterations provide the best performance.

5. CONCLUSION

In this work, we describe Unrolled iPPG, which unrolls a model-based sparse recovery algorithm for heart rate estimation from videos and achieves state-of-the-art results on the MMSE-HR dataset. We explicitly model the signal extracted from the face video as the sum of an underlying pulse signal and noise. Instead of handcrafted sparsity priors used in earlier algorithms, we define deep neural networks that learn implicit priors based on training data. Our models significantly reduce heart rate estimation errors compared with the previous state of the art.

6. REFERENCES

[1] Giulio Nittari, Demetris Savva, Daniele Tomassoni, Seyed Khosrow Tayebati, and Francesco Amenta, "Telemedicine in the covid-19 era: A narrative review based on current evidence," *Inter. Jour. of Environmental Research and Public Health*, vol. 19, no. 9, 2022.

[2] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, and Nicu Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *2016 Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2396–2404.

[3] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak, "Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity," in *2011 Federated Conf. on Computer Science and Information Systems (FedCSIS)*, 2011, pp. 405–410.

[4] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. on Biomedical Eng.*, vol. 58, no. 1, pp. 7–11, 2011.

[5] Gerard de Haan and Vincent Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Trans. on Biomedical Eng.*, vol. 60, no. 10, pp. 2878–2886, 2013.

[6] Youngjun Cho, Nadia Bianchi-Berthouze, and Simon J. Julier, "DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings," in *2017 Seventh Intern. Conf. on Affective Computing and Intelligent Interaction (ACII)*. oct 2017, IEEE.

[7] Dangdang Shao, Yuting Yang, Chenbin Liu, Francis Tsow, Hui Yu, and Nongjian Tao, "Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time," *IEEE Trans. on Biomedical Eng.*, vol. 61, no. 11, pp. 2760–2767, 2014.

[8] Kaito Iuchi, Ryogo Miyazaki, George C. Cardoso, Keiko Ogawa-Ochiai, and Norimichi Tsumura, "Remote estimation of continuous blood pressure by a convolutional neural network trained on spatial patterns of facial pulse waves," in *2022 Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 2138–2144.

[9] 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 Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1353–135309.

[10] Ewa M. Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan, "Near-infrared imaging photoplethysmography during driving," *IEEE Trans. on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3589–3600, 2022.

[11] Chihiro Takano and Yuji Ohta, "Heart rate measurement based on a time-lapse image," *Medical Eng. Physics*, vol. 29, no. 8, pp. 853–857, 2007.

[12] Guha Balakrishnan, Fredo Durand, and John Guttag, "Detecting pulse from head motions in video," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 3430–3437.

[13] Weixuan Chen and Daniel McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proc. of the european Conf. on computer vision (ECCV)*, 2018, pp. 349–365.

[14] Zhaodong Sun and Xiaobai Li, "Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast," in *Computer Vision–ECCV 2022: 17th European Conf.* Springer, 2022, pp. 492–510.

[15] Ewa M. Nowara, Daniel McDuff, and Ashok Veeraraghavan, "The benefit of distraction: Denoising camera-based physiological measurements using inverse attention," in *Proc. of the Inter. Conf. on Computer Vision (ICCV)*, October 2021, pp. 4955–4964.

[16] Armand Comas, Tim K. Marks, Hassan Mansour, Suhas Lohit, Yechi Ma, and Xiaoming Liu, "Turnip: Time-series u-net with recurrence for nir imaging ppg," in *2021 IEEE Inter. Conf. on Image Processing (ICIP)*, 2021, pp. 309–313.

[17] Zheng Zhang, Jeff M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[18] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[19] Gili Dardikman-Yoffe and Yonina C Eldar, "Learned sparcom: unfolded deep super-resolution microscopy," *Optics express*, vol. 28, no. 19, pp. 27736–27763, 2020.

[20] Xiaoshuai Zhang, Yiping Lu, Jiaying Liu, and Bin Dong, "Dynamically unfolding recurrent restorer: A moving endpoint control method for image restoration," 2018.

[21] Suhas Lohit, Dehong Liu, Hassan Mansour, and Petros T. Boufounos, "Unrolled projected gradient descent for multi-spectral image fusion," in *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7725–7729.

[22] Karol Gregor and Yann LeCun, "Learning fast approximations of sparse coding," in *Proc. of the 27th International Conf. on Machine Learning*, Madison, WI, USA, 2010, ICML'10, p. 399–406, Omnipress.

[23] Morteza Mardani, Hatef Monajemi, Vardan Papyan, Shreyas Vasanawala, David Donoho, and John Pauly, "Recurrent generative adversarial networks for proximal learning and automated compressive image recovery," *arXiv preprint arXiv:1711.10046*, 2017.

[24] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan, "Algorithmic principles of remote ppg," *IEEE Trans. on Biomedical Eng.*, vol. 64, no. 7, pp. 1479–1491, 2017.

[25] G de Haan and A van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, pp. 1913, aug 2014.

[26] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li, "Faceboxes: A cpu real-time face detector with high accuracy," in *2017 IEEE Inter. Joint Conf. on Biometrics (IJCB)*. IEEE, 2017, pp. 1–9.

[27] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng, "Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood," in *2020 Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8233–8243.

[28] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] 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*, vol. 18, no. 10, pp. 10762–10774, May 2010.