Supplementary Material
RADIANT: Better rPPG estimation using signal embeddings and Transformer

Anup Kumar Gupta Rupesh Kumar Lokendra Birla Puneet Gupta
Indian Institute of Technology Indore
{msrphd2105101002, ms2004101008, phd1901201001, puneet}@iiti.ac.in

1. Face ROI Extraction

The landmark points corresponding to different face regions are depicted in Figure 1 (a) in blue and yellow colors. The landmark points in yellow are used for defining the facial region with the significant rPPG information. The extracted face region is presented in Figure 1 (b). The divided face ROIs are depicted using red squares in Figure 1 (c).

(a) Landmark Detection (b) Facial Region (c) ROI Division

Figure 1. Example of ROI extraction. Face landmark points are shown in (a) with blue color and the landmarks used for obtaining the convex hull in green color, the extracted face region is shown in (b), and the extracted ROIs are shown in (c) using red color squares.

2. Chrominance signals

The chrominance subspace consists of projections of RGB temporal signals into two orthogonal vectors [2]. For obtaining the chrominance signals, the filtered RGB signals, $\tilde{\boldsymbol{r}}_j$, $\tilde{\boldsymbol{g}}_j$ and $\tilde{\boldsymbol{b}}_j$ are normalized. The normalized RGB signals are represented by $\bar{\boldsymbol{r}}_j$, $\bar{\boldsymbol{g}}_j$ and $\bar{\boldsymbol{b}}_j$. These normalized signals are then projected into orthogonal vectors $\boldsymbol{\eta}_j$ and $\boldsymbol{\mu}_j$ obtained by:

$$\begin{aligned} \boldsymbol{\eta}_j &= 3 * \bar{\boldsymbol{r}}_{\boldsymbol{j}} - 2 * \bar{\boldsymbol{g}}_{\boldsymbol{j}} \\ \boldsymbol{\mu}_j &= 1.5 * \bar{\boldsymbol{r}}_{\boldsymbol{j}} + \bar{\boldsymbol{g}}_{\boldsymbol{j}} - 1.5 * \bar{\boldsymbol{b}}_{\boldsymbol{j}} \end{aligned} \tag{1}$$

Kindly note that the coefficients associated with each of the color channels intensitites $\bar{\boldsymbol{r}}_j$, $\bar{\boldsymbol{g}}_j$ and $\bar{\boldsymbol{b}}_j$ are empirically derived by large scale experiments described in [2]. The obtained vectors, $\boldsymbol{\eta}_j$ and $\boldsymbol{\mu}_j$ are fed to the bandpass filter, $\psi_{bp}(\cdot)$ for obtaining the vectors $\tilde{\boldsymbol{\eta}}_j$ and $\tilde{\boldsymbol{\mu}}_j$. Eventually, the chrominance signal, $\boldsymbol{c}_j$ is obtained by:

$$\boldsymbol{c}_j = \tilde{\boldsymbol{\eta}}_j - \alpha * \tilde{\boldsymbol{\mu}}_j \text{ ,where } \alpha = \frac{\sigma(\tilde{\boldsymbol{\eta}}_j)}{\sigma(\tilde{\boldsymbol{\mu}}_j)} \tag{2}$$

where, $\sigma(\cdot)$ represents the standard deviation operator.

3. Synthetic Temporal Signals

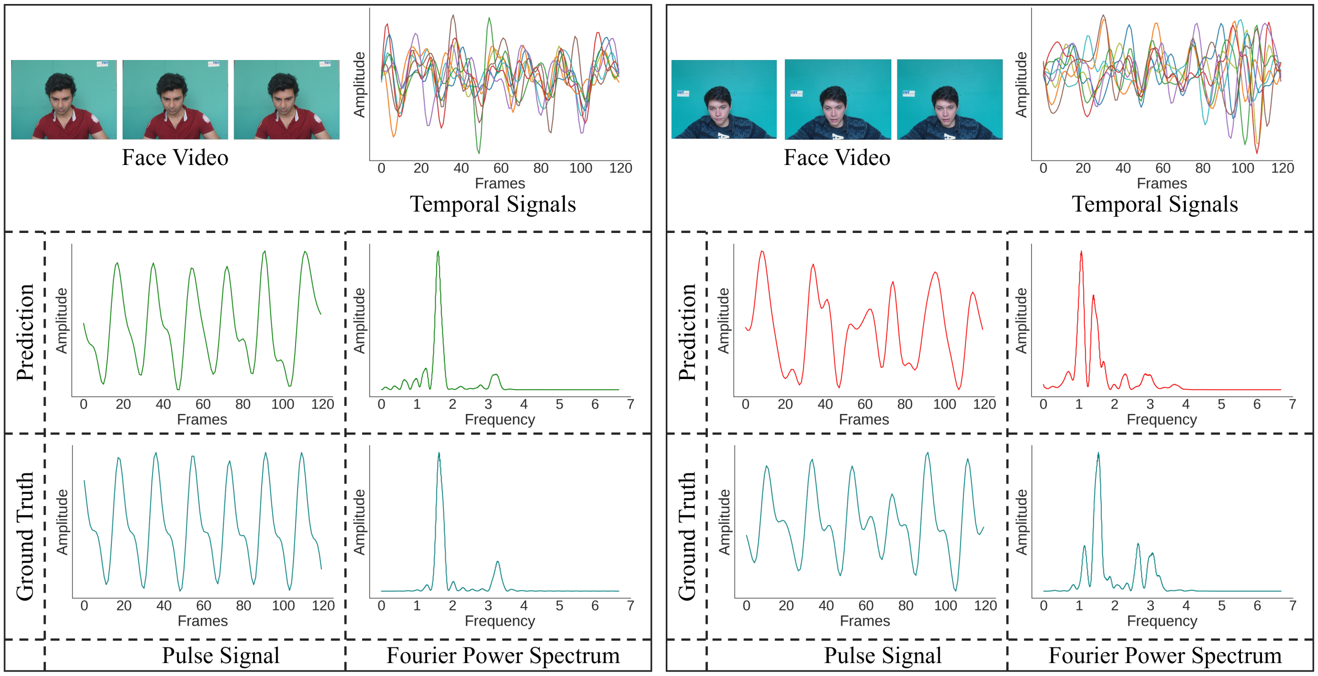
The synthetic temporal signals used for pre-training our architecture are generated using sine waves and noise functions as described in [3]. To mimic the systolic and diastolic peaks in the synthetic temporal signals, we have used two waves with the same time period, corresponding to the pulse, with one of the waves having twice the amplitude of the other wave. Further, another wave, having the time period corresponding to the respiratory signal, is used. For adding the effect of noise, we have used a step function and the Gaussian noise function. The synthetic temporal signal $\boldsymbol{s}_{syn}$ is given by:

$$\begin{aligned} \boldsymbol{s}_{syn} = {} & \kappa_1 * \sin(\omega_1 t + \phi) + \kappa_2 * \sin(\omega_2 t + \theta) + \\ & 0.5 * \kappa_1 * \sin(2\omega_1 t + \phi) + N(t) \\ & p_1 * \text{step}(t - t_1) + p_2 * \text{step}(t - t_2) \end{aligned} \tag{3}$$

where, κ_1 and κ_2 are amplitudes of the sine waves sampled randomly from $[0, 1]$; ω_1 and ω_2 are the HR and respiration frequencies, respectively. Random phase ϕ and θ are sampled randomly from $[0, \pi]$. Also, $\text{step}(t)$ denotes the step function used to add noise and the values t_1 and t_2 are chosen between $[0, T]$ randomly, where T is the video clip length. Furthermore. *N* represents the Gaussian noise. The values p_1 and p_2 are derived from Bernoulli distribution.

4. Effect of size of the ROIs used for rPPG estimation

The effect of different block sizes considered for ROI divison on the performance of our method is provided in Table 1. The experiments are performed by changing the



(a) Successful HR estimation

(b) Unsuccessful HR estimation

Figure 2. Example of HR estimation by our proposed method *RADIANT* over the UBFC-rPPG dataset [1].

Table 1. Performance analysis of *RADIANT* with varying blocks for obtaining the ROIs. All the values are in BPM and all the metrics represent better performance if they have lower values.

	UBFC-rPPG			COHFACE		
Blocks	σ	MAE	RMSE	σ	MAE	RMSE
8	05.54	03.68	06.65	10.11	09.81	11.10
9	04.32	03.51	05.12	08.99	09.02	10.56
10	**03.45**	**02.91**	**04.52**	**07.41**	**08.01**	**10.12**
11	05.22	04.61	06.69	09.12	08.65	11.32
12	06.81	05.25	08.59	11.78	10.18	12.06

number of non-overlapping blocks in the horizontal direction. Hence, when we increase the number of blocks, the size of the patch decreases. Likewise, on decreasing the number of blocks, the patch size increases. Initially, when we decrease the size of the patch, the performance of our method improves. However, the performance saturates for an optimum size of the patches, after which a decrease in the performance is observed. As we go on decreasing the patch size, the total number of temporal signals obtained increases, hence, providing a better combination of rPPG information. However, the smaller patch sizes are susceptible to noise, which affects the quality of rPPG information, leading to performance degradation. *RADIANT* obtains optimal performance when 10 horizontal blocks are considered.

5. Qualitative Results

The figures 2 (a) and 2 (b) depict examples for HR estimation by our proposed method. The face videos and temporal signals are presented in the first row; the second row shows the estimated pulse signal and its Fourier Power Spectrum, and the third row presents the ground truth signal and its Power Spectrum. In Figure 2 (a), an example of successful HR estimation is provided with the estimated and ground truth pulse rate to be 98 BPM. The quality score for the temporal signals is 5.19 for this sample. Further, a higher correlation is observed in the estimated and ground truth pulse signal. Similarly, the Fourier Power Spectrum of the estimated pulse and the ground truth pulse signal show single peaks denoting less amount of noise in the estimated pulse signal. An example of unsuccessful HR estimation can be observed in Figure 2 (b) where the estimated HR is 63 BPM, and the ground truth is 92 BPM. The quality score of the temporal signals is 3.80. Further, there is little correlation between the estimated pulse signal and the ground truth pulse signal. Likewise, the Fourier Power Spectrum of the estimated pulse signal shows multiple peaks indicating a higher noise content resulting in incorrect HR estimation.

References

[1] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue seg-

mentation for remote photoplethysmography. *Pattern Recognition Letters*, pages 82–90, 2019.

[2] Gerard De Haan and Vincent Jeanne. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering*, pages 2878–2886, 2013.

[3] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a Deep Heart Rate Estimator from General to Specific. In *International Conference on Pattern Recognition*, pages 3580–3585. IEEE Computing Society, 2018.