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

Abstract

Remote photoplethysmography can provide non-contact heart rate (HR) estimation by analyzing the skin color variations obtained from face videos. These variations are subtle, imperceptible to human eyes, and easily affected by noise. Existing deep learning-based rPPG estimators are incompetent due to three reasons. Firstly, they suppress the noise by utilizing information from the whole face even though different facial regions contain different noise characteristics. Secondly, local noise characteristics inherently affect the convolutional neural network (CNN) architectures. Lastly, the CNN sequential architectures fail to preserve long temporal dependencies. To address these issues, we propose RADIANT, that is, rPPG estimation using Signal Embeddings and Transformer. Our architecture utilizes a multi-head attention mechanism that facilitates feature subspace learning to extract the multiple correlations among the color variations corresponding to the periodic pulse. Also, its global information processing ability helps to suppress local noise characteristics. Furthermore, we propose novel signal embedding to enhance the rPPG feature representation and suppress noise. We have also improved the generalization of our architecture by adding a new training set. To this end, the effectiveness of synthetic temporal signals and data augmentations were explored. Experiments on extensively utilized rPPG datasets demonstrate that our architecture outperforms previous well-known architectures. Code: https://github.com/Deep-Intelligence-Lab/RADIANT.git

1. Introduction

The heart continuously pumps blood through the capillaries and induces periodic cardiovascular pulse throughout the body. The number of pulses induced in a minute is known as heart rate (HR). HR estimation is important to measure a person's health [21]. It is a primary indicator of heart-related problems and mental health states ranging from stress, depression, anxiety, and excitement [21].

The non-invasive HR estimation techniques comprise of electrocardiogram (ECG), ballistocardiogram (BCG) and photoplethysmography (PPG). These techniques require contact with the skin surface and cause discomfort for continuous HR estimation [61]. Hence, they offer limited applicability for skin-damaged patients, patients suffering from severe skin infections, exercise monitoring, and neonates monitoring [21, 29]. In contrast, remote Photoplethysmography (rPPG) is a non-contact HR estimation method utilizing non-contact face videos for estimating cardiovascular pulse. Since, it avoids any contact between sensors and skin, it can be used in applications where contact-based PPG methods cannot be used such as, drowsy driver detection [53], deepfake detection [25, 54], face anti-spoofing [4, 5], micro-expression recognition [16], micro-expression spotting [20, 17], lie detection [59] and stress monitoring [49, 38]. In the ongoing SARS-CoV-2 pandemic rPPG can be used for automated HR estimation and provide support to the patients requiring urgent and critical telehealthcare. It motivates us to propose an accurate rPPG-based HR estimation method.

The rPPG method analyzes the variations in the blood flow volume in the carotid arteries beneath the facial skin [21]. These variations induce subtle skin color changes imperceptible to the human eyes [34]. However, a video camera is capable of capturing these variations. The face videos acquired in controlled environments provide relevant rPPG information, resulting in correct HR estimation. However, in real scenarios, these videos are influenced by noise due to facial movements, illumination variations [51], and other artifacts and thus, result in spurious HR estimation [22].

The rPPG-based HR estimation first requires the extraction of the relevant face region, known as the region of interest (ROI). Usually, the color variations are analyzed in the ROI and the corresponding signals are referred to as temporal signals. The temporal signals are obtained by either averaging pixel values in the face ROI [39], or subtracting the pixel values from the consecutive video frames [9]. Eventually, the cardiovascular pulse present in the temporal signals is extracted using domain-specific knowledge. To this end, [45] and [2] have employed blind source separation (BSS) and maximum periodicity criteria for HR estimation [45].

Unfortunately, their HR estimations are erroneous when the temporal signals contain periodic noise.

Apart from domain-specific knowledge, rPPG-based HR estimation can be performed using deep learning. For instance, [9, 43] have utilized Convolutional Neural Network (CNN) based architectures for rPPG estimation. Further, AND-rPPG [3] have used temporal convolution network for pulse estimation. Further, a combination of CNN and sequential architecture is used in [39] for rPPG estimation. Unfortunately, the performance of these architectures can be affected by slight facial movements persisting in small facial regions, even for a short duration. Such behaviour is attributed to the local feature encoding of the CNNs [31]. Moreover, the sequential architectures are incompetent to provide correct HR estimation as these architectures fail to model long temporal dependencies [13, 37]. These issues are alleviated by employing dual-stream Transformer architecture in [30]. However, the background color variations used for denoising, fail to provide effective noise representation [32]. Additionally, many features are required by [30] when temporal signals are extracted using the difference of frames. The proper training of the corresponding architectures necessitates large-scale datasets to alleviate the problem of underfitting [7]. Unfortunately, the datasets available for training have limited training data, which restricts the applicability of frame difference-based temporal signal extraction.

We propose a novel rPPG-based HR estimation method, *RADIANT*, that is, betteR rPPG estimAtion methoD using signal embeddIngs ANd Transformer. We obtain the temporal signals by averaging the skin color variations in ROIs that provides efficient rPPG feature representation allowing convergence over limited training data. Furthermore, we utilize the Transformer architecture for estimating pulse as it can learn global context and effectively mitigate local noise [56]. Our main contributions are:

(1) Our proposed rPPG architecture utilizes Multilayer perceptron (MLP) for projecting the temporal signals into signal embeddings and the attention processing ability of the Transformer architecture. Linear projection using MLP provides the proper learning of relevant feature representation for rPPG information while the Transformer architecture effectively performs denoising and cardiovascular pulse estimation. (2) To mitigate the problem of limited training data, we explore the possibility of pre-training our Transformer architecture using synthetic temporal signals [39] and data augmentation [42]. Both are performed in a time-efficient manner, and it is observed that they improve the performance by allowing domain adaptation. (3) Our experimental results demonstrate that we obtain state-of-art results on publicly available datasets.

2. Related Works

2.1. Domain knowledge-based rPPG methods

The rPPG-based HR estimation can be performed by utilizing domain knowledge. Such estimation involves the following steps: ROI detection, spatial filtering, temporal signal extraction, and pulse signal estimation. Usually, color variations are employed to define the temporal signals. Amongst the RGB colors, the green color channel is shown to be the aptest for rPPG information extraction in [57]. On the contrary, a chrominance subspace transformation of the RGB color signals is used for pulse estimation in [10]. Further, BSS is used in [45] for pulse estimation. Eventually, HR is given by the peak frequency in the Fourier power spectrum of the pulse signal [3]. The above-discussed methods are incompetent to distinguish pulse signal and noise because they utilize handcrafted representations to model noise, and they lack the appropriate supervision to understand the noise attributes induced by facial movements [61].

2.2. CNN and sequential architectures for rPPG

CNN architectures have been used extensively in rPPG-based HR estimation because they allow feature subspace mapping with minimal requirement of domain-specific knowledge. For instance, [27] feeds the time-frequency representation of chrominance signals into the VGG15 [50] for HR estimation. A depthwise separable CNN architecture is used in [46] for pulse signal estimation from the Spatio-temporal feature representation of the color variations. Recurrent Neural Network (RNN) is employed by [40] for modeling time dependencies between the temporal signals to estimate HR. ETA-rPPGNet [28] utilizes the time domain subnet to address the problem of redundant rPPG information and noise induced by slight facial deformations. Moreover, the rPPG methods in [9, 35] have utilized end-to-end architectures for extracting relevant rPPG information from face videos. The method in [9] feeds the normalized frame difference into a CNN for HR estimation. Similarly, encoder-decoder architecture is employed in [41] for learning noise and rPPG information. Training these huge architectures is challenging, and it tends to underfitting over small-scale datasets [7]. Temporal difference CNN is used in [36] for capturing temporal color variations and generating appropriate signal representation. The importance of synthetic temporal signal generation in rPPG-based HR estimation has been explored in [39]. The synthetic signals can be efficiently generated for pre-training using periodic sine curves and random noise.

The CNN or sequential architectures require diverse datasets in considerable quantities to alleviate underfitting [7]. Further, they fail to learn global information from the facial regions and fail to model the long temporal dependency [18] if sequential architectures are employed [40].

2.3. Attention mechanism and Transformer

The Transformer architecture is proposed in [56] for Natural Language Translation. It models the complex global contextual dependencies between the words in a sentence using multi-head attention mechanism. The huge success of Transformer architecture due to the appropriate modeling enables their applicability in various natural language processing tasks [11, 14, 15]. The applicability of these architectures has been further explored for the image classification task using the Vision Transformer (ViT) [12], and performance improvements have been examined in this direction. It proliferates the research in unraveling the effectiveness of the Transformer architectures for computer vision applications [55, 8]. The rPPG-based HR estimation is no different. It aims to estimate pulse signals from several temporal signals, where each temporal signal contains mainly the pulse signal corrupted by noise. Thus, the pulse signal results in a strong correlation between the temporal signals, which the Transformer architecture can easily learn. For instance, TransPPG proposed in [30] feeds the temporal signals extracted from frame differences to the Transformer architecture for estimating the pulse signal. Similarly, [47] have utilized the Transformer architecture [56] for estimating pulse signal from facial video.

Kindly note that these estimated pulse signals can be erroneous when the corresponding temporal signals contain significant noise. Furthermore, these methods require large parameter learning using limited training data for relevant pulse signal estimation. This situation results in an underfitting problem and thereby degrade the efficacy [7].

3. Proposed method

This section presents our proposed rPPG-based HR estimation method, *RADIANT*. Figure 1 shows the flow diagram of our proposed method. Initially, the face region is identified and divided into multiple ROIs. Subsequently, the temporal signals are extracted from these ROIs, and chrominance subspace transformation [10] is applied to alleviate the effects of motion and luminance. Subsequently, an MLP layer is applied to project the resultant temporal signals into signal embeddings. Eventually, the Transformer architecture utilizes the signal embeddings for pulse estimation.

3.1. Video Clip Extraction

For HR estimation, the video is divided into multiple clips. We have utilized a non-overlapping window over the video clips with a window of 4 seconds for HR estimation. Such division overcomes loss of information due to lack of complete pulse signal in a small time interval [28]. The mean of the HRs estimated for 5 consecutive video clips is obtained for obtaining the HR for a 20 second video.

3.2. Temporal Signal Acquisition

3.2.1 ROI Extraction

Our first step is to define a fixed ROI that provides relevant rPPG information from an input video clip. The relevant rPPG information is present in the facial skin regions. Thus, we employ facial landmark points that outline the boundaries of the face, including its subparts like eyes, lips, and nose. These landmark points are extracted using the CLNF Openface 2.2.0 landmark detector [62] for the first video frame because computing landmark points for each video frame is time expensive [18]. It provides 68 landmark points. Since the regions near the eyes tend to be easily influenced by facial expressions, we avoid them for defining the ROI [3]. Similarly, the forehead region is avoided as it is usually covered with hairs to induce noise in the corresponding temporal signals [61]. We mainly use the region below the eyes containing the cheeks for obtaining the temporal signals. The desired face region containing relevant rPPG information is obtained by computing the convex hull of the landmark points: 1) $2, 3, 4, 5, 6$ of the left cheek, 2) $12, 13, 14, 15, 16$ of the right cheek, 3) 29 of the nose, and 4) $7, 8, 9, 10, 11$ of the chin. It is observed in [3] that the face boundaries are prone to significant temporal variations from minimal facial deformations. Thus, we employ the morphological operation to remove the boundary pixels similar to [62]. For a better visualization, kindly refer to the supplementary material.

We can extract the temporal signal from the entire face region, but noise-induced in a small facial region affects the resultant temporal signal. Thus, it is recommended in [18] to divide the region into smaller ROIs, extracting temporal signals from those ROIs, and then consolidate the signals for pulse estimation. Following this similar path, we divide the obtained facial region into small non-overlapping square blocks of the same size. Amongst the square blocks, we consider only those square blocks as ROI whose all the pixels belong to the skin pixels. The skin pixels are detected using the method described in [44]. We have used the method described in [22] for choosing the optimum size of our square blocks while alleviating the effect of scale difference among different faces.

3.2.2 Temporal Signal Extraction

The temporal signal extraction is performed in two steps: RGB signal extraction and projection into chrominance signals. The RGB signals are obtained by averaging the pixel values of the red, green, and blue channels from the face ROI. Mathematically, the temporal signal $\boldsymbol{r}_j$, denoting the red color channel signal for j^{th} ROI is given by:

$$\boldsymbol{r}_j = \left(\frac{\sum_k r^k_{j,1}}{p_{j,1}}, \frac{\sum_k r^k_{j,2}}{p_{j,2}}, \ldots, \frac{\sum_k r^k_{j,f}}{p_{j,f}} \right) \quad (1)$$

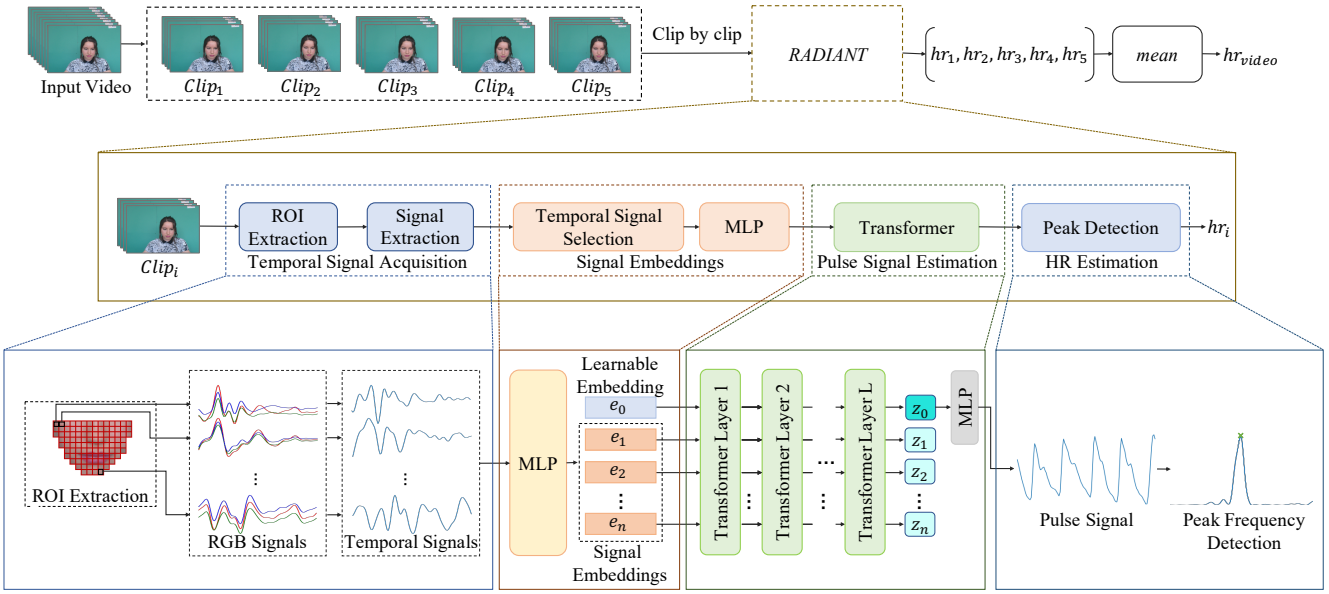

Figure 1. The flow diagram of our proposed method *RADIANT*.

where, $r^k_{j,i}$ refers to the red color channel intensity of the k^{th} pixel in the j^{th} ROI of the i^{th} frame. Similarly, the temporal signals $\boldsymbol{g}_j$ and $\boldsymbol{b}_j$ are obtained from the green and blue color channel intensities respectively. The total number of pixels in j^{th} ROI of i^{th} frame is represented by $p_{j,i}$ and the total number of frames in the video clip is represented by f. The obtained RGB signals are then passed through a bandpass filter to obtain the filtered signals as $\tilde{\boldsymbol{r}}_j$, $\tilde{\boldsymbol{b}}_j$ and $\tilde{\boldsymbol{g}}_j$. Mathematically:

$$\tilde{\boldsymbol{r}}_j = \psi_{bp}(\boldsymbol{r}_j),\ \tilde{\boldsymbol{g}}_j = \psi_{bp}(\boldsymbol{g}_j),\ \tilde{\boldsymbol{b}}_j = \psi_{bp}(\boldsymbol{b}_j) \qquad (2)$$

where, $\psi_{bp}(\cdot)$ is a Butterworth bandpass filter of order 4 [60] that suppresses any signal components that correspond to frequencies outside the HR range (0.7 Hz to 4.2 Hz) [3].

These RGB signals are then projected into chrominance signals to minimize the noise artifacts and suppress specular reflections [10] [1]. More details on chrominance transformation are provided in the supplementary material. The resultant temporal signal is filtered using a detrending filter to mitigate noise due to illumination variations [21].

3.3. Signal embeddings

3.3.1 Temporal signal selection

Assume that $(\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots \boldsymbol{c}_m)$ are the temporal signals obtained from m facial ROIs. Kindly note that different video clip results in a different number of ROIs. That is, the value of m depends on the input clip. Usually, deep learning architectures require fixed-size input. Thus, we select n temporal signals amongst the m extracted temporal signals. The value of n is selected such that each video clip contains at least n ROIs. For selecting the temporal signals, we use the observation that some extracted temporal signals are affected by noise due to facial movements. For instance, motion caused due to smiling usually affects the temporal signals obtained from the regions near the lips while leaving the other ROIs unaffected. Thus, we choose those n temporal signals that are least affected by facial deformations. To this end, we leverage the intuition that the face regions with facial movements will have large skin color variations, resulting in large standard deviation in the amplitude of the extracted temporal signals. Thus, temporal signals with less standard deviation will provide better rPPG information. Hence, we select those top-n temporal signals which have the higher quality scores, where the quality score corresponding to temporal signal $\boldsymbol{c}_j$ is given by:

$$quality_j = \frac{1}{\sigma(\boldsymbol{c}_j)} \qquad (3)$$

where σ refer to the standard deviation operator.

3.3.2 MLP

Even though the chosen temporal signals are least affected by noise, they can still result in erroneous HR estimation because they contain noise. Thus, we project the temporal signals into signal embeddings $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ of higher dimensions using an MLP. This projection allows adequate representation subspace for learning relevant rPPG features and performing denoising. Furthermore, we will utilize the

[1] Code `https://github.com/phuselab/pyVHR`

Transformer architecture for pulse estimation. The architecture consists of multiple Transformer layers that require input vectors of the same dimensions d. We want to benefit from effective weight initialization by pre-training with ImageNet [48] dataset. Hence, we have set the dimensions to 768 as the pre-trained weights require the input of size 768. We utilize a learnable MLP layer with 768 output nodes for projecting a temporal signal $\boldsymbol{c}_j$ into an embedding $\boldsymbol{e}_j$, for $j = 1, 2, \ldots, n$. Kindly note that our projection ensures that rPPG information in an embedding only depends on the corresponding temporal signal while remaining unaffected by other temporal signals. The reason for choosing such a projection is that when an embedding depends on several temporal signals, then information from a temporal signal with higher noise can easily affect the embedding, leading to the incorrect HR estimation [3]. Finally, inspired by the use of classification token by the architectures in [12, 11], we have prepended a learnable embedding $\boldsymbol{e}_0 \in \mathbb{R}^{1\times d}$ that learns the feature representation of the pulse signal.

3.4. Pulse signal estimation

3.4.1 Transformer Architecture

Each signal embedding contains a composition of features from pulse signal, other physiological parameters, and noise. Thus, we need to consolidate and filter the rPPG information from the signal embeddings for correct pulse signal estimation. To this end, we employ the Transformer architecture. The Transformer architecture utilizes the self-attention mechanism for consolidating feature information from the input components [31]. The self-attention mechanism learns the contextual dependencies between the signal embeddings, allowing our architecture to learn the correlations for contributing to rPPG information. Subsequently, the consolidated rPPG features are transformed using a two-layer MLP [31]. This combination allows the Transformer architecture to denoise and consolidates rPPG features for correct pulse signal estimation.

The self-attention mechanism is the building block of the Transformer. For each signal embedding $\boldsymbol{e}_j$, it computes a corresponding latent vector $\boldsymbol{z}_j$ which is the weighted sum of all the signal embeddings. Here, $\boldsymbol{z}_j$ is of the same dimensions as $\boldsymbol{e}_j$. For obtaining these corresponding weights, the signal embeddings $\boldsymbol{e}_j$ are projected into query ($\boldsymbol{q}_j \in \mathbb{R}^{1\times d_q}$), key ($\boldsymbol{k}_j \in \mathbb{R}^{1\times d_k}$) and value ($\boldsymbol{v}_j \in \mathbb{R}^{1\times d_v}$) vectors using learnable weights. Alternatively, the signal embedding matrix $\boldsymbol{E} \in \mathbb{R}^{n\times d}$ containing all the signal embeddings is projected into matrices: $\boldsymbol{Q} \in \mathbb{R}^{n\times d_q}$, $\boldsymbol{K} \in \mathbb{R}^{n\times d_k}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times d_v}$ corresponding to the queries, keys and values, respectively. That is,

$$\begin{aligned} &\boldsymbol{Q} = \boldsymbol{E}\cdot\boldsymbol{W^Q},\ \boldsymbol{K} = \boldsymbol{E}\cdot\boldsymbol{W^K},\ \text{and}\ \boldsymbol{V} = \boldsymbol{E}\cdot\boldsymbol{W^V} \\ &\boldsymbol{W^Q} \in \mathbb{R}^{d\times d_q},\ \ \boldsymbol{W^K} \in \mathbb{R}^{d\times d_k},\ \ \boldsymbol{W^V} \in \mathbb{R}^{d\times d_v} \end{aligned} \tag{4}$$

where, $\boldsymbol{W^Q}$, $\boldsymbol{W^K}$ and $\boldsymbol{W^V}$ are the learnable weights.

The correlations between the signal embeddings are reflected by the attention scores obtained by:

$$\boldsymbol{SA} = softmax\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^T}{\sqrt{d_q}}\right)\boldsymbol{V} \tag{5}$$

where, the dot product between its query and all the keys is calculated for a given signal embedding. The resulting value is scaled by the square root of d_q followed by a softmax operation. The obtained scores associated with each embedding transform them into a weighted sum of the features from all the signal embeddings.

There are multiple relationships among the signal embeddings contributing toward rPPG information. For encapsulating such relationships, the multi-headed self-attention (MSA) mechanism is employed by the Transformer architectures, as in natural language processing [11] and computer vision [12] applications. The MSA mechanism transforms the queries, keys, and values into multiple learned linear projections, corresponding to multiple heads, h, for modeling multiple relationships. Subsequently, independent self-attention computation is performed in the respective heads using the equation 5 resulting in $\boldsymbol{SA}_0, \boldsymbol{SA}_1, \ldots, \boldsymbol{SA}_{h-1}$. Kindly note that, here $\boldsymbol{SA}_i$ for $i = 0, 1, \ldots, h-1$ represents the output from i^{th} head. Finally, for consolidating the information, the output obtained from each head is concatenated and projected into d dimensional vectors using a learnable matrix $\boldsymbol{W^O}$. That is, the matrix containing the latent vectors, $\boldsymbol{Z'}$ is given by:

$$\boldsymbol{Z'} = [\boldsymbol{SA}_0, \boldsymbol{SA}_1, \ldots, \boldsymbol{SA}_{h-1}]\cdot\boldsymbol{W^O},\ \ \boldsymbol{W^O} \in \mathbb{R}^{(h\cdot d_v)\times d} \tag{6}$$

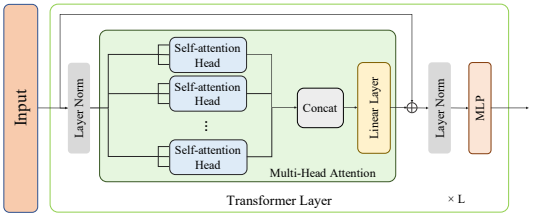

Figure 2. An illustration of a Transformer layer.

The output $\boldsymbol{Z'}$ is then fed to a LayerNorm (LN) layer [1] followed by a two layer MLP with GELU activation function [24] for feature transformations. A residual connection is placed after the MLP output. The architecture of a single Transformer layer is depicted in Figure 2. Kindly note that there are multiple Transformer layers and the output from a Transformer layer l is given by:

$$\begin{aligned} &\boldsymbol{Z}_l = MLP\left(LN\left(MSA\left(\boldsymbol{Z}_{l-1}\right)\right)\right) + \boldsymbol{Z}_{l-1}, \text{where} \\ &\boldsymbol{Z_0} = [\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_n] \end{aligned} \tag{7}$$

Assume that the output of the last Transformer layer $\boldsymbol{Z}_L$ is $(\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$. Among these vectors, the vectors $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)$ contain the refined correlations that contribute to the rPPG information. However, the vector $\boldsymbol{z}_0$ contains the pulse information extracted from the correlations learned by the other latent vectors $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)$. We use the mean square error loss function during training between the estimated pulse and the ground truth. It necessitates that the dimensions of the estimated and ground-truth pulse signal should be the same. Thus, an MLP head is attached to $\boldsymbol{z}_0$ that transforms it into the pulse signal of the same dimensions as the ground truth.

3.4.2 Pre-training

Transformers show poor generalization capabilities in the computer vision domain due to the lack of their inductive bias [12]; thus, pre-training is employed to improve their generalization. Following this, we employ pre-training in two stages. The first stage performs the pre-training over ImageNet dataset [48]. While the second stage uses synthetic temporal signals generated using the method described in [39]. For brevity, the synthetic signals are obtained using sine waves. Since the waves should represent periodic cardiovascular pulses, their frequencies are set between 0.7 Hz to 4.2 Hz, corresponding to the normal HR range. In addition, the generated sine waves are superimposed with another sine wave simulating breathing rhythm with periodicity between 5 beats per minute (BPM) to 20 BPM, random step signals and Gaussian noise. It stimulates the noisy variations. Further details are provided in supplementary material. After pre-training on the synthetic dataset, we perform fine-tuning on public rPPG datasets.

3.4.3 Data Augmentation

In the rPPG datasets, most samples contain HR ranges from 60 to 90 BPM, resulting in uneven data distribution. Training on such dataset will make our architecture biased towards samples containing HR in this range. Addressing this issue [42] have described a novel data augmentation method comprising temporal upsampling and downsampling of the video for providing proper HR coverage in the training dataset. Alternatively, temporal signal interpolation can be employed to obtain the same effect, and such interpolation prevents the time-consuming steps of processing video frames. Thus, we have utilized this technique to perform our data augmentation. Specifically, we generate samples with higher HR ranges by downsampling the temporal signals by 2 and 3 times. Similarly, we generate the samples with lower HR values by upsampling the temporal signals twice and thrice. Also, we have discarded the augmented samples, whose HR values lie outside the human HR range of $40 - 240$ BPM [3].

3.5. Heart Rate Estimation

The raw pulse signal ($\boldsymbol{y}$) obtained from the last Transformer layer is bandpass filtered for removing any signal components whose frequencies lie outside the normal HR range. We have applied the bandpass filter $\psi_{bp}(\cdot)$ (described in 3.2.2) over $\boldsymbol{y}$ to obtain the clean pulse signal $\tilde{\boldsymbol{y}}$. Subsequently, the pulse spectrum ($PS[freq]$) is obtained by applying Fast Fourier transform (FFT) over the pulse signal $\tilde{\boldsymbol{y}}$. Note that $PS[freq]$ is the amplitude of the pulse spectrum at frequency $freq$. For a video clip i, the HR is given by:

$$hr_i = \underset{freq}{\operatorname{argmax}}\, PS\,[freq] \times 60 \tag{8}$$

kindly note that as stated in section 3.1, the HR for a 20 seconds video is obtained by averaging the HR obtained from the consecutive 5 short video clips. Hence, the final HR estimate for a 20 seconds video is given by:

$$hr_{video} = mean(hr_1, hr_2, hr_3, hr_4, hr_5) \tag{9}$$

4. Experimental Results

4.1. Dataset and Metrics

We have provided evaluation results for the public UBFC-rPPG [6] and COHFACE [26] datasets. The UBFC-rPPG dataset consists of facial videos of 2 minutes from 42 subjects. Videos are recorded in a resolution of 640×480 in 8-bit RGB format at the frame rate of 30 frames per second. We have split the dataset into training and testing sets with videos from 28 subjects into training set and 14 subjects into the testing set. The COHFACE dataset contains face videos of 40 subjects alongwith their physiological information. Each video is of 1 min recorded at 20 frames per second. For reporting our evaluation results, we have reported mean absolute error (MAE), standard deviation (σ) and root mean squared error (RMSE) between the ground truth HR and estimated HR.

4.2. Choice of Training Parameters

We have used Adam optimizer with a learning rate of 3×10^{-4}. We have used a batch size of 32 for pre-training our architecture using the synthetic temporal signals. However, for fine-tuning, we have used a batch size of 4. We have performed pre-training and fine-tuning over 20 and 50 epochs, respectively. We have used the mean squared error loss function for training our architecture.

4.3. Comparative evaluation

We have provided the comparison of our method, *RADIANT*, with previous rPPG methods on UBFC-rPPG and COHFACE datasets. Note that we have used publicly available codes and experimental settings for comparisons. Furthermore, we have used the standard training and testing split

Table 1. Performance evaluation of *RADIANT* for Average HR variation per video. All the values are in BPM and all the metrics represent better performance if they have lower values.

	UBFC-rPPG			COHFACE		
	σ	MAE	RMSE	σ	MAE	RMSE
[58]	17.89	15.95	11.65	22.30	20.97	25.98
[45]	12.80	06.95	13.60	13.83	08.89	14.55
[10]	05.21	03.21	06.14	11.61	10.15	12.69
[18]	07.00	06.15	07.92	08.10	08.27	11.31
[19]	06.02	05.08	07.42	07.98	08.97	10.84
[33]	08.00	06.54	09.11	11.52	09.31	12.27
[9]	08.73	06.27	10.82	09.01	08.25	14.71
[52]	05.21	04.90	05.89	09.46	08.10	10.80
[47]	08.18	11.28	13.94	11.24	19.66	22.65
Ours	**03.45**	**02.91**	**04.52**	**07.41**	**08.01**	**10.12**

as described in [52, 26] for a fair comparison. Kindly note that the InstTrans [47] architecture provides HR estimation heart rate for 100 video frames. Thus, for obtaining HR from a video clip of the desired duration, we obtain an average of the HR estimates for multiple smaller video clips, each of 100 frames. Table 1 shows our results. It indicates that our method outperforms earlier works consisting of 2SR [58], ICA [45] and chrominance-rPPG [10] because they have used BSS for estimating the pulse signal. These techniques are incapable of denoising because they utilize handcrafted representations to model noise and the lack of supervision limits their capability to understand the noise features induced by facial movements [61].

Similarly, our method provides better performance than AHRE [18] and Fusion-EL [19] because they have also used BSS for estimating pulse signals which limit their ability to extract pulse signal attributes. Additionally, they use the same constraints for denoising the temporal signals obtained from all the face ROIs. However, different face regions have local noise sources [3]. This issue is mitigated by the signal embeddings employed in our proposed method since the signal embedding from a particular face region remains unaffected by other facial regions. For the same reasons as above, we obtain better performance than deep learning based approaches Deepphys [9], and HR-CNN [52] due to better representation modeling of physiological sources. Additionally, the aforementioned CNN-based architectures are affected by the noise induced for a short duration due to small facial movements. However, our Transformer-based architecture is able to mitigate such issues. This behavior is attributed to the global processing capabilities of Transformers. Our method outperforms META-rPPG [33] because they use Long Short-Term Memory (LSTM) network for modeling rPPG information, and it is observed that LSTM architectures are prone to information loss for long sequences [13].

The results indicate that our method outperforms Transformer-based InstTrans architecture [47] because they utilize difference of frames for obtaining the temporal signals and a dual-stream architecture for identifying facial regions with significant rPPG information. Thus, it requires large-scale datasets for overcoming underfitting [7]. Moreover, it has not utilized pre-training for alleviating the effect of poor inductive bias of the Transformer architecture [12]. In contrast, our efficient temporal signal embeddings, pre-training, and data augmentation techniques, followed by the Transformer-based architecture, allows our method to overcome these issues and provide correct HR estimation.

Table 2. Performance evaluation of our proposed method for different number of ROIs.All the values are in BPM

	UBFC-rPPG			COHFACE		
ROIs	σ	MAE	RMSE	σ	MAE	RMSE
8	07.64	05.45	09.39	09.40	11.33	14.72
10	04.99	04.55	06.75	07.67	09.60	12.29
12	**03.45**	**02.91**	**04.52**	**07.41**	**08.01**	**10.12**
14	04.19	03.05	05.19	07.12	10.14	12.39
16	05.00	03.23	05.96	08.04	10.85	13.73

4.4. Ablation Study

This subsection provides the impact of training parameters on our proposed architecture, *RADIANT* and the importance of different components of our architecture. Initially, we used different numbers of ROIs for pulse signal estimation in our proposed method, and the results are reported in Table 2. It can be observed that efficacy first improves when the number of ROIs increases. Increasing the number of ROIs increases the number of temporal signals, allowing better feature representation learning for the physiological signals. Whereas including more facial ROIs will also cause an increase in noise components, which will affect the performance of our architecture. Thus, we obtain optimal results for 12 facial ROIs and observe a decrease in performance when more than 12 ROIs are used.

Table 3. Performance evaluation of the proposed method for different experimental settings. All the values are in BPM and all the metrics represent better performance if they have lower values.

	UBFC-rPPG			COHFACE		
	σ	MAE	RMSE	σ	MAE	RMSE
Ours	**03.45**	**02.91**	**04.52**	**07.41**	**08.01**	**10.12**
GS	13.74	48.50	50.41	11.26	17.76	21.03
RGB	13.76	46.46	48.45	11.34	27.06	29.34
NP	08.13	07.73	11.22	10.83	13.43	17.26
NS	07.40	06.11	07.98	11.01	13.40	18.61
$Conv$	07.73	06.25	08.74	11.31	13.82	19.65

Table 3 presents the results for understanding the importance of different components of our method. These experiments are formed by modifying the proposed method, *RADIANT*. The experiments GS and RGB are formed by

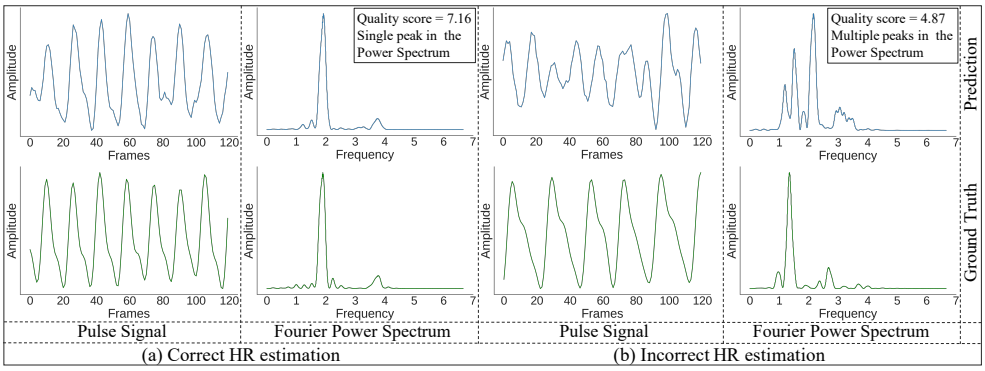

Figure 3. (a) Example of successful HR estimation and (b) example of unsuccessful HR estimation.

replacing the chrominance signals with the green and RGB channel's temporal signals, respectively. The results show that the chrominance signals outperform the other representations because they suppress the effects of motion and luminance variations [10]. The NP experiment is created by replacing the mean squared error loss function with the negative pearson loss function. The results show that our proposed method outperforms NP experiment, indicating that the mean squared error loss function performs better than the negative pearson loss function. In the experiment NS, we have used our architecture without pre-training. The results demonstrate that we obtain better performance when our Transformer is pre-trained with synthetic temporal signals because synthetic signals provide the necessary domain adaptation capability to our architecture for pulse estimation. Similarly, the experiment $Conv$ is formed by replacing the MLP projection layer by the input Convolutional layer of kernel (7×7), stride (2×2) and padding of 3 as used in ResNet-18 [23] input. Utilizing such projection leads to embeddings of d_c dimensions. Thus, we have applied an additional max pooling layer with kernel and stride of $(1 \times d_c)$ so that we obtain signal embeddings of the required dimensions by the Transformer layer. The results show that we perform poorly when we utilize information from nearby temporal signals for signal embeddings using the Convolutional layer. It happens because the noise from other temporal signals interferes with the rPPG information.

4.5. Discussion

We depict an example of HR estimation by our proposed method in Figure 3. The figure compares the predicted and ground truth pulse for success and failure cases. The first row represents the estimated pulse and its Fourier power spectrum, and the second row shows the ground truth. For the success case in Figure 3 (a), we observe that the estimated pulse signal correlates well with the ground truth. However, we observe a deviation in the estimated pulse signal from the ground truth signal for certain video frames because these frames contain slight noise due to facial movements. In contrast, the effect of large facial movements on the estimated pulse signal can be observed in Figure 3 (b). Comparison of the quality scores for the respective temporal signals suggests that it is an important parameter determining the quality of pulse estimation.

5. Conclusion

The existing deep learning based rPPG approaches have suffered from underfitting on limited datasets and failed in modeling long temporal dependencies leading to incorrect HR estimation. To mitigate these issues, we have provided a Transformer-based pulse estimation method, *RADIANT*. The architecture has benefitted from the attention processing capabilities of the Transformer that allows modeling of long temporal dependencies and effective denoising for correct pulse estimation. Further, it has utilized an MLP projection for obtaining the signal embeddings that provide an adequate subspace for rPPG feature representation. Our experiments justified the isolation of rPPG information in the signal embeddings. We have investigated the possibility of utilizing pre-training and efficient data augmentation techniques to improve generalization capabilities. The experimental results have demonstrated that our architecture provides better results when pre-trained on the synthetic dataset. Also, it indicates that data augmentation allows our architecture to generalize well over the normal HR range. Our results on extensively utilized datasets have shown that our architecture outperforms previous well-known rPPG methods. In the future, we will be working on providing an end-to-end HR estimation network by automating the temporal signal extraction process.

Acknowledgement: This research is supported partially by SERB, DST and project number is SRG/2020/001383. The work of Anup Kumar Gupta is partially supported by Prime Minister's Research Fellowship (PMRF), the Ministry of Education, Government of India (2101306).

References

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, abs/1607.06450, 2016.

[2] Guha Balakrishnan, Frédo Durand, and John V. Guttag. Detecting Pulse from Head Motions in Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3430–3437. IEEE, 2013.

[3] Lokendra Birla and Puneet Gupta. AND-rPPG: a novel denoising-rPPG network for improving remote heart rate estimation. *Computers in Biology and Medicine*, page 105146, 2021.

[4] Lokendra Birla and Puneet Gupta. PATRON: Exploring respiratory signal derived from non-contact face videos for face anti-spoofing. *Expert Systems with Applications*, 187:115883, 2022.

[5] Lokendra Birla, Puneet Gupta, and Shravan Kumar. SUNRISE: Improving 3d mask face anti-spoofing for short videos using pre-emptive split and merge. *IEEE Transactions on Dependable and Secure Computing*, 2022.

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

[7] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video. *Applied Sciences*, 9(20):4364, 2019.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[9] Weixuan Chen and Daniel McDuff. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In *European Conference on Computer Vision*, pages 349–365. Springer, 2018.

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

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2018.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*, 2020.

[13] Anup Kumar Gupta, Puneet Gupta, and Esa Rahtu. FATALRead-Fooling visual speech recognition models. *Applied Intelligence*, pages 1–16, 2021.

[14] Anup Kumar Gupta, Vardhan Paliwal, Aryan Rastogi, and Puneet Gupta. TRIESTE: translation based defense for text classifiers. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2022.

[15] Anup Kumar Gupta, Aryan Rastogi, Vardhan Paliwal, Fyse Nassar, and Puneet Gupta. D-NEXUS: Defending Text Networks Using Summarization. *Electronic Commerce Research and Applications*, 2022.

[16] Puneet Gupta. MERASTC: Micro-expression Recognition using Effective Feature Encodings and 2D Convolutional Neural network. *IEEE Transactions on Affective Computing*, pages 1–1, 2021.

[17] Puneet Gupta. PERSIST: Improving micro-expression spotting using better feature encodings and multi-scale gaussian tcn. *Applied Intelligence*, pages 1–15, 2022.

[18] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. Accurate heart-rate estimation from face videos using quality-based fusion. In *International Conference on Image Processing*, pages 4132–4136. IEEE, 2017.

[19] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. Serial fusion of Eulerian and Lagrangian approaches for accurate heart-rate estimation using face videos. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2834–2837. IEEE, 2017.

[20] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. Exploring the Feasibility of Face Video Based Instantaneous Heart-Rate for Micro-Expression Spotting. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1316–1323. Computer Vision Foundation / IEEE, 2018.

[21] Puneet Gupta, Brojeshwar Bhowmick, and Arpan Pal. MOMBAT: heart rate monitoring from face video using pulse modeling and Bayesian tracking. *Computers in Biology and Medicine*, 121:103813, 2020.

[22] Puneet Gupta, Brojeshwar Bhowmik, and Arpan Pal. Robust Adaptive Heart-Rate Monitoring Using Face Videos. In *IEEE Winter Conference on Applications of Computer Vision*, pages 530–538. IEEE, 2018.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

[24] Dan Hendrycks and Kevin Gimpel. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. *arXiv preprint arXiv:1606.08415*, 2016.

[25] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation. *Handbook of Digital Face Manipulation and Detection - From DeepFakes to Morphing Attacks*, pages 255–273, 2021.

[26] Guillaume Heusch, André Anjos, and Sébastien Marcel. A Reproducible Study on Remote Heart Rate Measurement. *arXiv preprint arXiv:1709.00962*, 2017.

[27] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Dep learning with time-frequency representation for pulse estimation from facial videos. In *IEEE International Joint Conference on Biometrics*, pages 383–389. IEEE, 2017.

[28] Min Hu, Fei Qian, Dong Guo, Xiaohua Wang, Lei He, and Fuji Ren. ETA-rPPGNet: Effective Time-Domain Attention Network for Remote Heart Rate Measurement. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

[29] Bin Huang, Weihai Chen, Chun-Liang Lin, Chia-Feng Juang, Yuanping Xing, Yanting Wang, and Jianhua Wang. A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks. *Engineering Applications of Artificial Intelligence*, 106:104447, 2021.

[30] Jiaqi Kang, Su Yang, and Weishan Zhang. TransPPG: Two-stream Transformer for Remote Heart Rate Estimate. *arXiv preprint arXiv:2201.10873*, 2022.

[31] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey. *ACM Computing Surveys*, 2021.

[32] Antony Lam and Yoshinori Kuno. Robust Heart Rate Measurement from Video Using Select Random Patches. In *IEEE International Conference on Computer Vision*, pages 3640–3648. IEEE, 2015.

[33] 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*, pages 392–409. Springer, 2020.

[34] Menghan Li, Bin Huang, and Guohui Tian. A comprehensive survey on 3d face recognition methods. *Engineering Applications of Artificial Intelligence*, 110:104669, 2022.

[35] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. *Advances in Neural Information Processing Systems*, pages 19400–19411, 2020.

[36] Xinhua Liu, Wenqian Wei, Hailan Kuang, and Xiaolin Ma. Heart Rate Measurement Based on 3D Central Difference Convolution with Attention Mechanism. *Sensors*, 22:688, 2022.

[37] Saumya Mishra, Anup Kumar Gupta, and Puneet Gupta. DARE: Deceiving audio–visual speech recognition model. *Knowledge-Based Systems*, 232:107503, 2021.

[38] Mina Chookhachizadeh Moghadam, Ehsan Masoumi, Samir Kendale, and Nader Bagherzadeh. Predicting hypotension in the ICU using noninvasive physiological signals. *Computers in Biology and Medicine*, page 104120, 2021.

[39] 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.

[40] 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, 2019.

[41] 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*, pages 295–310. Springer, 2020.

[42] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust Remote Heart Rate Estimation from Face Utilizing Spatial-temporal Attention. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8. IEEE, 2019.

[43] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention. *arXiv preprint arXiv:2010.07770*, 2020.

[44] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai. A novel skin color model in YCbCr color space and its application to human face detection. In *International Conference on Image Processing*, volume 1, pages I–I. IEEE, 2002.

[45] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, pages 10762–10774, 2010.

[46] Ying Qiu, Yang Liu, Juan Sebastian Arteaga-Falconi, Haiwei Dong, and Abdulmotaleb El-Saddik. EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video. *IEEE Transactions on Multimedia*, 21(7):1778–1787, 2018.

[47] Ambareesh Revanur, Ananyananda Dasari, Conrad S Tucker, and Laszlo A Jeni. Instantaneous Physiological Estimation using Video Transformers. *arXiv preprint arXiv:2202.12368*, 2022.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[49] Lizawati Salahuddin, Jaegeol Cho, Myeong Gi Jeong, and Desok Kim. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4656–4659. IEEE, 2007.

[50] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*, 2014.

[51] Rencheng Song, Senle Zhang, Juan Cheng, Chang Li, and Xun Chen. New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method. *Computers in Biology and Medicine*, 116:103535, 2020.

[52] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual Heart Rate Estimation with Convolutional Neural Network. In *British Machine Vision Conference*, pages 3–6. BMVA Press, 2018.

[53] Toshiyo Tamura. Current progress of photoplethysmography and SPO2 for health monitoring. *Biomedical Engineering Letters*, pages 21–36, 2019.

[54] Ruben Tolosana, Sergio Romero-Tapiador, Ruben Vera-Rodriguez, Ester Gonzalez-Sosa, and Julian Fierrez. Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. *Engineering Applications of Artificial Intelligence*, 110:104673, 2022.

[55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates Inc., 2017.

[57] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, pages 21434–21445, 2008.

[58] 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, 2015.

[59] Xusheng Wang, Xing Chen, and Congjun Cao. Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, page 115831, 2020.

[60] M Yang, J Liu, Y Xiao, and H Liao. 14.4 nw fourth-order bandpass filter for biomedical applications. *Electronics letters*, 46(14):973–974, 2010.

[61] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-Video-Based Physiological Signal Measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021.

[62] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection. In *International Conference on Computer Vision Workshops*, pages 2519–2528. IEEE Computer Society, 2017.