The Benefit of Distraction: Denoising Camera-Based Physiological Measurements using Inverse Attention

Ewa M. Nowara*, Daniel McDuff†, Ashok Veeraraghavan*
*Rice University, Houston, TX
†Microsoft Research, Redmond, WA
{emn3,vashok}@rice.edu, damcduff@microsoft.com

Abstract

Attention networks perform well on diverse computer vision tasks. The core idea is that the signal of interest is stronger in some pixels ("foreground"), and by selectively focusing computation on these pixels, networks can extract subtle information buried in noise and other sources of corruption. Our paper is based on one key observation: in many real-world applications, many sources of corruption, such as illumination and motion, are often shared between the "foreground" and the "background" pixels. Can we utilize this to our advantage? We propose the utility of inverse attention networks, which focus on extracting information about these shared sources of corruption. We show that this helps to effectively suppress shared covariates and amplify signal information, resulting in improved performance. We illustrate this on the task of camera-based physiological measurement where the signal of interest is weak and global illumination variations and motion act as significant shared sources of corruption. We perform experiments on three datasets and show that our approach of inverse attention produces state-of-the-art results, increasing the signal-to-noise ratio by up to 5.8 dB, reducing heart rate and breathing rate estimation errors by as much as 30 %, recovering subtle waveform dynamics, and generalizing from RGB to NIR videos without retraining.

1. Introduction

Attention mechanisms have been successfully applied in many areas of machine learning and computer vision [25, 45], including object detection [32], activity recognition [37], language tasks [1, 49], machine translation [2], and camera-based physiological measurement [5]. Attention networks often perform well because they can identify pixels that are most likely to contain strong signals of interest. By focusing on pixels useful for the task of interest and ignoring the remaining regions, attention networks are often robust to diverse sources of variations in a video.

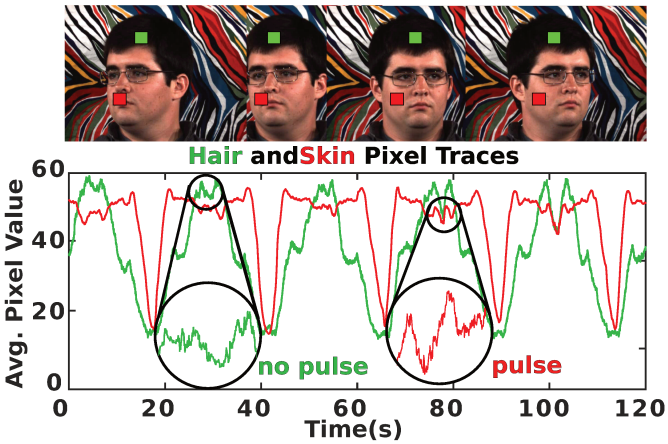

Figure 1. Temporal changes of hair (green) and skin (red) pixel intensities in a video are often correlated, e.g., when large head motion is present. Physiological signals are very subtle and strongest in the skin but are easily corrupted. We propose an approach using the regions ignored by most attention mechanisms (such as hair) to provide estimates of these corruptions and learn a denoising mapping to remove them from the physiological signal of interest.

In this paper, we refer to the regions containing the signal of interest as the "foreground" and the other regions as the "background". We focus on a counter-intuitive question – is there important information contained within the "background" regions that are typically ignored by attention models? And, can we exploit the information in those regions to improve the quality of estimation for the underlying signals of interest in the "foreground"? If noise or variations not related to the signal of interest are present in the video, they will likely corrupt the signal of interest. If these corruptions are random, then keeping as many "foreground" pixels as possible and ignoring the noisy "background" pixels is sufficient for a model to work well. However, these variations are often not random, but rather they are caused by a specific source which likely similarly affects multiple regions in the video.

To illustrate the effectiveness of using the inverse atten-

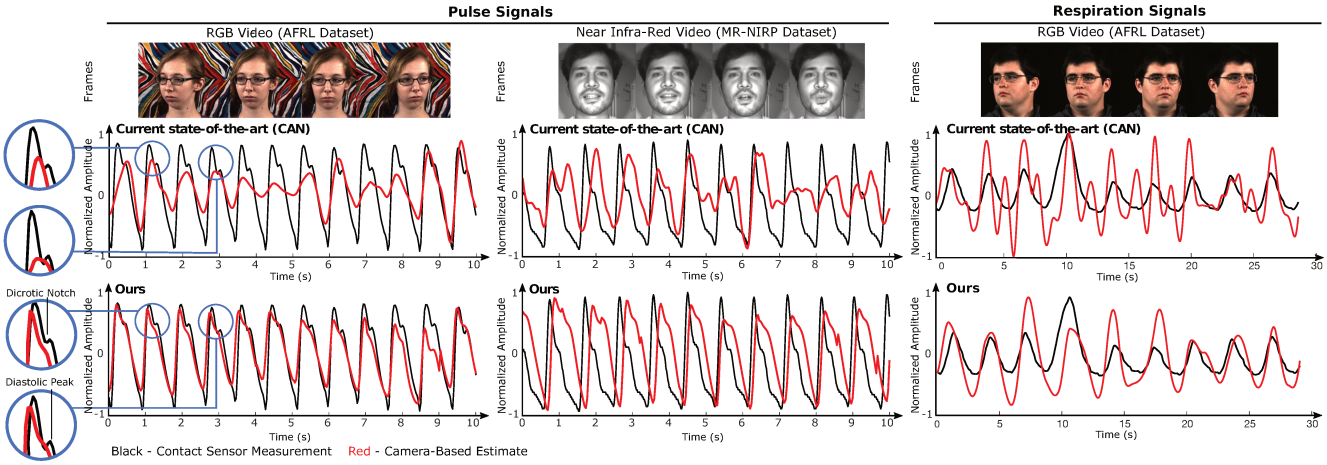

Figure 2. Pulse and breathing signals output by a state-of-the-art CAN network and our denoising method (both shown in red). Our method produces cleaner signals, free from motion artifacts (still present in the benchmark method), better matching the ground truth signal's (shown in black) subtle dynamics and shape. Notice the zoomed-in portions in the pulse waveforms with easily identifiable dicrotic notch and diastolic peaks in our outputs which are absent in the benchmark output.

tion for denoising temporal signals, we focus on the prediction problem of camera-based physiological measurement as an exemplar application for our approach. Physiological signals derived from a video are extremely subtle and are easily corrupted by any variation in a video that may alter the recorded image intensities. Therefore, this is a good and challenging application to illustrate the denoising capability of our method. For example, large head motion will often similarly affect skin regions in the "foreground" as well as several regions in the "background", such as the hair or the wall behind the person (see Fig. 1 for an illustration.) By using the "background" regions, not containing the signal, we can learn about the sources of these corruptions and use that information to suppress changes we are not interested in in the "foreground" pixels. We use an inverse of the attention mask to select the "background" regions and to learn an estimate of the corruptions present in a video.

Our application scenario is motivated by how the SARS-CoV-2 (COVID-19) pandemic has rapidly changed the face of healthcare [3, 38]. Recent research in computer vision has led to the development of camera-based physiological measurement techniques that leverage cameras and computer vision algorithms [40, 46, 34, 7, 48, 5]. Camera-based vital signs could improve the current telemedicine technology, and also enable applications where wearing contact devices for extended periods may be infeasible, such as long-term human-computer-interaction (HCI) studies [23], driver monitoring [30], or face anti-spoofing [17, 31]. Convolutional networks currently provide state-of-the-art performance on heart rate (HR) and breathing rate (BR) measurement from video [5, 50, 18]. While the convolutional neural networks may be able to accurately learn what features in the image are important for finding the physiological signals, they may not be able to learn a good representation of all other variations which may be present in the video but are not related to the signal of interest because of a wide range of factors that can constitute these corruptions. We refer to any variations not related to the physiological signals as "corruptions" because they all degrade the signal quality. In the context of camera-based physiology, these corruptions can be caused by head motion [9], facial expressions [52], speech, ambient light variations [30], video compression artifacts [50, 29], and camera sensor noise [14]. Such corruptions may also vary greatly across videos and datasets. Therefore, it is hard for any model to explicitly capture a good representation of such diverse variations and to remove them from the signals of interest. While the sources of corruptions may not be identical in the "foreground" and in the "background" regions, the variations in different regions of a video are often highly correlated because they are often caused by the same source (e.g., illumination variations from a flickering light bulb or video compression artifacts affecting all regions in a frame).

The key observation we make is that regions ignored by an attention mechanism in a network likely contain information about sources of corruption that are also present in the regions used by the attention mechanism to compute the physiological signals. Using these "distraction" regions that were ignored by the attention masks offers a way to estimate these variations independently for each video without making assumptions about the nature of the sources of the corruptions. The only assumption we make is that most regions ignored by the attention masks do not contain the signals of interest and consequently contain the corruptions that we want to suppress. This assumption should hold true as long as the attention mechanisms are able to segment the video to some degree, which is usually the case.

We demonstrate that regions outside of the attention

mask can be used to estimate the irrelevant intensity variations which corrupt the signal of interest. Once we have an estimate of those variations, we can learn a denoising mapping to remove them from the recovered signals. Our approach outperforms state-of-the-art methods on three datasets across a range of HR and BR error measures and also generalizes well to new data, even data recorded with different imaging modalities, such as near-infrared (NIR), without any additional training. Our proposed approach can even recover very subtle waveform dynamics, such as the clearly visible dicrotic notch and diastolic peaks, shown in Fig. 2, which is currently challenging for video-based methods. Obtaining clean and more accurate waveforms is useful for determining important health metrics, such as blood pressure [8], which is infeasible with most existing methods. Our method also obtained cleaner breathing signals compared to the baseline (Fig. 2). The idea of using the inverse attention regions is likely very useful in a wide range of vision tasks, where the attention networks are used to make temporal predictions, such as activity recognition or video deblurring. However, in this work, we only focused on the physiological measurement application.

The core contributions of this paper are to: (1) propose the use of inverse attention masks for generating estimates of variations which corrupt the signals of interest, (2) present a novel method for denoising camera-based physiological measurement using this approach, (3) evaluate our method on three datasets showing state-of-the-art performance on pulse and breathing measurement, (4) demonstrate that our approach generalizes to NIR data without further training. Supplementary material, including code, models, video examples, and additional experimental results, are provided with this submission.[1]

2. Related Work

Attention Mechanisms. Attention mechanisms provide a way for a model to learn which parts of an image or a video "are relevant for the task at hand and attach a higher importance to them" [37]. During training, the attention weights are learned reflecting the importance of the embedding features. Recently, transformer models, based solely on attention mechanisms, have become popular [45]. In convolutional neural networks (CNNs) these attention mechanisms typically form a spatial mask. These masks can help practitioners understand the decision-making process of a network [11]. In certain cases, the "fixations" of the attention masks generated by computer models and by human observers were very similar [32]. Attention mechanisms can be used to connect layers; for example, one focuses on temporal information (e.g., trained on flows) and another focuses on spatial information (e.g., trained on RGB frames). Prior work has found that these crosslink layers guide the spatial-stream to pay more attention to the human foreground areas and can be less affected by background clutter [43]. In physiological measurement, two-layer networks have been found to be effective as both color and motion information are valuable for extracting the subtle physiological signal in the presence of corruptions [5]. While attention mechanisms often work well, they are a simple representation of which regions are important. However, pixels outside these regions may provide useful context or a strong prior about the corruptions present.

Camera-Based Physiology. Volumetric changes in blood over time lead to subtle changes in light reflected from the skin and subtle motion variations which can be measured with a camera [40, 46]. The physiological signal obtained from a video can be used to recover several metrics and vital signs, including heart rate [34], heart rate variability [35], breathing rate [35], blood oxygenation [41] and pulse transit time [36]. NIR [30, 4] and thermal [12, 33] cameras have also been successfully used for measuring physiological signals in the dark. While there has been great progress in measuring cardio-pulmonary signals in the visible range, estimating these can still be more accurate using thermal cameras [10, 6]. Unfortunately, the signals of interest in camera-based physiological measurement are often very subtle and can be easily corrupted by noise due to body motions and ambient lighting changes. Early work in camera-based physiology used properties of the physiological signal, e.g., the periodic nature [34] and hemoglobin absorption spectra [7, 48] to recover the underlying physiological signal via de-mixing methods [16, 19, 20, 44]. Some of these unsupervised methods make simple assumptions that the pulse signal should be periodic (non-Gaussian) and that any other source signals are noise (e.g., ICA [34]). Others, such as POS [48], assume that the plane orthogonal to skin contains the pulsatile physiological signal and non-orthogonal planes contain specular reflections and noise. Others have used physical skin models to learn a mapping from color changes [24]. In these methods, the corruptions affecting the signals were not modeled explicitly. Recently, several groups have demonstrated that deep learning models free from heuristic assumptions about the signal structure can perform better, especially in presence of large motion and other corruptions [5, 51, 39, 22, 26, 27, 50, 15]. These end-to-end methods did not explicitly define the corruptions either but rather learned to recover the physiological signal in a fully supervised manner. We show that the performance of a state-of-the-art model is significantly improved by using the distraction regions as explicit corruption estimates.

3. Benefiting from Distraction

Intuition. Let us consider a situation where we want to recover a subtle temporal signal, $p(t)$, from a video that has

[1] https://github.com/ewanowara/benefitofdistraction

many additional sources of pixel changes. Each pixel either belongs to a "foreground" region and it contains the signal of interest, $p(t)$, or it belongs to the "background" region and it does not contain $p(t)$. If a pixel is in the "foreground", we can write the intensity of the i^{th} pixel, $y_i(t)$, as:

$$y_i(t) = a_{i,0}(t) + \alpha_i * p(t) + \beta_i * q(t) + \gamma_i * n(t)$$

where a_0 is the base intensity of the video, $p(t)$ is the signal of interest, $q(t)$ are the corruptions correlated in the "foreground" and in the "background", and $n(t)$ is random camera sensor noise. α, β, and γ modulate the strength of the signal $p(t)$, of the correlated corruptions, $q(t)$, and of the random noise, $n(t)$, respectively.

The "foreground" in our application predominantly refers to skin pixels on the face with the physiological signal, $p(t)$. The signal, $p(t)$, is not present in each pixel of the video with the same strength, e.g., some facial regions may be occluded by facial hair or they may have changes resulting from body motions (e.g., eyes during blinking and mouth during talking) [30, 14]. In the context of convolutional attention networks, the strength of the signal, $p(t)$, at each pixel, α, is equivalent to weights in the learned attention mask for all pixels, showing which regions in the video contain the signal of interest. We may not always know in advance which pixels belong to the "foreground" and which belong to the "background". However, we can assume that all pixels with α larger than a specified threshold in the attention mask should belong to the "foreground".

In addition to the physiological signal, $p(t)$, the intensity of the "foreground" also changes due to other variations, not related to $p(t)$ but affecting the quality of the recovered signal, $p(t)$. These variations may include the changing illumination, or motion of the camera or the person, $q(t)$, and camera sensor noise, $n(t)$. Camera sensor noise, $n(t)$, is random and is usually independent and identically distributed across all pixels. However, corruption $q(t)$ is usually not random nor is it uniformly distributed in the video frame. Instead, it is often statistically correlated with the variations caused by the same source in the "background".

On the other hand, if the pixel belongs to the "background", it will contain similar intensity variations as the "foreground" with the exception that it will not contain the signal of interest, $p(t)$. We consider the "background" to encompass all regions not containing $p(t)$:

$$y_i(t) = a_{i,0}(t) + \beta_i * q(t) + \gamma_i * n(t)$$

The physiological signal strength present in the "foreground" of the video is very small, with sub-pixel level amplitude. So, to extract it we need to identify the presence of the signal in many pixels and combine them into a single estimate to improve the SNR. If we can identify the "foreground" pixels which contain $p(t)$ and ignore other pixels, as is done by the attention networks, we might obtain a good estimate. The SNR of $p(t)$ obtained from the "foreground" regions in this manner will depend on the strength of $p(t)$ measured by α and the amount of corruption and random noise measured by β and γ:

$$SNR(p) = \frac{\alpha_i}{\beta_i + \gamma_i}$$

It is usually hard to remove $q(t)$ directly from the "foreground" regions selected by the attention masks because this corruption can be caused by diverse sources which are hard to model and suppress. But it is easier to estimate the related $q(t)$ present in the "background", which we can define to be any variation in the video that is not related to $p(t)$. The corruptions in the "foreground" and in the "background" may not be identical because there may be different variations in these regions of the video. However, $q(t)$ present in the "foreground" and in the "background" are often caused by the same source (e.g., the motion of the head affecting the skin, considered to be the "foreground", and hair pixels, considered to be the "background") and their variations will be similar. Therefore, if we can use only the "background" pixels to estimate the correlated variations $q(t)$ and their strength, β, we could suppress those variations in the "foreground", thereby increasing the SNR of $p(t)$ which now will be predominantly affected by the random noise:

$$SNR(p) \approx \frac{\alpha_i}{\gamma_i}$$

See Fig. 3 for an example of a signal denoised with our approach jointly using the attention and inverse attention masks compared to a signal obtained with a baseline using only the attention masks. While the corruptions, $q(t)$, in the "foreground" and in the "background" are highly correlated, their relationship may be non-linear and it is difficult to model it explicitly, but it can be learned with a deep learning model. We use an LSTM network to learn to suppress the corruptions, $q(t)$, in a video, given an estimate of the corruptions present in the "background". The proposed architecture is shown in Fig. 4. In practice, the correlation between $q(t)$ in the "foreground" and in the "background" is not perfect and β cannot be perfectly estimated. Therefore, the network can be trained to estimate these as well as possible, but it will not be able to perfectly estimate and remove all variations caused by motion and illumination.

Physiology and Corruption Encoder. Convolutional attention network (CAN) [5] serves as an encoder in our architecture and it provides an estimate of the physiological signal obtained from "foreground" regions and the estimate of the corruptions obtained from the "background" regions. The CAN network consists of two components working together – the appearance and the motion models. The appearance model is trained directly on the input video frames. It

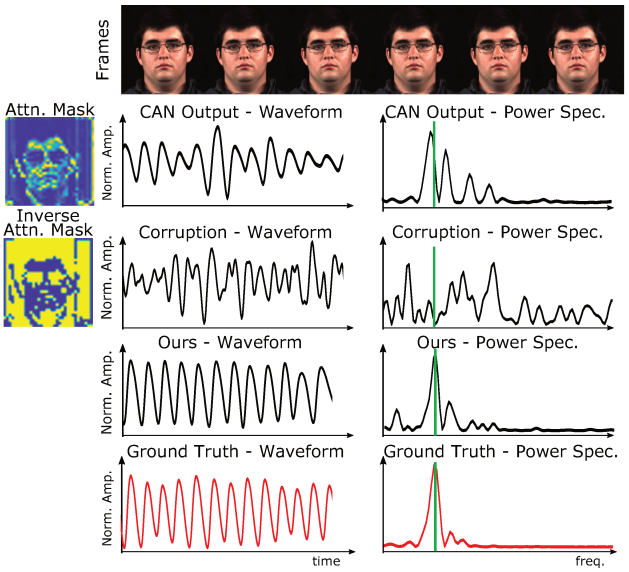

Figure 3. These are examples of attention masks and inverse attention masks used to obtain the initial pulse estimates, estimates of corruptions (only shown from the green camera channel), and the final denoised physiological signals. Higher weights in the masks are indicated by yellow and lower by blue. The ground truth heart rate frequency is shown as the vertical green lines.

learns from the color and texture information which regions in the video are likely to contain strong physiological signals. The motion model is trained on the difference of two consecutive video frames to differentiate between the intensity variations present in the video caused by the characteristic physiological variations from the variations caused by other sources. The attention mask then reflects a heatmap of the strength of the pulsatile physiological signal in each region of the frame. As shown in the first row of Fig. 3, the attention masks mostly focus on skin regions known to have strong physiological signals, while ignoring other regions, such as the eyes, hair, and background regions (see supplementary material for an example of an average attention mask computed over all stationary videos.) The CAN normally outputs a single one-dimensional (1D) physiological signal estimate. However, we perform an element-wise multiplication of the original input frame with the inverse of the attention mask weights to compute a secondary corruption estimate.

The "foreground" pixels can be easily found using the attention masks output by the network because the "foreground" regions are the pixels that the network primarily focuses on to make predictions. In order to estimate the correlated corruptions, $q(t)$, in the "background", we have to find all pixels which belong to the "background". The "background" regions are all the pixels that do not belong to the "foreground", so we can obtain the "background" pixels by creating an inverse of the attention mask. We compute the corruptions at each time step by multiplying the inverse attention masks with each channel of each video frame in an element-wise manner. We then spatially average the resulting weighted pixel intensities:

$$Q_{c,t} = \frac{1}{H}\frac{1}{W}\sum_{x=1}^{H}\sum_{y=1}^{W} I_{x,y,t} \circ M_{x,y,t} \tag{1}$$

where I_t and M_t are the frame and mask at time t. $Q_{c,t}$ is the corruption estimate from each [R, G, B] camera channel c at time t, and H and W are the image height and width, respectively. The attention and the inverse attention masks were 34 $\times$ 34 pixels and the video frames were downsampled to the same size using bicubic interpolation. We normalize the attention mask elements to a range between 0 and 1. To obtain a corruption estimate, we set all values larger than a fixed threshold, T, to 0 and everything else to 1, creating a binary mask. Based on the experiments, we found a threshold of T = 0.1 worked well. This binary inverse attention mask ignores regions in the video initially used to compute the physiological signals and keeps all other regions. Examples of inverse attention masks are shown in the second row of Fig. 3.

Denoising Model. Our denoising model is formed with a long short-term memory (LSTM) network with the encoder providing physiology and corruption inputs at each time step. The goal is to learn a denoising function to clean the physiological estimates, given the estimates of corruptions. As input to the denoising LSTM, we stacked the physiological signal and the corruption signals generated by the encoder. The contact physiological signal (e.g., finger pulse oximeter) was used as the ground truth signal for training. The corruption estimates guide the LSTM to learn which waveform features are related to the irrelevant variations and which ones are related to the physiological signal of interest. The LSTM was able to learn to suppress the diverse corruptions present in the physiological signal and it outputs a cleaner waveform matching the ground truth physiological signal better (see the third row of Fig. 3). See the video provided in the supplementary material for more examples of denoised signals.

In our experiments, we used a two-layer bidirectional LSTM, with 128 hidden units, trained for 10 epochs with Adam optimizer [13] and MSE loss. Because the LSTM tends to work better on shorter sequences, we split each video into sequences of 60 samples, with 50 % overlap between time windows, which corresponded to two seconds for the 30 frames per second (fps) videos. Physiological datasets are often relatively small due to the complexity associated with collecting carefully synchronized physiological signals and high-quality videos. Therefore, we implemented the CAN and the denoising LSTM as two separate networks to reduce the number of training parameters.

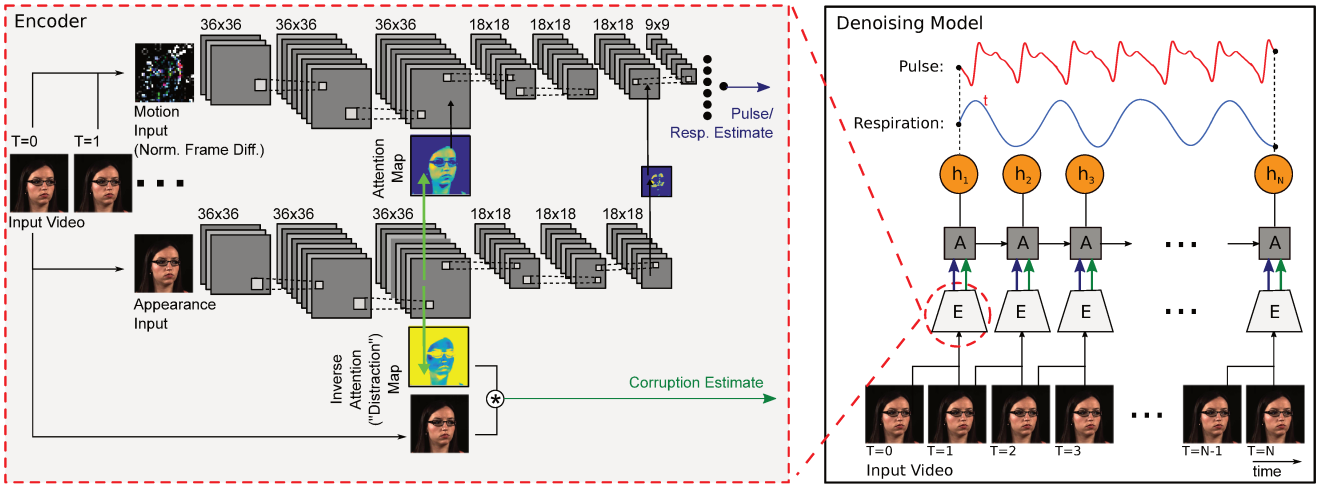

Figure 4. Proposed denoising architecture. The encoder provides the initial physiological signal and the corruption estimates to the LSTM at each time step which outputs a denoised physiological signal.

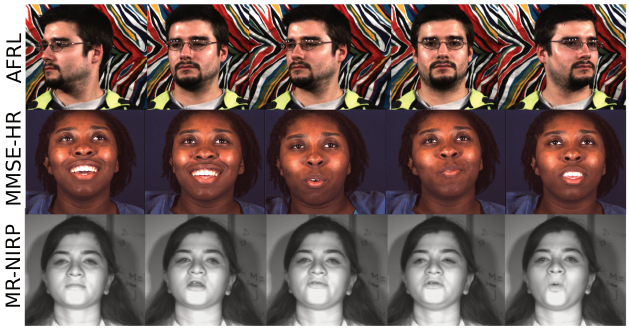

Figure 5. Examples of images used to evaluate our approach.

However, the proposed architecture could be implemented end-to-end, given sufficient training data.

Datasets. We evaluated our approach on two RGB and one NIR video dataset. Example images from each dataset are shown in Fig 5. **AFRL** [9] has 300 videos of 25 participants recorded at 120 fps. Each participant was recorded 12 times in each five-minute experiment with varying motion (increasing from Task 1 to Task 6) and two different backgrounds. We center-cropped the ARFL video frames to 492×492 pixels to remove the blank background areas. **MMSE-HR** [52] has 102 videos of 40 participants recorded at 25 fps during spontaneous emotion elicitation experiments. This dataset is challenging because of the sudden facial motions and rapidly changing heart rate. **MR-NIRP (NIR)** [30] has eight participants recorded with a NIR camera at 30 fps. Each participant was recorded twice, once sitting still and once performing motion tasks involving talking and randomly moving the head. This dataset is particularly challenging because the physiological signals are very weak in NIR [21, 47]. More details about the datasets are provided in the supplementary material.

4. Training Details

Training the Encoder. Due to a large number of parameters, we pretrained the encoder on the largest dataset (AFRL [9]) and locked its weights. When training the encoder, the loss was calculated as the mean squared error (MSE) between the physiological estimate and the ground truth. We performed training and testing separately for each of the six motion tasks from the AFRL dataset with participant-independent cross-validation, leaving out 20% of the participants in each validation split. For experiments on the MMSE-HR and MR-NIRP datasets, we used the trained model from Task 2 as these contained the most similar head position. To maximize the diversity of the participants that this model was trained on to improve its generalizability to new datasets, we instead used subject-dependent cross-validation, using four minutes of each video for training and one minute for testing.

Training the Denoising Model. When evaluating on the AFRL dataset we trained the denoising model with the same subject-independent procedure as for the encoder on AFRL. The MMSE-HR dataset has fewer videos than the AFRL dataset; therefore, we used leave-one-subject-out cross-validation where we left out all videos of one subject and trained the model on all remaining videos, repeating this for each subject. The MR-NIRP dataset was small and not suited for training the networks, so we used the LSTM trained on the AFRL dataset. This allowed us to test the cross-dataset generalization ability of our model.

We detrended [42] and bandpass filtered the signals using a frequency passband range of [0.7 Hz, 2.5 Hz] for HR and [0.08 Hz, 0.5 Hz] for BR. We normalized the signals by subtracting the temporal mean, dividing by the temporal standard deviation in each video, and we normalized their amplitudes to -1 and 1. We resampled all sequences to 30 fps. We estimated HR and BR within 30-second non-

overlapping time windows for signals from each video by finding the frequencies with maximum spectral energy in the respective passbands. We evaluated the performance of our proposed denoising approach across all time windows using mean absolute error (MAE), root mean square error (RMSE), Pearson's correlation coefficient (ρ) between the estimated HR and the ground truth HR, SNR of the estimated physiological signals [7], and waveform mean absolute error (WMAE) computed between the estimated and the ground truth signal. See the supplementary material for the definitions of the error metrics.

5. Results and Discussion

We compared four variants of our proposed approach to nine state-of-the-art methods for recovering the pulse signal [34, 7, 16, 44, 28, 48, 20, 5, 19] and two methods for recovering the breathing signal [5, 41] (see the supplementary material for implementation details). We compared training our model with the corruption estimates obtained from the "background" regions ("Distraction") and without the corruption estimates as input ("No Corr."). We can also directly subtract the corruption estimate from the signal estimate either in the time domain ("Wave. Sub."), or compute the power spectrum of the estimated corruption and signal and subtract the corruption spectrum from the signal spectrum ("Freq. Sub.").

Heart Rate Estimation. Our method achieved better performance compared to previous approaches, including lower HR MAE, RMSE, and waveform MAE and higher HR correlation (ρ) and SNR (see Table 1). On the AFRL dataset, the MAE was reduced from 2.93 beats per minute (BPM) to 2.25 BPM (25% reduction in error), and on the MMSE-HR dataset, the MAE was reduced from 3.74 BPM to 2.27 BPM (39 % reduction in error). This shows that information excluded by the attention mask can be successfully leveraged to remove diverse corruptions, leading to substantial improvements in signal quality. Moreover, the proposed denoising approach is able to recover the subtle waveform dynamics, reducing the waveform MAE by more than 50% on MMSE-HR. While simply subtracting the corruptions from the signals in the frequency domain often improved the SNR, it did not usually improve the heart rate estimates. Subtracting the corruption signal in the time domain performed even worse and often had a negative impact on the SNR. All results were statistically significant ($p < 0.01$) – see supplementary material for F-test results.

Breathing Rate Estimation. In addition to estimating HR, which is based on intensity variations in the skin, our method can also be used to estimate BR which is based on motion variations and it may be more challenging in presence of body motions. Only the AFRL dataset [9] had ground truth breathing signals, therefore we were not able to evaluate our BR results on the other datasets. Our method achieved a reduction in MAE from 3.68 BPM to 2.44 BPM (a 34% error reduction) over the baselines and an increase in SNR by 5.87 dB (Table 1).

True Benefit of Distraction Regions. Using our model without the corruption estimates works well when the signals do not change much over time and when the corruption in the training and test sets is similar. For example, training and testing on AFRL (Table 1) was not very difficult because the head motion was predictable. However, including the distraction regions yielded improvements in both HR and BR estimates when the physiological signal varied abruptly over time or there was a large domain gap between the training and test sets. For example, distraction regions improved the performance on MMSE-HR which has sudden pulse variations, uncontrolled motion, and the presence of facial expressions; and on the more challenging NIR MR-NRIP dataset (Table 1). Moreover, including the distraction regions improved the HR and BR estimation accuracy when we trained our model only on the stationary videos of AFRL (Task 1) which were free of major corruptions and tested on videos with large random motions (Task 6), as shown in Table 2. The SNR was often higher in the "No Corr." condition because the LSTM simply produced a smoother signal leading to greater sparsity in the frequency domain and higher SNR. However, the dominant frequency of that signal was often erroneous, resulting in worse MAE, RMSE, and ρ. These results show that the corruption estimates are useful beyond including an initial signal estimate alone to the model.

Transfer Learning. NIR videos of MR-NIRP are more challenging than RGB because the physiological signal is an order of magnitude weaker in the NIR range compared to the visible range, making it very prone to motion artifacts. When trained solely on RGB videos (AFRL dataset) without any fine-tuning, our method outperformed all the baselines across all five metrics on the NIR videos from the MR-NIRP dataset. As shown in Table 1, the MAE dropped from 7.78 BPM to 2.34 BPM (70% reduction in error). Other baseline methods require multiple color channels and therefore could not be compared on NIR videos.

Varying Head Motion. Our method also showed improvements on videos across all head motions of AFRL [9] (see Table 3). For instance, on videos with an angular head rotation of 30 deg/sec (Task 4) the HR MAE was reduced from 2.82 BPM to 1.94 BPM (30% reduction in error) and BR MAE was reduced from 4.85 BPM to 2.88 BPM (41 % reduction in error).

Performance by Different Skin Type. We have also broken down the results on MMSE-HR by skin type. Dark skin types (V - VI) are more challenging because they have lower iPPG SNR (see supplementary materials). Our method achieves better performance across all skin types and especially darker skin types (MAE [BPM] on skin type

Table 1. Including the "distraction" regions improves heart rate (HR) and breathing rate (BR) results estimation.

	Heart Rate															Breathing Rate				
	AFRL					MMSE-HR					MR-NIRP(NIR)					AFRL				
Method	MAE	RMSE	SNR	ρ	WMAE	MAE	RMSE	SNR	ρ	WMAE	MAE	RMSE	SNR	ρ	WMAE	MAE	RMSE	SNR	ρ	WMAE
Distraction	2.25	5.68	6.44	0.87	0.21	**2.27**	**4.90**	**5.00**	**0.94**	**0.19**	**2.34**	**4.46**	2.27	**0.85**	0.45	**2.44**	**4.23**	**14.20**	**0.35**	0.28
No Corr.	**2.12**	**5.37**	**6.86**	**0.88**	**0.21**	2.80	6.36	4.30	0.90	0.21	2.56	5.23	**2.28**	0.80	0.40	2.49	4.26	14.06	0.34	**0.27**
Freq. Sub.	2.92	6.67	3.66	0.82	0.24	3.97	9.93	4.49	0.76	0.57	8.58	17.59	-4.56	-0.11	**0.31**	5.03	7.45	7.78	0.12	0.31
Wave. Sub.	2.92	6.66	3.09	0.82	0.24	6.09	10.84	-4.75	0.71	0.55	8.83	17.00	-4.69	-0.17	**0.31**	4.98	7.40	7.76	0.12	0.30
MAICA [19]	–	–	–	–	–	3.91	–	–	0.86	–	–	–	–	–	–	–	–	–	–	–
RhythmNet [28]	–	–	–	–	–	–	5.49	–	0.84	–	–	–	–	–	–	–	–	–	–	–
PVM [20]	–	–	–	–	–	4.38	–	–	0.82	–	–	–	–	–	–	–	–	–	–	–
CAN [5]	2.93	6.69	3.36	0.82	0.23	4.06	9.51	0.63	0.77	0.52	7.78	16.8	-3.24	-0.03	0.36	4.86	7.32	8.33	0.10	0.27
POS [48]	4.36	9.45	0.73	0.74	0.45	3.90	9.61	2.33	0.78	0.39	–	–	–	–	–	–	–	–	–	–
Tulyakov [44]	–	–	–	–	–	–	11.37	–	0.71	–	–	–	–	–	–	–	–	–	–	–
Li [16]	–	–	–	–	–	–	19.95	–	0.38	–	–	–	–	–	–	–	–	–	–	–
Tarassenko [41]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.68	5.52	-6.22	0.29	0.29
CHROM [7]	4.07	9.72	0.29	0.72	0.41	3.74	8.11	1.90	0.82	0.37	–	–	–	–	–	–	–	–	–	–
ICA [34]	5.78	11.80	0.42	0.58	0.43	5.44	12.0	3.03	0.66	0.42	–	–	–	–	–	–	–	–	–	–

Table 2. Training on AFRL Task 1 and testing on Task 6. The ignored regions help when the training and test set are very different.

	Heart Rate					Breathing Rate				
Method	MAE	RMSE	SNR	ρ	WMAE	MAE	RMSE	SNR	ρ	WMAE
Distraction	**5.29**	**9.33**	-2.07	**0.70**	0.32	**4.28**	**6.00**	5.93	**0.10**	0.34
No Corr.	5.61	9.72	**-1.91**	0.67	0.32	4.38	6.15	**5.96**	0.07	0.34

Table 3. Motion increasing from 1 to 6 on AFRL

	Heart Rate MAE						Breathing Rate MAE					
Method	1	2	3	4	5	6	1	2	3	4	5	6
Distraction	**1.06**	2.11	**1.79**	**1.94**	2.50	4.78	**1.42**	**1.86**	1.88	**2.88**	**2.87**	**4.15**
No Corr.	1.14	1.90	1.80	3.39	**2.04**	**4.52**	1.47	1.95	**1.68**	2.96	2.99	**4.15**
Freq. Sub.	1.52	2.62	2.51	3.00	2.58	5.30	4.30	5.35	4.89	5.27	5.09	5.26
Wave. Sub.	1.57	2.59	2.53	3.03	2.72	5.09	4.31	5.24	4.88	5.17	5.08	5.19
CAN [5]	1.52	2.61	2.51	3.00	2.62	5.34	4.24	5.17	4.58	5.09	4.92	5.15
POS [48]	1.42	**1.52**	2.84	3.86	6.33	10.16	–	–	–	–	–	–
CHROM [7]	1.33	1.62	2.87	2.82	3.91	11.86	–	–	–	–	–	–
ICA [34]	2.18	2.64	4.74	4.93	7.02	13.18	–	–	–	–	–	–
Tarassenko [41]	–	–	–	–	–	–	2.51	2.53	3.19	4.85	4.22	4.78

VI: Ours = 1.57, CAN = 8.77).

Inverse Mask Definition. We tested computing the inverse attention mask used to estimate the corruptions as continuous or as binary values after thresholding. We also compared using all three and individual RGB channels to estimate the corruptions. However, we obtained comparable results with different variants of the inverse attention masks (see supplementary materials).

Importance of Different Distraction Regions. Certain regions in the video may contain more useful information about the sources of corruptions than others. For example, regions closer to the face may contain more information about the motion of the participant. We compared separately using distraction regions closer to the face (center of the frames) and further from the face (edges of the frames). When the motion was small, all regions contributed similarly to denoising (MAE = 1.08 BPM with center regions and MAE = 1.07 BPM with edges). But when there was large head motion, regions close to the head (center of the frames) helped more (MAE = **6.53** BPM with center regions and MAE = 8.74 BPM with edges). See supplementary materials for detailed results.

Performance on Subjects with Glasses. Interestingly, we observed that our method performed very well on subjects who wore glasses. The attention masks for subjects with and without glasses were comparably good. However, CAN performed worse on subjects with glasses and our approach offered a large improvement on those videos (MAE [BPM] with glasses: Ours = **2.17**, CAN = 3.33, and without glasses: Ours = **2.55**, CAN = 2.57). See supplementary materials for example attention masks and additional results.

6. Conclusion

We have presented a novel approach for generating corruption estimates from inverse attention masks to improve camera-based physiological signal measurements. We hypothesized that the corruptions affecting regions used by the attention masks to compute the signal of interest would likely be present in other regions in the video that are typically ignored by the attention masks. Our proposed denoising method outperformed all state-of-the-art methods in heart rate and breathing rate estimation from videos. The recovered physiological signals were sufficiently clean to recover even subtle waveform dynamics present in the ground truth contact signals, including the dicrotic notch and the diastolic peaks. Moreover, our approach trained only on RGB videos showed strong cross-dataset and cross-modality generalizability, outperforming the existing methods on challenging NIR videos.

Acknowledgments

Ewa Nowara and Ashok Veeraraghavan were partially supported by the NSF SaTC Award CNS-1801372, NSF Expeditions Award CCF-1730574, and NSF PATHS-UP Award EEC-1648451.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[3] Bastiaan R Bloem, E Ray Dorsey, and Michael S Okun. The coronavirus disease 2019 crisis as catalyst for telemedicine for chronic neurological disorders. *JAMA neurology*, 2020. 2

[4] Weixuan Chen, Javier Hernandez, and Rosalind W Picard. Estimating carotid pulse and breathing rate from near-infrared video of the neck. *Physiological measurement*, 39(10):10NT01, 2018. 3

[5] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 1, 2, 3, 4, 7, 8

[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 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 456–463. IEEE, 2017. 3

[7] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 3, 7, 8

[8] Mohamed Elgendi, Richard Fletcher, Yongbo Liang, Newton Howard, Nigel H Lovell, Derek Abbott, Kenneth Lim, and Rabab Ward. The use of photoplethysmography for assessing hypertension. *NPJ digital medicine*, 2(1):1–11, 2019. 3

[9] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1462–1469. IEEE, 2014. 2, 6, 7

[10] Jin Fei and Ioannis Pavlidis. Thermistor at a distance: unobtrusive measurement of breathing. *IEEE Transactions on Biomedical Engineering*, 57(4):988–998, 2009. 3

[11] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019. 3

[12] Marc Garbey, Nanfei Sun, Arcangelo Merla, and Ioannis Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE transactions on Biomedical Engineering*, 54(8):1418–1426, 2007. 3

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[14] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015. 2, 4

[15] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. *arXiv preprint arXiv:2007.06786*, 2020. 3

[16] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3, 7, 8

[17] Siqi Liu, Xiangyuan Lan, and PongChi Yuen. Temporal similarity analysis of remote photoplethysmography for fast 3d mask face presentation attack detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2608–2616, 2020. 2

[18] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020. 2

[19] Richard Macwan, Yannick Benezeth, and Alamin Mansouri. Heart rate estimation using remote photoplethysmography with multi-objective optimization. *Biomedical Signal Processing and Control*, 49:24–33, 2019. 3, 7, 8

[20] Richard Macwan, Serge Bobbia, Yannick Benezeth, Julien Dubois, and Alamin Mansouri. Periodic variance maximization using generalized eigenvalue decomposition applied to remote photoplethysmography estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1332–1340, 2018. 3, 7, 8

[21] Luis F Corral Martinez, Gonzalo Paez, and Marija Strojnik. Optimal wavelength selection for noncontact reflection photoplethysmography. In *22nd Congress of the International Commission for Optics: Light for the Development of the World*, volume 8011, page 801191. International Society for Optics and Photonics, 2011. 6

[22] Daniel McDuff. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1367–1374, 2018. 3

[23] Daniel McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4000–4004. ACM, 2016. 2

[24] Daniel J McDuff, Ethan B Blackford, Justin R Estepp, and Izumi Nishidate. A fast non-contact imaging photoplethysmography method using a tissue-like model. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*, volume 10501, page 105010Q. International Society for Optics and Photonics, 2018. 3

[25] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014. 1

[26] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to

specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3580–3585. IEEE, 2018. 3

[27] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018. 3

[28] 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. 7, 8

[29] Ewa Nowara and Daniel McDuff. Combating the impact of video compression on non-contact vital sign measurement using supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 2

[30] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1353–135309. IEEE, 2018. 2, 3, 4, 6

[31] Ewa Magdalena Nowara, Ashutosh Sabharwal, and Ashok Veeraraghavan. Ppgsecure: Biometric presentation attack detection using photoplethysmograms. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 56–62. IEEE, 2017. 2

[32] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003. 1, 3

[33] I Pavlidis, M Dcosta, S Taamneh, M Manser, T Ferris, R Wunderlich, E Akleman, and P Tsiamyrtzis. Dissecting driver behaviors under cognitive, emotional, sensorimotor, and mixed stressors. *Scientific reports*, 6:25651, 2016. 3

[34] Ming-Zher Poh, Daniel McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 2, 3, 7, 8

[35] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 3

[36] 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 Transactions on Biomedical Engineering*, 61(11):2760–2767, 2014. 3

[37] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015. 1, 3

[38] Xuan Song, Xinyan Liu, and Chunting Wang. The role of telemedicine during the covid-19 epidemic in chinaexperience from shandong province, 2020. 2

[39] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference, Newcastle, UK*, pages 3–6, 2018. 3

[40] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007. 2, 3

[41] L Tarassenko, M Villarroel, A Guazzi, J Jorge, DA Clifton, and C Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014. 3, 7, 8

[42] Mika P Tarvainen, Perttu O Ranta-Aho, and Pasi A Karjalainen. An advanced detrending method with application to hrv analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002. 6

[43] An Tran and Loong-Fah Cheong. Two-stream flow-guided convolutional attention networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3110–3119, 2017. 3

[44] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2396–2404, 2016. 3, 7, 8

[45] 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, 2017. 1, 3

[46] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2, 3

[47] Vytautas Vizbara. Comparison of green, blue and infrared light in wrist and forehead photoplethysmography. *BIOMEDICAL ENGINEERING 2016*, 17(1), 2013. 6

[48] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 2, 3, 7, 8

[49] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 1

[50] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 151–160, 2019. 2, 3

[51] Qi Zhan, Wenjin Wang, and Gerard de Haan. Analysis of cnn-based remote-ppg to understand limitations and sensitivities. *arXiv preprint arXiv:1911.02736*, 2019. 3

[52] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 2, 6