The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention: Supplementary Materials

ACM Reference Format:
. 2021. The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention: Supplementary Materials. In *Proceedings of (CHIL '21).* ACM, New York, NY, USA, 4 pages.

1 EVALUATION METRICS

To evaluate the performance of our proposed approach we used the following four standard error measures (MAE, RMSE, Correlation, SNR), and we defined a new measure (Waveform MAE) to measure the waveform dynamics.

Mean absolute error (MAE):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |R_i - \widehat{R}_i|}{N} \tag{1}$$

where N is the total number of time windows, R_i is the ground truth heart rate (HR) measured with a contact sensor for each 30 second time window and $\widehat{R}_i$ is the estimated HR from the video.

Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (R_i - \widehat{R}_i)^2}{N}} \tag{2}$$

Pearson's Correlation Coefficient (ρ): computed between HR estimates from each time window $\widehat{R} = [\widehat{R})(1), ..., \widehat{R}(N)]$ and the ground truth HR measurements $R = [R(1), ..., R(N)]$.

Signal-to-noise ratio (SNR): calculated as the ratio of the area under the curve of the power spectrum around the first and the second harmonic of the ground truth HR frequency divided by the rest of the power spectrum within the physiological range of 42 to 240 bpm [5]:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{42}^{240}((U_t(f)S(f))^2}{\sum_{42}^{240}((1-U_t(f))S(f))^2}\right) \tag{3}$$

where S is the power spectrum of the estimated iPPG signal, f is the frequency in beats per minute (BPM) and $U_t(f)$ is equal to one for frequencies around the first and second harmonic of the ground truth HR (HR-6 bpm to HR+6 bpm and 2*HR-6 bpm to 2*HR+6 bpm), and 0 everywhere else.

Waveform Mean Absolute Error (WMAE):

$$\mathrm{WMAE} = \frac{\sum_{i=1}^{N} |W_i - \widehat{W}_i|}{N} \tag{4}$$

where W_i is the ground truth pulse waveform obtained with the contact sensor for each 30 second time window and $\widehat{W}_i$ is the estimated pulse waveform from the video.

2 BASELINE METHODS

We compared the performance of our proposed approach to state-of-the-art supervised method using a convolutional attention network (CAN) and three unsupervised methods described below.

For the CHROM, ICA and POS methods face detection was first performed using MATLAB's face detection (`vision.CascadeObjectDetector()`). This was fixed for all methods, to avoid the influence of the face detector on performance. For the CAN method following the implementation in [4] we did not use face detection but rather we passed the full frame to the network after cropping the center portion to make the frame a square with W=H.

CHROM [5]. This method uses a linear combination of the chrominance signals obtained from the RGB video. The [x_R, x_G, x_B] signals are filtered using a zero-phase, 3rd-order Butterworth bandpass filter with pass-band frequencies of [0.7 2.5] Hz. Following this, a moving window method of length 1.6 seconds (with overlapping windows and a step size of 0.8 seconds) is applied. Within each window the color signals are normalized by dividing by their mean value to give [$\bar{x_r}$, $\bar{x_g}$, $\bar{x_b}$]. These signals are bandpass filtered using zero-phase forward and reverse 3rd-order Butterworth filters with pass-band frequencies of [0.7 2.5] Hz. The filtered signals [y_r, y_g, y_b] are then used to calculate S_{win}:

$$S_{win} = 3(1-\frac{\alpha}{2})y_r - 2(1+\frac{\alpha}{2})y_g + \frac{3\alpha}{2}y_b \tag{5}$$

Where α is the ratio of the standard deviations of the filtered versions of A and B:

$$A = 3y_r - 2y_g \tag{6}$$

$$B = 1.5y_r + y_g - 1.5y_b \tag{7}$$

The resulting outputs are scaled using a Hanning Window and summed with the subsequent window (with 50% overlap) to construct the final blood volume pulse (BVP) signal.

ICA [29]. This approach involves spatial averaging the pixels by color channel in the region of interest (ROI) for each frame to form time varying signals [x_R, x_G, x_B]. Following this, the observation

CHIL '21, ACM Conference on Health, Inference, and Learning, April 8–10, 2021,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00

signals are detrended. A Z-transform is applied to each of the detrended signals. The Independent Component Analysis (ICA) (JADE implementation) is applied to the normalized color signals.

POS [42]. The intensity signals $[x_R, x_G, x_B]$ are computed. A moving window of length 1.6 seconds (with overlapping windows and with a step size of one frame), is applied. For each time window, the signal is divided by its mean to give $[\bar{x_r}, \bar{x_g}, \bar{x_b}]$. Following this, X_s and Y_s are calculated where:

$$X_s = \bar{x}_g - \bar{x}_b \tag{8}$$

$$Y_s = -2\bar{x}_r + \bar{x}_g + \bar{x}_b \tag{9}$$

X_s and Y_s are then used to calculate S_{win}, where:

$$S_{win} = X_s + \frac{\sigma(X_s)}{\sigma(Y_s)} Y_s \tag{10}$$

The resulting outputs of the window-based analysis are used to construct the final BVP signal in an overlap add fashion.

CAN [4] Supervised convolutional attention neural network described in detail in the main text [4]. Following the implementation in that paper we did not use face detection but rather we pass the full frame to the network after cropping the center portion to make the frame a square with W=H.

Signal Pre-processing. We bandpass filtered the physiological signals and noise estimates to 0.7 Hz - 2.5 Hz range and detrended them [37] before feeding them into the LSTM. We set the detrending parameter λ for each dataset based on the video frame rate (λ = 500 for AFRL [7] and λ = 50 for MMSE-HR [47] and MR-NIRP [25].). We normalized the signals and noise estimates with AC/DC normalization by subtracting the temporal mean and dividing by the temporal standard deviation computed for each video. We additionally normalized the amplitude range of the signals, noise estimates and the ground truth signals to -1 and 1. Finally, we resampled all sequences to 30 fps.

Statistical Significance. We computed F-tests to verify that our errors had significantly lower variance (spread) than the baselines. For AFRL and MR-NIRP which had longer videos, we computed the error metrics for each video, and for the shorter MMSE-HR, we computed them for all time windows in the dataset. In addition to lower mean errors, for all datasets our approach led to significantly lower spread in the MAE and RMSE. AFRL (300 videos): MAE: F = 0.54, $p < 0.01$, RMSE: F = 0.56, $p < 0.01$, MMSE-HR (131 windows): MAE: F = 0.26, $p < 0.01$, RMSE: F = 3.92 $p < 0.01$, MR-NIRP (15 videos): MAE F = 7.94 $p < 0.01$, RMSE F = 6.63, $p < 0.01$.

3 COMPARISON OF NOISE ESTIMATION

Noise Signal Definition. We compared the performance of our proposed denoising framework with noise channels computed from a single red, green or blue camera channel to using all three R, G, B channels. We hypothesized that the blue channel might be the best one for the noise representation for the physiological signals because the hemoglobin present in blood has the lowest absorption in the blue light spectrum and its intensity variations would be least related to blood flow. Conversely, the green channel could also be a useful noise representation, because it would contain information most similar to the physiological signals since the hemoglobin has the largest absorption in the green spectrum. However, we found that there is not a large difference between using any one of the single channels or all three channels. We report the detailed results in Table 1 on the AFRL dataset [7].

Inverse Mask Definition. We also compared computing noise using a binary and a continuous inverse attention mask. The continuous mask was computed as a matrix of continuous values in which each element of the inverse mask M, $M_{i,j}$, was 1 - $A_{i,j}$ where A is the attention mask weights normalized from 0 to 1. The binary mask was computed by thresholding these values, where $A'_{i,j} = 1$, if $A_{i,j}$ >T, where T is a threshold from 0 to 1. We found that we obtained comparable results with the binary and continuous masks as shown in Table 1.

Table 1: Participant independent performance of pulse measurement on AFRL [7]. There was no systematic benefit of using R, G, B or RGB inputs or using the binary vs. continuous mask. We used the binary mask with RGB inputs for the results shown in the main paper.

	AFRL (All Tasks) [7]				
Method	MAE	RMSE	SNR	ρ	WMAE
Ours (LSTM RGB Binary Mask)	2.25	5.68	6.44	0.87	0.21
Ours (LSTM Red Binary Mask)	2.09	5.19	6.70	**0.89**	0.21
Ours (LSTM Green Binary Mask)	**2.04**	**5.11**	**6.84**	**0.89**	0.21
Ours (LSTM Blue Binary Mask)	2.18	5.27	6.59	0.88	0.21
Ours (LSTM RGB Continuous Mask)	2.10	5.61	**7.11**	0.87	**0.20**

Different Distraction Regions. We compared separately using noise estimates from distraction regions closer to the face ("Center" of the frames) and further from the face ("Edges" of the frames). We used an LSTM model trained on all ignored regions for this experiment. When motion was small, all regions contributed similarly to denoising. But when there was large head motion, regions close to the head (center of the frames) helped the most. See Table 2.

Table 2: Different Distraction Regions on AFRL [7]

	MAE						BVP SNR					
Region	1	2	3	4	5	6	1	2	3	4	5	6
Edges	**1.07**	**2.10**	1.92	2.10	2.68	8.74	**10.52**	7.23	8.59	6.04	3.07	-5.83
Center	1.08	2.11	**1.75**	**2.00**	**2.43**	**6.53**	10.50	**7.28**	**8.72**	**6.33**	**3.89**	**-4.47**

Effect of Glasses. We compared the performance of our denoising approach and the baseline CAN method on subjects with and without glasses. We found that our method offers largest improvements on subjects with glasses, as shown in Table 3. However, the attention masks output by CAN on subjects with and without glasses were comparable, as shown in Figure 1. Nine of the 25 subjects in the AFRL dataset were wearing glasses. No subjects in the MMSE-HR or MR-NIRP datasets were wearing glasses.

Table 3: Effect of Glasses on AFRL [7]

Method	MAE	RMSE	SNR	ρ	WMAE
Ours (LSTM) with Glasses	**2.17**	**4.55**	**7.33**	**0.87**	0.21
CAN with Glasses	3.33	6.56	3.80	0.76	0.24
Ours (LSTM) no Glasses	2.55	5.79	4.68	0.59	**0.20**
CAN no Glasses	2.57	5.13	2.50	0.66	0.22

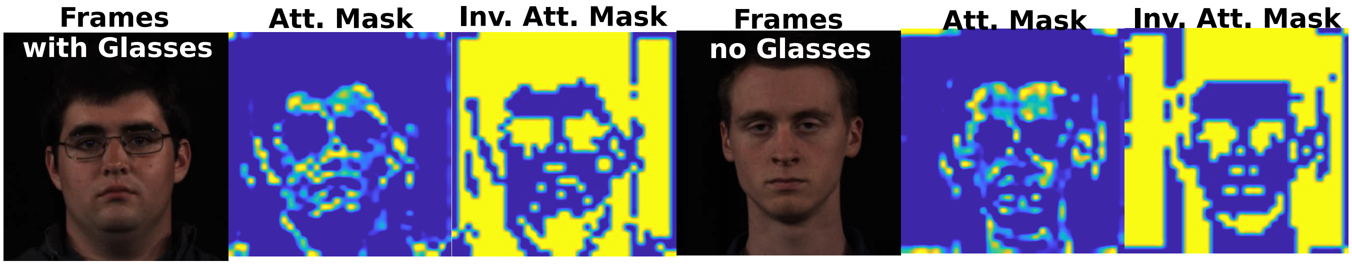

Figure 1: Comparison of attention masks and inverse attention masks on a video with and without glasses.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*. 6077–6086.

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

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

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

[5] Gerard De Haan and Vincent Jeanne. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2878–2886.

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

[7] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. 2014. 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)*. IEEE, 1462–1469.

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

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

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

[12] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. 2014. Remote Heart Rate Measurement From Face Videos Under Realistic Situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Siqi Liu, Xiangyuan Lan, and PongChi Yuen. 2020. Temporal Similarity Analysis of Remote Photoplethysmography for Fast 3D Mask Face Presentation Attack Detection. In *The IEEE Winter Conference on Applications of Computer Vision*. 2608–2616.

[14] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. 2020. Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. *arXiv preprint arXiv:2006.03790* (2020).

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

[16] Richard Macwan, Serge Bobbia, Yannick Benezeth, Julien Dubois, and Alamin Mansouri. 2018. 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*. 1332–1340.

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

[18] Daniel McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. 2016. 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*. ACM, 4000–4004.

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

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

[21] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. 2018. Synrhythm: Learning a deep heart rate estimator from general to specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 3580–3585.

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

[23] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. 2019. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing* 29 (2019), 2409–2423.

[24] Ewa Nowara and Daniel McDuff. 2019. 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*.

[25] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. 2018. 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)*. IEEE, 1353–135309.

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

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

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

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

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

[31] Valentina O Puntmann, M Ludovica Carerj, Imke Wieters, Masia Fahim, Christophe Arendt, Jedrzej Hoffmann, Anastasia Shchendrygina, Felicitas Escher, Mariuca Vasa-Nicotera, Andreas M Zeiher, et al. 2020. Outcomes of cardiovascular magnetic resonance imaging in patients recently recovered from coronavirus disease 2019 (COVID-19). *JAMA cardiology* (2020).

[32] Dangdang Shao, Yuting Yang, Chenbin Liu, Francis Tsow, Hui Yu, and Nongjian Tao. 2014. Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time. *IEEE Transactions on Biomedical Engineering* 61, 11 (2014), 2760–2767.

[33] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* (2015).

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

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

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

[37] Mika P Tarvainen, Perttu O Ranta-Aho, and Pasi A Karjalainen. 2002. An advanced detrending method with application to HRV analysis. *IEEE Transactions*

on Biomedical Engineering 49, 2 (2002), 172–175.

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

[39] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. 2016. 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*. 2396–2404.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

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

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

[43] Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, Lei Huang, Chao Zhang, Shuhong Liu, Peng Zhao, Hongxia Liu, Li Zhu, et al. 2020. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet respiratory medicine* 8, 4 (2020), 420–422.

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

[45] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. 2019. 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*. 151–160.

[46] Qi Zhan, Wenjin Wang, and Gerard de Haan. 2019. Analysis of CNN-based remote-PPG to understand limitations and sensitivities. *arXiv preprint arXiv:1911.02736* (2019).

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

[48] Ying-Ying Zheng, Yi-Tong Ma, Jin-Ying Zhang, and Xiang Xie. 2020. COVID-19 and the cardiovascular system. *Nature Reviews Cardiology* 17, 5 (2020), 259–260.