Supplementary Material

Contrast-Phys: Unsupervised Video-based Remote Physiological Measurement via Contrastive Learning

1 3DCNN Model Architecture

We modify the pooling layer of PhysNet [1] in the last stage of the model so that the model can output an ST-rPPG block with shape $T \times S \times S$ including spatiotemporal information. We use the following table to show our model architecture. For the output ST-rPPG block, the largest spatial dimension length S is 8 due to the previous layer's spatial dimension.

Table 1: 3DCNN Model Architecture

layer	Output Size	Kernel Size	Stride	Pad
conv3d+BN+ELU	(32,T,128,128)	(1,5,5)	(1,1,1)	(0,2,2)
AvgPool3d	(32,T,64,64)	(1,2,2)	(1,2,2)	(0,0,0)
conv3d+BN+ELU	(64,T,64,64)	(3,3,3)	(1,1,1)	(1,1,1)
conv3d+BN+ELU	(64,T,64,64)	(3,3,3)	(1,1,1)	(1,1,1)
AvgPool3d	(64,T/2,32,32)	(2,2,2)	(2,2,2)	(0,0,0)
conv3d+BN+ELU	(64,T/2,32,32)	(3,3,3)	(1,1,1)	(1,1,1)
conv3d+BN+ELU	(64,T/2,32,32)	(3,3,3)	(1,1,1)	(1,1,1)
AvgPool3d	(64,T/4,16,16)	(2,2,2)	(2,2,2)	(0,0,0)
conv3d+BN+ELU	(64,T/4,16,16)	(3,3,3)	(1,1,1)	(1,1,1)
conv3d+BN+ELU	(64,T/4,16,16)	(3,3,3)	(1,1,1)	(1,1,1)
AvgPool3d	(64,T/4,8,8)	(1,2,2)	(1,2,2)	(0,0,0)
conv3d+BN+ELU	(64,T/4,8,8)	(3,3,3)	(1,1,1)	(1,1,1)
conv3d+BN+ELU	(64,T/4,8,8)	(3,3,3)	(1,1,1)	(1,1,1)
Interpolate	(64,T/2,8,8)	-	-	-
conv3d+BN+ELU	(64,T/2,8,8)	(3,1,1)	(1,1,1)	(1,0,0)
Interpolate	(64,T,8,8)	-	-	-
conv3d+BN+ELU	(64,T,8,8)	(3,1,1)	(1,1,1)	(1,0,0)
AdaptiveAvgPool3d	(64,T,S,S)	-	-	-
conv3d	(1,T,S,S)	(1,1,1)	(1,1,1)	(0,0,0)

2 rPPG Spatiotemporal Sampling

We show the detailed procedures of rPPG spatiotemporal sampling in the following algorithm.

Algorithm 1 Spatiotemporal Sampling

Input: ST-rPPG block: P with shape $T \times S \times S$, Number of rPPG samples per spatial location: K, The default rPPG sample length length $\Delta t = T/2$

Initialze an empty list H for storing all rPPG samples
for $h = 1, ..., S$ **do** ▷ Loop over height
  for $w = 1, ..., S$ **do** ▷ Loop over width
    for $k = 1, ..., K$ **do** ▷ K rPPG samples per spatial location
      Randomly choose a starting time t between 0 and $T - \Delta t$
      Append the rPPG sample $P(t \rightarrow t + \Delta t, h, w)$ into the list H
    end for
  end for
end for

Output: The list H containing rPPG samples

3 Datasets

3.1 PURE

There are 10 subjects (8 male, 2 female) in the PURE dataset [2]. The face videos were recorded in 6 different setups for each subject, including steady, talking, slow translation, fast translation, small rotation, and medium rotation. Therefore, there are a total of 60 1-min RGB videos. The videos are uncompressed with 640x480 resolution and 30 fps. A finger pulse oximeter measures the ground truth physiological signal.

3.2 UBFC-rPPG

The UBFC-RPPG (Univ. Bourgogne Franche-Comté Remote PhotoPlethysmoGraphy) database [3] was recorded by a webcam with 640x480 resolution and 30 fps in uncompressed 8-bit RGB format. They recorded the contact PPG waveform and heart rates by a pulse oximeter. Each subject sits 1m away from the camera, and the face is at the center of the video. The lighting includes sunlight and indoor illumination. There are 42 participants, and each participant has one 1-min video.

3.3 OBF

Oulu Bio-face (OBF) dataset [4] has 100 healthy subjects. For each subject, there are two 5-min RGB videos. One video was recorded at a resting status, and the other was recorded after 10 minutes of exercise to cover a wider range of heart rates. Participants are sitting still without head motion and expressions. The contact PPG was recorded simultaneously by a pulse oximeter. All the videos have a resolution of 1920×1080 and a frame rate of 60 fps.

3.4 MR-NIRP

MR-NIRP dataset [5] has near-infrared (NIR) videos from eight participants. The videos were recorded at 640×640 resolution and 30 fps. A pulse oximeter measured the contact PPG signals as the ground truth signals. There are two videos for each participant. One was recorded with the participant sitting still, and another one was recorded with motion tasks, including talking and random head motion. This dataset is difficult because the physiological signals are weak in NIR videos [6, 7]. In addition, the head motion also brings some challenges to measuring rPPG

signals. MR-NIRP is currently the smallest rPPG dataset [8], which makes it challenging for model training.

3.5 MMSE-HR

MMSE-HR [9] has 102 RGB videos from 40 participants with 25 fps and 1040×1392 resolution. Each participant showed facial expression during the recording. The face also has some sudden large motion, and the heart rate will change rapidly. They record blood pressure signals, which can be used as the ground truth physiological signals.

4 Evaluation Metrics

We have a list of measurement results $\hat{m}$ such as heart rate, and a list of ground truth results m. We can also get the error list $e = \hat{m} - m$. We define the following metrics to evaluate the measurement accuracy.

4.1 Mean Absolute Error (MAE)

$$\text{MAE} = \sum_{i=1}^{N} |e_i|/N \tag{1}$$

4.2 Root Mean Squared Error (RMSE)

$$\text{MAE} = \sqrt{\sum_{i=1}^{N} e_i^2/N} \tag{2}$$

4.3 standard deviation (STD)

$$\text{STD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} e_i\right)^2} \tag{3}$$

4.4 Pearson Correlation Coefficient (R)

Pearson correlation is between -1 and +1. When Pearson correlation is close to +1, the measurement is close to the ground truth.

$$L_{rppg}(\hat{y}, y) = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{m}_i m_i - \frac{1}{N}\sum_{i=1}^{N} \hat{m}_i \frac{1}{N}\sum_{i=1}^{N} m_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \hat{m}_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} \hat{m}_i\right)^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N} m_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} m_i\right)^2}} \tag{4}$$

5 Irrelevant Power Ratio (IPR)

IPR is used in [8] to evaluate rPPG signal quality during training. This metric does not need ground truth signal, which can be used as an unsupervised signal quality metric. When we have power spectral density (PSD) S, we can get IPR by dividing the power in the relevant heartbeat interval (40-250 bpm) with the whole power (0-F bpm).

$$\text{IPR} = \frac{\sum_{f=40}^{250} S(f)}{\sum_{f=0}^{F} S(f)} \tag{5}$$

6 Saliency Maps

We use the gradient-based method to get saliency maps for our method and Gideon2021 [8]. We have a trained model G_θ with input video v. We calculate Pearson correlation $l(s, \hat{s})$ between the ground truth signal s and the predicted signal $\hat{s} = G_\theta(v)$, and get the gradient of Pearson correlation with respect to the input video v. The saliency map is obtained by $\nabla_v l(s, G_\theta(v))$. Since the saliency map has three channels and the shape is $128 \times 128 \times 3$, we only show the green channel since green channel has the largest rPPG information [10].

References

[1] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, p. 277, BMVA Press, 2019.

[2] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062, IEEE, 2014.

[3] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019.

[4] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, and G. Zhao, "The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 242–249, IEEE, 2018.

[5] E. Magdalena Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1272–1281, 2018.

[6] L. F. C. Martinez, G. Paez, and M. Strojnik, "Optimal wavelength selection for noncontact reflection photoplethysmography," in *22nd Congress of the International Commission for Optics: Light for the Development of the World*, vol. 8011, p. 801191, International Society for Optics and Photonics, 2011.

[7] V. Vizbara, "Comparison of green, blue and infrared light in wrist and forehead photoplethysmography," *BIOMEDICAL ENGINEERING 2016*, vol. 17, no. 1, 2013.

[8] J. Gideon and S. Stent, "The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3995–4004, 2021.

[9] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3438–3446, 2016.

[10] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.