Multi-Task Learning for Simultaneous Video Generation and Remote Photoplethysmography Estimation

Yun-Yun Tsou, Yi-An Lee, and Chiou-Ting Hsu

National Tsing Hua University, Hsinchu, Taiwan
tsou0320@gmail.com, s107062576@m107.nthu.edu.tw, cthsu@cs.nthu.edu.tw

1 Network architecture

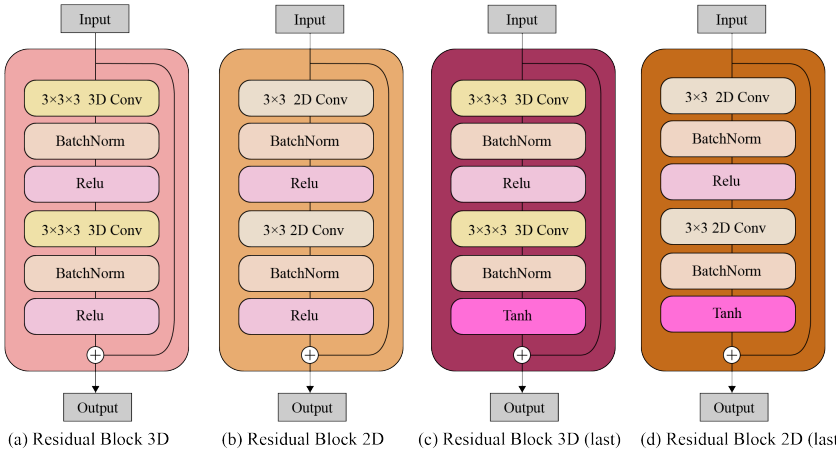

Fig. 1. The residual blocks used in our networks: (a) the ordinary 3D residual block; (b) the ordinary 2D residual block; (c) the 3D residual block used in the last layer; and (d) the 2D residual block used in the last layer.

We follow the residual architecture [1] to develop the encoder E, the rPPG removal network F, and the reconstruction network G in terms of residual blocks. Figure 1 shows different residual blocks adopted in our networks, where E is composed of four consecutive 2D residual blocks, F has four 3D residual blocks, and G has three 3D residual blocks. Detailed network architectures of E, F, and G are given in Tables 2, 3, and 4, respectively.

Table 1. rPPG network architecture.

Input Size	Layer Type	Filter Shape
256 x 80 x 80 x 3	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	AvgPool 3D	1 x 2 x 2
256 x 40 x 40 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 40 x 40 x 16	BatchNorm 3D	-
256 x 40 x 40 x 16	Relu	-
256 x 40 x 40 x 16	AvgPool 3D	1 x 2 x 2
256 x 20 x 20 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 32
256 x 20 x 20 x 32	BatchNorm 3D	-
256 x 20 x 20 x 32	Relu	-
256 x 20 x 20 x 32	AvgPool 3D	1 x 2 x 2
256 x 10 x 10 x 32	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 64
256 x 10 x 10 x 64	BatchNorm 3D	-
256 x 10 x 10 x 64	Relu	-
256 x 10 x 10 x 64	AvgPool 3D	1 x 2 x 2
256 x 5 x 5 x 64	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 128
256 x 5 x 5 x 128	BatchNorm 3D	-
256 x 5 x 5 x 128	Relu	-
256 x 5 x 5 x 128	AvgPool 3D	1 x 2 x 2
256 x 3 x 3 x 128	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 256
256 x 3 x 3 x 256	BatchNorm 3D	-
256 x 3 x 3 x 256	Relu	-
256 x 3 x 3 x 256	Global Average Pooling	-
256 x 1 x 1 x 256	Conv 3D (Stride 1/Padding 0)	1 x 1 x 1 x 1

Table 2. Network architecture of the encoder E in Image-to-Video network

Input Size	Layer Type	Filter Shape
80 x 80 x 3	Conv 2D (Stride 1/Padding 1)	3 x 3 x 16
80 x 80 x 16	BatchNorm 2D	-
80 x 80 x 16	Relu	-
80 x 80 x 16	Conv 2D (Stride 1/Padding 1)	3 x 3 x 16
80 x 80 x 16	BatchNorm 2D	-
80 x 80 x 16	Relu	-
80 x 80 x 16	Conv 2D (Stride 1/Padding 1)	3 x 3 x 16
80 x 80 x 16	BatchNorm 2D	-
80 x 80 x 16	Relu	-
80 x 80 x 16	Conv 2D (Stride 1/Padding 1)	3 x 3 x 16
80 x 80 x 16	BatchNorm 2D	-
80 x 80 x 16	Relu	-
80 x 80 x 16	Conv 2D (Stride 1/Padding 1)	3 x 3 x 16
80 x 80 x 16	BatchNorm 2D	-
80 x 80 x 16	Relu	-
80 x 80 x 16	Conv 2D (Stride 1/Padding 1)	3 x 3 x 16
80 x 80 x 16	BatchNorm 2D	-
80 x 80 x 16	Relu	-
80 x 80 x 16	Conv 2D (Stride 1/Padding 1)	3 x 3 x 16
80 x 80 x 16	BatchNorm 2D	-
80 x 80 x 16	Relu	-
80 x 80 x 16	Conv 2D (Stride 1/Padding 1)	3 x 3 x 3
80 x 80 x 3	BatchNorm 2D	-
80 x 80 x 3	Tanh	-

Table 3. Network architecture of the rPPG removal network F in Video-to-Video network.

Input Size	Layer Type	Filter Shape
256 x 80 x 80 x 3	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 3
256 x 80 x 80 x 3	BatchNorm 3D	-
256 x 80 x 80 x 3	Tanh	-

Table 4. Network architecture of the reconstruction network G.

Input Size	Layer Type	Filter Shape
256 x 80 x 80 x 3	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 16
256 x 80 x 80 x 16	BatchNorm 3D	-
256 x 80 x 80 x 16	Relu	-
256 x 80 x 80 x 16	Conv 3D (Stride 1/Padding 1)	3 x 3 x 3 x 3
256 x 80 x 80 x 3	BatchNorm 3D	-
256 x 80 x 80 x 3	Tanh	-

2 Training protocals

For the two datasets: COHFACE and PURE, we follow the same protocol and setting of [2] to conduct the experiments. That is, the training and testing split of COHFACE is determined by a protocol "all", which is introduced by Heusch et al. [3] in the bob.rppg.base Python package. As to the PURE dataset, we also follow the protocol defined in [2] to split the training and testing sets.

As to UBFC dataset, because there is no publicly available protocol for experimental comparison, we randomly split the training and testing dataset by keeping the same subject's videos either in the training or in the test sets, but not in both. That is, we ensure that a testing subject's videos have never been included in the training stage.

3 More Examples

More examples synthesized by Image-to-Video network and Video-to-Video network are given in Figures 2-7. All the synthesized videos well preserve the visual

appearance of the source data; and the estimated rPPG signals are accurately aligned with the corresponding source or target rPPG signals.

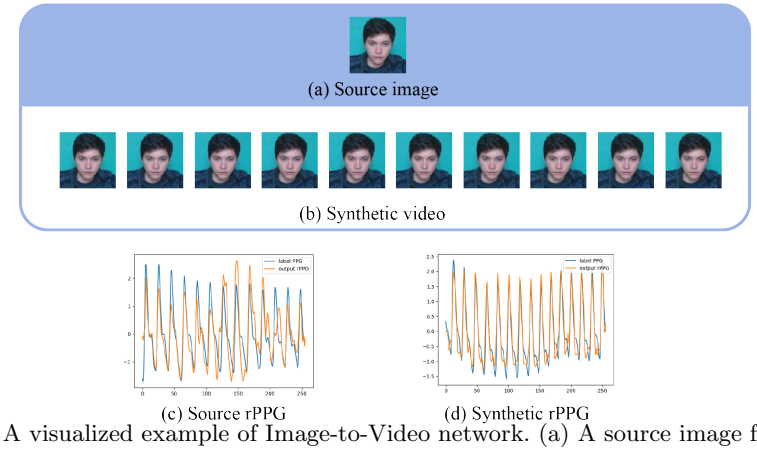

Fig. 2. A visualized example of Image-to-Video network. (a) A source image from the UBFC-RPPG dataset; (b) The synthesized videos; (c) The source PPG label (blue) and the predicted rPPG (orange); (d) The target rPPG (blue) and the predicted rPPG (orange)

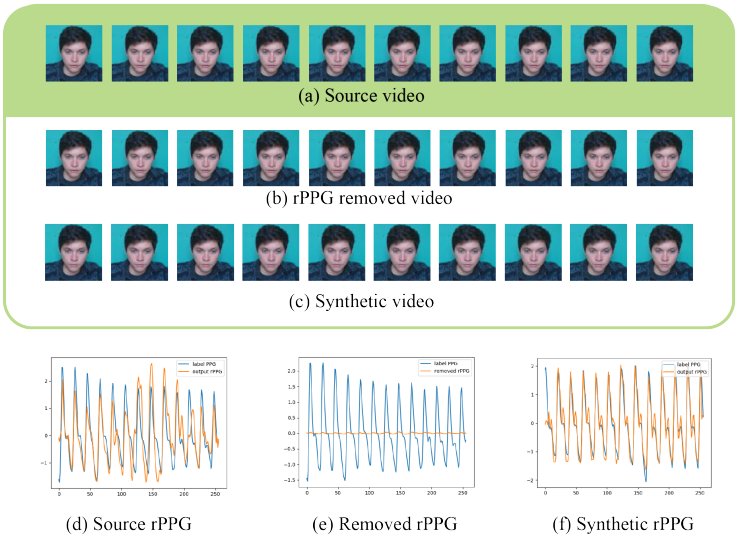

Fig. 3. A visualized example of Video-to-Video network. (a) Source video frames from the UBFC-RPPG dataset; (b) The synthesized videos; (c) The source PPG label (blue) and the predicted rPPG (orange); (d) The target rPPG (blue) and the predicted rPPG (orange)

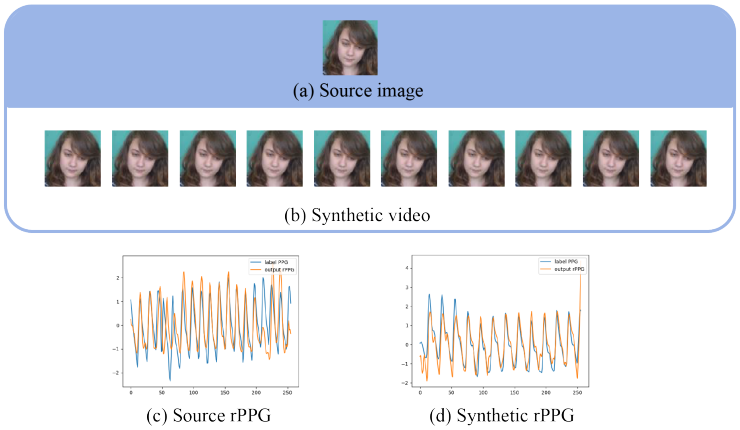

Fig. 4. A visualized example of Image-to-Video network. (a) A source image from the UBFC-RPPG dataset; (b) The synthesized videos; (c) The source PPG label (blue) and the predicted rPPG (orange); (d) The target rPPG (blue) and the predicted rPPG (orange)

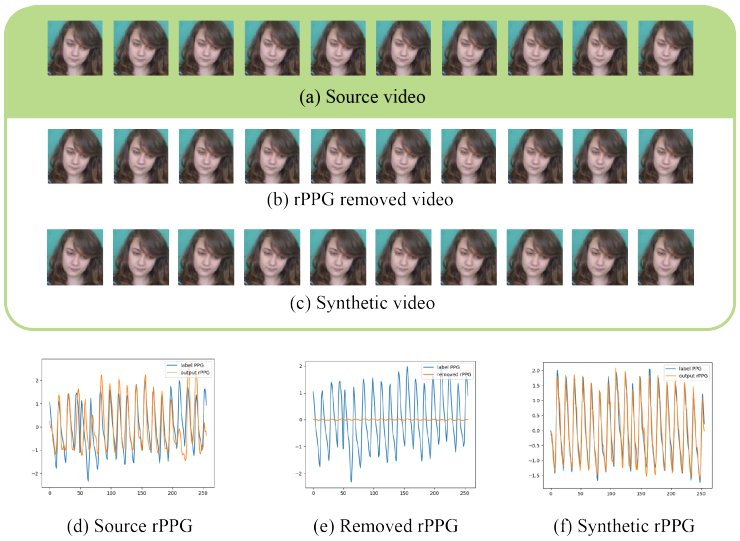

Fig. 5. A visualized example of Video-to-Video network. (a) Source video frames from the UBFC-RPPG dataset; (b) The synthesized videos; (c) The source PPG label (blue) and the predicted rPPG (orange); (d) The target rPPG (blue) and the predicted rPPG (orange)

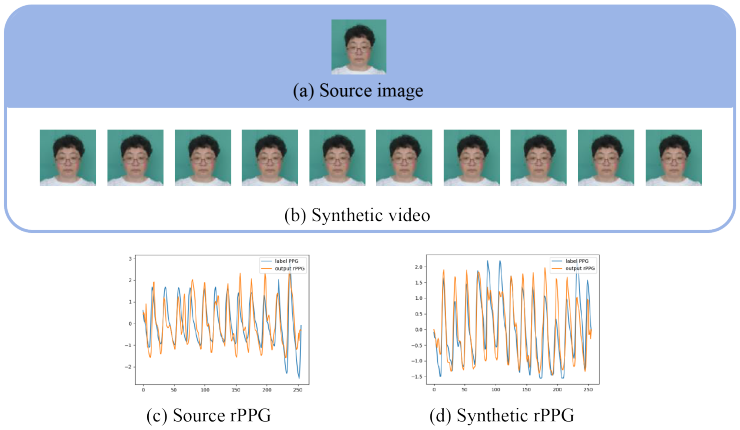

Fig. 6. A visualized example of Image-to-Video network. (a) A source image from the UBFC-RPPG dataset; (b) The synthesized videos; (c) The source PPG label (blue) and the predicted rPPG (orange); (d) The target rPPG (blue) and the predicted rPPG (orange)

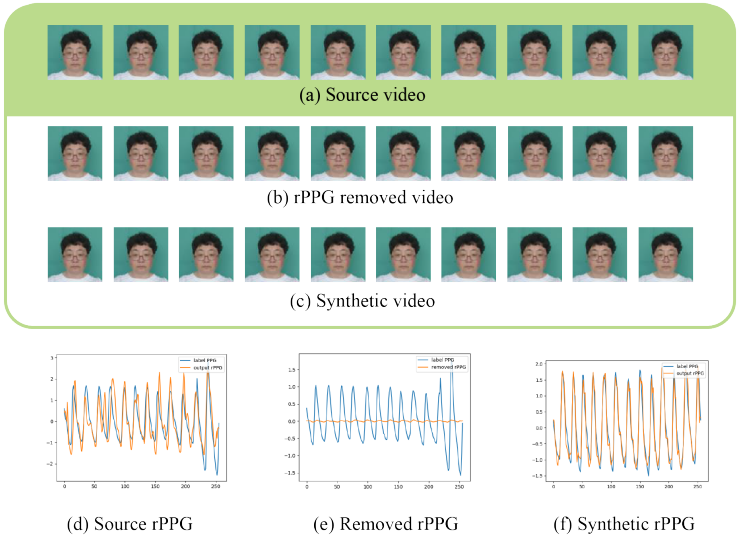

Fig. 7. A visualized example of Video-to-Video network. (a) Source video frames from the UBFC-RPPG dataset; (b) The synthesized videos; (c) The source PPG label (blue) and the predicted rPPG (orange); (d) The target rPPG (blue) and the predicted rPPG (orange)

References

1. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arxiv (2016)
2. Špetlík, R., Franc, V., Čech, J., Matas, J.: Visual Heart Rate Estimation with Convolutional Neural Network. In: Proceedings of British Machine Vision Conference. (2018)
3. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. CoRR **abs/1709.00962** (2017)