Deep Video Denoising for Facial Signal Processing

CSC2529 Computational Imaging 2022 - Final Project

Xiaoyan Li

Abstract—Video denoising is a more demanding task than still image denoising since it requires not only spatial denoising but also temporal coherency. On the other hand, video frames contain more information than a single image, which is helpful in the video denoising process. Thanks to the success of deep convolutional neural networks (CNNs), video denoising has recently achieved impressive performance using either supervised learning or unsupervised learning-based methods. In this project, we investigated two state-of-the-art algorithms, i.e., FastDVDnet and UDVD, and applied them to our real facial videos. The performance of the two algorithms was evaluated by PSNR and SSIM. Besides, the denoised videos are further used to extract image photoplethysmography (iPPG), a new noninvasive method to reveal the change of blood flow. We hypothesize that the iPPG extracted from the denoised facial video reflects the change of blood flow better than the iPPG extracted from the noisy facial video. This process can either work as the further application of facial video denoising or a metric to evaluate the results of facial video denoising when ground truth video is not available.

Index Terms—Deep video denoising, imaging photoplethysmography, explicit motion compensation.

1 Introduction

Benefited from the success of image denoising, video denoising has become an active research field in past years. Especially many deep CNN-based video denoising algorithms [1], [2], [3], [4] have achieved outstanding results and outperforming patch-based methods such as [1], [3], [4], [5], [6], [7], [8], [9]. Contrary to image denoising, video denoising is more demanding, requiring reasonable temporal coherency. But on the other hand, a video contains much more information than an image, which could help in the denoising process [3]. Generally, video denoising can be categorized as spatial video denoising, temporal video denoising, and spatial-temporal video denoising [10]. In spatial video denoising methods, denoising is applied to each frame individually; in temporal video denoising methods, denoising is applied to successive frame pairs; and spatial-temporal video denoising methods (i.e., 3D denoising) use a combination of spatial and temporal denoising.

FastDVDnet is one of these advanced spatial-temporal video denoising methods, introduced in [3]. FastDVDnet exhibits several desirable properties, including a small memory footprint and the ability to handle a wide range of noise levels with a single network model. With these attractive properties, in this project, FastDVDnet was explored and evaluated on our facial video data. However, similar to other start-of-the-art video denoising methods, FastDVDnet is based on supervised learning. [4] introduced an unsupervised video denoising method, called UDVD, achieving quite a similar performance as FastDVDnet. Therefore, we also included UDVD as one main architecture for our experiments and compared it to FastDVDnet using our in-house facial video data.

- *X. Li is with the Department of Computer Science, University of Toronto, Canada.*
  E-mail: xiaoy.li@mail.utoronto.ca

Based on [11], facial video could be used to simulate finger PPG (finger photoplethysmography), which is called an imaging PPG (iPPG). Therefore, in this project, as an extended application of the denoised facial video, the iPPG is extracted from the denoised facial videos and compared to finger PPG. This can also work as an evaluation metric of video denoising when PSNR and SSIM can not be calculated due to the unavailability of ground truth videos.

2 Related Work

2.1 FastDVDnet

Introduced in [3], FastDVDnet improves DVDnet [2] by replacing single-scale denoising blocks with multi-scale denoising blocks. Figure 1 shows the architecture of FastDVDnet. Its denoising process includes two steps. In the first denoising step, three denoising blocks share the same weights and take three consecutive frames as inputs. All denoising blocks in two steps have the same architecture (a modified U-Net), which is shown in Figure 2.

Instead of explicitly including motion estimation and/or compensation stage, FastDVDnet implicitly handles the motion by multi-scale modified U-Net [12] blocks which have been shown to be able to learn misalignment [13], [14]. Based on [3], FastDVDnet is trained end-to-end without optical flow alignment, which avoids distortions and artifacts due to erroneous flow.

Similar to [2], [15], a noise map is also included as a separate input in FastDVDnet to allow for the processing of spatially varying noise. The noise map provides information to the network about the distribution of the noise at the input, which is encoded as the expected per-pixel standard deviation of this noise.

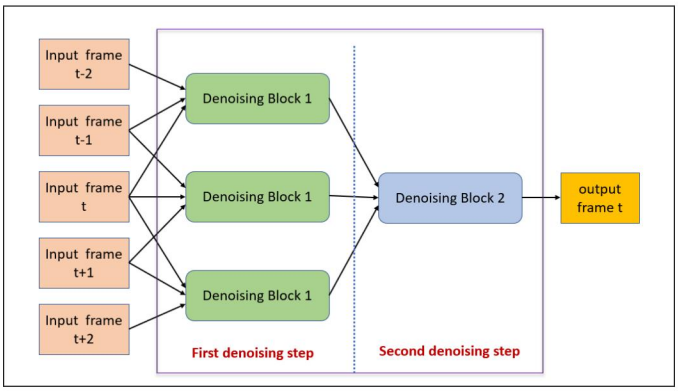

Fig. 1. Architecture of FastDVDnet. (We drew this figure based on [3].)

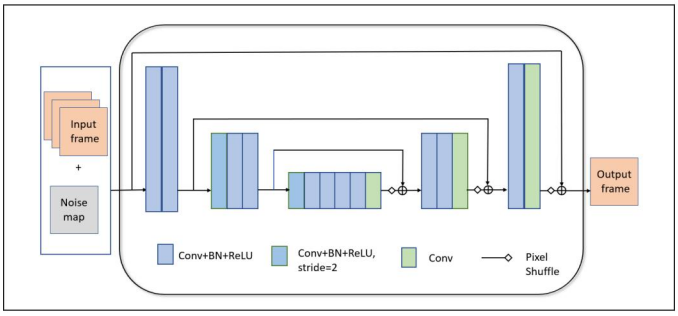

Fig. 2. The denoising blocks of FastDVDnet are composed of a modified multi-scale U-Net. (We drew this figure based on [3].)

The training dataset of FastDVDnet consists of noisy input and clean output pairs,

$$I_t^j = \{((S_t^j, \boldsymbol{M}_j), \boldsymbol{P}_t^j)\}_{j=0}^{m_t}, \tag{1}$$

where m_t is the number of training samples. $S_t^j = (\tilde{\boldsymbol{P}}_{t-2}^j, \tilde{\boldsymbol{P}}_{t-1}^j, \tilde{\boldsymbol{P}}_t^j, \tilde{\boldsymbol{P}}_{t+1}^j, \tilde{\boldsymbol{P}}_{t+2}^j)$ is five spatial patches cropped at the same location in successive frames, and $\boldsymbol{P}^j$ is the clean central patch of the sequence. $\boldsymbol{M}^j$ denotes the noise map with all its elements equal to the standard deviation of the noise, σ. The loss function of FastDVDnet is defined as

$$\mathcal{L}(\theta) = \frac{1}{2m_t} \sum_{j=1}^{m_t} \| \hat{\boldsymbol{P}}_t^j - \boldsymbol{P}_t^j \|^2, \tag{2}$$

where $\hat{\boldsymbol{P}}_t^j = \mathcal{F}((S_t^j, \boldsymbol{M}^j); \theta)$ is the output of the network, and θ is the set of all learnable parameters.

2.2 UDVD

Inspired by the "blind-spot" technique for unsupervised still image denoising, in which a CNN is trained to estimate each noisy pixel from the surrounding spatial neighborhood without taking into account the pixel itself, [4] introduced unsupervised video denoising (UDVD) method. The architecture of UDVD is shown in Figure 3. Input frames in UDVD are rotated by four different angles: $0°$, $90°$, $180°$ and $270°$ and the rotated frames are then processed in four separate branches with shared parameters, each containing asymmetric convolutional filters that are vertically causal. As a result, the branches produce outputs that only depend on the pixel above ($0°$ rotation), to the left ($90°$ rotation), below ($180°$ rotation), and to the right ($270°$ rotation). To produce the final output frame, the outputs from four different branches are then derotated and linearly combined through a 1×1 convolutions followed by a ReLU nonlinearity. Each branch of UDVD includes two denoising stages. The first stage in each branch consists of three U-Net blocks with shared parameters, and the second stage from each branch uses another U-Net.

Similar to other deep CNN-based video denoising methods, such as [1], [3], [16], UDVD performs motion compensation without explicitly estimating optical flow. Instead, the optical flow is automatically estimated from the network gradients, even though the network architectures are not designed to account for this.

Given the noisy neighbourhood Ω_y, the distribution of the three color channels of a pixel $x \in \mathcal{R}^3$ is modeled as $p(x|\Omega_y) = \mathcal{N}(\mu_x, \Sigma_x)$, where $\mu_x \in \mathcal{R}^3$ and $\Sigma_x \in \mathcal{R}^3$ denote the mean vector and covariance matrix respectively. Let the observed noisy pixel be defined as $y = x + \eta$, $\eta \sim \mathcal{N}(0, \sigma^2 I_3)$. Based on [4], by computing the mean of the posterior $p(x|y, \Omega_y)$, the information in the noisy pixel is fused with the output of UDVD, given by $E[x|y] = (\Sigma_x^{-1} + \sigma^{-2} I)^{-1} (\Sigma_x^{-1} \mu_x + \sigma^{-2} y)$. The mean and covariance of this distribution at each pixel are estimated by maximizing the log likelihood of the noisy data:

$$\mathcal{L}(\mu_x, \Sigma_x) = \frac{1}{2}[(y - \mu_x)^T (\Sigma_x + \sigma^2 I)^{-1} (y - \mu_x)] + \frac{1}{2} \log |\Sigma_x + \sigma^2 I|,$$

When the noise process is unknown, the objective becomes minimizing the MSE between the denoised output and noisy video, ignoring the center pixel.

2.3 Imaging Photoplethysmography

In the past decade, many studies have been done in the field of blood pressure (BP) estimation using Photoplethysmography (PPG) signal recorded by attaching a device (such as a light transmitter and a light receiver) to the body of a participant [17]. However, since this method resorts to professional measurement devices, it is considered not convenient and difficult to record PPG signals in long-term monitoring and estimate BP values from them. Recently, a new non-contact and camera-based method has been introduced to record the PPG signals [17], [18]. The PPG signals are extracted from images of a camera, which is called the iPPG signal. More specifically, the iPPG signal is extracted from facial video frames and based on changes in facial skin color. Hence, it could reveal blood volume changes. According to [19], the forehead is a good area for extracting the iPPG signal. The authors of [17] introduced an iPPG signal extraction algorithm which is illustrated in Figure 4. The average intensity of the extracted three (R, G, B) color channel pixels from each video frame is first calculated, and then a detrended algorithm is used to remove the low frequencies from these averaged intensity signals. After that, Independent Component Analysis (ICA) is applied to three signals to extract the three independent components. Finally, the Fast Fourier Transform (FFT) of the components whose first harmonic frequency is in the range of 0.5 to 5 is considered as the iPPG signal [20].

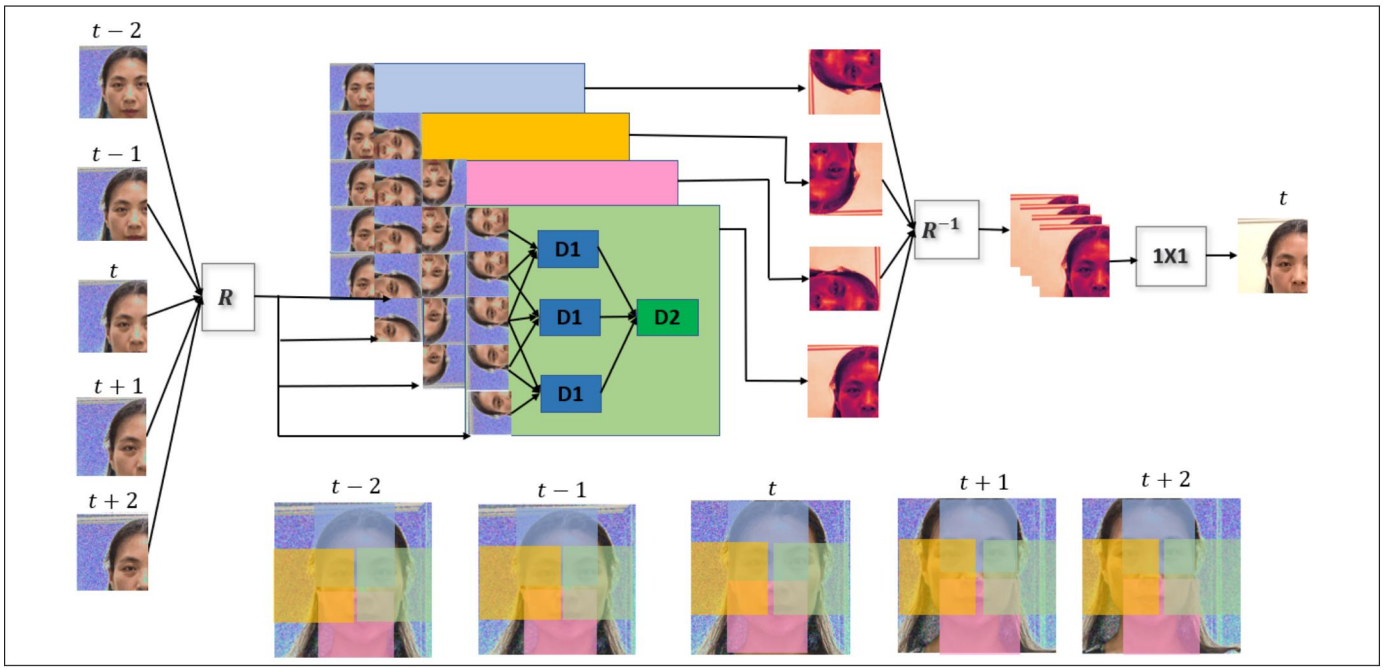

Fig. 3. Unsupervised Deep Video Denoising (UDVD) Network Architecture. (We drew this figure based on the idea in [4].)

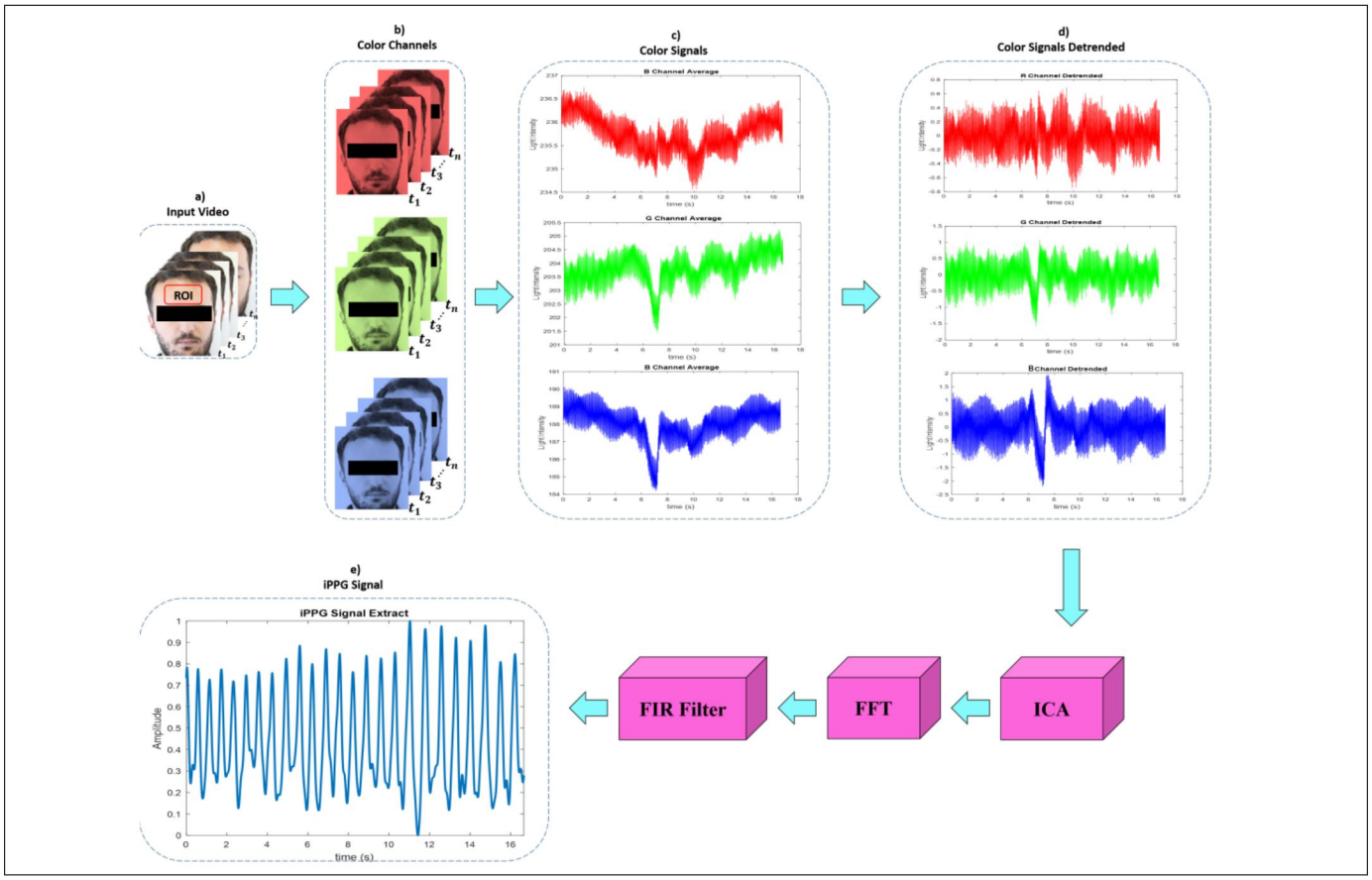

Fig. 4. Steps of extracting the iPPG signal. Source: [17].

3 Proposed Method

In this project, we propose to apply two state-of-the-art video denoising algorithms, i.e., supervised learning-based FastDVDnet and unsupervised learning-based UDVD, to our facial videos, so that the denoised facial video could be used to extract a non-contact iPPG signal that better matches the contact finger PPG signal. We use three facial videos recorded from three individuals by using an iPhone 6s, and contact finger PPGs corresponding to each participant

were simultaneously collected by using a finger PPG record system (Biopac MP160 monitoring system).

To save computational time, we use the pre-trained FastDVDnet and UDVD to denoise the facial videos. Both algorithms were pre-trained on the DAVIS [21] dataset. It is supposed that a noisy video is only corrupted with Gaussian noise, while according to [3] and [4], both FastDVDnet and UDVD can be used to denoise videos with other kinds of noise (e.g. Poissonian). UDVD was pre-trained on noisy images with standard deviation $\sigma = 30$. However, since we assume that the noisy level of facial video is not high, we fine-tune the pre-trained UDVD on the same DAVIS training data but with a smaller noise standard deviation $\sigma = 15$. By contrast, FastDVDnet allows a wide range of noise levels ($\sigma \in [5, 50]$), thus we directly use the pre-trained FastDVDnet without the fine-tuning process. We evaluated the performance of FastDVDnet with that of UDVD using two standard metrics: Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), and by comparing the extracted iPPGs from the noisy facial videos and the denoised facial videos.

To extract iPPG, OpenCV is used to detect faces from each frame, and a mediapipe module is used to locate the region of interest. A mask is applied to each facial frame to cover the eyes and mouth. For simplicity, the spatial average of all pixel values in the green color channel will serve as the iPPG. Following [6], we used Gaussian process regression (GPR) to further clean the iPPG signal. One output of GPR, i.e., predicted mean values, is then applied to a 6-th order Butterworth bandpass filter with cutoff frequencies 0.4 Hz and 3 Hz to produce the final iPPG signal. There are 60 frames per second, which means the facial video will produce a 60HZ iPPG signal. Figure 5 shows the process of iPPG extraction in our experiments.

4 Experimental Results

4.1 Results Analysis of Video Frame Denoising

Figures 6, 7, 8 and Table 1 show the denoising results of one frame from each noisy facial video. As aforementioned, we assume the videos are corrupted with Gaussian noise with standard deviation $\sigma = 15$. Based on these results, one can see that FastDVDnet and UDVD methods achieve quite similar results with FastDVDnet having slightly higher PSNRs on three noisy videos. However, considering FastDVDnet is a supervised learning method while UDVD is an unsupervised learning method, UDVD might be preferable when the ground truth video is not available. On the other hand, as reported in [3], FastDVDnet is robust to varying noise and a wider range of noise levels (i.e., $\sigma \in [5, 50]$), whereas UDVD is trained on specific noise levels. Therefore, in some cases, FastDVDnet might be more efficient and convenient.

TABLE 1
Comparison results on one frame of three individual noisy videos.

	FastDVDnet		UDVD	
	PSNR	SSIM	PSNR	SSIM
Video one	41.248	0.968	40.956	0.968
Video two	41.164	0.971	40.918	0.970
Video three	41.654	0.969	41.436	0.969

4.2 Results Analysis of iPPG Extraction

As mentioned in Section 3, to evaluate the extracted iPPG from facial videos, we compared it to contact finger PPG. Finger PPG is considered one typical means to measure the change in blood flow but is invasive and inconvenient [17]. We hypothesize that the iPPG extracted from the clean facial video should be closer to the finger iPPG. Figures 9, 10, 11 show iPPGs extracted from noisy, denoised, and ground truth facial videos of participant one, which are compared to the finger PPG. In this experiment, the noisy videos were first denoised by using the UDVD algorithm. All iPPGs in these figures were extracted from the green channel of the denoised video frames. By comparing Figure 9 with Figure 10, it seems that the iPPG extracted from the denoised facial video of participant one matches the finger PPG better than the iPPG extracted from the noisy facial video. Unfortunately, from Figure 11, one can be seen that even the iPPG extracted from the ground truth video cannot perfectly match the figure PPG. There are several possible reasons resulting in the discrepancy between them. First, it is possible that finger PPG doesn't exactly reveal the change of blood flow as there might be measurement noise. Second, the iPPG is based on average color intensity, however, some regions of the face (e.g., forehead) might reveal blood flow better than other regions, and averaging all pixel values of the facial image might reduce its performance.

Figures 12, 13, 14 show the comparison of iPPGs extracted from red channels of the video frames and the finger PPG of participant one. However, by comparing them with iPPG extracted from the green channel of video frames, red channel-based iPPGs have more discrepancies with finger PPG than green channel-based iPPGs. A similar conclusion can be made for blue channel-based iPPGs (see Figures 15, 16, 17). The reason for this phenomenon needs to be further investigated. One possible reason might be hemoglobin in the blood flow is more sensitive to green light than red or blue light.

Figures 18, 19, 20 and Figures 21, 22, 23 show the iPPGs extracted from green channels of other two participants. From these figures, we can see the iPPGs extracted from the facial videos including the ground truth video have bigger discrepancies with finger PPGs than that of participant one. The reasons need to be inspected in the future.

5 Discussion

In this project, we proposed to apply two deep CNN-based video denoising methods: FastDVDnet (supervised learning) and UDVD (unsupervised learning) to three noisy facial videos corrupted by Gaussian noise with standard deviation $\sigma = 15$. To compare the denoising performance of two algorithms, we used two typical metrics PSNR and SSIM. After the noisy facial video is denoised, we then used the denoised facial video to extract an iPPG signal. According to [17], the extracted iPPG signal should match invasive finger PPG. This process could be considered as the further application of facial video denoising, but also an evaluation

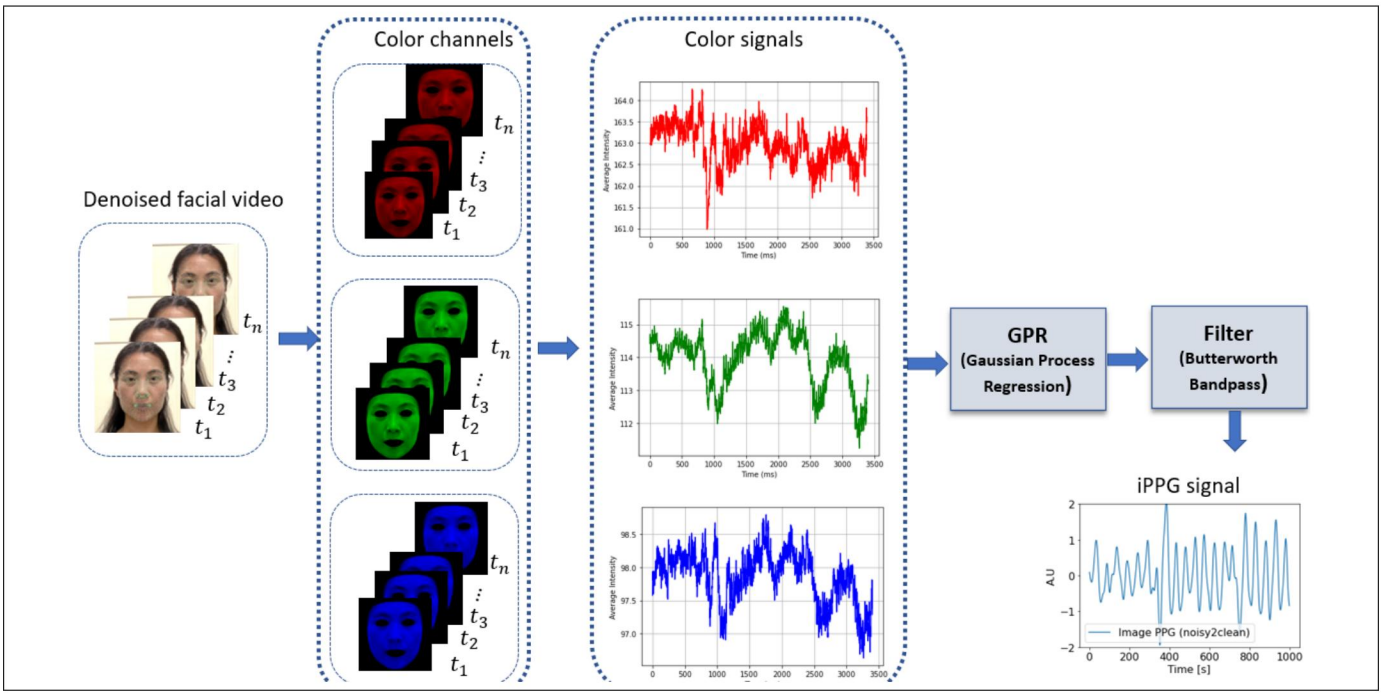

Fig. 5. Steps of extracting the iPPG signal from facial video images. Source: [17]

metric when PSNR and SSIM cannot be calculated due to the unavailability of the ground truth video.

Based on experimental results in Section 4, both FastDVDnet and UDVD achieved impressive results with PSNR values around 40 and SSIM above 0.96. As a supervised method, FastDVDnet performed a little bit better than UDVD on our three facial videos, whereas in some practical situations where ground truth is unavailable, the unsupervised method UDVD might be more preferable. Furthermore, as FastDVDnet is more robust to varying noise levels, it is allowed to directly use the pre-trained model for different video denoising tasks, while UDVD is trained on the specific noise level. When dealing with noisy video with different noise levels we might need to retrain or fine-tune the model.

From the aforementioned iPPG extraction results, we can observe that the denoised facial video produced a better iPPG signal than the noisy facial video. However, the iPPGs, including the iPPG extracted from the ground truth facial video, could not exactly match finger PPG. There are several possible reasons. First, finger PPG could not perfectly reveal the change in blood flow due to measurement noise. Second, for simplicity, we extracted iPPGs based on the average color intensity of the whole facial image, however, this method might not be as accurate as using the color intensity of specific regions of the face (e.g., forehead) which might mirror the change of blood flow better. Third, different camera exposure might result in lower performance. We might need to apply white balancing and color calibration approaches to the video frames, such that the color intensity of frames could be more accurate.

6 Conclusion

Two state-of-the-art video denoising methods, i.e., FastDVDnet and UDVD, were explored and evaluated on our in-house real facial videos in this project. Both methods are deep CNN-based (more specifically U-Net based), therefore they can automatically estimate optical flow without explicitly including a motion compensation module. According to [4], accurate optical-flow estimation is possibly realized from the network gradient. Our experimental results show that both FastDVDnet and UDVD have their own advantages. FastDVDnet achieved slightly better results than UDVD and is robust to varying noise levels but is a supervised learning method, while UDVD is a unsupervised learning method and is preferable when the ground truth video is not available. In addition to using PSNR and SSIM to evaluate the performance of two video denoising methods, we also used the denoised videos to extract iPPGs, which were then compared to finger PPG. We hypothesized that the iPPG extracted from the clean facial video matched the finger PPG better than the iPPG extracted from the noisy facial video. This hypothesis is supported by our experimental results. The IPPG extraction process can be viewed as a further application of facial video denoising or can be used as an evaluation metric when a ground truth video is unavailable.

Acknowledgments

I would like to express my greatest gratitude to Prof. David Lindell for his fascinating lectures, well-prepared course materials, and his instructive feedback on each student's questions. I also would like to thank my project mentor, Mian Wei for his useful insights and suggestions, and other course instructors for their well-prepared problem session and quick responses to our questions. I really benefit a lot from this course.

Fig. 6. Comparison of results of participant one.

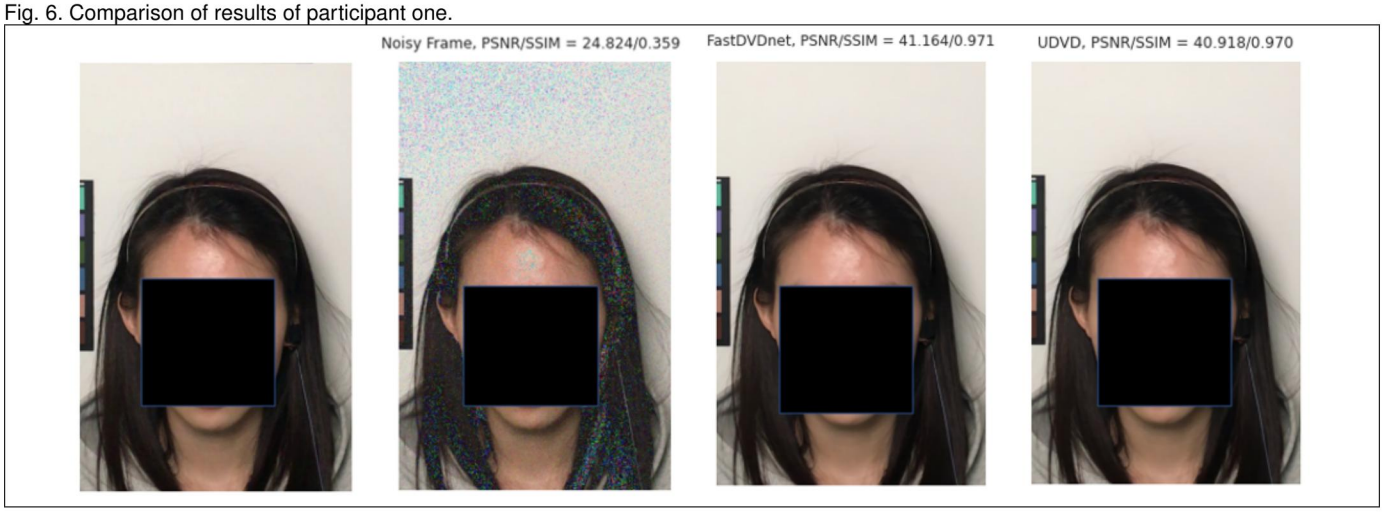

Fig. 7. Comparison of results of participant two.

Fig. 8. Comparison of results of participant three.

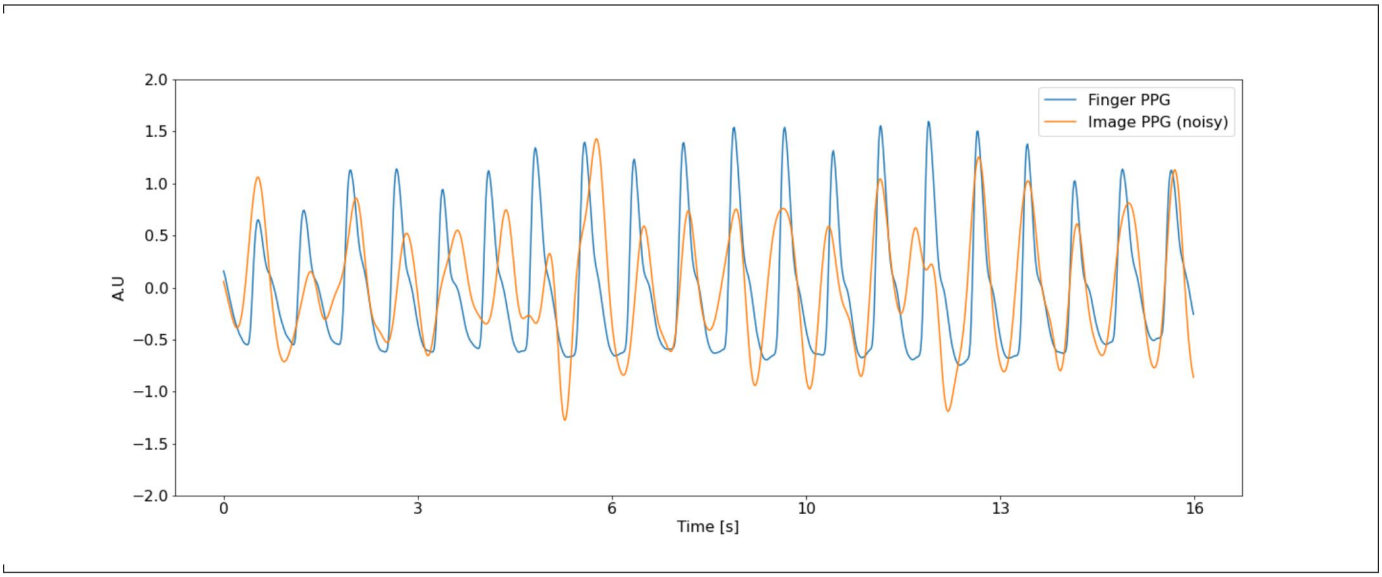

Fig. 9. Comparison of iPPG extracted from noisy facial video (green channel) of participant one and finger PPG.

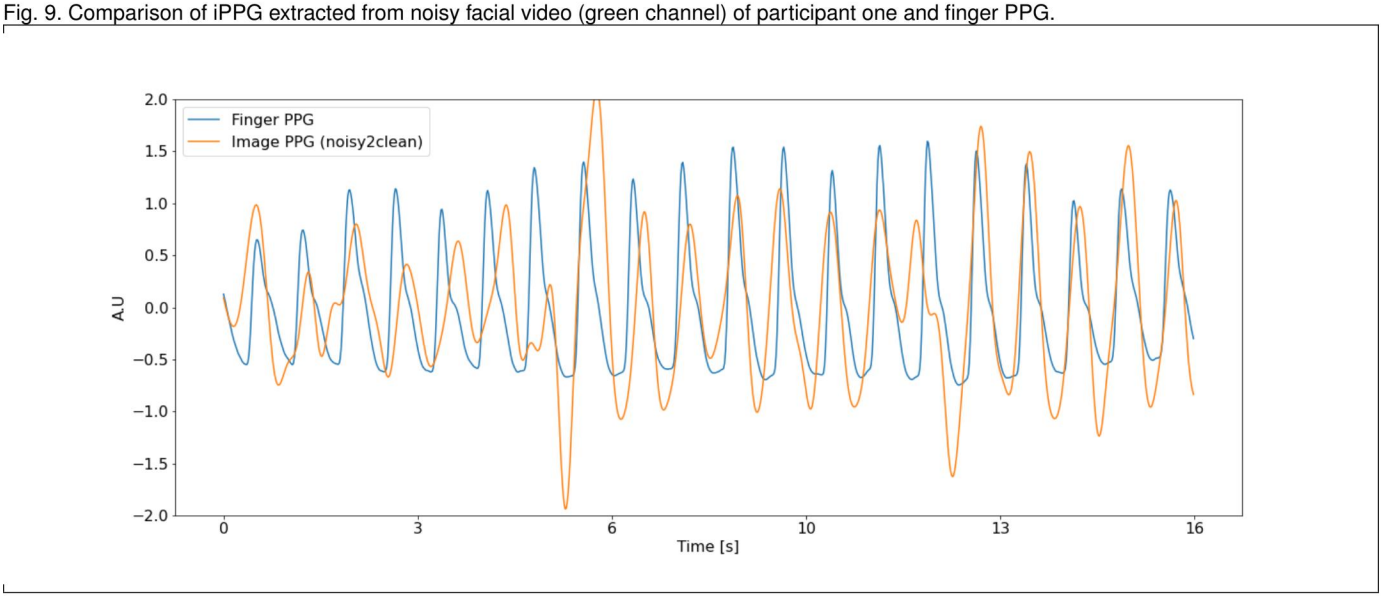

Fig. 10. Comparison of iPPG extracted from denoised facial video (green channel) of participant one and finger PPG.

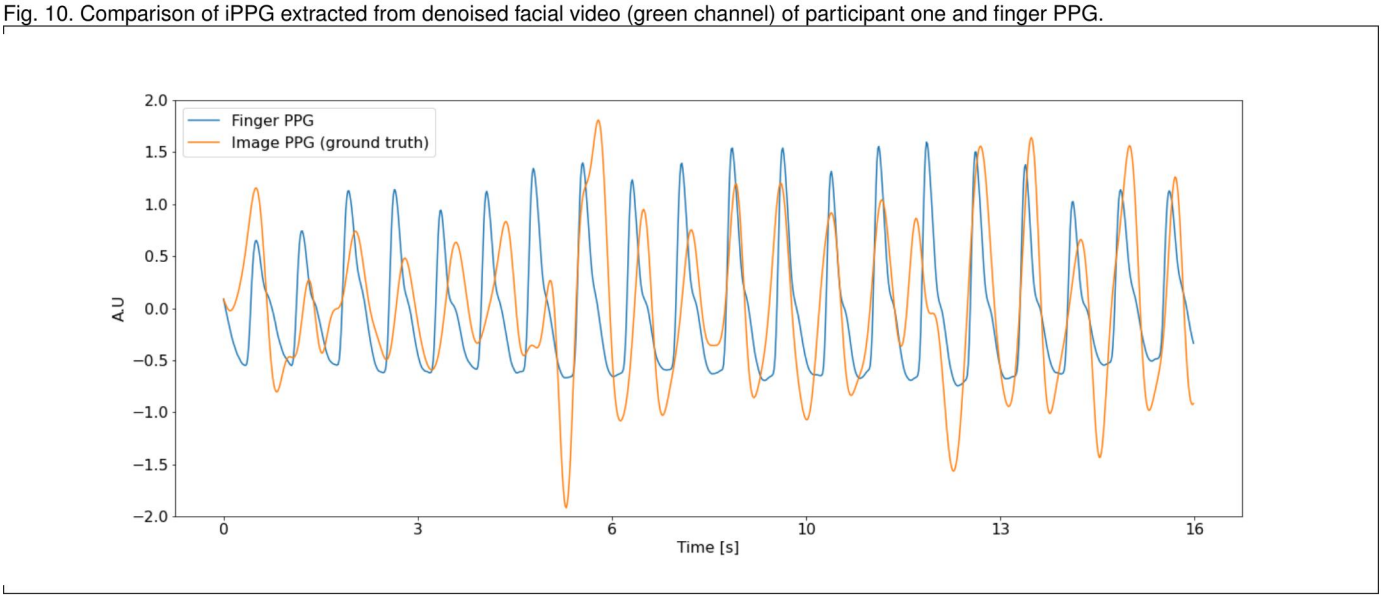

Fig. 11. Comparison of iPPG extracted from ground truth facial video (green channel) of participant one and finger PPG.

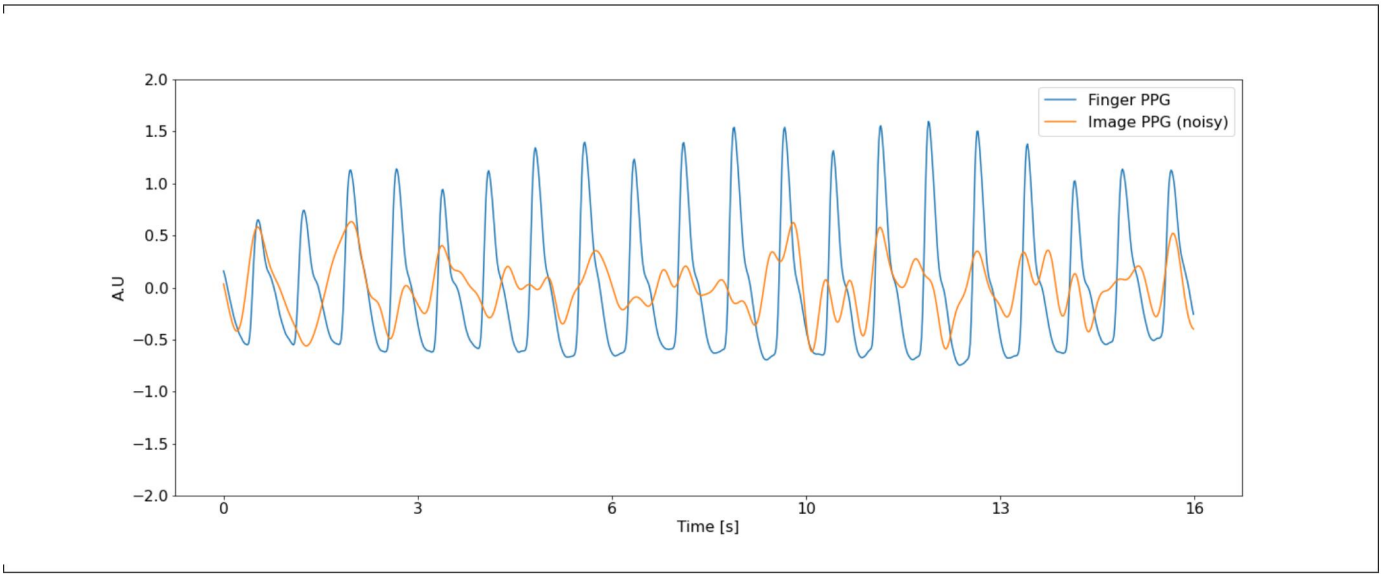

Fig. 12. Comparison of iPPG extracted from noisy facial video (red channel) of participant one and finger PPG.

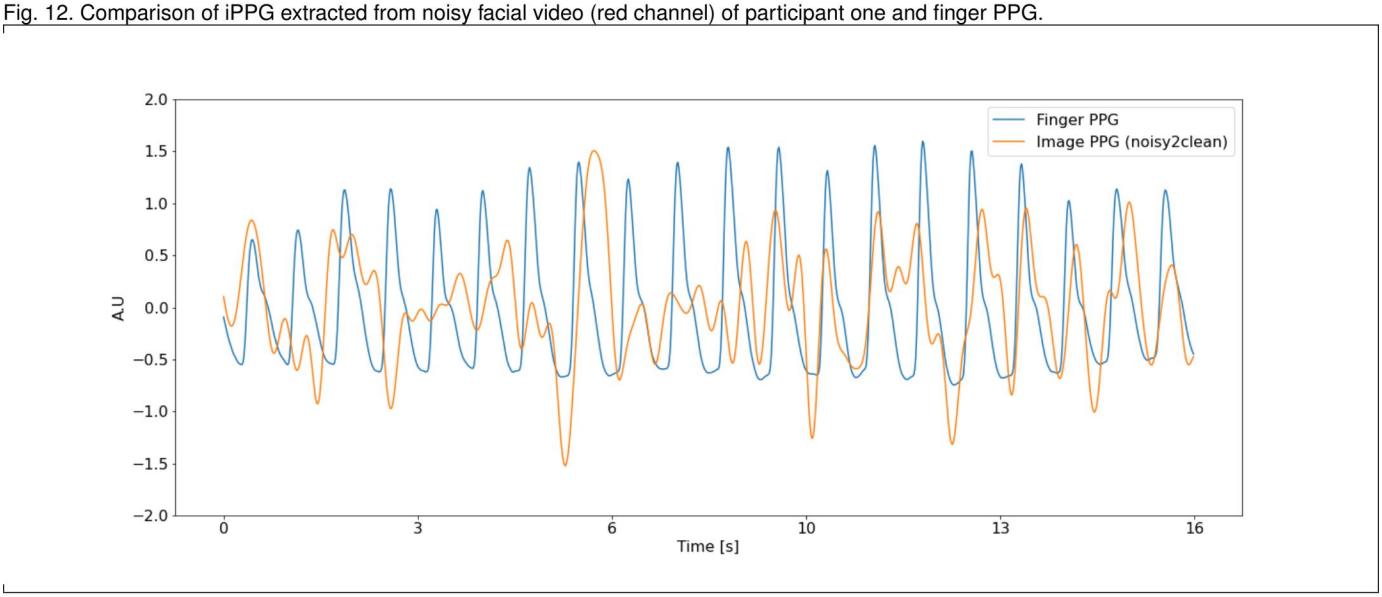

Fig. 13. Comparison of iPPG extracted from denoised facial video (red channel) of participant one and finger PPG.

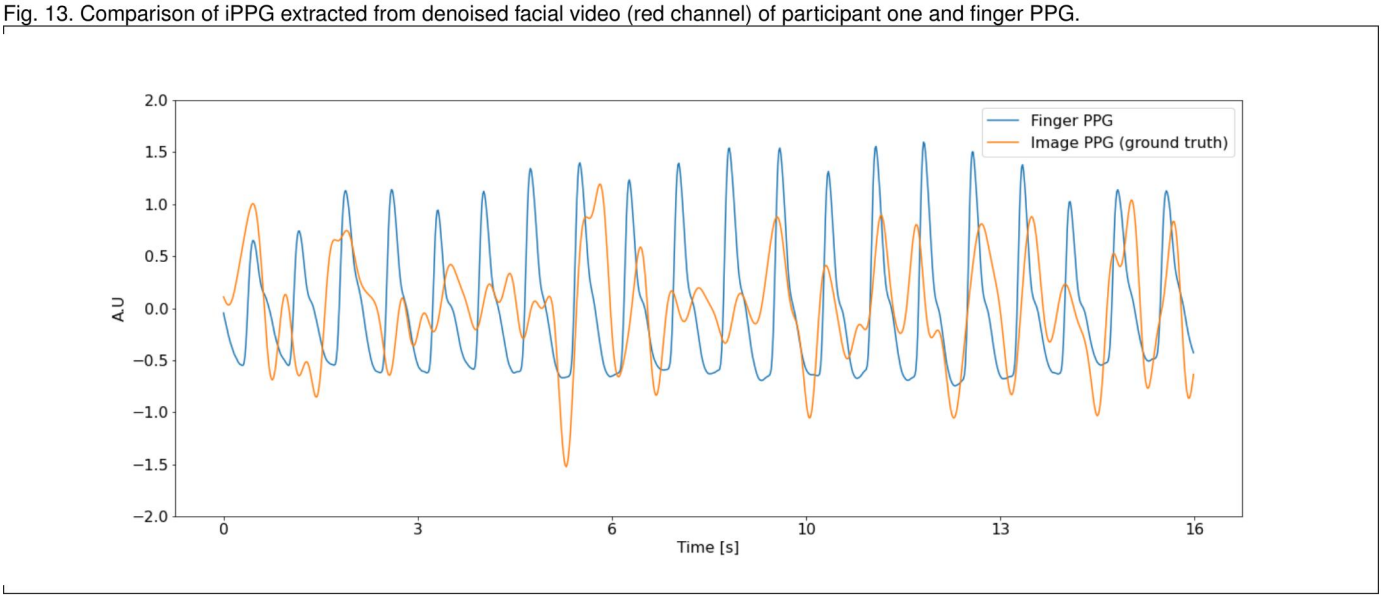

Fig. 14. Comparison of iPPG extracted from ground truth facial video (red channel) of participant one and finger PPG.

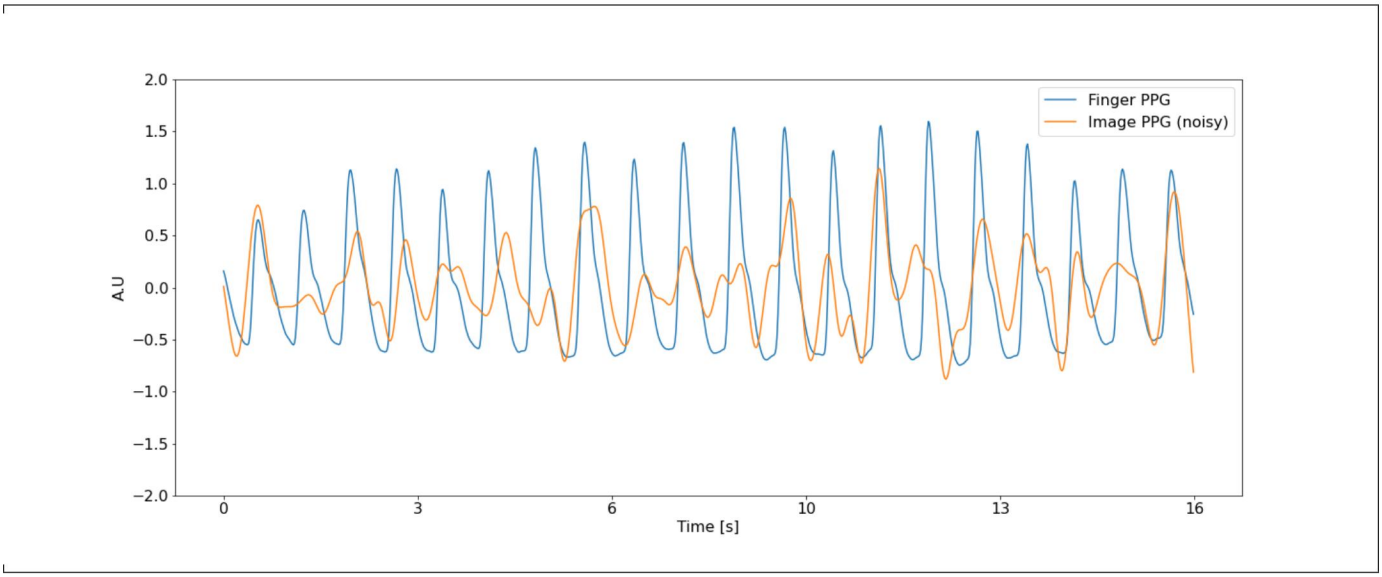

Fig. 15. Comparison of iPPG extracted from noisy facial video (blue channel) of participant one and finger PPG.

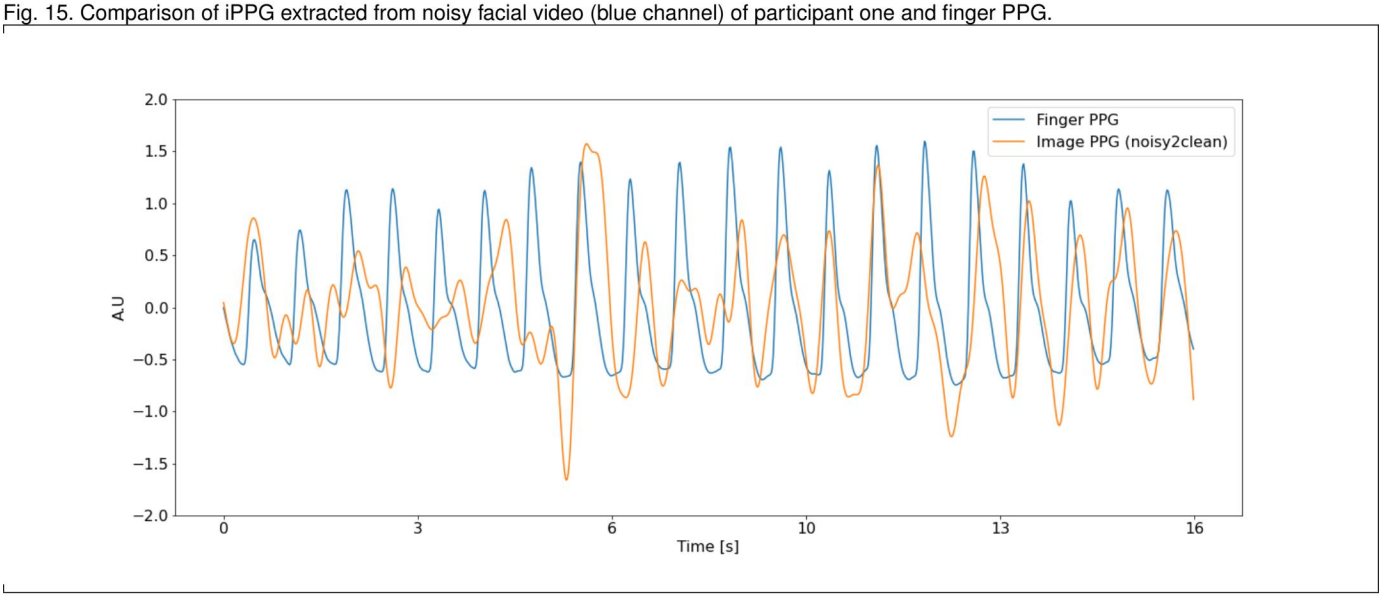

Fig. 16. Comparison of iPPG extracted from denoised facial video (blue channel) of participant one and finger PPG.

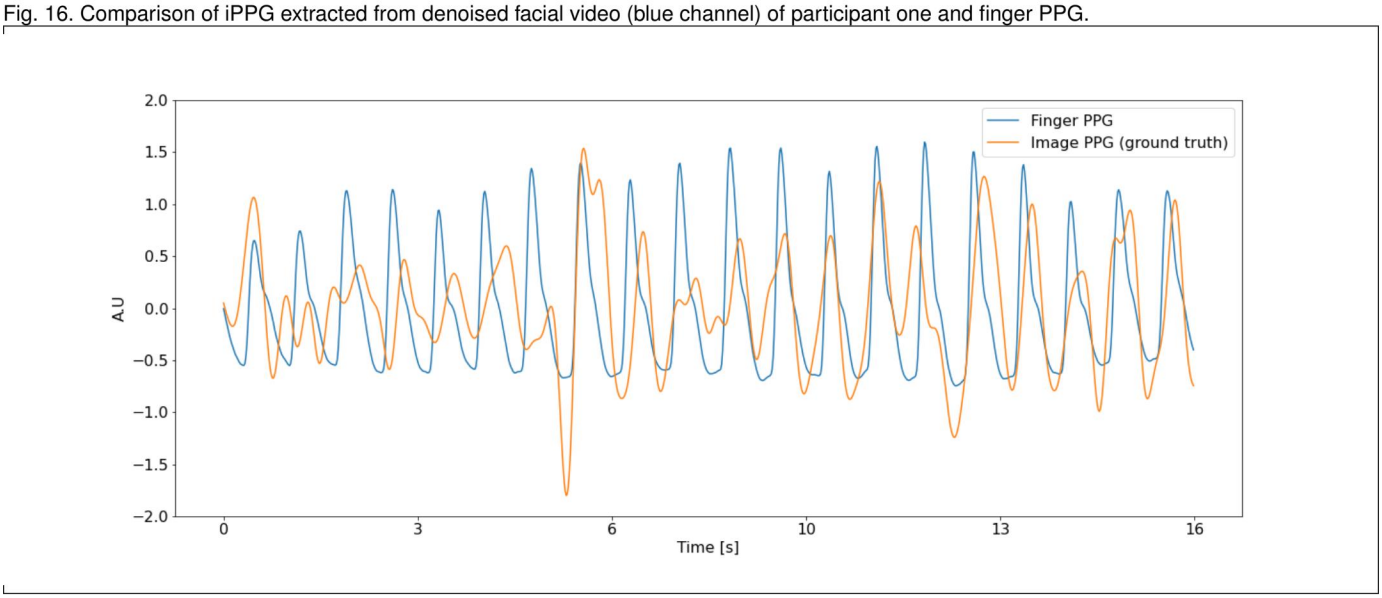

Fig. 17. Comparison of iPPG extracted from ground truth facial video (blue channel) of participant one and finger PPG.

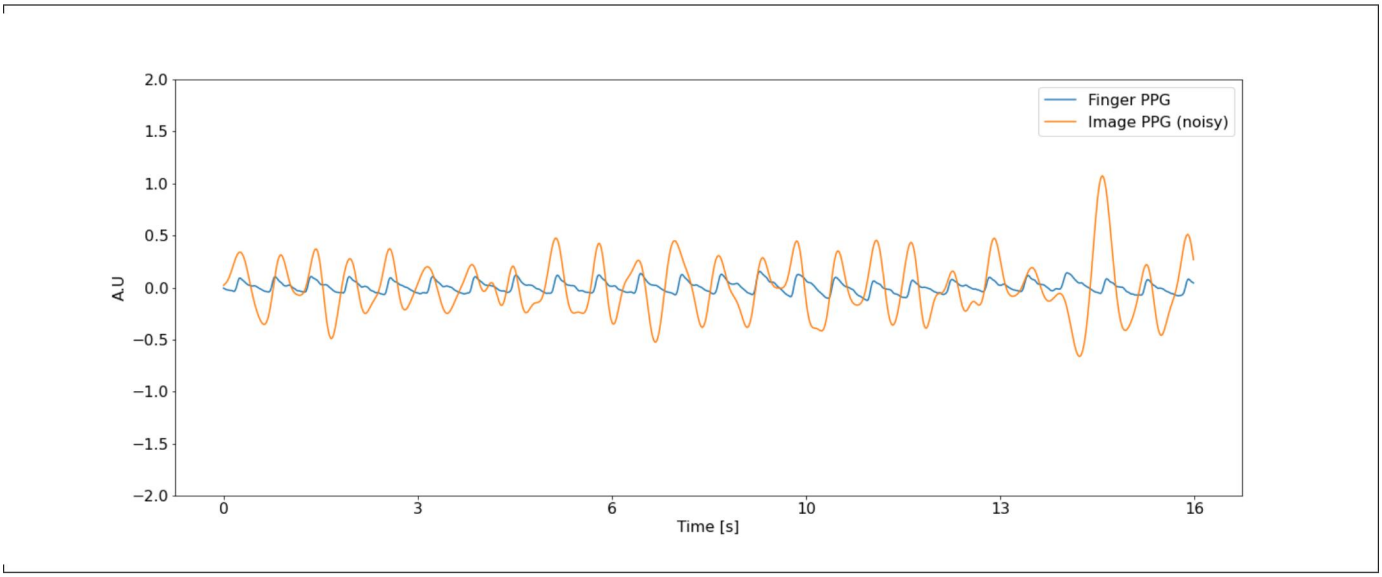

Fig. 18. Comparison of iPPG extracted from noisy facial video (green channel) of participant two and finger PPG.

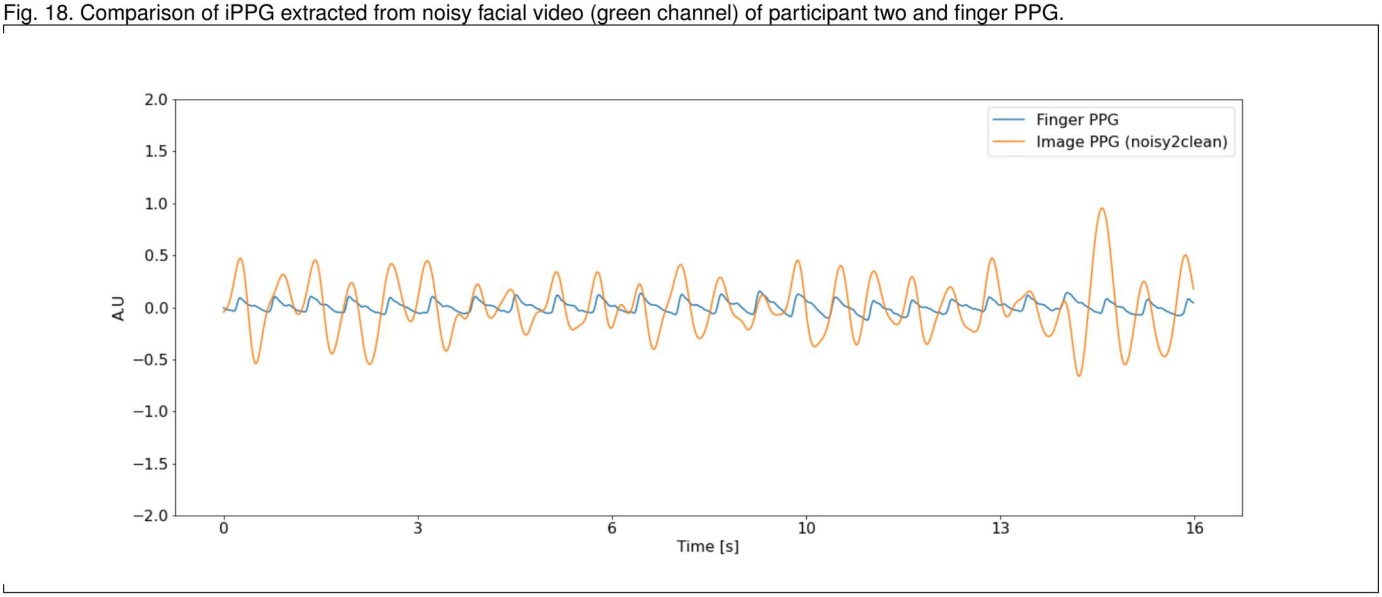

Fig. 19. Comparison of iPPG extracted from denoised facial video (green channel) of participant two and finger PPG.

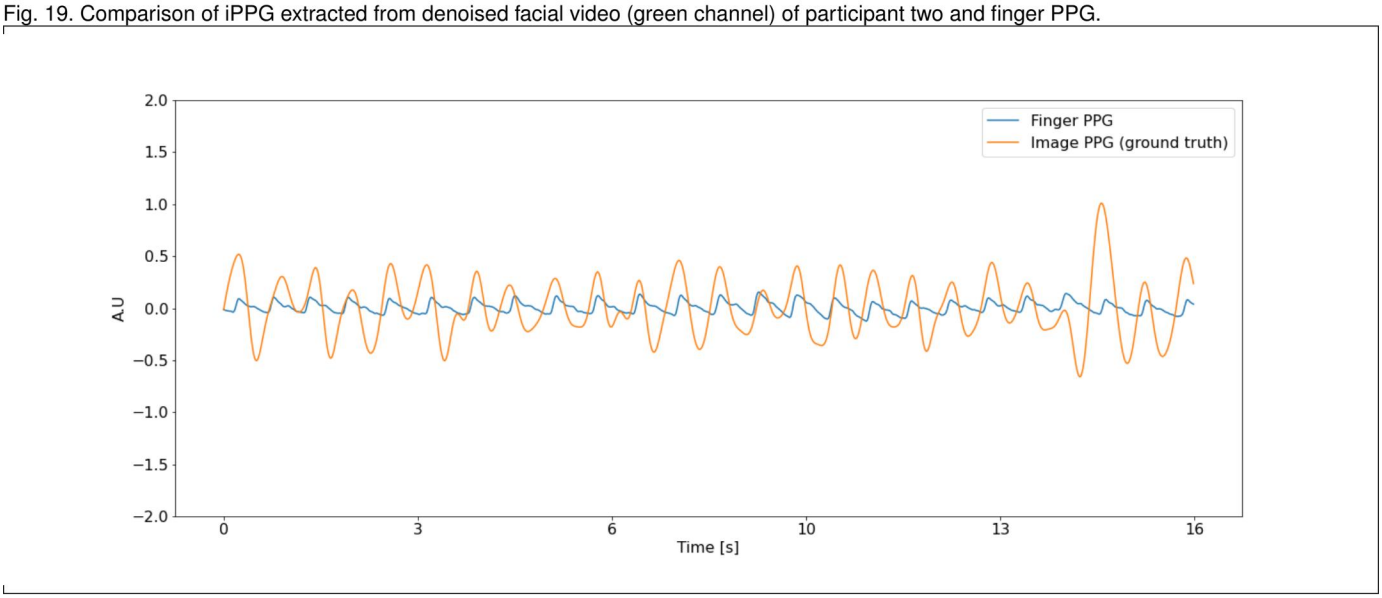

Fig. 20. Comparison of iPPG extracted from ground truth facial video (green channel) of participant two and finger PPG.

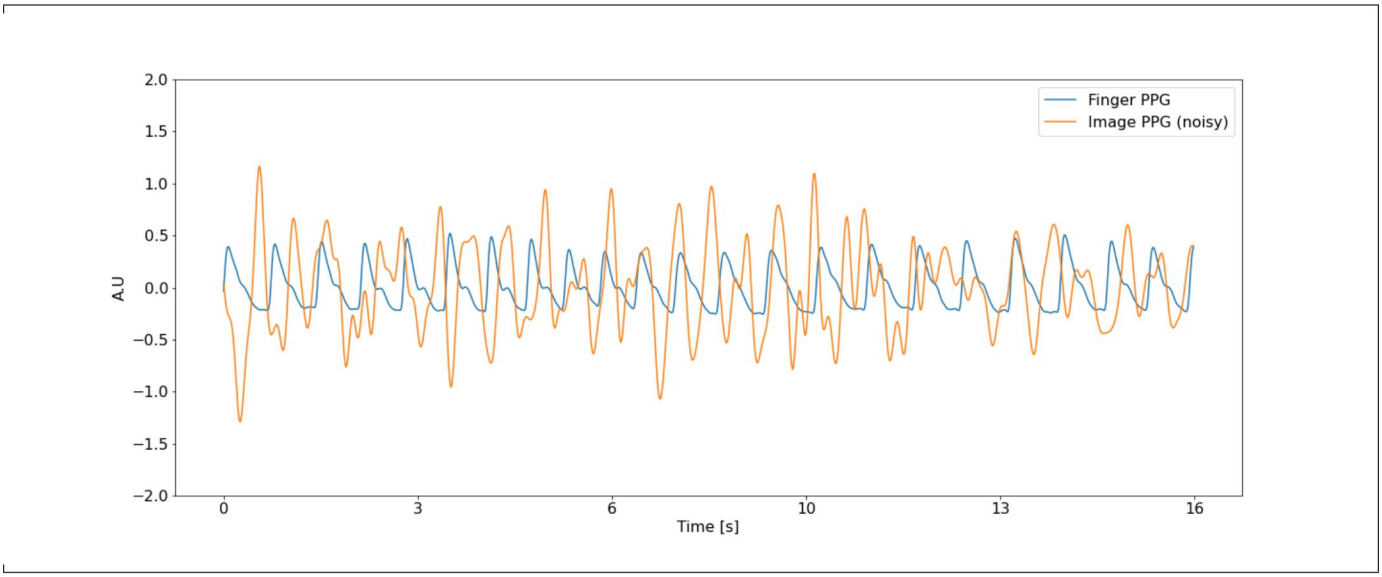

Fig. 21. Comparison of iPPG extracted from noisy facial video (green channel) of participant three and finger PPG.

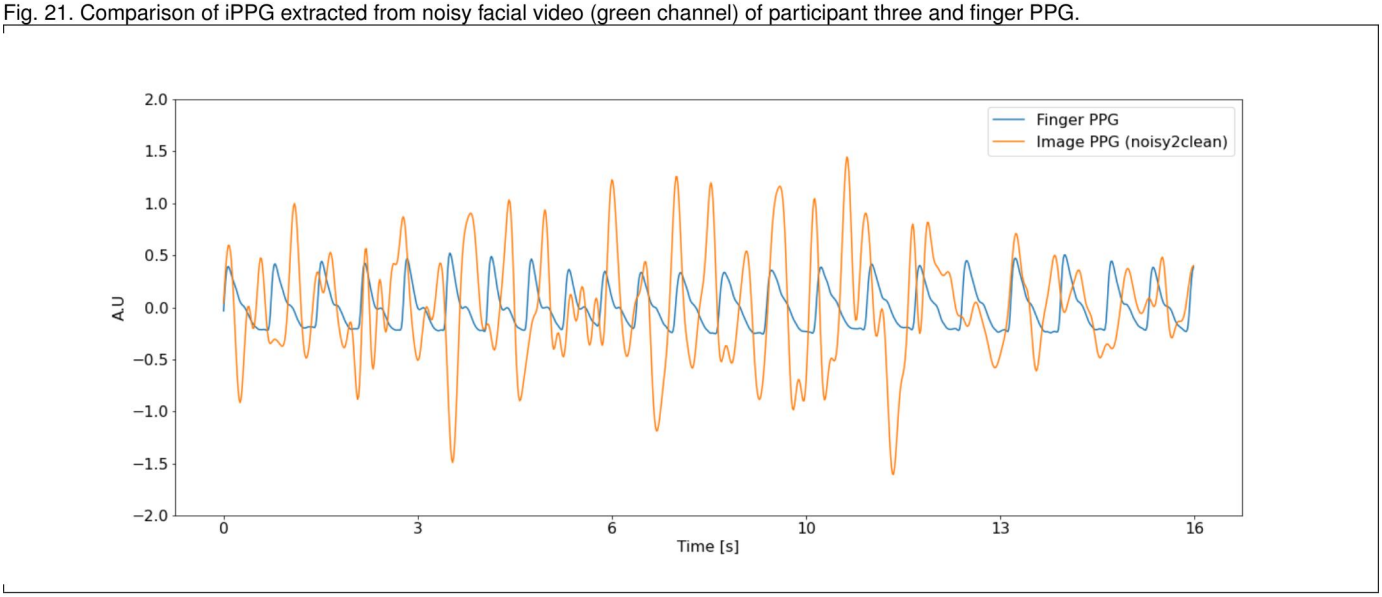

Fig. 22. Comparison of iPPG extracted from denoised facial video (green channel) participant three and finger PPG.

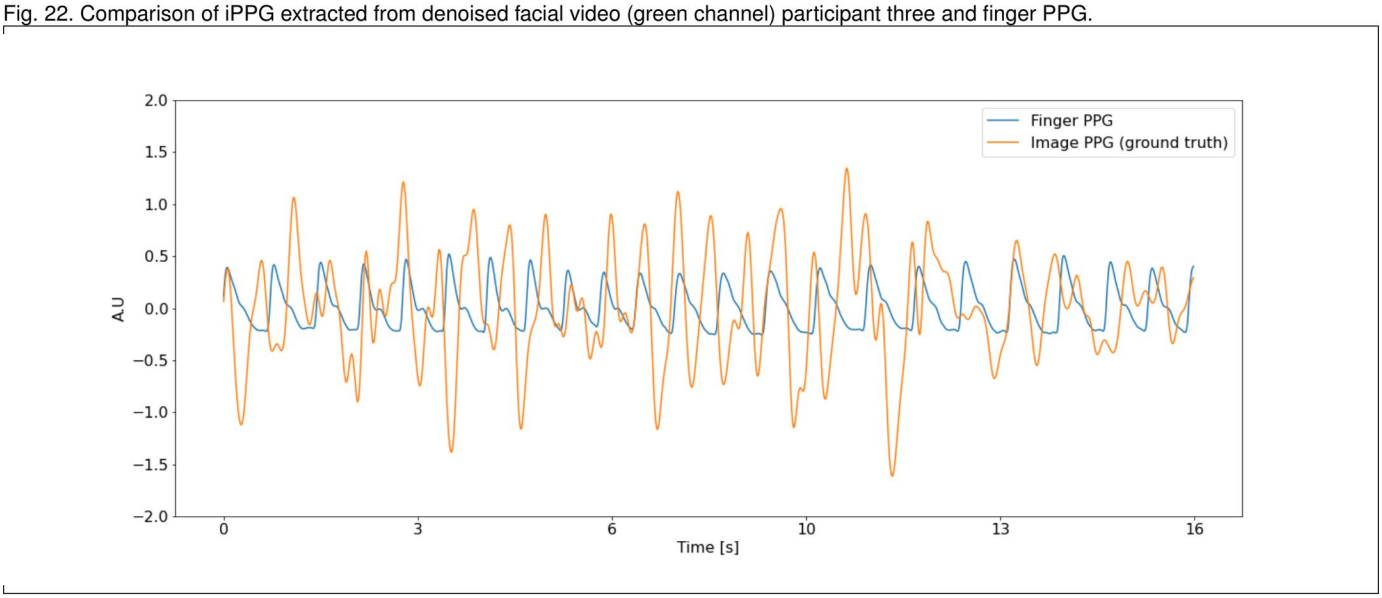

Fig. 23. Comparison of iPPG extracted from ground truth facial video (green channel) participant three and finger PPG.

REFERENCES

[1] M. Claus and J. van Gemert, "Videnn: Deep blind video denoising," in *Proceedings - 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2019*, 2019, pp. 1843–1852.

[2] M. Tassano, J. Delon, and T. Veit, "Dvdnet: A fast network for deep video denoising," 09 2019, pp. 1805–1809.

[3] ——, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1351–1360.

[4] D. Y. Sheth, S. Mohan, J. Vincent, R. Manzorro, P. A. Crozier, M. M. Khapra, E. P. Simoncelli, and C. Fernandez-Granda, "Unsupervised deep video denoising," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

[5] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *Trans. Img. Proc.*, vol. 26, no. 7, p. 3142–3155, jul 2017.

[6] J. Cao, Q. Wang, J. Liang, Y. Zhang, K. Zhang, and L. Van Gool, "Practical real video denoising with realistic degradation model," 2022.

[7] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. Van Gool, "Recurrent video restoration transformer with guided deformable attention," *arXiv preprint arXiv:2206.02146*, 2022.

[8] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *arXiv preprint arXiv:2201.12288*, 2022.

[9] C. Qi, J. Chen, X. Yang, and Q. Chen, "Real-time streaming video denoising with bidirectional buffers," in *ACM MM*, 2022.

[10] Video denoising, "Video denoising — Wikipedia, the free encyclopedia," 2022, [Online; accessed 15-November-2022]. [Online]. Available: https://en.wikipedia.org/wiki/Video_denoising

[11] Y. Sun, S. Hu, V. Azorin Peris, S. Greenwald, J. Chambers, and Y. Zhu, "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise," *Journal of biomedical optics*, vol. 16, p. 077010, 07 2011.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," vol. 9351, 10 2015, pp. 234–241.

[13] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," 04 2015.

[14] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *European Conference on Computer Vision*, 2018.

[15] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics*, vol. 35, pp. 1–12, 11 2016.

[16] A. Davy, T. Ehret, J.-M. Morel, P. Arias, and G. Facciolo, "A non-local cnn for video denoising," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2409–2413.

[17] R. H. Goudarzi, S. Somayyeh Mousavi, and M. Charmi, "Using imaging photoplethysmography (ippg) signal for blood pressure estimation," in *2020 International Conference on Machine Vision and Image Processing (MVIP)*, 2020, pp. 1–6.

[18] Y. Sun, S. Hu, V. Azorin Peris, R. Kalawsky, and S. Greenwald, "Noncontact imaging photoplethysmography to effectively access pulse rate variability," *Journal of biomedical optics*, vol. 18, p. 61205, 06 2013.

[19] S. Fallet, V. Moser, F. Braun, and J.-M. Vesin, "Imaging photoplethysmography: What are the best locations on the face to estimate heart rate?" in *2016 Computing in Cardiology Conference (CinC)*, 2016, pp. 341–344.

[20] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[21] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbelaez, A. Sorkine-Hornung, and L. V. Gool, "The 2017 DAVIS challenge on video object segmentation," *CoRR*, vol. abs/1704.00675, 2017. [Online]. Available: http://arxiv.org/abs/1704.00675

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.