De-identification of facial videos while preserving remote physiological utility

Marko Savic
marko.savic@oulu.fi
Guoying Zhao*
guoying.zhao@oulu.fi

Center for Machine Vision and Signal Analysis (CMVS)
University of Oulu
FI-90014, Finland

Abstract

Remote photoplethysmography has great potential in future telemedicine and affective computing, but contains sensitive biometric data, making privacy protection essential for future applications. Privacy related research concerning rPPG is severely limited and does not handle utility-preserving de-identification. In this paper we propose the first learning based method for facial video de-identification that preserves the rPPG signal and visual appearance, thus keeping the utility of the data for remote rPPG measure while protecting users' privacy. Our proposed semi-adversarial framework processes an input video by adding unobtrusive perturbations that remove biometric privacy while keeping the rPPG signal quality high. The framework is trained via learning-based constraints that leverage pre-trained biometric recognition networks and rPPG predictors. Furthermore, we propose a novel loss term that improves biometric de-identification by lowering downstream recognition confidence. We systematically evaluate our proposed method on two public datasets and with varied face identification and rPPG extraction methods, and provide a novel benchmark for future research in this direction. Our code is available at: https://github.com/marukosan93/De-id_rPPG.

1 Introduction

The development of information and communication technology (ICT) has led to new remote medical and interaction applications, especially during the COVID-19 pandemic, when a high demand for telemedicine and teleconferences was observed [35]. Heart rate (HR), heart rate variability (HRV), respiratory frequency (RF) and oxygen saturation (SpO2) are widely utilised as important healthcare parameters and psychological indicators since they change accordingly with our physical well-being and emotional states. Commonly, electrocardiography (ECG) or photoplethysmography (PPG) based contact devices are employed to measure physiological signals. However, they can be uncomfortable in long term monitoring scenarios, especially with neonates or patients with skin problems, and very inconvenient in natural interactions. Remote photoplethysmography (rPPG) [36] is a non-contact method that can lead to increased monitoring, and improve early detection rates for diseases such as Atrial fibrillation (AF), or help emotional interactions [30]. Similar to PPG optical sensors, common RGB cameras can capture subtle changes in skin colour that correspond to periodical variations of optical absorption of tissue caused by the cardiac cycle. Compared to

* indicates corresponding author

contact devices, in addition to the weak rPPG signal, cameras capture overwhelming environmental noise caused by lighting changes, subject movement and camera sensor variations, making accurate and robust rPPG a challenging task. An illustration of rPPG, Contact-PPG and a simplified extraction method are shown in Fig. 1. A major concern in future applica-

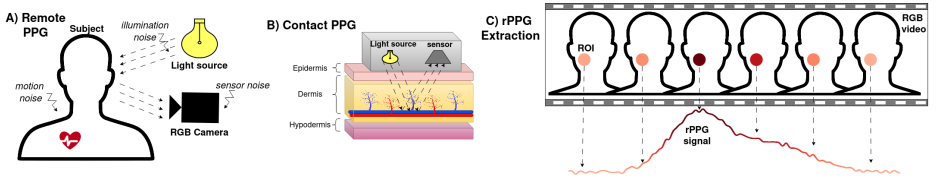

Figure 1: A) rPPG: Cardiac activity causes periodical variations in the reflected light intensity, noise is also captured (from e.g., lighting, motion and sensor). B) PPG: Uses simple optical sensor to capture strong signal, due to short distance and lesser noise. C) rPPG extraction: a coarse rPPG signal is obtained by selecting an ROI (e.g., cheeks or forehead) and averaging the pixels from each frame to extract a temporal signal.

tions of rPPG is data privacy [29], as it involves both personal physiological data and facial videos, that embed crucial pieces of private information such as identity, gender and race. Handling big data in the health care field becomes increasingly important [27], therefore data anonymisation is necessary to ensure fair treatment of subjects and compliance with regulations such as the General Data Protection Regulation [11] and the EU AI Act [10]. According to the GDPR, facial images are regarded as biometric data and as such should be subject to special restrictions. Recent advances in machine learning make it possible to easily identify a subject and automatically extract sensitive information that could lead to intentional or unintentional unethical practises. Therefore, digital de-identification that preserves the physiological features, while removing biometric identity, is fundamental. Despite the importance of protecting privacy in rPPG, the current research is severely limited. Several works deal with removing or modifying the underlying rPPG signal [2, 32], but to our knowledge, only one work [3] distinctly focuses on biometric de-identification for rPPG. However, it was evaluated with outdated methods and produces outputs that lose most of the visual information in the facial area, making it unsuitable for tasks such as emotion analysis and for comprehensive patient monitoring. In this work, we propose the first learning based facial video de-identification method that preserves the rPPG signal and visual fidelity. Our method significantly deteriorates the performance of identity recognition on the modified videos, making them unrecognizable to machines, therefore protecting privacy and avoiding potential misuse in big data sharing and processing. Moreover, the data's utility is preserved as the underlying rPPG signals retain their quality, and the visual appearance is kept, as the information contained in the facial area can be useful for e.g., emotion analysis or can provide visual cues to authorised human users. We evaluate our method on two public domain datasets and with different state-of-the-art biometric recognition networks and rPPG predictors, providing a benchmark for future research on this new topic.

Here we summarize our contributions. Firstly, we propose the first learning based method for rPPG preserving de-identification that maintains data utility. Secondly, we further improve de-identification performance with a new output probability regularisation term, which focuses on reducing the downstream recognition confidence. Thirdly, we systematically evaluate our method and provide a new benchmark for future research.

2 Related Work

Remote physiological signals. rPPG measurement was first established in [36], by extracting a signal from a single channel via averaged facial video pixels. Several traditional non-learning based methods were introduced that relied on optical/physiological considerations expressed through mathematical models (CHROM [8], POS [37], PBV [9], LGI [25]) or on blind source separation approaches (ICA [26] and PCA [17]). However, their assumptions and mathematical models did not hold in less constrained environments and were surpassed by deep learning methods. 2D-CNN models that extracted the HR from two adjacent frames were proposed in HR-CNN [42] and DeepPhys [5], that only take spatial information into account. End-to-end spatial-temporal 3D-CNN models such as PhysNet [39], rPPGNet [40] have been used to exploit the temporal information. Another way to exploit temporal context and suppress the information unrelated to HR signal is computing spatial-temporal maps as input, like in RhythmNet [23], CVD [24], Dual-GAN [20] and BVPNet [7]. Recently, transformer based networks such as EfficientPhys [19] and PhysFormer [41] have shown promising improvements by leveraging self-attention for better temporal modeling.

Image de-identification. In traditional de-identification methods, pixelation, masking and blurring were commonly applied [12]. These simple methods conceal the sensitive identity related information directly, but severely limit the utility of the images as they lead to loss of information. Recent studies have exploited the generative capabilities of deep learning to manipulate facial features for de-identification. L2M-GAN [38] manipulates the latent space of a GAN network to edit facial attributes. Semi-adversarial approaches like PrivacyNet [22], and CIAGAN [21] attempt to edit certain attributes while constraining the others. PrivacyNet aims to obfuscate gender, age and race while retaining biometric recognition performance. On the other hand, VGAN [7] and CIAGAN [21] retain facial expressions while altering identities. Depth information has also been used to improve de-identification while preserving facial expressions, gender and ethnicity [6]. Nonetheless, none of the aforementioned works considered de-identification that preserves underlying physiological signals, which are in much finer level and require spatial-temporal consistency in de-identification.

Privacy in rPPG. Privacy is a significant hurdle for remote physiological signal measurement, but the current research is limited, especially regarding de-identification. There are a few works about rPPG signal related to face video manipulation. PulseEdit [4] modifies the rPPG signal in facial videos by solving an optimisation problem to compute a perturbation to apply to the video data, but was tested on simple data and proved less effective than deep learning methods. PrivacyPhys [33] leverages a pre-trained 3D-CNN model to alter rPPG signals in videos, which proved to be more effective than the previous baseline. However, both methods superimpose a target rPPG signal on a video but are unable to de-identify a video while keeping the rPPG related information intact. In [3] the authors de-identify facial videos while preserving rPPG via a spatial-frequency decomposition and face structure standardization. Their method blurs the facial area, preserving the overall average pixel value. However, most of the information contained in the face is lost, making it unsuitable for other applications such as emotion analysis and limiting its clinical utility. Furthermore, it has only been evaluated with outdated methods, such as the traditional method POS [37] for rPPG, and simple machine learning methods for biometrics (Gaussian mixture models and support vector machines), therefore it could be ineffective with modern deep learning techniques in both reliable de-identification and accurate rPPG measurement. Despite the amount of recent works dealing with facial de-identification and rPPG measure, there are no effective methods focusing on preserving physiological signals and visual utility.

3 Methods

The overview of our method is shown in Fig. 2. The biometric recognition network and rPPG predictor are pre-trained and their weights are fixed during the de-identification autoencoder's training. The autoencoder is trained with the following objectives: perform a faithful reconstruction of the input video, de-identify the video so that the recogniser cannot identify the face correctly and have the same underlying rPPG signal as the original.

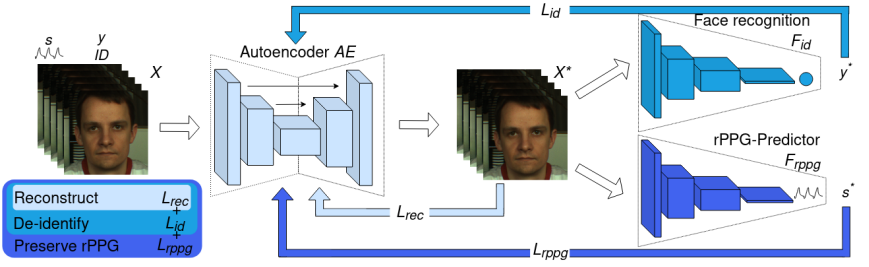

Figure 2: Framework for de-identification that preserves rPPG signal and visual appearance. *AE* is trained to reconstruct the original video while being guided by the de-identification (defined through F_{id}) and rPPG preservation (defined through F_{rppg}) constraints.

3.1 Video reconstruction

Inspired by semi-adversarial approaches [2, 6, 21, 22], we employ an autoencoder with constraints to reconstruct videos. Unlike the previous works that deal with facial expressions and soft biometrics (gender, ethnicity, etc.) that rely on pronounced spatial features, we aim at preserving the underlying physiological signals that depend on subtle spatial-temporal features. This introduces a new set of challenges as the underlying signal is temporally dependant and delicate, meaning that the applied perturbations need to be temporally consistent and not detrimental to the subtle visual cues that compose the rPPG signal. Therefore, inter-frame dependencies need to be taken into account when introducing an rPPG preserving constraint, this requires spatial-temporal modelling, which is why we choose a 3D-CNN for our autoencoder as shown in Tab. 1. The features are only spatially upsampled and downsampled as to not degrade the delicate temporal information. The autoencoder *AE* takes a video X (with T image frames $\{x_t\}_{t=1}^{T}$) as input, and returns the de-identified video X^*.

AE Encoder ↓		*AE* Decoder ↑
3x3x3x32 Conv, LReLU	— skip connection –>	3x3x3x3 Conv, Tanh
2x2x1 AvgPool		2x2x1 Upsample
3x3x3x64 Conv, LReLU	— skip connection –>	3x3x3x32 Conv, LReLU
2x2x1 AvgPool	— x3x3x64 Conv, LReLU –>	2x2x1 Upsample

Table 1: Autoencoder Architecture

It is trained via several constraints: reconstruction, de-identification and rPPG. Firstly, the reconstruction constraint is defined in Eq. 1, and ensures faithful reconstruction of the input X. The loss function is composed of a $L2$ term that minimises the square pixel difference between the input and output videos, and a structural similarity index measure ($SSIM$) calculated over W windows of the T frames to maximise perceived visual quality.

$$L_{rec}(X,X^*) = |X - X^*|_2 + \frac{1}{T}\sum_{t=1}^{T}\frac{1}{W}\sum_{i=1}^{W} SSIM(w_i x_t, w_i x_t^*) \quad (1)$$

3.2 De-identification

We define a de-identification constraint that adds perturbations aimed at deceiving the face recogniser, working at odds with the reconstruction. A pre-trained biometric recognition network F_{id} is used to impose the de-identification constraint, and during evaluation to compare the identification performance from the original and de-identified videos. The video clips used for rPPG measure are usually quite short, e.g., less than 10 seconds, so identity related features in the videos are temporally redundant, representing the same face in similar conditions. Since the identity related features are not time dependant, we choose to use 2D classification models for face recognition, as in previous studies [2, 6]. To make our method more generalisable and to study differences in face recognition models, we consider some widely used deep learning models such as GoogleNet [34], ResNet [15] and DenseNet [14]. Considering the relatively small size of rPPG datasets, for fast and robust convergence, we make use of transfer learning via Imagenet pre-trained weights. We only adapt the last fully-connected layer to fit our task, and then fine-tune the models on our data. We also include two advanced and commonly used face recognition models, i.e., FaceNet with Inception-Resnet backbone [28], and SE-ResNet-50 [13] (SENet). Both are pre-trained on VGGFace2 [1], with only the last Linear layer re-trained on our data. For evaluation, we run the recogniser on each frame of the original and de-identified videos. For training, we exploit the temporal redundancy to lighten the implementation by selecting $K = 8$ random frames of the video and pass to the recogniser, averaging the calculated per-frame losses for the de-identification loss, as shown in Eq. 2. We define the per-frame loss in Eq. 3, in which the 1st term encourages incorrect predictions via negative cross entropy (NCE). Nonetheless, with only the 1st term the downstream face recogniser will still be confident about its inaccurate predictions, thus misrecognize it to another person. To avoid this and make faces unrecognizable, *AE* needs to be penalized for creating output videos that lead to confident identity recognition predictions. We introduce a new output probability regulariser (Reg_{id}), the 2nd term in Eq. 3 that favours perturbations leading to F_{id} softmax outputs that are closer to a uniform distribution, resulting in output videos with less recognisable identity features.

$$L_{id}(X^*, F_{id}, y) = \frac{1}{K} \sum_{x_k^* \in X_K^*} l_{id}(x_k^*, F_{id}, y) \quad where \quad X_K^* = \{X^*(i)\}_{i=1}^K \tag{2}$$

$$l_{id}(x_k^*, F_{id}, y) = \frac{1}{N} \sum_{i=1}^{N} y_i log(F_{id}(x_k^*)) + \lambda \frac{\|F_{id}(x_k^*) - [(1/N)_{\times N}]\|_2}{\|F_{id}(x_k^*)\|_2 + \|[(1/N)_{\times N}]\|_2} \tag{3}$$

3.3 rPPG preservation

The rPPG signal is derived from subtle periodical variations in facial colour, meaning that any perturbation, even if weak in magnitude, can deteriorate the underlying rPPG signal's quality. To mitigate the effect that de-identification has on the rPPG signal, we incorporate a rPPG based constraint. For this, we use a rPPG predictor, that similar to the identity recogniser, imposes a constraint while training the de-identification framework and extracts rPPG signals from original and de-identified facial videos during evaluation. With s and s^* being the rPPG signals extracted from the original video X and reconstructed video X^*, respectively, we define a Pearson correlation based loss, as in Eq. 4. To study generalizability over different F_{rPPG} , we choose methods that are representative of the past and current state of rPPG research as they include a long-established traditional method (CHROM [8]), a widely used CNN method (PhysNet [39]) and recent best performing method (Physformer [41]).

$$L_{rppg} = \frac{\sum_i (s_i - \bar{s})(s_i^* - \bar{s_i}^*)}{\sqrt{\sum_i (s_i - \bar{s})^2 (s_i^* - \bar{s_i}^*)^2}} \quad i \in [0, T[\tag{4}$$

The final training loss is defined by the weighted sum of all losses, as shown in Eq. 5.

$$L = \alpha L_{rec} + \beta L_{id} + \gamma L_{rppg} \tag{5}$$

Figure 3: Loss functions physical meaning visualisation

In Fig. 3, we visualise the physical meaning of our reconstruction (1st column), de-identification (2nd column) and rPPG (3rd column) losses. Reconstruction with the L2 loss leads to unnatural appearing perturbations, with only SSIM global visual features are ignored, but combined they provide subtler perturbations. De-identification with NCE causes inaccurate recogniser predictions, but that are still confident. The Reg_{id} term flattens the output probabilities, leading to less confident wrong predictions. Finally, Pearson correlation (L_{rppg}) maximises trend similarity and produces accurately reconstructed temporal signals.

4 Experiments and Results

We evaluate our proposed method on the PURE [31] and OBF [18] public domain datasets. Experimental results are provided regarding the effectiveness on different biometric recognition networks and generalizability over different rPPG predictors.

4.1 Experimental Setup

Datasets: OBF [18] is a rPPG dataset captured in a controlled environment with stable lighting and minimal movement of the subjects. It contains 200 five-minute-long constant frame rate RGB videos recorded from 100 subjects, with corresponding ground truth ECG and BVP. It was selected due to its large number of diverse subjects and presence of both resting and elevated heart rates. PURE [31] is a rPPG dataset containing 60 one-minute-long videos from 10 subjects. The subjects were recorded under different movement conditions (steady, talking, slow translation, fast translation, slow rotation, medium rotation), with ambient lighting. It was chosen because it contains moderate subject movement and a more complex background.

Evaluation protocol: As utility-preserving de-identification in rPPG is a novel research direction we consider a closed set biometric recognition problem, meaning that each identity is seen during training. Thus, we establish a subject inclusive protocol. For the PURE dataset we utilise the 'steady' videos for the testing set and the other five videos, i.e., 'talking', 'slow translation', 'fast translation', 'slow rotation', and 'medium rotation', for training. For the OBF dataset, we divide each video into five equal size segments, out of which four are used for training and the left one is for testing. In this way, the training and testing data have an equal split for each subject, ensuring a fair biometric evaluation.

Metrics: To measure the global HR information, we extract the mean HR for 30s segments and calculate mean absolute error (MAE) and root-mean-square error (RMSE) between the original and de-identified data. To evaluate the similarity between the rPPG signals, we compare 30s segments by using Pearson's correlation coefficient (R) and Spectral similarity (SS) [3]. For de-identification performance, we use accuracy and equal error rate (EER). To validate the visual acceptance of the reconstructed video outputs, we provide conventional reconstruction quality metrics PSNR in dB and SSIM.

Implementation details: All videos are sampled at 30 fps and each frame is cropped and resized to $128x128$ pixel images, including a tight crop of the facial area. Inputs are $T = 64$ video segments, and no data augmentation is used. The rPPG predictor networks PhysNet [39] and PhysFormer [41] are pre-trained on the OBF and PURE datasets by following the authors' implementation and training parameters. For the traditional method CHROM [8] we re-implement the original method so that it can be used for backpropagation, and landmarks are calculated via Dlib [16] and stabilised with a 5-point moving average filter. For the identification networks GoogleNet [34], ResNet [15] and DenseNet [14], we fine-tune the Imagenet pre-trained weights on our data. For FaceNet [28], and SENet [13] we fine-tune only the last linear layer of the VGGFace2 [1] pre-trained weights. While training the Autoencoder, AE all pre-trained network weights from F_{id} and F_{rPPG} are frozen, and used only to calculate the constraints. AE is trained for 20 epochs using the AdamW optimizer with $epsilon = 1e-8$, $betas = (0.9, 0.99)$, $learning\_rate = 3e-4$, $weight\_decay = 0.05$, $batch\_size = 8$. Loss parameters are set empirically at $\alpha = 1.5$, $\beta = 0.3$, $\gamma = 0.1$, $\lambda = 0.1$ and kept the same for all experiments.

4.2 Experimental Results

To the best of our knowledge, our work is the first learning based method preserving rPPG signals and visual appearance in face de-identification, thus there are no similar methods to compare with directly. Nonetheless, we also assess performance with traditional de-identification methods, that do not preserve visual utility. We perform pixelation by downsampling the original image to $10x10$ and then nearest-neighbor upsampling back to $128x128$. We also implement a recent method that has the objective of preserving the rPPG signal [3]. We use the strongest level of their spatial frequency decomposition, which corresponds to Gaussian blurring. We will refer to if as "blur" and implement it by performing a Gaussian low-pass filter with $\sigma = 12$ on the facial area.

Firstly, we assess the de-identification and rPPG preservation for diverse biometric recognition networks. Tab. 2 shows experiments with the different biometric recognition networks while keeping PhysFormer as the rPPG predictor. In col. 1 is the performance with the original video, the HR and rPPG metrics are perfect as we compare the reference to itself, the biometric performance for each network is high, meaning that they all perform strongly on the data, making them suitable for evaluating de-identification. In col. 2 and 3, both pixelation and blur [3] are not sufficient for de-identification as the overall accuracy is well above random guessing (10% for PURE and 1% for OBF). We notice that pixelation results in both inaccurate mean HR values and signals, while blur has good performance. Blurring preserves the overall average ROI pixel value, often used to extract coarse rPPG signals, thus resulting in good rPPG preservation as it keeps useful low-spatial frequency components. In col. 4 we show the results of de-identification without the rPPG constraint and regularisation term, the de-identification is successful as the accuracy is below 10% and the EER is high, but the perturbations also affect the extracted physiological data negatively resulting in

	Dataset		OBF PURE	PURE					OBF				
	Method		1. Orig.	2. Pixel	3. Blur [8]	4. De-id w/o reg	5. De-id	**6. De-id rPPG**	7. Pixel	8. Blur [8]	9. De-id w/o reg	10. De-id	**11. De-id rPPG**
GoogleNet	HR	MAE ↓	0.00	2.60	0.05	4.74	2.16	**0.04**	1.95	0.49	0.33	0.24	**0.12**
		(MSE) ↓	(0.00)	(10.3)	(0.07)	(15.2)	(7.78)	**(0.05)**	(6.78)	(2.67)	(1.70)	(1.37)	**(1.04)**
	rPPG	R ↑	1.00	0.72	0.95	0.75	0.80	**0.98**	0.57	0.75	0.92	0.95	**0.99**
		(SS) ↑	(1.00)	(0.85)	(0.98)	(0.86)	(0.90)	**(0.99)**	(0.84)	(0.94)	(0.97)	(0.98)	**(0.99)**
	ID%	Acc ↓	100	30.1	21.2	0.00	0.39	**8.79**	19.9	22.0	1.05	1.35	**0.08**
		(EER) ↑	(0.00)	(26.8)	(25.0)	(35.7)	(34.5)	**(37.7)**	(19.5)	(22.7)	(13.6)	(24.3)	**(20.6)**
	Rec.	PSNR[dB] ↑	∞	24.5	24.3	35.3	36.5	**36.2**	19.6	21.1	30.9	30.4	**31.3**
		(SSIM) ↑	(1.00)	(0.63)	(0.68)	(0.97)	(0.97)	**(0.96)**	(0.52)	(0.57)	(0.97)	(0.97)	**0.97**
ResNet	HR	MAE ↓	0.00	2.60	0.05	1.39	2.46	**0.03**	1.95	0.49	0.62	0.28	**0.10**
		(MSE) ↓	(0.00)	(10.3)	(0.07)	(5.54)	(8.09)	**(0.04)**	(6.78)	(2.67)	(5.23)	(1.28)	**(0.67)**
	rPPG	R ↑	1.00	0.72	0.95	0.82	0.80	**0.99**	0.57	0.75	0.91	0.91	**0.99**
		(SS) ↑	(1.00)	(0.85)	(0.98)	(0.92)	(0.89)	**(1.00)**	(0.84)	(0.94)	(0.97)	(0.97)	**(1.00)**
	ID%	Acc ↓	100	48.0	37.0	9.89	0.29	**0.41**	11.3	31.9	2.05	0.03	**1.07**
		(EER) ↑	(0.00)	(29.1)	(25.9)	(28.7)	(45.7)	**(38.7)**	(27.8)	(18.7)	(15.7)	(22.6)	**(25.4)**
	Rec.	PSNR[dB] ↑	∞	24.5	24.3	38.7	38.3	**38.8**	19.6	21.1	31.5	32.4	**32.7**
		(SSIM) ↑	(1.00)	(0.63)	(0.68)	(0.98)	(0.98)	**(0.98)**	(0.52)	(0.57)	(0.98)	(0.98)	**(0.98)**
DenseNet	HR	MAE ↓	0.00	2.60	0.05	10.4	12.9	**0.04**	1.95	0.49	0.25	0.27	**0.11**
		(MSE) ↓	(0.00)	(10.3)	(0.07)	(16.6)	(18.3)	**(0.04)**	(6.78)	(2.67)	(1.18)	(1.07)	**(0.65)**
	rPPG	R ↑	1.00	0.72	0.94	0.04	0.15	**0.98**	0.57	0.75	0.87	0.85	**0.98**
		(SS) ↑	(1.00)	(0.85)	(0.98)	(0.64)	(0.56)	**(0.99)**	(0.84)	(0.94)	(0.96)	(0.95)	**(0.99)**
	ID%	Acc ↓	100	44.5	59.2	1.56	8.99	**10.3**	14.8	29.2	1.08	1.12	**1.15**
		(EER) ↑	(0.00)	(29.1)	(24.3)	(30.9)	(47.8)	**(39.9)**	(21.8)	(21.0)	(8.16)	(11.8)	**(11.2)**
	Rec.	PSNR[dB] ↑	∞	24.5	24.3	36.3	35.7	**35.3**	19.6	21.1	30.9	30.9	**30.9**
		(SSIM) ↑	(1.00)	(0.63)	(0.68)	(0.97)	(0.97)	**(0.96)**	(0.52)	(0.57)	(0.98)	(0.98)	**(0.97)**
FaceNet	HR	MAE ↓	0.00	2.60	0.05	0.06	0.06	**0.03**	1.95	0.49	0.11	0.09	**0.07**
		(MSE) ↓	(0.00)	(10.3)	(0.07)	(0.07)	(0.07)	**(0.04)**	(6.78)	(2.67)	(0.66)	(0.58)	**(0.65)**
	rPPG	R ↑	1.00	0.72	0.95	0.92	0.93	**0.98**	0.57	0.75	0.98	0.98	**0.99**
		(SS) ↑	(1.00)	(0.85)	(0.98)	(0.96)	(0.97)	**(0.99)**	(0.84)	(0.94)	(0.99)	(0.99)	**(1.00)**
	ID%	Acc ↓	100	9.89	9.89	0.21	0.04	**0.02**	1.02	1.02	0.48	1.05	**0.87**
		(EER) ↑	(0.00)	(48.0)	(49.7)	(51.7)	(79.3)	**(72.6)**	(48.2)	(48.7)	(29.5)	(28.7)	**(29.3)**
	Rec.	PSNR[dB] ↑	∞	24.5	24.3	38.0	38.3	**37.4**	19.6	21.1	38.1	38.3	**38.2**
		(SSIM) ↑	(1.00)	(0.63)	(0.68)	(0.98)	(0.98)	**(0.97)**	(0.52)	(0.57)	(0.99)	(0.99)	**(0.99)**
SENet	HR	MAE ↓	0.00	2.60	0.05	4.91	2.94	**0.02**	1.95	0.49	0.23	0.19	**0.11**
		(MSE) ↓	(0.00)	(10.3)	(0.07)	(12.8)	(8.24)	**(0.03)**	(6.78)	(2.67)	(1.26)	(0.89)	**(0.89)**
	rPPG	R ↑	1.00	0.72	0.95	0.47	0.46	**0.99**	0.57	0.75	0.97	0.96	**0.99**
		(SS) ↑	(1.00)	(0.85)	(0.98)	(0.79)	(0.84)	**(0.99)**	(0.84)	(0.94)	(0.99)	(0.99)	**(1.00)**
	ID%	Acc ↓	100	15.9	0.58	0.00	0.00	**0.00**	4.55	0.81	0.17	0.14	**0.86**
		(EER) ↑	(0.00)	(42.0)	(49.1)	(42.8)	(49.8)	**(48.7)**	(40.3)	(51.6)	(29.4)	(32.3)	**(27.9)**
	Rec.	PSNR[dB] ↑	∞	24.5	24.3	37.0	35.2	**37.2**	19.6	21.1	34.9	34.5	**34.5**
		(SSIM) ↑	(1.00)	(0.63)	(0.68)	(0.98)	(0.97)	**(0.97)**	(0.52)	(0.57)	(0.99)	(0.99)	**(0.99)**

Table 2: rPPG, biometric and reconstruction performance evaluation of our de-identification method with different constraints, and comparison with traditional de-identification.

inaccurate rPPG signals. In col. 5 the output probability regularisation term is applied (comparing to col. 4), resulting overall in better EER, meaning that the term helps in lowering the downstream recognition confidence and thus improves de-identification performance. In col. 6 we add the rPPG preservation constraint, which results in accurate HR measures and rPPG signals while keeping comparable de-identification performance. For the OBF dataset, the outcome is analogous and shown in columns 7 to 11. Blurring in OBF has less satisfactory rPPG preservation, likely due to the more varied data in the much larger dataset. Regarding visual acceptance, our method achieves $PSNR > 30dB$ and $SIM \approx 0.97$ indicating that it is hard to tell the difference between the original and de-identified videos.

Secondly, we test whether the rPPG preserving features can generalise over different predictors. In Tab. 3, we cross test rPPG predictors by using one for the training constraint and another for the evaluation. We fix the biometric recognition network as DenseNet as it creates the most disturbing noise for rPPG. The overall performance is satisfactory even when evaluating with a different predictor with very low HR error and high R, meaning that

the features learned can lead to accurate signals with any rPPG prediction method.

Dataset	Evaluation rPPG Method / Train De-id rPPG Method	CHROM HR MAE ↓ (MSE) ↓	CHROM rPPG R ↑ (SS) ↑	Physnet HR MAE ↓ (MSE) ↓	Physnet rPPG R ↑ (SS) ↑	Physformer HR MAE ↓ (MSE) ↓	Physformer rPPG R ↑ (SS) ↑
PURE	De-id rPPG CHROM	0.016 (0.023)	0.989 (0.995)	0.032 (0.044)	0.980 (0.992)	0.067 (0.083)	0.931 (0.974)
	De-id rPPG PhysNet	0.028 (0.036)	0.956 (0.982)	0.021 (0.031)	0.989 (0.995)	0.057 (0.083)	0.950 (0.983)
	De-id rPPG PhysFormer	0.031 (0.042)	0.943 (0.978)	0.039 (0.059)	0.983 (0.993)	0.037 (0.050)	0.981 (0.993)
OBF	De-id rPPG CHROM	0.499 (3.534)	0.984 (0.993)	0.226 (1.117)	0.937 (0.980)	0.267 (1.412)	0.930 0.982
	De-id rPPG PhysNet	1.186 (5.349)	0.888 (0.948)	0.178 (0.985)	0.974 (0.990)	0.088 (0.533)	0.959 (0.989)
	De-id rPPG PhysFormer	1.361 (6.684)	0.896 (0.953)	0.179 (0.947)	0.943 (0.985)	0.111 (0.648)	0.978 (0.991)

Table 3: Evaluation over different rPPG predictors

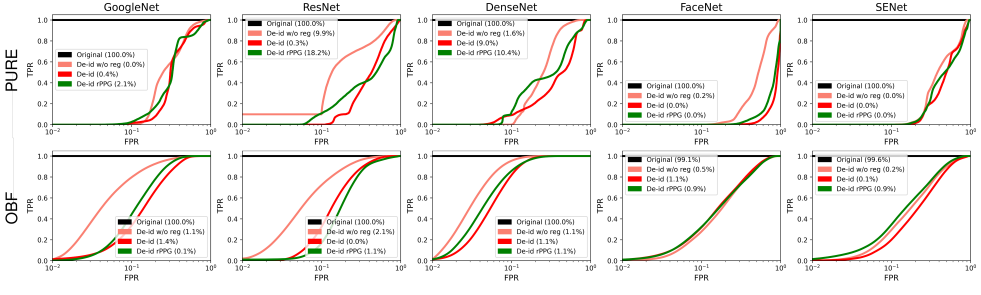

Figure 4: De-identification performance ROC, accuracy in parentheses

Furthermore, in Fig. 4 we show the de-identification results in further detail by plotting receiver operating characteristic (ROC) curves. The addition of the output probability regularisation (red curve vs. pink curve) improves de-identification performance, consistently with the results from Tab. 2, where overall the EER was higher after applying the term. Incorporating the rPPG constraint in most cases results in only a slight drop in de-identification performance.

4.3 Visualisation

We visualise the feature embeddings via PCA of all five biometric recognition networks with the original and de-identifed videos from the PURE dataset, as shown in Fig. 5, and calculate their intra-class and inter-class distance ratios. The ratios between Euclidean distances are calculated from embeddings belonging to the same class and the other classes, and then averaged. The original features are all clearly clustered according to each identity, but when applying de-identification (with or without the rPPG constraint) the feature embeddings for each identity come closer and are more difficult to distinguish, resulting in an increased intra-class over inter-class distance ratio. There are noticeable differences based on the recognition network, as with the same training parameters, DenseNet seems the most challenging to deceive.

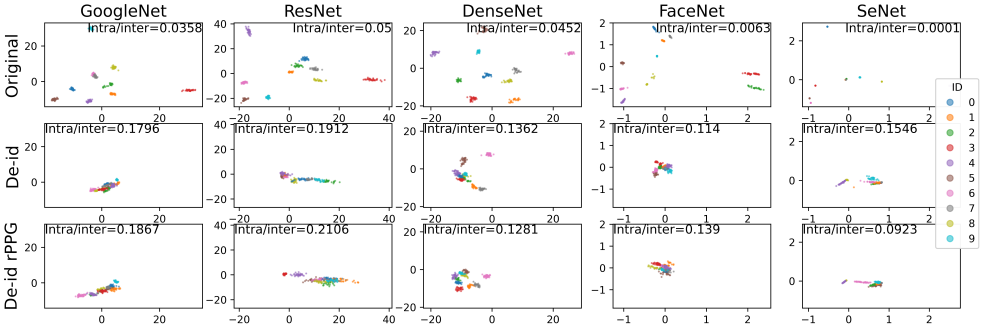

Figure 5: PCA visualisation of recogniser feature embeddings

In Fig. 6 we show a pixel level visualisation of the de-identification and rPPG signals extracted from a 64 frame segment. In contrast to the traditional de-identification methods, our proposed de-identification maintains the overall visual appearance by slightly perturbing the original video. As for the extracted rPPG signals we show that pixelation (i.e., 'Pixel' in Fig. 6) and de-identification without the rPPG constraint (i.e., 'De-id') result in noisy and inaccurate signals, while blurring (i.e., 'Blur') and de-identification with the rPPG constraint ('De-id rPPG') preserve most of the signals' shape.

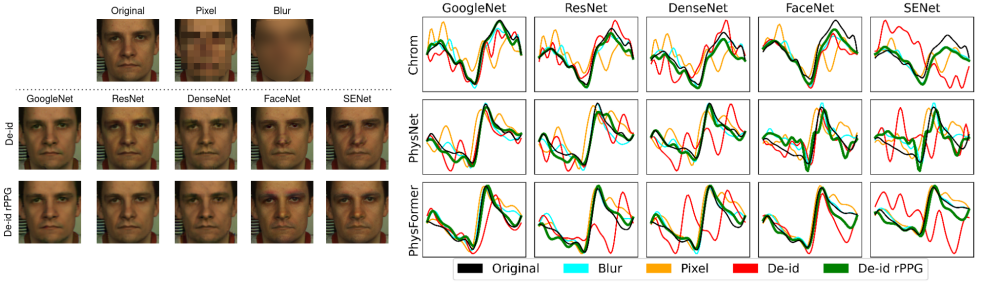

Figure 6: Pixel level and rPPG signal visualisation

5 Conclusion

We propose the first learning based method for facial video de-identification that preserves the physiological and visual fidelity, while protecting user's privacy from machines. Experimental results on two public datasets show that our framework is effective in significantly deteriorating biometric identification while keeping a high quality of the rPPG signal. Additionally, our new output probability regularisation term is shown to aid de-identification by improving EER and ROC. Future work will include more challenging biometric attack scenarios and removal of soft biometrics while preserving rPPG.

6 Acknowledgement

This work was supported by the Research Council of Finland (former Academy of Finland) for Academy Professor project EmotionAI (grants 336116, 345122) and ICT 2023 project TrustFace (grant 345948), and the University of Oulu & Research Council of Finland Profi 7 (grant 352788). As well, the authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[2] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1570–1579, 2018.

[3] Mingliang Chen, Ashira Jayaweera, Chau-Wai Wong, and Min Wu. Identity-Privacy Protection for Facial-rPPG Based Smart Health Research, July 2022. URL https://www.techrxiv.org/articles/preprint/Identity-Privacy_Protection_for_Facial-rPPG_Based_Smart_Health_Research/20352843/1.

[4] Mingliang Chen, Xin Liao, and Min Wu. PulseEdit: Editing Physiological Signals in Facial Videos for Privacy Protection. *IEEE Transactions on Information Forensics and Security*, 17:457–471, 2022. ISSN 1556-6021. doi: 10.1109/TIFS.2022.3142993. Conference Name: IEEE Transactions on Information Forensics and Security.

[5] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proc. ECCV*, pages 349–365, 2018.

[6] Kevin HM Cheng, Zitong Yu, Haoyu Chen, and Guoying Zhao. Benchmarking 3d face de-identification with preserving facial attributes. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 656–660. IEEE, 2022.

[7] Abhijit Das, Hao Lu, Hu Han, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation. In *FG 2021*, pages 01–08, 2021. doi: 10.1109/FG52635.2021.9666996.

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

[9] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014.

[10] EU AI Act. The Artificial Intelligence Act, February 2021. URL https://artificialintelligenceact.eu/the-act/.

[11] GDPR. EUR-Lex - 32016R0679 - EN - EUR-Lex, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

[12] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando De la Torre, and Simon Baker. Face de-identification. *Protecting privacy in video surveillance*, pages 129–146, 2009.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[15] He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 770. IEEE, 2016.

[16] Davis King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 07 2009. doi: 10.1145/1577069.1755843.

[17] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *FedCSIS 2011*, pages 405–410, 2011.

[18] Xiaobai Li et al. The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *FG 2018*, pages 242–249, 2018.

[19] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5008–5017, 2023.

[20] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proc. IEEE/CVF CVPR*, pages 12404–12413, 2021.

[21] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020.

[22] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Privacynet: Semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29: 9400–9412, 2020.

[23] 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, 2020. ISSN 1941-0042. doi: 10.1109/TIP.2019.2947204.

[24] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *ECCV 2020*, pages 295–310, 2020.

[25] Christian S. Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local Group Invariance for Heart Rate Estimation from Face Videos in the Wild. In *2018 IEEE/CVF CVPRW*, pages 1335–13358, June 2018. doi: 10.1109/CVPRW.2018.00172. ISSN: 2160-7516.

[26] 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.

[27] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3, February 2014. ISSN 2047-2501. doi: 10.1186/2047-2501-2-3. URL https://doi.org/10.1186/2047-2501-2-3.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[29] Dangdang Shao, Chenbin Liu, and Francis Tsow. Noncontact Physiological Measurement Using a Camera: A Technical Review and Future Directions. *ACS sensors*, 6(2): 321–334, February 2021. ISSN 2379-3694. doi: 10.1021/acssensors.0c02042.

[30] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. Atrial Fibrillation Detection From Face Videos by Fusing Subtle Variations. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2781–2795, August 2020. ISSN 1558-2205. doi: 10.1109/TCSVT.2019.2926632. Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.

[31] Ronny Stricker, Steffen Müller, and Horst-Michael 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*, pages 1056–1062, 2014.

[32] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 492–510. Springer, 2022.

[33] Zhaodong Sun and Xiaobai Li. Privacy-Phys: Facial Video-Based Physiological Modification for Privacy Protection. *IEEE Signal Processing Letters*, 29:1507–1511, 2022. ISSN 1558-2361. doi: 10.1109/LSP.2022.3185964. Conference Name: IEEE Signal Processing Letters.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[35] Akito Tohma, Maho Nishikawa, Takuya Hashimoto, Yoichi Yamazaki, and Guanghao Sun. Evaluation of Remote Photoplethysmography Measurement Conditions toward Telemedicine Applications. *Sensors*, 21(24):8357, January 2021. ISSN 1424-8220. doi: 10.3390/s21248357. URL https://www.mdpi.com/1424-8220/21/24/8357. Number: 24 Publisher: Multidisciplinary Digital Publishing Institute.

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

[37] 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, 2016.

[38] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2951–2960, 2021.

[39] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *30th British Machine Visison Conference: BMVC 2019. 9th-12th September 2019, Cardiff, UK*. The British Machine Vision Conference (BMVC), 2019.

[40] 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 *Proc. IEEE/CVF ICCV*, pages 151–160, 2019.

[41] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: facial video-based physiological measurement with temporal difference transformer. In *Proc. IEEE/CVF CVPR*, pages 4186–4196, 2022.

[42] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *BMVC 2018*, pages 3–6, 2018.