RGB, a Surrogate of Infrared Facial Videos for Physiological Signs Estimations in Dark

Ankit Gupta, Antonio G. Ravelo-García, and Fernando Morgado Dias

***Abstract*—Physiological signs are key indicators of cardiovascular health, which can be estimated using remote photoplethysmography. Their estimations in dark environments are particularly important, where infrared based methods were predominantly applied, since they are illumination resistant. However, the extracted signals have poor pulsatile strength with low signal-to-noise ratio, eventually resulting in spurious estimates. Conversely, RGB based methods exhibits stronger pulsatile strength, but hindered by poor illumination. To overcome these limitations, we propose 2E1D-Net, trained using a self-created database acquired in a dark environment with marginal illuminance $\leq$ 1 lux. It comprises dual encoders that take paired input images captured at different exposure levels, and project them to a latent. The decoder then, elevates the noise (darkness) component from the dark image, followed by multiscale feature fusion, to produce enhanced images. 2E1D-Net was trained using a linear combination of multiscale structured-similarity-index, L1 and L2 losses, respectively. Subsequently, RGB heart rate and oxygen saturation methods cascaded to trained 2E1D-Net, were tested on self-created and public databases. Experimental results proved the superiority of 2E1D-Net, over state-of-the-art, which ensured the extended ability of RGB methods for physiological measurements in dark, thereby proposing RGB as reliable and clinically relevant alternative to infrared methods without performance compromise.**

***Index Terms*—Dark environments, Deep learning, Independent component analysis, Physiological parameters estimations, Remote photoplethysmography.**

I. Introduction

PHysiological signs are critical indicators of the physiological state of an individual. Their monitoring is vital for various applications, including disease diagnosis, tracking the immediate or long-term effects of surgery, medication therapy, early identification of fatal disorders, and sleep analysis [1]. A widely used approach for their estimation is remote photoplethysmography (rPPG) due to its non-contact nature, which makes it suitable for unobtrusive monitoring over prolonged periods in scenarios such as Neonatal Intensice Care Units (NICU), skin sensitivity, non-contact sleep monitoring, and nighttime driving. Typically, most rPPG methods use face videos to measure subtle temporal colour variations that occur due to blood flow in the arteries, synchronised with the cardiac pulse.

Most real-time physiological sign estimation applications require dim light or dark environments with a significantly higher degree of motion, such as intensive care units and sleep monitoring laboratories. These conditions can be detrimental to the quality of rPPG signals, due to the challenges associated with capturing substantial facial details [2]. Consequently, the resultant signals have a weaker pulsatile amplitude with a low signal-to-noise ratio, which eventually leads to spurious estimates. To deal with such conditions, the infrared (IR) spectrum has proven to be the best choice, as it is resistant to illumination variations. However, the IR-extracted rPPG signal exhibits poor pulsatile strength and is also susceptible to motion artefacts (single channel) [3]. Increasing IR wavelength channels can provide promising results [4], but also increases associated costs and complexity.

Conversely, the RGB spectrum offer better pulsatile strength and motion robustness [4]; however, it is sensitive to illumination variations. In addition, poor illumination due to dark environments makes the estimation of physiological signs more challenging. To counter these challenges, the literature suggests enhancing facial region of interest (ROI) by fine-tuning camera exposure and gain [5], or using image enhancement methods [6]. However, these approaches are limited in the sense that not all cameras are equipped with these features, for instance, embedded cameras on portable devices. Additionally, selecting an efficient enhancement method is not trivial and depends on the illumination condition. Furthermore, the illumination conditions considered in the studies mentioned above range between 1 and 104 lux, which are still far from real-time scenarios, and are confined to the estimation of heart rate (HR) only, despite the equal importance of other physiological signs. Based on the above discussion, we aim to highlight the following key questions with a focus on physiological signs estimations in the dark real-time scenarios:

- What could be an optimal generalised light condition threshold covering most clinical and non-clinical real-time scenarios for physiological measurements in the RGB colour space?
- Is it possible to provide a suitable enhancement method

This manuscript is submitted to the journal in March 2024. This work is supported by the LARSyS Project–UIDP/50009/2020, FCT - Fundação da Ciência e Tecnologia, and Project 761-Smart Islands Hub, managed by Agência Regional para o Desenvolvimento da Investigação, Tecnologia e Inovação, and LERCO CZ.10.03.01/00/22_003/0000003 project via the Operational Programme Just Transition.

Ankit Gupta is with Department of Cybernetics and Biomedical Engineering,Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 17. listopadu 2172/15, 708 00 Ostrava 8, Madeira Interactive Technologies Institute (ITI/LARSyS/M-ITI), and University of Madeira Caminho da Penteada, Funchal 9020-105 Portugal email(ankit.gupta@iti.larsys.pt)

Antonio G Ravelo-García is with Madeira Interactive Technologies Institute (ITI/Larsys/M-ITI), 9020-105 Funchal, Portugal, and Institute for Technological Development and Innovation in Communications, Universidad de Las Palmas de Gran Canaria, 35001 Las Palmas de Gran Canaria, Spain email(antonio.ravelo@ulpgc.es).

Fernando Morgado-Dias is with Madeira Interactive Technologies Institute (ITI/LARSyS/M-ITI), 9020-105 Funchal, Portugal, and University of Madeira, Penteada Campus, 9020-105 Funchal, Portugal (email:morgado@staff.uma.pt).

that ensures substantial extraction of ROI details to extract the rPPG signal accurately?
- Is it possible to estimate physiological signs other than HR in dark environments?

Having identified the key questions, the next step is to find answers to each of these questions in the literature, as outlined in the following subsection.

A. Literature Review

As mentioned earlier, the feasibility of estimating physiological signs in dark environments depends on preserving the substantial facial ROI details from the captured dark videos. In practice, the combination of image enhancement and estimation of physiological signs can be a potentially viable solution for this task. Therefore, the state-of-the-art developments corresponding to both domains will be presented and analysed in this section.

1) Image Enhancement: Conventional image enhancement methods have inherent limitations, such as illumination blindness, under- or overexposure, poor visual perceptibility, susceptibility to colour distortions, and high noise. To counter these limitations, the first deep learning-based method, LLNet by Lore et al. [7], was developed for image denoising and contrast enhancement. Adopting a more principled approach, several deep learning methods based on retinex theory were also proposed in the literature, including KinD++ [8], PairLIE [9], and Self-Calibrated Illumination (SCI) [10]. The novelty of these methods stems from the unique capabilities of deep learning architectures for image decomposition, illumination enhancement, and reflectance restoration. However, their performance depends on the assumptions of extracting illuminance and reflectance components, such as piecewise smoothing of illumination maps and degradation-free reflectance components, except for KinD++, which does not hold in every scenario.

Additionally, several generative modelling attempts, such as EnlightenGAN [11], were also proposed to improve generalisability and overcome the overfitting problem. Similarly, LED-Net [12] was proposed to model different types of degradation for effective enhancement.

To counteract the issue of limited labelled data for enhancement, zero-shot learning was also explored for image enhancement, resulting in methods such as Zero-DCE++ [13] and BrightsightNet [14] (improved variants of Zero-DCE [15]). These methods primarily learn the mapping between low-light images and parameter maps based on quadratic curves. Recently, context-aware mapping methods [16], [17] have also been proposed to use text prompts for enhancement, which rely on the accurate identification of style embeddings [16] and text semantics [17], respectively.

All of the methods mentioned above belong to the category of low-light enhancements; however, the literature lacks a concrete definition of the term *low-light* which needs to be quantified for physiological sign estimations in dark conditions. To counteract this ambiguity, after conducting experiments under different light conditions, we found that a marginal illuminance value of 1 lux or less is suitable for designing estimation pipelines for applications ranging from sleeping environments to night-time driving, which answers the first question presented in the previous subsection.

2) Physiological signs estimations: IR spectra are robust to poor illumination or corresponding variations. Therefore, several methods such as TURNIP [18] and AutoSparsePPG [4] were used for dark driving scenarios at night and at NICUs, respectively. To account for weaker pulsatile strength and motion susceptibility, Wang, Woster, and Brinker [19] introduced multichannel IR spectra for robust motion-resistant HR estimations in dark environments. A more sophisticated approach, combining RGB and near infrared (NIR) was also proposed in numerous studies such as Lie et al. [20], Park et al. [21], Kado et al. [22], respectively. However, these approaches significantly increased the data dimensionality and complexity, resulting in sophisticated data processing pipelines.

Conversely, several RGB-based non-contact estimation methods were also proposed for physiological measurements in low-light environments. Concretely, the following studies [2], [5], [6], [23] proposed a cascaded combination of enhancement and estimation for HR estimations in illumination environments ranging between 1 and 400 lux. These studies were confined to HR estimations only. Furthermore, the results showed the suboptimal performance of the HR estimations under low illuminance conditions. Interestingly, Oxygen saturation (SpO2) estimations have not been attempted so far in dark / low light environments, possibly due to their dependence on infrared wavelengths. Therefore, this study does not only attempt to estimate HR but also SpO2 based on the conventional Ratio of Ratios (ROR) approach using red and blue channels, from videos acquired in dark environments.

B. Contributions

In line with the limitations presented in Sections I-A1 and I-A2, the main contributions of this work are as follows:

- Assuming an average illuminance $\leq$ of 1 lux, a new dark video data set was acquired using the data acquisition system presented in Fig. 6, with characteristics presented in Table I.
- A deep learning architecture two encoders-One decoder Network (2E1D-Net), comprising two encoders and a decoder with a weighted loss function, was proposed to enhance the image frames of dark videos.
- 2E1D-Net being the best performing enhancement method (see Table II, and Fig. 7) was cascaded to state-of-the-art (SOTA) non-contact estimation methods to estimate HR and SpO2 in the proposed dark environment.
- The combination of 2E1D-Net with U-LMA, and ROR outperformed all RGB and IR based HR, and SpO2 estimation methods.

II. Proposed Methodology

The notion of RGB-based physiological measurements in dark environments represents a relatively new concept, diverging from the conventional reliance on IR spectra to estimate physiological signs [4], [18], [19], [24], [25]. Furthermore, the absence of a publicly available database that addresses the

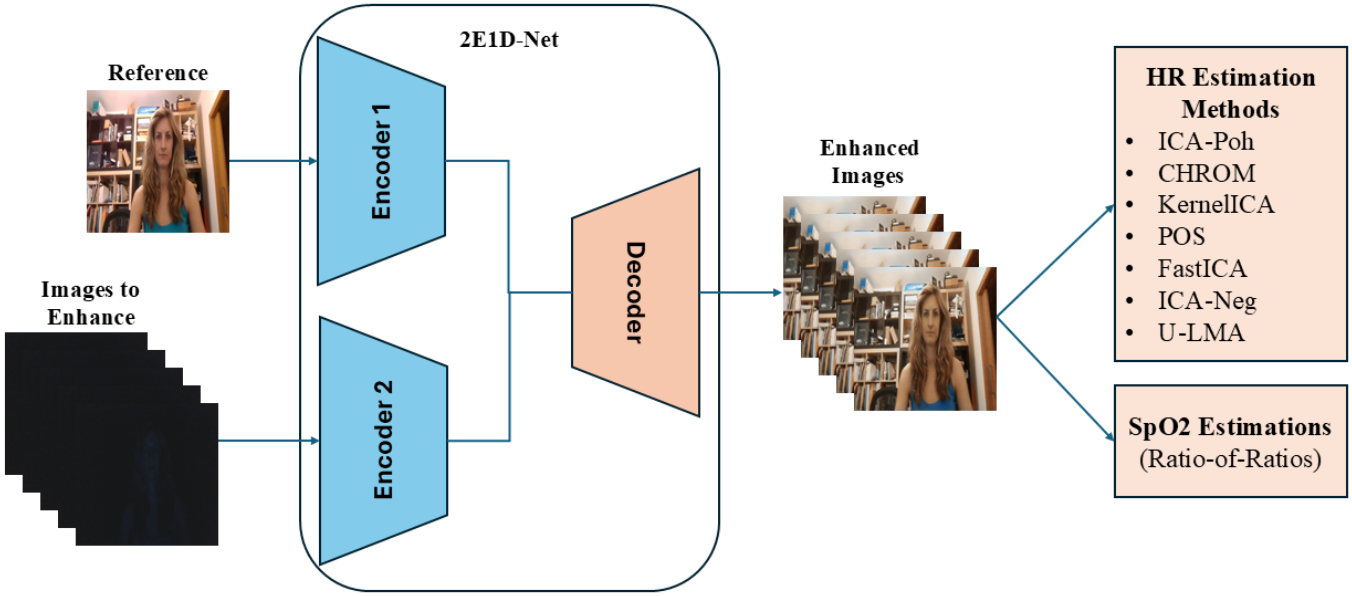

Fig. 1: A flow diagram of the proposed method for extracting HR and SpO2 estimations in dark environments.

specific conditions considered in this study adds to the novelty. Existing studies typically operate within an illuminance range of 1 to 400 lux, different from the assumed low-light conditions in our study. Consequently, we propose a new database to explore the feasibility and reliability of the proposed approach.

The proposed approach for physiological measurements in the assumed environment involves two main steps: 1) enhancing the quality of facial video frames and 2) estimating physiological signs (HR and SpO2). Fig. 1 visually outlines this process. Effective enhancement of image frames requires an efficient enhancer capable of revealing details hidden in darkness. Subsequently, a method for estimating physiological signs should accurately extract the rPPG signal, facilitating correct HR and SpO2 estimations.

Given the limitations of existing image enhancement methods in providing substantial improvements, we introduce a new deep learning-based image enhancement method called 2E1D-Net. This method uses a paired sample of images captured at different exposure levels for enhancement. The structure of 2E1D-Net is inspired by the concept of colour information propagation proposed by Welsh et al. [26], with the difference that instead of an image, the feature representations of the images at the multi-scale level are transferred for image enhancement. The trained model is then integrated with SOTA non-contact estimation methods for HR and SpO2, respectively. The subsequent sections present the details of the proposed approach.

A. Mathematical Formulation of the Enhancement Task

The videos acquired under the assumed dark conditions consist of a sequence of poorly illuminated frames, making it challenging to extract sufficient facial details, and consequently to track temporal colour variations due to pulsatile blood flow, required for extracting Photoplethysmography (PPG) signals for physiological measurements. Therefore, these image frames need to be enhanced for physiological measurements. Assuming darkness as D and cleaner image as I', poorly illuminated or dark image I^d, can be defined as follows:

$$I^d = I' + D \tag{1}$$

Since the distribution of darkness is unknown, approximating D and subtracting it from I^d, results in $\widetilde{I}$, which is an approximation of I'. To introduce this inversion, we will need a substantially cleaner version of I^d, that can provide chromatic and textural prior, to learn an enhancement function G, that reconstructs a noise suppressed and illumination-consistent image $\widetilde{I}$ to preserve temporal colour variations for extracting rPPG signal as:

$$\widetilde{I} = G(I^d \,\|\, I^a) \tag{2}$$

a) 2E1D-Net design rationale: Unlike generic enhancement frameworks optimised for perceptual aesthetics, our objective is to ensure that the reconstructed image frames should retain the subtle colour variations critical for rPPG, that requires constant global tone and contrast for stable facial region tracking and precise reconstruction of local fine-scale variations corresponding to periodic skin reflectance changes. An encoder-decoder architecture satisfies these requirements by supporting hierarchical abstraction and progressive reconstruction, i.e. coarse-scale features govern global illumination balance, while fine-scale features capture local physiologically relevant chromatic oscillations. We therefore propose a dual encoder, single decoder configuration (2E1D-Net), in which one encoder processes low signal-to-noise ratio (SNR) input I^d to extract structural and temporal cues, and the second encoder processes I^a to propagate illumination and colour priors.

b) Dual-encoder and multi-scale residual fusion: Let $G_{\mathrm{En}}^{(d)}$ and $G_{\mathrm{En}}^{(a)}$ denote encoders for I^d and I^a, respectively, and G_{De} the decoder. Furthermore, both encoders extract hierarchical features at multiple spatial resolutions, essential for

handling the non-uniform illumination and spatially varying noise due to dark environments. The coarse level captures global illumination structure, and fine features preserve texture, colour, and shape variations. The multi-scale feature representations for both encoders can be represented as:

$$F_{d,i} = G^{(d)}_{\mathrm{En},i}(I^d), \qquad F_{a,i} = G^{(a)}_{\mathrm{En},i}(I^a). \tag{3}$$

The decoder reconstructs $\widetilde{I}$ by progressive upsampling. After each transposed convolution $T_i(\cdot)$, illumination-context features from the reference encoder are injected through residual learning, followed by a convolutional refinement block $\psi_i(\cdot)$:

$$F_{\mathrm{dec},i} = \psi_i\left(T_i(F_{\mathrm{dec},i-1}) \bigoplus_{\mathrm{res}} F_{a,i}\right), \tag{4}$$

where $\bigoplus_{\mathrm{res}}$ denotes element-wise addition across number of kernels, and ψ_i comprises the convolution and activation layers. The multi-scale features extracted at each scale ($G^{(d)}_{\mathrm{En},i}$) was propagated to the decoder based on dimensionality constraints. After the n decoder stages, the enhanced image is obtained as:

$$\begin{aligned}\widetilde{I} &= G_{\mathrm{De}}(\{\, F_{\mathrm{dec},i} \,\}_{i=1}^{n}) \\ &= G_{\mathrm{De}}\left(\{\, \psi_i(T_i(F_{\mathrm{dec},i-1}) \bigoplus_{\mathrm{res}} F_{a,i}) \,\}_{i=1}^{n}\right)\end{aligned} \tag{5}$$

A low pass filtering (a few convolution operations) was also applied to counter the chromatic overcompensation (Fig. 9(b)), which occurred due to disproportionate amplification of red and green colour channels, then green.

c) Interpretation and architectural innovation: Equations (4)–(5) formalise the multi-scale residual fusion process, where features from the reference encoder progressively guide the decoder at different resolutions. This cross-branch conditioning enables illumination correction and noise suppression, while preserving the fine chromatic details that encode physiological pulsations. In particular, unlike Retinex-based or single-encoder U-Net architectures, 2E1D-Net performs physiologically constrained enhancement through residual cross-illumination fusion, explicitly designed to maintain both global photometric consistency and local temporal colour fidelity essential for rPPG signal estimation in dark environments.

d) Objective: The network parameters are optimised using a composite loss:

$$\min_{G} \; \mathcal{L}(G) = \sum_{j} \alpha_j \, \mathcal{L}_j, \tag{6}$$

where $\mathcal{L}_j$ represents individual loss components: L_1, L_2 and MS-SSIM, and α_j are the corresponding weighting coefficients.

B. Loss Function

In the absence of appropriate ground truth, a loss function with well-defined constraints is key to the generalisability and robustness of unsupervised deep learning models. Fortunately, a paired image sample with different exposure levels with carefully designed loss components can efficiently guide the network training process, which eventually must result in enhanced images. These images, when temporally aligned, should be able to preserve substantial ROI details, which would facilitate the temporal extraction of subtle colour variations for the extraction of the rPPG signal. Therefore, enhanced images should be able to preserve substantial image details, such as colour, shape, and texture.

Therefore, assuming that the paired image sample [I^d,I^a] shares a similar object of interest, the proposed loss function aims to exploit the similarity of image samples. Colour information can be preserved with channel-wise pixel intensity differences between the enhanced image, that is, $L_1 = \left\| \widetilde{I} - I^a \right\|_1$. Furthermore, the higher degree of distortions in I^d resulted in shape irregularity, as shown in Fig. 2(d) (the reflectance map is shown for a better representation), which can be alleviated by restoring the edges of the object. Based on a study by Zhao et al. [27], it was found that mean squared error (MSE) can efficiently preserve edges. Therefore, edge preservation can be achieved using MSE between the $\widetilde{I}$ and I^a as $L_2 = \left\| \widetilde{I} - I^a \right\|_2^2$. Finally, texture or structural similarity can be improved by calculating the structural similarity index, also known as structured similarity index metric (SSIM) [28], therefore the following equation $S_L = (1 - SSIM(\widetilde{I}, I^a))$ could be used to train the network. However, the flatter regions of the image cannot be improved by using the conventional SSIM, as the network could not preserve the local structure, and splotchy artefacts will also be reintroduced due to a substantially low standard deviation in those regions. This problem can be resolved using multi-scale SSIM [27]; therefore, the above equation is modified to $(1 - MS - SSIM(\widetilde{I}, I^a))$. Based on the above analysis, the loss function can be defined as:

$$\min_{G} loss(G) = (\alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 S_L) \tag{7}$$

where α_1, α_2, and α_3 are the balancing factors of L_1, L_2, and S_L. Eq. (7) is optimised with an Adamax algorithm [29] with a learning rate of $1e-4$, to minimise the loss function proposed to enhance the images.

C. 2E1D-Net

Based on section II-A, a two encoder, one decoder network named 2E1D-Net is proposed, which takes paired image samples, each captured at different exposure levels. In particular, 2E1D-Net relaxes the condition of equal pairwise low-high combinations for enhancement, as it needs only one reference image to enhance the whole dark video. Also, the reference image used for contextual information propagation is not the ground truth but an approximation with better perceptual visibility than dark image frames of the video. The schematic diagram illustrating the 2E1D-Net architecture is presented in Fig. 3. The encoder-decoder framework alleviates the dark component of the dark image. On the other hand, the additional encoder improves the enhancement process by extracting and transferring the contextual information at a multi-scale level from the high-exposure level image using residual learning. Additionally, the feature representations of the last convolution

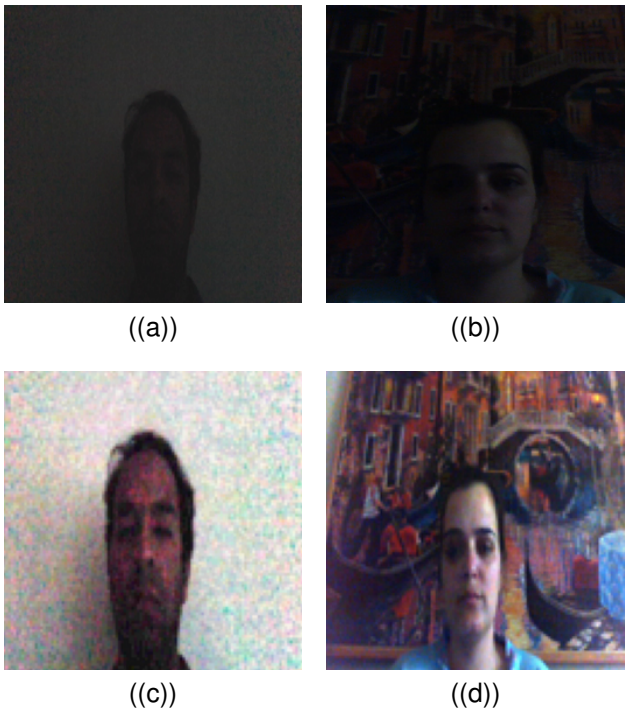

Fig. 2: Original images (a,b), and corresponding reflectance map (c,d) from fine-tuned Kind++. Note that the distortions are neither visible in original images nor illumination maps, so reflectance maps are used to show the distortions in the random image samples.

blocks of both encoders are fused before being passed to the decoder. Subsequently, the feature maps of the additional encoder, at different scales, are fused with the deconvoluted feature map of the respective deconvolution block through residual connections. The decoder also possessed a refinement block to avoid the effect of minute colour distortions causing chromatic overcompensation and preservation of image details, followed by a sigmoid activation to produce an enhanced image. A detailed explanation of various components of the proposed 2E1D-Net is presented in the following subsections.

1) Encoder architecture: The encoder aims to encode the images to their equivalent feature representations for image enhancement tasks: 1) to efficiently segregate the noise from the information of interest due to their different distributions, and 2) to ensure appropriate contextual information propagation for improved enhancement.

The 2E1D-Net encoders share a similar architecture consisting of five convolution blocks (CBs), as illustrated in Fig. 4. The first CB consists of four convolution layers, the last two incorporating Rectified Linear Unit (ReLU) activations. The convolution layers extract features, while ReLU activations help to learn complex patterns. In contrast, the other CBs consist of ReLU-activated convolution layers and a maximum pooling layer to sub-sample the prominent features of the feature map. All convolution layers have a kernel size of 3 with stride 1, while maximum pooling has a kernel size and stride of 2, respectively.

2) Decoder architecture: The decoder decodes the feature representations by hierarchically decoding the features from the encoders. The input to the decoder architecture is the aggregated set of features of both encoders. It consists of three deconvolution Blocks (DCBs), each consisting of a deconvolution followed by ReLU-activated convolution layers (Fig. 5). The deconvolution layer decodes feature representations by maintaining the same connectivity pattern as during encoding. The respective ablation study demonstrated that the encoder-decoder framework (ED-Net) could not preserve colour information and suffers from shape irregularity. Therefore, having shared the object of interest by paired image samples [I^d, I^a], the respective feature maps of I^a (from the additional encoder) are fused with the deconvoluted feature maps at different scales using residual learning. Subsequently, the fused feature map is passed through a ReLU-activated convolution layer of DCB for feature extraction. Following, a refinement block was also used to elevate the effect of minute colour distortions that cause chromatic overcompensation and preserve image details. It consists of four convolution layers with ReLU activation, except for the last, where a sigmoid activation is applied to generate the enhanced image. The contributions of these components of 2E1D-Net are demonstrated using carefully designed ablation studies in the following sections.

D. Physiological Signs Estimations

The trained 2E1D-Net is cascaded with SOTA RGB spectra-based non-contact physiological sign estimation methods for HR and SpO2 measurements in the dark environment (illuminance $\leq$ 1 lux) to achieve two objectives: 1) investigate the ability of 2E1D-Net to preserve substantial image details for accurate extraction of PPG information, and 2) test the conjunctions mentioned above for non-contact estimations of HR and SpO2, in dark environments. This work compares nine SOTA HR estimation methods and ROR method for SpO2 estimations, operating in the RGB colour space, respectively. Finally, the best conjunction was also compared with SOTA IR-based methods to investigate if the proposed conjunctions operating in the RGB colour space could provide performance comparable to IR-based methods, satisfying the darkness constraint. Since darkness conditions were not reported in the respective IR spectra-based studies, the relevant studies were selected if the proposed methods were tested in dark environments. In addition, the default conditions for each method were maintained for better performance and fair comparisons.

E. Dark-Video Dataset

Due to the entirely different and challenging real-time darkness conditions assumed in this work, a database, namely, the *Dark-Video* dataset, is proposed. This can be considered a step to alleviate the limitation of limited publicly available datasets for rPPG signal extraction with an emphasis on dark environments. The data set was collected considering factors such as age, ethnicity, sex, etc., which affect the quantity and quality of the extracted rPPG signal.

The proposed *Dark-video* dataset captured by the system presented in Fig. 6 consists of an image captured in ambient

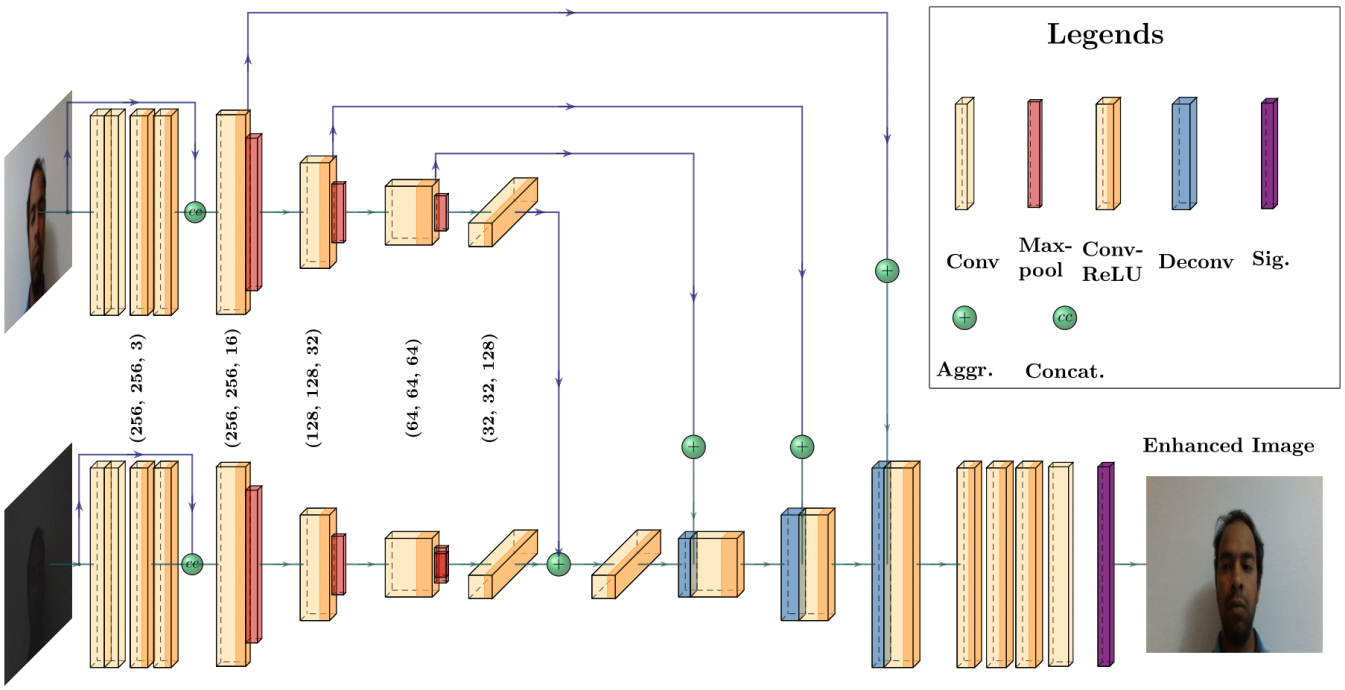

Fig. 3: Architecture of Proposed 2E1D-Net. The dimensions of feature maps of the decoder can be determined from Encoder, due to the requirement of same dimensions for feature aggregation.

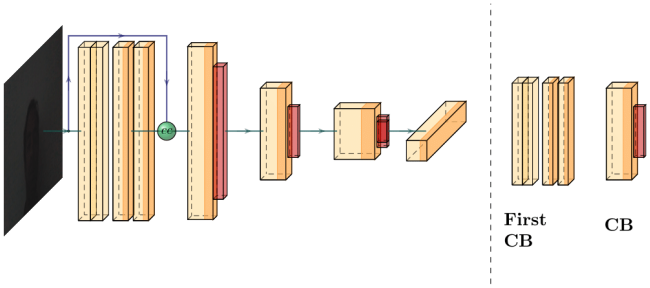

Fig. 4: Encoder architecture consists of two types of convolution blocks (1st CB) and (CB). 1st CB consists of 4 convolution layers with ReLU activations (Yellow with orange color); CB consists of a convolution layer with ReLU activation, followed by a max-pooling layer (dark orange).

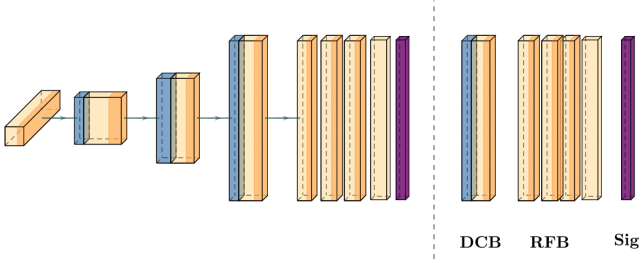

Fig. 5: Decoder architecture consists of deconvolution blocks (DCB) and refinement blocks (RFB). DCB consists of a transposed convolution layer (gray) and a ReLU-activated convolution layer. RFB has the first three ReLU-activated convolution layers, followed by sigmoid activation (purple).

light, a video for 90 seconds in a dark environment using a webcam embedded on a laptop, synchronised with ground-truth HR and SpO2 values using a *CMS60C* pulse oximeter for each subject. A schematic representation of the image and video acquisition system is shown in Figs. 6(a) and 6(b), respectively. The dark environment, characterised by an average illuminance value ≤ 1 lux, is measured and maintained using a lux metre *XFUK-881F* with a resolution of 0.1 Lux / Fc and a range of 1 to 400 000 lux. The proposed value of 1 lux was empirically found based on data collection under different illuminance values, ensuring dark environments, where the dark environment is defined as conditions infeasible to extract the face regions for physiological measurements. The data set will be made available to researchers upon email request, followed by signing the database usability agreement [1]. The data set comprises 57 compressed RGB images and videos collected from 55 subjects with diverse ethnic regions (45 European, 5 Asian and 5 African) and gender (41 males and 14 females), aged between 18 and 61 years. The video and image samples were captured using five different webcams: 720p Face Time HD camera (Apple Macbook Pro), HD webcam 720p (Asus Vivobook S15 S530F), Logitech C170-480p, Logitech hd720, to account for different camera characteristics and variable frame rates, respectively. Each participant was

[1] http://dark-video.biesalab.org/

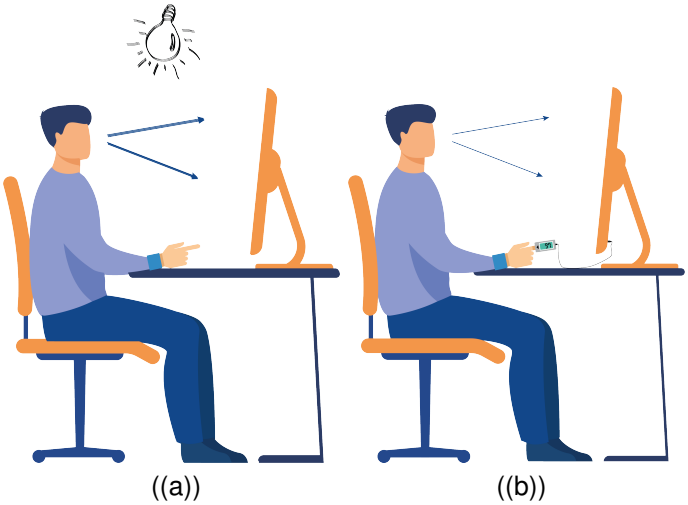

Fig. 6: Acquisition System: (a) Image acquisition consisting of a person facing the camera with illumination source for high exposure image capturing,(b) Video Acquisition system consisting of laptop camera with an internal source of light, and oximeter sensor attached to the index finger of the subject. (Video and high exposure image capturing share a common object of interest, i.e., facial region of the subject).

TABLE I: Proposed Database Summary

Features	Values
Participants	55
Age Range	19-61 years
Ethnicity	45 European, 5Asian, 5 African
Data	Image, Video (90 seconds), Heart rate and SpO2 values
Gender	14 females and 41 males
conditions	Ambient (image) and dark (video)
HR range	46-93 bpm
SpO2 range	94-99%

asked to sign the informed consent, approved by the ethical and data protection committee of the University of Madeira by the identification number 1 / CEUMAI / 2021, before participating in the data collection process. The summary of the proposed database is presented in Table I.

It is important to mention that the study by Xi et al. [2] has also proposed a data set that includes some video samples that partly correspond to the illumination condition considered in this study. Specifically, unlike Xi et al. [2], the video samples of the proposed Dark-Video dataset were collected in extremely dark environments ($\leq$ 1 lux), which correspond to sleeping and night-time driving environments, such as unlit highways or rural roads. Therefore, the novelty of the proposed data set can be explained based on the application-specific context that corresponds to driving environments at night and at sleep. Furthermore, the data set proposed by Xi et al. [2] consists of video samples ranging between 1 lux and 100 lux to analyse the effect of various illumination conditions on physiological measurements, while the proposed Dark-Video data set was explicitly created considering physiological measurements in extremely dark conditions.

III. Results

This section first presented details on the implementation of the proposed image enhancement method, as well as the HR and ROR methods. The analysis is divided into four parts: 1) a comparative analysis of image enhancement methods; 2) another comparative analysis of RGB-based HR and SpO2 estimation methods, cascaded to best image enhancement method; 3) analysis of the best enhancement-estimation combination under different illumination conditions using a publicly available dataset; and 4) compare the best combination from 2), with existing IR-based methods. In addition, a root mean square error (RMSE) analysis was also performed for all HR estimation methods, to provide insights about the estimation performance.

A. Implementation details

A data set comprising 57 videos was used to design experiments for image enhancement and estimation of physiological signs. For image enhancement, videos were transformed into sequential image frames, resulting in $43,534$ images, which were randomly sampled in the 50:25:25 ratio (%), to form training, validation, and testing data sets, ensuring subject independence. This resulted in a training set comprising $21,767$ images, while the validation and testing data set consisted of $10,884$ and $10,883$ images, respectively. The image enhancement network 2E1D-Net was trained using a batch size of 8 with image dimensions $256 \times 256 \times 3$ for 60 epochs using the $Adamax$ optimiser with a learning rate of $1e-4$. An early stopping criterion with a patience value of five was also used as a termination criterion to stop training. For HR and SpO2 measurements, the default conditions reported in the respective studies were maintained for better performance and fair comparisons.

B. Image Enhancement

The performance of 2E1D-Net is compared with its variant ED-Net (single encoder-decoder, taking only dark image as input) and nine SOTA image enhancement methods, which included zero-shot learning (Zero-DCE++ [13], BrightsightNet [14]), Retinex theory (PairLIE [9], and KinD++ [8]), LYT-Net [30], generative modelling (EnlightenGAN [11]), and full low-light image-based methods (NerCO [31], Self-Calibrated Illumination (SCI) [10], LEDNet [12]), respectively. The key component of the methods included for the comparative analysis was that they used the encoder-decoder framework as the baseline. All methods were analysed using the proposed Dark-Video data set, focussing on the specific problem addressed in this work, i.e., HR and SpO2 estimations in the dark environment considered using RGB videos.

All SOTA image enhancement methods were fine-tuned based on the default conditions reported in the respective articles. Subsequently, their performance was analysed qualitatively and quantitatively, using visual comparisons and the following metrics: peak signal-to-noise ratio (PSNR) [32], SSIM [33], naturalness image quality evaluator (NIQE) [34], and learned perceptual image patch similarity (LPIPS) [35]. Furthermore, different ablation studies were designed to further

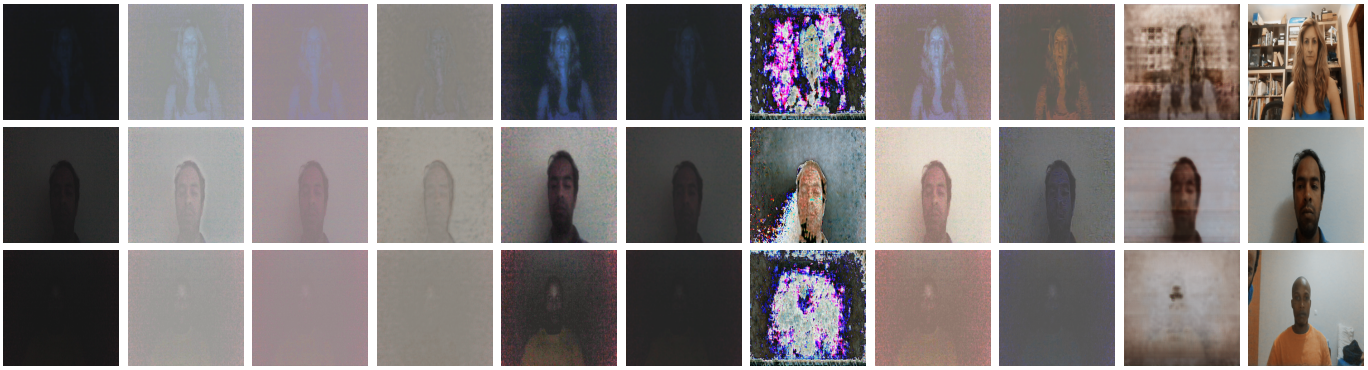

Fig. 7: Comparative analysis of image enhancement methods, depicting image samples from European (top), Asian (middle), and African (bottom) ethnic groups: (a) original, (b) Zero-DCE++, (c) BrightsightNet, (d) PairLIE, (e) KinD++, (f) EnlightenGAN,(g) NerCO, (h) SCI, and (i) LEDNet, and (j)2E1D-Net (Proposed).

investigate the performance of the constituting components of 2E1D-Net, and the visual results are presented in the following subsections.

The qualitative comparisons are presented in Fig. 7, while Table II presents the quantitative comparisons for all image enhancement methods. The suboptimal performance of zero-shot learning methods is attributed to parameter map estimation using quadratic light curves. Due to the presence of vivid colour and texture distortions in the low-light images, the estimated parameter maps, even after fine-tuning, could not provide substantial enhancement. An illustration in Fig. 8(a) illustrates this fact. Additionally, since BrightsightNet and Zero-DCE++ are both based on the same baseline, i.e., Zero-DCE, this finding is assumed to hold for both methods. Similarly, the Retinex theory-based methods (PairLIE and KinD++) could not provide substantial enhancement due to the underlying assumptions of the Retinex theory. Specifically, Retinex theory assumes a degradation-free reflectance map, which is not always feasible in practical situations, as also pointed out in the original KinD++ study [8]. Additionally, since the low-light images were highly distorted, the mechanisms used for illumination refinement and reflectance restoration in these methods could not substantially elevate the effect of these distortions, ultimately resulting in distorted images. Fig. 8(b) presents an illustration of the reflectance map, produced by layer decomposition Net of KinD++, fine-tuned by the proposed data set (the illuminance component is not shown since it is a full black image).

The substandard performance of EnlightenGAN, NerCO, SCI and LEDNet is also apparent in Fig. 7 and Table II, respectively. The suboptimal performance of EnlightenGAN is due to the poor performance of its self-regularised attention mechanism, which is dependent on the illuminance of the low-light image samples. Specifically, the attention mechanism was unable to accurately identify darker regions (only the facial and upper region were identified as dark regions), resulting in enhancement to only these regions, as shown in Fig. 8(c). However, the slight enhancement was due to carefully designed losses and a global-local discriminator. Upon critical analysis of the architecture of NerCO and LEDNet, it was

TABLE II: Comparative analysis of image enhancement methods.

Methods	PSNR	SSIM	NIQE	LPIPS
Zero-DCE++	9.96	0.28	0.18	5.5E-05
BrightsightNet	10.30	0.34	0.18	5.3E-05
PairLIE	10.81	0.39	0.28	6.1E-05
KinD++	10.17	0.28	0.57	6.2E-05
LYT-Net	16.63	0.61	0.40	5.3E-05
EnlightenGAN	10.31	0.37	0.74	3.37E-05
NeRCo	8.76	0.18	0.51	6.7E-05
SCI	10.72	0.37	0.37	5.8E-05
LEDNet	10.48	0.40	0.05	5.3E-05
ED-Net	10.83	0.35	0.14	3.27E-05
2E1D-Net	**34.19**	**0.97**	0.41	**1.00E-06**

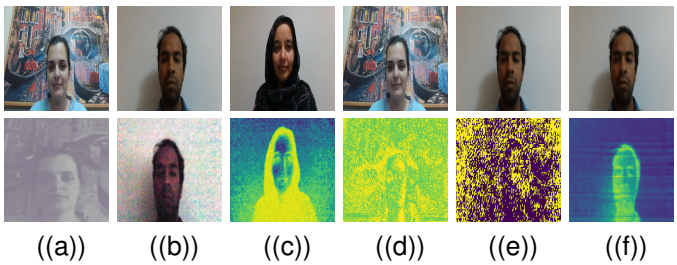

Fig. 8: Visual results of SOTA methods depicting potential reasons for suboptimal performances:(high images (top), and visual results (bottom)): (a) Parameter map from ZeroDCE++ Parameter Map (b) KinD++ reflectance map,(c) EnlightenGAN attention map, (d) NerCO's encoder feature map, and (e) LEDNet feature map, (f) LYT-Net feature map, combining all components.

found that their image enhancement capability is driven by efficient and accurate feature map extraction. Furthermore, feature map extractors of these methods are based on residual nets. The qualitative and quantitative performance of these methods is attributed to inefficient feature extraction. For instance, NerCO used a Resnet-based encoder for feature map extraction in Mask Extractor (ME) and Neural Representation Network (NRN), while LEDNet used Resnet blocks, comprising Residual downsampling and upsampling blocks. To further prove this reason, an illustration of feature maps extracted

from the fine-tuned versions of NerCO (feature map from the ME and NRN encoders) and LEDNet (LE encoder, post-feature fusion at multi-scales) were presented in Figs. 8(d) and 8(e), respectively, which shows inefficient feature extraction (extracted features are highlighted in yellow). Therefore, replacing Resnets with powerful deblurring networks might improve the performance of these methods.

On the other hand, SCI's consideration of enhancing low-light images based on a single enhancement block could not work well, which contradicts their consideration of substantial image enhancement with only one enhancement block, only [10]. It is apparent from Fig. 7(h) that the output looks distorted and visually unpleasing due to the presence of extreme distortions in the low-light images. In addition, the enhanced images also consisted of haze and halo artefacts. Therefore, adding subsequent image enhancement blocks might provide better enhancement results.

The above mentioned methods performed enhancement in RGB colour space. However, LYT-Net proposed by Brateanu et al. [30] used a more robust YUV colour space for enhancement, where the illuminance channel was used to extract relevant feature maps for enhancement, while the chrominance channels were denoised independently. Although its performance was better than other state-of-the-art methods, the enhanced samples showed shape irregularity and colour distortions, as shown in Fig. 7(j). To find potential reasons, the feature representations of the individual components of LYT-Net were investigated. We found that the colour and shape distortions were due to inability to extract robust and invariant relevant features, resulting from object obfuscation in extreme darkness. Consequently, multi-headed self-attention mechanism could enhance the portions corresponding to rich feature representations, followed by feature fusion from all components. A sample of feature representations after the fusion of all components is presented in Fig. 8(f), where the left portion of the image comprises fewer details than the other side.

In conclusion, the image samples of the proposed *Dark-video* dataset were intrinsically distorted in terms of texture, colour, and shape, which contributed to the suboptimal performance of SOTA image enhancement methods, and ED-Net. This observation is consistent with the proposed 2E1D-Net, which will be explained later in the ablation studies. Therefore, to alleviate the effect of these distortions, prior knowledge is required. For instance, PairLIE assumed that a pair of low-light images could provide better insight into enhancement tasks. This supported the fact that the enhancement process needed to be guided by contextual information, as mentioned in Section II-A. 2E1D-Net was designed based on this observation, thus taking image pairs captured at different exposure levels) as input to enhance low-light images. The multi-scale cross-illumination residual fusion mechanism provided by 2E1D-Net, ensured relatively better enhancement, as proved by its superior quantitative metrics and visual representations in Fig. 7 and Table II, respectively. For further investigation, two ablation studies corresponding to its architectural and loss function components are presented in the following subsections.

((a)) ((b)) ((c))

Fig. 9: Ablation studies analysing the contributions of loss functions components: (a)w/o second encoder,(b) w/o Refinement block, and (c) With all components.

C. Ablation studies

1) Architectural Components Ablation Study: This ablation study tested the contribution of an additional encoder and the refinement block of the 2E1D-Net for image enhancement. Therefore, 2E1D-Net was retrained after removing these components, one at a time, resulting in two models: one without an additional encoder (ED-Net) and the other without a refinement block. It is apparent from Fig. 9(a) that the additional encoder contributes significantly to the enhancement process by elevating the effect of distortion by knowledge transfer from similar high-exposure images. This observation also justified the suboptimal performance of SOTA methods due to the absence of prior information during the enhancement process. Fig. 9(b) presents the impact of the refinement block in 2E1D-Net on elevating the effect of colour distortions. As mentioned earlier, these distortions could be removed by applying low-pass filtering operations using a stack of convolution layers with non-linear activations (ReLU and sigmoid). Therefore, the components mentioned above contributed significantly to ensure superior performance (9(c)).

2) Loss Functions Ablation Study: A similar ablation study was also designed to investigate the contribution of various components of the proposed loss function. Specifically, 2E1D-Net was trained by removing one loss component at a time, resulting in three trained models. The visual illustrations of these models are presented in Figs. 10(a), 10(b), and 10(c), respectively. It is apparent from these figures that the absence of L_1 resulted in colour distortions, increased haze, and halo artefacts (rooftop region of the image), while the absence of L_2 caused illuminance degradation, resulting in poor contrast and motion blur. On the other hand, the absence of MS-SSIM resulted in severe colour and shape distortions, both locally and globally, with colour overspreading (chromatic overcompensation). The combination of the functions mentioned above ensured the superior performance of 2E1D-Net by alleviating the effect of the distortions mentioned above.

D. Physiological signs Estimation

1) HR Estimation: The performance of RGB spectra-based, non-contact SOTA HR and SpO2 estimation methods is dependent on the extraction of ROI details, which requires sufficient light conditions. Therefore, their applicability was limited in the case of low-light conditions. This work demonstrated their

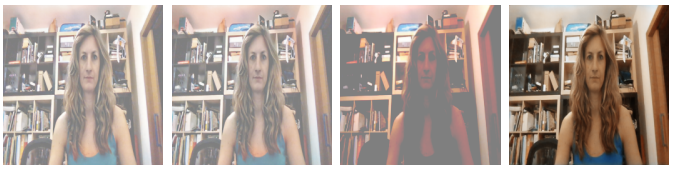

Fig. 10: Ablation studies analysing the contributions of loss functions components::(a) $w/o\ L_1$ loss, (b) $w/o\ L_2$ loss, (c) w/o MS-SSIM, (d) With full loss function.

ability to be applicable in extremely dark conditions when cascaded with an efficient image enhancement method.

Specifically, seven SOTA conventional HR estimation methods cascaded with 2E1D-Net were quantitatively compared using six performance metrics: RMSE, Mean Absolute Percentage Error (MAPE), mean error, standard deviation, accuracy, and Pearson correlation coefficients below the significance level of 0.01 (α). The following HR estimation methods, ICA-Poh [36], CHROM [37], POS [38], KernelICA [39], FastICA [22], and ICA-Neg and U-LMA [40], were included in this analysis, based on the study by Gupta et al. [40]. Additionally, the RMSE of the POS method reported in the study by Xi et al. [2] was also included for the comparative analysis (RMSE is reported only for the illumination condition partly resembling this study, i.e, 1.0 lux). We included only conventional methods for benchmarking physiological measurements in dark environments, providing a consistent baseline against which advanced rPPG methods designed for ambient environments can be compared. The conventional methods were chosen for several reasons:i) they are well-specified, widely implemented and yield consistent results across datasets and illumination conditions; ii) they offer a domain-neutral baseline that does not rely on training under bright conditions, allowing us to isolate the contribution of our enhancement step under ≤ 1 lux; iii) they remain standard benchmarks in both classical and modern literature, facilitating fair comparisons; and iv) their transparent failure behaviour (e.g. substantial ROI details, colour preservation), is particularly valuable for feasibility studies in extreme low-light conditions.

Table III demonstrates the average performance metrics of the methods mentioned above. The substandard performance of ICA-Poh is attributed to video compression and similar pulse and artefact spectra magnitudes, which resulted in a corrupted rPPG signal [36]. The same observation was also proved by the CHROM method, as its alpha-tuning procedure suffered due to similar pulse artefact spectra [37]. However, POS performed relatively better than CHROM due to different projection planes based on physiological information, unlike CHROM, where projection planes were based on specular components (challenging to estimate) [37], [38]. The performance of the aforementioned methods was notably impacted by the significantly similar magnitudes of the pulse signal and artefact spectra. It is potentially due to the prevalence of colour distortions resulting from non-rigid head movement and illumination variations, resulting from the inevitable effect

TABLE III: Performance metrics for HR estimation methods.

Methods	RMSE	MAPE	SD*	μ	Accuracy	ρ*
ICA-Poh [36]	27.52	30.51	19.15	19.92	19.30	0.09
CHROM [37]	15.25	17.39	10.39	11.25	35.09	0.41
KernelICA [39]	12.26	12.52	9.37	8.00	56.14	0.56
POS [38]	10.84	11.46	7.02	8.30	42.10	0.74
FastICA [22]	5.87	6.29	4.09	4.25	70.17	0.86
ICA-Neg [40]	4.36	3.16	3.63	2.46	87.72	0.94
Xi et al. [2]	12.42	-	-	-	-	-
U-LMA	**3.50**	**2.90**	**2.91**	**1.98**	**91.23**	**0.95**

μ: Mean error; SD*: Standard Deviation;r*:Pearson correlation is. calculated at the 0.001 significance level; Accuracy is defined as the percentage of achieving the error difference with ± 5 bpm.Accuracy and MAPE metrics are average percentage values, while other metrics are reported in terms of bpm. a stands for RMSE value corresponding to the illuminance conditions of this study (illuminance value of 1 lux)

of the darkness component in the original videos. In contrast, the study by Xi et al. [2] demonstrated the suboptimal performance of the POS under poor illuminance conditions despite enhancement. The better performance in this study is attributed to the robust image enhancement process of 2E1D-Net. Specifically, it was proved that under a complex dark environment, as considered in this study, it is equally important to address the distortions present in the reflectance components of the image (Fig. 2). However, the enhancement procedure applied in the study by Xi et al. [2] focused on improving the illumination component of the image frames without addressing distortions in the reflectance components, which resulted in the poor performance of POS in their study.

In contrast, KernelICA performed better than the methods mentioned above, due to its kernel density-based Independent Component Analysis (ICA) [41], which is resistant to similar pulse artefact spectra. However, the higher error metrics reported by the method could possibly be due to the assumption of smoothness and continuity of independent components, which is not always possible in practical scenarios, especially in dark environments, despite substantial enhancements.

On the other hand, entropy-based ICAs performed significantly better, since entropy ensured better statistical independence than kurtosis [42]. Specifically, the negentropy-based cost functions of FastICA and ICA-Neg ensured their better performances, unlike other methods. However, ICA-Neg performed slightly better than FastICA, especially in terms of error metrics. This proved that the entropy maximization of cumulative distributions of raw RGB signals ensures better statistical independence than signals themselves. Finally, U-LMA outperformed other methods in all performance metrics due to the observation mentioned above, with support of robust optimisation and faster convergence provided by Levenberg-Marquardt Algorithm (LMA) for entropy maximisation. Considering the recent extension of large language models in remote health care, we have also included Period-LLM by Zhang et al. [43], which is fine-tuned to learn periodic properties from the videos for HR estimations.

For further insights into the superior performance of the 2E1D-Net and U-LMA combination, its performance was also analysed using Bland-Altman analysis, as presented in Fig. 11. The mean bias reported by the Bland-Altman plot (Fig. 11) is -0.4737 beats per minute (bpm), while the upper and

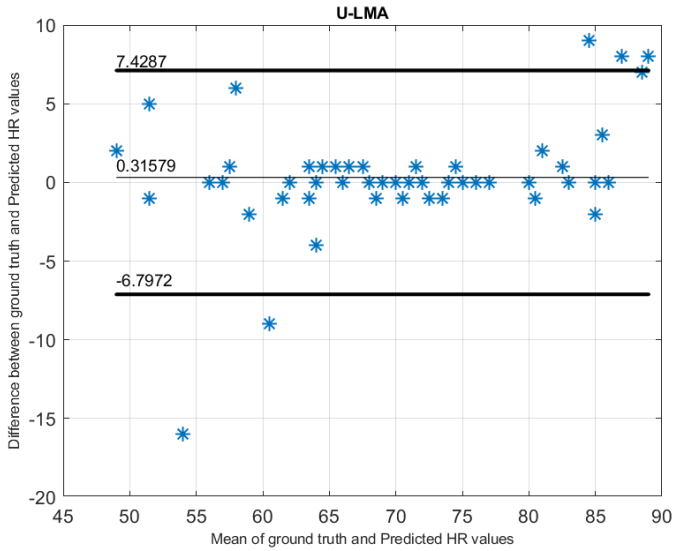

Fig. 11: Bland-Altman plot of 2E1D-Net cascaded to U-LMA.

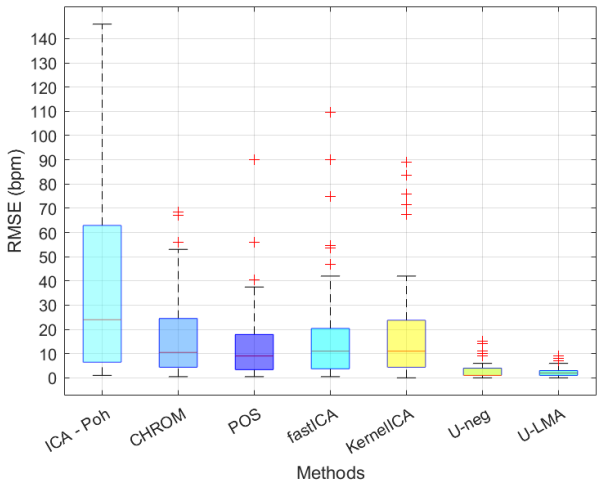

Fig. 12: RMSE analysis of HR estimations study.

TABLE IV: Comparative analysis results of 2E1D-Net-ULMA with IR spectra-based HR estimation methods.

Studies	RMSE	μ	Accuracy
Nowara et al. [45]	11.2	-	64.7
Comas et al. [38]	4.8	-	-
Wang et al. [46]	4	4.04	-
Guo et al. [47]	-	-	82
Van Gastel et al.[48]	-	1.5	87
Jinji et al. [44]	2.09	-	-
2E1D-Net-ULMA	**3.50**	**1.98**	**91.23**

lower statistical limits defined by $\mu \pm 1.96std$ are -7.4645 and 6.5172, respectively. Furthermore, the agreement between ground truth and estimated values is slightly lower, covering 91% of data points within the statistical thresholds, according to the Bland-Altman analysis. This agreement can be improved by considering more video samples for analysis. Additionally, the Pearson correlation coefficient (ρ) of 0.9504, closer to 1 at 0.01% significance, justified the superior performance of U-LMA in conjunction with 2E1D-Net.

2) RMSE Analysis: Fig. 12 presents the RMSE analysis of HR estimation methods used in this study. From the methods, it is apparent that ICA-Poh performed worst, producing the highest magnitude of RMSE. In contrast, CHROM, fastICA, and KernelICA performed relatively better than ICA-Poh. In addition, these methods have shown almost similar performances, as can be seen by almost similar error medians. However, the difference lies in the third and fourth quartile, which can be explained by the reasons listed in Section III-D1. POS has shown slightly better performance than the above methods, due to its robustness in low illumination environments compared to other methods.

Furthermore, U-neg and U-LMA performed substantially better due to their robust cumulative density-based objective function, allowing better statistical dependence of the resultant independent component. However, slightly better performance by U-LMA was due to efficient and robust unmixing matrix updates by customised LMA-based method. Although all methods were prone to outliers, due to inevitable illumination artefacts even after enhancement, the difference lies in the error magnitudes of these outliers. Specifically, U-LMA managed to keep RMSE magnitudes within 10 bpm (including outliers), while other methods failed to do so. Although U-LMA outperformed other methods, its error magnitude proved that it is not possible to fully remove distortions by image enhancement, which raises the need for further improvements.

3) Comparative Analysis with IR-based Methods: Since IR spectra are robust to illumination variations, they have been conventionally used in dark environments. To prove the applicability of RGB spectra in such environments, the best combination, that is, 2E1D-Net and U-LMA, was also compared with SOTA IR spectra-based HR estimation methods. These methods were selected on the basis of their applicability in dark conditions. Therefore, following HR estimation studies by Nowara et al. [4], Comas et al. (TURNIP) [18], Wang, Vosters and Brinker [19], Guo et al. [24], Van Gastel et al. [25] and Jinji et al. [44] were included. Specifically, Table IV presents the performance comparison of IR-based HR estimation methods with the proposed combinations based on reported RMSE, mean error, and accuracy, respectively. The comparable performance of the proposed combination is apparent from the respective table, which suggests the applicability of RGB spectra as an alternative to IR spectra without compromising the performance.

4) SpO2 Estimations: Due to the predominance of the ROR method for SpO2 estimations, it was also cascaded with 2E1D-Net to demonstrate the possibility of non-contact SpO2 estimations in extremely dark environments, considered in this study. However, this analysis was restrictive to a normal SpO2 value range, i.e., 94-99%, due to the associated complexity in acquiring abnormal SpO2 value ranges. Additionally, to overcome the problem of limited samples for two SpO2 values, 94% and 95%, nine samples corresponding to these values, were also included from the VIPL-HR database [49]. Subsequently, a regression model was trained to map the ROR values to the ground truth SpO2 values for which the performance metrics are reported in Table V. Furthermore, the

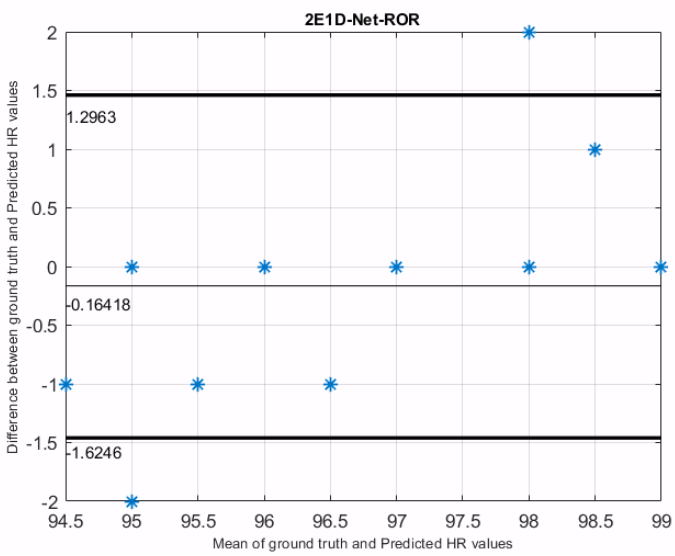

Fig. 13: Bland-Altman Plot of 2E1D-Net cascaded to ROR.

TABLE V: Comparative analysis of contactless SpO2 estimation methods.

Studies	Spectra	RMSE	$\rho*$	MAE
Casalino et al. [22]	RGB	1.64	0.53	1.33
Shao et al. [54]	IR	1.3	0.94	-
Akamatsu et al. [55]	RGB	0.88	0.45	0.55
Guazzi et al. [32]	RGB	-	0.81	2.08
2E1D-Net-ROR	**RGB**	**0.76**	**0.92**	**0.43**

performance of the proposed combination is also demonstrated using the Bland-Altman plot, presented in Fig. 13.

Noably, one data point in the figure corresponds to more than one video sample due to the restrictive SpO2 value range. From Fig. 13, the mean bias between ground truth and estimated values is -0.1343 with upper and lower statistical limits ($\mu \pm 1.96 * std$) as 1.3791 and -1.6477, respectively. Furthermore, 95% of the data points lie between the statistical bounds, which is consistent with the recommendation of the Bland-Altman analysis [50]. In particular, the statistical thresholds were fairly lower than the acceptable error difference, which is critical for the commercial viability of the method. In addition, the Pearson correlation value of 0.92 at the significance level of 0.01% justified a stronger correlation between ground truth and estimated SpO2 values. Hence, the proposed combination of 2E1D-Net and ROR demonstrated the potential for accurate non-contact SpO2 measurements in dark environments.

5) SpO2 comparative analysis: The proposed non-contact SpO2 measurement combination was also compared with SOTA RGB and IR-based estimation methods. The following methods Akamatsu, Onishi and Imaoka [51], Casalino, Castellano, and Zaza [52], and Guazzi et al. [53] were compared using MSE, Pearson's correlation, and mean absolute error (MAE). Table V shows that the proposed combination performed much better than other methods. However, the better performance could be due to the restrictive normal range considered in this work. Therefore, it is important to test this combination using a wider range of SpO2 values, to reach a conclusion.

TABLE VI: Effect of illumination conditions on HR estimation

Metrics	1.0 lux	1.6 lux	2.5 lux
RMSE	8.78	7.25	3.33
MAPE	7.12	6.83	3.03
ME	4.54	4.69	2.23
SD	7.83	5.76	2.59
Correlation	0.69	0.78	0.95
Accuracy	84.62	76.92	92.31

E. Effect of illuminance conditions on HR/SpO2 estimations

To investigate the effect of illuminance conditions on the accuracy and robustness of HR and SpO2 estimations, this study used a publicly available database proposed by Xi et al. [2], which comprises video samples collected under various illuminance conditions (measured in lux ranging between $10^{0.0}$ and $10^{2.0}$). To match the context of the study, we empirically tested the conditions under which ROI extraction was infeasible without enhancement. Therefore, we used video samples collected under three different illumination conditions ($10^{0.0}$,$10^{0.2}$,$10^{0.4}$), i.e., illuminance value with lux 1.0, 1.6, and 2.5 lux, respectively. Since 2E1D-Net requires a reference image, the first image frame of the video sample collected at 100 lux was extracted and used as a reference image for enhancement purposes. The performance metrics of U-LMA cascaded to 2E1D-Net are presented in table VI. It is apparent that the proposed combination achieved higher error and lower accuracy and correlation values in lower illuminance, respectively, while these metrics improved at higher illuminance values. This is attributed to the presence of minute but inevitable distortions under lower illuminance, which could not be alleviated or removed even by the enhancement procedure. Alternatively, these distortions were relatively less under higher illuminance, resulting in lower error metrics and higher correlation.

Similarly, a performance analysis of SpO2 estimation was also conducted, and the performance metrics were presented in Table VII. It is important to mention that accuracy has not been reported for SpO2 analysis due to the unavailability of metrics related to clinical relevance. However, an error difference of $\pm$ 2-3% is considered for commercial viability. All error differences between subjects were within this range and, therefore, were not considered for this analysis. The results show a similar trend for SpO2 estimations as HR, for the same reason.

TABLE VII: Effect of illumination conditions on SpO2 estimation

Metrics	1.0 lux	1.6 lux	2.5 lux
RMSE	0.68	0.48	0.28
MAPE	0.32	0.24	0.08
ME	0.31	0.23	0.08
SD	0.63	0.44	0.28
Correlation	0.95	0.94	0.97

F. Key Observations and Limitations

Extensive experiments demonstrated that in addition to IR, RGB spectra can also be a viable solution to estimate HR and

SpO2 in scenarios such as NICUs and sleeping environments. However, the illumination environment, i.e., illuminance ≤ 1 lux assumed in this work, is empirically selected and does not necessarily conform to the environments mentioned above. As mentioned earlier, several previous attempts, employing RGB spectra, had been made to test the feasibility of HR estimations in dark environments [6], [5], [23]; however, this work has proven its novelty in terms of illumination conditions, and demonstrating the ineffectiveness of existing SOTA image enhancement methods for extremely dark scenarios. Furthermore, this work also demonstrated the extended ability of conventional non-contact RGB-based HR / SpO2 estimations in conjunction with an efficient and robust image enhancement method.

However, this work has certain limitations: first, although 2E1D-Net was able to enhance dark images substantially, it has a dependence on the quality of its slightly illuminated counterparts; second, HR and SpO2 estimations in some scenarios require additional image processing for accurate assessments; and third, due to the complexity associated with SpO2 values acquisition, this study considered the normal SpO2 range $94 - 99\%$; and finally, deep learning models like rPPG-MAE [56], and PhysFormer++ [57], CodePhys [58], and Period-LLM [43], and approaches integrating explicit and implicit prior knowledge [59] have been developed and validated primarily under well-lit or ambient conditions. Adapting such models to extreme low-light conditions ($\leq$1 lux) would require substantial architectural and training modifications; therefore, using them naively in this study could led to misleading comparisons. Addressing these limitations forms the basis of our future research directions.

IV. Conclusion

This study proved the feasibility and reliability of the RGB colour space to estimate HR and SpO2 estimations in extremely dark environments (luminance ≤ 1 lux) by cascading an efficient and robust image enhancement method with conventional HR and SpO2 estimation methods. Identifying the reasons for the suboptimal performance of existing image enhancement methods in the proposed illumination condition, a two-encoder and one-decoder architecture was proposed and trained with a novel loss function (weighted combination of L_1, L_2 and multi-scale SSIM). The encoder-decoder framework aimed to alleviate the darkness component in low-light images, while the feature representations of the slightly better exposed counterpart, extracted from the additional encoder, were propagated and fused post-deconvolution (in the decoder) at a multi-scale level. Subsequently, 2E1D-Net was cascaded with conventional SOTA RGB-based HR and SpO2 estimation methods and compared and analysed to demonstrate their applicability in extremely dark environments, with an additional Bland-Altman analysis for the best combinations. In addition, the study also proved the reliability and efficacy of the best HR and SpO2 combinations compared to IR-based methods in dark environments.

Acknowledgements

The authors also thank IT4Innovations for providing GPU infrastructure and anonymous reviewers for their suggestions and recommendations to improve this manuscript for publication in the journal.

References

[1] V. Rideout and J. Beneken, "Parameter estimation applied to physiological systems," *Mathematics and Computers in Simulation*, vol. 17, no. 1, pp. 23–36, 1975.

[2] L. Xi, W. Chen, C. Zhao, X. Wu, and J. Wang, "Image enhancement for remote photoplethysmography in a low-light environment," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 1–7, IEEE, 2020.

[3] J. Hu, Y. He, J. Liu, M. He, and W. Wang, "Illumination robust heart-rate extraction from single-wavelength infrared camera using spatial-channel expansion," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3896–3899, IEEE, 2019.

[4] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Near-infrared imaging photoplethysmography during driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3589–3600, 2020.

[5] I. Odinaev, J. W. Chin, K. H. Luo, Z. Ke, R. H. So, and K. L. Wong, "Optimizing camera exposure control settings for remote vital sign measurements in low-light environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6085–6092, 2023.

[6] S. Chen, S. K. Ho, J. W. Chin, K. H. Luo, T. T. Chan, R. H. So, and K. L. Wong, "Deep learning-based image enhancement for robust remote photoplethysmography in various illumination scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6076–6084, 2023.

[7] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[8] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *International Journal of Computer Vision*, vol. 129, pp. 1013–1037, 2021.

[9] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K.-K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22252–22261, 2023.

[10] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5637–5646, 2022.

[11] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE transactions on image processing*, vol. 30, pp. 2340–2349, 2021.

[12] S. Zhou, C. Li, and C. Change Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *European Conference on Computer Vision*, pp. 573–589, Springer, 2022.

[13] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4225–4238, 2021.

[14] Z. Chen, J. Yang, and C. Yang, "Brightsightnet: A lightweight progressive low-light image enhancement network and its application in "rainbow" maglev train," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 10, p. 101814, 2023.

[15] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1780–1789, 2020.

[16] S. Kosugi and T. Yamasaki, "Personalized image enhancement featuring masked style modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[17] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12918–12927, 2023.

[18] A. Comas, T. K. Marks, H. Mansour, S. Lohit, Y. Ma, and X. Liu, "Turnip: Time-series u-net with recurrence for nir imaging ppg," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 309–313, IEEE, 2021.

[19] W. Wang, L. Vosters, and A. C. den Brinker, "Continuous-spectrum infrared illuminator for camera-ppg in darkness," *Sensors*, vol. 20, no. 11, p. 3044, 2020.

[20] W.-N. Lie, D.-Q. Le, C.-Y. Lai, and Y.-S. Fang, "Heart rate estimation from facial image sequences of a dual-modality rgb-nir camera," *Sensors*, vol. 23, no. 13, p. 6079, 2023.

[21] S. Park, B.-K. Kim, and S.-Y. Dong, "Self-supervised rgb-nir fusion video vision transformer framework for rppg estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[22] S. Kado, Y. Monno, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Spatial-spectral-temporal fusion for remote heart rate estimation," *IEEE Sensors Journal*, vol. 20, no. 19, pp. 11688–11697, 2020.

[23] H. Wang and S. Zhang, "Non-contact human respiratory rate measurement under dark environments by low-light video enhancement," *Biomedical Signal Processing and Control*, vol. 85, p. 104874, 2023.

[24] K. Guo, T. Zhai, M. H. Purushothama, A. Dobre, S. Meah, E. Pashollari, A. Vaish, C. DeWilde, and M. N. Islam, "Contactless vital sign monitoring system for in-vehicle driver monitoring using a near-infrared time-of-flight camera," *Applied Sciences*, vol. 12, no. 9, p. 4416, 2022.

[25] M. van Gastel, B. Balmaekers, S. B. Oetomo, and W. Verkruysse, "Near-continuous non-contact cardiac pulse monitoring in a neonatal intensive care unit in near darkness," in *Optical diagnostics and sensing XVIII: Toward point-of-care diagnostics*, vol. 10501, pp. 230–238, SPIE, 2018.

[26] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 277–280, 2002.

[27] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[28] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1488–1499, 2011.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] A. Brateanu, R. Balmez, A. Avram, C. Orhei, and C. Ancuti, "Lytnet: Lightweight yuv transformer-based network for low-light image enhancement," *IEEE Signal Processing Letters*, 2025.

[31] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12918–12927, October 2023.

[32] A. J. Santoso, L. E. Nugroho, G. B. Suparta, and R. Hidayat, "Compression ratio and peak signal to noise ratio in grayscale image compression using wavelet," *International Journal of Computer Science and Technology*, vol. 2, no. 2, pp. 7–11, 2011.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[34] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[36] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[37] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[38] W. Wang, A. C. den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[39] R. Song, S. Zhang, J. Cheng, C. Li, and X. Chen, "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method," *Computers in Biology and Medicine*, vol. 116, p. 103535, 2020.

[40] A. Gupta, A. G. Ravelo-García, and F. M. Dias, "A motion and illumination resistant non-contact method using undercomplete independent component analysis and levenberg-marquardt algorithm," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 4837–4848, 2022.

[41] A. Chen, "Fast kernel density independent component analysis," in *International Conference on Independent Component Analysis and Signal Separation*, pp. 24–31, Springer, 2006.

[42] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[43] Y. Zhang, H. Lu, Q. Hu, Y. Wang, K. Yuan, X. Liu, and K. Wu, "Period-llm: Extending the periodic capability of multimodal large language model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29237–29247, 2025.

[44] Jinjitan, J. Huang, Q. He, M. Huang, Z. Wu, and Q. Chen, "Heart rate measurement in low-light conditions and human natural sleeping positions based on video," in *2023 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 501–507, 2023.

[45] G. d. Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[46] K. D. Fairchild, D. E. Lake, J. Kattwinkel, J. R. Moorman, D. A. Bateman, P. G. Grieve, J. R. Isler, and R. Sahni, "Vital signs and their cross-correlation in sepsis and nec: a study of 1,065 very-low-birth-weight infants in two nicus," *Pediatric research*, vol. 81, no. 2, pp. 315–321, 2017.

[47] T. J. Moss, D. E. Lake, J. F. Calland, K. B. Enfield, J. B. Delos, K. D. Fairchild, and J. R. Moorman, "Signatures of subacute potentially catastrophic illness in the intensive care unit: model development and validation," *Critical care medicine*, vol. 44, no. 9, p. 1639, 2016.

[48] K. M. van der Kooij and M. Naber, "An open-source remote heart rate imaging method with practical apparatus and algorithms," *Behavior research methods*, vol. 51, pp. 2106–2119, 2019.

[49] X. Niu, H. Han, S. Shan, and X. Chen, "Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video," in *Asian conference on computer vision*, pp. 562–576, Springer, 2018.

[50] D. Giavarina, "Understanding bland altman analysis," *Biochemia medica: Biochemia medica*, vol. 25, no. 2, pp. 141–151, 2015.

[51] Y. Akamatsu, Y. Onishi, and H. Imaoka, "Heart rate and oxygen saturation estimation from facial video with multimodal physiological data generation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1111–1115, IEEE, 2022.

[52] G. Casalino, G. Castellano, and G. Zaza, "Evaluating the robustness of a contact-less mhealth solution for personal and remote monitoring of blood oxygen saturation," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2022.

[53] A. R. Guazzi, M. Villarroel, J. Jorge, J. Daly, M. C. Frise, P. A. Robbins, and L. Tarassenko, "Non-contact measurement of oxygen saturation with an rgb camera," *Biomedical optics express*, vol. 6, no. 9, p. 3320, 2015.

[54] D. Shao, C. Liu, F. Tsow, Y. Yang, Z. Du, R. Iriya, H. Yu, and N. Tao, "Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1091–1098, 2015.

[55] C. G. Scully, J. Lee, J. Meyer, A. M. Gorbach, D. Granquist-Fraser, Y. Mendelson, and K. H. Chon, "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 303–306, 2011.

[56] X. Liu, Y. Zhang, Z. Yu, H. Lu, H. Yue, and J. Yang, "rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements," *IEEE Transactions on Multimedia*, vol. 26, pp. 7278–7293, 2024.

[57] Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, P. Torr, and G. Zhao, "Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1307–1330, 2023.

[58] S. Chu, M. Xia, M. Yuan, X. Liu, T. Seppänen, G. Zhao, and J. Shi, "Codephys: Robust video-based remote physiological measurement through latent codebook querying," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[59] Y. Zhang, H. Lu, X. Liu, Y. Chen, and K. Wu, "Advancing generalizable remote physiological measurement through the integration of explicit and implicit prior knowledge," *IEEE Transactions on Image Processing*, 2025.

Ankit Gupta is a researcher in Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava. He completed his Ph.D. and Master of Technology in Computer Science and Engineering from the University of Madeira and Maharishi Markandeshwar University, and a Bachelor of Technology in Bioinformatics from Jaypee University of Information Technology, India. His research interests include dimensionality reduction, deep learning, remote photoplethysmography, and computer vision.

Antonio Gabriel Ravelo García is an Associate Professor in the Department of Signal and Communications and Institute for Technological Development and Innovation in Communications at University of Las Palmas de Gran Canaria. He has participated in different research projects and has published numerous papers in scientific journals and conferences. His research interests include biomedical signal processing, non-linear signal analysis, data mining, remote photoplethysmography, and sensor-based systems.

Fernando Morgado-Dias received his Master's degree in Microelectronics from the University Joseph Fourier in Grenoble, France, in 1995 and his Ph.D. from the University of Aveiro, Portugal, in 2005 and is currently a full professor with Habilitation at the University of Madeira and Researcher at ITI/Larsys. His research interests include machine learning, sleep, digital hardware, remote photoplethysmography, and renewable energy.