

Implicit Neural Models to Extract Heart Rate from Video

Pradyumna Chari^{*1}, Anirudh Bindiganavale Harish^{*†2}, Adnan Armouti^{†3},
Alexander Vilesov¹, Sanjit Sardar¹, Laleh Jalilian¹, and Achuta Kadambi¹

¹ University of California Los Angeles, Los Angeles CA 90095, USA
{pradyumnac,vilesov,sanjitsarda}@ucla.edu, ljalilian@mednet.ucla.edu
achuta@ee.ucla.edu

² Rice University, Houston TX 77005, USA
anirudhbh@rice.edu

³ Cornell Tech, New York NY 10044, USA
armouti@cs.cornell.edu

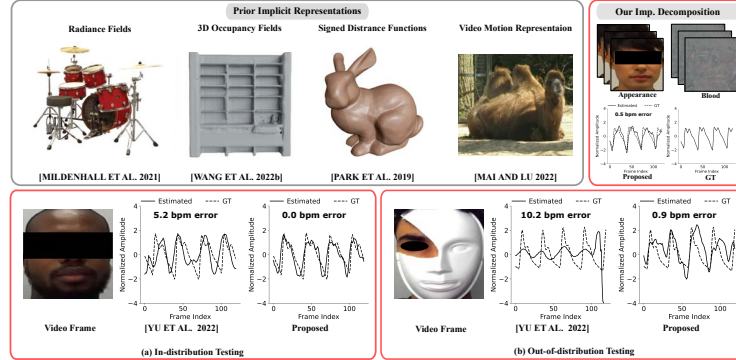


Fig. 1: Prior implicit neural models represent scenes for diverse applications. We propose an implicit neural representation (INR) to decompose face videos and isolate blood flow information. Our INR decomposes input videos into visual appearance and blood flow (“ $\mathcal{A}$ - $\mathcal{B}$ decomposition”). The decomposed data aids in estimating the remote photoplethysmography (rPPG) signal and heart rate.

Abstract. Scene representation networks, or implicit neural representations (INR) have seen a range of success in numerous image and video applications. However, being universal function fitters, they fit all variations in videos without any selectivity. This is particularly a problem for tasks such as remote plethysmography, the extraction of heart rate information from face videos. As a result of low native signal to noise ratio, previous signal processing techniques suffer from poor performance, while previous learning-based methods have improved performance but

* - Contributed Equally, †- Work done when at UCLA

suffer from hallucinations that mitigate generalizability. Directly applying prior INRs cannot remedy this signal strength deficit, since they fit to both the signal as well as interfering factors. In this work, we introduce an INR framework that increases this plethysmograph signal strength. Specifically, we leverage architectures to have selective representation capabilities. We are able to decompose the face video into a blood plethysmograph component, and a face appearance component. By inferring the plethysmograph signal from this blood component, we show state-of-the-art performance on out-of-distribution samples without sacrificing performance for in-distribution samples. We implement our framework on a custom-built multiresolution hash encoding backbone to enable practical dataset-scale representations through a 50x speed-up over traditional INRs. We also present a dataset of optically challenging out-of-distribution scenes to test generalization to real-world scenarios. Code and data may be found at <https://implicitppg.github.io/>.

Keywords: Implicit neural representations · Remote plethysmography

1 Introduction

Neural scene representations have gained significant prominence over recent years. The ability to use neural networks as function fitters and encode scene information within their weights has found considerable applications for tasks such as image representations [20, 27, 40, 44], scene representations [15, 39, 54, 61], radiance fields [18, 29, 30] and video representations [7, 9, 25]. Consistent among these applications is the need to fit all possible scene variations without selectivity.

However, for the task of remote plethysmography, the contactless monitoring of heart rate information from face videos via subtle skin color variations, selecting for the blood flow is required. Therefore, this indiscriminate fitting behavior of current implicit neural representations (INRs) is a significant drawback: both the signal and interfering factors will get fitted. Prior work [6, 16, 36, 49] has established the low Signal-to-Noise Ratio (SNR) of the measurement (in this case, the facial image) as the source of remote photoplethysmography (rPPG) performance degradation. Previous methods generally fall into two categories: signal processing-based or deep learning-based. These two classes have established a trade-off between domain generalizability and in-distribution performance.

Signal processing techniques improve signal strength by using model-based assumptions (such as band-pass filters or simple coloration models [11, 51]). These usually lead to relatively larger heart rate estimation errors. Recently, learning-based methods [8, 57, 58] for plethysmography have achieved state-of-the-art (SOTA) performance through end-to-end data-driven approaches. However, these methods suffer from generalization issues, where samples that are “out-of-distribution” (OOD) may show unspecified and often undesirable behavior due to overfitting.

We propose an INR-based framework to improve the plethysmograph signal strength in facial videos before estimating the blood volume pulse signal.

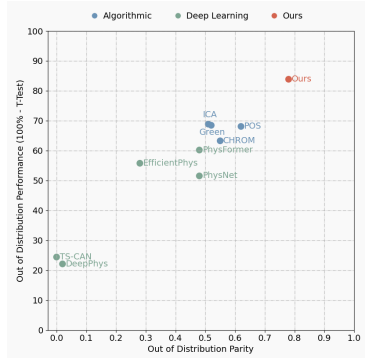


Fig. 2: Our implicit representation for rPPG achieves Pareto-optimality across out-of-distribution (OOD) performance and inter-distribution parity compared to prior algorithmic and learning-based methods. It performs better on OOD samples while maximizing parity between in-distribution and OOD performance. Table 1 shows metrics used for this plot. Higher is better along both axes.

Specifically, we introduce a new Appearance-Blood (or $\mathcal{A}\text{-}\mathcal{B}$) decomposition for facial video data. The decomposition is designed to improve the plethysmograph signal quality across the face. We show that the representation capacity of carefully designed INRs, with architectures motivated by light transport principles of the problem (namely the nature of subtle skin color variations due to the plethysmograph signal), can enable such a decomposition, allowing isolation of the relevant blood component. Conventional INRs, however, are slow to train, thereby making them intractable for dataset-scale analyses. We use multiresolution hash encodings inspired by [32] to facilitate fast $\mathcal{A}\text{-}\mathcal{B}$ decomposition. To the best of our knowledge, we propose the first camera-based rPPG method that uses implicit neural representations as a critical component of the estimation pipeline. Different from prior methods, we focus on both in and out-of-distribution performance (specifically, optically challenging OOD).

On a self-collected dataset of optically challenging OOD scenes, our method surpasses existing algorithmic and learning-based methods (Figure 2) with reliable plethysmograph and heart rate estimates. Notably, we achieve this OOD performance gain without compromising on in-distribution performance on rPPG datasets (Figure 1). As an additional outcome, we can obtain high-fidelity estimates of the plethysmograph signal strength (referred to as neural signal strength mask) across the face that generalize with high fidelity to OOD scenes and settings such as face paint, masks, glasses, and even reflection from windows. In summary, our contributions are as follows:

1. We formulate Appearance-Blood ($\mathcal{A}\text{-}\mathcal{B}$) decomposition as a means of enhancing plethysmograph signal strength. We show that carefully designed implicit neural representations can serve as efficient $\mathcal{A}\text{-}\mathcal{B}$ decomposers and propose a fast method to achieve this.

2. This decomposition is the foundation for our proposed rPPG estimation method. On our optically challenging OOD dataset, our method shows significant improvements in performance over prior algorithmic and learning-based methods (Figure 2) without losing out on in-distribution performance.
3. An optically challenging test dataset, consisting of 104 videos (≈ 1 hour of recorded data) across various optically challenging and OOD configurations, is collected and will be released upon acceptance of this work.

1.1 Scope

This work establishes the effectiveness of fast INRs to learn desired components of video data selectively - in this case, plethysmograph information. Other applications, such as video-based blood oxygenation and blood pressure estimation, require accurate plethysmography as a necessary first step. However, these applications are open research problems and is outside this work’s scope. Furthermore, motion-robust rPPG is an open challenge and requires extensive research on its own. While this is not our primary goal, and our proposed method is not designed to combat motion artifacts, we test performance on motion videos and find our method competitive (supplement). Finally, we acknowledge that OOD is multi-faceted. This work primarily focused on optically challenging scenes and achieved SOTA results while secondarily validating performance on real-world obfuscations such as talking (and associated motions).

2 Related Work

Implicit Neural Representations. Implicit neural representations (INRs) are an increasingly prevalent method to represent scenes. Examples settings include image representations [20, 27, 40, 44], video representations [7, 9, 25], and 3-D scene representations [15, 39, 54, 61]. Specifically, this work is interested in the class of video-based representations. These include methods that use convolutional networks [7], phase-based motion-adjustable representation [25], space-time super-resolution [9] and flow fields for space-time viewpoint synthesis [13, 19]. We take inspiration from these prior methods but have different goals. We aim to use the video representation capacity of implicit networks to isolate imperceptible physiological phenomena, such as plethysmograph signals.

Fast Implicit Neural Representations. A critical drawback of INRs is the time taken to train the networks for dataset scale experimentation and inference. Recent work, however, has taken a step toward alleviating this concern. [28] have proposed using a latent feature-based modulation network to generalize implicit representation models to model a broader range of images for faster convergence. Further, specialized toolboxes have been released to accelerate the training time of NeRFs [18, 32]. In this work, we are interested in employing [32]’s parametric multiresolution spatial-hash encodings as inputs to a shallow MLP. Using this foundation, we extend the framework to video INRs, specifically for

generalizable, robust camera-based plethysmography. This technique can achieve orders of magnitude improvements in training time for INRs.

Remote Plethysmography. Heart rate estimation using rPPG has been actively studied. Early methods were based on algorithmic principles of color variations [11, 51] or motion based on Newtonian reaction to blood flow [4]. Such methods are not limited to use with RGB cameras. Near Infrared (NIR) Imaging with active illumination has been employed to combat the effects of unreliable illumination in the visible (VIS) spectrum [24]. Augmented reality [14] is another avenue for this research. Deep learning approaches have also been utilized to attain SOTA results. [8] used an attention-based Convolutional Neural Network (CNN) while [57] introduced spatio-temporal CNNs. Other work has extended these architectures [35], including using transformers [58], incorporated meta-learning [17], improved PPG waveform characteristics [45], and augmented rPPG datasets with synthetic examples [3, 53]. Another class of methods focuses on improving equity between groups, such as participants of different skin tones [16, 36]. These include algorithmic methods [6] or multimodal fusion methods [49]. Methods for assessing and improving equity are not restricted to rPPG and extend to a range of computer vision problems [5, 16, 42, 52, 56, 59].

3 $\mathcal{A}$ - $\mathcal{B}$ Decomposition and Optimality

We begin with describing the mathematical underpinnings of rPPG and propose the notion of appearance-blood ($\mathcal{A}$ - $\mathcal{B}$) decomposition and its benefits.

Fundamentally, the rPPG signal is subtle, making its estimation difficult. While contemporary methods [8, 57, 58] circumvent this through deep networks, they fare poorly on OOD samples. This necessitates a generalizable method capable of extracting the rPPG signal without relying on domain-specific features that tend to overfit. This motivates our functional decomposition.

Given spatial coordinates $\mathbf{x} \in [-0.5, 0.5]^2$, and temporal coordinate $t \in [-0.5, 0.5]$, the face videos are interpreted as RGB color fields $\mathcal{C}(\mathbf{x}, t)$. This color field consists of specular reflections that act as interference and diffuse components arising from subsurface scattering containing plethysmograph information. We wish to decompose the color signal into a signal amplitude component, $p_m(\mathbf{x}, t)$, and a temporal plethysmograph signal $p_i(t)$, our desired signal. However, the relation between $p_m(\cdot, \cdot)$ and $p_i(\cdot)$ is non-linear in nature.

Under reasonable assumptions, we show color signals can be decomposed as:

$$\mathcal{C}(\mathbf{x}, t) = \mathcal{A}(\mathbf{x}, t) + \mathcal{B}(\mathbf{x}, t) + \mathbf{v}_n(\mathbf{x}, t), \quad (1)$$

where $\mathcal{A}(\mathbf{x}, t)$ represents the facial appearance (or $\mathcal{A}$ -function) - including specular highlights - which is interference. $\mathcal{B}(\mathbf{x}, t)$, the blood component (or $\mathcal{B}$ -function), contains the spatiotemporally varying color changes arising out of blood flow (plethysmography) that incorporates $p_m(\cdot, \cdot)$ and $p_i(\cdot)$. $\mathbf{v}_n(\cdot, \cdot)$ represents the measurement noise in the process. We will refer to this decomposition as the $\mathcal{A}$ - $\mathcal{B}$ decomposition. Then, the following Theorem holds.

Theorem 1: *The $\mathcal{A}$ – $\mathcal{B}$ decomposition results in a $\mathcal{B}$ -function with a Signal to Interference & Noise Ratio $SINR_{\mathcal{B}}(\mathbf{x}, t)$, such that $SINR_{\mathcal{B}}(\mathbf{x}, t) \geq SINR_{\mathcal{C}}(\mathbf{x}, t)$. That is, the $\mathcal{A}$ – $\mathcal{B}$ decomposition leads to an SINR gain.*

Note that Theorem 1 is an existence proof. It describes the existence of SINR benefits as a result of ideal decomposition, motivating its necessity in our real applications. Additionally, while SINR is computed across time, we also wish to model temporal variations due to lighting, pose changes, etc. This is what t -dependence for SINR denotes.

Corollary 1: *Given an $\mathcal{A}$ – $\mathcal{B}$ decomposed video field $\mathcal{C}(\mathbf{x}, t)$, the Maximal Ratio Combining (MRC)-optimal estimate for the plethysmograph $p_i(t)$ is given by,*

$$\hat{p}_i^*(t) = \int_{\mathbf{x} \in \Omega} SINR_{\mathcal{B}}(\mathbf{x}, t) \mathcal{B}(\mathbf{x}, t) d\mathbf{x}, \quad (2)$$

where $\Omega = [-0.5, 0.5]^2$, the domain of $\mathbf{x}$ as previously defined.

This is the notion of near-optimal plethysmography we aim to achieve. Detailed derivations for this section can be found in the supplement.

3.1 Optimal Plethysmography and Uncertainty

Equation 2 can also be interpreted in the context of uncertainty minimization [38]. In a discrete setting with N pixel locations $\{\mathbf{x}\}_{i=1}^N$, each pixel is viewed as a sensor and the signal strength, $SINR_{\mathcal{B}}(\mathbf{x}, t)$ is a proxy for the uncertainty or unreliability. Then the problem of estimating $p_i^*(t)$ can be interpreted as Bayesian inference. With assumptions on prior and posterior distributions, it can be shown that the posterior $\hat{p}_i(t)$ is Gaussian with mean $\frac{\sum_{i=1}^N SINR_{\mathcal{B}}(\mathbf{x}_i, t) \cdot \mathcal{B}(\mathbf{x}_i, t)}{\sum_{i=1}^N SINR_{\mathcal{B}}(\mathbf{x}_i, t)}$ (best posterior estimate, a discretized and normalized version of Equation 2) and variance $1/(\sum_{i=1}^N SINR_{\mathcal{B}}(\mathbf{x}_i, t))$. Since Equation 2 represents maximal ratio combining, $\sum_{i=1}^N SINR_{\mathcal{B}}(\mathbf{x}_i, t) \propto SINR_{\hat{p}_i(t)}$ ([43], Equation 31, 32), then, uncertainty (or the variance in the posterior) is inversely proportional to SINR of $\hat{p}_i(t)$. Optimal plethysmography therefore becomes akin maximization of SINR, or minimization of uncertainty of $\hat{p}_i^*(t)$. Please refer to the supplement for derivations.

4 $\mathcal{A}$ – $\mathcal{B}$ Decomposition Using INRs

This section explores conditions for $\mathcal{A}$ – $\mathcal{B}$ decomposition function approximators. We first establish necessary conditions that $\mathcal{A}$ – $\mathcal{B}$ decomposers must meet and then explore how INRs can be designed to perform such decomposition.

4.1 Functional Decomposition

Consider two functions, $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{A}})$ and $\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{B}})$, parameterized by $\boldsymbol{\Theta}_{\mathcal{A}}$ and $\boldsymbol{\Theta}_{\mathcal{B}}$, that aim to represent $\mathcal{A}(\mathbf{x}, t)$ and $\mathcal{B}(\mathbf{x}, t)$ respectively. To perform an $\mathcal{A}$ – $\mathcal{B}$ decomposition, we wish to find functions that satisfying:

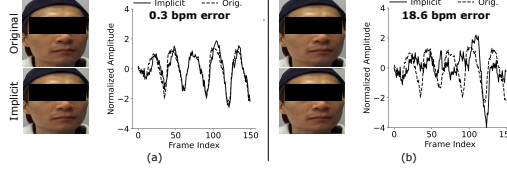


Fig. 3: Sinusoidal Representation Networks (SRNs) can represent both $\mathcal{A}$ and $\mathcal{B}$ -functions, while phase-based methods only represent the $\mathcal{A}$ -function. (a) SRNs, such as [44], can capture PPG color variations almost perfectly, while (b) phase-based motion representations, such as [25] are unable to capture it.

$$\begin{aligned}
 d_{\mathcal{A}}(\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{A}}^*), \mathcal{A}(\mathbf{x}, t)) &\leq \epsilon_{\mathcal{A}}, \\
 d_{\mathcal{B}}(\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{B}}^*), \mathcal{B}(\mathbf{x}, t)) &\leq \epsilon_{\mathcal{B}}, \\
 d_{\mathcal{B}}(\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{A}}^*), \mathcal{B}(\mathbf{x}, t)) &\geq \epsilon_{\mathcal{B}}.
 \end{aligned} \tag{3}$$

where $\boldsymbol{\Theta}_{\mathcal{A}}^*$ and $\boldsymbol{\Theta}_{\mathcal{B}}^*$ are optimal parameters, and $d_{\mathcal{A}}(\cdot, \mathcal{A}(\mathbf{x}, t))$ and $d_{\mathcal{B}}(\cdot, \mathcal{B}(\mathbf{x}, t))$ are some component-dependent metric distances. This quantifies the requirement that $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{A}})$ is able to represent the $\mathcal{A}$ -function but not the $\mathcal{B}$ -function.

We choose distance measures, $d_{\mathcal{A}}(\mathcal{F}, \mathcal{A}(\mathbf{x}, t)) = \|\mathcal{F} - \mathcal{A}(\mathbf{x}, t)\|_2, \forall \mathcal{F}$, such that the appearance metric distance rewards pixel-wise closeness. Similarly,

$$d_{\mathcal{B}}(\mathcal{F}, \mathcal{B}(\mathbf{x}, t)) = |H(\mathcal{F}) - H(\mathcal{B}(\mathbf{x}, t))|, \forall \mathcal{F}, \tag{4}$$

where $H(\cdot)$ is a heart rate estimator function. By finding functions that satisfy the constraints in Equation 3, we can achieve $\mathcal{A}$ - $\mathcal{B}$ decomposition by sequentially learning $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{A}})$ and $\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\mathcal{B}})$. A detailed mathematical analysis of this formulation is deferred to the supplement.

4.2 Identifying Implicit $\mathcal{A}$ - $\mathcal{B}$ Decomposers

Prior work on video magnification through Laplacian [55] and phase-based [50] pyramids is instructive. While the Laplacian pyramid-based approach [55] can magnify blood flow due to color changes, the phase-based method [50] cannot magnify blood flow. With this intuition, we use the metrics defined in the previous section to identify approximations for $\mathcal{A}$ & $\mathcal{B}$. We first look at phase-modulated INRs for motion modeling [25]. The key inductive bias here is parameterizing time as phase modulation in positional encodings [30, 46].

Empirical Claim 1 (based on observations): *Phase-based implicit models (models with phase-encoded input embeddings) cannot represent plethysmograph signals accurately, in terms of the metric in Equation 4.* (Figure 3(b))

While this can be viewed as a limitation of phase-based INRs, we use it to our advantage: phase-based models can be $\mathcal{A}$ -function estimators (Equation 3).

We next look at sinusoid representation networks (SRNs), such as [20, 40, 44]. **Empirical Claim 2:** *SRNs are good $\mathcal{A}$, $\mathcal{B}$ -function approximators.* (Figure 3(a))

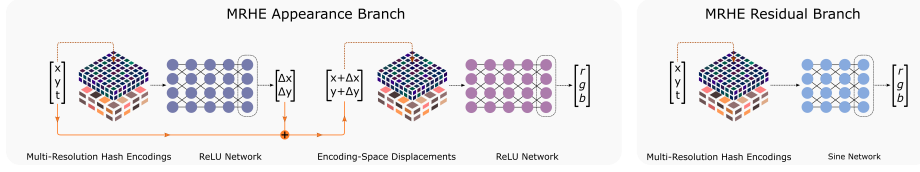


Fig. 4: To enable fast $\mathcal{A}$ - $\mathcal{B}$ decomposition, we use implicit neural representations as decomposing function fitters. Training is done sequentially: first, the cascaded appearance model learns the $\mathcal{A}$ -function. Then, the appearance model is frozen, and the residual model learns the $\mathcal{B}$ -function, thereby completing the decomposition. The use of multiresolution hash encodings makes dataset-scale decomposition viable.

However, being able to represent both $\mathcal{A}$ and $\mathcal{B}$ makes SRNs incapable of $\mathcal{A}$ - $\mathcal{B}$ decomposition. This pair exactly satisfy the constraints in Equation 3. Hence, **Empirical Claim 3:** *A phase-based model can serve as $\hat{\mathcal{A}}(\mathbf{x}, t; \Theta_{\mathcal{A}}^*)$, while an SRN can serve as $\hat{\mathcal{B}}(\mathbf{x}, t; \Theta_{\mathcal{B}}^*)$, when trained sequentially.*

Justifications for all claims may be found in the supplement. However, naive use of prior work [25, 44] for $\mathcal{A}$ - $\mathcal{B}$ decomposition poses a training time problem. Using variants of [25] and [44] as $\mathcal{A}$ and $\mathcal{B}$ -function estimators, respectively, we noted a training time of 20 minutes for a 2-second long video with a resolution of 128×128 . Such compute times are infeasible for dataset-scale experiments.

5 Hash Encodings for $\mathcal{A}$ - $\mathcal{B}$ Decomposition

In this section, we propose high-speed INR-based $\mathcal{A}$ - $\mathcal{B}$ decomposition. Specifically, to address the slower speeds of traditional INRs, we pivot to a faster framework and show design INRs for $\mathcal{A}$ - $\mathcal{B}$ decomposition within this framework.

Inspired by recent work in instant neural graphics primitives [32], we propose using a multiresolution hash input encoding (MRHE)-based framework for $\mathcal{A}$ - $\mathcal{B}$ decomposition. A critical aspect is using the spatial hash function for video fields. Spatial hash functions [34, 47] are typically 2-D and 3-D extensions to the prevalent hash tables. We use this principle to model the spatio-temporal coordinates (x, y, t) as a 3-D space. We posit that the 3-D spatial hash function can effectively model this space without violating any temporal constraints.

5.1 Cascaded Appearance Model $\hat{\mathcal{A}}(\mathbf{x}, t; \Theta_{\mathcal{A}})$

As discussed in Section 4.2, phase-based INRs [25], are efficient $\mathcal{A}$ -function representers. However, a direct translation of this concept to MRHEs is intractable, i.e., replacing the sinusoidal positional encodings with the faster MRHE makes phase incorporation infeasible. For sinusoidal encodings, phase modulations are akin to offsets. Thus, we propose to incorporate the time-to-phase conversion through learned spatiotemporal offset. The appearance model has two stages: the first maps spatio-temporal coordinates to positional offsets in the MRHE. The second is queried using the offset positions to generate video frames.

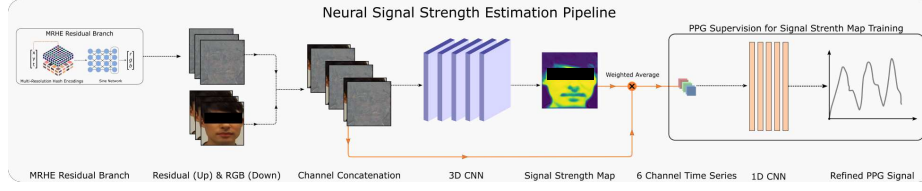


Fig. 5: Using the estimated $\mathcal{B}$ -function with the original video, we learn high-fidelity neural signal strength masks. The network takes the original RGB frames and the $\mathcal{B}$ -function estimate as inputs and returns a spatial strength mask. Training is supervised through an auxiliary 1-D CNN whose training target is the prediction of an accurate plethysmograph. The 1-D CNN is discarded post-training, and the learned mask model is used at inference time on the $\mathcal{B}$ -function to estimate rPPG.

Rather than considering a single offset vector in $\mathbb{R}^2$ for the entire frame (as [25] do for phase estimation), we estimate the offset, $\Delta \mathbf{x} \in \mathbb{R}^2$, for each point in the spatio-temporal grid. That is, $\Delta \mathbf{x} = f_{\mathcal{A}_1}(\mathbf{x}, t)$. The estimated offsets are added to the original spatial coordinates, which are then used to query a second model $f_{\mathcal{A}_2}(\mathbf{x} + \Delta \mathbf{x})$. Translating *Empirical Claim 1* from phase-based to offset-based models, we summarize our offset-adjustable cascaded appearance model as $\hat{\mathcal{A}}(\mathbf{x}, t; \Theta_{\mathcal{A}}) = f_{\mathcal{A}_2}(\mathbf{x} + f_{\mathcal{A}_1}(\mathbf{x}, t))$. The architecture and loss functions are elaborated on in the supplement. Figure 4 details this pipeline.

5.2 Residual Plethysmograph Model $\hat{\mathcal{B}}(\mathbf{x}, t; \Theta_{\mathcal{B}})$

We use a shallow sinusoidal model with MRHE, $f_{\mathcal{B}}(\cdot, \cdot)$, as our $\mathcal{B}$ -function estimator. In line with *Empirical Claim 3*, we train the residual plethysmograph network sequentially after the cascaded appearance model. While training, we freeze the cascaded appearance model and train the residual plethysmography network on the difference between the original video input and the output of the estimated appearance field. That is, $\hat{\mathcal{B}}(\mathbf{x}, t; \Theta_{\mathcal{B}}) = f_{\mathcal{B}}(\mathbf{x}, t)$.

The resulting residual plethysmograph model is shown to compensate for the cascaded appearance model’s inability to retain the plethysmograph signal. When queried, the model yields a spatio-temporal estimate of the plethysmograph signal. The supplement contains relevant implementation details.

6 rPPG Estimation

The previous section enables us to learn an estimator for $\mathcal{B}(\mathbf{x}, t)$, which is one part of the SINR optimality (and hence uncertainty minimization) from Equation 2. In this section, we proceed towards rPPG estimation from the $\mathcal{B}$ -function using high-fidelity neural signal masks.

Figure 5 highlights our proposed pipeline. Built on a 3-D CNN backbone, we indirectly supervise the spatial-attention-based network architecture with PPG waveforms. The model accepts the RGB input field concatenated with $\hat{\mathcal{B}}(\mathbf{x}, t)$ (a

combined 6-channel input) to estimate a mask. Here, $\mathcal{I}(\mathbf{x}, t) = [\mathcal{C}(\mathbf{x}, t), \hat{\mathcal{B}}(\mathbf{x}, t)]$, where $\mathcal{I}(\mathbf{x}, t)$ is the input to our SINR mask estimator. Limited by compute requirements, we relax the temporal component and only consider the first 64 frames (of the 300 frames) to compute a single mask for 10 seconds. End-to-end the network can be summarized as $SINR_{\mathcal{B}}(\mathbf{x}) = f_{SINR}(\mathcal{I}(\mathbf{x}, t); \Theta_{SINR})$.

It must be noted that while we only estimate a single mask for a 10-second video, our experiments show that our method is competitive even in the presence of natural motions like talking and head swaying (supplement). The SINR mask $SINR_{\mathcal{B}}(\mathbf{x})$ is directly used to compute $\mathbf{y}_1(t) = \frac{\sum_{\mathbf{x}} SINR_{\mathcal{B}}(\mathbf{x}) \cdot \mathcal{I}(\mathbf{x}, t)}{\sum_{\mathbf{x}} SINR_{\mathcal{B}}(\mathbf{x})}$, where $\mathbf{y}_1(t)$ is a time-series obtained from the 6-channel weighted spatial average of the input.

The 6-channel time-series signal is then passed through a 1-D CNN to yield a refined temporal estimate of the plethysmograph signal, $\hat{\mathbf{p}}(t)$. Note that this network is not used to generate the final estimates but to supervise the neural signal strength mask model indirectly. The whole pipeline is supervised end-to-end with the ground truth plethysmograph signals. Implementation details and further mathematical formulation are provided in the supplement. As a result of the architecture being based on the optimality constraint in Equation 2, the inferred masks are expected to be near SINR-optimal. This also means near optimality in terms of uncertainty minimization.

As in prior learning approaches, the neural regressor overfits the waveform, making it sub-optimal for OOD. Hence, we only retain $f_{SINR}(\cdot)$. The neural signal masks are combined with $\hat{\mathcal{B}}$ to estimate the rPPG signal. We follow [48] and utilize only the green channel. Advanced algorithms [11, 51] exist that utilize other channels, but they rely on the $\mathcal{A}$ -function [11, 51]. We perform detrending and Butterworth filtering for heart rate estimation as in [51].

7 Results

7.1 Experimental Setup

Datasets All methods are trained on two prior datasets [49, 53]. We also test our method on a self-collected Institutional Review Board (IRB) approved optically challenging OOD dataset. While we mainly focus on our dataset for OOD analysis, we also present results of cross-dataset inference. In our OOD dataset, we share full participant videos (without identifying information such as names), but only after recording contact details of dataset users. The dataset is released after potential users fill out a request form. Further details are in the supplement.

Evaluation Metrics We adopt the heart rate (HR) performance metrics from [49]. These include Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Pearson correlation coefficient (r) and the Threshold test (T-Test) [2, 31, 33]. Since our goal is OOD generalization, we propose “r-consistency” as the similarity metric between the Pearson coefficients of the in-distribution (ID) and OOD samples, calculated as a harmonic mean. A detailed discussion of this metric is present in the supplement.

Table 1: Performance on our OOD dataset considerably favors our method over prior work. We measure inter-distribution parity via the r-consistency metric. In-distribution values for r are given in Table 2. For algorithmic methods (non-learning), r-consistency on the datasets from [53] and [49] are shown in parenthesis. The best and second-best-performing numbers are highlighted in green and yellow, respectively.

Method	OOD Performance Metrics					Inter-Distribution Parity Metric
	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑	r-consistency ↑
Green [48]	31.54	7.80	9.84%	14.56	0.47	(0.54, 0.52)
POS [51]	31.92	8.02	11.06%	14.87	0.47	(0.60, 0.62)
CHROM [11]	36.73	9.32	12.62%	15.85	0.40	(0.53, 0.55)
ICA [41]	31.15	8.17	10.50%	15.32	0.41	(0.5, 0.51)
Train on [53]	DeepPhys [8]	78.27	18.64	23.19%	23.23	0.07
	TS-CAN [21]	77.12	18.46	23.24%	23.31	0.00
	EfficientPhys [22]	32.12	12.23	16.78%	23.40	0.31
	PhysNet [57]	33.65	8.64	11.09%	16.14	0.35
	PhysFormer [58]	48.46	10.41	13.26%	15.54	0.24
	Ours	18.08	5.10	6.81%	12.02	0.62
Train on [49]	DeepPhys [8]	77.88	17.94	22.44%	22.49	0.07
	TS-CAN [21]	75.58	17.65	22.02%	22.02	0.01
	EfficientPhys [22]	44.23	13.85	18.28%	22.84	0.19
	PhysNet [57]	48.46	11.24	14.27%	17.47	0.32
	PhysFormer [58]	39.81	9.58	12.59%	15.14	0.32
	Ours	16.15	4.61	6.07%	11.65	0.65

Evaluation Configuration We gauge performance of all methods via HR estimations from the power-spectral density. These estimations are carried over 300-sample windows (10 seconds), with a stride of 128 samples. Following [49], we validate⁴ the learning-based methods by performing six-fold cross-validation while averaging across the entire dataset for algorithmic methods. Similarly, we construct three folds for [53] to evaluate the learning methods. Please refer to the supplement for details, including visualization of signal strength masks.

Figures 6 and 7 show the results of our algorithm against the baseline methods. For ease of interpretability, we correct for random phase lag between the face (site of signal measurement) and finger (site of ground truth measurement).

7.2 Out of Distribution Plethysmography

We use our self-collected optically challenging dataset to benchmark OOD performance. While we primarily employ this dataset, we acknowledge that this is not the only notion of OOD. Hence, we also perform cross-dataset inference as part of the evaluation protocol to measure generalization capacity.

Optically Challenging OOD Evaluation Table 1 presents a quantitative analysis of OOD performance. Learning-based methods are trained on the datasets from [53] and [49], and the best fold is chosen for OOD inference.

Given their capacity to memorize distinguishable patterns, learning methods [57,58] handle input noise better than algorithmic methods for in-distribution

⁴ Baselines are configured, where possible, using the toolbox from [23].

Table 2: In-distribution and Cross-dataset OOD performance across two datasets - [49] and [53] shows comparable or superior performance compared to prior methods. Across in-distribution and cross-dataset validation, our method is the most consistent compared to SOTA methods. The best-performing and second-best-performing numbers are shown in **green** and **yellow**, respectively. There are 4 quadrants: 1^{st} - top left, 2^{nd} - top right, 3^{rd} - bottom right and 4^{th} - bottom left.

Method	Test on [53]					Test on [49]				
	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Green [48]	12.19	3.71	4.75%	10.27	0.64	Not Applicable				
POS [51]	6.19	2.03	2.84%	6.75	0.84					
CHROM [11]	9.66	2.74	3.81%	7.92	0.78					
ICA [41]	9.74	2.94	3.83%	9.15	0.71					
$T_{train on [53]}$ DeepPhys [8]	72.70	16.33	21.32%	20.43	0.02	76.26	18.21	22.79%	22.34	0.02
TS-CAN [21]	74.50	16.69	21.87%	20.69	0.01	77.17	18.28	23.17%	22.38	0.03
EfficientPhys [22]	2.26	1.02	1.59%	5.51	0.88	4.47	1.67	2.27%	5.90	0.89
PhysNet [57]	2.76	0.94	1.37%	4.69	0.92	3.58	1.19	1.59%	3.75	0.96
PhysFormer [58]	4.33	1.28	1.72%	5.03	0.90	25.08	6.50	8.58%	12.28	0.54
Ours	3.22	1.21	1.78%	5.61	0.88	4.39	1.66	2.17%	4.69	0.94
$T_{train on [49]}$ Green [48]	Not Applicable					19.58	6.27	7.53%	14.27	0.57
POS [51]						6.00	2.13	2.78%	6.54	0.89
CHROM [11]						8.07	2.38	3.07%	6.72	0.89
ICA [41]						17.15	5.13	6.00%	12.21	0.66
DeepPhys [8]	73.85	16.60	21.66%	20.55	0.00	77.43	18.46	23.26%	22.46	0.01
TS-CAN [21]	75.30	17.03	22.35%	20.96	0.03	77.36	18.61	23.49%	22.76	0.00
EfficientPhys [22]	11.45	4.44	6.43%	13.46	0.53	21.58	6.50	8.53%	14.08	0.53
PhysNet [57]	66.44	14.04	19.72%	17.81	0.06	4.72	1.48	1.87%	4.48	0.93
PhysFormer [58]	22.74	5.27	7.34%	9.93	0.62	2.96	1.06	1.37%	3.35	0.96
Ours	3.16	1.16	1.69%	5.33	0.89	3.34	1.22	1.61%	4.01	0.96

inference. However, this also leads to poorer generalizability on our OOD dataset, as emphasized by both the Pearson coefficient (r) and the “r-consistency” metrics. In contrast, algorithmic methods rely on deterministic heuristics. While this promotes good generalizability, it makes them extremely sensitive to factors such as lighting, occlusion, etc., leading to generally poorer performance.

Our proposed method maximizes signal strength while avoiding overfitting to the training distribution. We are, therefore, Pareto-optimal in both OOD performance and inter-distribution parity (Figure 2) regarding the T-Test value and “r-consistency” score. Our method is both the best and second-best performing method, outperforming Green (the best baseline) with 48.8% and 40.9% reduction in T-Test and MAE values. Clinically, this implies an almost 2x reduction in the number of participants on whom the method fails following the AAMI standard. Further granular analysis can be found in the supplement.

Cross-Dataset OOD and In-Distribution Evaluation Another notion of OOD is a cross-dataset generalization. Domain shifts between datasets result from differences in hardware and acquisition conditions. Learning-based methods tend to overfit the training data. Our experiments use datasets from [53] and [49] for their size and diversity. Table 2 indicates that while deep learning methods perform well in-distribution (1^{st} and 3^{rd} quadrant), they cannot maintain the same level of performance for cross-dataset inference (2^{nd} and 4^{th} quadrant). Algorithmic methods have no notion of cross-dataset evaluation. When trained

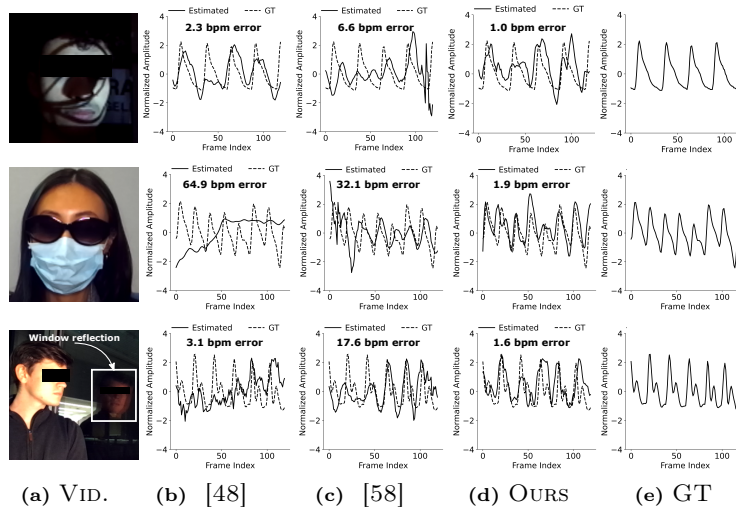


Fig. 6: Across challenging OOD optical settings, the proposed method can capture details of the plethysmograph waveform compared to prior methods. Results shown use models trained on the dataset proposed in [49], where applicable. Additional results are presented in the supplement.

on [49] and tested on [53] (4^{th} quadrant), we outperform all prior work. In the reverse scenario (2^{nd} quadrant), we are the second-best, marginally behind [57].

Despite OOD being our central focus, our model offers results comparable to SOTA algorithms even for in-distribution evaluation. Our method is within 0.3 bpm MAE compared to all top baselines (a difference that is not clinically significant). Additionally, our method shows a sub 0.5 bpm standard deviation of MAE across various folds used for evaluation in all configurations. Overall, our method is the most consistent across all quadrants while yielding errors within 2 bpm MAE, which most likely would be clinically accurate.

7.3 Qualitative Comparison

In Figure 6, we assess the qualitative waveforms generated on our optically challenging OOD dataset. We compare our method against the best-performing algorithmic and deep learning methods from Table 1, i.e., Green [48] and PhysFormer [58] respectively. For the representative samples, our method is superior in performance and can retain the shape of the plethysmography signal. The second scene, with the mask and glasses, especially highlights these gains. We also show superior rPPG estimation for extreme scenes (window reflections).

In Figure 7, we assess qualitative waveforms from in-distribution inference samples from [49]. We compare our method with PhysFormer [58] and POS [51] as the best-performing in-distribution deep learning and algorithmic methods. Our method retains both the frequency information and salient features of the waveform across diverse skin tones in the dataset.

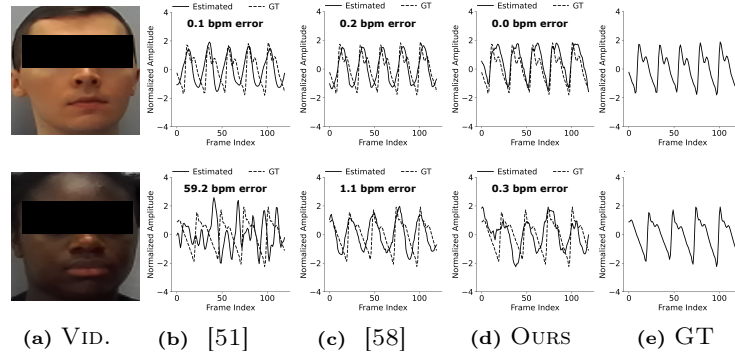


Fig. 7: In distribution inference on a diverse dataset. Our method better captures rPPG waveform details across skin tones for in-distribution evaluation on [49] dataset. While our method and [58] both perform reasonably well across skin tones, [58] does so with poorer OOD performance (Figure 6). More results in the supplement.

8 Discussion

We propose an INR for camera-based rPPG. We introduce the concept of $\mathcal{A}\mathcal{B}$ decomposition and propose an INR architecture to achieve this at scale. The proposed model is evaluated on a dataset of optically challenging OOD participants in addition to existing datasets. Our model performs best on the OOD dataset while comparable to SOTA for in-distribution validation.

8.1 Limitations

Being the first method using INRs for rPPG, limitations exist. First, while we reduce runtime over traditional INRs (as discussed in the supplement), and some rPPG methods report comparable runtimes [26], this is an avenue for future improvement. Second, while the proposed method is competitive for challenging scenes like talking and motion (supplement), these problems remain open.

8.2 Societal Impacts and Ethical Considerations

We focus on rPPG in optically challenging scenes (Section 1.1) due to factors like lighting, facial occlusions (masks, glasses, medical equipment), etc. Robustness to such settings enables applicability in settings like hospital rooms, classrooms, driving, etc. Additionally, prior work [1, 37] has shown promise in using rPPG for tasks like biometrics and forensics, and OOD-robustness will bring these applications one step closer to real-world deployment. Finally, these methods can also potentially be applied to other modalities and vital signs [10, 12, 60].

Technological solutions enabling healthcare for all, like this work, are relevant for social good. However, technology with clinical impact should only be deployed once the method is provably robust. Our method takes a step in that direction, but a pipeline of similar papers is needed to enable real-world rPPG deployment.

Acknowledgements

We thank the Visual Machines Group (VMG) at UCLA for feedback and support. A.K. was supported by a National Science Foundation (NSF) CAREER award, IIS-2046737, Army Young Investigator Program Award, Defense Advanced Research Projects Agency (DARPA) Young Faculty Award, and a Cisco Research Award. P.C. was partially supported by a Cisco Research Award.

References

1. Al Masri, A., Jasra, S.K.: The forensic biometric analysis of emotions from facial expressions, and physiological processes from the heart and skin. *Journal of Emerging Forensic Sciences Research* **1**(1), 61–77 (2016)
2. Association, C.T.: Physical activity monitoring for heart rate, ansi/cta-2065 (2018)
3. Ba, Y., Wang, Z., Karınca, K.D., Bozkurt, O.D., Kadambi, A.: Overcoming difficulty in obtaining dark-skinned subjects for remote-ppg by synthetic augmentation. *arXiv preprint arXiv:2106.06007* (2021)
4. Balakrishnan, G., Durand, F., Guttag, J.: Detecting pulse from head motions in video. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3430–3437 (2013)
5. Chari, P., Ba, Y., Athreya, S., Kadambi, A.: Mime: Minority inclusion for majority group enhancement of ai performance. In: *European conference on computer vision*. pp. 326–343. Springer (2022)
6. Chari, P., Kabra, K., Karınca, D., Lahiri, S., Srivastava, D., Kulkarni, K., Chen, T., Cannesson, M., Jalilian, L., Kadambi, A.: Diverse r-ppg: Camera-based heart rate estimation for diverse subject skin-tones and scenes. *arXiv preprint arXiv:2010.12769* (2020)
7. Chen, H., He, B., Wang, H., Ren, Y., Lim, S.N., Shrivastava, A.: Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems* **34**, 21557–21568 (2021)
8. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 349–365 (2018)
9. Chen, Z., Chen, Y., Liu, J., Xu, X., Goel, V., Wang, Z., Shi, H., Wang, X.: Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2047–2057 (2022)
10. Chen, Z., Zheng, T., Cai, C., Luo, J.: Movi-fi: Motion-robust vital signs waveform recovery via deep interpreted rf sensing. In: *Proceedings of the 27th annual international conference on mobile computing and networking*. pp. 392–405 (2021)
11. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering* **60**(10), 2878–2886 (2013)
12. Del Regno, K., Vilesov, A., Armouti, A., Harish, A.B., Can, S.E., Kita, A., Kadambi, A.: Thermal imaging and radar for remote sleep monitoring of breathing and apnea. *arXiv preprint arXiv:2407.11936* (2024)
13. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5712–5721 (2021)

14. Hurter, C., McDuff, D.: Cardiolens: remote physiological monitoring in a mixed reality environment. In: ACM siggraph 2017 emerging technologies, pp. 1–2 (2017)
15. Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., Funkhouser, T., et al.: Local implicit grid representations for 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6001–6010 (2020)
16. Kadambi, A.: Achieving fairness in medical devices. *Science* **372**(6537), 30–31 (2021)
17. Lee, E., Chen, E., Lee, C.Y.: Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In: European Conference on Computer Vision. pp. 392–409. Springer (2020)
18. Li, R., Tancik, M., Kanazawa, A.: Nerfacc: A general nerf acceleration toolbox. arXiv preprint arXiv:2210.04847 (2022)
19. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
20. Lindell, D.B., Van Veen, D., Park, J.J., Wetzstein, G.: Bacon: Band-limited coordinate networks for multiscale scene representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16252–16262 (2022)
21. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems* **33**, 19400–19411 (2020)
22. Liu, X., Hill, B., Jiang, Z., Patel, S., McDuff, D.: Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5008–5017 (January 2023)
23. Liu, X., Zhang, X., Narayanswamy, G., Zhang, Y., Wang, Y., Patel, S., McDuff, D.: Deep physiological sensing toolbox. arXiv preprint arXiv:2210.00716 (2022)
24. Magdalena Nowara, E., Marks, T.K., Mansour, H., Veeraraghavan, A.: Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1272–1281 (2018)
25. Mai, L., Liu, F.: Motion-adjustable neural implicit video representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10738–10747 (2022)
26. Maity, A.K., Wang, J., Sabharwal, A., Nayar, S.K.: Robustppg: camera-based robust heart rate estimation using motion cancellation. *Biomedical Optics Express* **13**(10), 5447–5467 (2022)
27. Martel, J.N., Lindell, D.B., Lin, C.Z., Chan, E.R., Monteiro, M., Wetzstein, G.: Acorn: Adaptive coordinate networks for neural scene representation. arXiv preprint arXiv:2105.02788 (2021)
28. Mehta, I., Gharbi, M., Barnes, C., Shechtman, E., Ramamoorthi, R., Chandraker, M.: Modulated periodic activations for generalizable local functional representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14214–14223 (2021)
29. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: Nerf in the dark: High dynamic range view synthesis from noisy raw images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16190–16199 (2022)

30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
31. Monitors, C.: Heart rate meters, and alarms. ANSI/AAMI Standard EC13 (2002)
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* **41**(4), 102:1–102:15 (Jul 2022). <https://doi.org/10.1145/3528223.3530127>, <https://doi.org/10.1145/3528223.3530127>
33. Nelson, B.W., Allen, N.B.: Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. *JMIR mHealth and uHealth* **7**(3), e10828 (2019)
34. Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)* **32**(6), 1–11 (2013)
35. Niu, X., Shan, S., Han, H., Chen, X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing* **29**, 2409–2423 (2019)
36. Nowara, E.M., McDuff, D., Veeraraghavan, A.: A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 284–285 (2020)
37. Nowara, E.M., Sabharwal, A., Veeraraghavan, A.: Ppgsecure: Biometric presentation attack detection using photoplethysmograms. In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. pp. 56–62. IEEE (2017)
38. Owthadi, H., Scovel, C., Sullivan, T.J., McKerns, M., Ortiz, M.: Optimal uncertainty quantification. *Siam Review* **55**(2), 271–345 (2013)
39. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 165–174 (2019)
40. Peters, H., Ba, Y., Kadambi, A.: pcon: Polarimetric coordinate networks for neural scene representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023)
41. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express* **18**(10), 10762–10774 (2010)
42. Ramaswamy, V.V., Kim, S.S., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9301–9310 (2021)
43. Schulz, P., Scheuvens, L., Fettweis, G.: A new perspective on maximal-ratio combining. In: *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. pp. 1–7. IEEE (2023)
44. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems* **33**, 7462–7473 (2020)
45. Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., Chen, X.: PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics* **25**(5), 1373–1384 (2021)

46. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems* **33**, 7537–7547 (2020)
47. Teschner, M., Heidelberger, B., Müller, M., Pomerantes, D., Gross, M.H.: Optimized spatial hashing for collision detection of deformable objects. In: *Vmv*. vol. 3, pp. 47–54 (2003)
48. Verkruyse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. *Optics express* **16**(26), 21434–21445 (2008)
49. Vilesov, A., Chari, P., Armouti, A., Harish, A.B., Kulkarni, K., Deoghare, A., Jalilian, L., Kadambi, A.: Blending camera and 77 ghz radar sensing for equitable, robust plethysmography. *ACM Transactions on Graphics (TOG)* **41**(4), 1–14 (2022)
50. Wadhwa, N., Rubinstein, M., Durand, F., Freeman, W.T.: Phase-based video motion processing. *ACM Transactions on Graphics (TOG)* **32**(4), 1–10 (2013)
51. Wang, W., Den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering* **64**(7), 1479–1491 (2016)
52. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8919–8928 (2020)
53. Wang, Z., Ba, Y., Chari, P., Bozkurt, O.D., Brown, G., Patwa, P., Vaddi, N., Jalilian, L., Kadambi, A.: Synthetic generation of face videos with plethysmograph physiology. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20587–20596 (2022)
54. Wang, Z., Zhou, S., Park, J.J., Paschalidou, D., You, S., Wetzstein, G., Guibas, L., Kadambi, A.: Alto: Alternating latent topologies for implicit 3d reconstruction. *arXiv preprint arXiv:2212.04096* (2022)
55. Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., Freeman, W.: Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)* **31**(4), 1–8 (2012)
56. Xu, T., White, J., Kalkan, S., Gunes, H.: Investigating bias and fairness in facial expression recognition. In: *Computer Vision—ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI* 16. pp. 506–523. Springer (2020)
57. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419* (2019)
58. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P.H., Zhao, G.: Physformer: facial video-based physiological measurement with temporal difference transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4186–4196 (2022)
59. Zhao, E.Q., Vilesov, A., Athreya, S., Chari, P., Merlos, J., Millett, K., Cyr, N.S., Jalilian, L., Kadambi, A.: Making thermal imaging more equitable and accurate: resolving solar loading biases. *arXiv preprint arXiv:2304.08832* (2023)
60. Zheng, T., Chen, Z., Zhang, S., Cai, C., Luo, J.: More-fi: Motion-robust and fine-grained respiration monitoring via deep-learning uwb radar. In: *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. pp. 111–124 (2021)
61. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15838–15847 (2021)