Title

On Computational Imaging in the Era of Neural Sensing: the Sensor, the Data and the Algorithm

Author

Chari, Pradyumna Venkatesh

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

On Computational Imaging in the Era of Neural Sensing: the Sensor, the Data and the Algorithm

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Pradyumna Venkatesh Chari

2024

ABSTRACT OF THE DISSERTATION

On Computational Imaging in the Era of Neural Sensing:
the Sensor, the Data and the Algorithm

by

Pradyumna Venkatesh Chari
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2024
Professor Achuta Kadambi, Chair

In recent years, sensing and perception techniques have evolved to be heavily reliant on learning-based pipelines. There is a specific need to explore computational imaging (joint design of hardware and software) in the era of AI. This work bridges this gap by understanding what we term as "neural sensing" through three pillars: the sensor, the data, and the learning algorithm. In the context of contactless heart rate monitoring of humans using visual sensors and beyond, we show that each of these three pillars pose specific, critical problems with the current state of the art: equity across demographic groups, lack of scalable, diverse data, and low signal to noise ratio in sensor measurements inhibiting accurate vital sign monitoring. We explore each pillar with the aim of addressing these limitations and demonstrate how a fundamental understanding and treatment of each of this pillars is critical towards building an operational perception systems. Through this thesis, we make contributions towards understanding the various pillars of neural sensing for and beyond contactless heart rate sensing, while also advancing the state of the art in remote plethysmography.

The dissertation of Pradyumna Venkatesh Chari is approved.

Bolei Zhou

Jonathan Chau-Yan Kao

Stefano Soatto

Achuta Kadambi, Committee Chair

University of California, Los Angeles

2024

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to begin by acknowledging my advisor Prof. Achuta Kadambi. It has been through his guidance, motivation, kind help and hours or conversation regardless of the time, that this thesis has taken the exciting shape and direction it currently holds and for that, I am thankful to him.

I would like to thank Prof. Stefano Soatto, Prof. Jonathan Chau-Yan Kao and Prof. Bolei Zhou as part of my thesis committee, for their support of this thesis. I am grateful for their time, help and insights in completing this work.

I would next like to thank Prof. Laleh Jalilian. Her guidance and help in setting together as daunting a task as collecting physiological data on diverse participants was critical in us achieving the healthcare applications explored in this thesis.

I would like to thank Alexander Vilesov, Adnan Armouti, and Anirudh Harish for all the additional hours of work that we have together put into the work in this thesis, and for their immense friendship. It has been my honor to have such good friends and colleagues.

I would like to thank Yunhao Ba, Chinmay Talegaonkar and Guangyuan Zhao for their support, help and friendship at the beginning of my PhD.

I would also like to thank all my other co-authors and colleagues on the projects included in this thesis. Additionally, all the members of the Visual Machines Group have played a significant role in my academic life and to all of them I am extremely thankful.

I would finally like to thank all my family, friends and colleagues, for their support throughout. To my wife, Kareena, thank you for your incredible support through the highs and the lows of the PhD and for always being there for me. I cannot thank you enough, but I know I don't have to. To my parents, Arundhati and Venkatesh, thanks for letting me do what I love, enabling me with all the opportunities that you have, and for your unconditional love and support. To my aunts, uncles and cousins, thanks for being my backbone and my support system. I am privileged to enjoy the love and care from all of you.

VITA

2015–2019	Bachelors of Technology (Electrical Engineering) Indian Institute of Technology, Madras.
2019–2021	M.S. (Electrical and Computer Engineering, UCLA)
2023	Computational Imaging Research Intern, Snap Inc.
2024	Research Intern, Vayu Robotics
2019–present	Ph.D. (Electrical and Computer Engineering, UCLA)

PUBLICATIONS

Vilesov, A.*, Chari, P.*, Armouti, A.*, Harish, A. B., Kulkarni, K., Deoghare, A., ... & Kadambi, A. (2022). Blending camera and 77 GHz radar sensing for equitable, robust plethysmography. ACM Trans. Graph., 41(4), 36-1.

Pei, S., Chari, P., Wang, X., Yang, X., Kadambi, A., & Zhang, Y. (2022, October). Forcesight: Non-contact force sensing with laser speckle imaging. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (pp. 1-11).

Wang, Z.*, Ba, Y.*, Chari, P., Bozkurt, O. D., Brown, G., Patwa, P., ... & Kadambi, A. (2022). Synthetic generation of face videos with plethysmograph physiology. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 20587-20596).

Chari, P., Ba, Y., Athreya, S., & Kadambi, A. (2022, October). Mime: Minority inclusion for majority group enhancement of ai performance. In European Conference on Computer Vision (pp. 326-343). Cham: Springer Nature Switzerland.

Zhou, S., Chang, H.*, Jiang, S.*, Fan, Z., Zhu, Z., Xu, D., Chari, P., ... & Kadambi, A. (2024). Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Chari, P.*, Harish, A.*, Armouti, A., Vilesov, A., Sarda, S., Jalilian, L., & Kadambi, A. (2024) Implicit Neural Models to Extract Heart Rate from Video. In European Conference on Computer Vision.

CHAPTER 1

Introduction

Computational imaging refers to the joint design of hardware and algorithms for visual tasks. Such control over both hardware and software has enabled several âĂIJsuperhumanâĂİ results in the past: depth from defocus [17], deblurring through coded exposure [18], seeing around corners [19], depth using polarization cues [20], and many more such results. In recent years, however, sensing and perception techniques have evolved to be heavily reliant on learning-based pipelines. With large models, trained on increasingly large datasets, the paradigm of what is practically possible with monocular RGB images is rapidly changing. Tasks like depth estimation [21] are increasingly moving away from the realm of computational imaging into the realm of large-scale computer vision. With this context, there is a specific need to explore, understand, and in certain cases redefine the scope of computational imaging in the era of AI. This thesis works towards understanding this new notion of computational imaging and AI, which we term as âĂIJneural sensingâĂİ, through the three crucial pillars of imaging in the era of AI: the sensor, the data, and the learning algorithm.

The **first pillar** is the design of task-specific sensors. With the context of patient monitoring tasks such as contactless heart rate (HR) estimation, we explore settings where careful consideration towards sensor design is critical, and where scale alone is not sufficient on its own: specifically, in the context of equity across skin tone groups. We find that naively designed RGB cameras can disadvantage darker skin tones. To alleviate this, we designed a novel sensor, that combined a camera and a radar to provide high-accuracy measurements while being skin tone equitable. Chapter 2 discusses this aspect in further detail. We also

explore other types of sensors. In Chapter 3, we explore the problem of detecting subtle touches on surfaces in the wild. This is an important and challenging problem towards realizing the potential of ubiquitous sensing and human computer interaction. Existing methods require either contact-based sensors, limiting usability, or camera-based methods, reducing accuracy. We propose a speckle-imaging based solution that is deployable in the wild.

Data at scale is arguably the most important component of a learning-based pipeline, and hence for neural sensing. However, collecting large datasets is a challenging task, especially while fairly representing demographic groups. In the **second pillar**, we discuss what we can do if obtaining minority group data samples at scale is infeasible, again in the context of equitable patient monitoring. For remote HR estimation, for instance, existing real datasets mostly contain light skin tone samples. Augmenting datasets to include dark skin tones is not as simple as changing the skin color: physically realistic color changes because of blood flow need to also be incorporated. We proposed such a method for synthetic, bio-realistic face video generation, leading to improved performance and fairness. Chapter 4 further discusses this aspect. We also explore solutions when it is not possible to acuqire minority samples, either in the form of real or synthetic data samples. In Chapter 5, both theoretically and experimentally, we make the surprising discovery that to maximize majority group classification accuracy, a majority-only training set is not optimal; having some minority samples is better. While the broader ML community agrees that minority inclusion in training sets benefits minority as well as overall performance, our observation enforces the benefit of minority inclusion for all test-time stakeholders. This means that naively creating equal-representation datasets may not be optimal for all stakeholders, motivating future research in optimal minority inclusion ratios.

The **third pillar** is the learning algorithm. In conjunction with the first pillar, we advanced the field of multimodal sensor fusion algorithms to also consider skin tone equity as a figure of merit for sensor fusion (discussed in Chapter 2). In conjunction with the second pillar, we developed rendering algorithms that consider the light transport principles

of human blood flow, while leveraging the benefits of learning-based methods (discussed in Chapter 4). Separate from these, we also explore more fundamental contributions to neural sensing and scene representations, in the context of healthcare. We show that implicit neural representations, which have shown great promise in tasks such as novel view synthesis, can also be used as selective function fitters. This has great relevance in extracting extremely low SNR signals from sensor measurements based on the physical properties of the said signal, such as in the case of camera-based heart rate estimation. This is discussed in Chapter 6. We also explore ways to leverage scene representations as a means of enabling 3D perception from 2D foundation models. In Chapter 7, we show that 3D representations such as Gaussian Splatting can be leveraged to distill features from 2D foundation models into 3D space. This enables 3D tasks such as 3D editing, 3D segmentation and so on from 2D models, which are easier to train and better performing.

Collectively, this body of work aims to advances all three pillars of neural sensing, while also taking a step towards equitable health sensing techniques for contactless monitoring of patients, and the development of health sensing systems while being constrained by sparse, biased training data. This thesis also makes contributions towards novel methods of sensing, as well as low-level computer vision.

1.1 Some Useful Definitions

A major portion of this thesis explores concepts of inequity (sometimes referred to as bias) across demographic (or other) groups. Please find below some definitions that help this understanding, which aim to be consistent with the International Vocabulary of Metrology [22].

True Value: The true value of a quantity is the oracle-known, actual value of a quantity. Note that the true value of a quantity cannot be known in practice.

Reference Value: The existing best measure of the quantity of interest that is currently available. In practice, 'ground truth' refers to this reference value (which may itself have

errors when compared with the true value).

Error: Let $\mathbf{p}$ be a subset of the global population $\mathbf{P}$, $\mathbf{p} \subseteq \mathbf{P}$. The error of a model $\mathbf{f}(\cdot)$ against a reference value $\mathbf{r}(\cdot)$ is given by,

$$e(\mathbf{p}) = \mathop{\mathbb{E}}_{s \in \mathbf{p}} \left[l \left(\mathbf{f}(s), \mathbf{r}(s) \right) \right],$$

where $l(\cdot, \cdot)$ is an appropriate distance metric.

Inequity: Let $\mathcal{G} = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_k\}$ be a set of exclusive and exhaustive partitions of the global population $\mathbf{P}$. Then, we define the inequity of an estimator $\mathbf{f}(\cdot)$ over the partition $\mathcal{G}$ as,

$$i(\mathcal{G}) = \max_{\mathbf{u}, \mathbf{v} \in \mathcal{G}} \left[|e(\mathbf{u}) - e(\mathbf{v})| \right].$$

CHAPTER 2

Building an Equitable Sensor for Remote Plethysmography

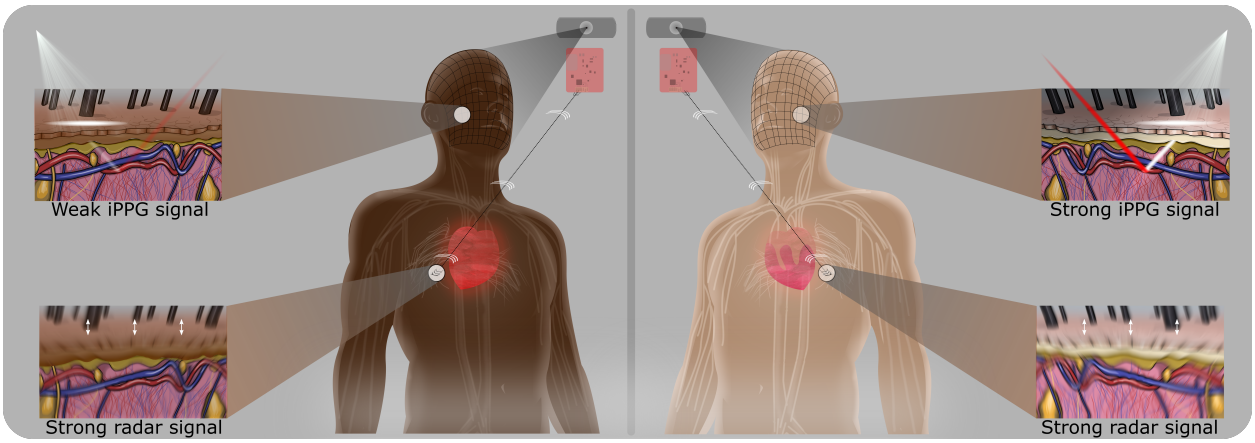


Figure 2.1: **The camera-based iPPG method, which uses subtle skin color changes as a function of blood flow to measure heart rate, shows inequity between (left) dark skin tones and (right) light skin tones. The radar-based method is more resistant to this skin tone inequity due to primarily observing chest motion.** Our proposed fusion method incorporates the complementary performance of the two methods and fairness properties of the radar to achieve performance and fairness gains over the iPPG method.

2.1 Introduction

Remote estimation of vital signs has gained growing relevance in recent years. The COVID-19 pandemic has further emphasized the need for rapid and reliable monitoring of health indicators. Remote plethysmography, the non-contact monitoring of blood volume flow in

the human body, is a critical technological step in this direction.

Camera-based remote plethysmography is a rapidly developing field. Most methods utilize small changes in the facial skin color as a function of dermal blood volume to capture pulse rate trends [23, 24]. Over the years, a broad range of methods have been proposed, ranging from physics-based approaches [25, 26, 27], blind source separation-based approaches [28, 29] and more recently, data-driven learning-based approaches [30, 31]. Through these advances, the heart-rate estimation performance has steadily approached levels of clinical accuracy. However, more recently, it has been established that most methods for imaging-based remote photoplethysmography (iPPG) are biased in performance against dark skin tone participants [32]. This points towards two potential problems: biases in datasets used for algorithmic evaluation, and potential fundamental biases in the physics of camera-based remote plethysmogaphy.

Another comparable modality for heart-rate estimation is the use of radio frequency (RF) devices. These devices capture the variations in chest displacement through cardiac cycles to estimate the frequency of the heart beat signal. Approaches use different types of radars and include signal processing [33, 34], as well as deep learning-based methods [35]. Ren et al. [36] performed a comparison study between a camera and a doppler stepped-frequency continuous wave radar on one subject and showed that both modalities perform nearly equally under ideal conditions for extracting heart rates. Additionally, it is noted from our experiments on multiple subjects that while iPPG and radar are comparable, iPPG slightly outperforms radar. However, since radar systems primarily capture displacement signals [37], radar techniques do not theoretically show performance correlation with skin tone and are therefore fairer.

In this work, we propose a fresh look at multi-modal fusion, from the perspective of inequity removal. We show that combining iPPG, an inequitable modality, with radar, an equitable modality, results in a better performing algorithm compared to the unimodal methods with only small trade-offs in fairness over the equitable method. That is, we show

that iPPG's Pareto frontier can be improved upon through carefully designed multi-modal fusion. To evaluate optimality, we establish a comprehensive set of performance and fairness metrics tailored to the task of remote plethysmography, evaluated on our novel multi-modal remote plethysmography dataset. This proposed fusion method, along with the existing iPPG and radar-based methods, constitute viable remote heart-rate detection approaches with differing performance and fairness trade-offs that an end-user may select from.

2.1.1 Contributions

The goal of this work is to use camera and 77 GHz radar fusion to create a higher performing and more equitable remote plethysmograph technique. We make three specific contributions:

Contribution A: Existing unimodal remote plethysmography methods show a Pareto tradeoff between performance and fairness. We show that through carefully chosen modalities, multi-modal fusion can improve the Pareto frontier for this tradeoff, enabling improvements in both performance and fairness.

Contribution B: To the best of our knowledge, we present the first RGB and radar plethysmograph multi-modal fusion technique incorporating inequity cues as part of a novel discriminative learning framework.

Contribution C: We open-source the first and largest multi-modal remote plethysmography dataset with representation across skin tones and other demographic markers.

Our code, dataset, and hardware tutorial may be accessed from https://github.com/UCLA-VMG/EquiPleth.

2.1.2 Scope

This work aims to establish the importance of multi-modal fusion towards achieving high performing and fair algorithms for vital sign sensing. We do not consider or incorporate

other confounding effects, such as motion, resolution, and compression. These are relevant engineering aspects that need to be considered when looking at deployability of the technology. In this work, however, we constrain our focus on the analysis and mitigation of skin tone inequity (as opposed to other kinds of biases).

2.2 Related Work

Image-photoplethysmography is biased against darker skin tones. To improve fairness and performance of unimodal iPPG, we fuse it with another sensing modality (radar). In what follows, we expand on background context.

Image Photoplethysmography Heart-rate estimation using image-photoplethysmography (iPPG) has been actively studied since the early 2000s [38, 39, 40]. Typically, early methods observed and took advantage of changes in optical absorption of hemoglobin molecules at the surface of the skin during a blood volume pulse with a RGB camera. The work that followed focused on reducing error due to motion with region of interest (ROI) alignment and clever modeling of physical properties of light reflectance [27, 26]. Remote heart-rate estimation is also achieved with Ballistocardiogram (BCG) methods which extract motion information due to the Newtonian reaction of a blood volume pulse [41]. Color analysis and BCG 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 spectrum [42], despite having a worse signal to noise ratio (SNR) to RGB cameras [43]. Infrared (IR) or thermal imaging has used BCG [44] and temporal temperature differentials [45]. Other work focused on visualizations of the blood volume pulse with Eulerian magnification [23] and augmented reality [46].

More recently, deep learning approaches have been utilized to attain state of the art results. [30] used an attention-based Convolutional Neural Network (CNN) to explicitly

fuse skin-reflection and motion information. [31] introduced spatio-temporal CNNs to iPPG to enable temporal context-aware networks. Other work has extended these architectures [47], incorporated meta-learning [48], improved PPG waveform characteristics [49], and augmented iPPG datasets with synthetic examples [50, 1]. Our work builds on previous iPPG work through multi-modal fusion with Frequency Modulated Continuous Wave (FMCW) radar to reduce inequity across skin tones while improving performance.

Radar Plethysmography Vital sensing using radar was pioneered in the 1970s for respiratory-rate detection [51]. Today, radar research has diverse applications in respiratory-rate, heart-rate, and blood-pressure detection. For heart-rate estimation, various hardware setups are used, including FMCW [33], Ultra Wide Band (UWB) Impulse [52], and Continuous Wave Doppler radars [53]. Vital sign detection is performed by observing millimeter (mm) level displacements in the chest. The average adult has chest displacements for breathing and heart pulses of 1-12 mm and 0.01-0.5 mm, respectively [54]. Through single-subject analysis, [36] note that both camera and radar-based methods perform nearly equally under ideal conditions. That is, both are potentially viable methods for remote heart-rate estimation. Since heart-rate detection is more prone to noise, applications and experimental results are often done with subjects laying down to avoid interference due to motion. In contrast, our work assumes that the participant is sitting. Our work also presents a deep learning method for learning plethysmograph signals using FMCW radar.

Fairness in iPPG Fairness in machine learning has been a rapidly growing area of research in the last decade. It has spanned ensuring fairness in classification [55], word embeddings [56], and computer vision [57]. Dataset inequity is a common problem and it has been shown that performance is lower for women and darker skin tones in machine learning problems due to underrepresentation [58]. In the iPPG field, fairness has been less studied. However, [32, 1] show that dataset inequity as well as lower SNRs result in darker skin tones producing poorer performance than lighter skin tones. In this work, we introduce a skin tone representative

dataset and propose to reduce inequity across skin tones through multi-modal fusion.

Multi-modal Fusion Multi-modal fusion is the process of combining two or more modalities to achieve better performance for a given task than any singular modality on its own. In deep learning, architectures either fuse modalities in a middle latent space or at a late stage once each modality independently gives a prediction. For mid-level fusion, Restricted Boltzmann machines [59] are a common choice in popular architectures such as Deep Belief Networks [60] and Stacked Autoencoders [61]. In late-level or decision-level fusion, predictions from various modalities are simply aggregated using majority voting, weighted voting, or a meta-classifier. These architectures and formulations of multi-modal fusion have achieved great success in classification-based problems [62]. Unfortunately, they do not easily translate to a regression-based problem such as plethysmography. Nonetheless, several works have attempted fusion such as RGB+Mid-Infrared (Thermal) [63] and RGB+Near-Infrared (NIR) [64]. Our work proposes the first machine-learning-based fusion of RGB+radar modalities in order to boost performance and reduce inequity across skin tones.

2.3 Problem Formulation

The goal of this work is to fairly estimate a plethysmograph signal of blood volume pulses from non-contact sensing data. In particular, a human subject is sensed by a non-contact sensing method, f, that processes data from M modalities (e.g. images, radar matrices, etc.). For the i-th training example, let this data be represented as an irregular list $\mathbf{x}_i = [\mathbf{v}_i^1, \mathbf{v}_i^2, ..., \mathbf{v}_i^M]$, with each $\mathbf{v}_i^m \in \mathbb{R}^{N_m}$. This subject has a ground truth and predicted plethysmograph signal, denoted as vectors $\mathbf{y}_i \in \mathbb{R}^K$ and $f(\mathbf{x}_i) = \hat{\mathbf{y}}_i \in \mathbb{R}^K$, respectively. The corresponding heart rate of the plethysmograph signal is $\mathbf{h}_i \in \mathbb{R}^+$. The sensor data, plethysmograph signal, and heart rate are drawn from distributions $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{H}$ respectively. Unique to this work's formulation (as compared to others in iPPG literature), the subject

also has a protected attribute $\mathbf{a}_i \in A$. This attribute describes skin tone categories, such that $A = \{\text{light}, \text{medium}, \text{dark}\}$. Subjects are labeled according to a modified Fitzpatrick skin tone scale [1] as light for I/II, medium for III/IV, and dark for V/VI. For brevity, the sample indexing is dropped here onward; $\mathbf{x}$, $\mathbf{y}$, $\mathbf{h}$, and $\mathbf{a}$ denote $\mathbf{x}_i$, $\mathbf{y}_i$, $\mathbf{h_i}$ and $\mathbf{a_i}$ respectively.

2.3.1 Performance

To assess general performance, we use heart rate prediction metrics as in previous iPPG work [31, 47]. Table 2.1 summarizes metrics for heart-rate accuracy. Several results are reported using the absolute percent error (APE) metric due to the Association for Advancement of Medical Instruments (AAMI) defining a threshold heart-rate error as no greater than 10% relative error [65], as well as its prevalence in evaluating heart monitors and physical monitoring devices [66, 67]. We note that all performance measures are evaluated over the entire testing dataset, consisting of samples from the light, medium and dark skin tone groups.

2.3.2 Fairness

To evaluate fairness, we adopt standard metrics from the fairness community [68]. A plethysmograph method can be considered fair if the outcome for the general population is the same as for a sub-population with a given attribute, $\mathbf{a}$. In our case, to facilitate experimental analysis, we evaluate fairness in terms of similarity of outcome between the light and dark skin tone groups. The following definitions introduce metrics for quantifying fairness and are summarized in Table 2.1.

2.3.2.1 Threshold Test

The threshold test [69] is a notion of sufficiency of fair performance. The test shows the proportion of test samples in the light (L) and dark (D) categories that fall outside of the

Table 2.1: **We use multiple heart-rate performance and fairness metrics for evaluation.** Performance and fairness are not necessarily correlated properties. A complete analysis of algorithms requires a two dimensional comparison.

	Metric	Expression
Performance	Mean Absolute Error (MAE) ↓	$\frac{1}{N}\sum_{i=1}^{N}\lvert h_i - \hat{h}_i\rvert$
	Root Mean Square Error (RMSE) ↓	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(h_i - \hat{h}_i)^2}$
	Mean Absolute Error (MAPE) ↓	$\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert h_i - \hat{h}_i\rvert}{h_i}$
	Pearson Corr. Coefficient (R) ↑	$\frac{\sum_{i=1}^{N}(\hat{h}_i - \mu_{\hat{h}})(h_i - \mu_h)}{\sum_{i=1}^{N}(\hat{h}_i - \mu_{\hat{h}})^2 \sum_{i=1}^{N}(h_i - \mu_h)2}$
Fairness	Threshold Test ↓	$\mathbb{P}_{\mathbf{X}}[APE(\hat{\mathbf{H}}) > 10\% \mid \mathbf{a}]\ ,\forall\ \mathbf{a} \in \{\mathrm{L}, \mathrm{D}\}$
	Performance inequity ↓	$\lvert D_{\mathbf{a}=\mathrm{D}}(\mathbf{H}, \hat{\mathbf{H}}) - D_{\mathbf{a}=\mathrm{L}}(\mathbf{H}, \hat{\mathbf{H}})\rvert$

threshold defined by AAMI.

2.3.2.2 Performance Inequity

In order to measure inequity through performance metrics, we follow [70] by evaluating a performance metric, D on attribute groups and taking a pairwise difference.

2.4 Plethysmography and Skin Tone Inequity

This section is composed of two parts. *First*, we establish that *iPPG performance is fundamentally inequitable* as a function of skin tone. *Second*, we show that fusing a better performing inequitable modality with a worse performing equitable modality leads to *improvements in both overall performance and inequity* over the inequitable modality.

Table 2.2: **Notation used for light transport modeling of iPPG.** The left column shows the notation used and the right column describes the notation.

Notation	Description
$\mu_{a,eum}(\lambda)$	Eumelanin absorption coefficient
$\mu_{a,phm}(\lambda)$	Phomelanin absorption coefficient
$\mu_{a,der}(\lambda)$	Dermal absorption coefficient
$\mu_{s,der}(\lambda)$	Dermal scattering coefficient
$\mu_{a,bld}(\lambda)$	Blood absorption coefficient
$\mu_{a,ski}(\lambda)$	Skin absorption coefficient
$\mu_{oxy}(\lambda)$	Oxygenated hemoglobin abs. coefficient
$\mu_{dox}(\lambda)$	Deoxygenated hemoglobin abs. coefficient
f_{mel}	Skin melanin fraction
f_{eum}	Epidermal eumelanin fraction
f_{bld}	Dermal blood volume fraction
f_{oxy}	Oxygenated hemoglobin fraction in blood

2.4.1 Motivation: iPPG has Skin Tone Inequity

We utilize existing biorealistic graphical rendering models [71, 72] and extend them to iPPG. A two layer skin model is assumed. The incident light undergoes attenuation while passing through the epidermis, while it undergoes scattering driven reflection at the dermis. Table 2.2 describes and summarizes the various symbols and notations used.

We start with describing the epidermal transmission. Following the Beer-Lambert Law,

$$T_{epi}(\lambda) = e^{-\mu_{a,epi}(\lambda)}, \tag{2.1}$$

where $\mu_{a,epi}(\lambda)$ is the absorption coefficient of the epidermis. Typically, this is modeled as a convex combination of skin tissue and melanin absorption,

$$\mu_{a,epi}(\lambda) = f_{mel}\mu_{a,mel}(\lambda) + (1 - f_{mel})\mu_{a,skin}(\lambda). \tag{2.2}$$

$\mu_{a,skin}(\lambda)$, the skin tissue absorption coefficient, is a biological parameter which is known. $\mu_{a,mel}(\lambda)$ may be defined as,

$$\mu_{a,mel}(\lambda) = f_{eum}\mu_{a,eum}(\lambda) + (1 - f_{eum})\mu_{a,phm}(\lambda), \tag{2.3}$$

where $\mu_{a,eum}(\lambda)$ is the absorption coefficient of eumelanin and $\mu_{a,phm}(\lambda)$ is the absorption coefficient of pheomelanin (both known biophysical parameters). By combining Equations 2.1, 2.2, and 2.3, the epidermal transmission may be accurately modeled.

We move towards describing the dermal reflection. This model follows the Kubelka-Munk theory for scattering-dependent reflection. Specifically, the fraction of reflected light, as a function of wavelength, is given by,

$$R_d(\lambda) = \frac{(1 - \beta(\lambda))^2(e^{K(\lambda)d_{der}} - e^{-K(\lambda)d_{der}})}{(1 + \beta(\lambda))^2 e^{K(\lambda)d_{der}} - (1 - \beta(\lambda))^2 e^{-K(\lambda)d_{der}}}. \tag{2.4}$$

Here, d_{der} is the dermal skin depth [72]. Also, $\beta(\lambda)$ and $K(\lambda)$ are deterministically related to $\mu_{a,der}(\lambda)$ (dermal absorption coefficient) and $\mu_{s,der}(\lambda)$ (reduced dermal scattering coefficient [73]), as given in [71, 72]. The dermal absorption coefficient and the blood absorption coefficient are understood as convex combinations shown below:

$$\mu_{a,der}(\lambda) = f_{bld}\mu_{a,bld}(\lambda) + (1 - f_{bld})\mu_{a,ski}(\lambda), \tag{2.5}$$

$$\mu_{a,bld}(\lambda) = f_{oxy}\mu_{oxy}(\lambda) + (1 - f_{oxy})\mu_{dox}(\lambda). \tag{2.6}$$

Here, various factors include blood reflection, skin baseline reflection, oxygenated blood reflection and deoxygenated blood reflection respectively. Given the expressions for epidermal transmission and dermal reflection, the expression for overall reflection is given by,

$$R(\lambda) = T^2{}_{epi}.R_d(\lambda). \tag{2.7}$$

Then, the overall intensity captured in channel $\mathbf{c}$ of the camera is given by,

$$I_c = \int_\lambda E(\lambda)S_c(\lambda)R(\lambda)d\lambda, \tag{2.8}$$

where $E(\lambda)$ is the source spectral distribution and $S_c(\lambda)$ is the camera spectral response for channel c.

iPPG Signal Strength The iPPG signal arises out of a variation in the blood volume fraction, f_{bld} under the skin. Our interest is in the signal strength across camera channels, Σ_c, which can be defined as *the maximum variation in the captured intensity* (proportional to signal amplitude). Mathematically,

$$\Sigma_c = \Delta I_c \approx \left| \frac{\partial I_c}{\partial f_{bld}} \right| \cdot \Delta f_{bld}. \tag{2.9}$$

Since $R(\lambda)$ is the only term dependent on f_{bl},

$$\Sigma_c \approx \left| \int_\lambda E(\lambda) S_c(\lambda) \frac{\partial R}{\partial f_{bld}} \Big|_{\overline{f_{bld}}} d\lambda \right| \cdot \Delta f_{bld}, \tag{2.10}$$

where $\overline{f_{bld}}$ is the average blood volume fraction, typically around 0.05. This approximation holds true since f_{bld} only varies by a small amount, typically around 0.05.

This plethysmographic signal variation occurs in addition to the average skin tone color, given by

$$\Gamma_c = \int_\lambda E(\lambda) S_c(\lambda) R(\lambda) \Big|_{\overline{f_{bl}}} d\lambda. \tag{2.11}$$

Since Σ_c and Γ_c are both dependent on f_{mel}, we refer to these as $\Sigma(f_{mel})$ and $\Gamma(f_{mel})$ subsequently.

Effect of Imaging Noise on iPPG Imaging noise refers to the inherent noise that arises due to the image capture process in a commercial camera. This arises due to various effects related to photon arrival processes, thermal noise in electronics and the quantization noise associated with digitally capturing images [74]. For pixels below the saturation level, the noise can be modeled as follows:

$$\sigma^2_{pixel} = \frac{\Phi t}{g^2} + \frac{\sigma_r^2}{g^2} + \sigma_q^2, \tag{2.12}$$

where Φ is the radiant power of light collect, t is the exposure time, g is the sensor gain (a constant for a given image), and σ_r and σ_q are camera noise parameters (also constant).

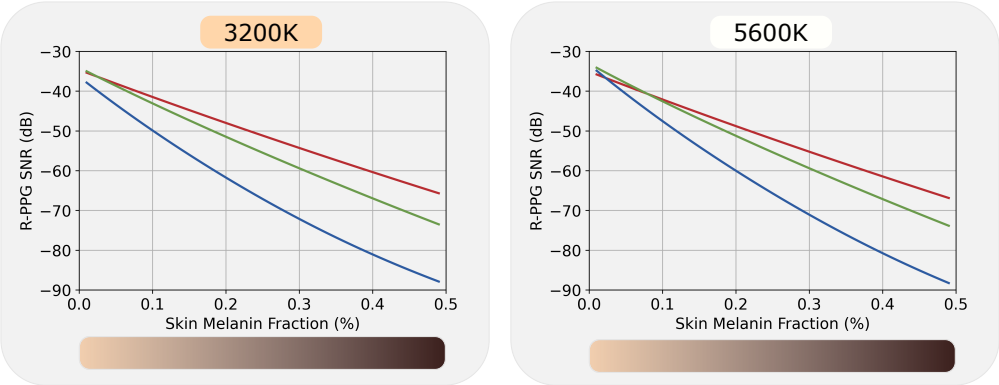


Figure 2.2: **The iPPG per-pixel SNR drastically worsens with increasing skin melanin fraction.** Using biophysical skin reflectance models, we estimate the iPPG signal strength as a function of skin melanin fraction. Along with the monotonic decrease in signal strength with melanin fraction, we also note finer trend differences between the SNR for the red, green and blue channels, as well as dependence on the spectral properties of the light source.

Using this noise model, we can the estimate the iPPG signal to noise ratio (SNR) for a pixel of a particular intensity and color channel c as follows:

$$SNR_c = \frac{\Sigma_c t}{g\sqrt{\frac{\Gamma_c t}{g^2} + \frac{\sigma_r^2}{g^2} + \sigma_q^2}}. \tag{2.13}$$

Here, we assume that the radiant power of light collected Φ is equal to the average skin tone color.

Figure 2.2 shows the iPPG per-pixel SNR plots for the three camera color channels, across two lighting conditions (indicated by the light source 'temperature'). We use average camera response functions $\mathbf{S_c}(\boldsymbol{\lambda})$ to identify responsiveness of each of the channels to incident light, as well as exemplar camera noise parameters. Specifically, we used $\sigma_r = 140.7$, $\sigma_q = 0.08$ and $g = 1.06$. These parameters are representative and are calculated for a typical cell phone camera. Their specific values do not affect the trends and hence are not of primary importance. We note that the SNR monotonically decays with increasing skin melanin fraction. This trend is consistent across color channels and scene lighting conditions. That

is, given a fixed scene and camera configuration, the underlying physical signal is poorer for dark skin tones.

2.4.2 Resisting inequity through Sensor Fusion, a Proof

In this section, we prove that sensor fusion, even if some of the individual sensors are inequitable, can lead to overall resistance of inequity. Consider a temporal signal $\mathbf{s} \in \mathbb{R}^n$. Without loss of generality, we assume normalized signals such that $\|\mathbf{s}\|_2 = 1$. Our sensing setup consists of modalities $m \in \{M_1, M_2\}$. Each modality m captures a noisy observation of the signal $\mathbf{y}_m \in \mathbb{R}^n$.

An additional property of a sample is its attribute $\mathbf{a} \in \{\text{light}, \text{dark}\}$. The observed signal is then given by $\mathbf{y}_m^{\mathbf{a}}$. Both modalities behave differently for different attributes. For example, a modality may have different expected performance for samples with different attributes.

Let the theoretical Signal to Noise Ratio (SNR) of a signal $\mathbf{y}_m^{\mathbf{a}}$ be denoted by the operator $\mathcal{S}(\mathbf{y}_m^{\mathbf{a}})$. We use the SNR to define modality-wise fundamental performance and fairness. Without loss of generality, we assume that the modality M_1 is better performing as compared to the modality M_2 (in our practical setting, M_1 would correspond to the RGB modality while M_2 would correspond to the radar modality). That is,

$$\mathop{\mathbb{E}}_{\mathbf{a},\, \mathbf{y}_{M_1}^{\mathbf{a}}} \left[\mathcal{S}(\mathbf{y}_{M_1}^{\mathbf{a}})\right] > \mathop{\mathbb{E}}_{\mathbf{a},\, \mathbf{y}_{M_2}^{\mathbf{a}}} \left[\mathcal{S}(\mathbf{y}_{M_2}^{\mathbf{a}})\right]. \tag{2.14}$$

Additionally, we note that according to our required conditions, the modality M_1 is inequitable in terms of attribute $\mathbf{a}$, while the modality M_2 is equitable. That is,

$$\begin{aligned} &\left| \mathop{\mathbb{E}}_{\mathbf{y}_{M_1}^{\text{Light}}} \left[\mathcal{S}(\mathbf{y}_{M_1}^{\text{Light}})\right] - \mathop{\mathbb{E}}_{\mathbf{y}_{M_1}^{\text{Dark}}} \left[\mathcal{S}(\mathbf{y}_{M_1}^{\text{Dark}})\right] \right| > \epsilon, \text{ and} \\ &\left| \mathop{\mathbb{E}}_{\mathbf{y}_{M_2}^{\text{Light}}} \left[\mathcal{S}(\mathbf{y}_{M_2}^{\text{Light}})\right] - \mathop{\mathbb{E}}_{\mathbf{y}_{M_2}^{\text{Dark}}} \left[\mathcal{S}(\mathbf{y}_{M_2}^{\text{Dark}})\right] \right| < \epsilon, \end{aligned} \tag{2.15}$$

for some suitable small ϵ. We also assume without loss of generality that the 'dark' attribute

is the worse performing attribute group on average. That is,

$$\mathop{\mathbb{E}}_{\mathbf{y}_m^{\text{Light}}}\left[\mathcal{S}(\mathbf{y}_m^{\text{Light}})\right] \geq \mathop{\mathbb{E}}_{\mathbf{y}_m^{\text{Dark}}}\left[\mathcal{S}(\mathbf{y}_m^{\text{Dark}})\right], \forall\, m \in \{M_1, M_2\}. \tag{2.16}$$

We wish to characterize the improvement in signal quality arising as a result of combining observations from the two modalities M_1 and M_2. That is, we wish to understand the performance and inequity properties of a combined measurement signal $\mathbf{y}_{\text{comb}}^{\mathbf{a}}$ that is optimal in the SNR sense, as follows:

$$\begin{aligned} \mathcal{C}^* &= \arg\max_{\mathcal{C}} \mathbb{E}\left[\mathcal{S}(\mathcal{C}(\mathbf{y}_{M_1}^{\mathbf{a}}, \mathbf{y}_{M_2}^{\mathbf{a}}))\right], \forall\, a. \\ \mathbf{y}_{\text{comb}}^{\mathbf{a}} &= \mathcal{C}^*(\mathbf{y}_{M_1}^{\mathbf{a}}, \mathbf{y}_{M_2}^{\mathbf{a}}), \forall\, \mathbf{a}. \end{aligned} \tag{2.17}$$

Here, $\mathcal{C}(\cdot, \cdot)$ is an appropriately chosen combining operator.

We wish to quantify the benefit of multi-modal combination in terms of the gains obtained over the modality M_1. The quality gain factor $Q_{\mathbf{a}}$ is therefore given by,

$$Q_{\mathbf{a}} = \frac{\mathbb{E}\left[\mathcal{S}(\mathbf{y}_{\text{comb}}^{\mathbf{a}})\right]}{\mathbb{E}\left[\mathcal{S}(\mathbf{y}_{M_1}^{\mathbf{a}})\right]}. \tag{2.18}$$

We have established the required terminology for our result.

Theorem 1: Let Q_{Light} and Q_{Dark} be the quality gain factors for the light and dark attributes respectively. Then, optimally combining observations $\mathbf{y}_{M_1}^{\mathbf{a}}$ from a better performing but inequitable modality M_1 and $\mathbf{y}_{M_2}^{\mathbf{a}}$ from a worse performing but equitable modality M_2, where $\mathbf{a} \in \{\text{light}, \text{dark}\}$, ensures,

$$Q_{\text{Dark}} > Q_{\text{Light}}. \tag{2.19}$$

That is, the worse performing attribute sees a greater relative gain in signal strength.

Proof: *The optimal combination method for two signals with known SNRs is the Maximal Ratio Combining [75]. The resulting expected SNR is given by,*

$$\mathbb{E}\left[\mathcal{S}(\mathbf{y}_{comb}^{\mathbf{a}})\right] = \mathbb{E}\left[\mathcal{S}(\mathbf{y}_{M_1}^{\mathbf{a}})\right] + \mathbb{E}\left[\mathcal{S}(\mathbf{y}_{M_2}^{\mathbf{a}})\right], \forall\, \mathbf{a}. \tag{2.20}$$

Then, the quality gain factor is given by,

$$\begin{aligned} Q_{\mathbf{a}} &= \frac{\mathbb{E}\left[\mathcal{S}(\mathbf{y}^{\mathbf{a}}_{M_1})\right] + \mathbb{E}\left[\mathcal{S}(\mathbf{y}^{\mathbf{a}}_{M_2})\right]}{\mathbb{E}\left[\mathcal{S}(\mathbf{y}^{\mathbf{a}}_{M_1})\right]} \\ &= 1 + \frac{\mathbb{E}\left[\mathcal{S}(\mathbf{y}^{\mathbf{a}}_{M_2})\right]}{\mathbb{E}\left[\mathcal{S}(\mathbf{y}^{\mathbf{a}}_{M_1})\right]}. \end{aligned} \tag{2.21}$$

Then, from Equations 2.15 and 2.21, we can infer that,

$$Q_{Dark} > Q_{Light}. \tag{2.22}$$

This completes the proof. ■

We therefore establish that fusion with a worse performing but less inequitable modality is in fact beneficial in terms of inequity mitigation.

2.4.3 Overall Inferences

We now summarize the inferences from the theory section, which serve as a motivation for our multi-modal fusion hypothesis for inequity alleviation.

1. The RGB modality is inequitable against darker skin tone samples. This arises fundamentally as a result of poor Signal to Noise Ratio.

2. Combining the RGB modality with a poorer but equitable modality results in larger improvements for the darker skin tone samples as compared to the lighter skin tone samples.

2.5 Implementation of Fusing RGB Camera and Radar for Plethysmography

In the previous section, we discussed inequity in the context of plethysmography. We now discuss the specific fusion of camera and 77 Ghz radar to resist inequity. Referring to notation

from Section 2.3, let $\mathbf{v}^1$ denote an RGB modality measurement, and let $\mathbf{v}^2$ denote a FMCW radar modality measurement. Then, $\mathbf{x} = [\mathbf{v}^1, \mathbf{v}^2]$, our overall measurement. Our goal is to learn a robust functional mapping f from the measurement space to the plethysmograph space,

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}) \\ &= f([\mathbf{v}^1, \mathbf{v}^2]). \end{aligned} \tag{2.23}$$

Figure 2.5 describes our overall pipeline. We use a late fusion parameterization for the mapping f. The unimodal plethysmography signal estimates are first obtained. A fusion architecture combines these to obtain the final fused plethysmography estimate. That is,

$$\mathbf{y} = f([\mathbf{v}^1, \mathbf{v}^2]) = g_f(g_1(\mathbf{v}^1), g_2(\mathbf{v}^2)), \tag{2.24}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are the modality specific estimators for RGB and radar respectively. The function $g_f(\cdot)$ is the late fusion model. We describe each of these components in the following text.

2.5.1 RGB Camera

The RGB measurements $\mathbf{v}^1 \in [0,1]^{T \times C \times H \times W}$ are tensor-valued. Here, T is the numbers of frames, C is the number of image channels (3), and H and W are the image height and width respectively. We use the Physnet spatio-temporal network by [31] as one of the inputs for fusion. The PPG network $g_1(\cdot)$ estimates the PPG waveform $\hat{\mathbf{y}}_{RGB}$ from video inputs $\mathbf{v}^1$ with $T = 64$ frame samples as input.

The PPG network is updated using a negative Pearson loss between the estimated waveform $\hat{\mathbf{y}}$ and the ground truth waveform $\mathbf{y}$ to enforce waveform reconstruction, similar to previous work [31]. This is given by,

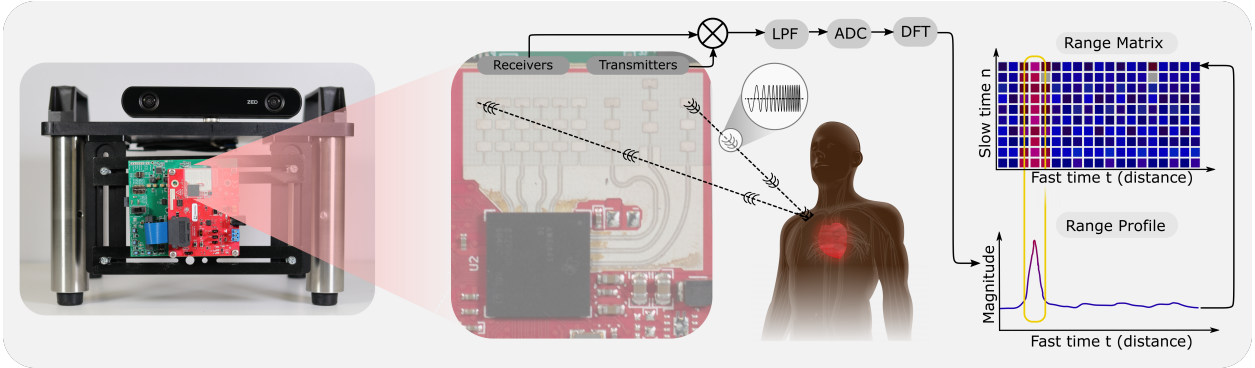


Figure 2.3: **We use a 77 Ghz FMCW radar setup for non-contact radar plethysmography.** Chirp signals are bounced off the subject's chest in order to capture subtle motion. By exploiting the dependency of the phase on the distance of flight, we are able to measure this motion.

$$
\begin{aligned}
L_P(\mathbf{y}, \hat{\mathbf{y}}_{RGB}) =& 1 - \frac{1}{\sqrt{a_1 \times a_2}} \left(N \sum_{i=1}^{N} \mathbf{y}_i \hat{\mathbf{y}}_i - \sum_{i=1}^{N} \mathbf{y}_i \sum_{i=1}^{N} \hat{\mathbf{y}}_i \right), \\
a_1 =& \left(N \sum_{i=1}^{N} \mathbf{y}_i^2 - \left(\sum_{i=1}^{N} \mathbf{y}_i \right)^2 \right) \\
a_2 =& \left(N \sum_{i=1}^{N} (\hat{\mathbf{y}}_i)^2 - \left(\sum_{i=1}^{N} \hat{\mathbf{y}}_i \right)^2 \right),
\end{aligned}
\tag{2.25}
$$

where N is the length of $\mathbf{y}$ and $\hat{\mathbf{y}}_{RGB}$. The overall loss function used, L_{RGB} is given by,

$$
L_{RGB} = L_P(\mathbf{y}, \hat{\mathbf{y}}_{RGB}). \tag{2.26}
$$

2.5.2 Radar

FMCW radar emits and receives (reflected) chirps, which are linearly frequency modulated electromagnetic (EM) waves, enabling the estimation of the distance travelled by the chirp before reflection. Aardal et al. [37] showed that one contributor to detecting a heartbeat is

a large reflection at the air/skin interface and experimentally demonstrated that heartbeat detection is primarily reliant on physical displacements of the chest. Therefore, we assume radar is not directly affected by skin tone. The transmitted and received signal, $s(t)$ and $u(t)$ respectively, can be modeled as:

$$s(t) = A_s cos(2\pi f_c t + \pi k t^2), 0 < t < T_c. \tag{2.27}$$

$$u(t) = A_u cos(2\pi f_c (t - t_d) + \pi k (t - t_d)^2), t_d < t < T_c. \tag{2.28}$$

where k is the frequency slope (the rate of change of frequency of the chirp), f_c is the starting frequency of the chirp, T_c is the duration of the chirp transmission, and t_d is the time delay between the start of transmission and initial reception of the reflected wave. Then, the bandwidth of the signal, the difference between the maximum and minimum frequencies, is given by:

$$B = f_{max} - f_{min} = (f_c + kT_c) - f_c = kT_c, \tag{2.29}$$

and the time delay is proportional to the round trip distance, $t_d = \frac{2R}{c}$, where R and c are the range of the object and speed of light respectively. Figure 2.4 indicates these values on a FMCW chirp sequence.

Upon receiving a reflected chirp, the radar mixes the received chirp with the still transmitting signal. The mixed signal is proportional to $s(t) \cdot u(t)$ and contains 2 components: a beat signal component with a frequency equal to the frequency difference of $s(t)$ and $r(t)$, $\Delta f = kt_d$, and a high frequency component situated near $4\pi f_c$. The higher frequency component is filtered out by a low pass filter (LPF) to prevent aliasing, generating $m(t)$. Concretely, the radar samples in-phase and quadrature (IQ) components such that:

$$m(t) \propto \text{LPF}[s_I(t) \cdot u(t)] + j \text{ LPF}[s_Q(t) \cdot u(t)], \tag{2.30}$$

where $s_I(t) \cdot u(t)$ and $s_Q(t) \cdot u(t)$ denote the in-phase and quadrature components respectively. The in-phase component is comprised of the transmitted signal $s(t) = s_I(t)$ multiplied with

the received signal $u(t)$. The quadrature component is derived by multiplying the received signal $u(t)$ with a copy of the transmitted signal shifted by a phase of $-90°$, $s_Q(t)$. However, IQ details are excluded for brevity and the following equations are represented with just the in-phase component. The IF signal, $m(t)$, can then be written as:

$$m(t) \propto A_m cos(2\pi f_c t_d + 2\pi(kt_d)t + \pi k t_d^2), t_d < t < T_c. \tag{2.31}$$

The $\pi k t_d^2$ term is several orders of magnitude smaller than the other terms and is thus negligible. Equation 2.31 can be rewritten into a more succinct form:

$$m(t) \propto A_m cos\left(\omega t + \phi\right), t_d < t < T_c. \tag{2.32}$$

$$\omega = 4\pi\frac{kR}{c}, \quad \phi = 4\pi\frac{R}{\lambda}. \tag{2.33}$$

The phase and frequency of the resulting signal depend on the range R and can be extracted through a discrete Fourier transform (DFT) of the signal after passing it through the analog to digital converter (ADC). The frequency term, $\omega = 2\pi\Delta f$, provides the range through the following relation $R = c\frac{\Delta f}{2k}$. Therefore, the maximum unambiguous range an object can

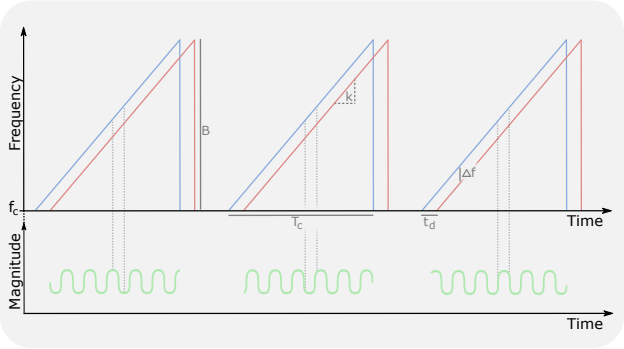


Figure 2.4: **A FMCW Chirp Sequence.** The blue and red signal are the transmitted and received chirps plotted with their frequency content as a function of time. The green signal denotes the mixed signal whose phase changes while the frequency remains relatively constant.

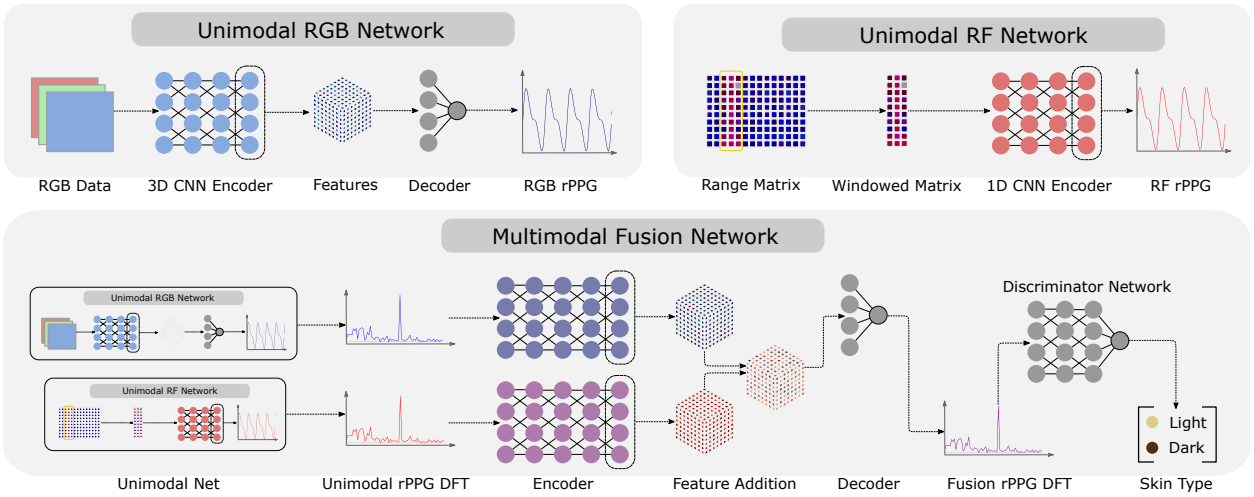


Figure 2.5: **The proposed approach uses a novel adversarial discriminative training-based approach for skin tone debiasing in the modality fusion module.** We follow a two-step training process for our pipeline - we first train the unimodal networks to estimate the plethysmograph waveform. The fusion network operates in the frequency domain using an alternating waveform reconstruction and adversarial losses.

be placed from the radar follows from the Nyquist sampling theorem and ADC sampling rate f_s, to give $R_{max} = c\frac{f_s}{4k}$. The range resolution is $c\frac{f_s}{4kN}$, where N is the number of ADC samples. The phase term ϕ is inversely proportional to the wavelength of the radar, $\lambda = \frac{c}{f_c}$.

The range of an object can be parameterized as $R(t) = R_o + r(t)$, where $r(t)$ models changes due to vibrations (for example, heart beat) around the average range R_o. To extract a heart rate, $r(t)$ needs to be sampled with multiple chirps. Note that the frequency term cannot be used to extract the sub millimeter displacement of a heart beat; the frequency resolution is on the order of centimeters. Instead, we use the highly sensitive phase to determine the oscillations of $r(t)$. The reader may note that in reality the phase extracted from the digital signal would be wrapped between $[-\pi, \pi]$. This can be solved using a standard phase unwrapping algorithm.

Practically, for the transmission and processing of the nth chirp, the ADC samples, $m_n[i]$,

are converted into the frequency domain or a single range profile, $M_n[f]$. To observe the periodic movements of the body due to a heart beat, we sample the range profile in time to construct a range matrix, $\mathbf{M} = [M_1[f], M_2[f], ...M_n[f]]^T$, such that a range bin is indexed by a fast time or range axis and a given chirp is indexed by the slow time axis (fast time refers to a chirp's ADC samples, while slow time refers to chirp samples). Figure 2.3 shows the processing pipeline as well as the range matrix and a range profile where the amplitude strengths determine a person's location.

To extract the heart rate, a frequency analysis can be performed on the phase of the central range bin (the range bin with the maximum power occupancy)[33]. However, as [34, 76] note, the phase is very sensitive to movement and body background reflection which can diminish the signal or cause interference due to the harmonics of the respiratory rate. To mitigate this, we were inspired by [76] work in UWB radar on extracting fine-grained respiratory signals by learning from raw IQ data. We employ a deep learning approach to estimate $g_2(\cdot)$, a mapping from $\mathbf{v}^2$ to $\hat{\mathbf{y}}_{Radar}$. Specifically, $\mathbf{v}^2$ consists of a window of range profiles around the central bin that are then processed by a series of 1D CNNs. The output is an estimate of the ground truth PPG signal even though the radar is only measuring vibrations. During training, we apply a data augmentation technique of rotating the complex components of the data [76] and a loss consisting of a Negative Pearson Loss L_P and a Signal to Noise Ratio (SNR) loss. The SNR loss is defined as follows:

$$L_{SNR}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\int_{f_0-w}^{f_0+w} |\hat{\mathbf{Y}}(f)|^2 \, df}{\int_{-\infty}^{f_0-w} |\hat{\mathbf{Y}}(f)|^2 \, df + \int_{f_0+w}^{\infty} |\hat{\mathbf{Y}}(f)|^2 \, df}, \tag{2.34}$$

$$f_0 = \arg\max_f \mathbf{Y}(f),$$

where $\mathbf{Y}(f)$ and $\hat{\mathbf{Y}}(f)$ are the respective Fourier transforms of $\mathbf{y}$ and $\hat{\mathbf{y}}$ and w is the chosen window size.

The overall loss function used is given by,

$$L_{Radar}(\mathbf{y}, \hat{\mathbf{y}}_{Radar}) = L_P(\mathbf{y}, \hat{\mathbf{y}}_{Radar}) + \lambda_{Radar} L_{SNR}(\mathbf{y}, \hat{\mathbf{y}}_{Radar}). \tag{2.35}$$

2.5.3 Fusion

Our fusion pipeline uniquely tries to incorporate performance properties from both the RGB and radar modality, while retaining the desirable fairness properties of the radar modality in our fused output. To achieve this, we use a novel discriminative training-based approach. The fusion model $g_f(\cdot, \cdot)$ learns a mapping from the unimodal plethysmograph estimates $(\hat{\mathbf{y}}_{RGB}, \hat{\mathbf{y}}_{Radar})$ to the fused multi-modal plethysmograph estimate $\hat{\mathbf{y}}$. A discriminator $d_f(\cdot)$ aims to classify plethysmograph signals as belonging to light or dark skin tone groups. Figure 2.5 shows this fusion network.

The discriminative training aims to minimize the mutual information between the estimated waveform $\hat{\mathbf{y}}$ and and skin tone attribute $\mathbf{a} \in A$. That is, we wish to minimize $\mathcal{I}(\mathbf{a}, \hat{\mathbf{y}})$, given by,

$$\mathcal{I}(\mathbf{a}, \hat{\mathbf{y}}) = H(\mathbf{a}) - H(\mathbf{a}|\hat{\mathbf{y}}). \tag{2.36}$$

As shown in [77], this can be approximated as a 2 agent minimax problem between the fusion network and the discriminator.

We establish modality fusion in the frequency domain, as opposed to the time domain. This allows for effective periodic structure capture and alleviates inaccuracies due to synchronization non-idealities between modalities. Specifically, the fusion model learns the following mapping:

$$|\hat{\mathbf{Y}}(f)| = g_f(|\hat{\mathbf{Y}}_{RGB}(f)|, |\hat{\mathbf{Y}}_{Radar}(f)|). \tag{2.37}$$

where upper-case characters indicate the Fourier transform of the corresponding lower-case The fusion model $g_f(\cdot, \cdot)$ consists of a two branch encoder comprised of three sets of 1D convolutional layers, followed by batch normalization [78] and ReLU activation. The final activations from the two branches are added to obtain the input to the decoder, consisting

of three branches, similar to the encoder, followed by a final convolutional block.

The discriminator model $d_f(\cdot)$ is a fully connected network with ReLU activations. The final layer has a sigmoid activation to classify between light and dark skin tone samples. Mathematically,

$$Pr\left[\hat{\mathbf{Y}}(f) \in \text{Dark}\right] = d_f(|\hat{\mathbf{Y}}(f)|) \tag{2.38}$$

The training step consists of three weight updates. *First*, the fusion model is updated using the squared negative Pearson loss L_{ppg} between the estimate waveform $\hat{\mathbf{Y}}$ and the ground truth waveform $\mathbf{Y}$ to enforce waveform reconstruction. The exponent factor in the loss function is an addition to previous work [31] enabling better reconstruction for difficult samples. The loss function is given as follows:

$$L_{ppg}(|\mathbf{Y}|, |\hat{\mathbf{Y}}|) = L_P(|\mathbf{Y}|, |\hat{\mathbf{Y}}|)^2. \tag{2.39}$$

Even though the input to the loss function is the magnitude spectrum, the Pearson loss is found to help spectrum reconstruction. *Second*, the discriminator model is updated using a binary cross entropy loss between the ground truth skin tone labels and the discriminator output $Pr\left[\hat{\mathbf{Y}}(f) \in \text{Dark}\right]$. *Third*, the generator model is also updated using a binary cross entropy loss. For this update, all training samples are assigned light skin tone labels. The loss is again evaluated against the discriminator output, $Pr\left[\hat{\mathbf{Y}}(f) \in \text{Dark}\right] = Pr\left[g_f(|\hat{\mathbf{Y}}_{RGB}(f)|, |\hat{\mathbf{Y}}_{Radar}(f)|) \in \text{Dark}\right]$.

To recover the time domain fused plethysmograph estimate $\hat{\mathbf{y}}$, we require phase information in addition to the magnitude information. We note that since the camera and radar modalities measure the plethysmograph signal at different locations on the body, a phase difference may exist between the two. Therefore, we choose to directly use phase information from the camera modality as an approximation of the phase of the fusion output. This has no effect on the heart-rate estimates, which are purely a property of the magnitude spectrum. That is,

$$\hat{\mathbf{y}} = \mathcal{F}^{-1}(|\hat{\mathbf{Y}}| \angle \hat{\mathbf{Y}}_{RGB}), \tag{2.40}$$

where $\mathcal{F}^{-1}(\cdot)$ is the inverse Fourier transform operator, and $\angle\cdot$ in the argument operator for a complex-valued variable.

2.6 Results

We perform experiments evaluating heart-rate performance and skin tone inequity on a self-collected multi-modal dataset consisting of RGB videos and Radar IQ data. Our results are compared to several approaches in RGB and a FFT-based method in radar.

Table 2.3: **Across baselines spanning the radar and camera modalities, the proposed fusion model shows performance and fairness improvements over the unimodal iPPG modality.** The performance metrics measure the average performance across the entire dataset. The pairwise difference between light and dark groups being bracketed and the sign shows direction of inequity - ideally the absolute value of this inequity should be low. The fairness threshold test measures the percent of the light and dark populations failing the AAMI standard. The best performing numbers are bolded between the fusion, RF, and PhysNet backbone.

	Performance (Fairness)				Fairness
Method	MAE ↓ (↓)	MAPE ↓ (↓)	RMSE ↓ (↓)	r ↑ (↓)	T-Test (APE %)
Green [25]	11.61 (0.23)	15.57% (1.09%)	16.56 (-0.97)	0.23 (-0.12)	42.9,52.9
ICA [29]	8.38 (4.42)	11.65% (6.19%)	14.03 (3.15)	0.41 (-0.36)	19.9,46.9
CHROM [26]	7.45 (4.97)	10.57% (6.81%)	13.38 (4.17)	0.46 (-0.38)	14.5,42.6
BCG [41]	13.01 (-0.99)	15.03% (-1.05%)	20.66 (-1.25)	0.132 (0.05)	30.5,29.1
FFT-based RF [33]	13.51 (2.25)	1.66% (2.56%)	21.07 (2.47)	0.240 (-0.25)	39.1,44.5
PhysNet [31]	1.78 (2.22)	2.35% (2.63%)	5.26 (4.05)	0.91 (-0.25)	2.1,12.2
Our RF	2.18 (**0.51**)	3.05% (**0.69%**)	6.12 (**0.85**)	0.89 (-0.13)	5.1,8.4
Our Fusion	**1.12** (0.67)	**1.52%** (0.79%)	**3.42** (1.44)	**0.95** (**-0.10**)	**1.1**,**4.2**

2.6.1 Experiment Setup

2.6.1.1 Dataset

To conduct evaluations, we recruited 91 volunteers to participate in this study. The dataset contains 28 light, 49 medium, and 14 dark skin tone volunteers. The skin tones were labeled according to the Fitzpatrick scale [79]. The volunteers are primarily from a college background, with representation between genders. 6 recordings were taken for each volunteer. A recording lasts 30 seconds and consists of a RGB video and radar IQ data. Environmental factors such as lighting variations are left in the dataset to enable more robust skin tone inequity trend analysis. Figure 2.6 describes our data collection setup in detail. The data was taken using one camera from a ZED stereo camera and a Texas Instruments AWR1443 RF development board at a distance of 0.5-1m. The entire dataset consists of over 550 recordings and 18,000 unique beats of the hearts. The ground-truth plethysmograph signal was acquired using an IntelliVue MX800 clinical grade patient monitor. An external computer is used to synchronize the capture and storage of the two modalities and the ground truth. The entire data collection setup is mobile, enabling large scale data capture irrespective of location.

The RGB camera was used with default factory settings at 30 fps. Videos were processed to 128x128 crops using a MTCNN [80] to locate facial regions. The FMCW radar was set to emit 120 chirps per second with a frequency slope of 60 MHz/μs, starting frequency of 77 Ghz, bandwidth of 3.720 Ghz, and sampling rate of 5 Mhz using a single transmitter-receiver pair. The sampled IQ data was processed into a range matrix and data related to regions of interest extracted within a 25 cm window.

We note that several previous works [30, 1, 31] evaluate metrics over 30 second windows of the estimated waveform. In this work, we choose to evaluate metrics over 10 second widows with a stride of 128 samples (4.27 s) instead. This provides a more realistic setting for analysis (with lower latency), in addition to better highlighting the effects related to

inequity and fairness. All evaluations and metrics are evaluated over six independent data splits. For each split, we include 40 participants in the traning set, 12 participants in the validation set and all remaining participants in the test set. Note that to enable control over the participant skin tone representation in each split, we ensure that the train and validation set have equal number of participants from Fitzpatrick groups I, II, III and groups IV, V, VI.

2.6.1.2 Training Configuration

All data processing and training was implemented using Python and the Pytorch machine learning library on a Nvidia Tesla P100 GPU. All neural network architectures architectures were trained with an Adam Optimizer [81] with a learning rate of 10^{-4} for 30 epochs. The fusion model is trained for 100 epochs. The ground truth was downsampled to 30 Hz and training clips set to 64 frames.

2.6.2 Evaluation

We perform a quantitative analysis of general performance and fairness and a qualitative analysis of the resulting waveforms. We compare our fusion method to seven other unimodal approaches [29, 26, 25, 41, 33, 31] including our unimodal radar model. The 30-second recordings were divided into 10-second windows with a stride of 128 frames. We then evaluate the performance and fairness metrics described in section 2.3 to the windows. Ground truth and estimated heart rate were calculated as in prior work [30].

2.6.2.1 Qualitative Analysis

Analyzing the *generated photoplethysmograph waveforms* allows for visual inspection of the estimated waveforms. The heart-rate estimates are frequency estimates obtained out of these waveforms. Figure 2.7 shows estimated plethysmograph waveforms for randomly chosen

samples from the light and dark groups respectively. We compare the estimated waveforms from the best RGB and radar unimodal models with our fusion-based model. For the RGB only modality, a degradation in signal quality is observed from the light to the dark skin tones. This highlights the performance inequity in the modality. For the radar only modality, we note across the board noisy waveforms. However, there is minimal inequity across skin tones. Our fusion model infers the best qualitative waveforms across all three groups, while also reducing the skin tone inequity that is evident in the RGB only modality.

Our second set of qualitative resources for analysis are Bland-Altman plots [82] (Figure 2.8). These plots visualize the distribution of the heart-rate estimation error versus the ground truth heart rates. The plots highlight that ground truth heart rates are distributed over a large range. In terms of heart rate estimation accuracy, we note good performance for the RGB only modality. However, visibly significant inequity is present across skin tones, as visible from the error distributions. For the RF only modality, we note a larger spread in the distributions and a larger 1σ value. However, the skin tone inequity is minimal. Again, our fusion method show significant performance and inequity improvements, with lower and largely similar 1σ thresholds across skin tones.

2.6.2.2 Quantitative Performance

Table 2.3 highlights the performance measures for the various compared methods. The signal processing-based iPPG and radar-based methods reflect relatively poor performance on our dataset. We note that in general, the iPPG methods show better performance but the radar-based method is fairer.

The deep learning-based iPPG method [31] shows significant gains in performance over the signal processing methods. This highlights the benefit of data-driven nonlinear modeling. However, the performance inequity between groups becomes much clearer. This reinforces the existence of fundamental inequity in the iPPG modality. We also note that implementations for DeepPhys [30] and LSTM PhysNet [31] were tried on our dataset, however they did not

converge during training. We believe this may be due to small misalignments between the ground-truth and video, that only the 3D-CNN PhysNet can handle.

Our deep learning-based RF method follows a similar trend. A significant improvement in performance is observed when compared to the signal processing-based RF method. However, the overall performance is lower that that of the deep learning-based iPPG method. Additionally, we note the significantly lower performance inequity between the groups.

Our fusion method outperforms all previously listed methods. We see clear improvements in overall performance across all metrics. In addition, we also notice significant improvement in inequity measures when compared to the RGB unimodal methods.

Notably, despite the lower performance of the radar-based method compared to the iPPG and fusion methods, the performance still remains high (with an average MAE of 2.18 beats per minute). Therefore, from the perspective of our evaluations on average performance, all three methods (iPPG, radar and fusion) show acceptably high average performance. Readers may note that real-world factors such as motion, distance from sensor and so on may have relevant effects on the relative performance of the three methods, that are beyond the scope of evaluation of this work. Such a detailed study, which will determine a more general conclusion on the viability of the three methods, is deferred to future work.

2.6.2.3 Fairness

Table 2.3 highlights performance inequity measures for the various compared methods. We note modality specific trends for these measures. For the RGB-only modality, we note that the T-Test values for the light and dark groups, as well as the performance inequity measures, show significant inequity. This is consistent with our theoretical analysis and expectations. On the other hand, for the radar modality, the T-Test values show low inequity. This is also noted in the performance inequity measures. Finally, we note that our fusion method achieves significantly better performance inequity and T-Test scores compared to the RGB-

Table 2.4: **An adversarial network for skin tone estimation is a novel contribution that helps obtain a more equitable plethysmograph estimator across skin tone.** When compared with a fusion network trained without the adversarial network, significant improvements are noted across all performance fairness measures, at a small cost in performance measures.

	MAE ↓ (↓)	MAPE ↓ (↓)	RMSE ↓ (↓)	r ↑ (↓)
Fusion w/o AN	**1.09** (0.98)	**1.47%** (1.28%)	**3.31** (2.56)	**0.964** (-0.146)
Fusion w/ AN	1.12 (**0.67**)	1.52% (**0.79%**)	3.42 (**1.44**)	0.953 (**-0.102**)

only modality. Compared to the radar modality, we note slightly worse performance inequity - the fusion method shows better performance inequity in terms of some metrics, while the radar-only method shows better performance inequity in terms of other metrics.

The observations from the previous subsections set up a tradeoff between performance and fairness. The proposed fusion method improves on both fronts over the iPPG method, but not compared to the radar-based method. Therefore, both these methods (fusion and radar-based) are potentially deployable candidates. We discuss this in some detail in Section 2.7.

2.6.3 Benefit of the Skin Tone Discriminative Loss

To establish the importance of our proposed skin tone-based adversarial discriminator, we compare against a naive fusion regime, trained only with our squared Pearson loss. Table 2.4 highlights the benefit of the skin tone discriminative loss. We note that the addition of the discriminative loss significantly reduced the skin tone inequity of the fusion model, at a small performance cost. The model is encouraged to minimize the distributional gap between the estimated plethysmograph waveforms for light and dark skin tone groups.

Table 2.5: **Our proposed fusion plethysmography has a similar runtime as the unimodal camera-based method.** This is because 77 GHz radar does not have much processing time in comparison to an image. The values tabulated above have been averaged across multiple runs for a 30 sec recording. The runtime values were clocked on the same hardware configurations as used for the training.

	Time (s)		
Method	Pre-processing	Inference	Total
RGB Only (PhysNet [31])	3.856	0.846	4.702
Radar Only (Ours)	0.172	0.623	0.795
Modality Fusion (Ours)	4.0282	1.4754	5.5036

2.6.4 Runtime Analysis

Table 2.5 highlights details pertaining to runtime complexity of the proposed multi-modal fusion method in comparison with the best unimodal iPPG and radar methods. The numbers in the pre-processing section are representative of the time taken by the dataloader for a single sample (30 seconds). Due to the variations in the nature of data-capture from one researcher to another, we have not included the dataset preparation time. Overall, we note that the processing of the RF signals is faster, due it its 1D nature, when compared to the images, which employ 2D algorithms. We note that the multi-modal fusion framework has a runtime similar to that of the RGB and radar network combined. This is due to the fusion model receiving its inputs from the RGB and radar models.

2.7 Discussion and Limitations

In summary, we fuse data from camera streams and 77 GHz radar to create a higher performing and more equitable plethysmography technique. To encourage reproducible research, we make the dataset and reference designs available. While we have only fused camera and

radar data, these two modalities were chosen for a reason. Camera-based plethysmography is generally known to be high performing but exhibits high skin tone inequity. In contrast, from our experiments, we note that radar-based methods have relatively poorer performance but are more resistant to skin tone inequity. This work demonstrates that in both theory and practice, sensor fusion of RGB and radar modalities can improve performance and fairness.

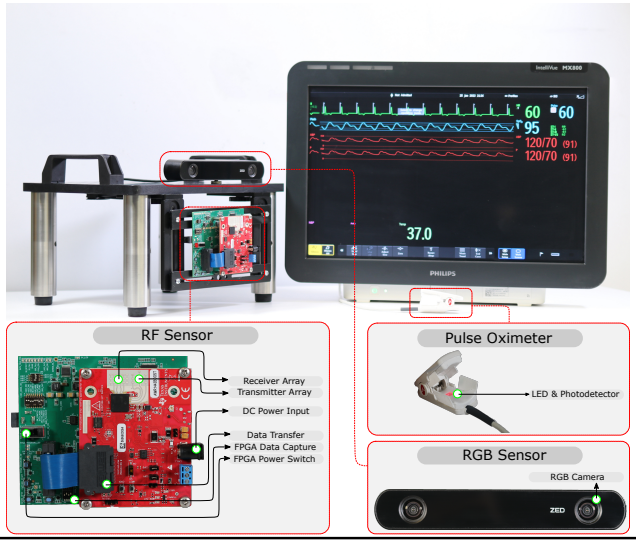
Comparing Fusion Results with Radar The goal of this work is to improve the fairness of camera-based plethysmography through fusion with radar measurements. Through such fusion, we show performance and fairness gains over the camera-based modality. An alternative analysis is comparison of the fusion method with the radar modality. This sets up a tradeoff-based selection: the fused modality shows sizeable performance gains over radar, albeit with a small reduction in fairness. Notably, this means that the dark skin tone performance of the fusion method is superior to both unimodal methods. This choice between performance and fairness depends on the end-user. We present the fusion-based method to the community as a viable alternative, for potential adoption. Follow-up works can attempt to improve the fusion-model fairness.

Limitations Constrained by relatively smaller data sizes, and imperfect skin tone balance in the dataset, we use a late fusion parameterization for our model, to enable better conditioned model training. End-to-end learning for multi-modal fusion may be explored in future works for improved performance gains. In addition, while our dataset is the largest dataset for camera-radar fusion plethysmography with a focus on demographic diversity, we note the need for continued effort towards dataset collection. Our definition of inequity is also specific in nature. We deal with inequity in terms of the signal to noise ratio (SNR). This is one potential interpretation of inequity, and future work can extend our tools and analysis to other definitions.

Future Work and Conclusion In follow-up work, it is possible to add further sensing modalities such as thermal, acoustic, near infrared, and polarization images to this dataset. These additional modalities have their own uses and can aid in the sensing of additional physiological information beyond the plethysmograph. We conclude by noting that this work is a small step in what might be a much bigger trend for the next-generation of internet of things (IoT) devices. Such next-generation IoT devices will move to incorporate both performance and equity as quality metrics.

2.8 Ethical Considerations

Although there is much algorithmic research on fairness, it is imperative for *devices* to also be equitable and not disadvantage segments of the population. This is particularly important for devices that may one day be used clinically.



Item Description	Model	Qty.	Manufacturer
RGB Sensor	Zed2	1	Stereolabs
Radar	AWR1443BOOST	1	Texas Instruments
Optical breadboard	SAB1012	1	Base Lab Tools
Mounting legs	PTB060	4	Base Lab Tools
Mobility handles	BHD010	2	Base Lab Tools
Vital sign monitor	MX800	1	Philips
FMS Rack	M8048	1	Phlips
Finger oxygenation probe	OEM 1863	1	Masimo

Figure 2.6: **A mobile multi-modal sensing platform was deployed to collect our remote plethysmography dataset. The parts list and reference designs may be found at https://github.com/UCLA-VMG/EquiPleth.** Key parts include a Zed2 RGB camera and Texas Instruments TI AWR1443 FMCW radar chip for signal measurement, in conjuction with a Philips MX800 clinical patient monitor and clinical peripheral hardware for ground-truthing.

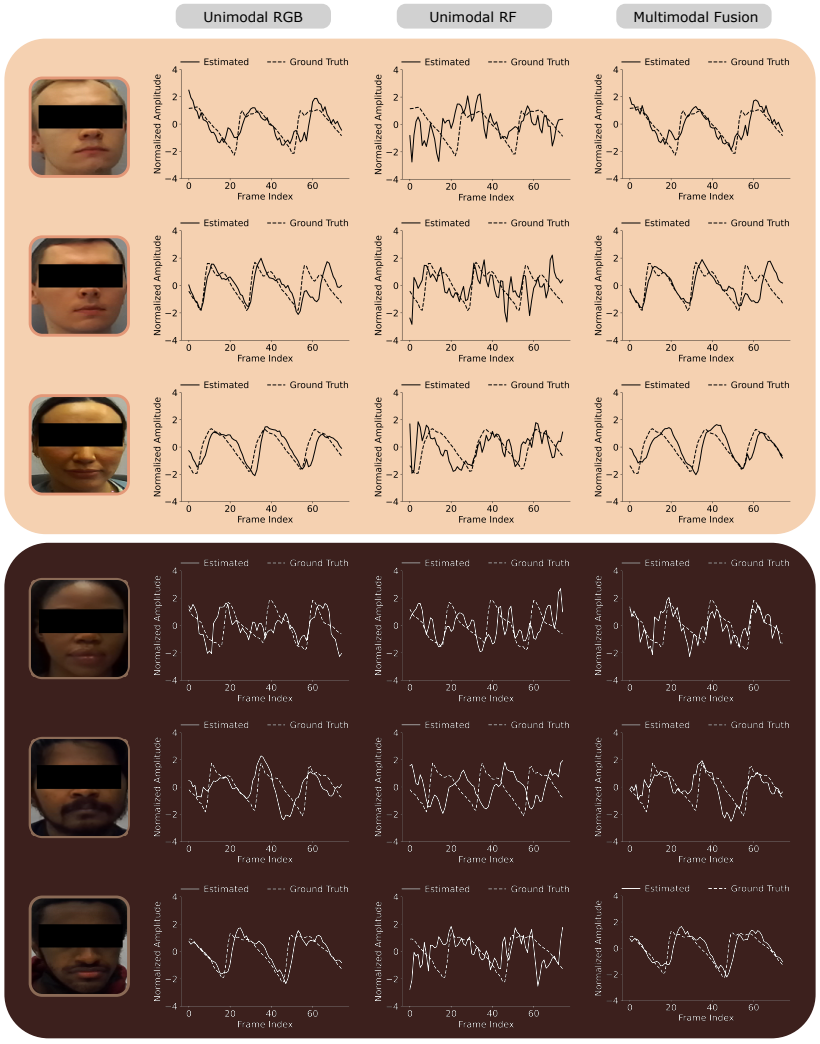


Figure 2.7: **Qualitative analysis of estimated waveforms indicates superior overall performance for the fusion model, with reduced group-wise inequity.** We highlight a randomly chosen snippet of the plethysmograph waveform to highlight qualitative differences. The RGB modality shows accurate reconstruction for the light skin tone group; however the waveform reconstruction for the dark participants is visually noisy. The radar modality shows poorer performance across the board compared to the RGB modality, but with reduced bias/inequity. Our proposed fusion model shows superior reconstruction as compared to both unimodal models. Additionally, the reconstruction for the dark skin tone participant is significantly better.

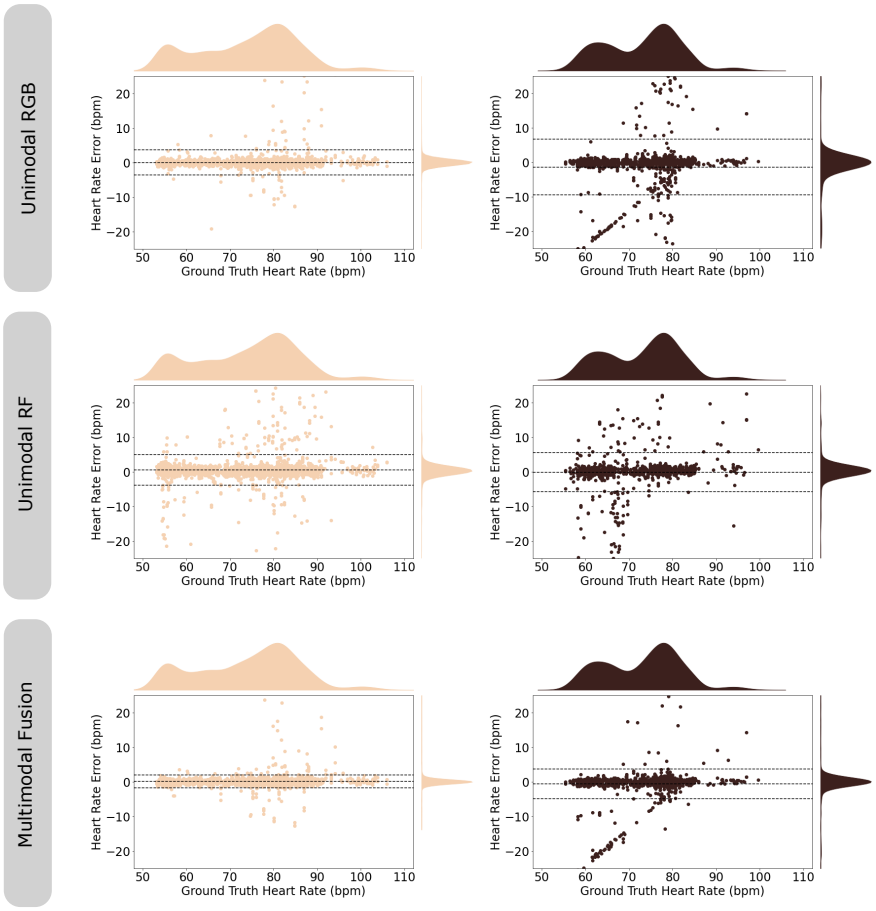


Figure 2.8: **Plotting the heart-rate estimation error versus the ground truth heart rate (Bland-Altman plots) emphasizes performance benefits of the proposed multi-modal fusion model.** Each plot highlights the distribution of the ground truth heart rates (top), distribution of the heart-rate estimation errors (right) and the plot of the estimation errors versus the ground truth heart rates. The ground truth heart rates cover a broad range for the two skin tones. In terms of error distribution, the RGB only model shows a visually distinguishable inequity (difference between the spread of the error distributions) between the light and dark skin tones. The radar only modality has a poorer overall performance but much lower inequity between groups. The proposed fusion model shows the best performance across skin tones, in addition to having inequity that is better than the iPPG modality.

CHAPTER 3

Building a Sensor for Contactless Touch Sensing in the Wild

3.1 Introduction

Force is a ubiquitous signal that occurs when objects are in contact. As a side product of human activities in environments, force reveals unique information and force sensing has a wide range of use cases in ubiquitous computing and human-computer interaction. For instance, touch interactions such as discrete button touches, swipes, and scrolling induce force between user fingers and interaction mediums such as buttons, glass panels, and skin. Robots rely on force as critical feedback for object manipulation. Moreover, the sensed force can be used to derive a rich set of second-order signals. For example, force applied to host surfaces by objects reveals their weights. Sensing the force between user fingers and contact surfaces adds an additional dimension to touch interactions. All these signals constitute rich information that intelligent vision-based sensing systems could leverage in addition to RGB and depth to become more robust, accurate, and even privacy-preserving.

In this chapter, we consider only normal force, applied to objects in contact perpendicular to the contacting surfaces. To sense this force, conventional approaches instrument sensors (e.g., Force Sensitive Resistor) on surfaces, or in between objects. This contact-based sensing approach either requires wiring which can be inflexible to deploy, or runs on battery-powered wireless sensor systems, which is costly to scale and maintain. Additionally, contact-based sensors could be sensitive to exposure of elements, and thus can be prone to error without

periodical calibrations. These inborn challenges of the contact-based approach eliminate sensing opportunities for a wide range of low-cost and passive objects such as 3D prints and room utilities (e.g., walls, tables, faucets). There are also scenarios where contact-based sensors might not be preferable such as on-body interactions, from a user experience perspective.

To address these challenges, we create a non-contact force sensing approach based on laser speckle imaging, a well-known imaging technique commonly used for medical applications (e.g., blood flow assessment) but now adapted to enable non-contact sensing for ubiquitous force signals that a wide array of interactive systems could leverage. Specifically, we detect minute deformations of surfaces when force is present. Our key observation is that laser speckles change significantly at surface deformations, even with very small magnitude. Because laser speckles are caused by scattered signals added constructively and destructively depending on their relative phases, surface deformations of the same order of magnitude as the laser wavelength (several hundred nanometers) can alter laser speckles significantly. During the course of surface deformations, the changes of laser speckles have structured spatial and temporal patterns that correlate with the amount of force applied. Our system, which mainly consists of a defocused camera, a laser source, and signal-processing algorithms, detects these structured patterns to infer the amount of force.

In this research, we first conducted a series of benchmark tests with common everyday materials and a high-precision force-sensing linear actuator to verify our sensing principle. Then we established a calibration process for later evaluation. Our core signal-processing algorithm features optic flow displacement tracking and denoised aggregation. We investigated two sensing configurations – one is the *diverged laser setting*, with a diverged laser beam covering a wide surface area in which force could happen anywhere inside; the other is the *focused laser setting*, which uses a focused laser beam to sense force at known locations. Finally, we conducted an evaluation that systematically investigated *ForceSight* with three common materials – wood, plastic, and metal of various sizes and thicknesses, and at various

distances with two calibration methods. We also investigated a wide spectrum of factors in supplemental studies to fully tease out the performance of our system. The results indicated a robust and accurate performance of our system, with all average errors across all materials and distances being less than 0.31 N. Finally, we demonstrate the applicability of our system with example applications. Overall, our contributions include:

- A theoretical model of laser speckle motion due to force-induced surface deformations.
- An end-to-end system including hardware and signal processing algorithms for non-contact force sensing based on laser speckle imaging.
- A systematic evaluation including two sensing configurations, two calibration procedures, and multiple series of tests to investigate the feasibility of the sensing approach.
- A representative set of example applications that demonstrate the expressivity of our proposed sensing approach.

3.2 Related Work

3.2.1 Laser Sensing for Interactive Systems

Laser is widely used in sensing systems for being collimated and coherent âĂŞ two unique properties that contribute to signal-to-noise ratio and high sensitivity respectively of laser-based sensing systems. Previous systems have leveraged collimated lasers (with low divergence) in creating interactive systems, e.g. Digits [83] uses angled line lasers to intersect fingers for finger position estimation and hand pose reconstruction. When modulated with RF frequencies as carrier waves, range-finding laser beams (i.e. LiDAR). have long been used to build interactive surfaces (e.g., The LaserWall [84, 85] and SurfaceSight [86]). A different object tracking principle using feedback loops featuring a movable mirror platform and a camera has been shown [87]. Lumitrack [88] used films in concert with lasers to have

structured light patterns on ambient optical sensors for 3D tracking. Due to the high coherence, constructive and destructive interferences between reflected laser wavefront result in light patterns of bright and dark dots respectively. This light pattern is called laser speckle, which has been thoroughly explained by Zizka et al. [89]. Next, we review prior work using this phenomenon, which *ForceSight* also leverages.

3.2.2 Laser Speckle Imaging

First, it is possible to have laser travel inside the transmission medium, which alters the laser path resulting in distinctive interference that encodes information into laser speckle patterns. For example, Li et al. [90] used laser speckles to detect perturbations of optic fibers. Kim et al. [91] used a similar principle to detect deformations of a scotch tape, through which pressure inside the cavity can be detected remotely. Note that it is possible to use non-laser optical approaches to detect surface deformations (e.g., [92, 93]), the use of laser by its nature of active sensing significantly improves the SNR and thus lowers the complexity of hardware and software.

Closer to our setup is prior work that detected laser speckles induced by the reflections of object surfaces. Prior work has demonstrated laser as carrier signals to reveal material type information [94, 95]. Jo et al. [96] and Smith et al. [97] leveraged the sensitivity of laser speckle to surface displacements to track objects in 3D space. With high-speed cameras, spatial correlations between speckle patterns in frames when objects are in motion can be preserved. Even fingertips can be tracked for micro-gesture input [98]. This sensing principle is akin to how a laser-based optical mouse tracks its position on 2D surfaces. SpeckleSense [89] and SpeckleEye [99] demonstrated low-cost and high-speed sensors in multiple configurations that enable rich interactive applications. It is also possible to detect second-order signals derived from this laser speckle shift caused by surface displacements. Shih et al. [100] demonstrated laser speckle imaging in surface tampering detection. Surface waves caused by in-air acoustic signals or vibrations from built-in motors can also be detected remotely

with laser speckle shifts [101, 102, 103, 104].

Closest to our work are Laser Speckle Imaging systems in the medical domain, with their capability to sense surface deformations over time at microscales. Researchers have used defused laser on body tissues to image blood flow. The slight deformations of microvasculature due to blood flow cause minute laser speckles movements. These movements generate blurred local regions on images [105, 106]. This sensing principle is easy to set up, low-cost to implement, and has shown a wide array of use cases in clinical settings (e.g., Dermatology [107], Ophthalmology [108], and Neurology [109]). For a complete review of laser speckle's clinical applications, we recommend Heeman et al. [110].

3.3 Modeling Laser Speckle

3.3.1 Laser Speckle Pattern on Rough Surfaces

When rough surfaces are illuminated by laser beams, a random interference pattern will be observed on the image plane, called laser speckle [111]. To elaborate, a whole diffuse surface can be regarded as being composed of massive independent scattering surface elements, which result in statistically independent phases of the reflected laser beams. These non-coherent beams add up constructively and destructively as they traverse in space, forming granular patterns of random distribution on the image plane.

To efficiently verify laser speckle forming, we built the model in simulation. We use a Gaussian beam to simulate an incident laser over a uniform random rough surface $\Phi(x, y)$. Beam distribution $g(x, y)$ on the rough surface can be described as Eq. 3.1 [112].

$$g(x,y) = \frac{\omega_0}{\omega} e^{\left[-(x^2+y^2)(\frac{1}{\omega^2}+\frac{ik}{2\rho})-ikd\right]} \tag{3.1}$$

where ω_0 and ω are the waist radius, and illuminated beam spot radius, respectively. (x, y) refers to the location on the rough surface. $k = \frac{2\pi}{\lambda}$ in which λ is the wavelength of Gaussian beam. ρ means the wave-front curvature radius. d is the shortest distance between

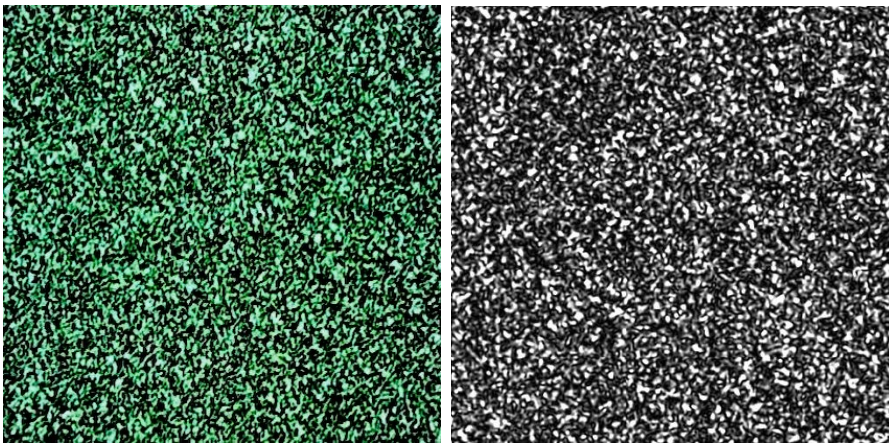


Figure 3.1: Left: Real speckles. Right: Simulated speckles.

the laser source and the rough surface. The back-scattered light can be modeled by the Fresnel diffraction [113].

To validate our simulation, we collected real-world laser speckles on a white wall using a USB camera with a resolution of 2592ÃŮ1944. The laser speckles were induced by a 10 mW 532 nm green laser (12-degree divergence), positioned 10 cm away from the wall. The real-world speckles and simulated speckles are shown in Fig. 3.1. The simulated speckles resemble real speckles in terms of their size and shape, though the overall distribution is sparser for the difference between the wall surface and the random rough surface.

With the subtle deformation of the object surface, speckle patterns of adjacent timeframes have high similarity, which allows speckle motion tracking. However, the speckle patterns can also "boi", meaning the speckles can tumble randomly fading in and out and the original spatial structure of patterns alters. In general, speckle motion appears as a combination of speckle translation and boiling, since the speckle deformation would occur inevitably [114]. To compensate for the boiling effect, as we will show later in the chapter, we designed our algorithm so that spatial continuity is not a prerequisite, i.e., we do not track the same set of speckles over long distances on the image plane.

3.3.2 Laser Speckle Motion Due to Surface Deformation

In this section, we derive a theoretical *speckle flow* model to explain how the laser speckle patterns change due to surface deformations in the presence of force.

3.3.2.1 Deformation Model

For simplicity of exposition without loss of generality, we frame the physical model as applying a concentrated load to the center of a rectangular plate with edges simply supported. Assuming the plate is isotropic and homogeneous, we can simplify the problem by looking at its transverse cross-section. As shown in Fig. 3.2 B, f is the point load actuated at the center of the beam, and x is the distance from the center to a point of interest. $\delta(x, f)$ and $\theta(x, f)$ are the plate deformation distance (i.e., deflection) and the angle in radians at the point of interest, respectively. δ_{max} is the maximum deflection which locates at the plate center. The θ is defined as 0 when the plate is not deformed. The equations of the deflection and angle are as follows [115, p. 330–331],

$$\delta(x, f) = \frac{f}{48EI_s}(L^3 - 6Lx^2 + 4x^3) \qquad 0 \leq x \leq L/2 \tag{3.2}$$

$$\delta(x, f)|_{x=0} = \delta_{max}(f) = \frac{fL^3}{48EI_s} \tag{3.3}$$

$$\theta(x, f) = \frac{f}{4EI_s}(Lx - x^2) \qquad 0 \leq x \leq L/2 \tag{3.4}$$

$$\theta(x, f)|_{x=\frac{L}{2}} = \theta_{max} = \frac{fL^2}{16EI_s} \tag{3.5}$$

where E is the Young's Modulus, I_s is the moment of inertia of the material plate, and L is the side length of the plate. The formulas indicate that the deflection, scales linearly with the magnitude of applied force, as verified in Section 3.3.3 (Fig. 3.3 Right).

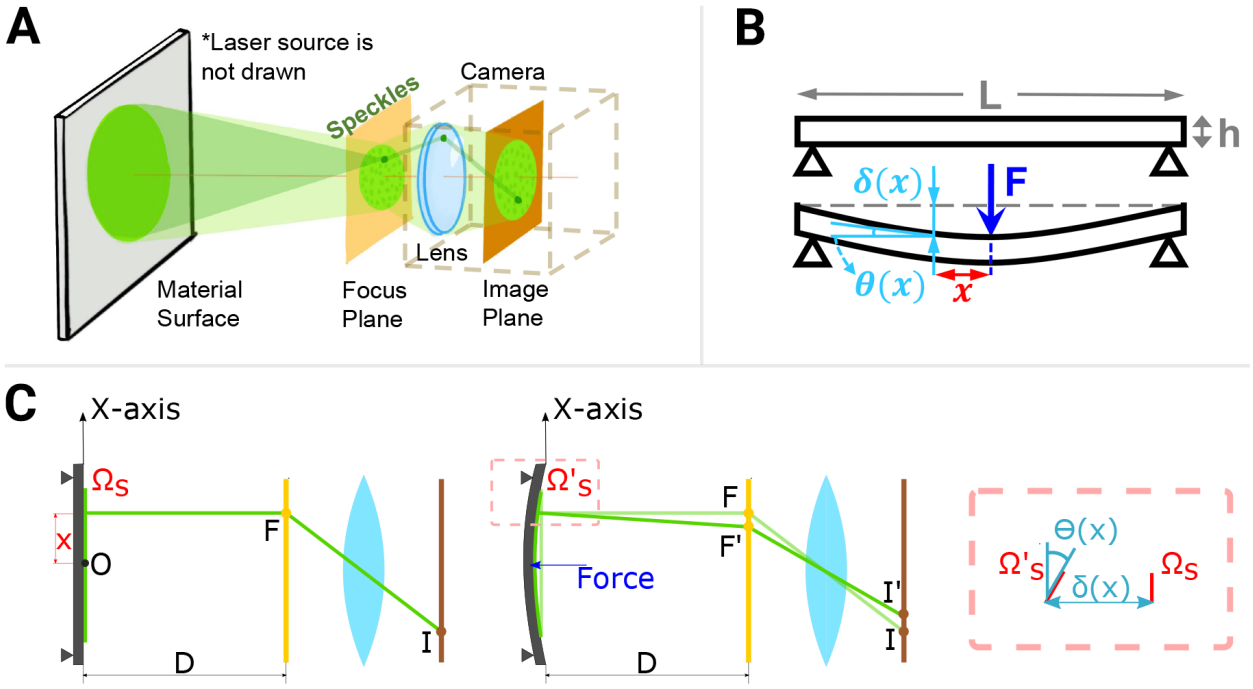


Figure 3.2: *ForceSight* Modeling. A: Configuration of laser speckle imaging. A defocused camera captures speckles formed by laser beams reflected from the material surface. B: Deformation model. C: Due to surface deformation at force, a laser beam reflected by the micro-surface Ω_s changes its imaging position from I to I' on the image plane. Left: no force applied. Center: force applied at O. Right: zoomed-in micro-surface.

3.3.2.2 Micro-Surface Hypothesis

We use the following hypothesis to approximate our surface for the sensing principle. A surface can be divided into multiple small sub-surfaces, as Fig. 3.2 C shows. When a sub-surface is small enough, its area becomes insignificant for our interest. We define such a sub-surface, as a *micro-surface*. If we look from the side, the surface plate is idealized as a polygonal line combining line segments of all micro-surfaces.

3.3.2.3 Speckle Motion due to Surface Deformation

As shown in Fig. 3.2 C, in 2D space, the location, deflection, and angle of a micro-surface are respectively represented by x (distance from plate center to the micro-surface), δ, and θ. Proof in 3D space is a symmetry-based extension of our discussion in 2D space, with 3D coordinates and normal vectors. Given that the deformation model is isotropic on the homogeneous plate, we will move on to prove it in 2D space which is more concise and clear.

Statement: The speckle motion goes towards the contact point in the presence of force. The motion displacement on the image plate can be described by

$$\Delta I = aD\frac{f_0}{4EI_s}(Lx - x^2) \qquad\qquad 0 \leq x \leq L/2 \tag{3.6}$$

where a is a scaling factor of focus-image projection model, D is the focus-surface distance, f_0 is actual force, L is the length, x is the $\Omega_s - O$ distance (from the micro-surface to the plate center).

Configuration: As shown in Fig. 3.2 A and C, our model consists of a laser, a material surface (i.e., the plate in the previous discussion), and a camera with a focus lens, a focus plane, and an image plane. The focus plane denotes the plane where objects are in focus, whose position can be deduced from the Thin Lens Equation. The camera is defocused, thus speckle motion is obvious while the imaging of surface is blurry, preserving the SNR of our setup by not letting surface textures register on the image plane. The coordinate system origin is O, which is set to the intersection of the material surface and optical axis of the

lens. The surface center is also configured at O. The original surface is D m away from the focus plane.

Proof: Suppose we have a micro-surface Ω_s which is x away from the origin O. A laser beam is reflected by Ω_s onto the focus plane at F. We call F a focus point (a point in focus, not a "focal point"). With no force, its deflection $\delta(x, f)|_{f=0} = 0$ and deformation angle $\theta(x, f)|_{f=0} = 0$. F registers a conjugate point I on the image plane.

Now, a small force is applied at the surface center. It pushes the micro-surface Ω_s all the way to Ω_s' with deflection δ and deformation angle θ. As a result, the focus point F shifts to F', and the conjugate point I moves to I' accordingly, toward the touch center. The deflection, angle, and speckle motion can be modeled as below,

$$\delta(x, f)|_{f=f_0} = \frac{f_0}{48EI_s}(L^3 - 6Lx^2 + 4x^3) \tag{3.7}$$

$$\theta(x, f)|_{f=f_0} = \frac{f_0}{4EI_s}(Lx - x^2) \tag{3.8}$$

$$\Delta I = a\Delta F = a|F' - F| = a(D + \delta(x, f_0)) \tan\theta(x, f_0) \tag{3.9}$$

Given our configuration where the D (m) is much larger than the surface deformation (from nm to mm), the model can be approximated as

$$\Delta I \approx aD \tan\theta(x, f) = aD\frac{f_0}{4EI_s}(Lx - x^2) \tag{3.10}$$

From this speckle motion model, we can draw several observations, which we also drew from validation in Section 3.3.3:

1. The speckle motion ΔI is linearly correlated with force f.
2. The speckle motion ΔI grows as the distance D increases.

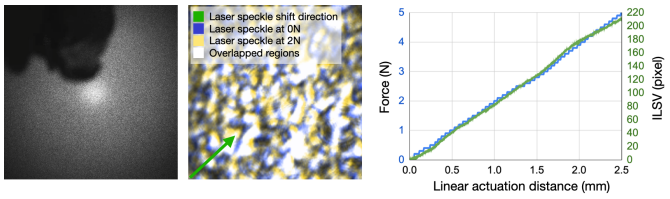


Figure 3.3: Sensing principle validated with a linear actuator setup. Left: raw laser speckle. Center: highlighted speckle shift due to the surface deformation caused by an applied force of 2 N. Right: Integrated Laser Speckle Velocity correlates with the applied force.

3. Given $I_s = \frac{wh^3}{12}$, where w and h are the width and thickness of the material surface plate, the speckle motion δI is proportional to the inverse of the cube of thickness h.

4. When the stiffness increases (i.e., the Young's Modulus E is larger), the speckle motion ΔI decreases.

3.3.3 Sensing Principle Validation

We collected data to verify our sensing principle and modeling. A motor-based linear actuator was used (see Fig. 3.5 Right) to actuate a 60.96 cm square metal surface which measured 1.59 mm ($\frac{1}{16}$") thick. Surface deformation was measured by counting motor steps (at 0.78 Âţm resolution) while the applied force was measured with the force meter affixed to the linear actuator's indenter. We bundled a camera with a diverged laser as shown in Fig. 3.5 Left (Details of this sensor bundle can be found in Section 3.4.1) and placed them above the surface. The linear actuator pushed the surface until the force reached 5 N. Data was streamed to a PC through USB.

Fig. 3.3 Left shows the raw laser speckles, while Fig. 3.3 Center highlights distinctive laser shifts (observed at 10 cm from the point of the applied force of 0 N vs. 2 N). The laser was focused during the data collection for better visualization of the laser speckle shifts, avoiding quantization in images (due to relatively small speckle sizes induced by diverged lasers). These shifts were due to small surface deformations caused by the applied force.

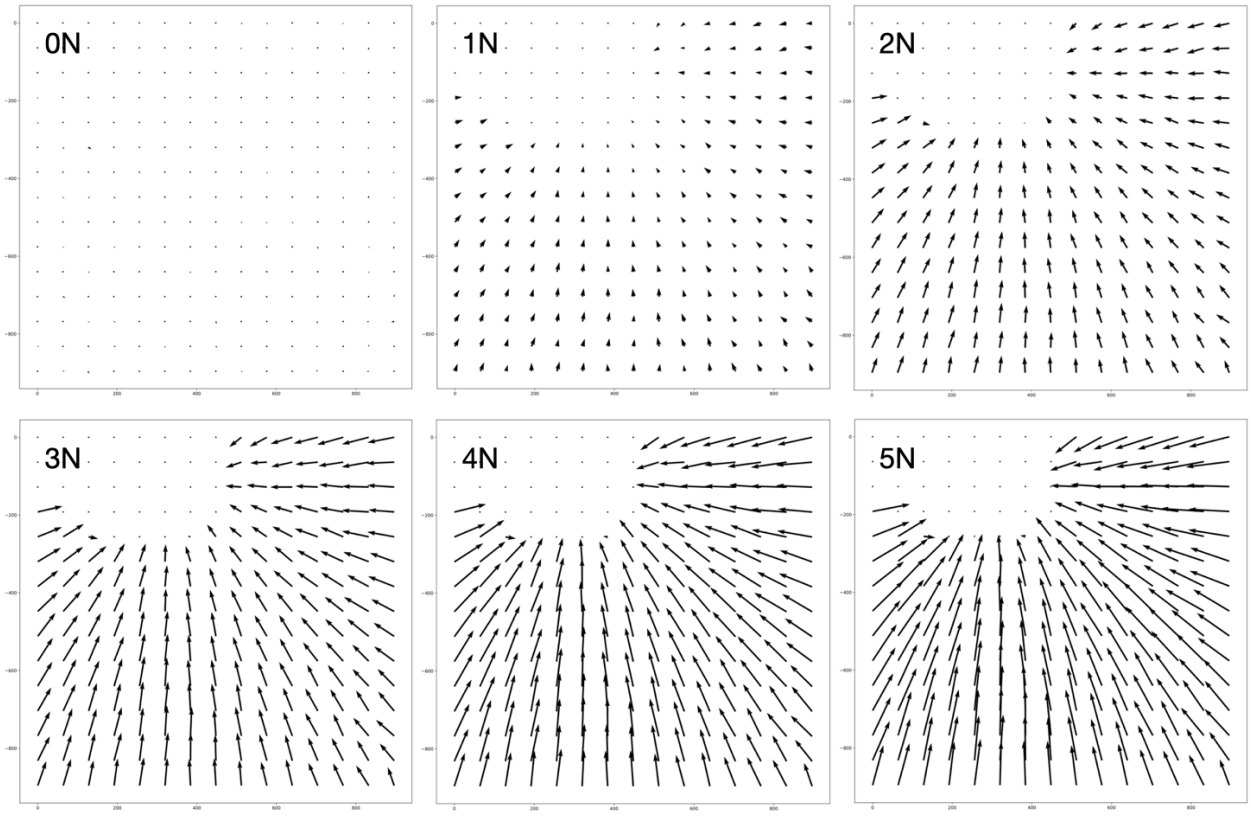


Figure 3.4: Fields of Integrated Laser Speckle Velocity in presence of different amounts of force, forming a centripetal pattern towards the force centers.

We use optical flow to calculate the distance of laser speckle shift between adjacent frames, called *laser speckle velocity* (LSV). Note that LSV was referred to as the speckle motion in our modeling section. Fig. 3.3 Right plots the integral of LSV (ILSV) and the applied force over linear actuation distance across the entire image frame excluding regions occluded by the sensor bundle. The ILSV correlates with the amount of force. We also plot out the ILSV across a larger region (900ÃŮ900 pixels) in the presence of 0 N, 1 N, 2 N, 3 N, 4 N, and 5 N forces respectively, as shown in Fig. 3.4. The length and direction of each quiver indicate the normalized magnitude and the direction of ILSV. It shows a centripetal pattern towards the force centers, with growing magnitudes as the force increases.

Overall, these results verify the sensing principle that the rest of this work builds upon.

Therefore, we can average ILSV magnitudes to get a robust indicator signal of *ForceSight* for force estimation. For higher accuracy, our force-sensing algorithm uses the magnitude projected onto the direction toward the force center for a weighted aggregation, which will be further described in Section 3.4.2.2.

3.3.4 Calibration Exploration

The test surfaces are simplifications of real-world objects which are often complex (e.g., uneven surfaces, irregular shapes, varying thicknesses, and heterogeneous material compositions). Modeling this level of complexity requires precise sensory systems (e.g., 3D scanners) and intense calculations. In comparison, *calibration* is a more viable path for its simple setup process so long as the signal has high repeatability or the algorithm can cope with shifts in signals over time and configuration changes. In fact, calibration is a common technique in Laser Speckle Imaging – for example, Laser Speckle Contrast Imaging requires captured data as a baseline [100]. Calibration is also common in force-sensing applications. For example, once FSR is inserted, it needs calibration to map its resistance to the amount of force. Therefore, we set out to design *ForceSight* with this empirical approach, to develop algorithms with minimal calibration needed in practical force-sensing applications.

3.4 Implementation

3.4.1 Sensor Bundle

Our sensor (Fig. 3.5 Left) consists of a camera (FLIR GS3-U3-32S4M-C 1/1.8" GrasshopperÂő3 1536ÃŮ2048) and a 532 nm (green) 100 mW point laser projector (from Civil Laser). We use the camera at its highest frame rate 121 fps with a fixed 4 mm/F1.8 lens (Edmund Optics) throughout the evaluation, with the camera out-of-focus such that its working distance from front housing is adjusted to 0 mm. A 532 nm camera filter is attached to the

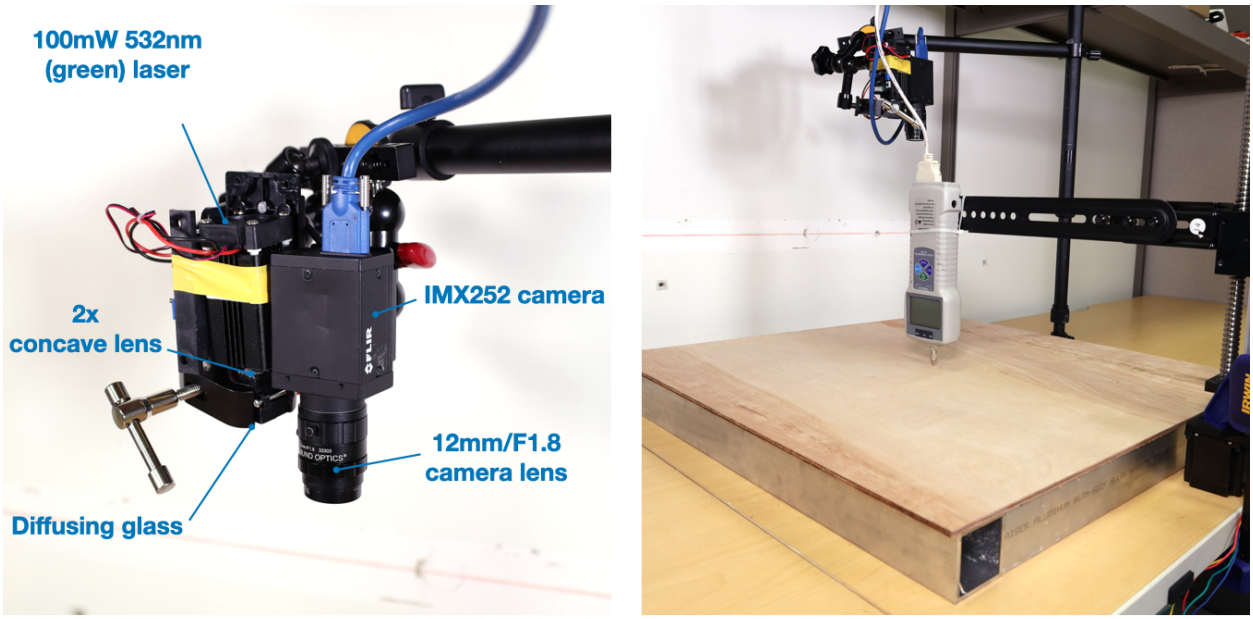


Figure 3.5: Left: *ForceSight* sensor bundle. Right: evaluation setup with the force gauge mounted on a linear actuator.

camera for better SNR. The camera and the laser projector are bundled, pointing in the same direction. The camera can capture speckles from the diffuse reflection of the laser on an object's surface.

We explore two configurations for the laser in our sensor bundle, the diverged mode, and the focused mode. In diverged mode, the laser is diverged and expanded with three concave lenses (two LD2568-A with -9.0 mm focal length and one LD2060-A with -15.0 mm focal length from Thorlabs) and one optical diffuser (HOLO 80 Deg 12.5mm from Edmund Optics), so the green laser can spread over a whole surface. In focused mode, the laser beam remains as a dot when it is landed on the object's surface, concentrating energy for long-distance sensing applications.

3.4.2 Algorithm

The output of our sensor setup is an ordered stack of video frames $\left\{\mathbf{v}^k\right\}_{k=0}^{N-1}$, where N is the total number of frames. We assume that the video is captured at a frame rate f. Our goals from this frame-stack are twofold: first, we want to reliably estimate speckle velocity fields; and second, to estimate the applied force in real-time. These aspects are discussed below sequentially.

3.4.2.1 Speckle Velocity Fields

The speckle frames have distinctive structures. Qualitatively, as a result of applied force, the speckle patterns show distinctive centripetal displacement. On smaller time scales, these can be approximated as local pattern translations. However, across larger timeframes, scale differences may also be observed in local patterns. Given these observations, we set up the velocity field estimation problem as a flow estimation problem across small timeframes. That is, we estimate flow displacement across every two adjacent timeframes, thereby obtaining a correlated metric to the flow velocity. The flow displacement is used as a proportional estimate for laser speckle velocity. Algorithm 1 includes pseudocode for this simple algorithm.

3.4.2.2 Real-time Force Estimation

Qualitatively, the applied force on the surface and temporal integral of the speckle velocity are directly correlated. Therefore, given the material and its physical configuration, a mapping may be learned to infer applied force from the integral of the estimated speckle velocity. The estimate speckle velocity is calculated by averaging projected lengths of all vectors within the image frame, towards the estimated force center. Note that this gives us a signed measure for the estimated speckle velocity. A cumulative sum of (i.e., integral) the estimated speckle velocity over time is then directly used by regression models to estimate the instantaneous

Algorithm 1: Speckle velocity field estimation

Input:

Frame stack $\left\{\mathbf{v}^k\right\}_{k=0}^{N-1}$, number of frames N, video frame rate f, Optical flow operator $\text{OpticalFlow}\left\{\cdot,\cdot\right\}$

Output:

speckle velocity stack $\left\{\mathbf{r}^k\right\}_{k=0}^{N-2}$

Initialize $\left\{\mathbf{r}^k\right\} \leftarrow 0 \;\forall\; k \in \{0, 1 \ldots N-2\}$

for $i = [0,\, N-2]$ **do**

  $\mathbf{r}^i \leftarrow \text{OpticalFlow}\left\{\mathbf{v}^i, \mathbf{v}^{i+1}\right\}$

end for

return $\left\{\mathbf{r}^k\right\}_{k=0}^{N-2}$

applied force.

3.5 Evaluation

3.5.1 Apparatus

As shown in Fig. 3.5 Right, a force curve gauge (resolution 0.1 N) is mounted on a linear actuator (resolution $7.8125x10^{-7}$ m/step), pointing towards the object surface. The sensor bundle was placed above the surface. The surface was supported on its edges by an aluminum frame base. The ground truth force reading from the force curve gauge and the raw data from the camera were streamed to a computer. The linear actuator was also connected to the computer for control.

3.5.2 Test Materials

Our test apparatus involved sheets of three types of materials (wood, acrylic, and metal) which are common to find in daily settings. These square sheets measured 60.96 cm long

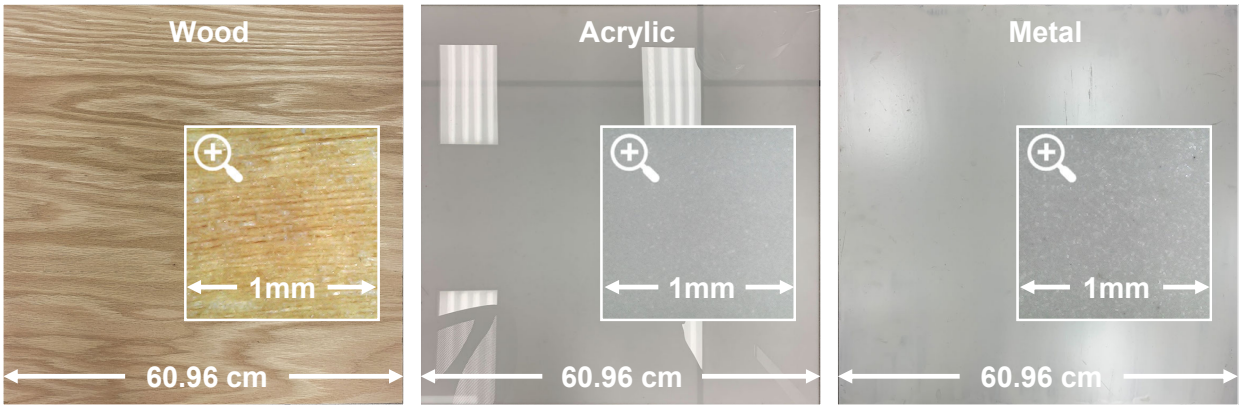


Figure 3.6: Photos and microscopic images of materials. The actual side lengths of global photos and zoom-in images are 60.96 cm and 1 mm respectively.

and of a variety of thicknesses, which are common building materials purchased from home improvement retailers [116, 117, 118]. To measure the roughness of these materials, we conducted a friction test using a 3D printed PLA instrument with a force gauge to measure the coefficients of friction. For the three type of materials we tested, their coefficients of friction measured 0.334, 0.417, and 0.301 for wood, acrylic, and metal. Fig. 3.6 shows a closer view of them.

- wood: 5 mm, 5.56 mm (7/32"), 6.35 mm (1/4")
- acrylic: 1.59 mm (1/16"), 3.18 mm (1/8"), 6.35 mm (1/4")
- metal: 0.79 mm (1/32"), 1.59 mm (1/16"), 3.18 mm (1/8")

3.5.3 Data Collection Procedures

We describe the procedures for one complete trial of data collection in this section. The first step was pre-collection preparation. Placed on top of the aluminum frame base, the sheet was simply supported by its four edges. The sensor bundle was then adjusted carefully for the correct working distance and laser coverage (i.e., diverged vs. focused modes). We also

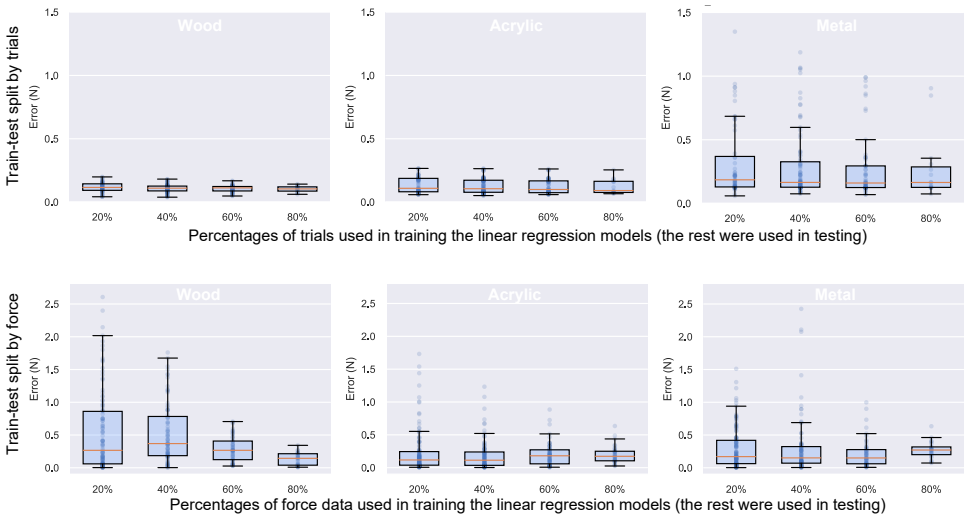


Figure 3.7: Evaluation results on sheets of three materials (wood, acrylic, metal) of various thicknesses.

adjusted the camera's exposure time and gain in software to ensure a clear view of laser speckles. Besides, we set the indenter of the force curve gauge to hover above the centroid of the sheet. The reading of the force curve gauge was 0.0 N at the beginning of each data collection trial.

We started data collection once the setup was ready. As the linear actuator went downwards at a speed of 6.720810^{-2} mm/s, the indenter of the force curve gauge approached the sheet surface with a force reading of 0.0 N. Once the indenter got in contact with the surface, the force reading started to increase as the actuated force incremented. Once the force reached 5.0 N, the force curve gauge started to retract until the reading returned to 0.0 N. The speckle images (with a resolution of 1536ÃŮ2048 at 121 fps), force readings (with a resolution of 0.1 N at 10 fps), and indenter displacement (counted in steps) were saved during this push-release process with synchronized timestamps. After the process, we recorded the location of the indenter (x_c, y_c).

The next step was post-processing. We applied a mask to remove regions where speckles were induced on the linear actuator as opposed to the tested sheet. In real-world applications,

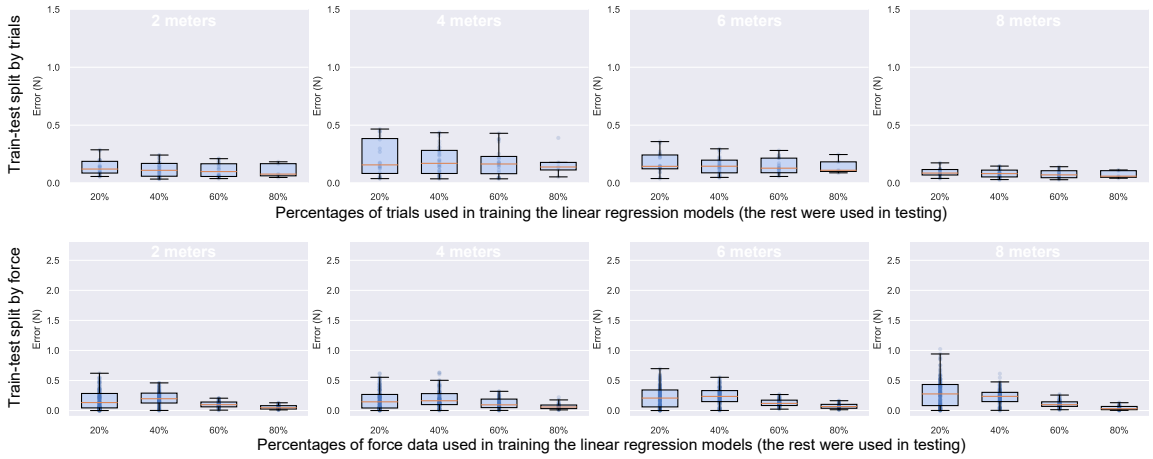


Figure 3.8: Evaluation results on four sensing distances (2 m, 4 m, 6 m, 8 m) tested on the metal sheet with a thickness of 1.59 mm (1/16").

this mask could be easily generated using depth cameras. We also set a threshold to get rid of regions that were too dark.

Following the procedure above, the collected data is called one "trial" for the given object sheet. Each trial took from 40 to 85 seconds to complete, depending on the elasticity of the material of the tested sheet. Five trials were collected per sheet. In total, we collected 2625 seconds of data with 317625 images, 2625 force readings, and 225830 linear actuator steps.

3.5.4 Train-Test/Calibration Procedures

We evaluated *ForceSight* in two procedures, each following a unique calibration process that could be used in real-world scenarios. Note that we use "train" and "test" to explain the data split in building and evaluating our regression models, though we did not use machine learning in *ForceSight*.

Procedure#1: Train-Test split by trials. In this procedure, we split the five trials into train trials and test trials with different split percentages. For example, the train-test split percentage is $1/(1+4) = 20\%$ when we build the regression model on one trial and test it

on the other four trials. Different combinations under the same percentages are grouped in an N-fold manner. This is to reflect a common real-world calibration process where sensors are calibrated with a full dynamic range of future signals to expect.

Procedure#2: Train-Test split by force. In this procedure, we first bucketed one trial of data (0-5 N) into five equal 1 N-range bins, and split the bins into train bin(s) and test bin(s) with different split percentages. For instance, the split percentage 40% indicates the regression model is built on forces in the first two bins and tested on the three remaining bins. Additionally, the train portion always starts from 0 N, and the test portion always follows the end of the train portion. It reflects another real-world scenario where sensors are calibrated with partial dynamic ranges of the future signals to expect. This is inherently challenging but could yield useful insights into the generalizability of the model.

In both procedures, we varied the amount of data in building the regression model, from 20 % to 80 % (i.e., 1-4 trials in Procedure#1, and 1-4 Newton range in Procedure#2). We evaluated our regression models with all train-test split combinations.

3.5.5 Results

We collected data in two settings, including one short sensing range with three materials (i.e., wood, acrylic, metal) using the diverged mode, and four long sensing ranges with one material (i.e., metal) using the focused mode. Additionally, we evaluated *ForceSight* with two train-test procedures. This evaluation process yielded four combinations, which we discuss in this section.

3.5.5.1 Short Range Sensing (Diverged Mode)

As Fig. 3.7 shows, *ForceSight* achieves an averaged error of 0.18 N (SD=0.11) and 0.31 N (SD=0.12) for the two train-test procedures, respectively.

Train-Test split by trials. Comparatively, the train-test split by trials (i.e., calibrating

the sensor with signals of full dynamic range) yielded better results. Among the three tested materials, *Wood* performs the best with the lowest averaged error of 0.11 N (SD=003) followed by *Acrylic* (error=0.13 N SD=0.06) and *Metal* (error=0.30 N SD=0.28). We found a significant source of error in the thickest metal sheet we tested (error=0.61 N, SD=0.31) for the small surface deformation resulting from the test force. Even with 5 N force, the surface deformation is almost invisible to naked eyes, though it can be detected by our sensor. We suspect that real-world applications with thick metal sheets would most likely involve stronger force, which could result in larger deformations and thus lower the errors (or percentage errors). When comparing between percentages of the training data, we did not find any major differences. This result indicates that *ForceSight* can be calibrated very efficiently with a small amount of data.

Train-Test split by force. Train-Test split by force (i.e., calibrating the sensor with partial dynamic ranges) yielded an average error of 0.31 N (SD=0.12) across all materials. Interestingly, *Wood* performs the worst, with an average error of 0.44 N (SD=0.49) among all materials. Based on our observations, this was due to the heterogeneous internal microstructure distribution inside the wood sheets, resulting in non-linearity, which makes it harder for the regression model to generalize for unseen signals. When comparing between percentages of the training data, we did not find any major differences, pointing us again to the insight that *ForceSight* can be calibrated very efficiently with a small amount of data.

3.5.5.2 Long Range Sensing (Focused Mode)

Fig. 3.8 shows that *ForceSight* achieved an averaged 0.18 N (SD=0.05) and 0.18 N (SD=0.03) error for the two train-test procedures, respectively.

Train-Test split by trials. Again, the train-test split by trials yielded better results among the two procedures, which suggest calibration with signals of full sensing dynamic range for real-world applications. Among the four tested distances (2 m, 4 m, 6 m, 8 m), *ForceSight* yielded average errors of 0.12 N (SD=0.06), 0.20 N (SD=0.13), 0.16 N (SD=0.08), and 0.08 N

(SD=0.04) respectively. We did not observe a clear correlation between distance and sensing performance, indicating the feasibility of *ForceSight* in long-range sensing. However, during the data collection, we observed more oscillations (i.e., noise) of laser speckles at longer sensing distances due to ambient vibration (e.g., airflow from HVAC, appliances running) and our algorithm is robust to these noises. We are cautious that severe vibrations from a longer sensing distance might require a superior denoise algorithm to process. Additionally, we did not find having more data in building regression models improves our sensing accuracy. This result is consistent with the outcome of the previous tests.

Train-Test split by force. Among the four tested distances (2 m, 4 m, 6 m, 8 m), results indicate average errors of 0.16 N (SD=0.13), 0.16 N (SD=0.14), 0.19 N (SD=0.15), and 0.21 N (SD=0.18) respectively. We did not find any linkage between distance and sensing performance. However, we found having more data in building the regression models improves our sensing accuracy.

3.5.6 Supplemental Studies

In supplemental studies, we investigated additional factors that could affect our sensing performance. Results from these additional factors further our understanding of this sensing technique and enrich its sensing vocabulary.

Angle of incidence. The angle of incidence has been a major factor in laser sensing performances due to the fact that reflected light energy increases as the laser gets perpendicular to the sensed surface. In this test, we collected data from various angles of incidence (from 0 to 40 degrees with a 10-degree internal), with a diverged laser positioned 30 cm away from the intersection point of its principal axis and the surface (i.e., 1.59 mm thick metal sheet), following the data collection and evaluation procedure as in our main evaluation. Results are shown in Fig. 3.9.

Overall, we noted an average error of 0.15 N (SD=0.06) and 0.13 N (SD=0.14) from

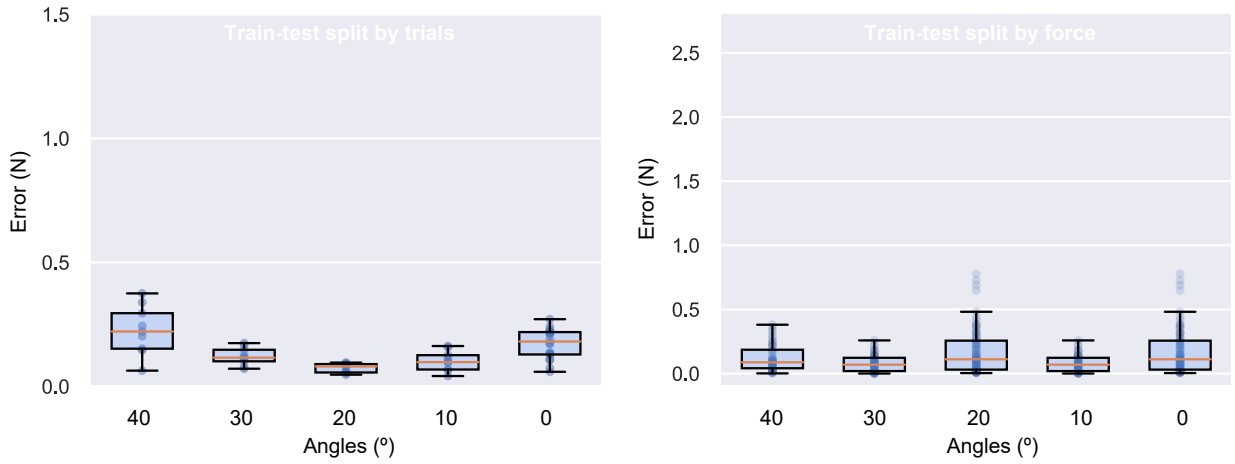


Figure 3.9: Evaluation results on the angle of incidence.

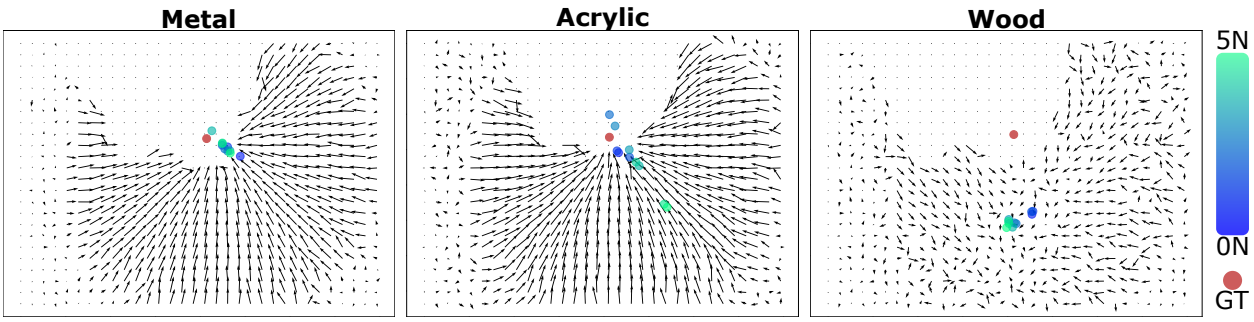


Figure 3.10: Detecting force location on different materials using *ForceSight*. The ground truth force location is shown in red. Speckle velocity is shown in a log scale.

calibration procedures #1 and #2 respectively. We did notice slight differences between performances when the sensor bundle is oriented with different angles of incidence, however, we did not see a trend that indicates a higher angle of incidence lowering the sensing performance. Though promising, we are cautious that more material types including ones that are more specular should be included in the test set.

Force location estimation. As a result of the centripetal displacement of the speckle patterns in response to force (e.g., a touch), the contact location is the common center for the estimated velocity vectors. We propose a center-estimation algorithm, which is essentially

solving a distance-minimization problem. For a given center estimate, the error value is the sum of perpendicular distances of the point from all the estimated laser speckle velocity vectors for a given velocity field. The point with a minimum error value is the best center estimate. We initialized a random center and then used gradient descent optimization. The algorithm is applied to a single frame with a learning rate of $\mu = 0.01$ and steps $T = 10,000$ in PyTorch. The final force location estimates are obtained by averaging over 10 random initialization and runs.

Figure 3.10 shows qualitative results on force location detection on the three materials (with the medium thicknesses). We note that we were able to approximately detect the force location for the metal and acrylic materials with mean Euclidean errors of 4.86 cm (STD of 1.66 cm) and 8.38 cm (STD of 6.00 cm) respectively. These results serve as proof of concept for *ForceSight* being a viable tool for not just force sensing but force location estimation as well. Notably, force location performs poorly on wood, as a result of its heterogeneous internal structure with a mean Euclidean error of 19.38 cm (STD of 1.09 cm). This establishes that the force location is limited in accuracy by the nature of material structures and resulting speckle motion features.

A wide array of materials. In this test, we included a wider set of materials and objects, including a book, pillow, package box, foam board, acrylic, wood, metal, and silicone. We collected one trial (0-5 N) of data for each material with a focused laser (i.e., focused mode) 30 cm away from surfaces and using the same procedure as previous tests. Force was applied 10 cm away from the laser dot on the tested surface. We built regression models that minimize errors (maximizing R^2) but included both linear and quadratic regression models in our search. Fig. 3.13 shows our results which indicate that simple models well fit data collected from these materials. We found *Book* to be the only object that requires a second-degree term among the test objects/material sheets. The distinctive coefficients across these materials can be used to identify material types. In this use case, *ForceSight* becomes a sensing instrument that yields elasticity of surfaces if the applied force is known.

3.6 Example Applications

3.6.1 On-world Touch Sensing

Projected touch interfaces create ubiquitous interaction experience, which much prior work has investigated [119, 120, 121]. With depth cameras, touch sensing on everyday surfaces has never been easier. And yet, commodity depth cameras cannot sense fine-grained touch with small finger movements (sub-centimeter), as shown and discussed in prior work [122, 119]. However, being able to segment touch from minute motions without having users exaggerate their movements to accommodate for sensor inaccuracy is critical to fully utilize the expressive and natural interactions provided by touch. In this regard, *ForceSight* creates a potential solution using force as an additional signal to aid touch segmentation (touch vs. no touch). Fig. 3.11 shows the integrated laser speckle velocity field on office partitions when a user touches them at forces similar to ones on touchscreens. Note that we used Google MediaPipe [123] pose tracking to exclude regions of user bodies so that the detection pipeline is robust against interference from users' motion. *ForceSight* also works with a broader array of everyday surfaces including a fabric couch arm, a wood table, walls, and a fridge door.

3.6.2 3D Printing Interactivity

ForceSight also provides a viable path to 3D printing interactivity as many previous systems aim to achieve [124, 125, 126]. To achieve this, we embedded a lite version of *ForceSight* consisting of a 3 mW laser and a low-end webcam as in Fig. 3.12 C. The lite sensor bundle costs less than $20 to make. Fig. 3.12 D and E show example interactions enabled by the 3D printed controller with embedded *ForceSight*. *ForceSight* senses and recognizes the discernible surface deformations due to the applied force when users press the buttons and tilt the joystick in different directions. Since *ForceSight* sensor bundles are installed at the controller base, a user can easily switch controller top plates for applications that demand different interactions.

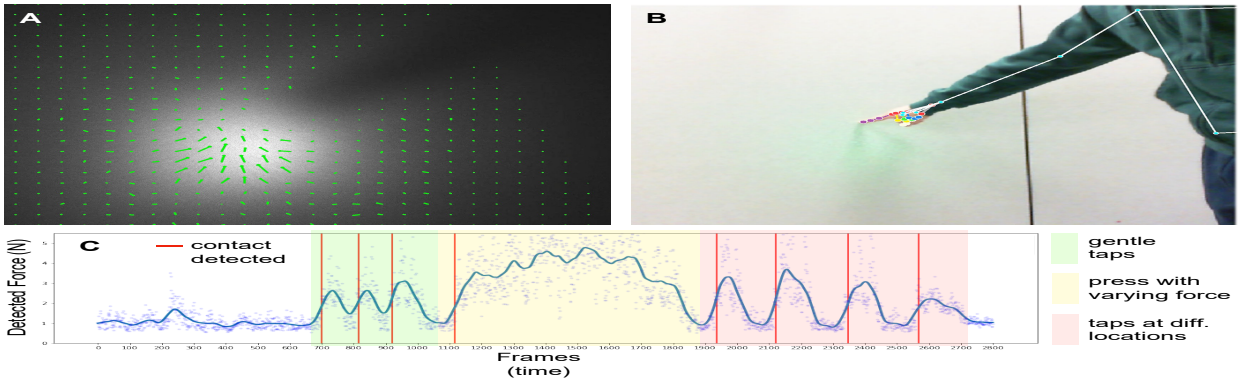


Figure 3.11: On-world true-force touch sensing. A: Integrated Laser Speckle Velocity Field overlaid on raw laser speckles. B: An RGB image captured by a webcam. C: Detected force from *ForceSight*. Of note that, to avoid optical flows induced by user motions, sensing is turned off at regions that are recognized as user body by MediaPipe pose tracking.

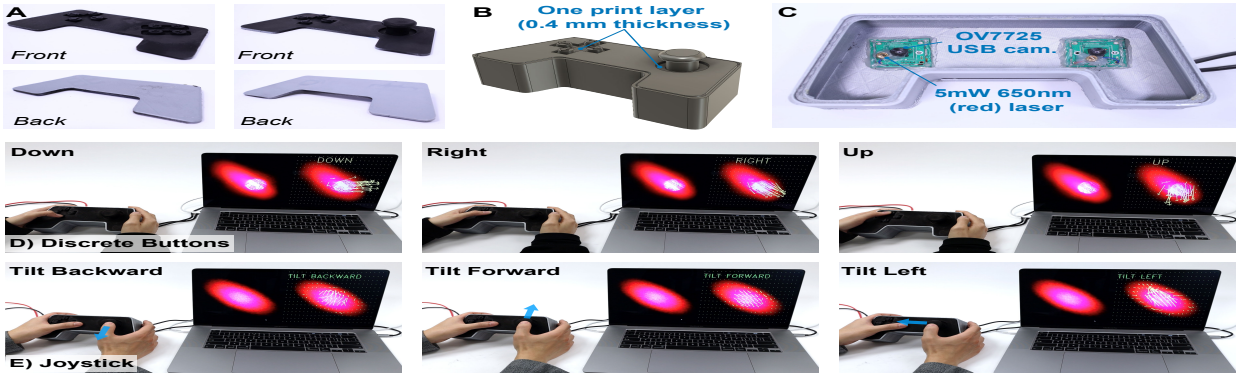


Figure 3.12: Interactive 3D prints using embedded *ForceSight* systems. A: Two designs of thin top plates that can transform user interactions into discernable plate deformations. B: 3D models of a controller. C: Two low-cost lite *ForceSight* bundles are embedded inside the controller. The rest of the figure shows live detection results of user interactions featuring discrete buttons and the joystick.

3.6.3 Force-based Material/Object Identification

Material identification has shown practical uses in HCI, as prior works demonstrated ID-enabled interactions [94] and material-aware laser cutting [95]. We notice that different materials exhibit distinguishable deformations in response to force due to variance in density and internal microstructures. For example, hard materials (e.g., wood) have a wider and shallow "footprint" whereas soft materials (e.g., silicone) deform locally around the force point resulting in a narrow and deep "footprint". The footprint geometry reveals much information about materials.

Another approach is to use regression model parameters as classifier features, which essentially convey Young's Modulus and the moment of inertia. Fig. 3.13 shows differences in parameters learned from our supplemental study *A Wide Array of Materials*, which can be leveraged for identification. We believe this force-based material identification can have broader applications in digital fabrications (e.g., water jetting) as well as object handling (i.e., robot arms can apply less amount of force when handling delicate materials).

3.6.4 Force-Aware Object Manipulation

Handling delicate objects requires force-sensitive mechanisms. Conventional methods rely on contact-based force sensors on robot arms. *ForceSight* creates a different approach to

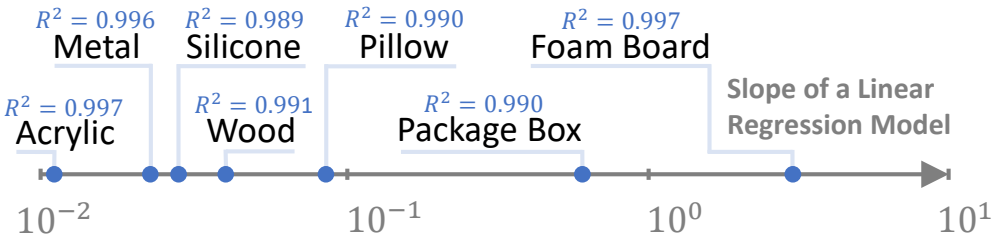


Figure 3.13: *ForceSight* builds a distinctive set of linear regression models for different materials/objects with high R^2. Coefficients of these models can in turn reveal the material type if the applied force is known, enabling material identification for richer applications.

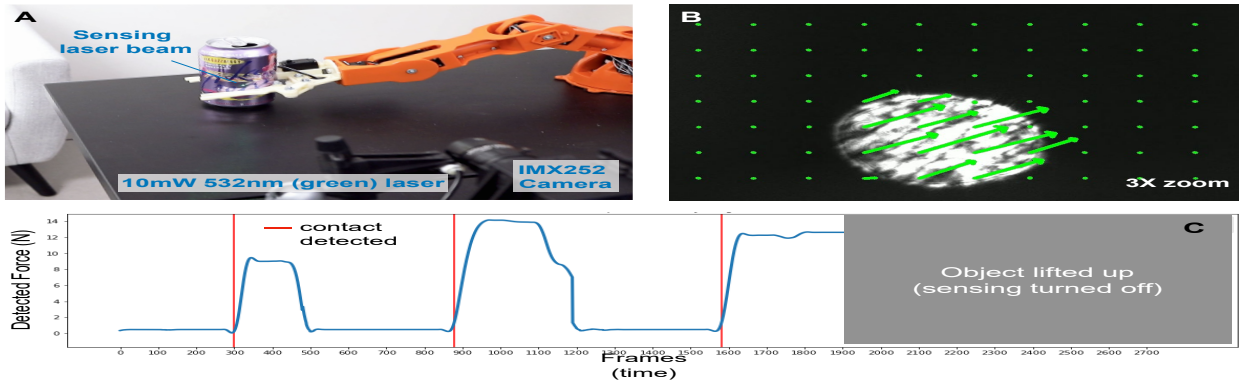


Figure 3.14: Remote force sensing for delicate object handling. A: Robotic arm grasps a soda can sequentially with three different forces – light, strong, and medium. B: Integrated Laser Speckle Velocity. C: Force detected by *ForceSight*.

facilitate remote sensing which can potentially turn into centralized sensing in which one sensor can serve multiple robot arms under its field of view (akin to the sensing scheme of security cameras). Fig. 3.14 shows *ForceSight* working with a focused 10 mW laser and a low-cost robot arm (Arduino Braccio) to sense the grasping force on a soda can as a test primitive. ILSV is shown in Fig. 3.14 B. Once the force reaches the desirable amount, the robot arm starts lifting up the object (Fig. 3.14 C).

3.7 Discussion

Laser safety The strongest laser used in *ForceSight* is 100 mW (Class III B) which is by itself hazardous for eye exposure. However, it is only used in diverged settings with wide divergence achieved by using three concave lenses concatenating with a diffusing glass. The divergence significantly shortens the Nominal Ocular Hazard Distance (NOHD) [127]. At our divergence (79.6 degrees), the NOHD is 5.09 cm. To further improve the safety of users, *ForceSight* could work with other sensing modalities such as RGB cameras and depth sensing – the laser can be turned off once users are too close. *ForceSight* could also use low-power guarding lasers [128] or deploy it at high installation/vantage locations, e.g., ceilings, to

improve safety.

Laser power and color During experiments and application developments, we used and tested the feasibility of a wide array of laser power levels (10, 20, 30, 50 mW) and colors (green, red). In this work, we predominantly demonstrated visible green lasers for ease of development and troubleshooting. In real-world applications, invisible infrared lasers can be used to minimize intrusiveness.

Different types of cameras Additionally, we tested a wide variety of cameras including the IDS Imaging U3-3060CP, ELP 5.0 megapixel, and 2.0 megapixel USB Camera. We found the high camera frame rate to be an important factor in capturing clearer speckles that are easier to track. Low-frame-rate cameras can be used for slower applications of force. To track sudden applications of force with low-frame-rate cameras, we can also use blur detection, which is commonly adopted for laser speckle contrast imaging in clinical applications. Even though blur detection focuses more on the presence of force, it still can enable use cases such as on-world touch segmentation.

Open source We open source our algorithms and dataset to facilitate others' use of *ForceSight*. We hope the joint force behind this technique could further advance it and enable an even more diverse set of applications with practical uses. The source code and data are available at `https://github.com/forcesight/ForceSight`.

3.8 Limitation

ForceSight has two main limitations which we plan to work on in future work. These limitations are around the compatibility of materials, and sensing range & resolution.

First, *ForceSight* works with many everyday surfaces with a few exceptions – plastically deformable materials, discontinuous materials, very stiff materials, and transparent materials. To begin with, *ForceSight* requires deformation delivery. Plastically deformable materials, e.g., Play-Doh, cannot transfer the deformation from the contact point to its sur-

roundings. Second, discontinuous materials like fur and polar fleece could not work with *ForceSight*, because the force applied at one point will not be passed on to its surrounding regions. Third, *ForceSight* cannot work with stiff surfaces that are too hard to deform, e.g., a thick wood table, or concrete floor. Finally, *ForceSight* does not work with transparent surfaces. Laser beams pass through them, generating extremely dim speckles beyond the sensitivity of our system.

We also plan to optimize *ForceSight* for 1) extreme large forces (e.g., car parking on the driveway) and 2) high sensing resolution (e.g., coin on the table). Achieving these requires us to have cameras with better performance (e.g., faster speed, denser pixels on the CCD sensor) and force meters that can provide more fine-grained data in future work.

3.9 Conclusion

We present *ForceSight*, a non-contact force sensing technique using laser speckle imaging. We derived models for both the formation and motion of laser speckles induced by the deformation of rough surfaces at force. We developed and evaluated our system with a series of tests featuring different materials, sensing distances, as well as calibration methods. Results indicate the high accuracy of *ForceSight* across test settings. We conclude the chapter with four applications showcasing the strength of *ForceSight* in different use cases. Overall, we believe *ForceSight* opens up new force sensing opportunities and novel interaction modalities, which could be readily integrated into many real-world applications and future computing systems.

CHAPTER 4

When Collecting Data at Scale is Infeasible: Generating Physiologically Realistic Synthetic Humans

4.1 Introduction

Traditional remote photoplethysmography (rPPG) methods either use Blind Source Separation (BSS) [29, 135, 28] or models based on skin reflectance [27, 26, 136] to separate out the pulse signal from the color changes on the face. These methods usually require pre-processing such as face tracking, registration and skin segmentation. More recently, deep learning and convolutional neural networks (CNN) have been more popular due to its expressiveness and flexibility [30, 31, 137, 47, 138, 139]. CNNs learn the mapping between the pulse signal and the color variations with end-to-end supervised training on the labeled dataset, thus achieving state-of-the-art performance on the vital sign detection. However, the performance of data-driven rPPG networks hinges on the quality of the dataset [32].

There are some efforts (as shown in Tab. 4.1) on collecting a large rPPG dataset for better physiological measurement. Nonetheless, there exists several practical constraints towards collecting real patient data for medical purposes. These include: (1) demographic biases (such as race biases) in society that translate to data. As pointed out in [1], a diverse rPPG dataset may not be accessible for some countries/regions due to geographical distribution of skin colors as reflected in their skin tone world map for indigenous people. (2) necessity of intrusive/semi-intrusive traditional methods for collection of data, (3) patient privacy concerns, and (4) requirement of medical-grade sensors to generate the data. Hence, there is

Dataset	# Subjects	# Videos	Demo. diversity	Orig. Videos Free Avail.
AFRL [129]	25	300	✗	✓
MMSE-HR [130]	40	102	✗	✗
UBFC-rPPG [131]	42	42	✗	✓
UBFC-Phys [132]	56	168	✗	✓
VIPL-HR [133]	107	3130	✗	✓
Dasari *et al.* [134]	140	140	✗	✗
Our synthetic method	480	480	High	✓

Table 4.1: **Comparison of rPPG real datasets and our proposed synthetic dataset.** Real datasets are limited by the number of subjects and videos and demographic diversity, while synthetic datasets have easy control of these attributes.

a pressing need for the concept of 'digital patients': physiologically accurate graphical renders that may assist development of algorithms and techniques for improvement of diagnostics and healthcare. We provide such a neural rendering instantiation in the rPPG field.

For decades, computer graphics has been a driving force for the visuals we see in movies and games. Imagine if we could harness computer graphics techniques to create not just photorealistic humans, but *physio-realistic* humans. We combine modalities of image and waveform to learn to generate a realistic video that can reflect underlying BVP variations as specified by the input waveform. We achieve this by an interpretable manipulation of UV albedo map obtained from the 3D Morphable Face Model (3DMM) [140]. Our model can generate rPPG videos with large variation of various attributes such as facial appearance and expression, head motions and environmental lighting as shown in Fig. 4.1.

4.1.1 Contributions

We summarize our contributions as follows:

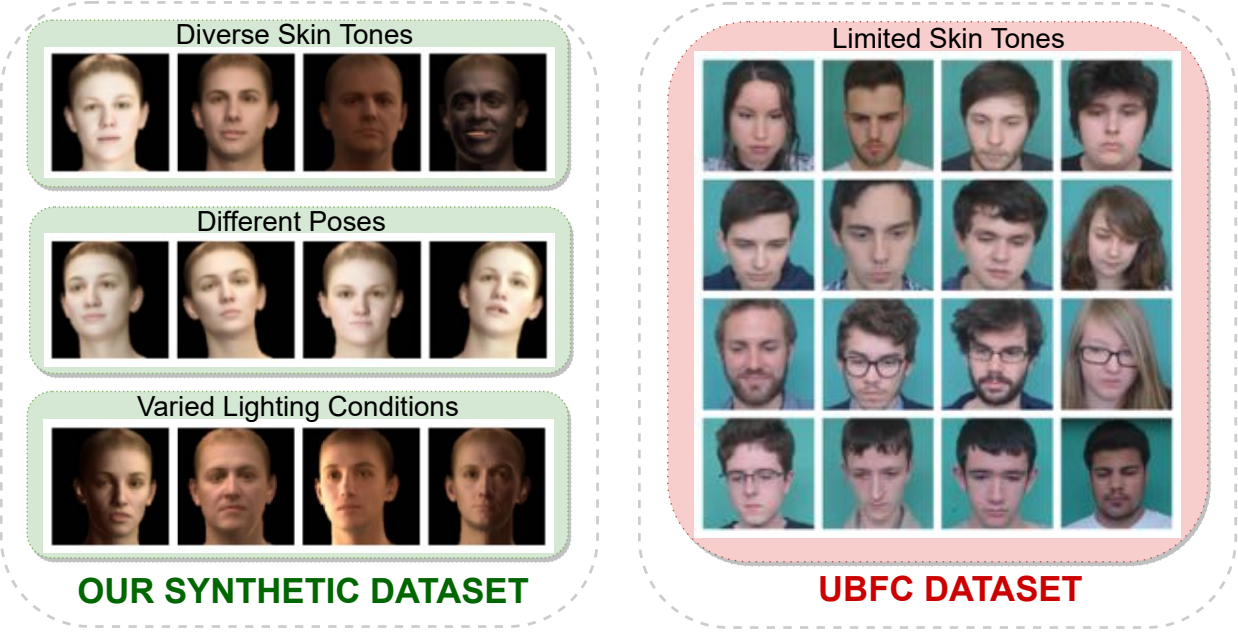


Figure 4.1: **Our proposed scalable model can generate synthetic rPPG videos with diverse attributes such as poses, skin tones and lighting conditions.** In contrast, existing real datasets (e.g. UBFC) only contain limited races.

- We propose a scalable physics-based learning model that can render realistic rPPG videos with high fidelity with respect to underlying blood volume variations.
- The synthetically generated videos can be directly utilized to improve the performance of the state-of-the-art deep rPPG methods. Notably, the corresponding rendering model can also be deployed to generate data for underrepresented groups, which provides an effective method to further mitigate the demographic bias in rPPG frameworks.
- To facilitate the rPPG research, we release a real rPPG dataset called UCLA-rPPG that contains diverse skin tones. This dataset can be used to benchmark performance across different demographic groups in this area.

4.2 Related Work

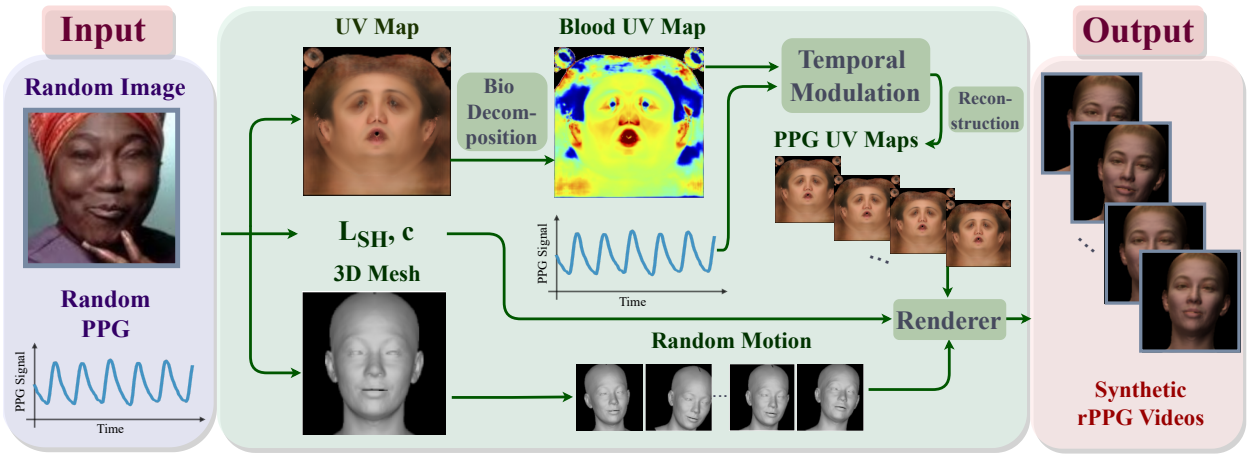


Figure 4.2: **Pipeline of our cross-modal synthetic generation model that can generate rPPG face videos given any face image and target rPPG signal as input.** The input image is encoded into UV albedo map, 3D mesh, illumination model L_{SH} and camera model c. We then decompose the UV albedo map into blood map, vary the UV blood map according to the target rPPG signal and generate the modified PPG UV maps. The modified PPG UV map that contains the target pulse signal variation is combined with L_{SH}, c to render the final frames with randomized motion.

rPPG methods: rPPG techniques aim to recover the blood volume change in the skin that is synchronous with the heart rate from the subtle color variations captured by a camera. Signal decomposition methods include [28] that utilizes Principal Component Analysis (PCA) on the raw traces and chooses the decomposed signal with the largest variance as the pulse signals and Independent Component Analysis (ICA) [29, 141] that demixes the raw signals and determines the separated signals with largest periodicity as the pulse. PCA and ICA are purely statistical approaches that do not use any prior information unique to rPPG problems. A chrominance-based method (CHROM) [26] is proposed to extract the blood volume pulse by assuming a standardized skin-color to white-balance the image and then lin-

early combine the chrominance signals. Plane Orthogonal to Skin-tone (POS) [27] projects the temporally normalized raw traces onto a plane that is orthogonal to the light intensity change, thus canceling out the effect of that. CNNs have achieved state-of-the-art results on vital sign detection due to their flexibility [30, 31, 137, 47, 138, 139, 1]. The representation for rPPG estimation can be efficiently learned in an end-to-end manner with the annotated datasets instead of handcrafted features for traditional methods. We use two representative work PhysNet [31] and PRN [1] in our experiments to demonstrate the performance of the rPPG models on both real and synthetic datasets.

Real rPPG datasets: There are many efforts on collecting real datasets for more accurate physiological sensing [129, 130, 131, 132, 133, 134]. However, these datasets are usually very limited in the number of subject participants and also inequitable towards certain demographic group. Some work includes subject with darker skin types, but the number is still very limited [130]. Making machine learning methods equitable is of increasing interest in medical domain [142, 143]. There is a lack of a benchmark dataset to measure the performance of various rPPG methods on diverse skin tones, especially dark skin tones in rPPG area. Dasari *et al.* [134] proposed a dataset that only contains dark skin tones. However, the actual videos are not shared but the color space values of skin region of interest. The current best-performing deep learning algorithms require sizeable input data. The rPPG model trained on such an inequitable dataset may easily disadvantage certain underrepresented groups in the dataset. The lack of such a benchmark dataset to systematically and rigorously evaluate various methods on diverse skin tones makes it hard to ensure that the rPPG methods deployed into the society would not cause inequities against certain groups that are underrepresented. Our real dataset represents a first step towards filling this gap.

Synthetic generation of rPPG videos: The real rPPG dataset construction is a laborious process and generally takes a large amount of time for collection and administrative work for Institutional Review Board (IRB) approval. Therefore, it is tempting to have a

scalable method that can generate large-scale synthetic rPPG datasets for data augmentation. Realizing the difficulty of this, there are a few groups working on generating synthetic rPPG facial videos to augment real data [144, 145, 1, 146]. Mcduff *et al.* [144] propose to render rPPG face videos using facial avatars and simulate the blood volume change with Blender. However, as discussed in the limitation of their method, the rendering of a frame is extremely slow (20 seconds per frame), thus preventing synthetic generation of large-scale videos. The initial overhead for creating the pipeline is also expensive and labor-intensive. A skin tone augmentation method is proposed in [1] where they use a generative neural network to transfer light skin tones to dark skin tones while retaining the pulsatile signals so that the performance on dark skin tones can be improved with the augmented dataset more balanced. Like the other augmentation method on rPPG signals [145], they are both limited as they can only be utilized on current datasets and have to be retrained with new datasets. In contrast, our synthetic generation method can generate diverse appearance with any in-the-wild image and target rPPG signal as input and the generation is merely a forward pass of the neural network.

4.3 Methods

In this section, we propose a scalable method that can generate synthetic dataset with any given reference image and target rPPG signal in Sec. 4.3.1. The generated videos can be used to train the state-of-the-art rPPG networks, which we introduce in Sec. 4.3.2.

4.3.1 Synthesizing Biorealistic Face Videos

We first describe the 3DMM model used to obtain the facial albedo maps and then demonstrate how to further obtain facial blood maps from the extracted albedo by analyzing light transport in the skin. Details about how to generate synthetic facial videos with the decomposed blood maps and the source of the input facial images and PPG waveforms are

also provided in this section. Please see Fig. 4.2 for an illustration of the entire synthetic generation pipeline.

Non-linear 3DMM: To generate faces with different poses, illuminations and desirable rPPG signal variations, we have to infer the 3D shape and albedo parameters of the face. We use DECA [140] to predict subject-specific albedo, shape, pose, and lighting parameters from an image. In details, it uses a statistical 3D head model FLAME [147] to output a mesh M with $n = 5023$ vertices. The camera model $\mathbf{c}$ is learned to map the mesh M to image space. Since there is no appearance model in FLAME, the linear albedo subspace of Basel Face Model (BFM) [148] is used and the UV layout of BFM is converted to be compatible with FLAME. It outputs a UV albedo map A with a learnable coefficient $\boldsymbol{\alpha}$. By expressing illumination model as the Spherical Harmonics (SH) [149], the shaded face image can be represented as the following equation:

$$B\left(\boldsymbol{\alpha}, \mathbf{l}, N_{uv}\right)_{i,j} = A(\boldsymbol{\alpha})_{i,j} \odot \sum_{k=1}^{9} \mathbf{l}_k H_k\left(N_{i,j}\right), \tag{4.1}$$

where H_k is the SH basis, $\mathbf{l}_k$ are the corresponding coefficients and $\odot$ denotes the Hadamard product. $N_{i,j}$ is the normal map expressed in the UV form. The final texture image is obtained by rendering the image using the mesh M, shaded image B, and the camera model $\mathbf{c}$ through a rendering function $\mathcal{R}(\cdot)$:

$$I_r = \mathcal{R}(M, B, \mathbf{c}). \tag{4.2}$$

As rPPG is essentially the change of blood volume in the face, our idea is to first obtain the spatial concentration of blood f_{blood} of the UV albedo A and then temporally modulate the UV blood albedo map in a way that is consistent with the rPPG signals. We will next show how this biophysically interpretable manipulation is achieved.

Light transport in the skin: In order to obtain blood map f_{blood} on the face, we first study light transport in the skin to build the connection between face albedo and f_{blood}. Fol-

lowing a spectral image formation model, the original UV face albedo A_c with $c \in \{R, G, B\}$ is reconstructed by integrating the product of the camera spectral sensitivities S_c, the spectral reflectance R, and the spectral power distribution of the illuminant E over wavelength λ [150]:

$$A_c = \int_{\lambda} E(\lambda) R(f_{\text{mel}}\,, f_{\text{blood}}\,, \lambda) S_c(\lambda) d\lambda. \tag{4.3}$$

An optical skin reflectance model [151] with hemoglobin f_{blood} and melanin map f_{mel} as parameters is utilized to define the wavelength-dependent skin reflectance $R(f_{\text{mel}}\,, f_{\text{blood}}\,, \lambda)$. Specifically, we assume a two-layer skin model that characterizes the transmission through the epidermis $T_{\text{epidermis}}$ and reflection from the dermis R_{dermis} :

$$R\left(f_{\text{mel}}, f_{\text{blood}}, \lambda\right) = T_{\text{epidermis}}\ \left(f_{\text{mel}}, \lambda\right)^2 R_{\text{dermis}}\ \left(f_{\text{blood}}, \lambda\right). \tag{4.4}$$

The transmittance in epidermis is modeled by Lambert-Beer law [152] as light not absorbed by the melanin in this layer is propagated to the dermis [72]:

$$T_{\text{epidermis}}(f_{\text{mel}}, \lambda) = e^{-\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda)}, \tag{4.5}$$

where $\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda)$ is the absorption coefficient of the epidermis. More specifically,

$$\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda) = f_{\text{mel}} \mu_{a.\text{mel}}(\lambda) + (1 - f_{\text{mel}}) \mu_{\text{skinbaseline}}(\lambda), \tag{4.6}$$

where $\mu_{a.\text{mel}}$ is the absorption coefficient of melanin and $\mu_{\text{skinbaseline}}$ is baseline skin absorption coefficient.

The reflectance in dermis can be modeled using the Kubelka-Munk theory [153], and the proportion of light remitted from a layer is given by [72]:

$$R_{\text{dermis}}\ (f_{\text{blood}}, \lambda) = \frac{(1 - \beta^2)\left(e^{K d_{\text{pd}}} - e^{-K d_{\text{pd}}}\right)}{(1 + \beta^2)\, e^{K d_{\text{pd}}} - (1 - \beta)^2 e^{-K d_{\text{pd}}}}, \tag{4.7}$$

where d_{pd} is the thickness of the dermis, and K and β are related to the absorption of the medium contained within the dermis (i.e. blood). For simplicity of notation, we drop the dependence of K and β on f_{blood} and λ in Eq. (4.7).

Biophysical decomposition and variation of UV albedo map: With the light transport theory of the skin, we follow a physics-based learning framework (BioFaceNet [150]) to obtain f_{blood} from albedo A. The wavelengths are discretized into 33 parts from 400nm to 720nm with 10nm equal spacing. We utilize an autoencoder architecture and use a fully-convolutional network as encoder to predict the hemoglobin and melanin maps and fully-connected networks to encode the parameters for lighting E and camera spectral sensitivities S_c. The model-based decoder is then to reconstruct the albedo with all the learned parameters according to Eq. (4.3).

Different from the previous work [150], we obtain biophysical parameters directly from the UV albedo maps instead of the facial images. This arrangement allows us to model the underlying blood volume changes more precisely regardless of the environmental illumination variations. Our model is trained to minimize the following loss function:

$$\mathcal{L} = w_1 \mathcal{L}_{\text{appearance}} + w_2 \mathcal{L}_{\text{CameraPrior}}, \tag{4.8}$$

where the appearance loss $\mathcal{L}_{\text{appearance}}$ is the $L2$ distance between the reconstructed UV map A_{linRecon} and the original one in the linear RGB space A_{linRGB}. We convert A to linear space by inverting the Gamma transformation with $\gamma = 2.2$. To make the problem more constrained, we also introduce the additional camera prior loss: $\mathcal{L}_{\text{CameraPrior}} = \|\mathbf{b}\|_2^2$, where $\mathbf{b}$ is the prior for the camera spectral sensitivities. w_1 and w_2 are the weights for the reconstructed loss and camera prior loss, respectively.

To reflect the change of the target rPPG signal on the face, we temporally vary the UV blood map f_{blood} linearly with the target rPPG signal in the test phase. Given the blood map of a reference UV map (e.g. the UV blood map of first frame), we generate the UV blood map of the consequent frames as the multiplication of the UV blood map of the reference frame and a ratio scalar that is calculated as the ratio of p_t (rPPG signal at time t) and p_{ref} (rPPG signal at the reference time). Then the modified UV blood map of each frame that contains the desired rPPG signal is reconstructed using the BioFaceNet decoder to get UV

map. The final image is rendered using the UV map combined with illumination and camera model according to Eq. (4.2).

For the purpose of simulating real-world scenarios where the subject might move in the collection process, we randomize the poses in the generation of the sequence of the frames by adding a small random value to the pose and expression parameter of the previous frame.

Face image dataset: To generate synthetic rPPG videos with diverse face appearances, we use the public in-the-wild face datasets BUPT-Balancedface [154]. It is categorized according to ethnicity (i.e. Caucasian, Indian, Asian and African). We use these images as the reference images for generating the synthetic videos as shown in Fig. 4.2.

PPG recordings: To synthesize videos of a given input PPG signal, we use PPG waveforms recordings from BIDMC PPG and Respiration Dataset [155]. It contains 53 8-minute contact PPG recordings with sampling frequency 125Hz. We sample it correspondingly with the video frame rate (30Hz) and the first sequences of time length L are used where L is the duration of the generated video.

4.3.2 Physiological Measurement Networks

We use two state-of-the-art deep rPPG networks PhysNet [31] and PRN [1] to benchmark the performance on both real and synthetic datasets. PhysNet and PRN both utilize 3D convolutional neural networks (3D-CNN) architecture to learn spatio-temporal representation of the rPPG videos and predict the rPPG signal in the facial videos. PRN differs in that it uses residual connection for convolutional layers. They take consecutive frames of length T as the input, and its output is the corresponding BVP value for each input frame. The Negative Pearson loss is used to measure the difference between the ground-truth PPG signal p and the estimated rPPG signal $\hat{p}$:

$$L_{ppg}(p,\hat{p}) = 1- \frac{T\sum_i p_i\hat{p}_i - \sum_i p_i \sum_i \hat{p}_i}{\sqrt{\left(T\sum_i p_i^2 - \left(\sum_i p_i\right)^2\right)\left(T\sum_i \hat{p}_i^2 - \left(\sum_i \hat{p}_i\right)^2\right)}}, \tag{4.9}$$

where all the summation is over the length of frames T.

Implementation details: For the training of BioFaceNet, we use 3000 face albedo images with 750 images in each race. We use 80% images for training and 20% for validation. The weight w_1 and w_2 for the loss is $1e^{-3}$ and $1e^{-4}$ respectively. The learning rate is set as $1e^{-4}$ and the number of epochs is 200. For the generation of synthetic videos, we set the length of generated frames L as 2100.

The bounding boxes of the videos are generated using a pretrained Haar cascade face detection model. For each video, one bounding box is detected and increased 60% in each direction before the frames are cropped. To be consistent with the original works, each frame is resized to 128×128 pixels using bilinear interpolation for PhysNet and 80×80 for PRN. The length of training clips T is 128 for PhysNet and 256 for PRN. The Adam optimizer is used and the learning rate is set as $1e^{-4}$. All the code is implemented in PyTorch [156] and trained on Nvidia V100 GPU.

4.4 Experiments

In this section, we introduce the datasets we use for the experiments and evaluation protocol in Sec. 4.4.1. We report and analyze the experimental results for our real dataset in Sec. 4.4.2 and UBFC-rPPG dataset in Sec. 4.4.3.

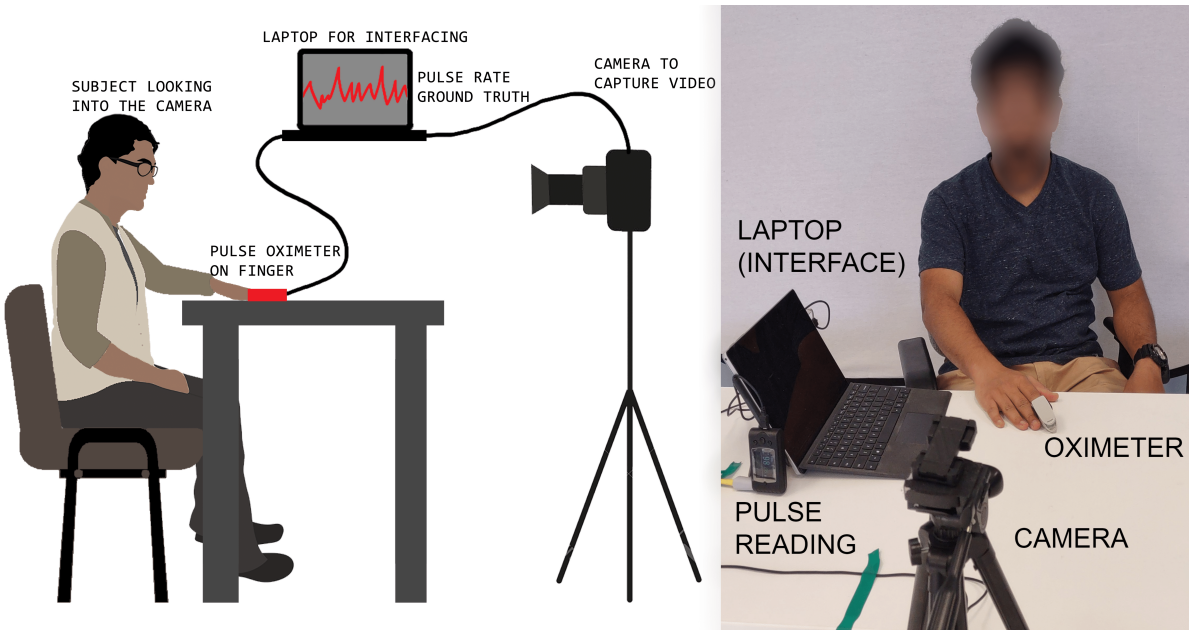


Figure 4.3: **Experimental setup of data collection.** The subject wears an oximeter on their finger and sits looking directly into the camera. The camera and the oximeter are connected to a laptop to get synchronous video and ground-truth pulse reading. Face blurred to preserve anonymity.

4.4.1 Datasets and Evaluation Protocol

Our real dataset UCLA-rPPG: In order to benchmark the performance of current rPPG estimation methods, we collect a real dataset of 104 subjects. The setting is faulty for two of them so we dropped their samples. Finally, the dataset consists of 102 subjects of various skin tone, age, gender, ethnicity and race. The Fitzpatrick (FP) skin type scale [157] of the subjects varies from 1-6. For each subject, we record 5 videos of about 1 minute each (1790 frames at 30fps). After removing erroneous videos we have total 503 videos. All the videos in our dataset are uncompressed and synchronized with the ground truth heart rate.

Fig. 4.3 illustrates the data collection process of our real dataset UCLA-rPPG. The left part of the figure is a cartoon illustration of the data collection process. The right part of the figure is a photo depicting the actual data collection process. The human subjects wear

an oximeter on finger and looks into the camera. Both the camera and the oximeter are connected to a laptop to get synchronous data.

UBFC-rPPG [131]: UBFC-rPPG database contains 42 front facing videos of 42 subjects and corresponding ground truth PPG data recorded from a pulse oximeter. The videos are recorded at 30 frames per second with a resolution of 640×480. Each video is roughly one minute long.

Metrics: To evaluate how the heart rate estimates compare with gold-standard heart rates obtained from gold-standard pulse waves, we use the following four metrics Mean absolute error (MAE), Root Mean Squared Error (RMSE), PearsonâĂŹs Correlation Coefficient (PCC) and Signal-to-Noise Ratio (SNR). PearsonâĂŹs Correlation Coefficient (PCC) and Signal-to-Noise Ratio (SNR) is defined as in [158].

For traditional baseline methods POS, CHROM and ICA we compare, we use iPhys toolbox [159] to get the estimated rPPG waveforms. The output rPPG signals are normalized by subtracting the mean and dividing by the standard deviation. We filter all the model outputs using a 6th-order Butterworth filter with cut-off frequencies 0.7 and 2.5 Hz. The filtered signals are divided into 30-second windows with 1-second stride and the above four evaluation metrics are calculated on these windows and averaged.

4.4.2 Performance on UCLA-rPPG

For the study of this work, we split the subjects into three skin tone groups based on the Fitzpatrick skin type [157]. They are light skin tones, consisting of skin tones in the FP 1 and 2 scales, medium skin tones, consisting of skin tones in the FP 3 and 4 scales, and dark skin tones, consisting of skin tones in the FP 5 and 6 scales. This aggregation helps compare experimental results on skin tones more objectively. Since our ultimate goal is to improve the performance on our dataset, we first train on all the synthetic data and then finetune

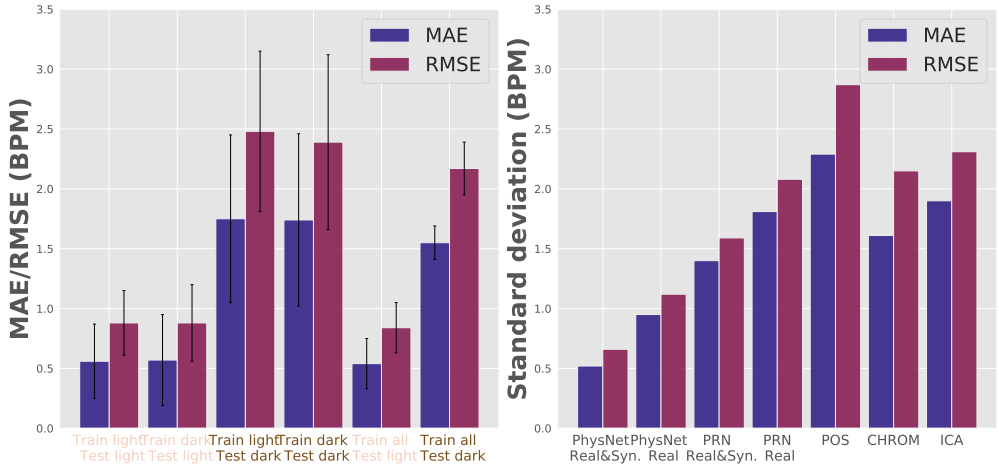


Figure 4.4: **Left: Ablation study.** The model pre-trained with all synthetic dataset outperforms these pre-trained on either light or dark skin tones alone. **Right: Inequity mitigation.** The standard deviation of MAE and RMSE of the deep rPPG models trained with real and synthetic dataset are smaller than real data alone and the traditional models.

on the real data for the models trained with both real and synthetic data. For training and testing deep rPPG networks PhysNet and PRN on real dataset, we randomly split all the subjects into training, validation and test set with 50%, 10% and 40% and all the test results are averaged on three random splits. The validation set is used to select the best epoch for testing the model.

We report results on the three groups and overall performance using evaluation metrics of MAE, RMSE, PCC and SNR in Tab. 4.2. In general, models trained with both real and synthetic data perform consistently better than using real data alone on all the skin tones for all evaluation metrics. PhysNet trained with both real and synthetic data achieved the best overall MAE result 0.71 BPM, with 33% reduction in error compared with PhysNet trained with only real data (1.06 BPM). Notably, the performance improvement is most significant on dark skin stones F5-6 group with 41% and 35% reduction in MAE and RMSE respectively for PhysNet. The same phenomenon is also observed for PRN, where the improvement is most noticeable for darker skin tones. We attribute this to the introduction of synthetic

videos we generate in Sec. 4.3.1. The other two metrics PCC and SNR also validate the superiority of the model trained with both real and synthetic datasets. The results for traditional methods POS, CHROM and ICA are far worse than the deep learning methods, as these methods usually takes the average of all the pixels and ignore the inhomogeneous spatial contribution of the pixels to pulsatile signals.

Inequity mitigation: To evaluate the inequity of various rPPG methods on subjects with diverse skin tones, we use the standard deviation of the MAE and RMSE results on three skin tone groups. From the right of Fig. 4.4, we can see the standard deviation of PhysNet with both real and synthetic dataset is the smallest and the MAE disparity among all the three groups are reduced by 45% (from 0.95 BPM to 0.52 BPM) compared with the model trained with only real dataset. Similarly, the standard deviations of both metrics MAE and RMSE for PRN are also reduced for the model trained with both real and synthetic datasets.

Ablation study: We first pre-train the PhysNet with either light skin tones (subjects with race Caucasian in the synthetic dataset) or dark skin tones (subjects with race African), then finetune the model on real dataset and test the model on real subjects with either light skin tones or dark skin tones. From the left of Fig. 4.4, we can see the model with the pre-trained rPPG network on diverse races are consistently better than these on a single race. The improvement is more obvious on dark skin tones test set. This demonstrates the benefits of a diverse synthetic dataset.

4.4.3 Performance on UBFC-rPPG

We use the model with best performance on our real dataset to test them on UBFC-rPPG dataset [131] along with the traditional methods. Since this is a cross-dataset evaluation for the model trained on UCLA-rPPG, we test the deep learning models on all the subjects in

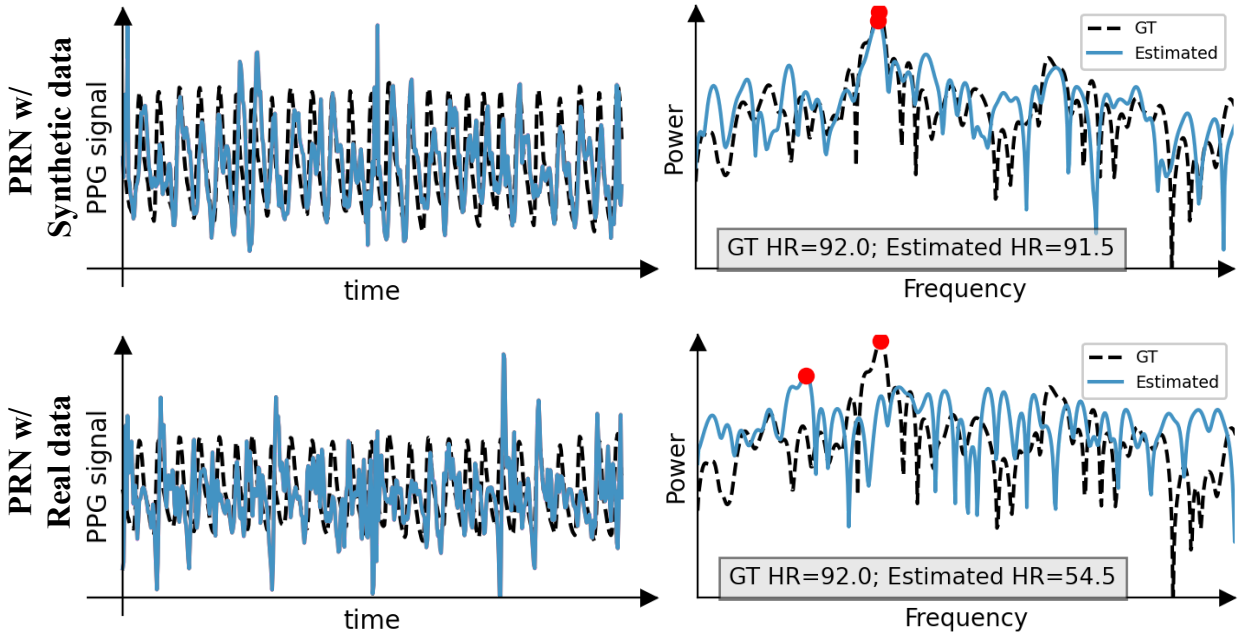


Figure 4.5: **The example shows that PRN [1] trained with synthetic data (above) generalizes better than PRN trained with real data (bottom) on UBFC-rPPG dataset.** The waves are more aligned with the ground-truth PPG wave (dashed black line) and the power spectrum plot is also more consistent with the ground-truth for the PRN trained with synthetic data.

UBFC-rPPG. All the results with four evaluation metrics are reported in Tab. 4.3. While the synthetic dataset performs worse than the models trained in our real dataset, the performance gain is more obvious in UBFC dataset. The MAE of PhysNet trained on synthetic dataset achieved the lowest MAE and RMSE (0.84 BPM and 1.76 BPM respectively). The explanation for this observation is that when the distribution of the dataset is similar to the distribution of the test data as in the intra-dataset setting in our real dataset, the benefits of synthetic datasets are not straightforward. The models trained on real dataset perform worse on generalizing to another dataset due to different environmental setting such as lighting. We also give a qualitative study in Fig. 4.5 that shows that the rPPG wave extracted using our synthetic dataset resemble more closely to the ground-truth than that using real

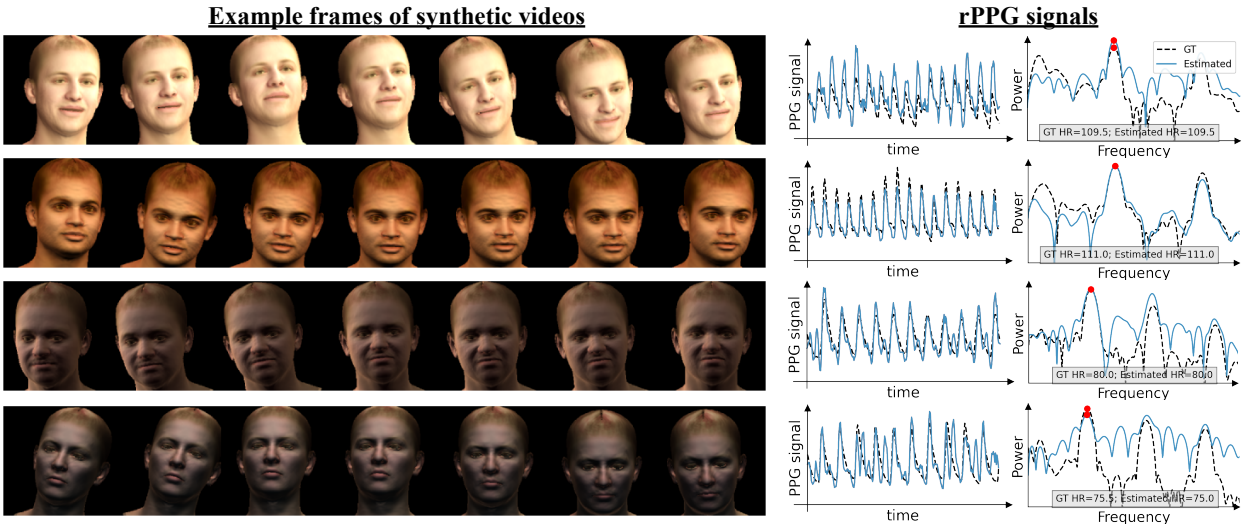


Figure 4.6: **Illustration of example frames of our generated synthetic videos.** Our proposed framework has successfully incorporated PPG signals into the reference image. The estimated pulse waves from PRN for generated synthetic videos are highly correlated to the ground-truth waves, and the heart rates are preserved as shown in the power spectrum plot.

dataset. As a result, it gives more accurate heart rate estimation.

4.4.4 Visualization

As shown in Fig. 4.6, our model can successfully produce synthetic avatar videos that reflect the associated underlying blood volume changes. Estimated pulse waves from the synthetic videos are closely aligned with the ground truth. The power spectrum of the PPG waves with a clear peak near the gold-standard HR value also validates the effectiveness of the incorporation of pulsatile signals.

4.5 Discussion

Limitations: Though our synthetic dataset could be used to achieve state-of-the-art results (on UBFC-rPPG datasets, it alone can generalize even better than the model trained

on real dataset) for heart rate estimation, the facial appearance is not photo-realistic, which may still degrade the performance due to sim2real gap. We are not focused on modeling the background in the generated videos in this work. However, it is found in [158] that the background can be utilized for better pulsatile signals extraction. Also we vary the UV blood map linearly according to the target rPPG signals in the synthetic generation method. While this yields reasonable empirical results, we believe biophysical model based manipulation of the UV blood map could further improve the performance of the synthetic generation.

Ethics Statement: This work's novelty is to generate synthetic face videos that are physiologically consistent with heartbeat, and we hope it can be a tool to address some social issues, such as inequities around race and gender in medicine. It should also be noted that even though the research here was solely used to improve remote health technologies, it might be used to fool rPPG-based deepfake detectors. We strongly advise against using this technology for such applications.

Conclusion: We propose a method to generate large-scale synthetic rPPG videos with high-fidelity to the underlying rPPG signals. The synthetic generation pipeline enables the scalable generation of rPPG facial videos with any given image and rPPG signal. We validate the effectiveness of the synthetic videos on UCLA-rPPG dataset we collect that contains diverse skin tones and UBFC-rPPG dataset. The experimental results show that the synthetic dataset can improve the performance on both datasets and help reduce the inequities among different demographic groups.

Method	F1-2		F3-4		F5-6		Overall	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
PhysNet [31] w/ Real&Synth	**0.54**	0.84	0.38	0.70	**1.55**	**2.17**	**0.71**	**1.10**
PhysNet [31] w/ Real	0.81	1.21	0.43	0.77	2.61	3.34	1.06	1.51
PhysNet [31] w/ Synth	1.06	1.52	1.16	1.66	4.96	6.20	2.06	2.73
PRN [1] w/ Real&Synth	**0.54**	**0.79**	**0.36**	**0.65**	3.41	4.09	1.15	1.53
PRN [1] w/ Real	0.65	1.02	0.40	0.71	4.35	5.26	1.43	1.90
PRN [1] w/ Synth	1.47	2.00	0.63	1.07	8.89	9.88	2.87	3.47
POS [27]	3.40	4.34	3.03	3.98	8.07	10.23	4.27	5.49
CHROM [26]	4.06	5.11	3.99	5.25	7.45	9.74	4.79	6.22
ICA [29]	3.75	4.73	3.26	4.19	7.51	9.34	4.35	5.50
	F1-2		F3-4		F5-6		Overall	
	PCC ↑	SNR ↑	PCC ↑	SNR ↑	PCC ↑	SNR ↑	PCC ↑	SNR ↑
PhysNet [31] w/ Real&Synth	**0.84**	**14.40**	**0.80**	**17.11**	**0.60**	**9.19**	**0.76**	**14.45**
PhysNet [31] w/ Real	0.81	13.13	0.77	15.83	0.59	6.54	0.74	12.84
PhysNet [31] w/ Synth	0.74	7.19	0.64	6.11	0.23	-3.33	0.57	4.10
PRN [1] w/ Real&Synth	0.81	12.24	0.79	14.61	0.57	4.84	0.74	11.59
PRN [1] w/ Real	0.77	10.73	0.77	13.22	0.48	2.38	0.70	9.91
PRN [1] w/ Synth	0.69	5.14	0.67	5.27	0.21	-5.81	0.56	2.53
POS [27]	0.50	-0.30	0.42	-0.09	0.27	-5.38	0.41	-1.34
CHROM [26]	0.41	-1.81	0.31	-1.60	0.26	-5.31	0.33	-2.49
ICA [29]	0.45	-0.60	0.38	-0.19	0.27	-5.24	0.37	-1.44

Table 4.2: **Heart rate estimation results on our real dataset UCLA-rPPG show that both PhysNet and PRN trained with real and synthetic datasets performs consistently better than the models trained with only real data.** The improved performance shows the benefit of the synthetic video dataset we generate.

Method	MAE ↓	RMSE ↓	PCC ↑	SNR ↑
PhysNet [31] w/ Real&Synth	0.90	1.80	**0.84**	6.28
PhysNet [31] w/ Real	1.42	2.74	0.78	5.64
PhysNet [31] w/ Synth	**0.84**	**1.76**	0.83	**6.70**
PRN [1] w/ Real&Synth	1.15	2.38	0.82	5.36
PRN [1] w/ Real	2.36	4.21	0.66	-1.24
PRN [1] w/ Synth	1.09	1.99	0.83	3.00
POS [27]	3.69	5.31	0.75	3.07
CHROM [26]	1.84	3.40	0.77	4.84
ICA [29]	8.28	9.82	0.55	1.45

Table 4.3: **Performance of HR estimation on UBFC-rPPG shows the superiority of the synthetic datasets.** Boldface font represents the preferred results.

CHAPTER 5

Minority Inclusion for Majority Group Enhancement of AI Performance

5.1 Introduction

Inclusion of minorities in a dataset impacts the performance of artificial intelligence (AI). Recent research has presented the value of inclusive datasets to improve AI performance on minorities and also for society-at-large [160, 58, 161, 162, 163, 164, 165, 166, 143]. A society-at-large consists of both majority and minority stakeholders. However, an objection (often silently posed) to minority inclusion efforts, is that the inclusion of minorities can diminish performance for the majority. This is based on a "rule of thumb" that AI performance is maximized when one trains and tests on the same distribution. A devil's advocate position against minority inclusion might be presented as: "In a fictitious society where we are absolutely certain that only blue-skinned humans will exist in the test set, why include out of distribution orange-skinned humans in the training set?".

In this work, we make the surprising finding that inclusion of minority samples improves AI performance not just for minorities, not just for society-at-large, but *even for majorities*. We refer to this effect as Minority Inclusion, Majority Enhancement (MIME), illustrated in Figure 5.2. Specifically, we note that including some minority samples in the train set improves majority group test performance. However, continued addition of minority samples leads to performance drop. The effect holds under statistical conditions that are represented in traditional computer vision datasets including FairFace [167], UTKFace [5], pets [168],

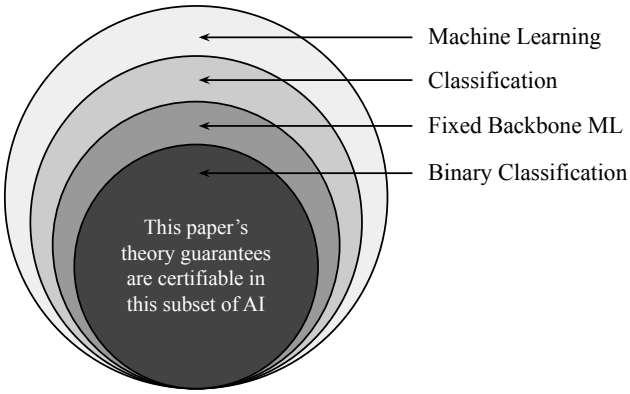


Figure 5.1: **This work proves* that including minorities improves majority performance.** *When do the provable guarantees hold? The guarantees are certifiable for fixed backbone binary classification (e.g. one uses a head network with pretrained weights and fine-tunes a downstream layer for classification). The fixed backbone ML is far from a toy scenario (it is considered SoTA by some authors [2]) and also enables provable certification - ordinarily it is hard to prove things for neural network settings.

medical imaging datasets [169] and even non-vision data [3]. Although deep learning is used for these problems, the flattening layer of a network can be empirically approximated to elementary distributions like Gaussian Mixture Models (GMMs). A GMM facilitates closed-form analysis to prove the existence of the MIME effect. Additionally, we show existence of MIME on general distributions. Classification experiments on neural networks validate using Gaussian mixtures: complex neural networks exhibit feature embeddings in flat layers, distributed with approximately Gaussian density, across six datasets, in and beyond computer vision, and across many realizations and configurations.

Fairness in machine learning is an exceedingly popular area, and our results benefit from several key papers published in recent years. Sample reweighting approaches recognize the need to preferentially weight difficult examples [15, 170, 171]. Active and online learning benefit from insights into sample "informativeness" (i.e. given a budget on the number of training samples, which would be the best sample to include [172, 173]). Domain random-

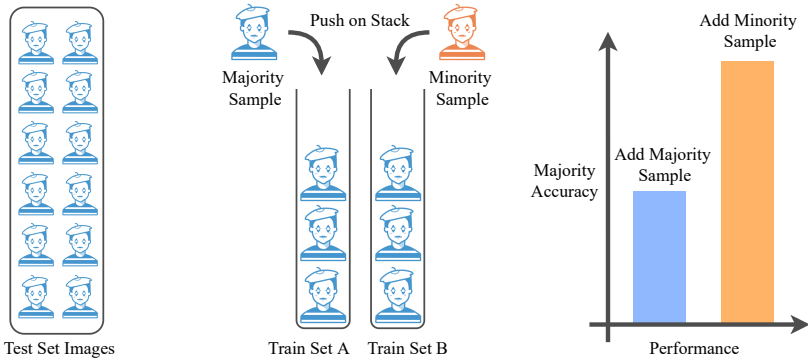


Figure 5.2: **Inclusion of minorities can improve performance for majorities.** We theoretically describe an effect called Minority Inclusion, Majority Enhancement (MIME). The figure depicts test classification of blue mimes, and an initial training stack, also of blue mimes. If allowed to add one more training sample, it can be better to push an orange mime onto the training stack rather than a blue mime. Test accuracy can increase by pushing orange, even though the test set consists of blue mimes alone.

ization literature indicates that surprising perturbations to the training set can improve generalization performance [174, 175, 176]. We extend some of these theoretical insights to the sphere of analyzing benefits of minority inclusion on majority performance.

5.1.1 Contributions

While some works [177, 161] have observed related phenomena for isolated tasks, to the best of our knowledge, characterizing benefits to majority groups by including minority data is largely unexplored theoretically. Our contributions are as follows:

- We introduce the Minority Inclusion Majority Enhancement (MIME) effect in a theoretical and empirical setting.
- Theoretically: we derive in closed form, the existence of the MIME effect both with and without domain gap (Key Results 1 and 2) and for general sample distributions (Key Result 3).

- Empirically: we test the MIME effect on six datasets, as varied as animals to medical images, and observe the existence of MIME consistent with theory.

5.1.2 Outline of Theoretical Scope

Figure 5.1 describes the theoretical scope. Through three key results (Theorem 1, Theorem 2 and Theorem 3), this chapter offers an existence proof of the MIME effect. An existence proof can leverage a tractable setting. As in Figure 5.2, training data is a stack of $K-1$ majority samples. Test data is all majority samples. We can push one additional training sample to increase the stack size to K. We are allowed the choice of having the K-th sample drawn from the minority or majority group. Theorem 1 proves that, under the assumptions in Section 5.3, pushing a minority sample is superior for majority group performance improvements. Theorem 2 generalizes this result to a more realistic scenario, with domain gap. Theorem 3 extends the existence proof to general sample distributions. Empirical results on real-world AI tasks offer validation for theoretical assumptions.

5.2 Related Work

Debiasing and fairness: It has been widely reported that biases in training data lead to inequitable algorithmic performance [56, 178, 58]. Work has been carried out in identifying and quantifying inequities [179, 180, 181] and a range of methods exist to address them [166, 164]. Early approaches suggest oversampling strategies [182, 183]. Other methods propose resampling based on individual performance [163]. Some works utilize information bottlenecks to disentangle inequitable attributes [184]. Still other methods propose inequity mitigation solutions based on adversarial learning [185] or include considerations like protected class-specific classifiers [57]. Generative models have also found use in creating synthetic datasets with debiased attributes [186]. Xu *et al.* [187] identify inherent inequity amplification as a result of adversarial training and propose a framework to mitigate these in-

equities. Our goals are different – while these aim to reduce test time performance inequities across groups, we analyze influence of minority samples on majority group performance.

Learning from multiple domains: Domain adaptation literature explores learning from multiple sources [188]. It could therefore be one potential way to analyze our problem of training on combinations of majority and minority data. In our setting, data arising from distinct domains is seen as being drawn from different distributions with a domain gap [189]. Between these domains, [190] establishes error bounds for learning from combinations of domains. However, these error estimates and bounds do not take into account the notion of majority and minority groups; therefore, describing the MIME effect is outside their scope.

Dataset diversity: An important push towards fairness is through analysis of dataset composition. Several works indicate the importance of diverse datasets [160, 165]. Ryu *et al.* [162] note that class imbalance in the training set leads to performance reduction. Wang *et al.* [181] highlight that perfectly balanced datasets may still not lead to balanced performance. For designing medical devices, [143] emphasizes the importance of diverse datasets. Through experiments on X-ray datasets, [161] observe that imbalanced training sets adversely affect performance on the disadvantaged group. They also observe that an unbiased training set shows the best overall accuracy. However, their inferences are related empirical observations on a few medical tasks and datasets. From an application perspective, the task of remote photoplethysmography enables analysis of the inequity problem. Prior work notes that camera-based heart rate estimation exhibits skin tone inequities [32], and [1, 191] propose synthetic augmentations to mitigate this. Additionally, [192, 8] establish that camera based heart rate estimation is fundamentally inequitable against dark skin tone subjects, establishing a notion of task complexity. While all these works recognize that data composition affects inequity, none to our knowledge describe the effect of varying minority group proportions on majority group accuracy.

5.3 Statistical Origins of the MIME Effect

For more concise exposition, we make assumptions in the derivation in the chapter and defer extended generality to the appendix. Assumptions include:

- Assumption 1: one-dimensional data samples and binary labels, $x \in \mathbb{R}$, $y \in \{1, 2\}$. This is relevant to modern classification problems since the final classification decision is based on a one dimensional projection of the feature representation of the sample with respect to the learnt hyperplane (discussed in Figure 5.1, Section 5.4). Additionally, existence proof of MIME holds for more general vectorized notation, as discussed in the appendix.

- Assumption 2: the binary classifier used is a perceptron: this assumption relates to real neural networks since the last layer is perceptron-like [193].

We now introduce some key definitions that follow from these assumptions.

Definition 1: (Task complexity): *For binary classification we define task complexity for a group of data θ as a continuous variable in $[0, 1]$, such that,*

$$\theta = \arg\min_{h \in H} \epsilon(h), \tag{5.1}$$

where $\epsilon(h)$ is the classification error for hypothesis h (the classifier), H is the space of feasible hypotheses. It is noted later that this is empirically equivalent to distributional overlap. This definition is not new. Hard-sample mining [15] establishes the of use performance measures as an indicator of difficulty.

Definition 2: (Majority Group): *Group class (i.e. group label $g = major$) on which the task performs better. Quantified by training a network only with majority group data and evaluating test performance: $\theta^{major} = \arg\min_{h \in H} \epsilon^{major}(h)$.*

Definition 3: (Minority Group): *Group class (i.e. group label $g = minor$) on which the task performs worse. Quantified by training a network only with minority class data and*

evaluating test performance: $\theta^{minor} = \arg\min_{h \in H} \epsilon^{minor}(h)$.

Definition 4: (Minority Training Ratio (β)): *Ratio of minority to majority samples in the data under consideration (training set, in the context of this work).*

Definition 5: (MIME Domain Gap): *Measure of how classification differs for minorities and majorities. Quantified as a difference between ideal hyperplanes. Note that this definition for domain gap could be different from other definitions. In this work, domain gap should be taken to mean MIME domain gap.*

Empirical observations on cutting-edge machine learning tasks demonstrate the real-world applicability of the assumptions above. We now discuss three key results. For ease of understanding, we make two simplifying assumptions for Key Results 1 and 2: (i) simplified distributions that follow a symmetric Gaussian Mixture Model, and (ii) equally likely class labels, i.e. $Pr(y = 1) = Pr(y = 2)$. These assumptions are relaxed in Key Result 3.

Key Result 1: A minority sample can be more valuable for majority classifiers than another majority sample

Our first key result shows that it can benefit performance on the majority group more if one adds minority data (instead of majority data). Consider a binary classification setting with data samples $x \in \mathbb{R}$ and labels $y \in \{1, 2\}$. Samples from the two classes are drawn from distributions with distinct means:

$$\begin{aligned} x|y = 1 &\sim p_1(x|\mu_1, \sigma_1) \\ x|y = 2 &\sim p_2(x|\mu_2, \sigma_2). \end{aligned} \tag{5.2}$$

Maximum likelihood (ML) can be used to estimate the label as

$$\widehat{y} = \arg\max_{y} \mathcal{L}(x|y). \tag{5.3}$$

An ideal hyperplane for ML $\mathcal{H}_{\text{ideal}}$ is a set of data samples such that:

$$\mathcal{H}_{\text{ideal}} = \big\{ x \;\big|\; \mathcal{L}(x|y = 1) = \mathcal{L}(x|y = 2) \big\}. \tag{5.4}$$

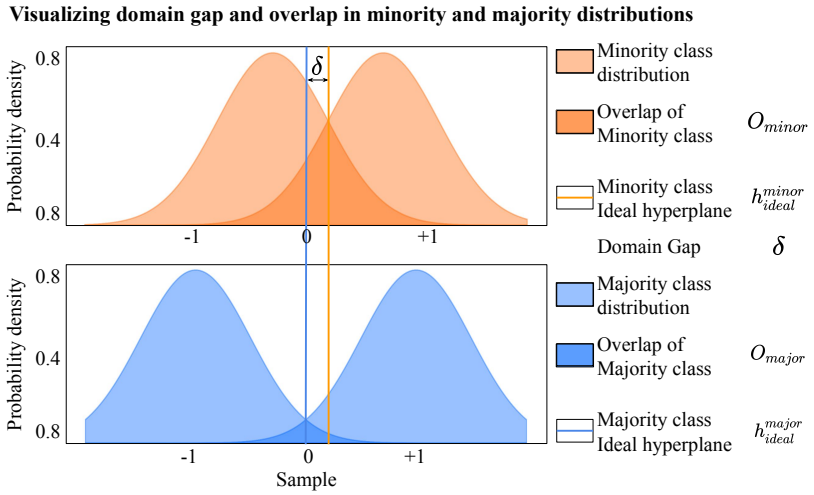


Figure 5.3: **Visualizating of Gaussian Mixture Model parameters.** We plot GMMs with different task complexities. The domain gap δ is visualized as the difference in the ideal threshold locations. The overlap/task complexity metric can be visually seen.

We consider the hyperplane's geometry to be linear in this one dimensional setting. Therefore the hyperplane can be represented as a normal vector: $\mathbf{h}_{\text{ideal}}$. The normalized hyperplane is represented by a two dimensional vector, $\mathbf{h} = [1\ b]^T$. Here, b is the offset/bias. In general, a hyperplane $\mathbf{h}$ may not be ideal. The accuracy of a hyperplane is based on a performance measure $\mathcal{P}\{\mathbf{h}\}$, where the operator $\mathcal{P}$ takes as input the hyperplane and outputs the closeness to the ideal hyperplane $\mathbf{h}_{\text{ideal}}$. A goal of a learning based classifier is to obtain:

$$\widehat{\mathbf{h}} = \arg\min_{\mathbf{h}} \mathcal{P}\{\mathbf{h}\} = \arg\min_{\mathbf{h}} \|\mathbf{h} - \mathbf{h}_{\text{ideal}}\|, \tag{5.5}$$

where $\widehat{\mathbf{h}}$ is the best learnt estimate of the ideal hyperplane. The ideal hyperplane is the global minimizer of this objective. Now, assume we are provided a finite training set of labelled data $\mathcal{D}_{K-1} = \{(x_i, y_i)\}_{i=1}^{K-1}$. Let the estimated hyperplane be $\mathbf{h}_{K-1}$, denoting that $K-1$ samples have been used to learn the hyperplane. If one additional data sample is made available, then the learnt hyperplane would be $\mathbf{h}_K$. From Equation 5.2, the k-th sample is

drawn from one of two distributions:

$$\begin{aligned} x_k|y=1 &\sim p_1(x|\mu_1,\sigma_1) \\ x_k|y=2 &\sim p_2(x|\mu_2,\sigma_2). \end{aligned} \tag{5.6}$$

We now introduce the notion of majority and minority sampling.

Introducing Majority/Minority Distributions: Suppose that the k-th data sample could be drawn for the same classification task from a minority or majority group. Let $g \in \{\text{major}, \text{minor}\}$ denote the group label (for the group class). Equation 5.2 can now be conditioned on the group label, such that there are four possible distributions from which the k-th sample can be drawn:

$$\left.\begin{aligned} x_k|g=\text{major}, y=1 &\sim p_1^{\text{major}}(x|\mu_1^{\text{major}},\sigma_1^{\text{major}}) \\ x_k|g=\text{major}, y=2 &\sim p_2^{\text{major}}(x|\mu_2^{\text{major}},\sigma_2^{\text{major}}) \end{aligned}\right\} \text{Majority group} \tag{5.7}$$

$$\left.\begin{aligned} x_k|g=\text{minor}, y=1 &\sim p_1^{\text{minor}}(x|\mu_1^{\text{minor}},\sigma_1^{\text{minor}}) \\ x_k|g=\text{minor}, y=2 &\sim p_2^{\text{minor}}(x|\mu_2^{\text{minor}},\sigma_2^{\text{minor}}) \end{aligned}\right\} \text{Minority group}$$

Overlap: Let the ideal decision hyperplane be located at $x = d_{\text{ideal}}$. Then, given equal likelihood of the two labels for y, the overlap for the majority group is defined as the probability of erroneous sample classification:

$$O_{\text{major}} = 0.5 \int_{x=-\infty}^{d_{\text{ideal}}} p_2^{\text{major}}(x)dx + 0.5 \int_{x=d_{\text{ideal}}}^{\infty} p_1^{\text{major}}(x)dx. \tag{5.8}$$

The same definition holds true for the minority class as well. Therefore, by definition, $O_{\text{major}} < O_{\text{minor}}$. The task complexities θ^{major} and θ^{minor} are empirical estimates of the respective overlaps. Hereafter, we assume that all four marginal distributions are Gaussian and symmetric (this is relaxed later for Key Result 3). Figure 5.3 visually highlights relevant parameters. $O_{\text{minor}} > O_{\text{major}}$ occurs through the interplay of component means and variances.

The expectation over the class label yields majority and minority sampling:

$$\begin{aligned} x_k^{\text{major}} &\triangleq x_k|g=\text{major} \sim \mathbb{E}_y\big[x_k|g=\text{major}, y\big] \\ x_k^{\text{minor}} &\triangleq x_k|g=\text{minor} \sim \mathbb{E}_y\big[x_k|g=\text{minor}, y\big], \end{aligned} \tag{5.9}$$

where we have defined x_k^{major} or x_k^{minor} as having the k-th sample come from the majority or minority distributions.

Armed with an expression for the k-th sample, we can consider a scope similar to active/online learning [194, 195, 196, 197, 198, 199, 173, 200]. Suppose a dataset of $K-1$ samples has been collected on majority samples, such that there exists a dataset stack $\mathcal{D}_{K-1}^{\text{major}} = \left\{(x_i^{\text{major}}, y_i^{\text{major}})\right\}_{i=1}^{K-1}$. A hyperplane $\mathbf{h}_{K-1}$ is learnt on this dataset and can be improved by expanding the dataset size. Consider pushing sample index K, denoted as x_K onto the stack. Now we have a choice of pushing x_K^{major} or x_K^{minor}, to create one of two datasets:

$$\begin{aligned} \mathcal{D}_K^+ &= \{\mathcal{D}_{K-1}^{\text{major}}, x_K^{\text{major}}\} \\ \mathcal{D}_K^- &= \{\mathcal{D}_{K-1}^{\text{major}}, x_K^{\text{minor}}\}, \end{aligned} \tag{5.10}$$

where $\mathcal{D}_K^-$ represents the interesting case where we choose to push a minority sample onto a dataset with all majority samples (e.g. adding a dark skinned sample to a light skinned dataset). Denote $\mathbf{h}_K^+$ and $\mathbf{h}_K^-$ as hyperplanes learnt on $\mathcal{D}_K^+$ and $\mathcal{D}_K^-$. We now arrive at the following result.

Theorem 1: *Let $\mathcal{P}^{major}\{\cdot\}$ be the performance of a hyperplane on the majority group. Let $\Delta = \mathcal{P}^{major}\{\mathbf{h}_{K-1}\}$. Assume that the minority group distribution has an overlap O_{minor} while the majority group has an overlap $O_{major} < O_{minor}$. Both have the same ideal hyperplane $\mathbf{h}_{ideal}$. Under the definitions of $\mathbf{h}_K^-$ and $\mathbf{h}_K^+$ as above, assuming Δ is sufficiently small and the group class distribution variances are not very large,*

$$\mathop{\mathbb{E}}_{x_K^{minor}} \mathcal{P}^{major}\{\mathbf{h}_K^-\} < \mathop{\mathbb{E}}_{x_K^{major}} \mathcal{P}^{major}\{\mathbf{h}_K^+\}, \tag{5.11}$$

stating that, perhaps surprisingly, expected performance for majorities improves more by pushing a minority sample on the stack, rather than a majority sample.

Proof (Sketch): *A sketch is provided, please see the appendix for the full proof. The general idea is to show that samples closer to $\mathbf{h}_{ideal}$ are more beneficial, and minority distributions*

may sample these with higher likelihood. Without loss of generality, we assume that $\mathbf{h}_{K-1}$ *is located, non-ideally, closer to the task class* $y = 2$ *(arbitrarily called the positive class) than* $\mathbf{h}_{ideal}$*. For our perceptron update rule, the improvement in the estimated hyperplane due to* x_K *is proportional to the difference between the false negative rate (FNR) and the false positive rate (FPR) for* $\mathbf{h}_{K-1}$*, with respect to the distribution of* x_K*. For sufficiently small* Δ*,* $FNR - FPR$ *can be approximated in terms of the likelihood* l *that* x_K *is on the ideal hyperplane. The likelihood* l *is directly proportional to* $FPR - FNR$*. Under the assumptions of the theorem, a direct relation is established between the overlap and* l *for each of the group classes. Then, it is shown that an additional minority sample, with overlap* $O_{minor} > O_{major}$ *leads to greater expected gains as compared to an additional majority sample, concluding the proof.* ■

Key Result 2: MIME holds under domain gap

In the previous key result we described the MIME effect in a restrictive setting where a minority and majority group have the same target hyperplane. However, it is rarely the case that minorities and majorities have the same decision boundary. We now consider the case with non-zero domain gap, to show that MIME holds on a more realistic setting. Domain gap can be quantified in terms of ideal decision hyperplanes. If $\mathbf{h}_{\text{ideal}}^{\text{major}}$ and $\mathbf{h}_{\text{ideal}}^{\text{minor}}$ denote ideal hyperplanes for the majority and minority groups respectively, then domain gap $\delta = \|\mathbf{h}_{\text{ideal}}^{\text{major}} - \mathbf{h}_{\text{ideal}}^{\text{minor}}\|$.

A visual illustration of domain gap is provided in Figure 5.3. Next, we define relative hyperplane locations in terms of halfspaces (since all hyperplanes in the one dimensional setting are parallel). We say two hyperplanes $\mathbf{h}_1$ and $\mathbf{h}_2$ lie in the same halfspace of a reference hyperplane $\mathbf{h}_0$ if their respective offsets/biases satisfy the condition $(b_1 - b_0)(b_2 - b_0) > 0$. For occupancy in different halfspaces, the condition is $(b_1 - b_0)(b_2 - b_0) < 0$. We now enter into the second key result.

Theorem 2: *Let $\delta \neq 0$ be the domain gap between the majority and minority groups. Assume that the minority group distribution has an ideal hyperplane $\mathbf{h}_{ideal}^{minor}$; while the majority group has an ideal hyperplane $\mathbf{h}_{ideal}^{major}$. Then, if $\delta < \Delta$, $\delta + \Delta$ is small enough, and the group class distribution variances are not very large, it can be shown that if either of the following two cases:*

1. *$\mathbf{h}_{K-1}$ and $\mathbf{h}_{ideal}^{minor}$ lie in different halfspaces of $\mathbf{h}_{ideal}^{major}$,*

   or

2. *$\mathbf{h}_{K-1}$ and $\mathbf{h}_{ideal}^{minor}$ lie in the same halfspace of $\mathbf{h}_{ideal}^{major}$, and if*

$$\frac{O_{major}}{O_{minor}} < (1 - \frac{\delta}{\Delta})f, \tag{5.12}$$

are true, then:

$$\mathop{\mathbb{E}}_{x_K^{minor}} \mathcal{P}^{major}\{\mathbf{h}_K^-\} < \mathop{\mathbb{E}}_{x_K^{major}} \mathcal{P}^{major}\{\mathbf{h}_K^+\}, \tag{5.13}$$

where f is a non-negative constant that depends on the majority and minority means and standard deviations for all the individual GMM components.

Proof (Sketch): *A sketch is provided, please see the appendix for the full proof. We prove independently for both cases.*

1. *When $\mathbf{h}_{K-1}$ and $\mathbf{h}_{ideal}^{minor}$ lie in different halfspaces of $\mathbf{h}_{ideal}^{major}$, it can be shown that the expected improvement in the hyperplane is higher for the minority group as compared to the majority group, using a similar argument as in Theorem 1. This proves the theorem for Case 1.*

2. *When $\mathbf{h}_{K-1}$ and $\mathbf{h}_{ideal}^{minor}$ lie in the same halfspace of $\mathbf{h}_{ideal}^{major}$, and assuming that $\mathbf{h}_{K-1}$ is located closer to the positive class, we approximate the $FNR - FPR$ value as function of δ, Δ and the likelihood l as defined for Theorem 1. Then, through algebraic manipulation, constraints can be established in terms of the two likelihoods l_{minor} and l_{major}. Under the assumptions of the theorem, a relation can be established between the ratios $\frac{l_{minor}}{l_{major}}$ and $\frac{O_{minor}}{O_{major}}$. This proves the theorem for Case 2, and concludes the proof.* ■

Key Result 3: MIME holds for general distributions

We now relax the symmetric Gaussian and equally likely labels requirements to arrive at a general condition for MIME existence. Let p_1^{major} and p_2^{major} be general distributions describing the majority group $y = 1$ and $y = 2$ classes. Additionally, $Pr(y = 1) \neq Pr(y = 2)$. Minority group distributions are described similarly. We define the signed tail weight for the majority group as follows:

$$T^{\text{major}}(x_d) = \pi^{\text{major}} \int_{x=-\infty}^{x_d} p_2^{\text{major}}(x)dx - (1 - \pi^{\text{major}}) \int_{x=x_d}^{\infty} p_1^{\text{major}}(x)dx, \tag{5.14}$$

where $\pi^{\text{major}} = Pr(x = 2)$ for the majority group. $T^{\text{minor}}(\cdot)$ is similarly defined. This leads us to our third key result.

Theorem 3: *Consider majority and minority groups, with general sample distributions and unequal prior label distributions. If,*

$$min\left\{T^{minor}(d_{ideal} + \Delta), -T^{minor}(d_{ideal} - \Delta)\right\} > max\left\{T^{major}(d_{ideal} + \Delta), -T^{major}(d_{ideal} - \Delta)\right\}, \tag{5.15}$$

then $\mathop{\mathbb{E}}_{x_K^{minor}} \mathcal{P}^{major}\{\mathbf{h}_K^-\} < \mathop{\mathbb{E}}_{x_K^{major}} \mathcal{P}^{major}\{\mathbf{h}_K^+\}$.

Proof (Sketch): *A sketch is provided, please see the appendix for the full proof. The perceptron algorithm update rule is proportional to $FNR - FPR$ (if $\mathbf{h}_{K-1}$ is located closer to the positive class) or the $FPR - FNR$ (if $\mathbf{h}_{K-1}$ is located closer to the negative class). The MIME effect exists in the scenario where the worst case update for the minority group is better than the best case update for the majority group (described in Equation 5.15). This proves the theorem.* ■

Generalizations of Theorem 3 to include domain gap are discussed in the appendix, for brevity. Theorems 1 and 2 are special cases of the general Theorem 3, describing MIME existence for specific group distributions.

Table 5.1: **Experimental measures of overlap and domain gap are consistent with the theory in Section 5.3**. Note that the majority group consistently has lower overlap. Domain gaps are found to be small. DS-1 is FairFace, DS-2 is Pet Images, DS-4 is Chest-Xray14 and DS-5 is Adult. DS-6 is the high domain gap gender classification experiment. DS-3 is excluded here since it deals with a 9 class classification problem.

Dataset (Task)	DS-1 [167] (Gender)	DS-2 [168] (Species)	DS-4 [169] (Diagnosis)	DS-5 [3] (Income)	DS-6 [5, 4] (Gender)
Major. overlap	0.186	0.163	0.294	0.132	0.09
Minor. overlap	0.224	0.198	0.369	0.208	0.19
Domain gap	0.276	0.518	0.494	0.170	1.62

5.4 Verifying MIME Theory on Real Tasks

In the previous section, we provide existence conditions for the MIME phenomenon for general sample distributions. However, experimental validation of the phenomenon requires quantification in terms of measurable quantities such as overlap. Theorem 2 provides us these resources. Here, we verify that the assumptions in Theorem 2 are validated by experiments on real tasks.

5.4.1 Verifying Assumptions

Verifying Gaussianity: Theorem 2 assumes that data x is drawn from a Gaussian Mixture Model. At first glance, this quantification may appear to be unrelated to complex neural networks. However, as illustrated at the top of Figure 5.4, a ConvNet is essentially a feature extractor that feeds a flattened layer into a simple perceptron or linear classifier. The flattened layer can be orthogonally projected onto the decision boundary to generate, in analogy, an x used for linear classification (Figure 5.1, fixed-backbone configuration). We

use this as a first approximation to the end-to-end configuration used in our experiments.

Plotting empirical histograms of these flattened layers (Figure 5.4) shows Gaussian-like distribution. This is consistent with the Law of Large Numbers – linear combination of several random variables follows an approximate Gaussian distribution. Hence, Theorem 2 is approximately related in this setting. Details about implementation and comparison to Gaussians are deferred to the appendix.

Verifying minority/majority definitions: The MIME proof linked minority and majority definitions to distributional overlap and domain gap. Given the histogram embeddings from above, it is seen that minority groups on all four vision tasks have greater overlap. There also exists a domain gap between majority and minority but this is small compared to distribution spread (except for the high domain gap experiment). This establishes applicability of small domain gap requirements. Quantification is provided in Table 5.1. Code is in the appendix.

5.4.2 MIME Effect Across Six, Real Datasets

Implementation: Six multi-attribute datasets are used to assess the MIME effect (five are in computer vision). For a particular *experiment*, we identify a task category to evaluate accuracy over (e.g. gender), and a group category (e.g. race). The best test accuracy on the majority group across all epochs is recorded as our accuracy measure. Each experiment is run for a fixed number of *minority training ratios* (β). For each minority training ratio, the total number of training samples remains constant. That is, the minority samples replace the majority samples, instead of being appended to the training set. Each experiment is also run for a finite number of *trials*. Different trials have different random train and test sets (except for the FairFace dataset [167] where we use the provided test split). Averaging is done across trials. Note that minority samples to be added are randomly chosen – the MIME effect is not specific to particular samples. For the vision datasets, we use a ResNet-34 architecture [201], with the output layer appropriately modified. For the non-visual dataset,

a fully connected network is used. Average accuracy and trend error, across trials are used to evaluate performance. Specific implementation details are in the appendix.

MIME effect on gender classification: The FairFace dataset [167] is used to perform gender classification ($y = 1$ is male, $y = 2$ is female). The majority and minority groups $g = \{\text{major}, \text{minor}\}$ are light and dark skin, respectively. Results are averaged over five trials. Figure 5.5 describes qualitative accuracy. The accuracy trends indicate that adding 10% of minority samples to the training set leads to approximately a 1.5% gain in majority group (light skin) test accuracy.

MIME effect on animal species identification: We manually annotate light and dark cats and dogs from the Pets dataset [168]. We classify between cats ($y = 1$) and dogs ($y = 2$). The majority and minority groups are light and dark fur color respectively. Figure 5.5 shows qualitative results. Over five trials, we see a majority group accuracy gain of about 2%, with a peak at β =10%.

MIME effect on age classification: We use a second human faces dataset, the UTKFace dataset [5], for the age classification task (9 classes of age-intervals). We pre-process the UTKFace age labels into class bins to match the FairFace dataset format. The majority and minority groups are male and female respectively. The proportion of task class labels is kept the same across group classes. Results are averaged over five trials. Figure 5.5 shows trends. We observe a smaller average improvement for the 10% minority training ratio. However, since these are average trends, this indicates consistent gain. Results on this dataset also empirically highlight the existence of the MIME effect beyond two class settings.

MIME effect on X-ray diagnosis Classification: We use the NIH Chest-Xray14 dataset [169] to analyze trends on a medical imaging task. We perform binary classification of scans belonging to 'Atelectasis' ($y = 1$) and 'Pneumothorax' ($y = 2$) categories. The male and female genders are the majority and minority groups respectively. Results are averaged over seven trials (due to noisier trends). From Figure 5.5, we observe noisy trends - specifically we see a performance drop for $\beta = 0.2$, prior to an overall gain for $\beta = 0.3$. The error bounds also

have considerably more noise. However, confidence in the peak and the MIME effect, as seen from the average trends and the error bounds, remains high.

MIME effect on income classification: For validation in a non-vision setting, we use the Adult (Census Income) dataset [3]. The data consists of census information with annual income labels (income less than or equal to \$50,000 is $y = 1$, income greater than \$50,000 is $y = 2$). The majority and minority groups are female and male genders respectively. Results are averaged over five trials. Figure 5.6(a) highlights a prominent accuracy gain for $\beta = 0.6$.

MIME effect and domain gap: Theorem 2 (Section 5.3) suggests that large domain gap settings will not show the MIME effect. We set up an experiment to verify this (Figure 5.6(b)). Gender classification among chickens (majority group) and humans (minority group) has a high domain gap due to minimal common context (validated by the domain gap estimates, Table 5.1). With increasing β, the majority accuracy decreases. This (and Figure 5.4, Table 5.1 that show low domain gap for other datasets) validates Theorem 2. Note that while this result may not be unexpected, it further validates our proposed theory.

5.5 Discussion

Secondary validation and analysis: Table 5.2 supplies additional metrics to analyze MIME. Across datasets, almost all trials show existence, with every dataset showing average MIME performance gain. Some readers may view the error bars in Figures 5.5 and 5.6 as large, however they are comparable to other empirical ML works [202, 203]; they may appear larger due to scaling. Reasons for error bars include variations in train-test data and train set size (Table B and C, appendix). Further analysis, including interplay with debiasing methods (e.g. hard-sample mining [15]) and reconciliation with work on equal representation datasets [160, 58, 161, 162, 163, 164, 165, 166, 143] is deferred to the appendix.

Optimality of inclusion ratios: Our experiments show that there can exist an optimal amount of minority inclusion to benefit the majority group the most. This appears true across

Table 5.2: **Additional evaluation metrics provide further evidence of MIME existence across all datasets.** The table highlights: (i) number of trials with MIME performance gain (i.e. majority accuracy at some $\beta > 0$ is greater than majority accuracy at $\beta = 0$), and (ii) the mean MIME performance gain across trials (in % points).

Dataset	DS-1 [167]	DS-2 [168]	DS-3 [5]	DS-4 [169]	DS-5 [3]
#MIME trials/Total trials	4/5	4/5	5/5	6/7	4/5
Avg. MIME perf. gain	0.72%	1.84%	0.70%	1.89%	0.98%

all experiments in Figures 5.5, 5.6. However, beyond a certain amount, accuracy decreases consistently, with lowest accuracy on majority samples observed when no majorities are used in training. This optimal β depends on individual task complexities, among other factors. Since identifying it is outside our scope (Section 5.1.1, 5.1.2), our experiments use 10% sampling resolution for β. Peaks at $\beta = 10\%$ for some datasets are due to this lower resolution; optimal peak need not lie there for all datasets (e.g. X-ray [169] & Adult [3]). Future work can identify optimal ratios through finer analysis over β.

Limitations: The theoretical scope is certifiable within fixed-backbone binary classification, which is narrower than all of machine learning (Figure 5.1). Should this theory be accepted by the community, follow-up work can generalize theoretical claims. Another limitation is the definition-compatibility of majority and minority groups. Our theory is applicable to task-advantage definitions; some scholars in the community instead define majorities and minorities by proportion. Our theory is applicable to these authors as well, albeit with a slight redefinition of terminology. Additional considerations are included in the appendix.

Conclusion: In conclusion, majority performance benefits from a non-zero fraction of inclusion of minority data given a sufficiently small domain gap.

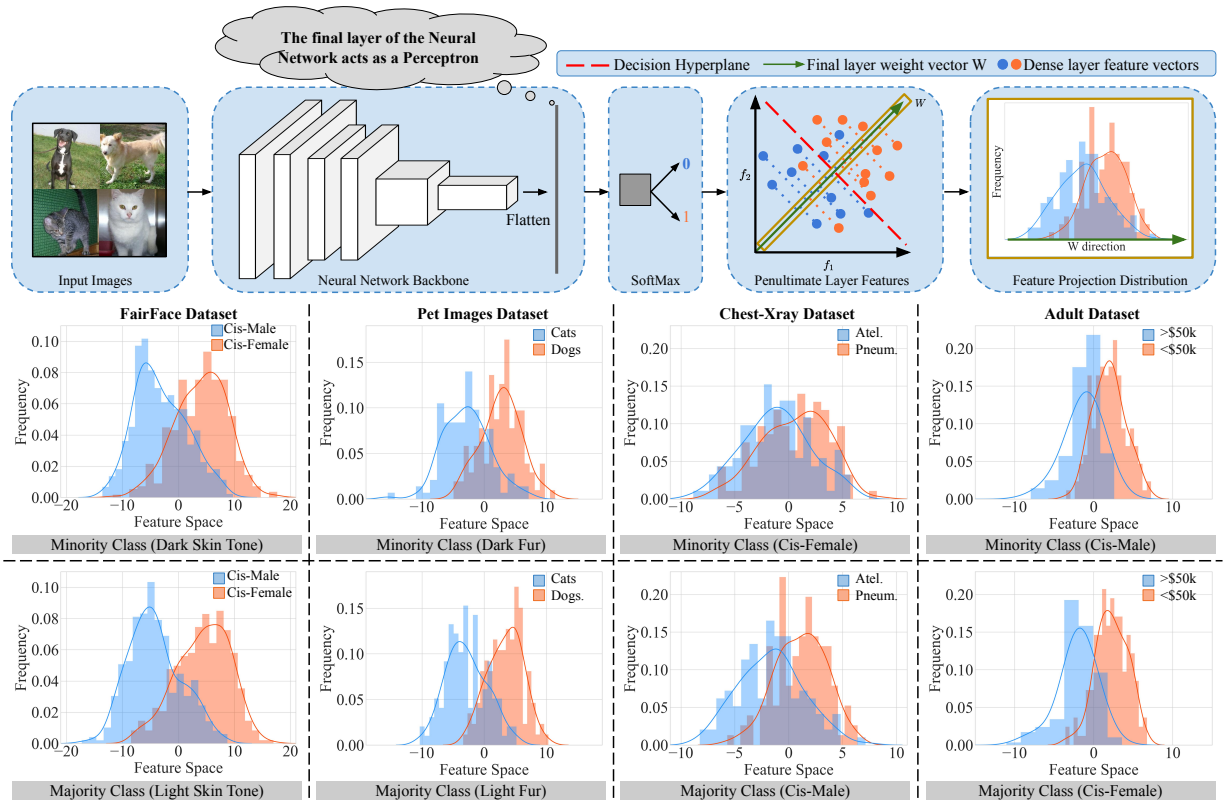


Figure 5.4: **The use of Gaussian mixtures to represent minority and majority distributions is consistent with behaviors in modern neural networks, on real-world datasets.** (top row) The last layer of common neural architectures is a linear classifier on features. Histograms of the penultimate layer projections are generated for models with $\beta = 0.5$. (middle row) Minority histograms: note the greater difficulty due to less separation of data. (bottom row) Majority histograms: note smaller overlap and easier classification. Figure can be parsed on a per-dataset basis. Within each column, the reader can compare the domain gap and overlap in the two histograms.

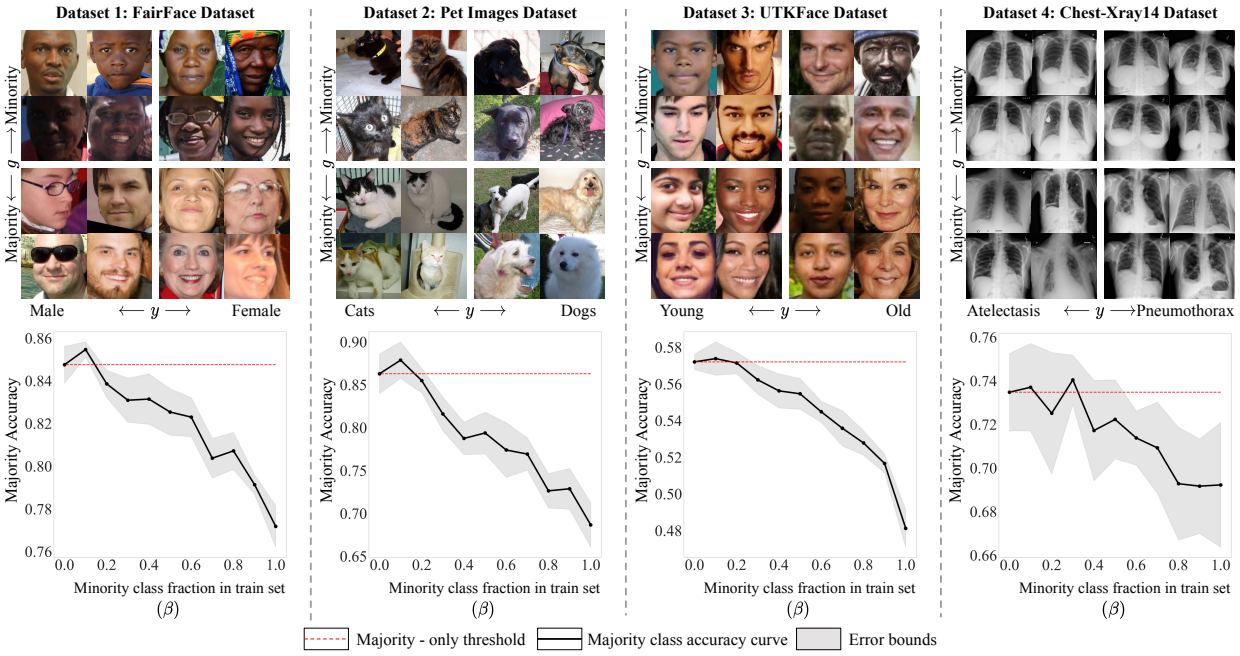


Figure 5.5: **When domain gap is small, the MIME effect holds.** On four vision datasets, majority performance is maximized with some inclusion of minorities. All experiments are run for several trials and realizations (described in Section 5.4.2).

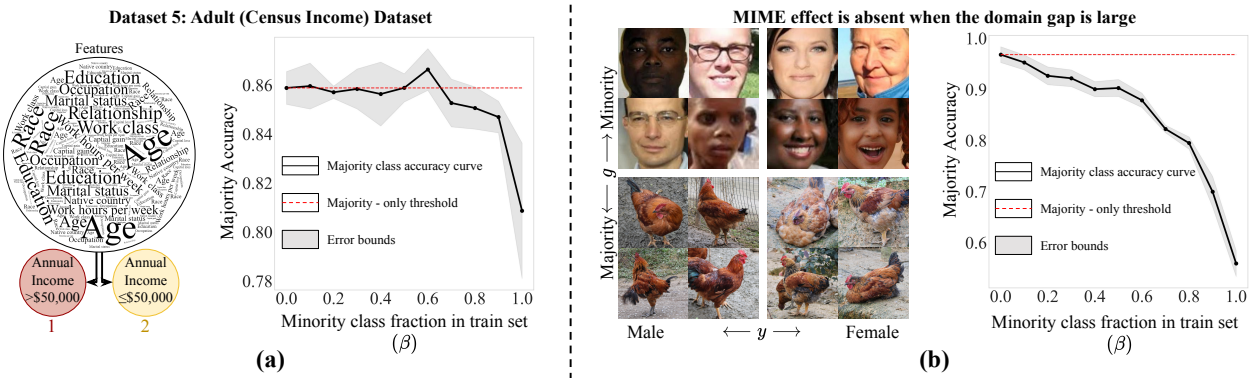


Figure 5.6: **MIME effect is observed in non-vision datasets, and is absent in the case of large domain gap.** (a) The Adult Dataset [3] uses Census data to predict an income label. (b) On dataset six, gender classification is rescoped to occur in a high domain gap setting. Majority group is chickens [4] and minority group is humans [5].

CHAPTER 6

Using Neural Implicit Video Representations to Enable Low-SNR rPPG

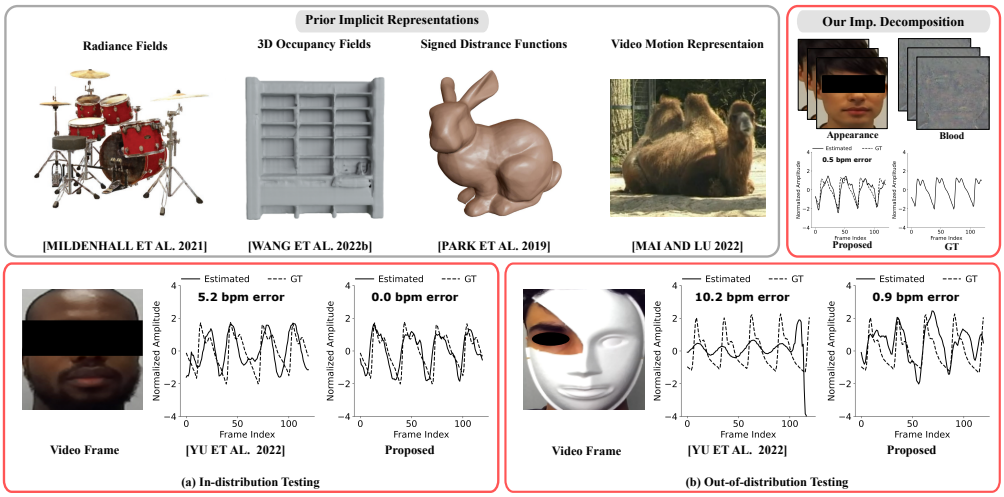


Figure 6.1: **Prior implicit neural models represent scenes for diverse applications. We propose an implicit neural representation (INR) to decompose face videos and isolate blood flow information.** Our INR decomposes input videos into visual appearance and blood flow ("$\mathcal{A}$-$\mathcal{B}$ decomposition"). The decomposed data aids in estimating the remote photoplethysmography (rPPG) signal and heart rate.

6.1 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 [6, 204, 205, 206], scene representations [207, 208, 209, 210], radiance fields [211, 212, 213] and video

representations [214, 215, 7]. 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 [192, 143, 32, 8] 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 [26, 27]). These usually lead to relatively larger heart rate estimation errors. Recently, learning-based methods [30, 31, 9] 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. Specifically, we introduce a new Appearance-Blood (or $\mathcal{A}$-$\mathcal{B}$) 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

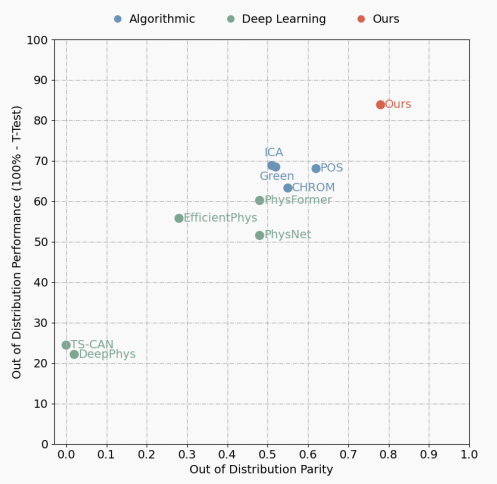


Figure 6.2: **Our implicit representation for rPPG achieves Pareto-optimality across out-of-distribution (OOD) performance and inter-distribution parity compared to prior algorithmic and learning-based methods.** It performs better on OOD samples while maximizing parity between in-distribution and OOD performance. Table 6.1 shows metrics used for this plot. Higher is better along both axes.

by [216] to facilitate fast $\mathcal{A}$-$\mathcal{B}$ 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 6.2) with reliable plethysmograph and heart rate estimates. Notably, we achieve this OOD performance gain without compromising on in-distribution performance on rPPG datasets (Figure 6.1). 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 ($\mathcal{A}$-$\mathcal{B}$) decomposition as a means of enhancing plethysmograph signal strength. We show that carefully designed implicit neural representations can serve as efficient $\mathcal{A}$-$\mathcal{B}$ 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 6.2) without losing out on in-distribution performance.

3. An optically challenging test dataset, consisting of 104 videos ($\approx$ 1 *hour* of recorded data) across various optically challenging and OOD configurations, is collected and will be released upon acceptance of this work.

6.1.1 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 (appendix). 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).

6.2 Related Work

Implicit Neural Representations. Implicit neural representations (INRs) are an increasingly prevalent method to represent scenes. Examples settings include image representations [6, 204, 205, 206], video representations [214, 215, 7], and 3-D scene representations [207, 208, 209, 210]. Specifically, this work is interested in the class of video-based representations. These include methods that use convolutional networks [214], phase-based motion-adjustable representation [7], space-time super-resolution [215] and flow fields for space-time viewpoint synthesis [217, 218]. 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, how-

ever, has taken a step toward alleviating this concern. [219] 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 [213, 216]. In this work, we are interested in employing [216]'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 [27, 26] or motion based on Newtonian reaction to blood flow [41]. 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 [42]. Augmented reality [46] is another avenue for this research. Deep learning approaches have also been utilized to attain SOTA results. [30] used an attention-based Convolutional Neural Network (CNN) while [31] introduced spatio-temporal CNNs. Other work has extended these architectures [47], including using transformers [9], incorporated meta-learning [48], improved PPG waveform characteristics [49], and augmented rPPG datasets with synthetic examples [1, 16]. Another class of methods focuses on improving equity between groups, such as participants of different skin tones [32, 143]. These include algorithmic methods [192] or multimodal fusion methods [8]. Methods for assessing and improving equity are not restricted to rPPG and extend to a range of computer vision problems [220, 143, 57, 186, 221, 222].

6.3 $\mathcal{A}$-$\mathcal{B}$ 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 [30, 31, 9] 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:

$$\mathcal{C}(\mathbf{x}, t) = \mathcal{A}(\mathbf{x}, t) + \mathcal{B}(\mathbf{x}, t) + \mathbf{v_n}(\mathbf{x}, t), \tag{6.1}$$

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,*

$$\widehat{p_i^*}(t) = \int_{\mathbf{x}\in\mathbf{\Omega}} SINR_{\mathcal{B}}(\mathbf{x},t)\mathcal{B}(\mathbf{x},t)d\mathbf{x}, \tag{6.2}$$

where $\mathbf{\Omega} = [-0.5, 0.5]^2$, 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 appendix.

6.3.1 Optimal Plethysmography and Uncertainty

Equation 6.2 can also be interpreted in the context of uncertainty minimization [223]. 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 $\widehat{p_i^*}(t)$ can be interpreted as Bayesian inference. With assumptions on prior and posterior distributions, it can be shown that the posterior $\widehat{p_i^*}(t)$ 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 6.2) and variance $1/(\sum_{i=1}^N SINR_{\mathcal{B}}(\mathbf{x}_i,t))$. Since Equation 6.2 represents maximal ratio combining, $\sum_{i=1}^N SINR_{\mathcal{B}}(\mathbf{x}_i,t) \propto SINR_{\widehat{p_i^*}(t)}$ ([224], Equation 31, 32), then, uncertainty (or the variance in the posterior) is inversely proportional to SINR of $\widehat{p_i^*}(t)$. Optimal plethysmography therefore becomes akin maximization of SINR, or minimization of uncertainty of $\widehat{p_i^*}(t)$. Please refer to the appendix for derivations.

6.4 $\mathcal{A}$-$\mathcal{B}$ 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.

6.4.1 Functional Decomposition

Consider two functions, $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{A}}})$ and $\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{B}}})$, 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:

$$\begin{aligned} d_{\mathcal{A}}(\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}^*_{\boldsymbol{\mathcal{A}}}), \mathcal{A}(\mathbf{x}, t)) &\leq \epsilon_{\mathcal{A}}, \\ d_{\mathcal{B}}(\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}^*_{\boldsymbol{\mathcal{B}}}), \mathcal{B}(\mathbf{x}, t)) &\leq \epsilon_{\mathcal{B}}, \\ d_{\mathcal{B}}(\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}^*_{\boldsymbol{\mathcal{A}}}), \mathcal{B}(\mathbf{x}, t)) &\geq \epsilon_{\mathcal{B}}. \end{aligned} \tag{6.3}$$

where $\boldsymbol{\Theta}^*_{\boldsymbol{\mathcal{A}}}$ and $\boldsymbol{\Theta}^*_{\boldsymbol{\mathcal{B}}}$ 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 $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{A}}})$ 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,

$$d_{\mathcal{B}}(\mathcal{F}, \mathcal{B}(\mathbf{x}, t)) = |H(\mathcal{F}) - H(\mathcal{B}(\mathbf{x}, t))|, \forall\, \mathcal{F}, \tag{6.4}$$

where $H(\cdot)$ is a heart rate estimator function. By finding functions that satisfy the constraints in Equation 6.3, we can achieve $\mathcal{A}$-$\mathcal{B}$ decomposition by sequentially learning $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{A}}})$ and $\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{B}}})$. A detailed mathematical analysis of this formulation is deferred to the appendix.

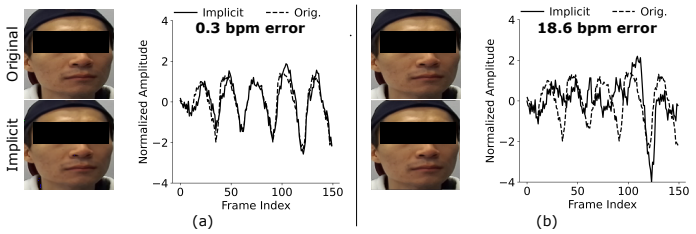


Figure 6.3: **Sinusoidal Representation Networks (SRNs) can represent both $\mathcal{A}$ and $\mathcal{B}$-functions, while phase-based methods only represent the $\mathcal{A}$-function.** (a) SRNs, such as [6], can capture PPG color variations almost perfectly, while (b) phase-based motion representations, such as [7] are unable to capture it.

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

Prior work on video magnification through Laplacian [23] and phase-based [225] pyramids is instructive. While the Laplacian pyramid-based approach [23] can magnify blood flow due to color changes, the phase-based method [225] 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 [7]. The key inductive bias here is parameterizing time as phase modulation in positional encodings [211, 226].

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 6.4.* (Figure 6.3(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 6.3).

We next look at sinusoid representation networks (SRNs), such as [6, 205, 206].

Empirical Claim 2: *SRNs are good $\mathcal{A}$, $\mathcal{B}$-function approximators.* (Figure 6.3(a))

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 6.3. Hence,

Empirical Claim 3: *A phase-based model can serve as $\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}^*_{\mathcal{A}})$, while an SRN can*

serve as $\hat{\mathcal{B}}(\mathbf{x}, t; \mathbf{\Theta}^*_{\mathcal{B}})$*, when trained sequentially.*

Justifications for all claims may be found in the appendix. However, naive use of prior work [6, 7] for $\mathcal{A}$-$\mathcal{B}$ decomposition poses a training time problem. Using variants of [7] and [6] 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.

6.5 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 [216], 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 [227, 228] 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.

6.5.1 Cascaded Appearance Model $\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}_{\mathcal{A}})$

As discussed in Section 6.4.2, phase-based INRs [7], 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

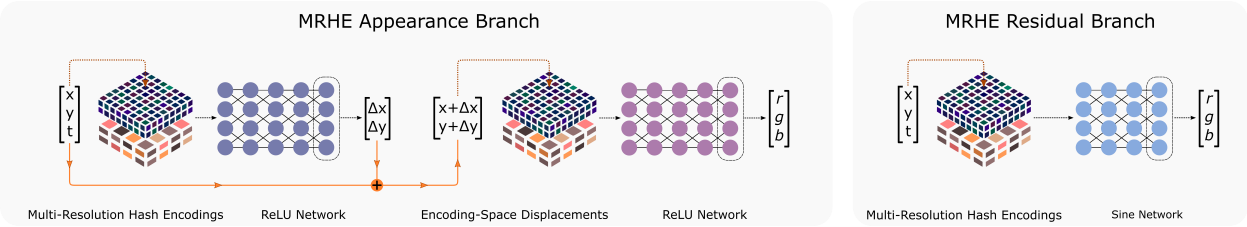


Figure 6.4: **To enable fast $\mathcal{A}$-$\mathcal{B}$ decomposition, we use implicit neural representations as decomposing function fitters.** Training is done sequentially: first, the cascaded appearance model learns the $\mathcal{A}$-function. Then, the appearance model is frozen, and the residual model learns the $\mathcal{B}$-function, thereby completing the decomposition. The use of multiresolution hash encodings makes dataset-scale decomposition viable.

MRHE. The second is queried using the offset positions to generate video frames.

Rather than considering a single offset vector in $\mathbb{R}^2$ for the entire frame (as [7] 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 $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{A}}}) = f_{\mathcal{A}_2}(\mathbf{x} + f_{\mathcal{A}_1}(\mathbf{x}, t))$. The architecture and loss functions are elaborated on in the appendix. Figure 6.4 details this pipeline.

6.5.2 Residual Plethysmograph Model $\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{B}}})$

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, $\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}_{\boldsymbol{\mathcal{B}}}) = f_{\mathcal{B}}(\mathbf{x}, t)$.

The resulting residual plethysmograph model is shown to compensate for the cascaded

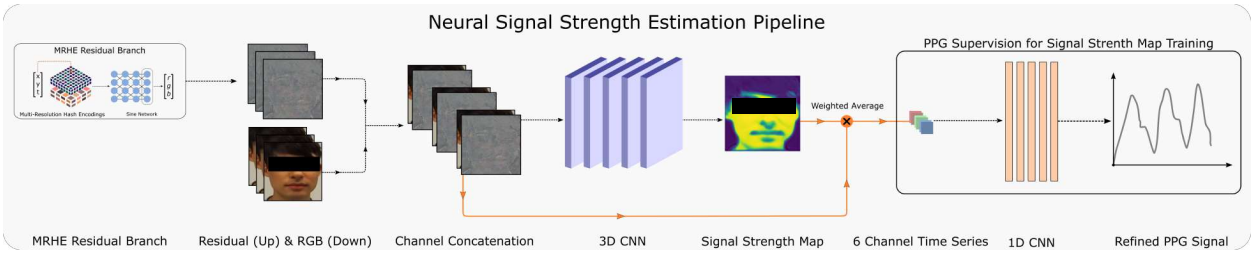


Figure 6.5: **Using the estimated $\mathcal{B}$-function with the original video, we learn high-fidelity neural signal strength masks.** The network takes the original RGB frames and the $\mathcal{B}$-function estimate as inputs and returns a spatial strength mask. Training is supervised through an auxiliary 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.

appearance model's inability to retain the plethysmograph signal. When queried, the model yields a spatio-temporal estimate of the plethysmograph signal. The appendix contains relevant implementation details.

6.6 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 6.2. In this section, we proceed towards rPPG estimation from the $\mathcal{B}$-function using high-fidelity neural signal masks.

Figure 6.5 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, $\mathcal{I}(\mathbf{x}, t) = \left[\mathcal{C}\left(\mathbf{x}, t\right), \hat{\mathcal{B}}(\mathbf{x}, t)\right]$, 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 $\widehat{SINR}_{\mathcal{B}}(\mathbf{x}) = f_{SINR}\left(\mathcal{I}\left(\mathbf{x}, t\right); \Theta_{SINR}\right)$.

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 (appendix). The SINR mask $\widehat{SINR}_{\mathcal{B}}(\mathbf{x})$ is directly used compute $\mathbf{y}_1(t) = \frac{\sum_{\mathbf{x}} \widehat{SINR}_{\mathcal{B}}(\mathbf{x}) \cdot I(\mathbf{x},t)}{\sum_{\mathbf{x}} \widehat{SINR}_{\mathcal{B}}(\mathbf{x})}$, 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 appendix. As a result of the architecture being based on the optimality constraint in Equation 6.2, 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 [25] and utilize only the green channel. Advanced algorithms [27, 26] exist that utilize other channels, but they rely on the $\mathcal{A}$-function [27, 26]. We perform detrending and Butterworth filtering for heart rate estimation as in [27].

6.7 Results

6.7.1 Experimental Setup

6.7.1.1 Datasets

All methods are trained on two prior datasets [16, 8]. We also test our method on a self-collected Institutional Review Board (IRB) approved optically challenging OOD dataset.

Table 6.1: **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 6.2. For algorithmic methods (non-learning), r-consistency on the datasets from [16] and [8] are shown in parenthesis. The best and second-best-performing numbers are highlighted in green and yellow, respectively.

	OOD Performance Metrics					Inter-Distribution Parity Metric
Method	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑	r-consistency ↑
POS [Wang et al. (2016)]	31.92	8.02	11.06%	14.87	0.47	0.62
PhysFormer [Verkruysse et al. (2008)]	39.81	9.58	12.59%	15.14	0.32	0.48
Ours	16.15	4.61	6.07%	11.65	0.65	0.78

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 appendix.

6.7.1.2 Evaluation Metrics

We adopt the heart rate (HR) performance metrics from [8]. 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) [65, 67, 66]. 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 appendix.

Table 6.2: **In-distribution and Cross-dataset OOD performance across two datasets - [8] and [16] 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.

	Test on Vilesov et al. (2022)				
Method	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑
POS [Wang et al. (2016)]	6.00	2.13	2.78%	6.54	0.89
PhysFormer [Verkruysse et al. (2008)]	2.96	1.06	1.37%	3.35	0.96
Ours	3.34	1.22	1.61%	4.01	0.96

6.7.1.3 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 [8], we validate[1] 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 [16] to evaluate the learning methods. Please refer to the appendix for details, including visualization of signal strength masks.

Figures 6.6 and 6.7 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).

[1]Baselines are configured, where possible, using the toolbox from [229].

6.7.2 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.

6.7.2.1 Optically Challenging OOD Evaluation

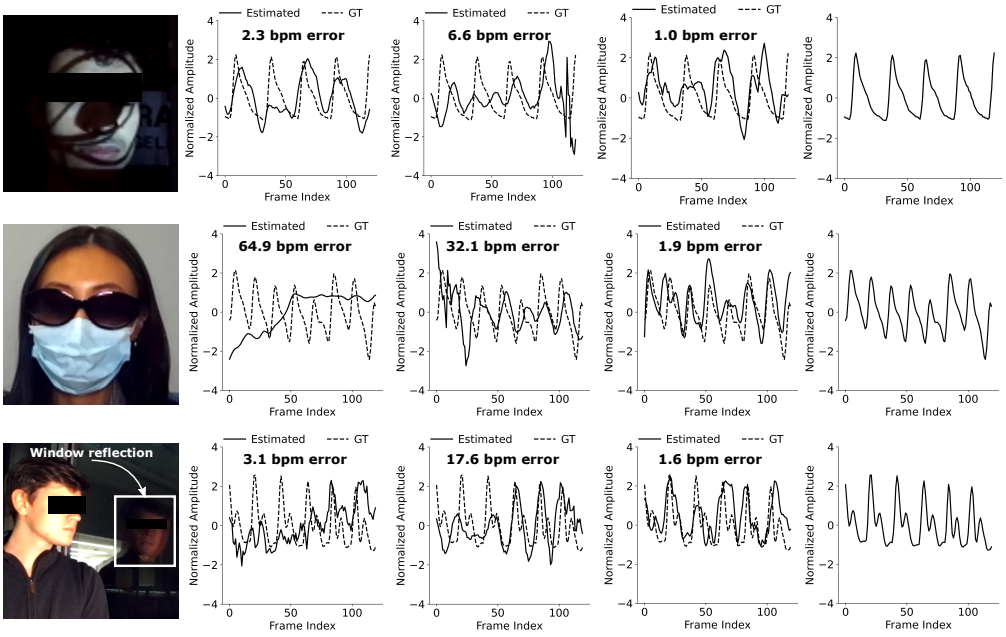
Table 6.1 presents a quantitative analysis of OOD performance. Learning-based methods are trained on the datasets from [16] and [8], and the best fold is chosen for OOD inference.

Given their capacity to memorize distinguishable patterns, learning methods [31, 9] handle input noise better than algorithmic methods for in-distribution 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 6.2) 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 appendix.

6.7.2.2 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

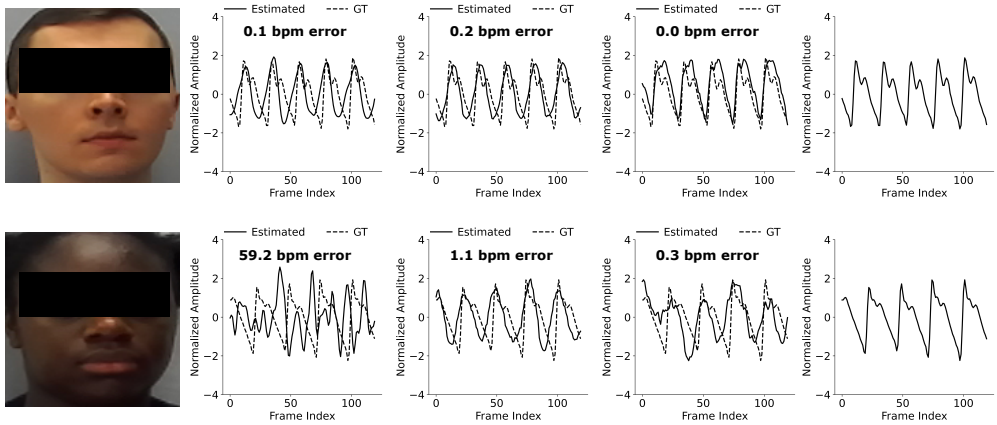


(a) Vid. (b) [25] (c) [9] (d) Ours (e) GT

Figure 6.6: **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 [8], where applicable. Additional results are presented in the appendix.

to overfit the training data. Our experiments use datasets from [16] and [8] for their size and diversity. Table 6.2 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 on [8] and tested on [16] (4^{th} quadrant), we outperform all prior work. In the reverse scenario ($2nd$ quadrant), we are the second-best, marginally behind [31].

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



(a) Vid. (b) [27] (c) [9] (d) Ours (e) GT

Figure 6.7: **In distribution inference on a diverse dataset.** Our method better captures rPPG waveform details across skin tones for in-distribution evaluation on [8] dataset. While our method and [9] both perform reasonably well across skin tones, [9] does so with poorer OOD performance (Figure 6.6). More results in the appendix.

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.

6.7.3 Qualitative Comparison

In Figure 6.6, 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 6.1, i.e., Green [25] and PhysFormer [9] 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 6.7, we assess qualitative waveforms from in-distribution inference samples from [8]. We compare our method with PhysFormer [9] and POS [27] 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.

6.8 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.

6.8.1 Limitations

Being the first method using INRs for rPPG, limitations exist. First, while we reduce runtime over traditional INRs (as discussed in the appendix), and some rPPG methods report comparable runtimes [230], this is an avenue for future improvement. Second, while the proposed method is competitive for challenging scenes like talking and motion (appendix), these problems remain open.

6.8.2 Societal Impacts and Ethical Considerations

We focus on rPPG in optically challenging scenes (Section 6.1.1) 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 [231, 232] 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 [233, 76, 234].

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.

CHAPTER 7

Enabling 3D Perception from 2D Foundation Models

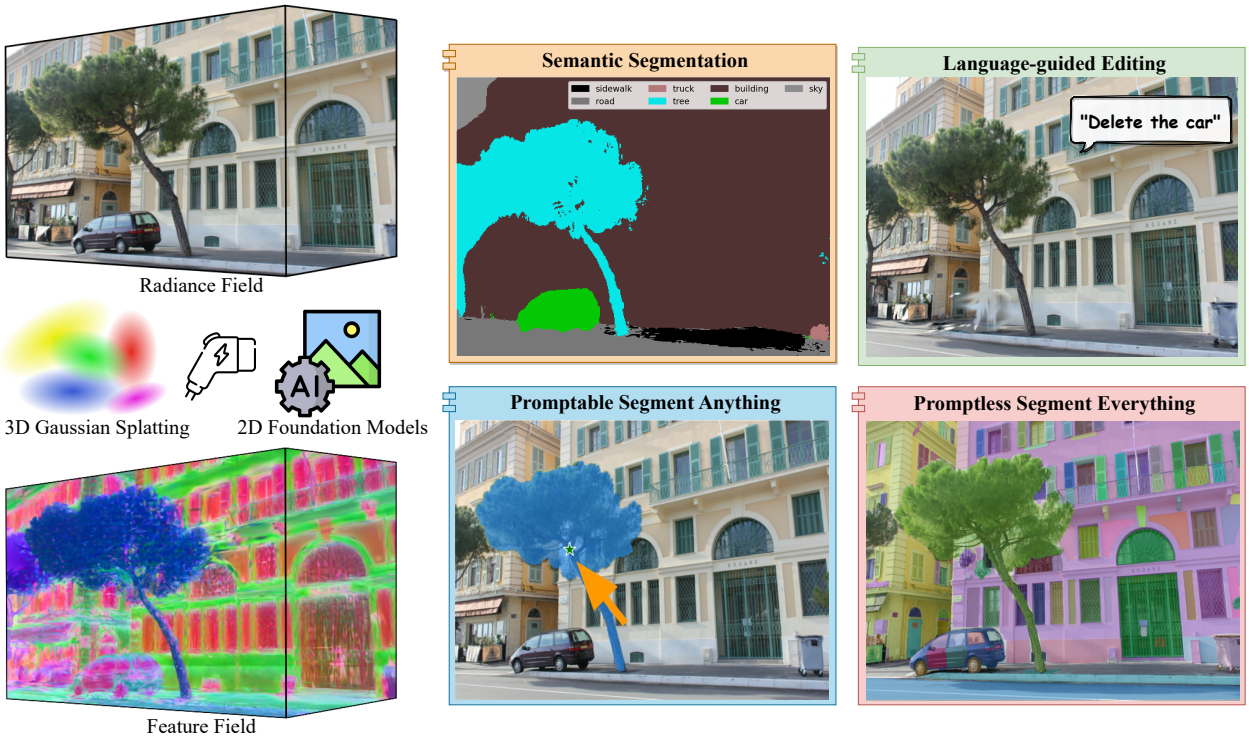


Figure 7.1: **Feature 3DGS**. We present a general method that significantly enhances 3D Gaussian Splatting through the integration of large 2D foundation models via feature field distillation. This advancement extends the capabilities of 3D Gaussian Splatting beyond mere novel view synthesis. It now encompasses a range of functionalities, including semantic segmentation, language-guided editing, and promptable segmentations such as “segment anything" or automatic segmentation of everything from any novel view. Scene from [10].

7.1 Introduction

3D scene representation techniques have been at the forefront of computer vision and graphics advances in recent years. Methods such as Neural Radiance Fields (NeRFs) [211], and works that have followed up on it, have enabled learning implicitly represented 3D fields that are supervised on 2D images using the rendering equation. These methods have shown great promise for tasks such as novel view synthesis. However, since the implicit function is only designed to store local radiance information at every 3D location, the information contained in the field is limited from the perspective of downstream applications.

More recently, NeRF-based methods have attempted to use the 3D field to store additional descriptive features for the scene, in addition to the radiance [13, 235, 14, 236]. These features, when rendered into feature images, can then provide additional semantic information for the scene, enabling donwstream tasks such as editing, segmentation and so on. However, feature field distillation through such a method is subject to a major disadvantage: NeRF-based methods can be natively slow to train as well as to infer. This is further complicated by model capacity issues: if the implicit representation network is kept fixed, while requiring it to learn an additional feature field (to not make the rendering and inference speeds even slower), the quality of the radiance field, as well as the feature field is likely to be affected unless the weight hyperparameter is meticulously tuned [13].

A recent alternative for implicit radiance field representations is the 3D Gaussian splatting-based radiance field proposed by Kerbl et al. [237]. This explicitly-represented field using 3D Gaussians is found to have superior training speeds and rendering speeds when compared with NeRF-based methods, while retaining comparable or better quality of rendered images. This speed of rendering while retaining high quality has paved the way for real-time rendering applications, such as in VR and AR, that were previously found to be difficult. However, the 3D Gaussian splatting framework suffers the same representation limitation as NeRFs: natively, the framework does not support joint learning of semantic features and radiance

field information at each Gaussian.

In this work, we present Feature 3DGS: the first feature field distillation technique based on the 3D Gaussian Splatting framework. Specifically, we propose learning a semantic feature at each 3D Gaussian, in addition to color information. Then, by splatting and rasterizing the feature vectors differentiably, the distillation of the feature field is possible using guidance from 2D foundation models. While the structure is natural and simple, enabling fast yet high-quality feature field distillation is not trivial: as the dimension of the learnt feature at each Gaussian increases, both training and rendering speeds drop drastically. We therefore propose learning a structured lower-dimensional feature field, which is later upsampled using a lightweight convolutional decoder at the end of the rasterization process. Therefore, this pipeline enables us to achieve improved feature field distillation at faster training and rendering speeds than NeRF-based methods, enabling a range of applications, including semantic segmentation, language-guided editing, promptable/promptless instance segmentation and so on.

In summary, our contributions are as follows:

- A novel 3D Gaussian splatting inspired framework for feature field distillation using guidance from 2D foundation models.
- A general distillation framework capable of working with a variety of feature fields such as CLIP-LSeg, Segment Anything (SAM) and so on.
- Up to **2.7**× faster feature field distillation and feature rendering over NeRF-based method by leveraging low-dimensional distillation followed by learnt convolutional upsampling.
- Up to **23%** improvement on mIoU for tasks such as semantic segmentation.

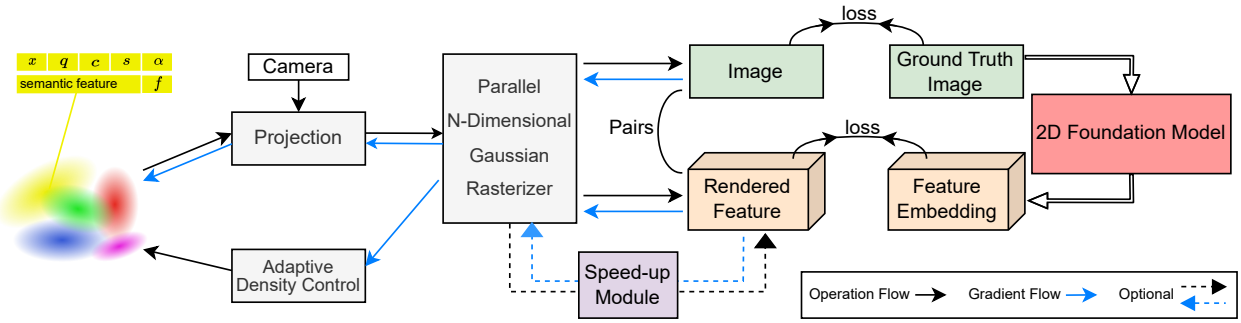


Figure 7.2: **An overview of our method.** We adopt the same 3D Gaussian initialization from sparse SfM point clouds as utilized in 3DGS, with the addition of an essential attribute: the *semantic feature*. Our primary innovation lies in the development of a Parallel N-dimensional Gaussian Rasterizer, complemented by a convolutional speed-up module as an optional branch. This configuration is adept at rapidly rendering arbitrarily high-dimensional features without sacrificing downstream performance.

7.2 Related Work

7.2.1 Implicit Radiance Field Representations

Implicit neural representations have achieved remarkable success in recent years across a variety of areas within the field of computer vision [210, 238, 239, 240, 211, 241]. NeRF [211] demonstrates outstanding performance in novel view synthesis by representing 3D scenes with a coordinate-based neural network. In mip-NeRF [241], point-based ray tracing is replaced using cone tracing to combat aliasing. Zip-NeRF [242] utilized an anti-aliased grid-based technique to boost the radiance field performance. Instant-NGP [216] reduces the cost for neural primitives with a versatile new input encoding that permits the use of a smaller network without sacrificing quality, thus significantly reducing the number of floating point and memory access operations. IBRNet [243], MVSNeRF [244], and PixelNeRF [245] construct a generalizable 3D representation by leveraging features gathered from various observed viewpoints. However, NeRF-based methods are hindered by slow rendering speeds

and substantial memory usage during training, a consequence of their implicit design.

7.2.2 Explicit Radiance Field Representations

Pure implicit radiance fields are slow to operate and usually require millions of times querying a neural network for rendering a large-scale scene. Marrying explicit representations into implicit radiance fields enjoys the best of both worlds. Triplane [246], TensoRF [247], K-Plane [248], TILED [249] adopt tensor factorization to obtain efficient explicit representation. InstantNGP [216] utilizes multi-scale hash grids to work with large-scale scenes. Block-NeRF [250] further extends NeRF to render city-scale scenes spanning multiple blocks. Point NeRF [251] uses neural 3D points for representing and rendering a continuous radiance volume. NU-MCC [252] similarly utilizes latent point features but focuses on shape completion tasks. Unlike NeRF-style volumetric rendering, 3D Gaussian Splatting introduces point-based α-blending and an efficient point-based rasterizer. Our work follows 3D Gaussians Splatting, where we represent the scene using explicit point-based 3D representation, *i.e.* anisotropic 3D Gaussians.

7.2.3 Feature Field Distillation

Enabling simultaneously novel view synthesis and representing feature fields is well explored under NeRF [211] literature. Pioneering works such as Semantic NeRF [253] and Panoptic Lifting [254] have successfully embedded semantic data from segmentation networks into 3D spaces. Their research has shown that merging noisy or inconsistent 2D labels in a 3D environment can yield sharp and precise 3D segmentation. Further extending this idea, techniques like those presented in [255] have demonstrated the effectiveness of segmenting objects in 3D with minimal user input, like rudimentary foreground-background masks. Beyond optimizing NeRF with estimated labels, Distilled Feature Fields [13], NeRF-SOS [235], LERF [14], and Neural Feature Fusion Fields [236] have delved into embedding pixel-aligned

feature vectors from technologies such as LSeg or DINO [256] into NeRF frameworks. Additionally, [257, 258, 259, 260, 261, 262, 263] also explore feature fusion and manipulation in 3D. Feature 3DGS shares a similar idea for distilling 2D well-trained models, but also demonstrates an effective way of distilling into explicit point-based 3D representations, for simultaneous photo-realistic view synthesis and label map rendering.

7.3 Method

NeRF-based feature field distillation, as explored in [13], utilizes two distinct branches of MLPs to output the color c and feature f. Subsequently, the RGB image and high-dimensional feature map are rendered individually through volumetric rendering. The transition from NeRF to 3DGS is not as straightforward as simply rasterizing RGB images and feature maps independently. Typically, feature maps have fixed dimensions that often differ from those of RGB images. Due to the tile-based rasterization procedure and shared attributes between images and feature maps, rendering them independently can be problematic. A naive approach is to adopt a two-stage training method that rasterizes them separately. However, this approach could result in suboptimal quality for both RGB images and feature maps, given the high-dimensional correlations of semantic features with the shared attributes of RGB.

In this section, we introduce a novel pipeline for high-dimensional feature rendering and feature field distillation, which enables 3D Gaussians to explicitly represent both radiance fields and feature fields. Our proposed parallel N-dimensional Gaussian rasterizer and speed-up module can effectively solve the aforementioned problems and is capable of rendering arbitrary dimensional semantic feature map. An overview of our method is shown in Fig. 7.2. Our proposed method is general and compatible with any 2D foundation model, by distilling the semantic features into a 3D feature field using 3D Gaussian splatting. In our experiments, we employ SAM [264] and LSeg [265], facilitating promptable, promptless

(zero-shot [266] [267]) and language-driven computer vision tasks in a 3D context.

7.3.1 High-dimensional Semantic Feature Rendering

To develop a general feature field distillation pipeline, our method should be able to render 2D feature maps of arbitrary size and feature dimension, in order to cope with different kinds of 2D foundation models. To achieve this, we use the rendering pipeline based on the differentiable Gaussian splatting framework proposed by [237] as our foundation. We follow the same 3D Gaussians initialization technique using Structure from Motion [268]. Given this initial point cloud, each point $x \in \mathbb{R}^3$ within it can be described as the center of a Gaussian. In world coordinates, the 3D Gaussians are defined by a full 3D covariance matrix Σ, which is transformed to Σ' in camera coordinates when 3D Gaussians are projected to 2D image / feature map space [269]:

$$\Sigma' = JW\Sigma W^T J^T, \tag{7.1}$$

where W is the world-to-camera transformation matrix and J is the Jacobian of the affine approximation of the projective transformation. Σ is physically meaningful only when it is positive semi-definite — a condition that cannot always be guaranteed during optimization. This issue can be addressed by decomposing Σ into rotation matrix R and scaling matrix S:

$$\Sigma = RSS^T R^T, \tag{7.2}$$

Practically, the rotation matrix R and the scaling matrix S are stored as a rotation quaternion $q \in \mathbb{R}^4$ and a scaling factor $s \in \mathbb{R}^3$ respectively. Besides the aforementioned optimizable parameters, an opacity value $\alpha \in \mathbb{R}$ and spherical harmonics (SH) up to the 3rd order are also stored in the 3D Gaussians. In practice, we optimize the zeroth-order SH for the first 1000 iterations, which equates to a simple diffuse color representation $c \in \mathbb{R}^3$, and we introduce 1 band every 1000 iterations until all 4 bands of SH are represented. Additionally, we incorporate the semantic feature $f \in \mathbb{R}^N$, where N can be any arbitrary

number representing the latent dimension of the feature. In summary, for the $i-$th 3D Gaussian, the optimizable attributes are given by $\Theta_i = \{x_i, q_i, s_i, \alpha_i, c_i, f_i\}$.

Upon projecting the 3D Gaussians into a 2D space, the color C of a pixel and the feature value F_s of a feature map pixel are computed by volumetric rendering which is performed using front-to-back depth order [270]:

$$C = \sum_{i \in \mathcal{N}} c_i \alpha_i T_i, \quad F_s = \sum_{i \in \mathcal{N}} f_i \alpha_i T_i, \tag{7.3}$$

where $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$, $\mathcal{N}$ is the set of sorted Gaussians overlapping with the given pixel, T_i is the transmittance, defined as the product of opacity values of previous Gaussians overlapping the same pixel. The subscript s in F_s denotes “student", indicating that this rendered feature is per-pixel supervised by the “teacher" feature F_t. The latter represents the latent embedding obtained by encoding the ground truth image using the encoder of 2D foundation models. This supervisory relationship underscores the instructional dynamic between F_s and F_t in our model. In essence, our approach involves distilling [271] the large 2D teacher model into our small 3D student explicit scene representation model through differentiable volumetric rendering.

In the rasterization stage, we adopted a joint optimization method, as opposed to rasterizing the RGB image and feature map independently. Both image and feature map utilize the same tile-based rasterization procedure, where the screen is divided into 16×16 tiles, and each thread processes one pixel. Subsequently, 3D Gaussians are culled against both the view frustum and each tile. Owing to their shared attributes, both the feature map and RGB image are rasterized to the same resolution but in different dimensions, corresponding to the dimensions of c_i and f_i initialized in the 3D Gaussians. This approach ensures that the fidelity of the feature map is rendered as high as that of the RGB image, thereby preserving per-pixel accuracy.

7.3.2 Optimization and Speed-up

The loss function is the photometric loss combined with the feature loss:

$$\mathcal{L} = \mathcal{L}_{rgb} + \gamma \mathcal{L}_f, \tag{7.4}$$

with

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\mathcal{L}_1(I, \hat{I}) + \lambda \mathcal{L}_{D-SSIM}(I, \hat{I}),$$
$$\mathcal{L}_f = \|F_t(I) - F_s(\hat{I})\|_1.$$

where I is the ground truth image and $\hat{I}$ is our rendered image. The latent embedding $F_t(I)$ is derived from the 2D foundation model by encoding the image I, while $F_s(\hat{I})$ represents our rendered feature map. To ensure identical resolution $H \times W$ for the per-pixel $\mathcal{L}_1$ loss calculation, we apply bilinear interpolation to resize $F_s(\hat{I})$ accordingly. In practice, we set the weight hyperparameters $\gamma = 1.0$ and $\lambda = 0.2$.

It is important to note that in NeRF-based feature field distillation, the scene is implicitly represented as a neural network. In this configuration, as discussed in [13], the branch dedicated to the feature field shares some layers with the radiance field. This overlap could potentially lead to interference, where learning the feature fields might adversely affect the radiance fields. To address this issue, a compromise approach is to set γ to a low value, meaning the weight of the feature field is much smaller than that of the radiance field during the optimization. [13] also mentions that NeRF is highly sensitive to γ. Conversely, our explicit scene representation avoids this issue. Our equal-weighted joint optimization approach has demonstrated that the resulting high-dimensional semantic features significantly contribute to scene understanding and enhance the depiction of physical scene attributes, such as opacity and relative positioning. See the comparison between Ours and Base 3DGS in Tab. 7.1.

To optimize the semantic feature $f \in \mathbb{R}^N$, we minimize the difference between the rendered feature map $F_s(\hat{I}) \in \mathbb{R}^{H \times W \times N}$ and the teacher feature map $F_t(I) \in \mathbb{R}^{H \times W \times M}$, ideally

with $N = M$. However, in practice, M tends to be a very large number due to the high latent dimensions in 2D foundation models (e.g. $M = 512$ for LSeg and $M = 256$ for SAM), making direct rendering of such high-dimensional feature maps time-consuming. To address this issue, we introduce a speed-up module at the end of the rasterization process. This module consists of a lightweight convolutional decoder that upsamples the feature channels with kernel size 1×1. Consequently, it is feasible to initialize $f \in \mathbb{R}^N$ on 3D Gaussians with any arbitrary $N \ll M$ and to use this learnable decoder to match the feature channels. This allows us to not only effectively achieve $F_s(\hat{I}) \in \mathbb{R}^{H\times W\times M}$, but also significantly speed up the optimization process without compromising the performance on downstream tasks.

The advantages of implementing this convolutional speed-up module are threefold: Firstly, the input to the convolution layer, with a kernel size of 1×1, is the resized rendered feature map, which is significantly smaller in size compared to the original image. This makes the 1×1 convolution operation computationally efficient. Secondly, this convolution layer is a learnable component, facilitating channel-wise communication within the high-dimensional rendered feature, enhancing the feature representation. Lastly, the module's design is optional. Whether included or not, it does not impact the performance of downstream tasks, thereby maintaining the flexibility and adaptability of the entire pipeline.

7.3.3 Promptable Explicit Scene Representation

Foundation models provide a base layer of knowledge and skills that can be adapted for a variety of specific tasks and applications. We wish to use our feature field distillation approach to enable practical 3D representations of these features. Specifically, we consider two foundation models, namely Segment Anything [264], and LSeg [265]. The *Segment Anything Model (SAM)* [264] allows for both promptable and promptless zero-shot segmentation in 2D, without the need for specific task training. LSeg [265] introduces a language-driven approach to zero-shot semantic segmentation. Utilizing the image feature encoder with the DPT architecture [272] and text encoders from CLIP [273], LSeg extends text-image associa-

tions to a 2D pixel-level granularity. Through the teacher-student distillation, **our distilled feature fields facilitate the extension of all 2D functionalities — prompted by point, box, or text — into the 3D realm.**

Our promptable explicit scene representation works as follows: for a 3D Gaussian x among the N ordered Gaussians overlapping the target pixel, i.e. $x_i \in \mathcal{X}$ where $\mathcal{X} = \{x_1, \ldots, x_N\}$, the activation score of a prompt τ on the 3D Gaussian x is calculated by cosine similarity between the query $q(\tau)$ in the feature space and the semantic feature $f(x)$ of the 3D Gaussian followed by a softmax:

$$s = \frac{f(x) \cdot q(\tau)}{\|f(x)\|\|q(\tau)\|}, \tag{7.5}$$

If we have a set $\mathcal{T}$ of possible labels, such as a text label set for semantic segmentation or a point set of all the possible pixels for point-prompt, the probability of a prompt τ of a 3D Gaussian can be obtained by softmax:

$$\mathbf{p}(\tau|x) = \text{softmax}(s) = \frac{\exp(s)}{\sum_{s_j \in \mathcal{T}} \exp(s_j)}. \tag{7.6}$$

We utilize the computed probabilities to filter out Gaussians with low probability scores. This selective approach enables various operations, such as extraction, deletion, or appearance modification, by updating the color $c(x)$ and opacity $\alpha(x)$ values as needed. With the newly updated color set $\{c_i\}_{i=1}^n$ and opacity set $\{\alpha_i\}_{i=1}^n$, where n is smaller than N, we can implement point-based α-blending to render the edited radiance field from any novel view.

7.4 Experiments

7.4.1 Novel view semantic segmentation

The number of classes of a dataset is usually limited from tens [274] to hundreds [275], which is insignificant to English words [276]. In light of the limitation, semantic features

Metrics	PSNR(±s.d.)↑	SSIM(±s.d.)↑	LPIPS(±s.d.)↓
Ours (w/ speed-up)	**37.012** (±0.07)	**0.971** (±5.3e-4)	**0.023** (±2.9e-4)
Ours	36.915 (±0.05)	0.970 (±5.7e-4)	0.024 (±1.1e-3)
Base 3DGS	36.133 (±0.06)	0.965 (±1.5e-4)	0.033 (±1.2e-3)

Table 7.1: **Performance on Replica Dataset.** (average performance for 5K training iterations, speed-up module rendered feature $dim = 128$). Boldface font represents the preferred results.

Metrics	mIoU↑	accuracy↑	FPS↑
Ours (w/ speed-up)	0.782	**0.943**	**14.55**
Ours	**0.787**	**0.943**	6.84
NeRF-DFF	0.636	0.864	5.38

Table 7.2: **Performance of semantic segmentation on Replica dataset compared to NeRF-DFF.** (speed-up module rendered feature $dim = 128$). Boldface font represents the preferred results.

empower models to comprehend unseen labels by mapping semantically close labels to similar regions in the embedding space, as articulated by Li et al [265]. This advancement notably promotes the scalability in information acquisition and scene understanding, facilitating a profound comprehension of intricate scenes. We distill LSeg feature for this novel view semantic segmentation task. Our experiments demonstrate the improvement of incorporating semantic feature over the naive 3D Gaussian rasterization method [237]. In Tab. 7.1, we show that our model surpasses the baseline 3D Gaussian model in performance metrics on Replica dataset [11] with 5000 training iterations for all three models. Noticeably, the integration of

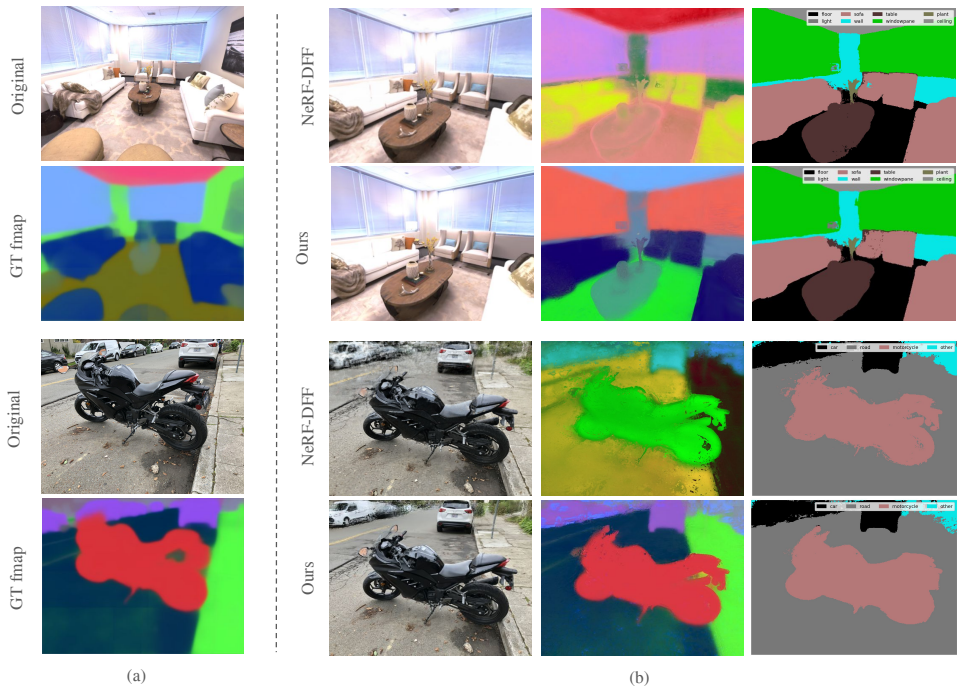


Figure 7.3: **Novel view semantic segmentation (LSeg) results on scenes from Replica dataset [11] and LLFF dataset [12].** (a) We show examples of original images in training views together with the ground-truth feature visualizations. (b) We compare the qualitative segmentation results using our Feature 3DGS with the NeRF-DFF [13]. Our inference is **1.66**× faster when rendered feature $dim = 128$. Our method demonstrates more fine-grained segmentation results with higher-quality feature maps.

the speed-up module to our model does not compromise the performance.

In our further comparison with NeRF-DFF [13] using the Replica dataset, we address the potential trade-off between the quality of the semantic feature map and RGB images. In Tab. 7.2, our model demonstrates higher accuracy and mean intersection-over-union (mIoU). Additionally, by incorporating our speed-up module, we achieved more than double the frame rate per second (FPS) of our full model while maintaining comparable performance. In Fig. 7.3 (b) the last column, our approach yields better visual quality on novel views and semantic segmentation masks for both synthetic and real scenes compared to

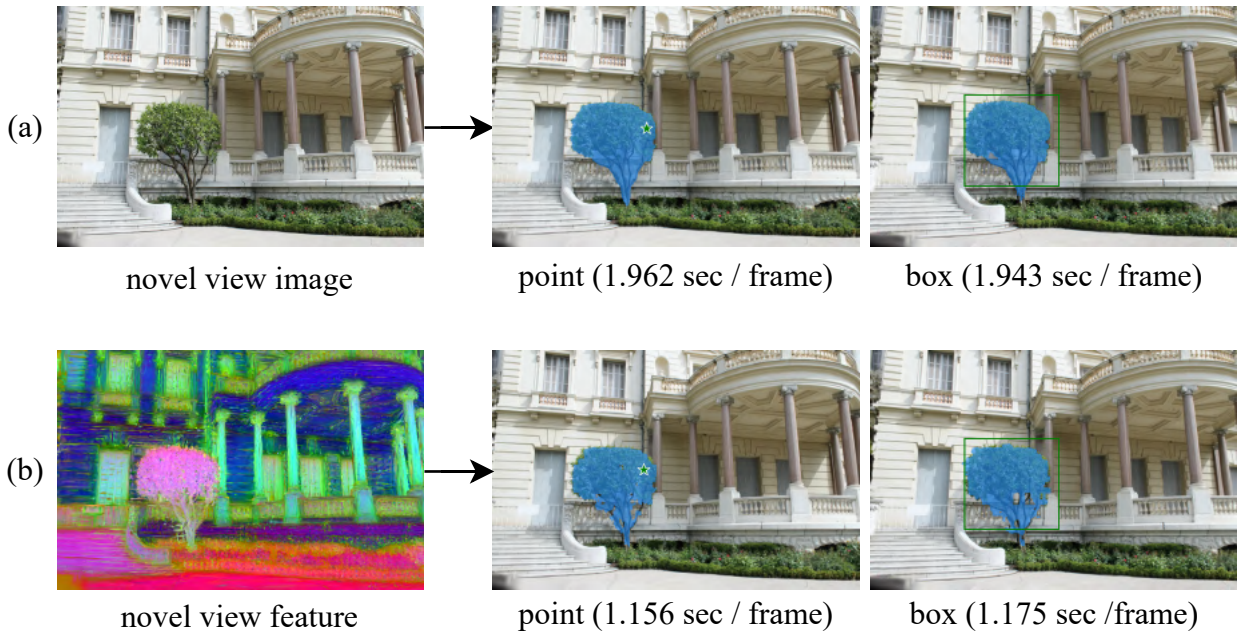


Figure 7.4: **Comparison of SAM segmentation results obtained by** (a) naively applying the SAM encoder-decoder module to a novel-view rendered image **with** (b) directly decoding a rendered feature. Our method is up to 1.7× faster in total inference speed including rendering and segmentation while preserving the quality of segmentation masks. Scene from [10].

NeRF-DFF.

7.4.2 Segment Anything from Any View

SAM excels in performing precise instance segmentation, utilizing interactive points and boxes as prompts to automatically segment objects in any 2D image. In our experiments, we extend this capability to 3D, aiming to achieve fast and accurate segmentation from any viewpoint. Our distilled feature field enables the model to render the SAM feature map directly for any given camera pose. As such, the SAM decoder is the only component needed to interact with the input prompt and produce the segmentation mask, thereby bypassing

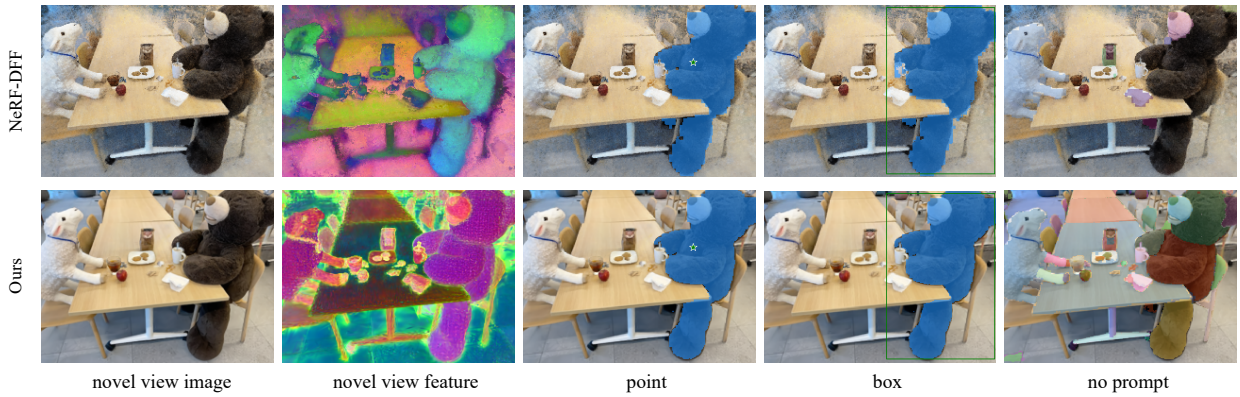


Figure 7.5: **Novel view segmentation (SAM) results compared with NeRF-DFF**. (Upper) NeRF-DFF method presents lower-quality segmentation masks - note the failure on segmenting the cup from the bear and the coarse-grained mask boundary on the bear's leg in box-prompted results. (Lower) Our method provides higher-quality masks with more fine-grained segmentation details. Scene from [14].

the need to synthesize a novel view image first and then process it through the entire SAM encoder-decoder pipeline. Furthermore, to enhance training and inference speed, we use the speed-up module in this experiment. In practice, we set the rendered feature dimension to 128, which is half of SAM's latent dimension of 256, maintaining the comparable quality of segmentation.

In Fig. 7.4, we compare the results of both point and box prompted segmentation on novel views using the naive approach (SAM encoder + decoder) and our proposed feature field approach (SAM decoder only). We achieve nearly equivalent segmentation quality, but our method is up to 1.7× faster. In Fig. 7.5, we contrast our method with NeRF-DFF [13]. Our rendered features not only yield higher quality mask boundaries, as evidenced by the bear's leg, but also deliver more accurate and comprehensive instance segmentation, capable of segmenting finer-grained instances (as illustrated by the more 'detailed' mask on the far right). Additionally, we use PCA-based feature visualization [277] to demonstrate that our high-quality segmentation masks result from superior feature rendering.

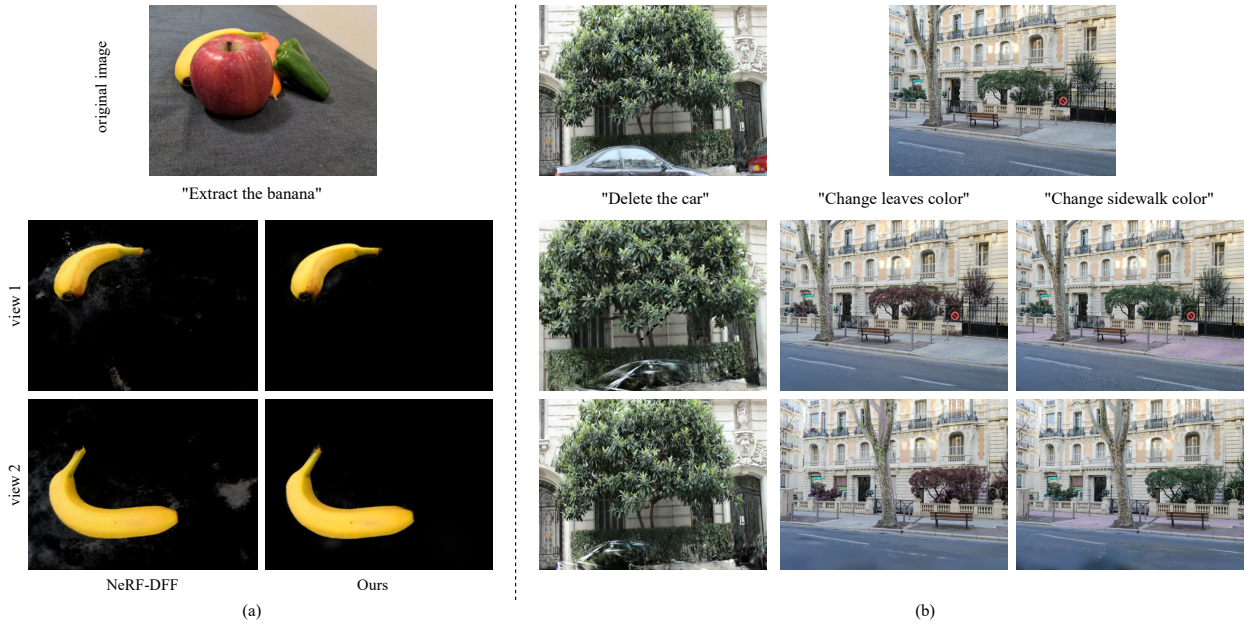


Figure 7.6: **Demonstration of results with various language-guided edit operations by querying the 3D feature field and comparison with NeRF-DFF** (a) We compare our edit results with NeRF-DFF method on the sample dataset provided by NeRF-DFF [13]. Note that our method outperforms NeRF-DFF method by extracting the entire banana hidden by an apple in the original image and with less floaters in the background. (b) We demonstrate results with deletion and appearance modification on different targets. Note that the car is deleted with background preserved, and the appearance of the leaves changes with the appearance of the stop sign remained the same.

7.4.3 Language-guided Editing

In this section, we showcase the capability of our Feature 3DGS, distilled from LSeg, to perform editable novel view synthesis. The process begins by querying the feature field with a text prompt, typically comprising an edit operation followed by a target object, such as “extract the car". For text encoding, we employ a ViT-B/32 CLIP encoder. Our editing pipeline capitalizes on the semantic features queried from the 3D feature field, rather than relying on a 2D rendered feature map. We compute semantic scores for each 3D Gaussian,

represented by a K-dimensional vector (where K is the number of object categories), using a softmax function. Subsequently, we engage in either soft selection (by setting a threshold value) or hard selection (by filtering based on the highest score across K categories). To identify the region for editing, we generate a “Gaussian mask" through thresholding on the score matrix, which is then utilized for modifications to color c and opacity α on 3D Gaussians.

In Fig. 7.6, we showcase our novel view editing results, achieved through various operations prompted by language inputs. Specifically, we conduct editing tasks such as extraction, deletion, and appearance modification on text-specified targets within diverse scenes. As illustrated in Fig. 7.6 (a), we successfully extract an entire banana from the scene. Notably, by leveraging 3D Gaussians to update the rendering parameters, our Feature 3DGS model gains an understanding of the 3D scene environment from any viewpoint. This enables the model to reconstruct occluded or invisible parts of the scene, as evidenced by the complete extraction of a banana initially hidden by an apple in view 1. Furthermore, compared with our edit results with NeRF-DFF, our method stands out by providing a cleaner extraction with little floaters in the background. Additionally, in Fig. 7.6 (b), our model is able to delete objects like cars while retaining the background elements, such as plants, due to the opacity updates in 3DGS, showcasing its 3D scene awareness. Moreover, we also demonstrate the model's capability in modifying the appearance of specific objects, like ‘sidewalk’ and ‘leaves’, without affecting adjacent objects’ appearance (e.g., the ‘stop sign’ remains red).

7.5 Discussion and Conclusion

In this work, we present a notable advancement in explicit 3D scene representation by integrating 3D Gaussian Splatting with feature field distillation from 2D foundation models, a development that not only broadens the scope of radiance fields beyond traditional uses but also addresses key limitations of previous NeRF-based methods in implicitly represented fea-

ture fields. Our work, as showcased in various experiments including complex semantic tasks like editing, segmentation, and language-prompted interactions with models like CLIP-LSeg and SAM, opening the door to a brand new semantic, editable, and promptable explicit 3D scene representation.

However, our Feature 3DGS framework does have its inherent limitations. The student feature's limited access to the ground truth feature restricts the overall performance, and the imperfections of the teacher network further constrain our framework's effectiveness. In addition, our adaptation of the original 3DGS pipeline, which inherently generates noise-inducing floaters, poses another challenge, affecting our model's optimal performance.

CHAPTER 8

Conclusion

In this dissertation, we explore the paradigm of "neural sensing": computatational imaging in the era of AI and large scale models. We explore each of the three pillars of neural sensing (the sensor, the data and the algorithm) in the context of specific challenges afforded in contactless patient vital signs monitoring and beyond.

We first show that aspects of equity (such as skin tone equity) are dictated by the sensing principle of the selected sensor. Inequities that originate at the sensor level cannot be eradicated regardless of the amount of data, or the inference algorithm in use. Therefore, design of equitable approaches requires rethinking sensor design itself. This philosophy of sensor-first design extends beyond equity, to new applications such as touch sensing as well.

We next explore the importance of data at scale for neural sensing pipelines. In the context of medical imaging, where data at scale is challenging to obtain, we find that carefully synthesised "physiorealistic" synthetic humans can pave the path for leveraging the benefits of data scale. We also explore solutions for when data size is difficult to scale.

Finally, we explore the inference algorithm. Specifically, we explore extreme sensing scenarios with very low signal to noise ratio and show that prudently designed algorithms are capable of elevating the performance of neural sensing pipelines. We also show contributions towards more general computer vision tasks such as 3D sensing.

Through all this, we show that understanding and exploring each of these three pillars is critical to enable deployment of a neural sensing pipeline. Future explorations as part of this thesis will aim to expand all three pillars to broader visual applications.

APPENDIX A

Supplemental Content: Minority Inclusion for Majority Enhancement of AI Performance

Supplementary Contents

This supplement is organized as follows:

- Section A.1 contains the proof for Theorem 1.
- Section A.2 contains the proof for Theorem 2.
- Section A.3 contains the proof for Theorem 3.
- Section A.4 discusses MIME existence beyond 1D.
- Section A.5 describes further details for the feature space analysis.
- Section A.6 contains implementation details across the six datasets.
- Section A.7 describes additional secondary analysis.
- Section A.8 describes our implementation of the hard mining comparison.
- Section A.9 describes our code.
- Section A.10 discusses potential negative ethical impacts of this work.

A.1 Proof for Theorem 1

We consider the one-dimensional linear classifier setting, trained using the Perceptron algorithm. Given any $x \in \mathbb{R}$, the classifier evaluates an output y given by,

$$y = wx + b, \tag{A.1}$$

where $w, b \in \mathbb{R}$. The decision threshold in this case is at $y = 0$. For simplification, we reduce the redundant parameter, as follows:

$$y' = x + b'. \tag{A.2}$$

Note that the decision threshold is unaffected by this conversion. For notational simplicity, we use $y = y'$ and $b = b'$ here onward. We consider the perceptron decision and update rule, modified for our case. That is, for any training sample (x_i, y_i), the predicted output is given by,

$$\hat{y}_i = \frac{sign(x_i + b) + 3}{2}, \tag{A.3}$$

where $sign(\cdot)$ is the sign function. Readers will notice the unconventional form of this decision rule. The additional terms map the conventional perceptron labels in $\{-1, 1\}$ to our chosen labels $\{1, 2\}$ respectively.

For an appropriately chosen learning rate γ, the parameter update rule for this setting is given by:

$$b \leftarrow \begin{cases} b + \gamma, & \text{if } \hat{y}_i \neq y_i \text{ and } y_i = 2 \\ b - \gamma, & \text{if } \hat{y}_i \neq y_i \text{ and } y_i = 1 \end{cases}. \tag{A.4}$$

Let $\mathbf{h}_{\text{ideal}} \triangleq [1,\, b_{\text{ideal}}]^T$ denote the ideal decision hyperplane. Under the current assumption of no domain gap, it can be shown that this ideal hyperplane is located at $x = d_{\text{ideal}}$ such that,

$$\begin{aligned} p_1^{\text{minor}}(d_{\text{ideal}}) &= p_2^{\text{minor}}(d_{\text{ideal}}) \\ p_1^{\text{major}}(d_{\text{ideal}}) &= p_2^{\text{major}}(d_{\text{ideal}}). \end{aligned} \tag{A.5}$$

This also implies that $b_{\text{ideal}} = -d_{\text{ideal}}$. Now, consider an initial training set of $K-1$ samples from the majority group, $\mathcal{D}_{K-1}^{\text{major}}$. A decision hyperplane $\mathbf{h}_{K-1}$ is learnt from these samples. Then, without loss of generality, we can assume that,

$$d_{K-1} = d_{\text{ideal}} + \Delta. \tag{A.6}$$

That is, the real hyperplane $\mathbf{h}_{K-1}$ is non-ideally located closer to the positve class ($y = 2$) than $\mathbf{h}_{\text{ideal}}$. Δ is a small positive value representing the error in the learnt decision hyperplane. Consider that the K-th sample is drawn from the majority group x_K^{major}. Recall that parameter updates for the Perceptron algorithm take place only in the event of incorrect label estimation $\hat{y}_K \neq y_K$. If we denote the change in the parameter b due to this sample as Δb, then three cases exist:

1. Sample from class 2 is classified as belonging to class 1 such that $x_K^{\text{major}} \sim p_2^{\text{minor}}(x),\ x_K^{\text{major}} < d_{\text{ideal}} - \Delta$. Associated $\Delta b = +\gamma$.

2. Sample from class 2 is classified as belonging to class 1 such that $x_K^{\text{major}} \sim p_2^{\text{minor}}(x),\ d_{\text{ideal}} - \Delta \leq x_K^{\text{major}} < d_{\text{ideal}} + \Delta$. Associated $\Delta b = +\gamma$.

3. Sample from class 1 classified as belonging to class 2 such that $x_K^{\text{major}} \sim p_1^{\text{minor}}(x),\ x_K^{\text{major}} \geq d_{\text{ideal}} + \Delta$. Associated $\Delta b = -\gamma$.

Let the expected change in b due to one majority group sample be denoted as Δb^{major}. Δd^{major} is similarly defined for the expected change in d. Then, the following holds true:

$$\Delta b^{\text{major}} = \mathop{\mathbb{E}}_{x_K^{\text{major}}} [\Delta b]\,. \tag{A.7}$$

Writing out the expectation over all three cases,

$$\Delta b^{\text{major}} = \gamma \int_{x=-\infty}^{d_{\text{ideal}}-\Delta} p_2^{\text{major}}(x)dx + \gamma \int_{x=d_{\text{ideal}}-\Delta}^{d_{\text{ideal}}+\Delta} p_2^{\text{major}}(x)dx - \gamma \int_{x=d_{\text{ideal}}+\Delta}^{+\infty} p_1^{\text{major}}(x)dx. \tag{A.8}$$

Similar expressions can be identified if the K-th sample is drawn from the minority group. Under the assumption that the mixture models under consideration are symmetric Gaussian mixture models,

$$\int_{x=-\infty}^{d_{\text{ideal}}-\Delta} p_2^{\text{major}}(x)dx = \int_{x=d_{\text{ideal}}+\Delta}^{+\infty} p_1^{\text{major}}(x)dx. \tag{A.9}$$

Then, using Equation A.8 and Equation A.9,

$$\Delta b^{\text{major}} = \gamma \int_{x=d_{\text{ideal}}-\Delta}^{d_{\text{ideal}}+\Delta} p_2^{\text{major}}(x)dx. \tag{A.10}$$

The region between $x = d_{\text{ideal}} - \Delta$ and $d_{\text{ideal}} + \Delta$ determines the expected change in the classification parameter. If Δ is small enough, $\Delta b^{\text{major}} \approx 2\gamma p_2^{\text{major}}(d_{\text{ideal}})\Delta$. Similarly, $\Delta b^{\text{minor}} \approx 2\gamma p_2^{\text{minor}}(d_{\text{ideal}})\Delta$.

We now identify a sufficient condition where $p_2^{\text{minor}}(x) > p_2^{\text{major}}(x)$ for $-\Delta \leq x \leq \Delta$, given that the overlaps satisfy the condition $O_{\text{minor}} > O_{\text{major}}$, as defined in the main text. Under the GMM assumption,

$$p_2^{\text{major}}(x) = \frac{1}{\sqrt{2\pi(\sigma_2^{\text{major}})^2}} exp\left(-\frac{(x-\mu_2^{\text{major}})^2}{2(\sigma_2^{\text{major}})^2}\right). \tag{A.11}$$

A similar expression exists for the minority group distribution as well. We wish to find the intersection point for the majority and minority distributions, that is $p_2^{\text{major}}(x) = p_1^{\text{major}}(x)$ for some x. This expression reduces to,

$$\frac{\left(x-\mu_2^{\text{major}}\right)^2}{\sigma^{\text{major}^2}} - \frac{\left(x-\mu_2^{\text{minor}}\right)^2}{\sigma^{\text{minor}^2}} = 2ln\left|\frac{\sigma^{\text{minor}}}{\sigma^{\text{major}}}\right|. \tag{A.12}$$

We want to ensure that this intersection point occurs for an $x > d_{\text{ideal}}$. This sets up a hyperbolic equation for the condition. For our purposes of proving existence, we qualitatively note that if the majority group variance is not very large (meaning the likelihood of sampling at the ideal hyperplane is low for the majority group), and the minority group variance is not

very large (such that it does not tend close to a uniform distribution), $p_2^{\text{minor}}(x) > p_2^{\text{major}}(x)$. Then,

$$\Delta b^{\text{minor}} > \Delta b^{\text{major}}. \tag{A.13}$$

$$\Delta d^{\text{minor}} < \Delta d^{\text{major}} < 0. \tag{A.14}$$

Our final task is to relate the expected change in the decision hyperplane over a choice of training sets $\mathcal{D}_K^+$ and $\mathcal{D}_K^-$, with associated learnt hyperplanes $\mathbf{h}_K^+$ and $\mathbf{h}_K^-$. As a reminder,

$$\begin{aligned} \mathcal{D}_K^+ &= \{\mathcal{D}_{K-1}^{\text{major}}, x_K^{\text{major}}\} \\ \mathcal{D}_K^- &= \{\mathcal{D}_{K-1}^{\text{major}}, x_K^{\text{minor}}\}, \end{aligned} \tag{A.15}$$

Consider a general training setting, where we use minibatches of size $M > 1$, over multiple epochs. Then, any minibatch containing the K-th sample can be split into the K-th sample and a random subset of $M-1$ samples from $\mathcal{D}_{K-1}^{\text{major}}$. Therefore, on average, the only difference to the sample updates would be due to the contributions of the K-th sample. This brings us to our final observations,

$$\begin{aligned} \mathop{\mathbb{E}}_{x_K^{\text{minor}}}[d_K^+] &= d_{K-1}^{\text{major}} + \Delta d^{\text{major}} \\ \mathop{\mathbb{E}}_{x_K^{\text{minor}}}[d_K^-] &= d_{K-1}^{\text{major}} + \Delta d^{\text{minor}}. \end{aligned} \tag{A.16}$$

From Equations A.14 and A.16,

$$\mathop{\mathbb{E}}_{x_K^{\text{minor}}}[d_K^-] < \mathop{\mathbb{E}}_{x_K^{\text{minor}}}[d_K^+], \text{ and} \tag{A.17}$$

$$\mathop{\mathbb{E}}_{x_K^{\text{minor}}}[|d_{\text{ideal}} - d_K^-|] < \mathop{\mathbb{E}}_{x_K^{\text{minor}}}[|d_{\text{ideal}} - d_K^+|]. \tag{A.18}$$

The above holds for small enough γ. Since we know the relationship between the decision hyperplane $\mathbf{h}$ and the associated d in our setting, the following equations hold true:

$$\mathop{\mathbb{E}}_{x_K^{\text{minor}}}\|\mathbf{h}_{\text{ideal}} - \mathbf{h}_K^-\| < \mathop{\mathbb{E}}_{x_K^{\text{minor}}}\|\mathbf{h}_{\text{ideal}} - \mathbf{h}_K^+\|, \tag{A.19}$$

$$\mathop{\mathbb{E}}_{x_K^{\text{minor}}} \mathcal{P}^{\text{major}}\{\mathbf{h}_K^-\} < \mathop{\mathbb{E}}_{x_K^{\text{major}}} \mathcal{P}^{\text{major}}\{\mathbf{h}_K^+\}. \blacksquare \tag{A.20}$$

A.2 Proof for Theorem 2

We follow a similar approach as in Theorem 1. Let $\mathbf{h}_{\text{ideal}}^{\text{major}} \triangleq [1, b_{\text{ideal}}^{\text{major}}]^T$ denote the ideal decision hyperplane for the majority group. Let $\mathbf{h}_{\text{ideal}}^{\text{minor}} \triangleq [1, b_{\text{ideal}}^{\text{minor}}]^T$ denote the ideal decision hyperplane for the minority group. Then, the ideal hyperplanes are located at $x = d_{\text{ideal}}^{\text{major}}$ and $x = d_{\text{ideal}}^{\text{minor}}$ respectively such that,

$$\begin{aligned} p_1^{\text{minor}}(d_{\text{ideal}}^{\text{minor}}) &= p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}}) \\ p_1^{\text{major}}(d_{\text{ideal}}^{\text{major}}) &= p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}}). \end{aligned} \tag{A.21}$$

This implies that $b_{\text{ideal}}^{\text{major}} = -d_{\text{ideal}}^{\text{major}}$ and $b_{\text{ideal}}^{\text{minor}} = -d_{\text{ideal}}^{\text{minor}}$. Consider an initial training set of $K-1$ samples from the majority group, $\mathcal{D}_{K-1}^{\text{major}}$. Then, without loss of generality, we can assume that $d_{K-1} = d_{\text{ideal}}^{\text{major}} + \Delta$, where $\Delta > 0$. Additionally, we consider the existence of domain gap in this case, that is, $d_{\text{ideal}}^{\text{minor}} = d_{\text{ideal}}^{\text{major}} + \delta$.

Let $\delta < \Delta$. Similar to the setting in Theorem 1 (Equation A.8), we can set up the equation for expected parameter change in the case of the majority and minority groups as follows:

$$\begin{aligned} \Delta b^{\text{major}} = \gamma \int_{x=-\infty}^{d_{\text{ideal}}^{\text{major}}-\Delta} p_2^{\text{major}}(x)dx + \gamma \int_{x=d_{\text{ideal}}^{\text{major}}-\Delta}^{d_{\text{ideal}}^{\text{major}}+\Delta} p_2^{\text{major}}(x)dx \\ - \gamma \int_{x=d_{\text{ideal}}^{\text{major}}+\Delta}^{+\infty} p_1^{\text{major}}(x)dx. \end{aligned} \tag{A.22}$$

$$\begin{aligned} \Delta b^{\text{minor}} = \gamma \int_{x=-\infty}^{d_{\text{ideal}}^{\text{minor}}-(\Delta-\delta)} p_2^{\text{minor}}(x)dx + \gamma \int_{x=d_{\text{ideal}}^{\text{minor}}-(\Delta-\delta)}^{d_{\text{ideal}}^{\text{minor}}+(\Delta-\delta)} p_2^{\text{minor}}(x)dx \\ - \gamma \int_{x=d_{\text{ideal}}^{\text{minor}}+(\Delta-\delta)}^{+\infty} p_1^{\text{minor}}(x)dx. \end{aligned} \tag{A.23}$$

Under the assumption that the mixture models under consideration are symmetric Gaussian mixture models,

$$\Delta b^{\text{major}} = \gamma \int_{x=d_{\text{ideal}}^{\text{major}}-\Delta}^{d_{\text{ideal}}^{\text{major}}+\Delta} p_2^{\text{major}}(x)dx, \tag{A.24}$$

$$\Delta b^{\text{minor}} = \gamma \int_{x=d_{\text{ideal}}^{\text{minor}}-(\Delta-\delta)}^{d_{\text{ideal}}^{\text{minor}}+(\Delta-\delta)} p_2^{\text{minor}}(x)dx. \tag{A.25}$$

If $\Delta + |\delta|$ is small enough,

$$\Delta b^{\text{major}} \approx 2\gamma p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}})\Delta, \tag{A.26}$$

$$\Delta b^{\text{minor}} \approx 2\gamma p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}})(\Delta - \delta). \tag{A.27}$$

By establishing the same conditions on group class variances as Theorem 1, we know that $p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}}) > p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}})$. We now identify conditions under which $\Delta b^{\text{minor}} > \Delta b^{\text{major}}$.

Case 1 - $\delta < 0$: Under the same conditions as Theorem 1, $(\Delta - \delta) > \Delta$, and $p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}}) > p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}})$. Therefore,

$$\Delta b^{\text{minor}} > \Delta b^{\text{major}}. \tag{A.28}$$

Case 2 - $\delta > 0$:

$$\Delta b^{\text{major}} \approx 2\gamma p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}})\Delta, \tag{A.29}$$

$$\Delta b^{\text{minor}} \approx 2\gamma p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}})(\Delta - \delta). \tag{A.30}$$

For $\Delta b^{\text{minor}} > \Delta b^{\text{major}}$,

$$p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}})(\Delta - \delta) > p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}})\Delta. \tag{A.31}$$

Rearranging Equation A.31,

$$\frac{p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}})}{p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}})} < \left(1 - \frac{\delta}{\Delta}\right). \tag{A.32}$$

Given the definitions of the majority and minority groups,

$$p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}}) < p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}}), \tag{A.33}$$

$$O_{\text{major}} < O_{\text{minor}}. \tag{A.34}$$

Since all four of these terms depend only on the means and variances of the Gaussian components, we can write,

$$\frac{O_{\text{major}}}{O_{\text{minor}}} = \frac{p_2^{\text{major}}(d_{\text{ideal}}^{\text{major}})}{p_2^{\text{minor}}(d_{\text{ideal}}^{\text{minor}})} f, \tag{A.35}$$

where f is a positive scalar constant that depends only on the component means and variances. From Equations A.32 and A.35,

$$\frac{O_{\text{major}}}{O_{\text{minor}}} < \left(1 - \frac{\delta}{\Delta}\right) f. \tag{A.36}$$

This proves the conditions in the theorem. Theorem 1 can now be used to show the existence of the MIME effect in the presence of domain gap, for these conditions. ■

A Note on the Theorems: Theorems 1 and 2 are existence theorems. That is, they show that there exist certain conditions under which the MIME effect can be observed. The theorems make these arguments based on the 'usefulness' of points close to the ideal hyperplane. The direct metric of correlation is the likelihood for a particular distribution to sample at the ideal hyperplane. However, since this cannot be easily measured in practice, we set up our proofs in terms of a correlated metric: the overlap.

A.3 Proof for Theorem 3

This Theorem considers distributions with general prior distributions. Therefore, for the majority group, let

$$\begin{aligned} p_2^{\text{major}'}(x) &= \pi^{\text{major}} p_2^{\text{major}}(x), \\ p_q^{\text{major}'}(x) &= (1-\pi^{\text{major}}) p_1^{\text{major}}(x). \end{aligned} \tag{A.37}$$

Similar definitions are made for the minority group as well. Then, assuming $d_{K-1} = d_{\text{ideal}}^{\text{major}} + \Delta$, $\Delta > 0$ (similar to Theorem 2), and $\delta = 0$ (for now), and drawing from Equation A.8), we can set up the equation for expected parameter change in the case of the majority group as follows:

$$\begin{aligned} \Delta b^{\text{major}} &= \gamma \int_{x=-\infty}^{d_{\text{ideal}}+\Delta} p_2^{\text{major}'}(x)dx - \gamma \int_{x=d_{\text{ideal}}+\Delta}^{+\infty} p_1^{\text{major}'}(x)dx \\ &= T^{\text{major}}(d_{\text{ideal}} + \Delta). \end{aligned} \tag{A.38}$$

A similar expression holds true for the minority group. Then, if $T^{\text{major}}(d_{\text{ideal}} + \Delta) < T^{\text{minor}}(d_{\text{ideal}} + \Delta)$, the MIME effect will hold true.

Similarly, if $d_{K-1} = d_{\text{ideal}}^{\text{major}} - \Delta$, $\Delta > 0$,

$$\begin{aligned} \Delta b^{\text{major}} &= -\gamma \int_{x=-\infty}^{d_{\text{ideal}}-\Delta} p_2^{\text{major}'}(x)dx + \gamma \int_{x=d_{\text{ideal}}-\Delta}^{+\infty} p_1^{\text{major}'}(x)dx \\ &= -T^{\text{major}}(d_{\text{ideal}} - \Delta). \end{aligned} \tag{A.39}$$

Then, if $-T^{\text{major}}(d_{\text{ideal}} - \Delta) < -T^{\text{minor}}(d_{\text{ideal}} - \Delta)$, the MIME effect will hold true.

Combining the two expressions, for a sufficient existence condition, we get,

$$\begin{aligned} \min\left\{T^{\text{minor}}(d_{\text{ideal}} + \Delta), -T^{\text{minor}}(d_{\text{ideal}} - \Delta)\right\} > \\ \max\left\{T^{\text{major}}(d_{\text{ideal}} + \Delta), -T^{\text{major}}(d_{\text{ideal}} - \Delta)\right\}. \end{aligned} \tag{A.40}$$

This completes the proof. ■

Note that the existence proof for Theorem 3 ignores the effect of domain gap δ, in the interest of readability and brevity. A very similar existence proof can be established with domain gap. We omit the derivation and provide the final condition below (under the constraints on δ and Δ as in Theorem 2, and using the same notation):

$$\min\left\{T^{\text{minor}}(d_{\text{ideal}}^{\text{major}}+\Delta), -T^{\text{minor}}(d_{\text{ideal}}^{\text{major}}-\Delta)\right\} > \max\left\{T^{\text{major}}(d_{\text{ideal}}^{\text{major}}+\Delta), -T^{\text{major}}(d_{\text{ideal}}^{\text{major}}-\Delta)\right\}. \quad \text{(A.41)}$$

A.4 MIME Existence Beyond 1D Settings

Consider $\mathbf{x} \in \mathbb{R}^n$. The perceptron decisions are based on the metric $y = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^n$, and $y, b \in \mathbb{R}$. Similar to Theorem 1, we consider the perceptron decision and update rule. That is, for any training sample $(\mathbf{x}_i, y_i)$, the predicted label is given by,

$$\hat{y}_i = \frac{sign(\mathbf{w}^T\mathbf{x}_i + b) + 3}{2}. \quad \text{(A.42)}$$

We can rewrite this in terms of a single decision hyperplane by defining $\tilde{\mathbf{w}} = [\mathbf{w}^T\ b]^T$ and $\tilde{\mathbf{x}} = [\mathbf{x}^T\ 1]^T$. For a small learning rate γ, the updated decision rule becomes,

$$\hat{\tilde{y}}_i = \frac{sign(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_i) + 3}{2}. \quad \text{(A.43)}$$

$$\tilde{\mathbf{w}} \leftarrow \begin{cases} \tilde{\mathbf{w}} + \gamma\tilde{\mathbf{x}}_i, & \text{if } \hat{y}_i \neq y_i \text{ and } y_i = 2 \\ \tilde{\mathbf{w}} - \gamma\tilde{\mathbf{x}}_i, & \text{if } \hat{y}_i \neq y_i \text{ and } y_i = 1 \end{cases}. \quad \text{(A.44)}$$

We now refer to the hyperplane $\tilde{\mathbf{w}}$ as the decision hyperplane. Let $\mathbf{h}_{\text{ideal}}$ be the ideal decision hyperplane. In this setting, any domain gap δ or error in real hyperplane estimation Δ manifests as a direction/angle error in the hyperplane normal vector (since the bias term b is subsumed in the hyperplane). The updates change the normal vector of the hyperplane through a linear combination with the sample $\tilde{\mathbf{x}}_i$, scaled by the learning rate γ.

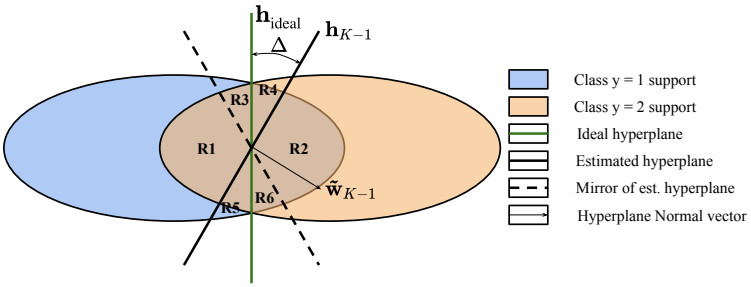


Figure A.1: **The MIME effect holds in a multidimensional setting as well.** We show the support for the two finite distributions. Weight vector updates arising out of samples from regions R3, R4, R5 and R6 lead to an update with a large vertical (corrective) component (favorable update). Updates arising out of regions R1 and R2 result in an overall update in the horizontal direction (unfavorable update).

We now provide a qualitative description for the existence of the MIME effect, in terms of the likelihood of a favorable update to $\tilde{\mathbf{w}}$. We consider a simplified 2D case with symmetric distributions and $\delta = 0$. A finite support is assumed for the majority and minority groups, for ease of understanding. Consider that the bias term b is known, and only the hyperplane direction is to be refined. Again, we denote the hyperplane from our finite training set $\mathcal{D}_{K-1}^{\text{major}}$ as $\mathbf{h}_{K-1}$. The error Δ in this case is now the angular error between the normals for $\mathbf{h}_{\text{ideal}}$ and $\mathbf{h}_{K-1}$. Figure A.1 indicates this setting. The learnt hyperplane $\tilde{\mathbf{w}}_{K-1}$ is shown as a black solid line. The black dashed line represents the mirror image of the learnt hyperplane, defined for aid in simplification. Recall that updates to the weight vector take place on misclassification. On average, the updates due to samples in regions R1 for ($y = 2$) and R2 (for $y = 1$) lead to a net horizontal (leftward) weight update. This is an unfavorable update that increases Δ. Therefore, the favorable updates on average are from regions R3 and R4 for $y = 2$, and R5 and R6 for $y = 1$. This is a net update with large vertical (upward) update. This is a favorable update that decreases Δ. These regions are described based on the small angular deviation Δ. Since the distributions have finite support along the direction parallel to the ideal hyperplane (vertical direction in Figure A.1), the requirement again reduces

to greater likelihood of sampling close to the ideal hyperplane (similar to Theorems 1 and 2), since Δ is small. That is, distributions that sample close to the ideal hyperplane with greater probability have a greater expected likelihood of a favorable update. Under similar conditions as Theorem 1, MIME effect holds in this case.

The extension to include the bias term b is straightforward. We follow the setting in Equation A.43 and subsume the bias as part of the weights. In this case, Δ includes the error in both the hyperplane normal direction as well as the bias. Extensions to greater number of dimensions can be done using the same arguments. Additionally, domain gap can also be introduced. We omit explicit mathematical expressions in the interest of brevity, and since our goal here is to establish existence.

A.5 Feature Space Analysis

Constructing the Projected Feature Histograms: Let f denote a feature vector, in the penultimate layer of a classification neural network. For example, in the case of ResNet-34 [201], $\mathbf{f} \in \mathbb{R}^{512}$. Similarly, let $\mathbf{w}$ be the final layer weights. In the case of multiple final layer hyperplanes, we choose any one of the hyperplanes (since for the two class classification task, the two projected variables are correlated when trained against the cross entropy loss for 2 classes). Then, we define $x \in \mathbb{R}$ as,

$$x = \mathbf{w}^T \mathbf{f}. \tag{A.45}$$

Classification decisions are made solely on the basis of the projected variable x. Therefore, we analyze the histogram distributions for x. Practically, for each dataset, we use the best performing (in terms of majority group performance) model trained using a minority training fraction (β) of 0.5. This is chosen in order to obtain histograms of x for all four distributions – the two task classes for both the majority and minority groups. The histograms are created using the test set samples.

Estimating the Overlap: The overlap is estimated from the histograms, using the following Python code snippet:

```
def histogram_intersection(h1, h2, bins):
    #INPUTS:
    #h1, h2: normalized histograms
    #bins: number of bins in the histograms (should be equal for the
                                            two histograms)
    #OUTPUTS:
    #sm: overlap fraction
    sm = 0
    for i in range(bins):
        sm += min(h1[i], h2[i])
    return sm
```

Estimating the Domain Gap: We follow a two step process to estimate the domain gap δ. First, the ideal decision hyperplanes for the majority and minority groups are estimated, using Equation A.21. We fit a fifth order polynomial to the two histograms. The central intersection point of the histograms (i.e the intersection point that lies between the means of the two classes) is then the location of the ideal decision threshold. The following Python code snippet describes this:

```
import numpy as np

def ideal_hyperplane(h1, h2, z, ref=5):
    #INPUTS:
    #h1, h2: the two histograms, of equal length and identical bins
    #z: a list of the histogram bin centers
    #ref: Search space for the intersection of the two histograms-
                                            default is from -5 to 5
    #OUTPUT:
    #z_dec: Ideal decision threshold between the two histogram
```

```
                                                    distributions
    z_dash = np.polyfit(z, h1, 5)
    f1 = np.poly1d(z_dash)
    # calculate polynomial
    z_dash = np.polyfit(z, f2, 5)
    f2 = np.poly1d(z_dash)
    new_z = np.linspace(-ref,ref,5000)
    new_f1 = f1(new_z)
    new_f2 = f2(new_z)
    id_dec = np.argmin(np.abs(new_f1-new_f2))
    z_dec = new_z[id_dec]
    return z_dec
```

The domain gap is the absolute difference between two ideal decision thresholds, for each of the two group classes. Figure 3 of the main paper may be referred to for a graphical visualization.

Notes on the Estimated Measures: The latent feature space analysis is not perfect. This is because the feature extraction part of the network is jointly learnt along with decision hyperplane. Histograms are plotted on the 50% minority training ratio so as to enable a fair domain gap and overlap comparison between the two group classes. Specifically, note that we define task complexity in the main paper in terms of the minority only and majority only train sets which deviates from the setting here. The estimates for overlap and domain gap are therefore approximate correlated estimates and not exact measures.

Analysis of Feature Space Gaussian-like Behavior: We set up the Chi-Squared goodness of fit test on all 20 distributions under consideration (i.e. across 5 datasets and 4 distributions each per dataset). These statistics correspond to the distributions in Table 1 and Figure 4 of the main paper. Python code for testing the hypotheses is given below. The number of bins are chosen so as to ensure ≥ 5 samples per bin on average.

```
from scipy.stats import chisquare
from scipy.stats import norm
from scipy import stats
import pandas as pd

def chi_square_stats(vals,no_bins)
    #INPUTS:
    #vals: a list of samples whose Gaussianity is to be tested
    #no_bins: number of bins (thumb rule: no_bins<len(vals)/5)

    tot_vals = len(vals)
    # mean and standard deviation of given data
    mean = np.mean(vals)
    std = np.std(vals)

    interval = []
    for i in range(1,no_bins+1):
      val = stats.norm.ppf(i/no_bins, mean, std)
      interval.append(val)
    interval.insert(0, -np.inf)

    lower = interval[:-1]
    upper = interval[1:]

    df = pd.DataFrame({'lower_limit':lower, 'upper_limit':upper})


    sorted_vals = list(sorted(vals))
    df['obs_freq'] = df.apply(lambda x:sum([i>x['lower_limit'] and i<=
                                        x['upper_limit'] for i in
                                        sorted_vals]), axis=1)
    df['exp_freq'] = tot_vals/no_bins
```

```
    statistic = stats.chisquare(df['obs_freq'], df['exp_freq'])

    p = 2    # number of parameters for 1D Gaussian
    DOF = len(df['obs_freq']) - p -1
    thresh = stats.chi2.ppf(0.95, DOF)

    return statistic, thresh
```

Table A.1 highlights the evaluated chi-square statistics, as well as related parameters. Note that a lower value of the statistic is better, and the null hypothesis is not rejected when the value of the statistic is lower than the critical value. We establish the null hypothesis at a 5% level of significance for each distribution to be that the samples are drawn from a Gaussian distribution. Distributions that are unable to reject the null hypothesis are indicated in bold. It can be seen that a large majority of the distributions indicate that the projected latent features follow a Gaussian-like distribution.

Table A.1: **Chi-Squared goodness of fit measures for all distributions.** Distributions with bolded values show the estimated statistics that are lower than the critical value, indicating that the null hypothesis (Gaussian distribution) cannot be rejected.

Dataset	No. of samples per group per class	No. of Bins	Critical Value	Majority Group		Minority Group	
				$y = 1$	$y = 2$	$y = 1$	$y = 2$
DS-1 [167]	379	15	21.03	**13.65**	28.69	**10.25**	**15.39**
DS-2 [168]	126	15	21.03	**7.81**	**12.10**	**11.62**	**9.24**
DS-4 [169]	126	15	21.03	**10.43**	**17.57**	**17.10**	**4.48**
DS-5 [3]	159	15	21.03	**11.09**	24.05	**5.74**	**14.40**
DS-6 [4, 5]	43	5	5.99	25.48	**5.02**	**5.72**	**4.79**

A.6 Implementation Details

Analysis measures: For each task, we estimate the test accuracy $a_p^i(\beta)$ as a function of minority group fraction in the train set $\beta \in [0, 1]$, for a trial $i \in \{1, ..., N\}$, for a group class g (e.g. dark skin tones). N is the total number of trials. Practically, we evaluate performance for a finite set of β values, represented by the set $B = \{0, 0.1, 0.2, \ldots, 1.0\}$. We now define the following measures.

Average accuracy: For a given minority training ratio β_0, and for a given group class g, we define the average accuracy,

$$\bar{a}_g(\beta_0) = \frac{1}{N} \sum_{i=1}^{N} a_g^i(\beta_0). \tag{A.46}$$

Error bounds: We also evaluate the *trend variation* among $a_g^i(\beta)$ for various i. That is, we want to evaluate if across all the trials (for a particular task-dataset combination), the relative trend (of majority group performance gain) holds true. One candidate measure for this is $std_i(a_g^i(\beta))$ for each β, where $std_i(\cdot)$ is the standard deviation operator, over i. However, this measure will include average changes in accuracy for all splits, for a particular trial (arising out of unrelated effects such as different train or test set samples). This is unnecessary in our case. Therefore, we define our error measure $\hat{\zeta}(\beta)$ as the β-mean subtracted standard deviation. That is,

$$\begin{aligned} \hat{\zeta}(\beta) &= std_i(a_g^i(\beta) - \bar{a}_g^i), \\ \bar{a}_g^i &= \frac{1}{|B|} \sum_{\beta \in B} a_g^i(\beta), \end{aligned} \tag{A.47}$$

where $|\cdot|$ is the cardinality operator representing the size of a set. In our graphs, we plot the average accuracy $\bar{a}_g(\beta)$ as well as the error bounds, from $\bar{a}_g(\beta) - \hat{\zeta}(\beta)$ to $\bar{a}_g(\beta) + \hat{\zeta}(\beta)$, $\forall \beta \in B$.

Network Architectures Used: For all the vision-related experiments, we use the ResNet-34 architecture [201]. We only modify the output layer of the network so as to match the number of task classes (9 for Dataset 3, and 2 for all other tasks). For the Adult (Census) Dataset [3], we use a fully connected network with sigmoid outputs. The PyTorch [156] implementation for the model is included below.

```
#Model


def act(x):
    return F.relu(x)

class Network(nn.Module):
    def __init__(self,):
        super().__init__()
        self.fc1 = nn.Linear(101, 50)
        self.fc2 = nn.Linear(50, 50)
        self.fc3 = nn.Linear(50, 50)
        self.fcLast = nn.Linear(50,2)

    def forward(self,x):

        x = act(self.fc1(x))
        # x = self.b1(x)
        x = act(self.fc2(x))
        x = act(self.fc3(x))
        x = torch.sigmoid(self.fcLast(x))
        return x
```

General Experiment Details: All experiments were carried out using PyTorch [156]. Table A.2 highlights the training parameters used for each dataset. We use different parameters for each of the datasets. These are experimentally chosen to maximize accuracy. All the models are trained using the AdamW optimizer [278] and the cross entropy loss.

Table A.2: **Training configuration and parameters for all datasets and experiments.** Parameters for each dataset are chosen so as to maximize performance.

Dataset (Task)	DS-1 [167] (Gender)	DS-2 [168] (Species)	DS-3 [5] (Age)	DS-4 [169] (Diagnosis)	DS-5 [3] (Income)	DS-6 [4, 5] (Gender)
Group class	Race	Skin tone	Gender	Gender	Gender	Species
Train set size	10900	1500	7700	1500	2600	750
Test set size (per group)	760	250	970	250	300	90
No. of trials	5	5	5	7	5	5
No. of epochs	35	60	65	40	250	20
Learning rate	0.0005	0.0006	0.0006	0.0006	0.0005	0.0005
Weight Decay	0.08	0.05	0.05	0.05	0.08	0.08
Input Shape/Config.	3x100x100	3x256x256	3x100x100	3x256x256	101x1	3x100x100

The train and test set sizes vary slightly across trials, due to different data splits. However, the train set size remains the same for all minority training ratios of a particular trial. A validation set is held out but given the small sample size of several datasets, we measure trends based on best test performance. This is to minimize the effect of sample specific performance gap in small datasets. Averaging of trends over multiple trials, and hence multiple train-test splits ensures that the trends do not overfit to a particular configuration. Each trial is run using a unique random seed. Table A.3 highlights the random seeds used for our experiments, which were randomly chosen. Input images are resized to the chosen input size for each dataset. For the Adult dataset [3], we use a one-hot encoding scheme for the input. The group class information is dropped from the input before passing to the network. For all the datasets, across all minority training ratios for a particular trial, we use a fixed model initialization to ensure that the changes in accuracy are completely attributable to the train data configuration.

Dataset Specific Information: To perform experiments on the **Pet Images Dataset**, we manually annotate light and dark fur cats and dogs from the larger dataset used in [168]. For the age classification task on the **UTKFace Dataset** [5], we pre-process the age labels

Table A.3: **Random seeds used for the trials.** Seeds were chosen at random for trials to generate average trends and error bounds.

Dataset (Task)	DS-1 [167] (Gender)	DS-2 [168] (Species)	DS-3 [5] (Age)	DS-4 [169] (Diagnosis)	DS-5 [3] (Income)	DS-6 [4, 5] (Gender)
Random Seeds	0, 1, 3, 5, 7	21, 42, 35, 28, 31	0, 55, 2, 15, 6	33, 42, 24, 36 54, 21, 28	13, 15, 17, 19, 21	0, 1, 3, 5, 9

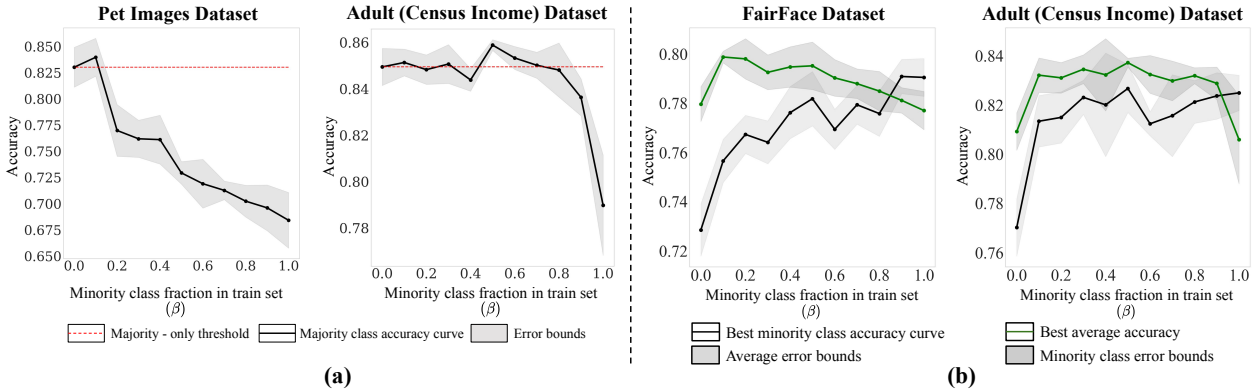


Figure A.2: **The MIME effect is complementary to data debiasing methods and consistent with research aimed at equal representation (ER) datasets.** (a) Training configurations using data debiasing methods [15] show the MIME effect. (b) While ER datasets are not optimal for the MIME effect (Figure 5 and 6, main paper), optimal overall performance is observed close to ER.

to match the annotation format for the FairFace dataset [167]. For the large domain gap gender classification task using the **UTKFace and Chicken Images Datasets** [5, 4], we perform gender classification over human and chicken groups. Therefore, this experiment is over a new, composite dataset.

A.7 Additional Secondary Analysis of MIME

MIME effect with debiasing methods: We now analyze the interaction of the MIME effect with existing debiasing methods. Specifically, while applying hard-sample mining [15] (as an exemplary case) across the task classes ($y = 1, 2$), we sweep across various minority training ratios. Figure A.2(a) shows results on two datasets (implementation details may be found in the following section). The MIME effect continues to be observed. Debiasing methods act on the task classes ($y = 1, 2$) in an effort to improve performance while MIME acts on majority and minority groups, regardless of the task class. Therefore, MIME is complementary to debiasing methods, rather than a competitor. In our experiments, hard-sample mining does not lead to significant performance gains since the task classes are balanced by experimental design. In other scenarios where this might not be the case, MIME and hard sample mining might together improve performance.

Reconciling MIME with existing equal representation (ER) datasets: In this paper, we focus only on majority group performance, for which ER training datasets are not optimal in general. In contrast, existing efforts [160, 58, 161, 162, 163, 164, 165, 166, 143] focus on ER datasets to maximize overall (majority+minority) performance. This need not be optimal but is a good thumb rule. This is because while majority group performance eventually reduces with minority training ratio, minority group performance increases (Figure A.2(b) highlights this).

A.8 Hard Mining Baseline Implementation

We implement a version of the method proposed in [15]. From a batch of 30 samples, 12 samples (6 of each task class) are retained and used in the training step. These are the samples with least confidence, with respect to ground truth targets. Code is shown below. Trial random seeds are the same as shown in Table A.3.

```
class compute_crossentropyloss_hardMine:
    """
    y0 is the vector with shape (batch_size,C)
    x shape is the same (batch_size), whose entries are integers from 0 to
                                            C-1
    In our case, C=2.
    """
    def __init__(self, ignore_index=-100) -> None:
        self.ignore_index=ignore_index

    def __call__(self, y0, x):
        loss = 0.
        eps = 1e-5
        K = 6
        n_batch, n_class = y0.shape
        pos_score = torch.ones(n_batch).to(device)
        neg_score = torch.ones(n_batch).to(device)
        ix_pos = 0
        ix_neg = 0
        for y1, x1 in zip(y0, x):
            class_index = int(x1.item())
            score = torch.exp(y1[class_index])/(torch.exp(y1).sum()+eps)
            if class_index == 0:
                neg_score[ix_neg] = score
                ix_neg+=1
            else:
                pos_score[ix_neg] = score
                ix_pos+=1

        pos_score,_ = torch.sort(pos_score,dim=0)
        neg_score,_ = torch.sort(neg_score,dim=0)
        pos_els = np.minimum(K,ix_pos)
        neg_els = np.minimum(K,ix_neg)
```

```
        for ix in np.arange(pos_els):
            loss = loss -torch.log(pos_score[ix])

        for ix in np.arange(neg_els):
            loss = loss -torch.log(neg_score[ix])

        loss = loss/(pos_els+neg_els)
        torch.cuda.empty_cache()
        return loss
```

A.9 Our Code

Our code may be accessed through the project webpage at `https://visual.ee.ucla.edu/mime.htm/`. We provide code and guidance to perform experiments on all six datasets. Due to specific requirements for each dataset, we provide six Jupyter notebooks. We also include details on setting up file structures and link to datasets wherever necessary. Please refer to the README file for further details.

A.10 Negative Impacts and Mitigation

This paper focuses on highlighting the existence of the MIME effect, and not optimal configurations for performance gain. Nevertheless, potential negative outcomes may occur if the results are misinterpreted as guidance on dataset construction with respect to certain stakeholder groups. The rigor of our theoretical results emphasizes this nuance to computer scientists, and future work in diverse venues can extend the notion of minority inclusion for majority group performance gains to broader audiences.

APPENDIX B

Supplemental Content: Using Neural Implicit Video Representations to Enable Low-SNR rPPG

This supplement is organized as follows:

1. Section 1 contains mathematical derivations for our implicit decomposition formulation.
2. Section 2 contains the architectural details and training configurations.
3. Section 3 verifies the effectiveness of the $\mathcal{A} - \mathcal{B}$ decomposition.
4. Section 4 contains runtime details.
5. Section 5 elaborates on the choice of our parity metric for qualifying the performance gaps due to out-of-distribution (OOD) evaluation.
6. Section 6 discusses the ablation experiments for the implicit representation architecture.
7. Section 7 discusses the out of distribution dataset and detailed scene-wise performance analysis.
8. Section 8 shows additional qualitative results.
9. Section 9 shows additional baselines.
10. Section 10 discusses potential future directions.

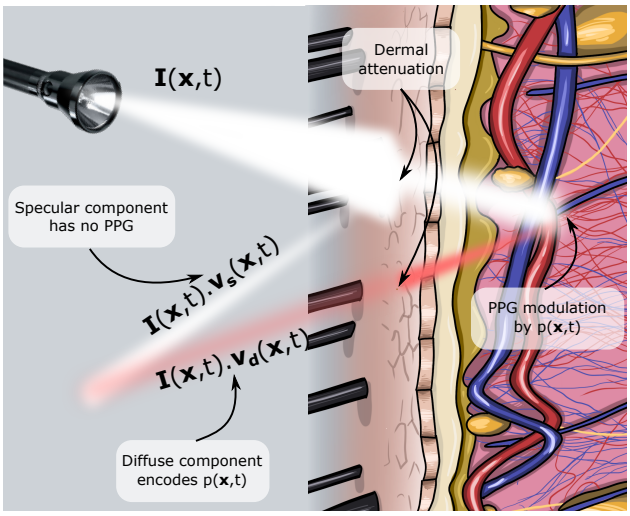


Figure B.1: **Light transport analysis provides insights on plethysmograph signal quality.** The reflected light arriving at the camera sensor consists of two components: (i) the specular component, shown in white, does not contain PPG information since it arises out of surface reflections, and (ii) diffuse reflection, shown in red, arising out of subsurface scattering, contains PPG information.

B.1 Mathematical Formulation

B.1.1 Light transport of Plethysmography

We begin with a mathematical description of light transport for rPPG estimation. This will set up our theoretical exposition in subsequent sections. For a more detailed treatment of modeling light transport for plethysmography and under the skin imaging, we direct readers to [27] and [279]. For a description of the theoretical foundations of optical bias, we direct readers to [8]. Fundamentally the rPPG signal is subtle, making its estimation difficult. While contemporary methods [30, 31, 9] circumvent this through the use of deep networks, they fare poorly on OOD samples. Hence, this necessitates a generalizable method capable of extracting the rPPG without relying on domain-specific features that tend to overfit. This motivates our functional decomposition.

Figure B.1 highlights the light transport of remote plethysmography. Given spatial coordinates $\mathbf{x} \in [-0.5, 0.5]^2$, and temporal coordinate $t \in [-0.5, 0.5]$, the RGB color field $\mathcal{C}(\mathbf{x}, t)$ is given by,

$$\mathcal{C}(\mathbf{x}, t) = I(\mathbf{x}, t) \cdot (\mathbf{v_s}(\mathbf{x}, t) + \mathbf{v_d}(\mathbf{x}, t)) + \mathbf{v_n}(\mathbf{x}, t), \tag{B.1}$$

where $I(\mathbf{x}, t)$ encompasses the illuminatory signal modulated by 2 reflective components, $\mathbf{v_s}(\cdot, \cdot)$ and $\mathbf{v_d}(\cdot, \cdot)$, representing the specular and diffuse components, and $\mathbf{v_n}(\cdot, \cdot)$ is the measurement noise.

These specular reflections are not constant throughout space and time and are governed by the geometry of the human body. Since this component is a scaled reflection of the light source from the skin surface and does not contain major plethysmograph information, it acts as an interference signal.

The diffuse component arising from subsurface scattering contains plethysmograph information. For our purposes, it can be resolved into a steady-state skin color vector $d_0(\mathbf{x}, t) \cdot \mathbf{u_d}$ (where $\mathbf{u_d}$ is the unit vector along the skin color direction), and a time-varying pulsatile component $p(\mathbf{x}, t) \cdot \mathbf{u_p}$ (where $\mathbf{u_p}$ is the unit colorspace 'direction' for the plethysmograph color changes and $p(t)$ is a unit-power signal). That is,

$$\mathbf{v_d}(\mathbf{x}, t) = d_0(\mathbf{x}, t) \cdot \mathbf{u_d} + p(\mathbf{x}, t) \cdot \mathbf{u_p}. \tag{B.2}$$

B.1.2 $\mathcal{A}$-$\mathcal{B}$ Decomposition and Optimality

We propose decomposing the color signal into a signal amplitude component, $p_m(\mathbf{x}, t)$, and a temporal plethysmograph signal $p_i(t)$, which forms our desired signal. However, the relation between $p(\cdot)$, $p_m(\cdot, \cdot)$, $p_i(\cdot)$ is non-linear in nature. Spatial variations manifest in the temporal component $p_i(\cdot)$ as a result of the pulse wave taking finite time to propagate, while motions may induce pixel shifts in the spatial component $p_m(\cdot, \cdot)$. These individual effects can be expressed through a non-linear function,

$$p(\mathbf{x}, t) = \mathcal{Z}(p_m, p_i, \mathbf{x}, t). \tag{B.3}$$

However, this function is intractable, making closed form analysis difficult. We therefore simplify the model to exclude motion effects (that is, $p_m(\cdot)$ is only a function of $\mathbf{x}$). That being said, empirically, we find our method to be capable of handling natural motion (as discussed later). We can now rewrite the pulsatile signal as

$$p(\mathbf{x}, t) \approx p_m(\mathbf{x}) \cdot p_i(t). \tag{B.4}$$

Furthermore, the scope of our work is constrained to assume constant phase across the entire face, similar to prior work [27, 25, 29, 26], thereby ignoring the effect of pulse transit time. This is a minor assumption, since transit time across the scale of a face is small.

The RGB color field for the video can be written as follows:

$$\mathcal{C}\left(\mathbf{x}, t\right) = I \cdot (\mathbf{x}, t)(\mathbf{v}_s(\mathbf{x}, t) + d_0\left(\mathbf{x}, t\right) \cdot \mathbf{u_d} + p_m(\mathbf{x}) \cdot p_i(t) \cdot \mathbf{u_p}) + \mathbf{v_n}\left(\mathbf{x}, t\right). \tag{B.5}$$

This can be succinctly decomposed into two parts, such that,

$$\mathcal{C}\left(\mathbf{x}, t\right) = \mathcal{A}(\mathbf{x}, t) + \mathcal{B}(\mathbf{x}, t) + \mathbf{v_n}\left(\mathbf{x}, t\right), \tag{B.6}$$

where,

$$\begin{aligned} \mathcal{A}(\mathbf{x}, t) &= I(\mathbf{x}, t) \cdot \mathbf{v_s}\left(\mathbf{x}, t\right) + d_0\left(\mathbf{x}, t\right) \cdot \mathbf{u_d}, \\ \mathcal{B}(\mathbf{x}, t) &= I(\mathbf{x}, t) \cdot p_m(\mathbf{x}) \cdot p_i(t) \cdot \mathbf{u_p}. \end{aligned} \tag{B.7}$$

$\mathcal{A}(\mathbf{x}, t)$, the appearance component (or $\mathcal{A}$-function), encompasses the base appearance of the face, as well as texture and other details. $\mathcal{B}(\mathbf{x}, t)$, the blood component (or $\mathcal{B}$-function), contains the spatio-temporally varying color changes arising out of blood flow (plethysmography). We will refer to this decomposition as the $\mathcal{A} - \mathcal{B}$ *decomposition*.

From Equation B.5, a Signal to Interference & Noise Ratio (SINR) for $\mathcal{C}\left(\mathbf{x}, t\right)$ can be given by,

$$SINR_{\mathcal{C}}\left(\mathbf{x}, t\right) = \frac{p_m(\mathbf{x})^2}{|\mathbf{v_s}\left(\mathbf{x}, t\right)|^2 + d_0(\mathbf{x}, t)^2 + |\mathbf{v_n}\left(\mathbf{x}, t\right)|^2 / I(\mathbf{x}, t)^2}. \tag{B.8}$$

This leads us to the following Theorem.

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,*

$$SINR_{\mathcal{B}}(\mathbf{x},t) \geq SINR_{\mathcal{C}}(\mathbf{x},t). \tag{B.9}$$

That is, the $\mathcal{A}-\mathcal{B}$ decomposition leads to an effective SINR gain.

Proof: *Assuming that ideal $\mathcal{A}-\mathcal{B}$ decomposition adds no additional noise to the estimate,*

$$\begin{aligned} SINR_{\mathcal{B}}\left(\mathbf{x},t\right) &\geq \frac{p_m(\mathbf{x})^2}{|\mathbf{v_n}\left(\mathbf{x},t\right)|^2/I(\mathbf{x},t)^2}, \\ &\geq \frac{p_m(\mathbf{x})^2}{|\mathbf{v_s}\left(\mathbf{x},t\right)|^2+d_0(\mathbf{x},t)^2+|\mathbf{v_n}\left(\mathbf{x},t\right)|^2/I(\mathbf{x},t)^2}, \\ &= SINR_{\mathcal{C}}\left(\mathbf{x},t\right). \end{aligned} \tag{B.10}$$

Therefore,

$$SINR_{\mathcal{B}}(\mathbf{x},t) \geq SINR_{\mathcal{C}}(\mathbf{x},t). \tag{B.11}$$

This completes the proof. ■

Note that Theorem 1 is an existence proof. That is, it describes the existence of SINR benefits as a result of ideal decomposition, thereby motivating its necessity in our real applications. Additionally, Theorem 1 has practical significance: to the best of our knowledge, prior state-of-the art methods for rPPG, such as [30, 31, 137, 9, 280], operate on undecomposed, raw RGB video frames. Through an $\mathcal{A}-\mathcal{B}$ decomposition, the ceiling for potential SINR gains is raised, making it an attractive avenue for exploration.

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

$$\widehat{p^*}(t) = \int_{\mathbf{x}\in\mathbf{\Omega}} SINR_{\mathcal{B}}(\mathbf{x},t)\mathcal{B}(\mathbf{x},t)d\mathbf{x}, \tag{B.12}$$

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

This is the notion of near-optimal plethysmography that we will aim to achieve in this work. Note that the two necessary components to achieve this, namely $SINR_{\mathcal{B}}(\mathbf{x}, t)$ and $\mathcal{B}(\mathbf{x}, t)$, are extremely difficult to evaluate analytically, due to dependence on multiple factors, and the extremely small signal amplitudes making it susceptible to noise.

B.1.3 Functional Decomposition

Consider two functions, $\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{A}}})$ and $\hat{\mathcal{B}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{B}}})$, parameterized by $\mathbf{\Theta}_{\mathcal{A}}$ and $\mathbf{\Theta}_{\mathcal{B}}$, that aim to represent $\mathcal{A}(\mathbf{x}, t)$ and $\mathcal{B}(\mathbf{x}, t)$ respectively. $\mathcal{R}[\cdot]$ is a range-space operator that returns the set of all possible functions that may be represented by a parameterized function. To perform an $\mathcal{A}$-$\mathcal{B}$ decomposition, we wish to find parameterized functions that satisfy the following requirements under ideal conditions:

$$\begin{aligned} \mathcal{A}(\mathbf{x}, t) &\in \mathcal{R}\left[\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{A}}})\right], \\ \mathcal{B}(\mathbf{x}, t) &\in \mathcal{R}\left[\hat{\mathcal{B}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{B}}})\right], \\ \mathcal{B}(\mathbf{x}, t) &\notin \mathcal{R}\left[\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{A}}})\right]. \end{aligned} \tag{B.13}$$

That is, $\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{A}}})$ is able to represent the $\mathcal{A}$-function but not the $\mathcal{B}$-function. If we are able to find functions that satisfy the constraints in Equation B.13, we can achieve $\mathcal{A}$-$\mathcal{B}$ decomposition by sequentially learning $\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{A}}})$ and $\hat{\mathcal{B}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{B}}})$.

For a real-world application, both the $\mathcal{A}$ and the $\mathcal{B}$-functions cannot be learnt exactly as a result of noise. In such a case, these parametric functions learn representations that minimize some component-dependent metric distances, represented as $d_{\mathcal{A}}(\cdot, \mathcal{A}(\mathbf{x}, t))$ and $d_{\mathcal{B}}(\cdot, \mathcal{B}(\mathbf{x}, t))$. That is,

$$d_{\mathcal{A}}(\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}^*_{\boldsymbol{\mathcal{A}}}), \mathcal{A}(\mathbf{x}, t)) \leq d_{\mathcal{A}}(\hat{\mathcal{A}}(\mathbf{x}, t; \mathbf{\Theta}_{\boldsymbol{\mathcal{A}}}), \mathcal{A}(\mathbf{x}, t)),\ \forall \mathbf{\Theta}_{\boldsymbol{\mathcal{A}}}, \tag{B.14}$$

$$d_{\mathcal{B}}(\hat{\mathcal{B}}(\mathbf{x},t;\mathbf{\Theta}^*_{\boldsymbol{\mathcal{B}}}),\mathcal{B}(\mathbf{x},t)) \leq d_{\mathcal{B}}(\hat{\mathcal{B}}(\mathbf{x},t;\mathbf{\Theta}_{\boldsymbol{\mathcal{B}}}),\mathcal{B}(\mathbf{x},t)),\ \forall \mathbf{\Theta}_{\boldsymbol{\mathcal{B}}}, \tag{B.15}$$

where $\mathbf{\Theta}^*_{\boldsymbol{\mathcal{A}}}$ and $\mathbf{\Theta}^*_{\boldsymbol{\mathcal{B}}}$ are optimal parameters for the functional representations. These distances can be used to rewrite the constraints in Equation B.13, as follows:

$$\begin{aligned} d_{\mathcal{A}}(\hat{\mathcal{A}}(\mathbf{x},t;\mathbf{\Theta}^*_{\boldsymbol{\mathcal{A}}}),\mathcal{A}(\mathbf{x},t)) &\leq \epsilon_{\mathcal{A}}, \\ d_{\mathcal{B}}(\hat{\mathcal{B}}(\mathbf{x},t;\mathbf{\Theta}^*_{\boldsymbol{\mathcal{B}}}),\mathcal{B}(\mathbf{x},t)) &\leq \epsilon_{\mathcal{B}}, \\ d_{\mathcal{B}}(\hat{\mathcal{A}}(\mathbf{x},t;\mathbf{\Theta}^*_{\boldsymbol{\mathcal{A}}}),\mathcal{B}(\mathbf{x},t)) &\geq \epsilon_{\mathcal{B}}. \end{aligned} \tag{B.16}$$

In our case, we can choose distance measures based on the requirements for the $\mathcal{A}$ and $\mathcal{B}$-functions. One choice is,

$$d_{\mathcal{A}}(\mathcal{F},\mathcal{A}(\mathbf{x},t)) = ||\mathcal{F}-\mathcal{A}(\mathbf{x},t)||_2,\ \forall\, \mathcal{F}, \tag{B.17}$$

such that the appearance metric distance rewards pixel wise closeness. Similarly,

$$d_{\mathcal{B}}(\mathcal{F},\mathcal{B}(\mathbf{x},t)) = |H(\mathcal{F}) - H(\mathcal{B}(\mathbf{x},t))|,\ \forall\, f\mathcal{F}, \tag{B.18}$$

where $H(\cdot)$ is a heart rate estimator function, enforcing similarity in estimated heart rates. This choice of functions allows us to evaluate and identify potential function representations for the purpose of $\mathcal{A}$-$\mathcal{B}$ decomposition.

Empirical Claim 1 *(based on observations): Phase-based motion models (models with phase encoded input embeddings) cannot represent plethysmograph signals accurately, in terms of the error metric defined in Equation B.18.*

Justification: *Figure 3, main paper highlights the performance of a phase-based INR on the task of fitting to a face video. While it is able reconstruct the video frame with high fidelity, the heart rate estimation error (using a pretrained PhysNet [31]) is high (that is,* $d_{\mathcal{A}}(\cdot,\mathcal{A}(\mathbf{x},t))$ *is low while* $d_{\mathcal{B}}(\cdot,\mathcal{B}(\mathbf{x},t))$ *is high). The model inductive biases prevent it from representing plethysmograph signals efficiently.*

This could be viewed as a limitation of phase-based INRs. However, we use it to our advantage: phase-based models can be an effective estimator for the $\mathcal{A}$-function, in accordance with Equations B.13 and B.16.

We next look at sinusoidal representation networks (SRNs), such as [6, 205, 206].

Empirical Claim 2: *Sinusoidal Representation Networks are effective $\mathcal{A}$ and $\mathcal{B}$-function approximators.*

Justification: *Figure 3, main paper again highlights the performance of an SRN for our task. In this case, while the appearance and semantic information remain accurately retained, the heart rate error is also very low. That is, both $d_{\mathcal{A}}(\cdot, \mathcal{A}(\mathbf{x}, t))$ and $d_{\mathcal{B}}(\cdot, \mathcal{B}(\mathbf{x}, t))$ are low.*

SRNs are therefore, effective $\mathcal{A}$ and $\mathcal{B}$-function estimators. However, being able to represent both $\mathcal{A}$ and $\mathcal{B}$ makes SRNs incapable of $\mathcal{A}$-$\mathcal{B}$ decomposition. Fortunately, this pair exactly satisfy the constraints in Equation B.16. Therefore,

Empirical Claim 3: *A phase-based model can serve as $\hat{\mathcal{A}}(\mathbf{x}, t; \boldsymbol{\Theta}^*_{\boldsymbol{\mathcal{A}}})$, while an SRN can serve as $\hat{\mathcal{B}}(\mathbf{x}, t; \boldsymbol{\Theta}^*_{\boldsymbol{\mathcal{B}}})$, when trained sequentially.*

Justification: *When sequentially trained, the phase-based model will represent $\mathcal{A}(\mathbf{x}, t)$ while ignoring $\mathcal{B}(\mathbf{x}, t)$. When trained on the residual of $\mathcal{C}(\mathbf{x}, t)$ and $\mathcal{A}(\mathbf{x}, t)$, the SRN will represent only $\mathcal{B}(\mathbf{x}, t)$.*

B.2 Architecture Details and Training Configuration

B.2.1 Cascaded Appearance

Our codebase builds on top of the resources from [281]. The overall cascaded appearance model includes 2 encoding layers and 2 network layers. We use one of each to map spatio-temporal coordinates to the spatial offsets and the other to map the spatial coordinates with the spatial offsets to the color field. Both the MRHE encoding layers are kept identical with the hyperparameters specified in Table B.1.

Table B.1: Hyperparameters for the multi-resolution hash-grid encodings.

Hyperparameter	**Symbol**	**Value**
Number of levels	L	8
Hash table size	T	24
Number of features per entry	F	2
Base resolution	N_{min}	16
Growth factor between 2 levels	b	1.5

For the network, we used a simple multilayer perceptron with 2 hidden layers, where each hidden layer consists of 64 neurons accompanied by a ReLU activation function [282]. The final layer's activation function for the first stage (coordinate to offset module) is chosen to be a $tanh(\cdot)$ function scaled by a factor of 0.5, while the final layer of the offset to color field module is paired with no activation function. For a given video, the cascaded appearance model $\hat{\mathcal{A}}$ is trained using a simple Mean Squared Error loss function on the original video.

B.2.2 Residual Plethysmograph

Echoing the architectural nuances of the Cascaded Appearance's encoding layer, we deploy an identical MRHE encoding layer for the Residual Plethysmograph model. These encodings are fed through a multilayer perceptron with 2 hidden layers, each consisting of 64 neurons

with sinusoidal activations. The output layer for the plethysmograph model was implemented with no activation functions. Again, $\hat{\mathcal{B}}$ is also trained using a Mean Squared Error loss only.

B.2.3 Training Configurations for the Implicit Representation

We run our experiments with a batch size of 2^{13}. A lower batch size increases the number of descent steps per epoch, resulting in faster convergence, albeit at the expense of a noisier descent. Our chosen batch size is sufficient to ensure faster convergence without hampering the smoothness of the descent.

The cascaded appearance model is trained with the Adam optimizer[81] with a learning rate of 10^{-2}. Adhering to the optimizer settings from [216], we choose $\epsilon = 10^{-15}, \beta_1 = 0.9, \beta_2 = 0.99$. Similarly, we use the Adam optimizer with a learning rate of 10^{-2} to train the residual plethysmograph model. Additionally, we include a l_2 regularization factor of 10^{-6} to the network in the residual plethysmograph model (and not the MRHE). The optimizer configurations are otherwise kept identical. We use the mean squared error as our metric for training the cascaded appearance and the residual plethysmograph models.

We train the Cascaded Appearance for a total of 10 epochs with the above-stated configuration. Once trained, we freeze this model and instantiate the Residual Plethysmograph model with the previously stated configurations. The Residual Plethysmograph is now trained for 5 epochs such that the sum of the outputs of the Cascaded Appearance model and the Residual Plethysmograph model together reconstruct the video. We use the mean squared error to supervise all training procedures.

B.2.4 Refinement Network

Our refinement network is deployed on top of the concatenated (channel-wise) raw RGB and queried residual frames. This network consists of two blocks, namely the 2-D SINR estimator and the 1-D plethysmograph regressor. The spatial attention module is not trained directly,

but rather indirectly through the plethysmograph regressor to maximize the signal strength. Hence the generated masks are known as neural signal strength masks. The hyperparameters for the refinement network have elaborated the architecture in Table B.2.

Table B.2: Hyperparameters for the refinement network.

	Conv Layer					
Module	**Kernel**	**Padding**	**Out Channels**	**Batch Norm**	**Activation**	**Max Pooling**
Spatial Signal Attention	[1,5,5]	[0,2,2]	16	✓	ReLU	-
	[3,3,3]	[1,1,1]	32	✓	ReLU	-
	[3,3,3]	[1,1,1]	64	✓	ReLU	Temporal (factor = 2)
	[3,3,3]	[1,1,1]	64	✓	ReLU	Temporal (global adaptive)
	[1,1,1]	[0,0,0]	1	-	Sigmoid	-
We multiply the output of the Spatial Signal Attention module with the input 6-channel video tensor and average it spatially						
The resultant time-series tensor is normalized with respect to mean and standard deviation						
1-D Pleth Regressor	9	4	64	✓	ReLU	-
	9	4	128	✓	ReLU	-
	9	4	128	✓	ReLU	-
	9	4	128	✓	ReLU	-
	9	4	128	✓	ReLU	-
	9	4	128	✓	ReLU	-
	9	4	64	✓	ReLU	-
	9	4	16	✓	ReLU	-
	1	0	1	✓	-	-

In terms of training parameters, the model is trained for 2 epochs, using an Adam optimizer with $lr = 10^{-4}$ and weight-decay of 10^{-6}. The refinement network is supervised with a squared negative Pearson loss similar to [31, 8]. Additionally, we use an SNR Loss between the estimated plethysmograph waveform $\hat{\mathbf{y}}$ and the ground truth waveform $\mathbf{y}$. These losses are carefully chosen to motivate better signal reconstruction by maximizing the signal-to-label correlation and SNR. To improve the consistency and fidelity of the learned neural signal strength masks, we also apply a total variation regularization on these masks. The three loss functions are weighted such that the Pearson correlation loss is weighted by a factor of 3, the SNR Loss by a factor of 1 and the TV loss by a factor of 5. Note that the

SNR loss term is in the decibels scale. These loss functions are mathematically delineated as follows:

$$L_P(\mathbf{y}, \hat{\mathbf{y}}) = \left[1 - \frac{1}{\sqrt{a_1 \times a_2}} \left(N \sum_{i=1}^{N} \mathbf{y}_i \hat{\mathbf{y}}_i - \sum_{i=1}^{N} \mathbf{y}_i \sum_{i=1}^{N} \hat{\mathbf{y}}_i \right) \right]^2, \tag{B.19}$$

$$a_1 = \left(N \sum_{i=1}^{N} \mathbf{y}_i^2 - \left(\sum_{i=1}^{N} \mathbf{y}_i \right)^2 \right) \tag{B.20}$$

$$a_2 = \left(N \sum_{i=1}^{N} (\hat{\mathbf{y}}_i)^2 - \left(\sum_{i=1}^{N} \hat{\mathbf{y}}_i \right)^2 \right), \tag{B.21}$$

$$L_{SNR}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\int_{f_0-w}^{f_0+w} |\hat{\mathbf{Y}}(f)|^2 \, df}{\int_{-\infty}^{f_0-w} |\hat{\mathbf{Y}}(f)|^2 \, df + \int_{f_0+w}^{\infty} |\hat{\mathbf{Y}}(f)|^2 \, df}, \tag{B.22}$$

$$f_0 = \arg\max_f \mathbf{Y}(f), \tag{B.23}$$

$$L_{TV}\left(\widehat{\Gamma}(\mathbf{x})\right) = \left|\left|\nabla \widehat{\Gamma}(\mathbf{x})\right|\right|_2^2, \tag{B.24}$$

$$L_{total}(\mathbf{y}, \hat{\mathbf{y}}) = L_P(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_{SNR} L_{SNR}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_{TV} L_{TV}(\widehat{\Gamma}(\mathbf{x})). \tag{B.25}$$

where N is the length of $\mathbf{y}$ and $\hat{\mathbf{y}}$. Furthermore, $\mathbf{Y}(f)$ and $\hat{\mathbf{Y}}(f)$ are the respective Fourier transforms of $\mathbf{y}$ and $\hat{\mathbf{y}}$ and w is the chosen window size.

B.3 Physical Significance of the $\mathcal{A}$ & $\mathcal{B}$-functions

B.3.1 $\mathcal{A}$-function

In the context of rPPG, $\mathcal{A}$-component the interference that contains little to no information of the plethysmograph. However, in a more general case, this component retains all appearance-related information. We corroborate our claim with the help of the POS [27] and CHROM [26] algorithm.

Given that the residual signal is theorized to have (almost) no appearance-based infor-

mation, it can prove to be detrimental to these algorithmic methods. This is in-line with our findings in Table B.3. Contrarily, the $\mathcal{A}$-component is postulated to have the appearance-related information. Specifically, for rPPG, the appearance information is encoded in the instantaneous relative values between the color channels. Hence, we can "add" the appearance information back by adding the spatial average of $\mathcal{A}$-component to the rPPG signal (i.e., the spatial average of the $\mathcal{B}$-component). Further, given that appearance-based information for rPPG only required relative values, we can scale the spatial average of the appearance information to suppress the interference that originally decomposed from the rPPG signal, i.e., the $\mathcal{B}$-component. In our experiments, we scale the spatial average of the appearance branch by a factor of 0.1 while keeping the residual estimate at its normal strength. Table B.3 shows that the now altered POS and CHROM variants are able to capture plethysmographic information, evidenced by the remarkable increase in performance. Hence, through this analysis, we verify that the Cascaded Appearance, i.e., the $\mathcal{A}$-component, does indeed contain appearance-based information.

While the decomposed residual signal is pivotal for rPPG estimation, it would not have much of an impact on other downstream computer vision tasks due to its low signal strength. Hence, while removing this information from decomposition can potentially enhance the appearance information, we believe that the change in signal quality in the purview of these popular downstream algorithms would not be significant.

B.3.2 $\mathcal{B}$-function

We now analyze the benefit of the $\mathcal{B}$-component. Figure B.2 shows two participants: with a party mask and with a beard. We plot the average pixel value in the regions corresponding to red and blue boxes. The blue boxes (skin region, with high sig. strength) show a strong periodic rPPG signal, while the red boxes (occluded/beard region, low sig. strength) show noise. Therefore, the $\mathcal{B}$-signal has significant spatial signal strength information and improves signal strength map estimation.

Table B.3: **Experiments to show the effectiveness of the implicit decomposition to separate the appearance from the plethysmograph.** We perform an algorithmic correction to restore a small amount of the appearance features back to the residual output to test its effectiveness on prior art. We use a scaling factor of 0.1 for the appearance network's output in analysis. We use our method **with the neural signal strength masks trained on [8]** for this analysis.

Method	Algorithm	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Ours with only the residual output	**POS**	9.14	11.36%	12.30	0.62
	CHROM	11.67	15.65%	15.50	0.48
Ours with appearance correction	**POS**	1.4	1.70%	4.36	0.95
	CHROM	1.80	2.25%	5.23	0.93

B.3.3 Optimal Plethysmography and Uncertainty Proofs

We now provide mathematical detail for the argument put forth in Section 3.1 of the main paper. Assuming that the signal of interest is the plethysmography signal, $y(t)$ (note that we denote the signal as y for the purpose of this derivation and not p to avoid confusion) and that we have a sequence of estimates of the plethysmography signal, $\{\mathcal{B}(x_i,t)\}_{i=1}^{N}$, we can formulate the posterior distribution as:

$$p(y(t)|\mathcal{B}(x_1,t),\mathcal{B}(x_2,t),...,\mathcal{B}(x_N,t) \propto p(y(t))\prod_{i=1}^{N} p(\mathcal{B}(x_i,t)|y(t)). \tag{B.26}$$

Here, we assume that the prior and each likelihood term are Gaussian:

$$y(t) \sim \mathcal{N}(0,\sigma_o^2) \qquad \mathcal{B}(x_i,t)|y(t) \sim \mathcal{N}(y(t),\sigma_i^2). \tag{B.27}$$

The variance of likelihood is then inversely proportional to the signal and interference noise ratio, $\sigma_i^2 = \frac{1}{SINR_{\mathcal{B}}(x_i,t)}$. Due to the Gaussian assumptions, we can rewrite the posterior as:

$$
\begin{aligned}
p(y(t)|\mathcal{B}(x_1,t),...) &\propto exp\left(-\frac{1}{2\sigma_o^2}y(t)^2\right)\prod_{i=1}^{N} exp\left(-\frac{1}{2\sigma_i^2}(\mathcal{B}(x_i,t)-y(t))^2\right) \\
&\propto exp\left(-\frac{1}{2\sigma_o^2}y(t)^2-\sum_{i=1}^{N}\frac{1}{2\sigma_i^2}(\mathcal{B}(x_i,t)^2-2\mathcal{B}(x_i,t)y(t)+y(t)^2))\right) \\
&\propto exp\left(\frac{-1}{2}\left[\left(\frac{1}{\sigma_o^2}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)y(t)^2-2\left(\sum_{i=1}^{N}\frac{\mathcal{B}(x_i,t))}{\sigma_i^2}\right)y(t)+constant\right]\right) \\
&\propto exp\left(\frac{-(y(t)-AB)^2}{2B}\right) \\
A=\sum_{i=1}^{N}\frac{\mathcal{B}(x_i,t))}{\sigma_i^2} &\qquad B=\frac{1}{\frac{1}{\sigma_o^2}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}.
\end{aligned} \tag{B.28}
$$

Therefore, the posterior is also Gaussian:

$$
y(t)|\mathcal{B}(x_1,t),\mathcal{B}(x_2,t),...,\mathcal{B}(x_N,t)\sim\mathcal{N}(AB,B) \tag{B.29}
$$

The posterior mean can be written as:

$$
\mu_{y(t)}=\frac{\sum_{i=1}^{N}\frac{\mathcal{B}(x_i,t)}{\sigma_i^2}}{\frac{1}{\sigma_o^2}+\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}=\frac{\sum_{i=1}^{N}SINR_{\mathcal{B}}(x_i,t)\mathcal{B}(x_i,t)}{\frac{1}{\sigma_o^2}+\sum_{i=1}^{N}SINR_{\mathcal{B}}(x_i,t)}. \tag{B.30}
$$

Under the assumption that the prior's variance is large compared to the sum of the SINRs that are extracted, we can then directly relate the mean to Eq. (B.12):

$$
\mu_{y(t)}\approx\frac{\sum_{i=1}^{N}SINR_{\mathcal{B}}(x_i,t)\mathcal{B}(x_i,t)}{\sum_{i=1}^{N}SINR_{\mathcal{B}}(x_i,t)}\propto\sum_{i=1}^{N}SINR_{\mathcal{B}}(x_i,t)\mathcal{B}(x_i,t). \tag{B.31}
$$

Additionally, the posterior variance is:

$$
\sigma_{y(t)}^2\approx\frac{1}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}=\frac{1}{\sum_{i=1}^{N}SINR_{\mathcal{B}}(x_i,t)}. \tag{B.32}
$$

B.4 Robustness to Random Initializations

We conducted an experiment with 4 videos from the OOD dataset. The $\mathcal{A}-\mathcal{B}$ decomposition is run on these videos 10 times with random initializations, giving us 10 test datasets of 4

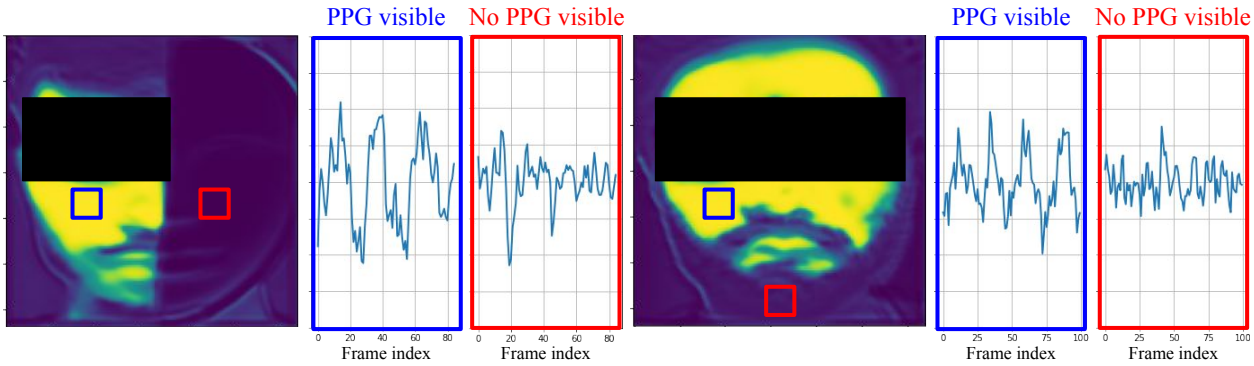


Figure B.2: **The $\mathcal{B}$-function contains information relevant to the estimation of the signal strength map.** Skin regions show PPG signal (albeit noisy) while occluded regions show only noise.

videos each. On average, across videos, the average heart rate (HR) error standard deviation across the 10 videos is 2.92 beats per minute. We also look at a second metric: the standard deviation of overall heart rate error across these 10 datasets, which is 1.29 beats per minute. One can expect these numbers to be inflated due to the small dataset used. Even so, we find that our decomposition is sufficiently robust.

B.5 Run-time and GPU Compute

From [8], we find that the typical time taken for deep learning architectures on their provided dataset is ≈ 800 *millisecs*, not including the time spent on preprocessing (cropping) the video frames. [280] have been able to decrease further the overall time taken, achieving a runtime of ≈ 40 *milli secs*.

Our method, however, being an implicit representation network, would need to be re-trained on each sample. Our framework, and hence our timing analysis, consists of two parts - the cascaded appearance model and the residual plethysmograph model. The cascaded appearance model takes an average time of 8.05 $secs(\pm 0.1$ *secs*) per epoch. Having been

trained on 10 epochs for our experiments, the total training time for the cascaded appearance model is 80.45 *secs*. The training of the residual plethysmograph model, on the other hand, takes 9.17 *secs*($\pm$0.055 *secs*) per epoch. With the model having been trained for 5 epochs, it takes a total of 45.83 *secs* to train the residual plethysmograph model. In summary, the total training time of our entire framework is 126.28 *secs*, i.e., a little over 2 *mins* for a 10-second video sample of size 128×3 sampled at 30 frames per second.

As part of our preliminary analysis, we found that Siren [6] takes $\approx$ 20 *mins* to train on a 2 *sec* video clip. Extrapolating these values, it would take a total of 100 *mins* to evaluate a 10 *secs* video clip. Therefore, despite our slow runtime compared to contemporary methods, we still achieve an acceleration of $\approx 50\times$ over other implicit methods.

By design, both models are run sequentially. For all our experiments, we use workstations with RTX 3090 GPUs and an Intel i9 CPU. From our analysis, we found that data loading is the most computationally expensive operation. Hence, we pre-load the whole video onto the GPU memory and sample the indices within our code. On our systems, we found this to provide a speed-up of $5\times$ when compared to individually loading each batch onto memory. On average, our net memory usage is 2.86 GB while training the appearance model and 2.73 GB while training the residual model. From our observations, these numbers are not significantly high enough to limit the model's deployment since most modern GPUs allow for much higher memory usage.

B.6 Choice of the Parity Metric (r-consistency)

Unlike prior methods [8], we cannot use the difference in performance as a parity metric for OOD evaluation. Difference-based metrics are incomplete. In that, we note that these metrics only take into account the absolute values. However, these values can be easily skewed. Consider an example with 2 underperforming values. Their difference would be small, however, this result would be an inaccurate representation as the values themselves

are sub-optimal. Hence, there is a need to identify a metric that can accurately represent the network's performance on the in distribution and OOD samples.

Taking inspiration from the F-1 score, we define a parity metric for OOD evaluation. Since the F-1 score is generally defined for precision and recall, whose values are between 0 to 1.0 (0 to 100%), we repurpose the metric to work with the Pearson correlation coefficient (r). Mathematically, this is given as:

$$\text{r-consistency} = \text{Harmonic Mean}(\text{r}_{\text{ID}}, \text{r}_{\text{OOD}}) = 2\frac{\text{r}_{\text{ID}} \times \text{r}_{\text{OOD}}}{\text{r}_{\text{ID}} + \text{r}_{\text{OOD}}} \tag{B.33}$$

By taking the Harmonic Mean of the correlation values for the in-distribution dataset and OOD dataset, we can concretely rank an algorithm's ability to generalize to unseen distributions

B.7 Ablation Analysis

B.7.1 Heart Rate Estimation with Different Appearance and Plethysmograph Model Configurations

In this section, we demonstrate the effectiveness of the Cascaded Appearance and Residual Plethysmograph networks for their respective tasks. In Table B.4, we present the results of our method with various configurations for the Cascaded Appearance and Residual Plethysmograph models.

Here, we make an important note. Given our aim to test the effectiveness of our method to best represent the plethysmographic information, we evaluate only the green channel of the Residual Plethysmograph's output. We do not use the neural signal strength masks to improve the SNR, but rather use all pixels in the green channel of the residual output. The values presented in Table B.4 are tested on the in-distribution samples from [8], using the six fold cross validation as proposed in the work.

We observe that a vanilla implementation of the XYT network with sinusoidal activa-

tion functions (that is, without the cascaded model) is able to retain the plethysmographic information, evidenced by the values of the first row. However, since this model does not possess a tractable structure we cannot isolate the plethysmograph signals from the reconstructed outputs - therfore, the performance metrics are poor. From the second row, we notice that adding another sinusoidal XYT model as the residual model adds no value to the reconstruction. In fact, it is detrimental to the estimation as there is minimal residual component to learn from, since the first model has already managed to learn all possible relevant plethysmographic information.

On fixing the appearance model to be our proposed Cascaded Appearance model, we note that the third row of Table B.4 corroborates our theory. That is, the Cascaded Appearance model performs very poorly and is unable to retain the plethysmograph waveform. Hence, in line with our theoretical formulations, our appearance model is able to efficiently perform as an $\mathcal{A}$ function estimator. Following these lines, rows 4 and 5 ablate over the activation function in the plethysmograph model and show the necessity of using the sinusoidal activation, leading the performance to increase by a factor 2.5 over using a ReLU activation.

B.7.2 Using the Difference of the Video and the Appearance in Place of the Plethysmograph Model

A reasonable consideration is whether the plethysmograph implicit model ($\mathcal{B}$-function approximator) is even necessary or if the plethysmograph component can be obtained through a simple difference of the appearance model output from the original video (without fitting a model for this). Table B.5 highlights this configuration (dubbed the difference model) and compares with our proposed method on in-distribution testing for the dataset proposed in [8]. As can be seen, this model performs worse than our proposed method. This empirical observation likely arises as a result of the denoising effect of implicit representations [226, 214], especially since the plethysmograph signal is a low dimensional signal encoded in the high dimensional residual model. This observation emphasizes the necessity of all components of

Table B.4: **Ablation analysis 1: design choice of the implicit decomposition pipeline.** We show the impact of each block on the overall performance of our pipeline. The numbers generated **do not use the neural signal strength masks**. In the table below, **Our Appearance Model** is the Cascaded Appearance model used by our method. **Sinusoidal XYT** is the Residual Plethysmograph model used by our method. Finally, **ReLU XYT** represents a simple architecture structurally identical to the Residual Plethysmograph, but with ReLU activations in place of the sinusoidal activation.

Appearance Model	Plethysmograph Model	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Sinusoidal XYT	None	4.94	6.93%	12.02	0.65
Sinusoidal XYT	Sinusoidal XYT	17.62	27.53%	21.67	0.04
Our Appearance Model	None	12.45	17.54%	18.25	0.21
Our Appearance Model	ReLU XYT	4.28	5.42%	9.83	0.76
Our Appearance Model	Sinusoidal XYT	1.57	2.03%	4.79	0.94

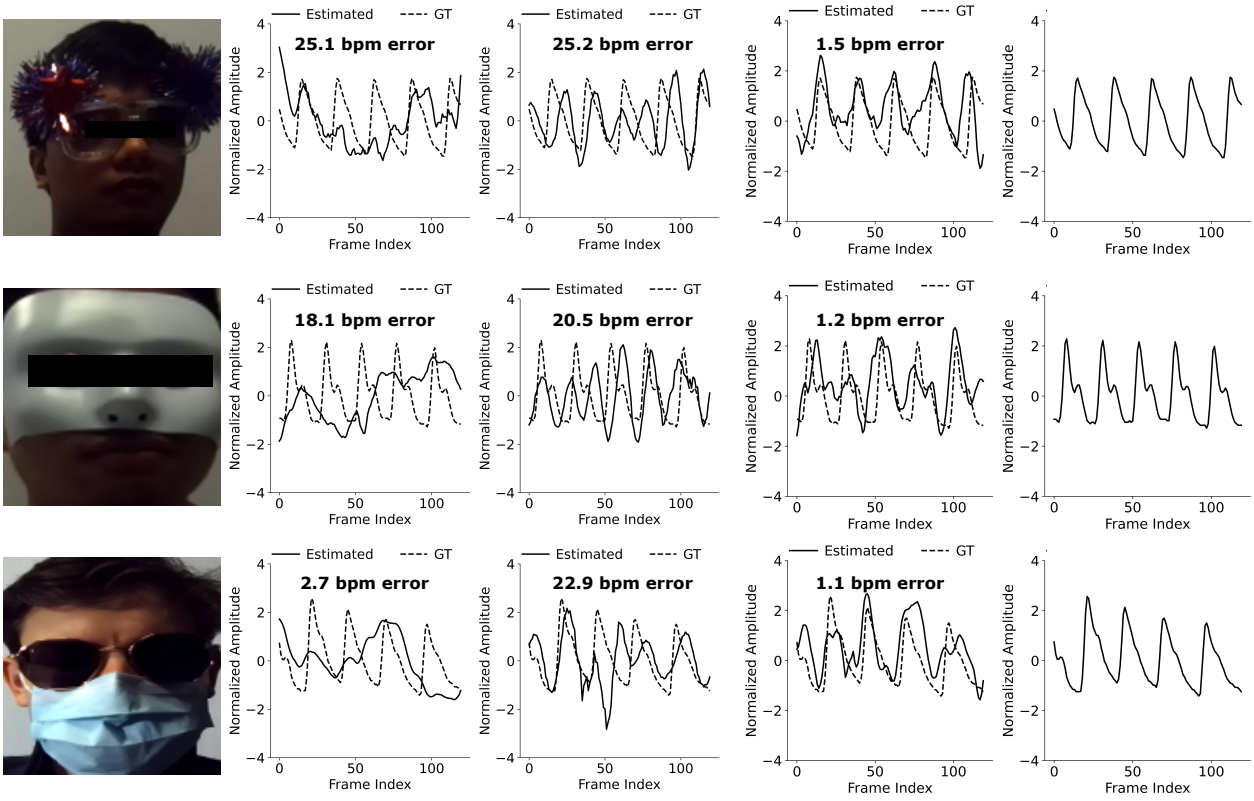
our pipeline.

B.8 Detailed Analysis of Out-of-distribution Performance

Section 6.2 of the main paper discusses overall qualitative and quantitative performance of the proposed and prior baseline methods on our optically challenging out-of-distribution dataset. In this section, we will analyze more granular stratified analysis of the proposed and baseline methods.

B.8.1 Dataset Description

We propose a novel optically challenging dataset (inference-only) to test the performance of remote plethysmography (rPPG) methods. Based on this philosophy, the dataset consists of 104 videos of around 30 second duration across 13 participants, with aligned contact



(a) Frame (b) [25] (c) [9] (d) Ours (e) Gr. Truth

Figure B.3: **The proposed method is a superior performer across low-light, optically challenging scenes.** We compare the best-performing algorithmic baseline, the best deep learning baseline and our proposed method (all trained on the dataset from [8] where applicable).

Table B.5: **Ablation analysis 2: we show the importance of the proposed model over a simple 'difference model'**, where the plethysmograph/blood component is extracted through a simple difference of the original video and the estimated appearance component. The metrics shown are **evaluated using neural signal strength masks** trained on the dataset from [8], with the numbers supplied being the in-distribution test results on the same dataset (please correlate with Table 2 in main paper, quadrant 3).

Configuration	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Difference model	1.51	1.98%	4.28	0.95
Ours	**1.22**	**1.61%**	**4.01**	**0.96**

plethysmograph ground truth data.

We follow a similar camera setup to [8] while collecting our OOD dataset. We use a monocular camera stream from a Zed2 camera from Stereolabs and a Contec CMS-60C pulse oximeter for ground truth.

Data samples in this dataset consist of various attributes, such as face occlusions (face mask, eye glasses, face paint etc.), low light, motion, talking and so on. A particular video may have one of more of these attributes. Since the primary focus is optically challenging data, most of the samples possess some kind of **face occlusion (77 samples)**. Additionally, we stratify the dataset according to **well lit (49 samples)** and **low light settings (55 samples)**. We define low light as settings where the ambient room lighting is dimmed, or an uncontrolled low light setting (such as the outdoor window reflection example from Figure 7, main paper).

Additionally, Figure 8 in the main paper highlights performance on natural motions, which are included as part of our dataset. Since this is a secondary focus (primary being optically challenging OOD), the dataset contains relatively fewer such samples: **12 in total**, with 6 samples having participants talking, and 6 samples having participants moving their

heads side to side. The dataset, along with metadata tags for attributes, will be released post acceptance.

In our OOD dataset, we are able to share full participant videos with users (without identifying information such as names), but only after having recorded contact details of dataset users. The dataset is released after potential users fill out a request form on the project webpage. This will enable us to keep track of all users of the dataset and rescind access if ever necessary. The videos are also audio-stripped, so apart from the video frames and the plethysmograph waveform itself, no other participant information is ever available to the dataset users.

B.8.2 Performance Across Lighting Configurations

Figure B.3 highlights selected qualitative results on low light samples from our OOD dataset. Across three different samples, the proposed method shows superior performance both in terms of waveform reconstruction and heart rate estimation. For dataset scale results, we can then look at Table B.6 and Table B.7, where methods are trained on the data from [8] and [16] respectively. The proposed method is the best performer, both for well lit and low-light configurations, by a large margin (1.5-2x better than next best performing method). Overall, our method, with the signal strength mask model trained on the dataset from [8] is the best performer across both configurations. Additionally, the reader may note from Table B.6 that the second best performing method for the two configurations ([27] for well lit, and [25] for low light) indicate that different baseline methods are better depending on the lighting condition, wheras our method is considerably better for both settings. This is an additional point of merit for our method.

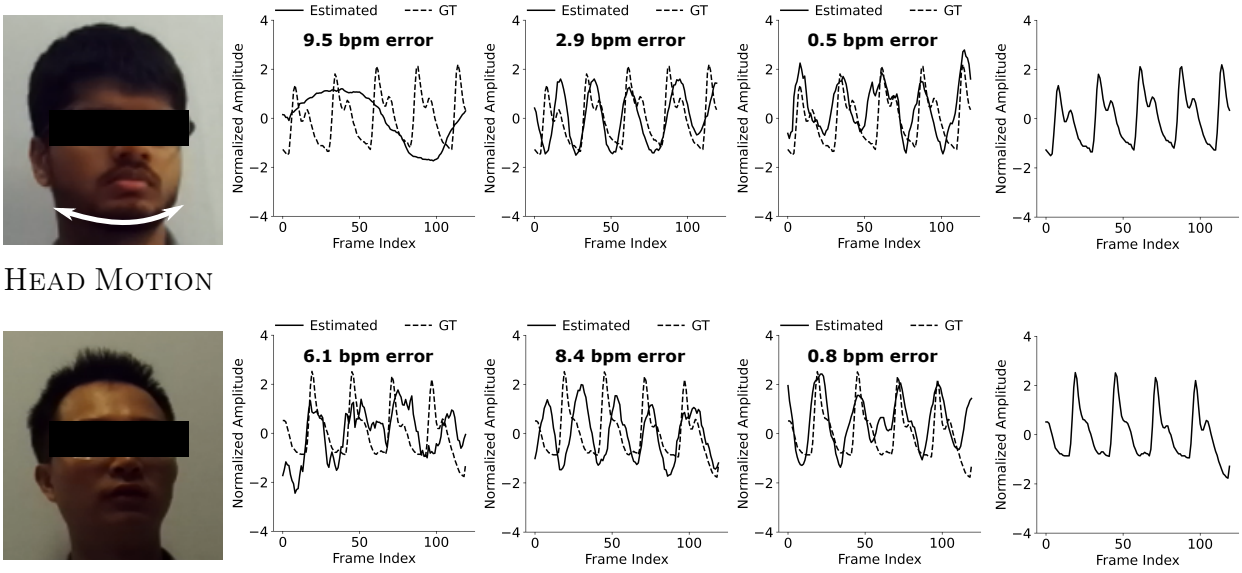
B.8.3 Performance on Face Occlusions

A majority of the videos in our OOD dataset consist of optically challenging face occlusion, making the rPPG problem challenging. Table B.8 and Table B.9 show relevant dataset-wide results. The proposed method is considerably better than all prior methods, by a margin of close to 1.5-2 times. That is, the proposed method is better performing and more efficient in these low light settings. Additionally, we note that the proposed method when trained on the dataset from [8] is the best performing.

B.8.4 Performance on Motion Videos

As an empirical test of robustness, we extend our OOD analysis to Figure B.4. While we previously dealt with specific optically challenging OOD examples, here we take a look at motion-induced obfuscations that can occur in a real-world setting. This includes 2 scenarios: (a) the volunteer shaking their head from side to side and (b) talking. While our method is not specifically designed to handle motion, we observe little degradation in terms of performance for the displayed samples. This is evident from the penultimate column of Figure B.4. Our method archives errors of under 2 bpm MAE for these samples while capturing salient features of the plethysmograph signal. While not being a clear best performer for this setting, our method is competitive, comparable, or better in performance than baseline methods. We must note that the number of samples for motion in our dataset is fewer than those for optically challenging scenes. However, this scale is sufficient for this validation that our method is able to handle motion.

Table B.8, B.9 show detailed quantitative metrics. As can be seen, even though not designed to handle motion specifically, our propsed method is very competetive when compared with prior state of the art methods, for the motion and talking scenarios. For example, in the talking setting, the proposed method is the best performer. While models trained on both [8] and [16] data are best performing, the model from [16] data is found to work better,

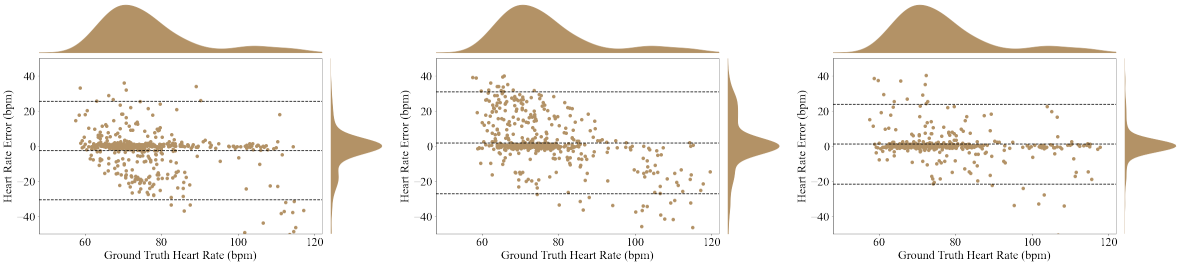


HEAD MOTION

TALKING

(a) FRAME (b) [25] (c) [9] (d) OURS (e) GR. TRUTH

Figure B.4: **The proposed method shows comparable or better performance across secondary OOD settings, such as talking and motion, when compared with prior art.** We compare the best-performing algorithmic baseline, the best deep learning baseline and our proposed method (all trained on the dataset from [8] where applicable) on scenes that are part of our OOD dataset.



(a) [25] (b) [9] (c) Ours

Figure B.5: **Bland-Altman plots are used to quantify heart rate performance in clinical literature as they span a range of heart rates [8].** The x-axis represents ground truth heart rate, while the y-axis represents heart rate estimation error. The horizontal lines mark the mean error and 1.96 times the standard deviation. A smaller vertical spread indicates a lower error and is desired behavior. A trend (as in (b)) indicates correlated errors, which is non-ideal.

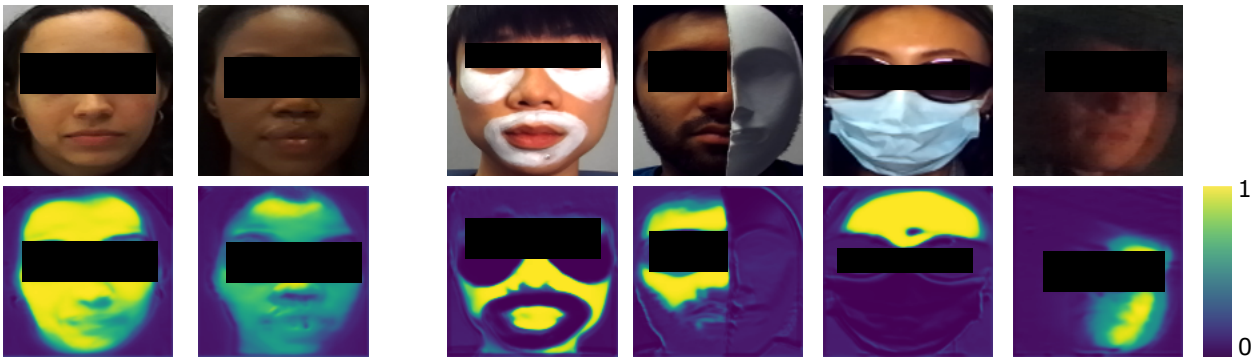
interestingly. The next best method is found to be [31], specifically when trained on [16] data.

In the case of side to side head motion, the proposed method is not the best performer. That being said, performance is not much worse than the best performing methods. Again, the model from [16] data is found to work better. Therefore, while not the best performer, when looking at all motion and talking videos in summary, the proposed method indeed is either comparably performing or superior, and is certainly able to reliably work in these challenging OOD conditions comparably to the best prior methods.

B.8.5 Comparing Bland-Altman Plots

We visualize sample-wise errors and group error trends through Bland-Altman (BA) plots. We plot the ground truth heart rate on the x-axis and the heart rate estimation error on the y-axis. Therefore, a smaller vertical spread in the BA plots indicates better perfor-

mance. Figure B.5 shows the BA plots for Green [25] (best baseline algorithmic method), PhysFormer [9] (best baseline learning method), and our proposed method on the dataset provided by [8]. Our method shows the smallest veritical spread of errors, correlating with the OOD performance metrics from Table 1, main paper.



(a) In-Distribution (b) Out-of-Distribution

Figure B.6: **Our predicted neural signal strength masks accurately generalize to OOD configurations, in addition to performing well on in-distribution test samples.** (a) On inference samples from the [8] dataset, our method is able to identify high-fidelity details such as eyes, hair and specular highlights. (b) This high-fidelity nature of the reconstruction continues in OOD inference, such as optically challenging samples. Unseen phenomena such as face paint, face masks, sunglasses, and even reflections through semi-transparent glass windows are appropriately handled.

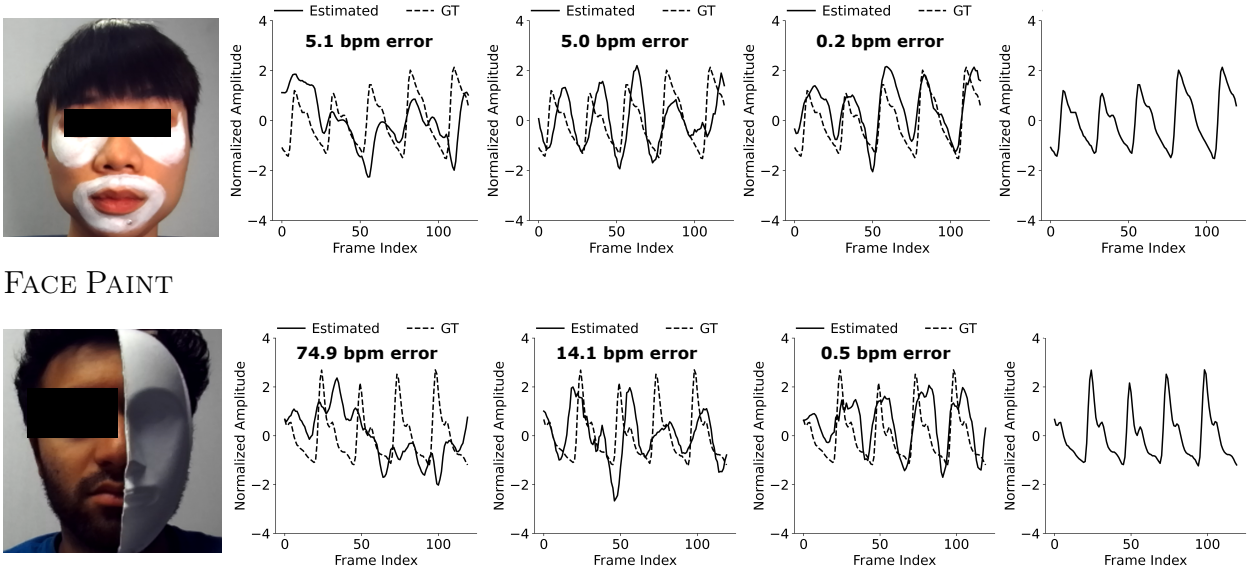
B.8.6 Neural Signal Strength Masks

The neural signal strength masks, generated as an intermediate signal in the neural refinement model, are advantageous in improving the signal strength of the plethysmograph estimates. Being the only supervised block of our entire framework, we test performance when trained on the two different datasets (discussed earlier). From Tables 1 and 2, main paper, we find that our framework is not significantly affected by the dataset used to train, therefore showing that the masks are easy to generalize.

Qualitatively, Figure B.6 shows example masks. Our generated masks are consistent with our understanding of light transport for plethysmography, as discussed in Section B.1.1. Therefore, the neural signal strength masks assign a lower priority to pixels associated with specular highlights (specular regions on the forehead in the fifth column). It also assigns lower

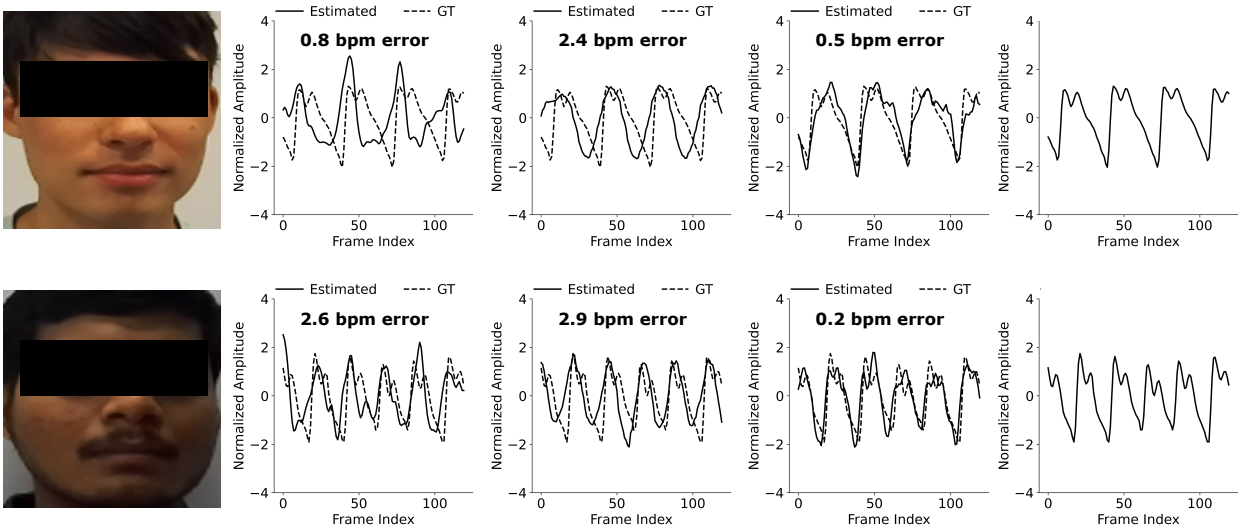
weights to the regions underneath the eye and near the nostrils compared to the cheeks and forehead. These masks are also able to ignore occlusions, such as face paint (third column) face masks and beards (fourth column), and surgical masks and sunglasses (fifth column).

One particularly interesting sample is the Window Reflection case (final column). The captured image has a dual reflection caused by the double-paned glass window. This sample is especially optically challenging due to attenuation and interference from reflections and multiple light sources. Our method can, however, distinguish between these reflections and appropriately weigh the second reflection less than the first reflection, with the regions around the cheeks being the strongest regions of interest.



(a) Frame (b) [25] (c) [9] (d) Ours (e) Gr. Truth

Figure B.7: **Additional challenging OOD optical settings.** Results shown use models trained on the dataset proposed in [8], where applicable.



(a) Frame (b) [27] (c) [9] (d) Ours (e) Gr. Truth

Figure B.8: **Additional results for in distribution inference on a skin tone diverse dataset.**

B.9 Additional Qualitative Results

Figures B.7 and B.8 show additional qualitative results for the OOD and in-distribution settings respectively. These add to the results in Figures 6 and 7 of the main paper.

B.10 Additional Baselines

We show quantitative performance metrics on additional baselines for both OOD and in-distribution settings in Table B.10 and B.11. The proposed method outperforms these both in-distribution and OOD.

B.11 Future Work

In the domain of plethysmography, the use of implicit representations is new to the best of our knowledge and, therefore, sets the stage for future work in architectural improvements, runtime optimization, and performance enhancement. Additionally, the use of implicit representation as functional decomposers can generally be applicable to a range of different applications such as reflection removal, low-level imaging tasks such as deraining and dehazing, and so on. Finally, other kinds of medical devices and complex problems, such as contactless pulse oximetry, may benefit from such an analysis-by-synthesis approach. Finally, more research into other kinds of OOD phenomena from rPPG is desirable.

Table B.6: **(Trained on [8] dataset) Performance on our OOD dataset across lighting intensity indicates state of the art performance over both well lit and low light conditions.** OOD performance metrics include T-Test (APE %), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Pearson correlation coefficient (r). The best and second-best-performing numbers are shown in green and yellow, respectively.

		OOD Performance Metrics				
	Method	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Well Lit	**Green** [25]	32.24	8.62	10.65%	16.50	0.13
	POS [27]	20.41	5.18	7.41%	11.69	0.63
	CHROM [26]	25.71	6.81	9.62%	13.35	0.58
	ICA [135]	23.67	6.66	8.61%	14.65	0.33
	DeepPhys [30]	79.59	16.81	21.72%	20.88	0.03
	TS-CAN [137]	72.24	15.59	20.07%	19.73	0.06
	EfficientPhys [280]	42.86	12.39	16.26%	20.68	-0.03
	PhysNet [31]	34.69	7.14	9.56%	12.91	0.53
	PhysFormer [9]	39.18	9.44	12.64%	15.36	0.33
	Ours	12.65	3.38	4.26%	9.99	0.68
Low Light	**Green** [25]	30.91	7.07	9.11%	12.58	0.66
	POS [27]	42.18	10.55	14.30%	17.22	0.33
	CHROM [26]	46.55	11.55	15.29%	17.78	0.23
	ICA [135]	37.82	9.51	12.19%	15.89	0.44
	DeepPhys [30]	76.36	18.95	23.08%	23.84	0.09
	TS-CAN [137]	78.55	19.49	23.77%	23.87	-0.02
	EfficientPhys [280]	45.45	15.15	20.08%	24.60	0.26
	PhysNet [31]	60.73	14.89	18.46%	20.70	0.19
	PhysFormer [9]	40.36	9.70	12.55%	14.94	0.30
	Ours	19.27	5.71	7.68%	12.95	0.61

Table B.7: **(Trained on [16] dataset) Performance on our OOD dataset across lighting intensity indicates state of the art performance over both well lit and low light conditions.** OOD performance metrics include T-Test (APE %), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Pearson correlation coefficient (r). The best-performing numbers are shown in green (second best excluded since algorithmic methods from Table B.6 are second best). Algorithmic methods excluded from this table since they are not trained on a particular dataset - numbers same as Table B.6.

		OOD Performance Metrics				
	Method	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Well Lit	**DeepPhys** [30]	76.73	17.83	23.05%	22.36	-0.13
	TS-CAN [137]	73.06	16.04	20.79%	20.86	0.08
	EfficientPhys [280]	30.61	10.29	14.40%	20.39	0.32
	PhysNet [31]	27.35	7.48	10.44%	15.53	0.47
	PhysFormer [9]	46.53	9.97	13.09%	15.31	0.13
	Ours	14.29	3.17	4.07%	7.92	0.78
Low Light	**DeepPhys** [30]	79.64	19.37	23.31%	23.98	-0.03
	TS-CAN [137]	80.73	20.61	25.42%	25.29	-0.06
	EfficientPhys [280]	33.45	13.95	18.90%	25.78	0.27
	PhysNet [31]	39.27	9.68	11.67%	16.67	0.28
	PhysFormer [9]	50.18	10.80	13.41%	15.73	0.32
	Ours	21.45	6.83	9.24%	14.74	0.51

Table B.8: **(Trained on [8] dataset) Performance on our OOD dataset across face occlusions, talking and motion. Our method is state of the art for occlusions, while being close to optimal or better for talking and motion.** OOD performance metrics include T-Test (APE %), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Pearson correlation coefficient (r). The best and second-best-performing numbers are shown in green and yellow, respectively.

		OOD Performance Metrics				
	Method	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Face Occlusions	**Green** [25]	30.91	7.71	9.90%	14.81	0.17
	POS [27]	25.97	7.16	10.44%	14.84	0.43
	CHROM [26]	32.99	8.82	12.68%	15.82	0.41
	ICA [135]	30.39	7.85	10.65%	15.38	0.25
	DeepPhys [30]	76.10	16.14	21.17%	20.01	0.06
	TS-CAN [137]	71.43	14.92	19.62%	18.80	0.07
	EfficientPhys [280]	39.48	12.40	16.85%	21.76	0.00
	PhysNet [31]	47.79	9.92	13.40%	15.31	0.36
	PhysFormer [9]	38.18	9.09	12.56%	14.59	0.30
	Ours	15.58	4.43	6.15%	12.03	0.51
Talking	**Green** [25]	73.33	13.95	17.46%	16.84	0.08
	POS [27]	43.33	5.97	7.43%	8.40	0.37
	CHROM [26]	43.33	6.87	8.50%	10.02	0.05
	ICA [135]	56.67	11.81	14.78%	14.73	0.03
	DeepPhys [30]	73.33	17.11	21.43%	20.83	0.28
	TS-CAN [137]	80.00	20.10	25.09%	23.14	0.05
	EfficientPhys [280]	50.00	13.42	16.90%	19.38	0.19
	PhysNet [31]	23.33	5.74	7.27%	10.22	0.23
	PhysFormer [9]	33.33	5.92	7.46%	10.20	0.08
	Ours	23.33	5.40	6.71%	9.04	0.51
	Green [25]	50.00	10.09	13.88%	13.30	0.41
	POS [27]	33.33	6.23	8.61%	10.05	0.39

Table B.9: **(Trained on [16] dataset) Performance on our OOD dataset across face occlusions, talking and motion. Our method is state of the art for occlusions, while being close to optimal or better for talking and motion.** OOD performance metrics include T-Test (APE %), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Pearson correlation coefficient (r). The best-performing numbers are shown in green (second best excluded due to Table B.6). Algorithmic methods excluded from this table since they are not trained on a particular dataset - numbers same as Table B.6.

		OOD Performance Metrics				
	Method	T-Test (APE %) ↓	MAE ↓	MAPE ↓	RMSE ↓	r ↑
Face Occlusions	**DeepPhys** [30]	74.29	16.28	21.38%	20.43	-0.07
	TS-CAN [137]	75.06	16.29	21.63%	20.58	0.05
	EfficientPhys [280]	32.47	12.97	18.56%	24.94	0.16
	PhysNet [31]	30.91	7.24	10.18%	14.16	0.46
	PhysFormer [9]	48.05	9.79	13.21%	14.59	0.17
	Ours	17.40	5.01	7.06%	12.50	0.50
Talking	**DeepPhys** [30]	83.33	17.23	21.39%	19.90	-0.08
	TS-CAN [137]	70.00	17.99	22.52%	21.67	-0.08
	EfficientPhys [280]	43.33	11.48	14.14%	18.82	0.47
	PhysNet [31]	26.67	5.21	6.50%	8.79	0.50
	PhysFormer [9]	36.67	7.62	9.50%	11.02	0.22
	Ours	23.33	5.06	6.32%	8.19	0.75
Motion	**DeepPhys** [30]	83.33	17.70	24.21%	19.41	0.16
	TS-CAN [137]	90.00	16.35	22.34%	18.47	-0.37
	EfficientPhys [280]	26.67	7.79	10.74%	15.19	0.24
	PhysNet [31]	33.33	6.45	8.89%	10.98	0.34
	PhysFormer [9]	33.33	5.86	8.07%	8.64	0.37
	Ours	16.67	4.41	6.02%	9.61	0.30

Table B.10: **Additional OOD baselines.** Train on [8] and test on our OOD dataset.

Method	MAE ↓	MAPE ↓	RMSE ↓
ContrastPhys+ (unsupervised) [283]	15.05	20.55%	17.34
ContrastPhys+ (semi-supervised) [283]	13.03	17.04%	15.29
ContrastPhys+ (fully-supervised) [283]	13.42	17.43%	15.63

Table B.11: **In-distribution performance for additional baselines.** Train and test on [8].

Method	MAE ↓	MAPE ↓	RMSE ↓
ContrastPhys+ (unsupervised) [283]	2.01	2.91%	2.54
ContrastPhys+ (semi-supervised) [283]	1.70	2.44%	2.16
ContrastPhys+ (fully-supervised) [283]	1.64	2.23%	2.21

REFERENCES

[1] Y. Ba, Z. Wang, K. D. Karinca, O. D. Bozkurt, and A. Kadambi, "Overcoming difficulty in obtaining dark-skinned subjects for remote-ppg by synthetic augmentation," *arXiv preprint arXiv:2106.06007*, 2021.

[2] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019.

[3] C. L. Blake and C. J. Merz, "Uci repository of machine learning databases, 1998," 1998.

[4] Y. Yao, H. Yu, J. Mu, J. Li, and H. Pu, "Estimation of the gender ratio of chickens based on computer vision: Dataset and exploration," *Entropy*, vol. 22, no. 7, p. 719, 2020.

[5] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5810–5818.

[6] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020.

[7] L. Mai and F. Liu, "Motion-adjustable neural implicit video representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 738–10 747.

[8] A. Vilesov, P. Chari, A. Armouti, A. B. Harish, K. Kulkarni, A. Deoghare, L. Jalilian, and A. Kadambi, "Blending camera and 77 ghz radar sensing for equitable, robust plethysmography," in *ACM Trans. Graph. (SIGGRAPH)*, 2022.

[9] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: facial video-based physiological measurement with temporal difference transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4186–4196.

[10] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow, "Deep blending for free-viewpoint image-based rendering," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 6, pp. 1–15, 2018.

[11] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira,

M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The replica dataset: A digital replica of indoor spaces," 2019.

[12] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[13] S. Kobayashi, E. Matsumoto, and V. Sitzmann, "Decomposing nerf for editing via feature field distillation," in *NeurIPS*, vol. 35, 2022. [Online]. Available: https://arxiv.org/pdf/2205.15585.pdf

[14] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 729–19 739.

[15] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1851–1860.

[16] Z. Wang, Y. Ba, P. Chari, O. D. Bozkurt, G. Brown, P. Patwa, N. Vaddi, L. Jalilian, and A. Kadambi, "Synthetic generation of face videos with plethysmograph physiology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 587–20 596.

[17] H. Tang, S. Cohen, B. Price, S. Schiller, and K. N. Kutulakos, "Depth from defocus in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2740–2748.

[18] R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: motion deblurring using fluttered shutter," in *Acm Siggraph 2006 Papers*, 2006, pp. 795–804.

[19] K. L. Bouman, V. Ye, A. B. Yedidia, F. Durand, G. W. Wornell, A. Torralba, and W. T. Freeman, "Turning corners into cameras: Principles and methods," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2270–2278.

[20] A. Kadambi, V. Taamazyan, B. Shi, and R. Raskar, "Polarized 3d: High-quality depth sensing with polarization cues," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3370–3378.

[21] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," *arXiv preprint arXiv:2401.10891*, 2024.

[22] P. De Bièvre, "The 2012 international vocabulary of metrology:âĂŸâĂŸvimâĂŹâĂŹ," *Chemistry International–Newsmagazine for IUPAC*, vol. 34, no. 3, pp. 26–27, 2012.

[23] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[24] M. A. R. Ahad, U. Mahbub, and T. Rahman, *Contactless Human Activity Analysis*. Springer, 2021.

[25] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[26] 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.

[27] 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.

[28] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcamâĂŤa non-contact method for evaluating cardiac activity," in *2011 federated conference on computer science and information systems (FedCSIS)*. IEEE, 2011, pp. 405–410.

[29] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[30] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 349–365.

[31] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," *arXiv preprint arXiv:1905.02419*, 2019.

[32] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "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*, 2020, pp. 284–285.

[33] M. Alizadeh, G. Shaker, J. C. M. De Almeida, P. P. Morita, and S. Safavi-Naeini, "Remote monitoring of human vital signs using mm-wave fmcw radar," *IEEE Access*, vol. 7, pp. 54 958–54 968, 2019.

[34] W. Lv, W. He, X. Lin, and J. Miao, "Non-contact monitoring of human vital signs using fmcw millimeter wave radar in the 120 ghz band," *Sensors*, vol. 21, no. 8, p. 2732, 2021.

[35] S. Wu, T. Sakamoto, K. Oishi, T. Sato, K. Inoue, T. Fukuda, K. Mizutani, and H. Sakai, "Person-specific heart rate estimation with ultra-wideband radar using convolutional neural networks," *IEEE Access*, vol. 7, pp. 168 484–168 494, 2019.

[36] L. Ren, L. Kong, F. Foroughian, H. Wang, P. Theilmann, and A. E. Fathy, "Comparison study of noncontact vital signs detection using a doppler stepped-frequency continuous-wave radar and camera-based imaging photoplethysmography," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 9, pp. 3519–3529, 2017.

[37] Ø. Aardal, Y. Paichard, S. Brovoll, T. Berger, T. S. Lande, and S.-E. Hamran, "Physical working principles of medical radar," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1142–1149, 2012.

[38] T. Wu, V. Blazek, and H. J. Schmitt, "Photoplethysmography imaging: a new non-invasive and noncontact method for mapping of the dermal perfusion changes," in *Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics*, vol. 4163. International Society for Optics and Photonics, 2000, pp. 62–70.

[39] T. Wu, "Ppgi: New development in noninvasive and contactless diagnosis of dermal perfusion using near infrared light," *J. GCPD eV*, vol. 7, no. 1, pp. 17–24, 2003.

[40] F. P. Wieringa, F. Mastik, and A. F. van der Steen, "Contactless multiple wavelength photoplethysmographic imaging: A first step toward âĂIJspo 2 cameraâĂİ technology," *Annals of biomedical engineering*, vol. 33, no. 8, pp. 1034–1041, 2005.

[41] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3430–3437.

[42] E. Magdalena Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "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*, 2018, pp. 1272–1281.

[43] V. Vizbara, "Comparison of green, blue and infrared light in wrist and forehead photoplethysmography," *BIOMEDICAL ENGINEERING 2016*, vol. 17, no. 1, 2013.

[44] C. Barbosa Pereira, M. Czaplik, V. Blazek, S. Leonhardt, and D. Teichmann, "Monitoring of cardiorespiratory signals using thermal imaging: a pilot study on healthy human subjects," *Sensors*, vol. 18, no. 5, p. 1541, 2018.

[45] Y. Kim, Y. Park, J. Kim, and E. C. Lee, "Remote heart rate monitoring method using infrared thermal camera," *International Journal of Engineering Research and Technology*, vol. 11, no. 3, pp. 493–500, 2018.

[46] C. Hurter and D. McDuff, "Cardiolens: remote physiological monitoring in a mixed reality environment," in *ACM siggraph 2017 emerging technologies*, 2017, pp. 1–2.

[47] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2019.

[48] E. Lee, E. Chen, and C.-Y. Lee, "Meta-rppg: Remote heart rate estimation using a transductive meta-learner," in *European Conference on Computer Vision*. Springer, 2020, pp. 392–409.

[49] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1373–1384, 2021.

[50] Z. Wang, Y. Ba, P. Chari, O. Bozkurt, G. Brown, P. Patwa, N. Vaddi, L. Jalilian, and A. Kadambi, "Synthetic generation of face videos with plethysmograph physiology," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[51] J. C. Lin, "Noninvasive microwave measurement of respiration," *Proceedings of the IEEE*, vol. 63, no. 10, pp. 1530–1530, 1975.

[52] L. Ren, H. Wang, K. Naishadham, O. Kilic, and A. E. Fathy, "Phase-based methods for heart rate detection using uwb impulse doppler radar," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 10, pp. 3319–3331, 2016.

[53] A. D. Droitcour, O. Boric-Lubecke, V. M. Lubecke, J. Lin, and G. T. Kovacs, "Range correlation and i/q performance benefits in single-chip silicon doppler radars for noncontact cardiopulmonary monitoring," *IEEE Transactions on Microwave Theory and Techniques*, vol. 52, no. 3, pp. 838–848, 2004.

[54] A. D. Droitcour, *Non-contact measurement of heart and respiration rates with a single-chip microwave doppler radar*. Stanford University, 2006.

[55] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 962–970.

[56] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, pp. 4349–4357, 2016.

[57] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8919–8928.

[58] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[59] N. Srivastava, R. Salakhutdinov *et al.*, "Multimodal learning with deep boltzmann machines." in *NIPS*, vol. 1. Citeseer, 2012, p. 2.

[60] M. Kaur and D. Singh, "Fusion of medical images using deep belief networks," *Cluster Computing*, vol. 23, no. 2, pp. 1439–1453, 2020.

[61] V. Singh, N. K. Verma, Z. U. Islam, and Y. Cui, "Feature learning using stacked autoencoder for shared and multimodal fusion of medical images," in *Computational Intelligence: Theories, Applications and Future Directions-Volume I*. Springer, 2019, pp. 53–66.

[62] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *arXiv preprint arXiv:2109.00412*, 2021.

[63] T. Negishi, S. Abe, T. Matsui, H. Liu, M. Kurosawa, T. Kirimoto, and G. Sun, "Contactless vital signs measurement system using rgb-thermal image sensors and its clinical screening test on patients with seasonal influenza," *Sensors*, vol. 20, no. 8, p. 2171, 2020.

[64] K. Matsumura, S. Toda, and Y. Kato, "Rgb and near-infrared light reflectance/transmittance photoplethysmography for measuring heart rate during motion," *IEEE Access*, vol. 8, pp. 80 233–80 242, 2020.

[65] C. Monitors, "Heart rate meters, and alarms," *ANSI/AAMI Standard EC13*, 2002.

[66] B. W. Nelson and N. B. Allen, "Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study," *JMIR mHealth and uHealth*, vol. 7, no. 3, p. e10828, 2019.

[67] C. T. Association, "Physical activity monitoring for heart rate, ansi/cta-2065," 2018.

[68] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[69] S. Barocas, M. Hardt, and A. Narayanan, "Fairness in machine learning," *Nips tutorial*, vol. 1, p. 2017, 2017.

[70] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," in *2013 IEEE 13th international conference on data mining*. IEEE, 2013, pp. 71–80.

[71] T. Igarashi, K. Nishino, and S. K. Nayar, *The appearance of human skin: A survey.* Now Publishers Inc, 2007.

[72] S. Alotaibi and W. A. Smith, "A biophysical 3d morphable model of face appearance," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 824–832.

[73] R. R. Anderson and J. A. Parrish, "The optics of human skin," *Journal of Investigative Dermatology*, vol. 77, no. 1, pp. 13–19, 1981. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022202X15461251

[74] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 553–560, iSSN: 1063-6919.

[75] P. Pai and M. Z. A. Khan, "Comparison of sc and mrc receiver complexity for three antenna diversity systems," in *2008 24th Biennial Symposium on Communications.* IEEE, 2008, pp. 302–305.

[76] T. Zheng, Z. Chen, S. Zhang, C. Cai, and J. Luo, "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*, 2021, pp. 111–124.

[77] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9012–9020.

[78] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning.* PMLR, 2015, pp. 448–456.

[79] S. Sachdeva *et al.*, "Fitzpatrick skin typing: Applications in dermatology," *Indian journal of dermatology, venereology and leprology*, vol. 75, no. 1, p. 93, 2009.

[80] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

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

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

[83] D. Kim, O. Hilliges, S. Izadi, A. D. Butler, J. Chen, I. Oikonomidis, and P. Olivier, "Digits: freehand 3d interactions anywhere using a wrist-worn gloveless sensor," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 2012, pp. 167–176.

[84] J. A. Paradiso, "The laserwall," *The LaserWall*, 1997. [Online]. Available: http://paradiso.media.mit.edu/SpectrumWeb/captions/Laser.html

[85] J. A. Paradiso, K.-Y. Hsiao, J. Strickon, and P. Rice, "New sensor and music systems for large interactive surfaces," in *ICMC*. Citeseer, 2000.

[86] G. Laput and C. Harrison, "Surfacesight: a new spin on touch, user, and object sensing for iot experiences," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[87] Á. Cassinelli, S. Perrin, and M. Ishikawa, "Smart laser-scanner for 3d human-machine interface," in *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1138–1139.

[88] R. Xiao, C. Harrison, K. D. Willis, I. Poupyrev, and S. E. Hudson, "Lumitrack: low cost, high precision, high speed tracking with projected m-sequences," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013, pp. 3–12.

[89] J. Zizka, A. Olwal, and R. Raskar, "Specklesense: fast, precise, low-cost and compact motion sensing using laser speckle," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 489–498.

[90] D. Li, X. Liu, Y. Liang, J. Fan, and L. Wang, "A low-cost portable nanophotonic sensor based on a smartphone: A system readily available for many applications," *IEEE Nanotechnology Magazine*, vol. 13, no. 3, pp. 6–12, 2019.

[91] K. Kim, H. Yu, J. Koh, J. H. Shin, W. Lee, and Y. Park, "Remote sensing of pressure inside deformable microchannels using light scattering in scotch tape," *Optics Letters*, vol. 41, no. 8, pp. 1837–1840, 2016.

[92] B. T. Feng, A. C. Ogren, C. Daraio, and K. L. Bouman, "Visual vibration tomography: Estimating interior material properties from monocular video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 231–16 240.

[93] A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, F. Durand, and W. T. Freeman, "Visual vibrometry: Estimating material properties from small motion in video," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 5335–5343.

[94] M. Sato, S. Yoshida, A. Olwal, B. Shi, A. Hiyama, T. Tanikawa, M. Hirose, and R. Raskar, "Spectrans: Versatile material classification for interaction with textureless, specular and transparent surfaces," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2191–2200.

[95] M. D. Dogan, S. V. Acevedo Colon, V. Sinha, K. Akşit, and S. Mueller, "Sensicut: Material-aware laser cutting using speckle sensing and deep learning," in *The 34th Annual ACM Symposium on User Interface Software and Technology*, 2021, pp. 24–38.

[96] K. Jo, M. Gupta, and S. K. Nayar, "Spedo: 6 dof ego-motion sensor using speckle defocus imaging," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4319–4327.

[97] B. M. Smith, M. O'Toole, and M. Gupta, "Tracking multiple objects outside the line of sight using speckle imaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6258–6266.

[98] B. M. Smith, P. Desai, V. Agarwal, and M. Gupta, "Colux: Multi-object 3d micro-motion analysis using speckle imaging," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[99] A. Olwal, A. Bardagjy, J. Zizka, and R. Raskar, "Speckleeye: gestural interaction for embedded electronics in ubiquitous computing," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 2012, pp. 2237–2242.

[100] Y. C. Shih, A. Davis, S. W. Hasinoff, F. Durand, and W. T. Freeman, "Laser speckle photography for surface tampering detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 33–40.

[101] N. Wu and S. Haruyama, "Real-time audio detection and regeneration of moving sound source based on optical flow algorithm of laser speckle images," *Optics Express*, vol. 28, no. 4, pp. 4475–4488, 2020.

[102] Y. Zhang, G. Laput, and C. Harrison, "Vibrosight: Long-range vibrometry for smart environment sensing," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 225–236.

[103] Y. Zhang, S. Mayer, J. T. Gonzalez, and C. Harrison, "Vibrosight++: City-scale sensing using existing retroreflective signs and markers," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[104] M. Sheinin, D. Chan, M. O'Toole, and S. G. Narasimhan, "Dual-shutter optical vibration sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 324–16 333.

[105] A. K. Dunn, H. Bolay, M. A. Moskowitz, and D. A. Boas, "Dynamic imaging of cerebral blood flow using laser speckle," *Journal of Cerebral Blood Flow & Metabolism*, vol. 21, no. 3, pp. 195–201, 2001.

[106] J. D. Briers and S. Webster, "Laser speckle contrast analysis (lasca): a nonscanning, full-field technique for monitoring capillary blood flow," *Journal of biomedical optics*, vol. 1, no. 2, pp. 174–179, 1996.

[107] Y.-C. Huang, T. L. Ringold, J. S. Nelson, and B. Choi, "Noninvasive blood flow imaging for real-time feedback during laser therapy of port wine stain birthmarks," *Lasers in Surgery and Medicine: The Official Journal of the American Society for Laser Medicine and Surgery*, vol. 40, no. 3, pp. 167–173, 2008.

[108] Y. Tamaki, M. Araie, E. Kawamoto, S. Eguchi, and H. Fujii, "Noncontact, two-dimensional measurement of retinal microcirculation using laser speckle phenomenon." *Investigative ophthalmology & visual science*, vol. 35, no. 11, pp. 3825–3834, 1994.

[109] A. B. Parthasarathy, E. L. Weber, L. M. Richards, D. J. Fox, and A. K. Dunn, "Laser speckle contrast imaging of cerebral blood flow in humans during neurosurgery: a pilot clinical study," *Journal of biomedical optics*, vol. 15, no. 6, p. 066030, 2010.

[110] W. Heeman, W. Steenbergen, G. M. van Dam, and E. C. Boerma, "Clinical applications of laser speckle contrast imaging: a review," *Journal of biomedical optics*, vol. 24, no. 8, p. 080901, 2019.

[111] D. Briers, D. D. Duncan, E. R. Hirst, S. J. Kirkpatrick, M. Larsson, W. Steenbergen, T. Stromberg, and O. B. Thompson, "Laser speckle contrast imaging: theoretical and practical limitations," *Journal of biomedical optics*, vol. 18, no. 6, p. 066018, 2013.

[112] W. J. Warren, E. A. Moro, M. E. Briggs, and E. B. Flynn, "Simulating translation-induced laser speckle dynamics in photon doppler velocimetry," *Applied Optics*, vol. 53, no. 21, pp. 4661–4668, 2014.

[113] C. Sheppard and M. Hrynevych, "Diffraction by a circular aperture: a generalization of fresnel diffraction theory," *JOSA A*, vol. 9, no. 2, pp. 274–281, 1992.

[114] N. Takai, T. Iwai, and T. Asakura, "Correlation distance of dynamic speckles," *Applied Optics*, vol. 22, no. 1, pp. 170–177, 1983.

[115] T. H. G. Megson, *Structural and stress analysis*. Butterworth-Heinemann, 2019.

[116] H. Depot, "Plywood – columbia forest products," 2022, last accessed 24 July 2022. [Online]. Available: https://www.homedepot.com/p/Columbia-Forest-Products-1-2-in-x-2-ft-x-2-ft-PureBond-Red-Oak-Plywood-Project-Panel-Free-Cu 204771237

[117] M. Supermarkets, "Metal – metal supermarkets," 2022, last accessed 24 July 2022. [Online]. Available: https://www.metalsupermarkets.com/product/aluminum-sheet-5052/

[118] S. M. Plastics, "Plastic – santa monica plastics," 2022, last accessed 24 July 2022. [Online]. Available: https://santamonicaplastics.com/shop/acrylic-sheets-more-cut-to-size/cut-to-size-clear-white/acrylic-sheets-cut-to-size-opaque-white-7328/

[119] R. Xiao, S. Hudson, and C. Harrison, "Direct: Making touch tracking on ordinary surfaces practical with hybrid depth-infrared sensing," in *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*, 2016, pp. 85–94.

[120] ——, "Supporting responsive cohabitation between virtual interfaces and physical objects on everyday surfaces," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. EICS, pp. 1–17, 2017.

[121] R. Xiao, C. Harrison, and S. E. Hudson, "Worldkit: rapid and easy creation of ad-hoc interactive applications on everyday surfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 879–888.

[122] Y. Zhang, W. Kienzle, Y. Ma, S. S. Ng, H. Benko, and C. Harrison, "Actitouch: Robust touch detection for on-skin ar/vr interfaces," in *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 2019, pp. 1151–1159.

[123] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 2019, 2019.

[124] K. Willis, E. Brockmeyer, S. Hudson, and I. Poupyrev, "Printed optics: 3d printing of embedded optical elements for interactive devices," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 2012, pp. 589–598.

[125] M. Schmitz, M. Khalilbeigi, M. Balwierz, R. Lissermann, M. Mühlhäuser, and J. Steimle, "Capricate: A fabrication pipeline to design and 3d print capacitive touch sensors for interactive objects," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 2015, pp. 253–258.

[126] V. Savage, C. Chang, and B. Hartmann, "Sauron: embedded single-camera sensing of printed physical user interfaces," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013, pp. 447–456.

[127] R. Henderson and K. Schulmeister, *Laser safety*. CRC Press, 2003.

[128] V. Iyer, E. Bayati, R. Nandakumar, A. Majumdar, and S. Gollakota, "Charging a smartphone across a room using lasers," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–21, 2018.

[129] J. R. Estepp, E. B. Blackford, and C. M. Meier, "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014, pp. 1462–1469.

[130] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.

[131] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019.

[132] R. Meziatisabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, 2021.

[133] 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*. Springer, 2018, pp. 562–576.

[134] A. Dasari, S. K. A. Prakash, L. A. Jeni, and C. S. Tucker, "Evaluation of biases in remote photoplethysmography methods," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[135] 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. 10 762–10 774, 2010.

[136] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "Distanceppg: Robust non-contact vital signs monitoring using a camera," *Biomedical optics express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[137] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *arXiv preprint arXiv:2006.03790*, 2020.

[138] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 295–310.

[139] H. Lu, H. Han, and S. K. Zhou, "Dual-gan: Joint bvp and noise modeling for remote physiological measurement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 404–12 413.

[140] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[141] D. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2593–2601, 2014.

[142] J. Zou and L. Schiebinger, "Ensuring that biomedical ai benefits diverse populations," *EBioMedicine*, p. 103358, 2021.

[143] A. Kadambi, "Achieving fairness in medical devices," *Science*, vol. 372, no. 6537, pp. 30–31, 2021.

[144] D. McDuff, J. Hernandez, E. Wood, X. Liu, and T. Baltrusaitis, "Advancing non-contact vital sign measurement using synthetic avatars," *arXiv preprint arXiv:2010.12949*, 2020.

[145] Y.-Y. Tsou, Y.-A. Lee, and C.-T. Hsu, "Multi-task learning for simultaneous video generation and remote photoplethysmography estimation," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[146] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "Combining magnification and measurement for non-contact cardiac monitoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, pp. 3810–3819.

[147] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[148] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *IEEE international conference on advanced video and signal based surveillance*, 2009, pp. 296–301.

[149] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment maps," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 497–500.

[150] S. Alotaibi and W. Smith, "Biofacenet: Deep biophysical face image interpretation," in *British Machine Vision Conference (BMVC)*, 2019.

[151] S. Alotaibi and W. A. Smith, "Decomposing multispectral face images into diffuse and specular shading and biophysical parameters," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3138–3142.

[152] S. J. Preece and E. Claridge, "Spectral filter optimization for the recovery of parameters which describe human skin," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 7, pp. 913–922, 2004.

[153] T. Igarashi, K. Nishino, and S. K. Nayar, *The appearance of human skin: A survey*. Now Publishers Inc, 2007.

[154] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 692–702.

[155] M. A. Pimentel, A. E. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2016.

[156] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[157] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.

[158] E. Nowara, D. McDuff, and A. Veeraraghavan, "The benefit of distraction: Denoising remote vitals measurements using inverse attention," *arXiv preprint arXiv:2010.07770*, 2020.

[159] D. McDuff and E. Blackford, "iphys: An open non-contact imaging-based physiological measurement toolbox," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 6521–6524.

[160] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018.

[161] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020.

[162] H. J. Ryu, H. Adam, and M. Mitchell, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," *arXiv preprint arXiv:1712.00193*, 2017.

[163] Y. Li and N. Vasconcelos, "Repair: Removing representation bias by dataset resampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9572–9581.

[164] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[165] E. S. Jo and T. Gebru, "Lessons from archives: Strategies for collecting sociocultural data in machine learning," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 306–316.

[166] Z. Gong, P. Zhong, and W. Hu, "Diversity in machine learning," *IEEE Access*, vol. 7, pp. 64 323–64 350, 2019.

[167] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558.

[168] P. Golle, "Machine learning attacks against the asirra captcha," in *Proceedings of the 15th ACM conference on Computer and communications security*, 2008, pp. 535–542.

[169] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[170] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4334–4343.

[171] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[172] J. Choi, I. Elezi, H.-J. Lee, C. Farabet, and J. M. Alvarez, "Active learning for deep object detection via probabilistic modeling," *arXiv preprint arXiv:2103.16130*, 2021.

[173] S. Dasgupta, "Two faces of active learning," *Theoretical computer science*, vol. 412, no. 19, pp. 1767–1781, 2011.

[174] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 969–977.

[175] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2100–2110.

[176] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6891–6902.

[177] M. Gwilliam, S. Hegde, L. Tinubu, and A. Hanson, "Rethinking common assumptions to mitigate racial bias in face recognition datasets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4123–4132.

[178] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 771–787.

[179] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona, "Towards causal benchmarking of biasin face analysis algorithms," in *Deep Learning-Based Face Analytics*. Springer, 2021, pp. 327–359.

[180] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[181] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5310–5319.

[182] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[183] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning under covariate shift." *Journal of Machine Learning Research*, vol. 10, no. 9, 2009.

[184] E. Tartaglione, C. A. Barbano, and M. Grangetto, "End: Entangling and disentangling deep representations for bias correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 508–13 517.

[185] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[186] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9301–9310.

[187] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 492–11 501.

[188] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, "A survey on domain adaptation theory: learning bounds and theoretical guarantees," *arXiv preprint arXiv:2004.11829*, 2020.

[189] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, "Analysis of representations for domain adaptation," *Advances in Neural Information Processing Systems*, vol. 19, p. 137, 2007.

[190] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[191] Z. Wang, Y. Ba, P. Chari, O. D. Bozkurt, G. Brown, P. Patwa, N. Vaddi, L. Jalilian, and A. Kadambi, "Synthetic generation of face videos with plethysmograph physiology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 587–20 596.

[192] P. Chari, K. Kabra, D. Karinca, S. Lahiri, D. Srivastava, K. Kulkarni, T. Chen, M. Cannesson, L. Jalilian, and A. Kadambi, "Diverse r-ppg: Camera-based heart rate estimation for diverse subject skin-tones and scenes," *arXiv preprint arXiv:2010.12769*, 2020.

[193] M. Mohri and A. Rostamizadeh, "Perceptron mistake bounds," *arXiv preprint arXiv:1305.0208*, 2013.

[194] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: active learning in imbalanced data classification," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 127–136.

[195] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *Advances in Neural Information Processing Systems*, vol. 23, pp. 892–900, 2010.

[196] B. Settles, "Active learning literature survey," 2009.

[197] J. Kremer, K. Steenstrup Pedersen, and C. Igel, "Active learning with support vector machines," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 4, pp. 313–326, 2014.

[198] S. Dasgupta, D. J. Hsu, and C. Monteleoni, *A general agnostic active learning algorithm.* Citeseer, 2007.

[199] A. Beygelzimer, D. J. Hsu, J. Langford, and T. Zhang, "Agnostic active learning without constraints," *Advances in Neural Information Processing Systems*, vol. 23, pp. 199–207, 2010.

[200] M.-F. Balcan, A. Broder, and T. Zhang, "Margin based active learning," in *International Conference on Computational Learning Theory.* Springer, 2007, pp. 35–50.

[201] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[202] T. Liu, G. Vietri, and S. Z. Wu, "Iterative methods for private synthetic data: Unifying framework and new methods," *Advances in Neural Information Processing Systems*, vol. 34, pp. 690–702, 2021.

[203] S. d'Ascoli, M. Gabrié, L. Sagun, and G. Biroli, "On the interplay between data structure and loss function in classification problems," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8506–8517, 2021.

[204] J. N. Martel, D. B. Lindell, C. Z. Lin, E. R. Chan, M. Monteiro, and G. Wetzstein, "Acorn: Adaptive coordinate networks for neural scene representation," *arXiv preprint arXiv:2105.02788*, 2021.

[205] D. B. Lindell, D. Van Veen, J. J. Park, and G. Wetzstein, "Bacon: Band-limited coordinate networks for multiscale scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 252–16 262.

[206] H. Peters, Y. Ba, and A. Kadambi, "pcon: Polarimetric coordinate networks for neural scene representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[207] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, T. Funkhouser *et al.*, "Local implicit grid representations for 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6001–6010.

[208] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 838–15 847.

[209] Z. Wang, S. Zhou, J. J. Park, D. Paschalidou, S. You, G. Wetzstein, L. Guibas, and A. Kadambi, "Alto: Alternating latent topologies for implicit 3d reconstruction," *arXiv preprint arXiv:2212.04096*, 2022.

[210] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[211] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[212] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron, "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*, 2022, pp. 16 190–16 199.

[213] R. Li, M. Tancik, and A. Kanazawa, "Nerfacc: A general nerf acceleration toolbox," *arXiv preprint arXiv:2210.04847*, 2022.

[214] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, "Nerv: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 557–21 568, 2021.

[215] Z. Chen, Y. Chen, J. Liu, X. Xu, V. Goel, Z. Wang, H. Shi, and X. Wang, "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*, 2022, pp. 2047–2057.

[216] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.

[217] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "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*, 2021, pp. 6498–6508.

[218] C. Gao, A. Saraf, J. Kopf, and J.-B. Huang, "Dynamic view synthesis from dynamic monocular video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5712–5721.

[219] I. Mehta, M. Gharbi, C. Barnes, E. Shechtman, R. Ramamoorthi, and M. Chandraker, "Modulated periodic activations for generalizable local functional representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 214–14 223.

[220] E. Q. Zhao, A. Vilesov, S. Athreya, P. Chari, J. Merlos, K. Millett, N. S. Cyr, L. Jalilian, and A. Kadambi, "Making thermal imaging more equitable and accurate: resolving solar loading biases," *arXiv preprint arXiv:2304.08832*, 2023.

[221] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 506–523.

[222] P. Chari, Y. Ba, S. Athreya, and A. Kadambi, "Mime: Minority inclusion for majority group enhancement of ai performance," in *European conference on computer vision*. Springer, 2022, pp. 326–343.

[223] H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz, "Optimal uncertainty quantification," *Siam Review*, vol. 55, no. 2, pp. 271–345, 2013.

[224] P. Schulz, L. Scheuvens, and G. Fettweis, "A new perspective on maximal-ratio combining," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2023, pp. 1–7.

[225] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-based video motion processing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, pp. 1–10, 2013.

[226] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[227] M. Teschner, B. Heidelberger, M. Müller, D. Pomerantes, and M. H. Gross, "Optimized spatial hashing for collision detection of deformable objects." in *Vmv*, vol. 3, 2003, pp. 47–54.

[228] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, pp. 1–11, 2013.

[229] X. Liu, X. Zhang, G. Narayanswamy, Y. Zhang, Y. Wang, S. Patel, and D. McDuff, "Deep physiological sensing toolbox," *arXiv preprint arXiv:2210.00716*, 2022.

[230] A. K. Maity, J. Wang, A. Sabharwal, and S. K. Nayar, "Robustppg: camera-based robust heart rate estimation using motion cancellation," *Biomedical Optics Express*, vol. 13, no. 10, pp. 5447–5467, 2022.

[231] E. M. Nowara, A. Sabharwal, and A. Veeraraghavan, "Ppgsecure: Biometric presentation attack detection using photopletysmograms," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 56–62.

[232] A. Al Masri and S. K. Jasra, "The forensic biometric analysis of emotions from facial expressions, and physiological processes from the heart and skin," *Journal of Emerging Forensic Sciences Research*, vol. 1, no. 1, pp. 61–77, 2016.

[233] K. Del Regno, A. Vilesov, A. Armouti, A. B. Harish, S. E. Can, A. Kita, and A. Kadambi, "Thermal imaging and radar for remote sleep monitoring of breathing and apnea," *arXiv preprint arXiv:2407.11936*, 2024.

[234] Z. Chen, T. Zheng, C. Cai, and J. Luo, "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*, 2021, pp. 392–405.

[235] Z. Fan, P. Wang, Y. Jiang, X. Gong, D. Xu, and Z. Wang, "Nerf-sos: Any-view self-supervised object segmentation on complex scenes," *arXiv preprint arXiv:2209.08776*, 2022.

[236] V. Tschernezki, I. Laina, D. Larlus, and A. Vedaldi, "Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representations," in *3DV*, 2022.

[237] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (ToG)*, vol. 42, no. 4, pp. 1–14, 2023.

[238] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 523–540.

[239] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470.

[240] Z. Wang, S. Zhou, J. J. Park, D. Paschalidou, S. You, G. Wetzstein, L. Guibas, and A. Kadambi, "Alto: Alternating latent topologies for implicit 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 259–270.

[241] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," *ICCV*, 2021.

[242] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Zip-nerf: Anti-aliased grid-based neural radiance fields," *ICCV*, 2023.

[243] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "Ibrnet: Learning multi-view image-based rendering," in *CVPR*, 2021.

[244] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 124–14 133.

[245] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelNeRF: Neural radiance fields from one or few images," in *CVPR*, 2021.

[246] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis *et al.*, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 123–16 133.

[247] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision*. Springer, 2022, pp. 333–350.

[248] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 479–12 488.

[249] B. Yi, W. Zeng, S. Buchanan, and Y. Ma, "Canonical factors for hybrid neural fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3414–3426.

[250] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8248–8258.

[251] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann, "Point-nerf: Point-based neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5438–5448.

[252] S. Lionar, X. Xu, M. Lin, and G. H. Lee, "Nu-mcc: Multiview compressive coding with neighborhood decoder and repulsive udf," *arXiv preprint arXiv:2307.09112*, 2023.

[253] S. Zhi, T. Laidlow, S. Leutenegger, and A. Davison, "In-place scene labelling and understanding with implicit scene representation," in *ICCV*, 2021.

[254] Y. Siddiqui, L. Porzi, S. R. Buló, N. Müller, M. Nießner, A. Dai, and P. Kontschieder, "Panoptic lifting for 3d scene understanding with neural fields," *arXiv preprint arXiv:2212.09802*, 2022.

[255] Z. Ren, A. Agarwala†, B. Russell†, A. G. Schwing†, and O. Wang†, "Neural volumetric object selection," in *CVPR*, 2022, († alphabetic ordering).

[256] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[257] K. Mazur, E. Sucar, and A. J. Davison, "Feature-realistic neural fusion for real-time, open set scene understanding," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8201–8207.

[258] N. Tsagkas, O. Mac Aodha, and C. X. Lu, "Vl-fields: Towards language-grounded neural implicit spatial representations," in *2023 IEEE International Conference on Robotics and Automation*. IEEE, 2023.

[259] R. Goel, D. Sirikonda, S. Saini, and P. Narayanan, "Interactive segmentation of radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4201–4211.

[260] K. Liu, F. Zhan, J. Zhang, M. Xu, Y. Yu, A. El Saddik, C. Theobalt, E. Xing, and S. Lu, "Weakly supervised 3d open-vocabulary segmentation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[261] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "Clip-fields: Weakly supervised semantic fields for robotic memory," *arXiv preprint arXiv:2210.05663*, 2022.

[262] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot language-guided manipulation," in *7th Annual Conference on Robot Learning*, 2023.

[263] J. Ye, N. Wang, and X. Wang, "Featurenerf: Learning generalizable nerfs by distilling foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8962–8973.

[264] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[265] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," 2022.

[266] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[267] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, "Context-aware feature generation for zero-shot semantic segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1921–1929.

[268] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[269] M. Zwicker, H. Pfister, J. Van Baar, and M. Gross, "Ewa volume splatting," in *Proceedings Visualization, 2001. VIS'01.* IEEE, 2001, pp. 29–538.

[270] G. Kopanas, J. Philip, T. Leimkühler, and G. Drettakis, "Point-based neural rendering with per-view optimization," in *Computer Graphics Forum*, vol. 40, no. 4. Wiley Online Library, 2021, pp. 29–43.

[271] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[272] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[273] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning.* PMLR, 2021, pp. 8748–8763.

[274] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[275] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[276] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, "Fss-1000: A 1000-class dataset for few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2869–2878.

[277] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[278] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[279] A. Bhandari, A. Kadambi, and R. Raskar, *Computational Imaging*. MIT Press, 2022.

[280] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 5008–5017.

[281] T. Müller, "tiny-cuda-nn," 4 2021. [Online]. Available: https://github.com/NVlabs/tiny-cuda-nn

[282] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[283] Z. Sun and X. Li, "Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.