Title
On Hybrid Methods that Blend Computer Vision and Physics

Author
Ba, Yunhao

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

On Hybrid Methods that Blend Computer Vision and Physics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Yunhao Ba

2023

ABSTRACT OF THE DISSERTATION

On Hybrid Methods that Blend Computer Vision and Physics

by

Yunhao Ba
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2023
Professor Achuta Kadambi, Chair

Deep learning has exhibited remarkable performance on various computer vision tasks. However, these models usually suffer from the generalization issue when the training sets are not sufficiently large or diverse. Human intelligence, on the other hand, is capable of learning with a few samples. One of the potential reasons for this is that we use other prior knowledge to generalize to new environments and unseen data, as opposed to learning everything from the provided training sets. We aim to enable machines with such capability. More specifically, we focus on integrating different types of prior physical knowledge and inductive biases into neural networks for various computer vision applications.

The core idea is to exploit physical models as inductive biases and design specific strategies to blend them with the neural network learning process. This problem is difficult since we need to consider both the fidelity of our prior knowledge and the quality of the training samples. To validate the effectiveness of the proposed blending strategies, extensive experiments have been conducted on multiple computer vision tasks, such as Shape from Polarization (SfP), remote photoplethysmography (rPPG), and single-image rain removal.

The dissertation of Yunhao Ba is approved.

Stefano Soatto

Cho-Jui Hsieh

Bolei Zhou

Laleh Jalilian

Achuta Kadambi, Committee Chair

University of California, Los Angeles

2023

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation is based on the accumulative work during my Ph.D. journey, and it would not have been possible without the support of my advisor, colleagues, friends, and family. My first thanks go to my advisor, Prof. Achuta Kadambi, for his invaluable guidance on how to conduct research on computational imaging and computer vision in general. The research discussions have always been inspiring, and it was a great fortune to be one of his first few students at UCLA.

I would also like to express my sincere gratitude to my committee members, Prof. Stefano Soatto, Prof. Cho-Jui Hsieh, Prof. Bolei Zhou, Prof. Laleh Jalilian, and Prof. Achuta Kadambi for their constructive feedback and time to serve on my committee. I would like to offer my special thanks to Prof. Soatto for his kind guidance on not just research but also lessons beyond pure academics.

For the past few years, I have enjoyed a lot for being a part of the UCLA Visual Machines Group (VMG). It has been such a privilege to work with so many talented and motivated VMG members. I would like to thank them for their kind support in both research and daily life. I also appreciate the timely support from the Office of Graduate Student Affairs at UCLA ECE Department since I was admitted into the program.

I am deeply grateful to all my co-authors: Achuta Kadambi, Akash Deep Singh, Akira Suzuki, Alex Gilbert, Alex Wong, Ankur Sarker, Arnold Pfahnl, Blake Gella, Boxin Shi, Celso M. de Melo, Chethan Chinder Chandrappa, Chinmay Talegaonkar, Cho-Jui Hsieh, Ethan Yang, Franklin Wang, Gianna Brown, Guangyuan Zhao, Henry Peters, Howard Zhang, Huan Zhang, Jinfa Yang, Kerim Doruk Karinca, Laleh Jalilian, Lei Yan, Mani Srivastava, Niranjan Vaddi, Oyku Deniz Bozkurt, Parth Patwa, Pradyumna Chari, Rishi Upadhyay, Rui Chen, Shijie Zhou, Shreeram Athreya, Sophia Chen, Stefano Soatto, Suya You, Varan Mehra, Yihan Wang, Yiqin Wang, and Zhen Wang. I thank them for their commitment, inspiration, helpful discussions, and solid contributions to the work. I would like to extend

my sincere thanks to Prof. Boxin Shi and Prof. Alex Wong for their insightful ideas and timely help during the collaboration. I also want to thank Prof. Wan-Chi Siu, my advisor at the Hong Kong Polytechnic University, who offered me research opportunities during my undergraduate study.

Finally, I want to express my deepest gratitude to my parents, for their unconditional support and love.

VITA

2017	Bachelor of Engineering (Honours) Electronic and Information Engineering The Hong Kong Polytechnic University
2019	Master of Science Electrical and Computer Engineering University of California, Los Angeles
2018 – 2022	Graduate Student Researcher Electrical and Computer Engineering University of California, Los Angeles

SELECTED PUBLICATIONS

Y. Ba†, H. Zhang†, E. Yang, A. Suzuki, A. Pfahnl, C. C. Chandrappa, C. M. de Melo., S. You, S. Soatto, A. Wong, and A. Kadambi. *Not Just Streaks: Towards Ground Truth for Single Image Deraining.* In European Conference on Computer Vision (ECCV). October 2022.

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 (ECCV). October 2022.

Y. Ba†, Z. Wang†, K. D. Karinca, O. D. Bozkurt, and A. Kadambi. *Style Transfer with Bio-realistic Appearance Manipulation for Skin-tone Inclusive rPPG.* In 2022 IEEE International

Conference on Computational Photography (ICCP). August 2022.

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). June 2022.

Y. Ba, A. Gilbert[†], F. Wang[†], J. Yang, R. Chen, Y. Wang, L. Yan, B. Shi, and A. Kadambi. *Deep shape from polarization.* In European Conference on Computer Vision (ECCV). August 2020.

Y. Ba[†], G. Zhao[†], and A. Kadambi. *Blending diverse physical priors with neural networks.* Preprint. October 2019.

[†] Equal contribution.

CHAPTER 1

Introduction

Human intelligence is capable of using learned knowledge for fast adaptation. By observing just a few samples, we can effortlessly conduct inference in a new scene for a new task by reutilizing what we have known previously. In contrast, the current leading deep learning models often have to learn everything from scratch, which makes them notoriously data-hungry. Unfortunately, a large and diverse dataset may not be available for many real-world applications, since data acquisition is usually an expensive and time-consuming process. Without sufficiently large datasets, these deep learning models may fail to generalize their performance on unseen samples and produce results that are inconsistent with real-world physical constraints.

Meanwhile, there exist vast amounts of already available physical models for many scientific problems. Due to the existence of physical model mismatches, solutions derived from these physical models may not be as accurate as the data-driven solutions when the testing distribution is sufficiently close to the training distribution. However, these physical models are able to provide robust solutions across various environmental conditions that are not fully covered by the data acquisition process.

Given that neither physics-based solutions nor purely data-driven models are considered sufficient to provide accurate and robust estimation for some complicated real-world applications, the community has started to explore the concept of physics-based learning (PBL), where these two types of approaches are integrated together in a synergistic and complemen-

Chapter	Content
Chapter 2	Introduction of physics-based learning (PBL) and AutoPhysics
Chapter 3	Blending physics and learning for Shape from Polarization (SfP)
Chapters 4 and 5	Blending physics and learning for remote photoplethysmography (rPPG)
Chapter 6	Blending physics and learning for single-image rain removal

Table 1.1: **Dissertation overview.**

tary manner.[1] This dissertation discusses how to blend some existing physical knowledge into the deep learning process. We successfully demonstrate that the generalizability of neural networks can be significantly improved by introducing additional prior knowledge for various tasks in the field of computational imaging and computer vision.

1.1 Overview

As illustrated in Tab. 1.1, there are four main contributions in this dissertation. We start with an introduction to PBL and propose an automated framework (AutoPhysics) to find suitable neural network architectures to blend physical knowledge given some training samples. We then dive into three specific applications in computational imaging and computer vision. These three applications are Shape from Polarization (SfP), remote photoplethysmography (rPPG), and single-image rain removal.

[1]Physics-based learning is also referred to as physics-guided learning, physics-informed learning, or physics-aware learning. Please see [WJX22] for a detailed list of various techniques to integrate prior knowledge with machine learning.

1.2 Summary and Contributions

In Chapter 2, we introduce the concept of *physics-based learning* (PBL) and propose an automated approach to blending various types of physical priors during the learning process. Machine learning in the context of physical systems merits a re-examination of the learning strategy. In addition to data, one can leverage a vast library of physical prior models (e.g., kinematics, fluid flow, etc) to perform more robust inference. The nascent sub-field of PBL studies the blending of neural networks with physical priors. While previous PBL algorithms have been applied successfully to specific tasks, it is hard to generalize existing PBL methods to a wide range of physics-based problems. Such generalization would require an architecture that can adapt to variations in the correctness of the physics, or in the quality of training data. Unfortunately, no such network architecture exists. In this chapter, we aim to generalize PBL, by making a first attempt to bring neural architecture search (NAS) to the realm of PBL. We introduce a new method known as automated physics-based learning (AutoPhysics) that is able to generate PBL topology with top performance across a diverse range of quality in the physical model and the dataset. This chapter revises [BZK19].

In Chapter 3, we investigate the Shape from Polarization (SfP) problem and make an attempt to bring the SfP problem to the realm of deep learning. The previous state-of-the-art methods for SfP have been purely physics-based. We see value in these principled models and blend these physical models as priors into a neural network architecture. This proposed approach achieves results that exceed the previous state-of-the-art on a challenging dataset we introduce. This dataset consists of polarization images taken over a range of object textures, paints, and lighting conditions. We report that our proposed method achieves the lowest test error on each tested condition in our dataset, showing the value of blending data-driven and physics-driven approaches. This chapter revises [BGW20].

In Chapter 4 and Chapter 5, we take remote photoplethysmography (rPPG) as an example to examine the performance bias from skin tone variations in non-contact heart rate

estimation. Accelerated by telemedicine, advances in rPPG are beginning to offer a viable path toward non-contact physiological measurement. Unfortunately, the datasets for rPPG are limited as they require videos of the human face paired with ground-truth, synchronized heart rate data from a medical-grade health monitor. Also troubling is that the datasets are not inclusive of diverse populations, i.e., current real rPPG facial video datasets are imbalanced in terms of races or skin tones, leading to accuracy disparities on different demographic groups. For example, MMSE-HR [ZGW16], AFRL [EBM14], and UBFC-RPPG [BMB19] only contain roughly 10%, 0%, and 5% of dark-skinned subjects respectively. In Chapter 4, we show a first attempt to overcome the lack of dark-skinned subjects by synthetic augmentation. A joint optimization framework is utilized to translate real videos from light-skinned subjects to dark skin tones while retaining their underlying blood volume variations. In the experiment, our method exhibits around 38% reduction in mean absolute error for the dark-skinned group and 49% improvement on bias mitigation, as compared with the previous work trained with just real samples. In Chapter 5, we propose a scalable biophysical learning-based method to generate physio-realistic synthetic rPPG videos given any reference image and target rPPG signal and shows that it could further improve the state-of-the-art physiological measurement and reduce the bias among different groups. We also collect a large-scale rPPG dataset (UCLA-rPPG) with diverse skin tones, in the hope that this dataset could serve as a benchmark in the field to ensure that advances in the technique can benefit all people for healthcare equity. Chapter 4 revises [BWK22], and Chapter 5 revises [WBC22].

In Chapter 6, we study the single-image rain removal problem and show how the performance of deraining can be improved by incorporating the knowledge of rain phenomena and human experts through a carefully designed data collection pipeline and a rain-robust loss function. We propose a large-scale dataset of real-world rainy and clean image pairs and a method to remove degradations, induced by rain streaks and rain accumulation, from the image. As there exists no real-world dataset for deraining, current state-of-the-art methods rely on synthetic data and thus are limited by the sim2real domain gap; moreover, rigorous

evaluation remains a challenge due to the absence of a real paired dataset. We fill this gap by collecting a real paired deraining dataset through meticulous control of non-rain variations. Our dataset enables paired training and quantitative evaluation for diverse real-world rain phenomena (e.g., rain streaks and rain accumulation). To learn a representation robust to rain phenomena, we propose a deep neural network that reconstructs the underlying scene by minimizing a rain-robust loss between rainy and clean images. Extensive experiments demonstrate that our model outperforms the state-of-the-art deraining methods on real rainy images under various conditions. This chapter revises [BZY22].

Finally, we discuss some potential future directions to conclude the dissertation in Chapter 7.

CHAPTER 2

Integrating Physical Priors as Inductive Biases into Deep Learning

2.1 Introduction

Advances in machine learning can transform the way physical calculations are performed and may even help with the discovery of physical equations [CTB19]. Many physical models are idealized and do not precisely match real-world data. An elementary example would be equations for projectile motion which do not account for air resistance. Using these idealized equations, a completely *physics-driven approach* would have large errors on real-world data. A separate approach is completely *data-driven*, e.g., one could repeatedly record real-world projectile tosses and use a regression model to estimate a future trajectory, or the physical expression [BSP02, FAL15]; unfortunately, this approach requires large datasets and lacks interpretability. To bridge this gap, the field of *physics-based learning* (PBL) aims to blend physical priors with data-driven inference, to combine the best of both worlds [GTH00].

Previous PBL architectures have achieved competitive performance on a wide variety of tasks in computational microscopy [SLL17, NWM18, NXL18, RWO19], low-level and high-level computer vision [GAL18, MSV18, BGW20, SLK19], medical imaging [JMF17, KMY17], and robot control [ASA18, ZSL19, ABW19, SSO19]. These seemingly diverse problem statements share a common thread: the presence of a partially known physical prior that can be blended with a neural network. Unfortunately, these existing PBL methods are typically designed for a specific task. Generalization would (as a first step) require

a PBL architecture capable of adapting to variations in the correctness of physics or the quality of training data. Our experiments show that no such architecture exists. Having a general recipe for blending physics and learning is an important step in adopting physics-based learning to encompass the wide range of physical problems, where priors are only approximate and training data can be sparse.

In this chapter, we approach the problem of PBL from a different angle. Inspired by work in neural architecture search (NAS) [ZL16, BGN16, LSY18, CZH18], we propose a first attempt to automatically find the optimal PBL topology. NAS for PBL is complicated because the search algorithm needs to learn characteristics of not just data, but also physics. This means the existing NAS framework can not be applied to PBL directly. To customize NAS for PBL problems, we find that three modifications must be made to the existing NAS framework: (1) the inclusion of physical inputs; (2) the inclusion of physical operation sets; and (3) edge weights to normalize variations in the degrees of freedom introduced by the inclusion of physical operators. As these modifications are specific to the PBL problem, we refer to our algorithm as AutoPhysics. The goal of AutoPhysics is to handle a diverse range of quality in the physical prior or data. Experiments in Sec. 2.4.3 offer support for this goal, where AutoPhysics outperforms previous PBL methods on multiple physical tasks across a range of physical priors and dataset conditions.

Our contributions to physics-based learning can be summarized as follows:

- We make a first attempt at bringing neural architecture search (NAS) into the realm of physics-based learning (PBL), which alleviates the burden of human expertise in designing high-quality PBL topology;
- We propose AutoPhysics as a novel hybrid differentiable NAS framework based on the combination of [LSY18] and [CZH18] with three customized modifications for PBL applications;
- We show in our experiments that AutoPhysics generalizes to a wider range of diversity

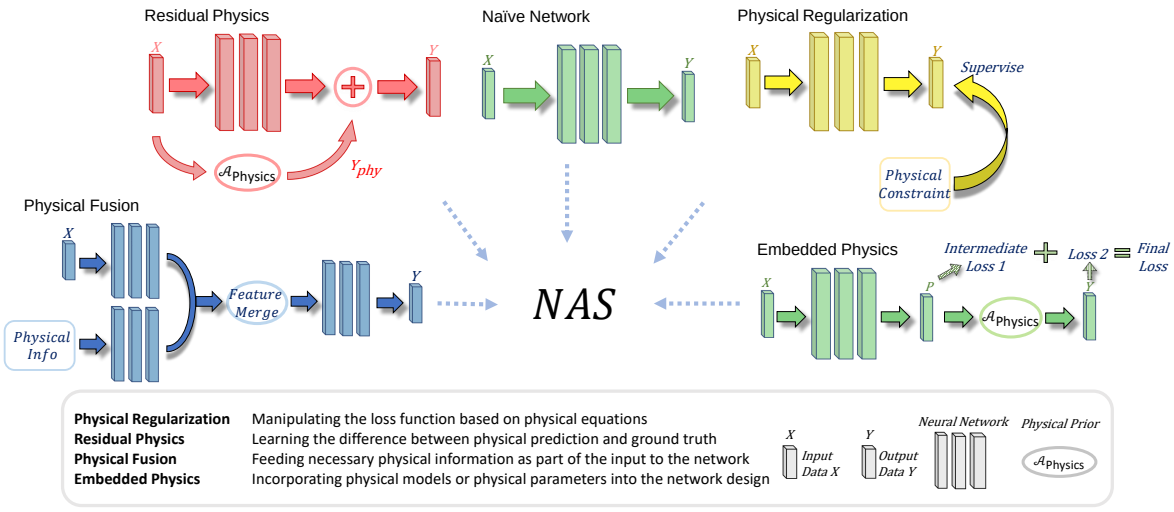


Figure 2.1: **An overview of proposed NAS-based blending approach.** Our Auto-Physics framework can take advantage of the existing methods of blending physical prior and is capable of generating new hybrid architectures for tasks under diversified physical environments.

in data and physical priors, as compared to manually designed PBL architectures.

Although our primary contributions are to PBL, it is worth noting that conventional differentiable NAS approaches [LSY18, CZH18] are not designed to incorporate physical priors; in developing this framework we found it necessary to modify such approaches to incorporate physical priors as both inputs and candidate operations.

2.2 Categorizing Prior Work in Physics-based Learning

There has been remarkable progress in blending physical priors with neural networks over the past few years. Here, we make a first attempt to group previous methods into the four categories as illustrated in Fig. 2.1:

- **Physical Fusion** feeds the solution from physics-based models as part of the in-

put [KWR17a, BGW20]. The solutions can be stacked with the original input, or additional, identical network branches can be used to extract features separately;

- **Residual Physics** is another way to improve the model-based solutions with deployments in robot control [ZSL19, ABW19] and medical imaging [JMF17, KMY17]. By adding the physical solution onto the network output, the neural networks only need to learn the mismatch between the model-based solution and the ground truth in this case;

- **Physical Regularization** harnesses the regularization term from a set of physical constraints to penalize the network solutions. The regularization term can be appended as part of the loss function explicitly [KWR17a, SE17, RPK17, Rai18, FWS19], or through a reconstruction process from physics [CLZ18, CGG18, PLD18];

- **Embedded Physics** takes the physical model inside the network optimization loop, where the physical model acts as a skeleton, and the network is in charge of learning parameters used in these models. Unrolled networks [GL10, DSH17, SDP18, KBR19, MYK19], PDE-Nets [LLM17], and variational networks [Cha16, HKK18] can all be classified into this category. During training, auxiliary intermediate losses can be inserted to guarantee the learned parameters indeed carry their corresponding physical meanings as well [HCS19, SF19, LCT19].

Continuing to propose new models for PBL is a viable direction, however, this may not address adaptability to diverse scenarios of physical model mismatch and sparsity in training data. AutoPhysics is a different tack, where we design basic operation sets inspired by PBL strategies, and allow networks to customize their architectures during training.

2.3 AutoPhysics

In what follows, we describe the AutoPhysics algorithm. In Sec. 2.3.1, we discuss the problem setup. We then describe the search algorithm in Sec. 2.3.2 and the detailed features of AutoPhysics in Sec. 2.3.3.

2.3.1 Problem Setup

In the PBL problem, we have access to a training set $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$, and a partially known physical operator $\mathcal{A}_{phy}$. Each sample within the training set is a data pair (x_i, y_i) formed by an input instance $x_i \in X$ and the corresponding output $y_i \in Y$. The objective is to learn a function $f(\cdot)$ that maps input space to output space ($X \to Y$). $f(\cdot)$ is approximated by a physics-based network from a search space $\mathcal{H}$ with hypotheses $\hat{f}(\omega, \alpha, \mathcal{A}_{phy})$, where ω denotes network parameters and α denotes architecture parameters. The learning algorithm searches inside $\mathcal{H}$ and tries to find a $\{\omega, \alpha\}$ that parameterizes the optimal $\hat{f}(\omega, \alpha, \mathcal{A}_{phy})$ for D_{train}. The challenge lies in finding a suitable method to incorporate $\mathcal{A}_{phy}$ into the network design under diverse regimes of physical model mismatch.

2.3.2 Search Algorithm

In contrast to NAS for vision tasks, we do not search the cell structures and apply these searched cells to some predefined meta-architectures, since the concatenation of multiple cells may undermine the underlying physical principles in PBL. Therefore, AutoPhysics tries to learn an architecture that directly links the network input and the target output. To achieve this goal, we develop AutoPhysics based on Differentiable ARchiTecture Search (DARTS) [LSY18] with binarized architecture parameters [CZH18]. To reduce the memory requirement and accelerate the searching process, we replace the weighted sum in DARTS with the gated sum using the binarized mask sampled from the softmax probability of architecture parameters. These binarized parameters are trained using gradient approximation as

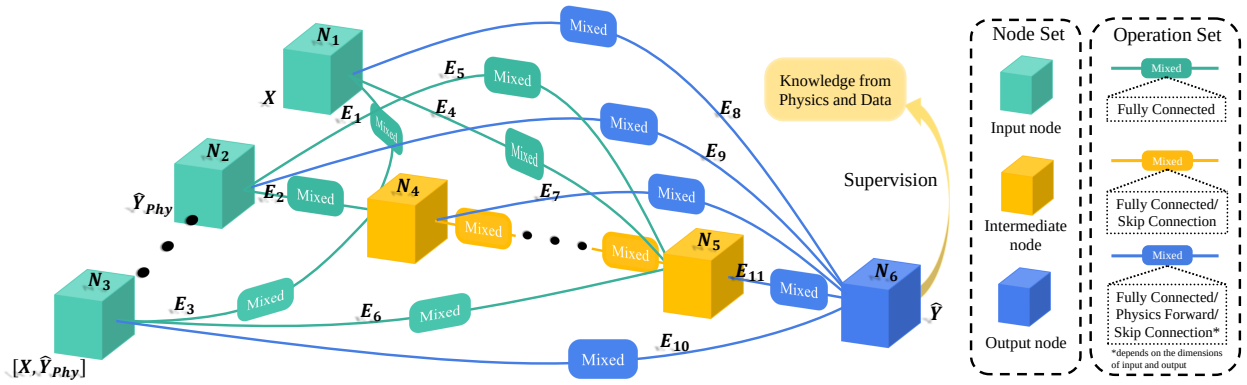


Figure 2.2: **Search space of our AutoPhysics.** In the proposed AutoPhysics, all the nodes are densely connected by mixed operators from predefined candidate operation sets. The hidden nodes can obtain information from the original inputs or from previous hidden nodes within this search setup. The training process is supervised by both ground truth and physical constraints.

described in BinaryConnect [CBD15] and ProxylessNAS [CZH18]. This binarized arrangement ensures that only the sampled paths are activated when searching the PBL topology, which reduces the gap between the architecture search stage and the architecture selection stage.

The search space of AutoPhysics is illustrated in Fig. 2.2, where the whole architecture is represented by a directed acyclic graph with nodes $\{N_i\}_{i=1}^{N}$ and edges $\{E_m\}_{m=1}^{M}$. Each edge connects two nodes (N_i, N_j) through a mixed operator, and each node corresponds to a type of input or a feature vector extracted from previous nodes through the mixed operators. The output of the mixed operator between (N_i, N_j) is the gated sum of all candidate operations $\{o_k\}_{k=1}^{K}$:

$$m_{ij}(n_i) = \sum_{k=1}^{K} g_{o_k} o_k(n_i), \tag{2.1}$$

where m_{ij} is the output of this mixed operator, n_i is the feature vector of node N_i, g_o is the binarized operation mask sampled from softmax probabilities of architecture parameters for different operations α_o, and K is the number of operations inside an edge, which depends

on the properties of node pair (N_i, N_j). The nodes are densely connected so that n_j (the output features of node N_j) is the gated sum of features from all its previous nodes:

$$n_j = \sum_{i=0}^{j-1} g_{e_i} m_{ij}(n_i) = \sum_{i=0}^{j-1} g_{e_i} \sum_{k=1}^{K} g_{o_k} o_k(n_i), \tag{2.2}$$

where g_e is the binarized edge mask sampled from the softmax probabilities of architecture parameters for different edges α_e, and N_j can either be an intermediate node or the output node.

During training, we retain two incoming edges for each node and one operation for each edge through the binary gate sampling. During the inference stage, we pick two candidate edges with the largest edge probabilities, and we select the operation with the maximum operation probability for each of the two edges. We choose two edges for each node to enable AutoPhysics to learn complicated structures, such as residual connection and multi-stream encoding. In order to learn both the network weights and the associated architecture parameters, we update these two sets of parameters alternately. In the architecture step, we freeze the network weights ω and minimize the validation loss $\mathcal{L}_{val}(\omega, \alpha)$ by updating α. In the network step, we update ω to minimize the training loss $\mathcal{L}_{train}(\omega, \alpha)$ with frozen α.

2.3.3 AutoPhysics Features

To incorporate priors into AutoPhysics, we make three customized modifications in the search process.

Physical Inputs: As a first step in blending physics into AutoPhysics, we need to prepare unique input nodes that take into account four categories of input information: (1) the data input X; (2) the duplicated data input X_{dup}, to verify whether physical information is indeed necessary since each node has to pick two edges; (3) the estimated solution from physics $\hat{Y}_{phy} = \mathcal{A}_{phy}(X)$; and (4) the concatenation of X and $\hat{Y}_{phy}$ to test at which stage to conduct the physical fusion.

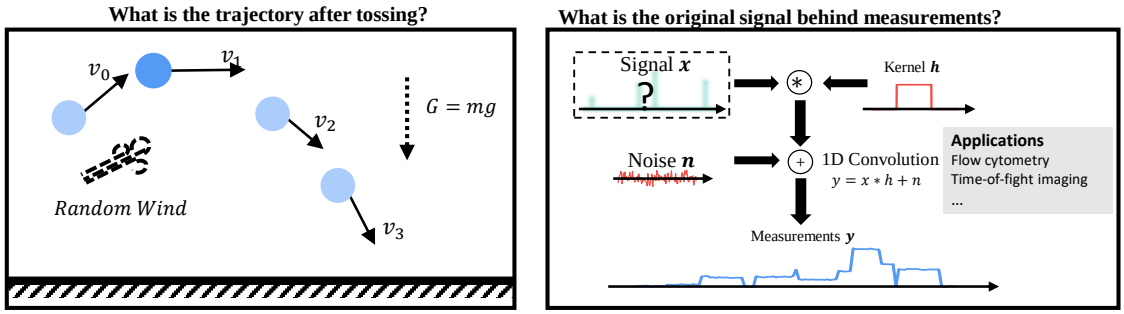


Figure 2.3: **We evaluate our method on a simulator of classical tasks.** The first task (left) is predicting the trajectory of a ball being tossed, and the second task (right) is one-dimensional deconvolution.

Physical Operations: To merge physical models inside the network, we create physics-informed operation sets $O = \{o_{NN_1}, ..., o_{NN_L}, o_{phy}\}$, where o_{NN_i} denotes the neural network operations (e.g., convolutional layer, skip connection) and o_{phy} denotes the physical forward operation. Specifically, for the physical forward module, a lightweight network layer, such as a single fully connected (FC) layer, might be useful to make the size of its input consistent with the parameter size required by the physical module.

Edge Weights: In AutoPhysics, not all edges are created with the same number of operations, since they are used to connect different types of nodes. Consequently, if we select the edges purely based on the operation probabilities, edges with fewer operations are naturally preferred due to the softmax probability, which causes a biased architecture selection. We solve this issue by introducing the edge weight as described in Eq. (2.2). After searching, we first pick a desired edge according to the edge weights and then select the desired operation for that edge based on the operation weights.

2.4 Experiments and Results

To comprehensively evaluate AutoPhysics, we simulate two representative physical tasks for which we can vary the model mismatch: (1) predicting trajectories of an object being tossed, and (2) recovering the original signal degraded by the convolution process and additive noise. Figure 2.3 illustrates these tasks; further details are provided in Sec. 2.4.1. Compared manually designed PBL architectures are described in Sec. 2.4.2. We evaluate AutoPhysics and provide a detailed analysis of the searched architectures in Sec. 2.4.3.

2.4.1 Description of Tasks

For the TOSSING TRAJECTORY PREDICTION task (see Fig. 2.3 for visualization), the initial three locations of an object with a fixed mass $\{l_1, l_2, l_3\}$ are given as input X, and our objective is to predict locations of this object in the following 15 timestamps, $\{l_4, l_5, ..., l_{18}\}$. We only consider the displacement within a 2D plane, therefore, the coordinates of each location can be represented by two numbers, i.e., $l_i = (l_{x_i}, l_{y_i})$. We adopt the following elementary free-falling equations as the prior and examine different methods under this inadequate physical prior:

$$\hat{Y}_{phy} : \begin{cases} l_{x_i} = l_{x_1} + v_x t_i \\ \\ l_{y_i} = l_{y_1} + v_y t_i - \frac{1}{2} g t_i^2 \end{cases}, \tag{2.3}$$

where l_{x_i} and l_{y_i} denote the object location at time t_i, l_{x_1} and l_{y_1} are the initial location of the object, v_x and v_y denote the initial velocities along horizontal and vertical directions respectively, and g is the fixed gravitational acceleration of $9.8 m/s^2$. We introduce two model mismatches: random acceleration from wind and an additional damping factor based on $F_{air} = k \times v^2$ to simulate air resistance. The future locations estimated by this mismatched prior are used as the physical input $\hat{Y}_{phy}$. As to the physical modules in Embedded Physics and AutoPhysics, we estimate parameters $\{\hat{l}_{x_1}, \hat{l}_{y_1}, \hat{v}_x, \hat{v}_y\}$, and substitute these parameters

into Eq. (2.3) as the physical operation. This physical module is only included in the edges that connect to the output node for this task.

We use ONE-DIMENSIONAL DECONVOLUTION task (see Fig. 2.3 for visualization) to further demonstrate the capability of AutoPhysics on real-world applications. One-dimensional deconvolution has been viewed as an important component for many sensing techniques, such as reflection seismology [Yil01], time-of-flight imaging [KWB13] and flow cytometry [AFE18]. In our examination, we aim to recover a one-dimensional signal of length 40 from its blurry observation. We generate random signal x as the ground truth and convolve it with a known point spread function (PSF) h under Gaussian noise corruption to generate the blurry observation y. We use different noise n variances to adjust the model mismatch levels. A box kernel of size 15 is selected as the PSF. For this task, the output of the Wiener filter is utilized as the physical solution $\hat{Y}_{phy}$, and we deploy the operation in the unrolled deblurring network [DSH17] as the Embedded Physics model. The unrolled operator can be expressed in the equation below:

$$x_{out} = \mathcal{F}^{-1}\left\{\frac{\overline{\mathcal{F}(h)}\mathcal{F}(y) + \gamma\mathcal{F}(x_{net})}{\overline{\mathcal{F}(h)}\mathcal{F}(h) + \gamma}\right\}, \tag{2.4}$$

where x_{out} is the output of the unrolled operation, $\mathcal{F}(\cdot)$ is the Fourier transform operator, $\mathcal{F}^{-1}(\cdot)$ is the inverse Fourier transform operator, $\overline{(\cdot)}$ is the complex conjugate operator, h is the PSF, γ is the learnable balance factor, and x_{net} is the estimated prior from a neural network. In [DSH17], there is a residual connection to the network prior, and we include this connection during comparison. In AutoPhysics, we incorporate Eq. (2.4) as the physical operation, where x_{net} is estimated from the previous nodes.

2.4.2 Manually Designed PBL Methods

For the sake of comparison, several manually designed architectures from Sec. 2.2 are also evaluated. In the TOSSING task, we use a three-layer multilayer perceptron (MLP) as the naive data-driven baseline, since it has sufficient expressiveness to fit this physical prob-

lem [Csa01]. Network structures for the Physical Regularization model and the Residual Physics model are the same as the naive model. The output of the Residual Physics model is a summation of $\hat{Y}_{phy}$ and the learned residual from the network. There is an additional regularization term in the loss function of the Physical Regularization model. Since only a partially correct physical prior is used, directly using physical solutions as the regularization will in turn aggravate the error. Thus, we introduce a ReLU-based regularization similar to [KWR17a]. The regularization loss penalizes the network solution based on the assumption that the object moves along one direction in the horizontal axis. In the Physical Fusion approach, two separate branches are utilized to extract features from X and $\hat{Y}_{phy}$ respectively, and each of them is a two-layer MLP. The extracted features will then be concatenated and fed into the output layer. The Embedded Physics model first estimates necessary parameters in Eq. (2.3) with a three-layer MLP, and then produces trajectory estimation based on the fixed physical process. All the above models are supervised by the ground truth with mean square error (MSE) loss, and the hidden dimension for FC layers is 128.

In the DECONVOLUTION task, we use convolutional neural networks (CNNs) to preserve the spatial relationship. The naive data-driven baseline is obtained with an eight-layer CNN. The architecture for the Residual Physics model and the Physical Regularization model is the same CNN as the pure data-driven baseline. We deploy a similar physical reconstruction process in [PLD18], where the network output is convolved with the PSF to reconstruct the blurry observation. We use the MSE loss between the real blurry observation and this reconstructed blurry observation as the additional physical regularization. For the Physical Fusion model, we use two identical branches to extract features from the Wiener output and the blurry observation. Each branch is an eight-layer CNN, and a convolutional layer is used to combine the extracted features at the end. As to the Embedded Physics model, we use the unrolled network in [DSH17]. Each unrolled block contains a five-layer CNN for x_{net} estimation, and the output of each unrolled block is calculated based on Eq. (2.4). We concatenate five unrolled blocks as the Embedded Physics model. Similarly, all the models

are supervised by the ground-truth signal with MSE loss. The number of hidden channels for convolution layers is 64, and the kernel size is set to be seven to guarantee the receptive field.

2.4.3 Results Analysis

Implementation Details: When searching, we split the training set into two subsets of the same size to update architecture parameters and network weights respectively. We validate the performance of the searched architectures on the architecture updating subset and pick the architecture with the best performance as the searched result. Please note that the test set is never used for architecture search or selection. After searching, we retrain the searched architectures with full training sets. We limit the number of learnable nodes in AutoPhysics to be 5 and 10 for the TOSSING and the DECONVOLUTION tasks respectively. In the TOSSING task, we deploy a single FC layer as the data-driven operation, and the skip connection operation is enabled when the input and output dimensions of the connected nodes match with each other. In the DECONVOLUTION task, we use a three-layer CNN with 64 hidden channels as the data-driven modules. To include the skip connections, we fix the number of output channels to be one for each node. ReLU is the only activation function used. For all the models described in Sec. 2.4.2, we tune their hyperparameters with three different hyperparameter sets for a fair comparison. When testing, we fixed the testing set size to 512. All the models are implemented in PyTorch [PGC17] and are trained using the Adam optimizer [KB14].

Random Search Baseline: We include the random search comparison as shown in Tab. 2.1 and Tab. 2.2. Similar to the settings in DARTS, we randomly sample 12 architectures from the search space and train them separately with the subset for the network weights updating. We then select the architecture with the lowest validation error and retrain it using the full training set.

Mismatch Level	**Low**		**High**	
Sample Amount	32	128	32	128
Naive Network	0.768	0.250	0.594	0.345
Physical Fusion	**0.295**	**0.161**	**0.332**	**0.239**
Residual Physics	0.317	0.185	0.564	0.279
Embedded Physics	0.801	0.213	0.718	0.446
Physics Reg.	0.564	0.252	0.688	0.306
Physics Only ($\hat{Y}_{phy}$)	0.613	0.613	1.002	1.002
AutoPhysics	**0.218**	**0.102**	**0.274**	**0.152**
Random Search	0.268	0.113	0.342	0.185

Table 2.1: **Testing performance on the tossing task.** We adopt the average Euclidean distance between the ground truth and the predicted locations as the evaluation metric (a lower distance is preferred). The low mismatch level corresponds to a small random initial acceleration range $[-1m/s^2, 1m/s^2]$ and a small damping factor of 0.2. The high mismatch level corresponds to a large acceleration range $[-3m/s^2, 3m/s^2]$ and a large damping factor of 0.5. The best model is marked in **red** and the sub-optimal manually designed PBL model is marked in **blue**.

Mismatch Level	**Low**		**High**	
Sample Amount	32	128	32	128
Naive Network	25.58	27.77	24.17	25.60
Physical Fusion	25.86	27.91	24.62	25.75
Residual Physics	23.07	27.96	22.79	25.35
Embedded Physics	**29.42**	**32.14**	**25.93**	**27.96**
Physics Reg.	25.69	28.04	24.26	25.69
Physics Only ($\hat{Y}_{phy}$)	10.90	10.90	9.48	9.48
AutoPhysics	**29.63**	**32.47**	**26.24**	**28.11**
Random Search	28.56	30.52	25.13	27.27

Table 2.2: **Testing performance on the deconvolution task.** We use the peak signal-to-noise ratio (PSNR) between the ground truth and the deconvolution results as the metric (a higher PSNR is preferred). The range of the signal is between zero and one. The low mismatch level corresponds to a random Gaussian noise with $\sigma^2 = 0.0004$, and the high mismatch level corresponds to a random Gaussian noise with $\sigma^2 = 0.001$. The best model is marked in **red** and the sub-optimal manually designed PBL model is marked in **blue**.

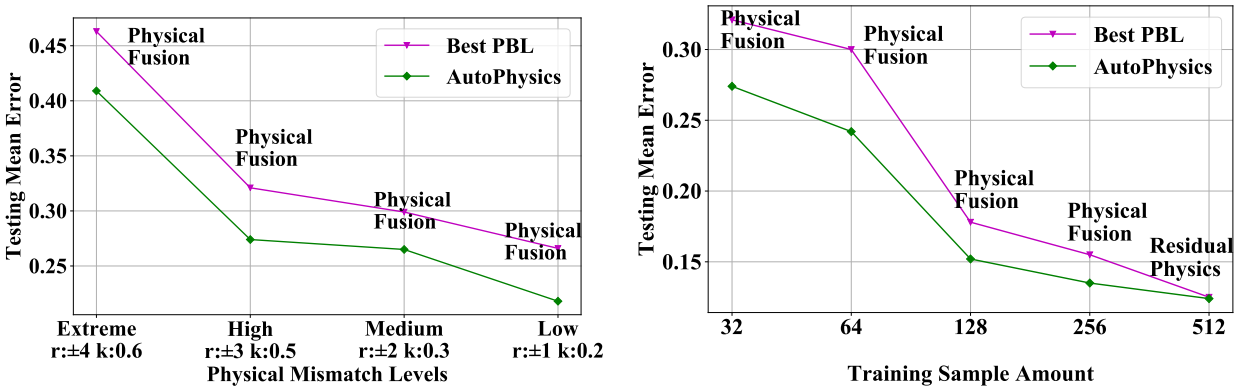
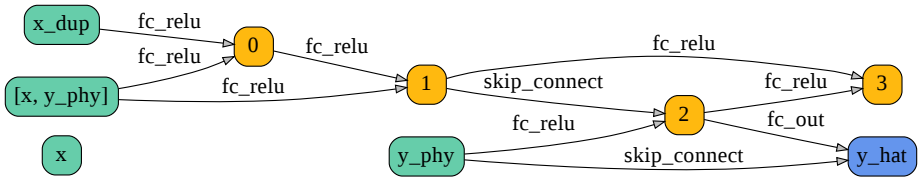
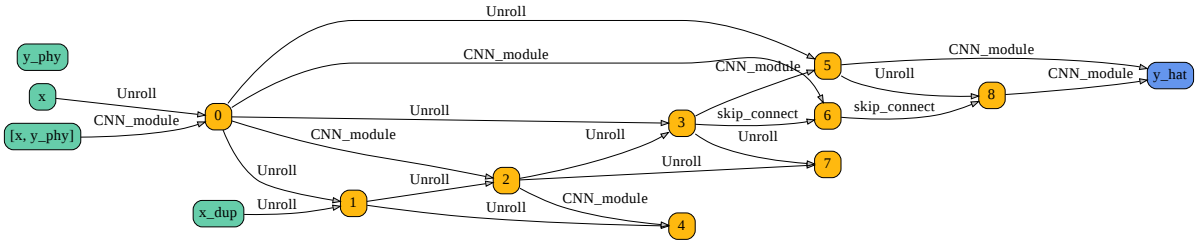


Figure 2.4: **AutoPhysics has lower errors as compared with the best PBL methods over a range of quality conditions in physics and data.** The left figure shows a comparison between the best PBL methods and AutoPhysics along different physical mismatch levels with 32 training samples. The physical mismatch levels are from extreme to low, respectively. Here $(r : \pm i,\ k : j)$ refers to the mismatch level of an initial acceleration range $[-im/s^2,\ im/s^2]$ and a damping factor j. Analogously, the right figure shows a comparison along different data amounts under a mismatch level of $(r : \pm 3,\ k : 0.5)$. The plots show the error; lower curves are preferred.

Performance Comparison: To fully evaluate the performance of the proposed approach, we vary the physical model mismatch and sparsity in a controlled manner and apply the proposed AutoPhysics to learn architectures embedded in the search space. The testing results of AutoPhysics and other existing PBL methods (as detailed in Sec. 2.4.2) are summarized in Tab. 2.1 and Tab. 2.2. As shown in these two tables, the performance of different PBL models varies based on the disparity of mismatch levels and training data sizes, while AutoPhysics is capable of generating architectures that consistently outperform these manually designed PBL models. Results in Fig. 2.4 further demonstrate this capability along the data dimension and the physics dimension in a fine-grained scale. In addition, AutoPhysics outperforms the random search baseline, validating the necessity of searching with sampled paths under limited computational resources. Our experiments also show that AutoPhysics



Tossing task, high mismatch, 128 samples



Deconvolution task, low mismatch, 128 samples

Figure 2.5: **Utilization of physics in AutoPhysics.** AutoPhysics is able to utilize physical inputs and physics-inspired operations based on the provided physical situations, which supports that AutoPhysics can learn characteristics of not just data, but also physics.

is able to perform inference on small training datasets: the physical prior reduces the demand for high-fidelity training samples. Moreover, the performance gap between AutoPhysics and the naive MLP method decreases as the number of training samples increases. This suggests that AutoPhysics is more favorable in scenarios where the number of training samples is limited.

Utilization of Physics: As illustrated in Fig. 2.5, AutoPhysics is capable of taking advantage of both physical inputs and physical operations and utilizing them properly based on the provided physical conditions. In particular, the inaccurate physical operation in the TOSSING task is not preferred, and AutoPhysics tries to refine the physical solutions using the Residual Physics strategy in the final searched architectures. For the DECONVOLUTION

task, multiple unrolled operators are utilized, since they could provide reliable estimations from physics. However, we can also observe that the unrolled operation alone is not sufficient. AutoPhysics includes additional CNN operations after unrolled operations to compensate for the errors from physics. The above results indicate that the proposed AutoPhysics framework is able to grasp the characteristics of both data and physics for PBL problems.

2.5 Conclusion

In conclusion, our experiments show that AutoPhysics can handle a wider range of physical models and data, as compared to the existing PBL methods. The main focus of our approach is to take a first attempt at increasing the diversity of PBL through architecture search. Ultimately, our hope is to apply AutoPhysics to problems as diverse as computational microscopy [BOS19], computer vision [VWG12], sensor fusion [ESS15, XAJ18] and astrophysics [BJZ16, AAA19], where it is important to handle variations in model mismatch and dataset quality across these problem domains.

CHAPTER 3

Using Polarization Physics as an Inductive Bias for Deep Surface Normal Estimation

3.1 Introduction

While deep learning has revolutionized many areas of computer vision, the deep learning revolution has not yet been studied in the context of Shape from Polarization (SfP). The SfP problem is fascinating because, if successful, shapes could be obtained in completely passive lighting conditions without estimating lighting direction. Recent progress in CMOS sensors has spawned machine vision cameras that capture the required polarization information in a single shot [Pol17], making the capture process more relaxed than photometric stereo.

This SfP problem can be stated simply: light that reflects off an object has a polarization state that corresponds to shape. In reality, the underlying physics is among the most optically complex of all computer vision problems. For this reason, previous SfP methods have high error rates (in the context of mean angular error (MAE) of surface normal estimation), and limited generalization to mixed materials and lighting conditions.

The physics of SfP are based on the Fresnel Equations. These equations lead to an underdetermined system—the so-called *ambiguity problem*. This problem arises because a linear polarizer cannot distinguish between polarized light that is rotated by π radians. This results in two confounding estimates for the azimuth angle at each pixel. Previous work in SfP has used additional information to constrain the ambiguity problem. For instance, Smith *et al.* [SRT16] use both polarization and shading constraints as linear equations when

Method	Inputs	Mean Angular Error	Robustness to Texture-Copy	Lighting Invariance
Miyazaki [MTH03]	Polarization Images	High	Strong	Moderate
Mahmoud [MEF12]	Polarization Images	High	Not Observed	Moderate
Smith [SRT18]	Polarization Images Lighting Estimate	Moderate	Strong	Moderate
Proposed	Polarization Images	Lowest	Strong	Strong

Table 3.1: **Deep SfP versus previous methods.** We compare the input constraints and the surface normal quality of the proposed hybrid of physics and learning compared to previous, physics-based SfP methods.

solving object depth, and Mahmoud *et al.* [MEF12] use shape from shading constraints to correct the ambiguities. Other authors assume surface convexity to constrain the azimuth angle [MTH03, AH06] or use a coarse depth map to constrain the ambiguity [KTS15, KTS17]. There are also additional binary ambiguities based on reflection type, as discussed in [AH06, MEF12]. Table 3.1 compares our proposed technique with prior work.

Another contributing factor to the underdetermined nature of SfP is the *refractive problem*. SfP needs knowledge of per-pixel refractive indices. Previous work has used hard-coded values to estimate the refractive index of scenes [MTH03]. This leads to a relative shape recovered with refractive distortion.

Yet another limitation of the physical model is particular susceptibility to *noise*. The polarization signal is very subtle for fronto-parallel geometries so it is important that the input images are relatively noise-free. Unfortunately, a polarizing filter reduces the captured light intensity by 50 percent, worsening the effects of Poisson shot noise, encouraging a noise tolerant SfP algorithm.[1]

In this chapter, we address these SfP pitfalls by moving away from a physics-only solution,

[1]For a detailed discussion of other sources of noise please refer to Schechner [Sch15].

toward the realm of data-driven techniques. While it is tempting to apply traditional deep learning models to the SfP problem, we find this approach does not maximize performance. Instead, we propose a physics-based learning algorithm that not only outperforms traditional deep learning but also outperforms three baseline comparisons to physics-based SfP. We summarize our contributions as follows:

- a first attempt to apply deep learning techniques to solve the SfP problem;
- incorporation of the existing physical model into the deep learning approach;
- demonstration of significant error reduction; and
- introduction of the first polarization image dataset with ground truth shape, laying a foundation for future data-driven methods.

Limitations: As a physics-based learning approach, our technique still relies on computing the physical priors for every test example. This means that the per-frame runtime would be the sum of the compute time for the forward pass and that of the physics-based prior. Future work could parallelize compute of the physical prior. Another limitation pertains to the accuracy inherent to SfP. Our average MAE on the test set is 18.5 degrees. While this is the best SfP performer on our challenging dataset, the error is higher than with a more controlled technique like photometric stereo.

3.2 Related Work

Polarization cues have been employed for various tasks, such as reflectometry estimation [GCP10], radiometric calibration [TSZ18], facial geometry reconstruction [GFT11], dynamic interferometry [MKS18], polarimetric spatially varying surface reflectance functions (SVBRDF) recovery [BJT18], and object shape acquisition [MHP07, GPD12, RRF17, ZS19]. Our approach is at the seamline of deep learning and SfP, offering unique performance tradeoffs

from prior work. Refer to Tab. 3.1 for an overview.

Shape from Polarization infers the shape (usually represented in surface normals) of a surface by observing the correlated changes of image intensity with the polarization information. Changes of polarization information could be captured by rotating a linear polarizer in front of an ordinary camera [Wol97, AE18] or polarization cameras using a single shot in real time (e.g., PolarM [Pol17] in [YTL18]). Conventional SfP decodes such information to recover the surface normal up to some ambiguity. If only images with different polarization information are available, heuristic priors such as the surface normals along the boundary and convexity of the objects are employed to remove the ambiguity [MTH03, AH06]. Photometric constraints from shape from shading [MEF12] and photometric stereo [DS01, NNT15, Atk17] complements polarization constraints to make the normal estimates unique. If multi-spectral measurements are available, the surface normal and its refractive index could be estimated at the same time [HRH10, HRH13]. More recently, a joint formulation of shape from shading and SfP in a linear manner is shown to be able to directly estimate the depth of the surface [SRT16, TSZ17, SRT18]. Our approach is the first attempt at combining deep learning and SfP.

Polarized 3D involves stronger assumptions than SfP and has different inputs and outputs. Recognizing that SfP alone is a limited technique, the Polarized 3D class of methods integrates SfP with a low-resolution depth estimate. This additional constraint allows not just recovery of shape but also a high-quality 3D model. The low-resolution depth could be achieved by employing two-view [MKI04, AH05, BVM17], three-view [CZS18], multi-view [MSB16, CGS17] stereo, or even in real time by using a SLAM system [YTL18]. These depth estimates from geometric methods are not reliable in textureless regions where finding correspondence for triangulation is difficult. Polarimetric cues could be jointly used to improve such unreliable depth estimates to obtain a more complete shape estimation. A depth sensor such as the Kinect can also provide coarse depth prior to disambiguate the ambiguous normal estimates given by SfP [KTS15, KTS17]. The key step that characterizes Polarized

3D is a holistic approach that rethinks both SfP and the depth-normal fusion process. The main limitation of Polarized 3D is the strong requirement of a coarse depth map, which is not true for our proposed technique.

Data-driven computational imaging approaches draw much attention in recent years thanks to the powerful modeling ability of deep neural networks. Various types of convolutional neural networks (CNNs) are designed to enable 3D imaging for many types of sensors and measurements. From single photon sensor measurements, a multi-scale denoising and upsampling CNN is proposed to refine depth estimates [LOW18]. CNNs also show advantages in solving phase unwrapping, multipath interference, and denoising jointly from raw time-of-flight measurements [MHM17, SHW18]. From multi-directional lighting measurements, a fully-connected network is proposed to solve photometric stereo for general reflectance with a pre-defined set of light directions [SSS17]. Then the fully convolutional network with an order-agnostic max-pooling operation [CHW18] and the observation map invariant to the number and permutation of the images [Ike18] are concurrently proposed to deal with an arbitrary set of light directions. Normal estimates from photometric stereo can also be learned in an unsupervised manner by minimizing reconstruction loss [TM18]. Other than 3D imaging, deep learning has helped solve several inverse problems in the field of computational imaging [STG17, TSR18, TSS18, LCL19]. Separation of shape, reflectance, and illuminance maps for wild facial images can be achieved with the CNNs as well [SKC18]. CNNs also exhibit potential for modeling SVBRDF of a near-planar surface [LDP17, YLD18, LSC18, DAD18], and more complex objects [LXR18]. The challenge with existing deep learning frameworks is that they do not leverage the unique physics of polarization.

3.3 Proposed Method

In this section, we first introduce basic knowledge of SfP and then present our physics-based CNN. Blending physics and deep learning improves the performance and generalizability of the method.

3.3.1 Physical Solution

Our objective is to reconstruct surface normals $\hat{\boldsymbol{N}}$ from a set of polarization images $\{\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, ..., \boldsymbol{I}_{\phi_M}\}$ with different polarization angles. For a specific polarization angle ϕ_{pol}, the intensity at a pixel of a captured image follows a sinusoidal variation under unpolarized illumination:

$$I(\phi_{pol}) = \frac{I_{max} + I_{min}}{2} + \frac{I_{max} - I_{min}}{2} \cos(2(\phi_{pol} - \phi)), \tag{3.1}$$

where ϕ denotes the phase angle, and I_{min} and I_{max} are lower and upper bounds for the observed intensity. Equation (3.1) has a π-***ambiguity*** in the context of ϕ: two phase angles, with a π shift, will result in the same intensity in the captured images. Based on the phase angle ϕ, the azimuth angle φ can be retrieved with $\frac{\pi}{2}$-***ambiguity*** as follows [CGS17]:

$$\phi = \begin{cases} \varphi, & \text{if diffuse reflection dominates} \\ \varphi - \frac{\pi}{2}, & \text{if specular reflection dominates} \end{cases}. \tag{3.2}$$

The zenith angle θ is related to the degree of polarization ρ, which can be written as:

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}. \tag{3.3}$$

When diffuse reflection is dominant, the degree of polarization can be expressed with the zenith angle θ and the refractive index n as follows [AH06]:

$$\rho_d = \frac{(n - \frac{1}{n})^2 \sin^2\theta}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2\theta + 4\cos\theta\sqrt{n^2 - \sin^2\theta}}. \tag{3.4}$$

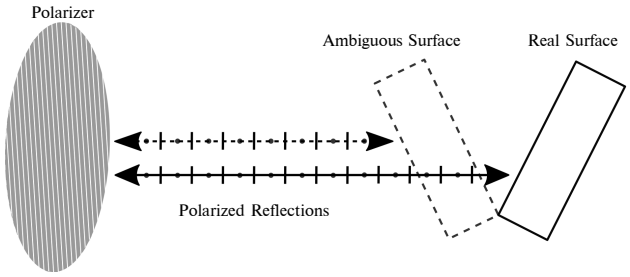


Figure 3.1: **SfP is underdetermined and one causal factor is the *ambiguity problem.*** Here, two different surface orientations could result in exactly the same polarization signal, represented by dots and hashes. The dots represent polarization out of the plane of the paper and the hashes represent polarization within the plane of the board. Based on the measured data, it is unclear which orientation is correct. Ambiguities can also arise due to specular and diffuse reflections (which change the phase of light). For this reason, our network uses multiple physical priors.

The dependency of ρ_d on n is weak [AH06], and we assume $n = 1.5$ throughout the rest of this chapter. With this known n, Eq. (3.4) can be rearranged to obtain a closed-form estimation of the zenith angle for the diffuse dominant case.

When specular reflection is dominant, the degree of polarization can be written as [AH06]:

$$\rho_s = \frac{2\sin^2\theta\cos\theta\sqrt{n^2-\sin^2\theta}}{n^2-\sin^2\theta-n^2\sin^2\theta+2\sin^4\theta}. \tag{3.5}$$

Equation (3.5) can not be inverted analytically, and solving the zenith angle with numerical interpolation will produce two solutions if there are no additional constraints. For real-world objects, specular reflection and diffuse reflection are mixed depending on the surface material of the object. As shown in Fig. 3.1, the ambiguity in the azimuth angle and uncertainty in the zenith angle are fundamental limitations of SfP. Overcoming these limitations through physics-based neural networks is the primary focus of this work.

3.3.2 Learning with Physics

A straightforward approach to estimating the normals, from polarization would be to simply take the set of polarization images as input, encode it into a feature map using a CNN, and feed the feature map into a normal-regression sub-network. Unsurprisingly, we find this results in normal reconstructions with higher MAE and undesirable lighting artifacts (see Fig. 3.7). To guide the network towards more optimal solutions from the polarization information, one possible method is to force our learned solutions to adhere to the polarization equations described in Sec. 3.3.1, similar to the method used in [KWR17b]. However, it is difficult to use these physical solutions for SfP tasks due to the following reasons: **1.** Normals derived from the equations will inherently have ambiguous azimuth angles. **2.** Specular reflection and diffuse reflection coexist simultaneously, and determining the proportion of each type is complicated. **3.** Polarization images are usually noisy, causing errors in the ambiguous normals, especially when the degree of polarization is low. Shifting the azimuth angles by π or $\frac{\pi}{2}$ could not reconstruct the surface normals properly for noisy images.

Therefore, we propose directly feeding both the polarization images and ambiguous normal maps into the network and leave the network to learn how to combine both of these inputs effectively from training data. The estimated surface normals can be structured as follows:

$$\hat{\boldsymbol{N}} = f(\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, ..., \boldsymbol{I}_{\phi_M}, \boldsymbol{N}_{diff}, \boldsymbol{N}_{spec1}, \boldsymbol{N}_{spec2}), \tag{3.6}$$

where $f(\cdot)$ is the proposed prediction model, $\{\boldsymbol{I}_{\phi_1},\ \boldsymbol{I}_{\phi_2},\ ...,\ \boldsymbol{I}_{\phi_M}\}$ is a set of polarization images, and $\hat{\boldsymbol{N}}$ is the estimated surface normals. We use the diffuse model in Sec. 3.3.1 to calculate $\boldsymbol{N}_{diff}$, and $\boldsymbol{N}_{spec1}, \boldsymbol{N}_{spec2}$ are the two solutions from the specular model. These ambiguous normals can implicitly direct the proposed network to learn the surface normal information from the polarization.

Our network structure is illustrated in Fig. 3.2. It consists of a fully convolutional encoder to extract and combine high-level features from the ambiguous physical solutions and the

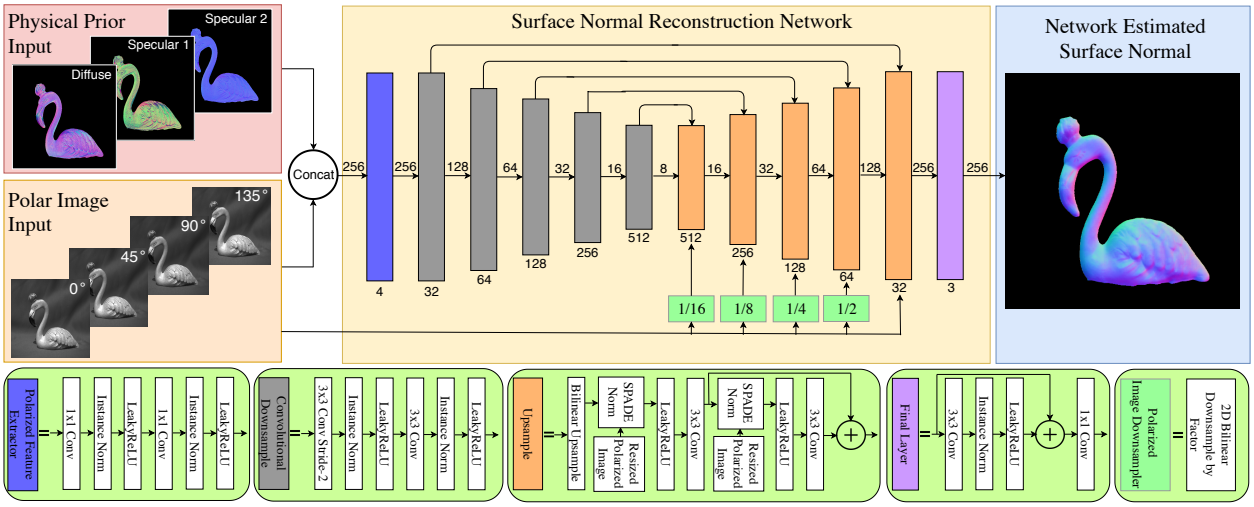


Figure 3.2: **Overview of our proposed physics-based neural network.** The network is designed according to the encoder-decoder architecture in a fully convolutional manner. The blocks comprising the network are shown below the high-level diagram of our network pipeline. We use a block based on spatially-adaptive normalization as previously implemented in [PLW19]. The numbers below the blocks refer to the number of output channels and the numbers next to the arrows refer to the spatial dimension.

polarization images, and a decoder to output the estimated normals, $\hat{\boldsymbol{N}}$. Although three polarization images are sufficient to capture the polarization information, we use images with a polarizer at $\phi_{pol} \in \{0°, 45°, 90°, 135°\}$. These images are concatenated channel-wise with the ambiguous normal solutions as the model input.

Note that the fixed nature of our network input is not arbitrary, but based on the output of standard polarization cameras. Such cameras utilize a layer of polarizers above the photodiodes to capture these four polarization images in a single shot. Our network design is intended to enable applications using this current single-shot capture technology. Single-shot capture is a clear advantage of our method over alternative reconstruction approaches, such as photometric stereo, since it allows images to be captured in a less constrained setting.

After polarization feature extraction, there are five encoder blocks to encode the input

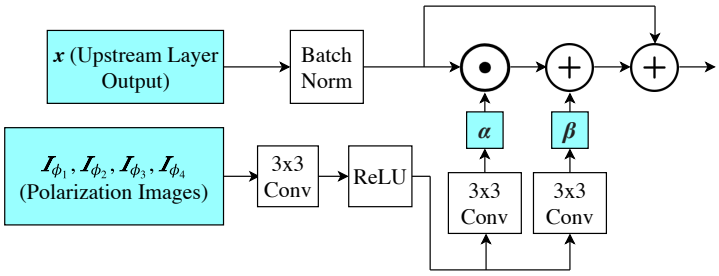


Figure 3.3: **Diagram of SPADE normalization block.** We use the polarization images to hierarchically inject back information in upsampling. The SPADE block, which takes a feature map $\boldsymbol{x}$ and a set of downsampled polarization images $\{\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, \boldsymbol{I}_{\phi_3}, \boldsymbol{I}_{\phi_4}\}$ as the input, learns affine modulation parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The circle dot sign represents elementwise multiplication, and the circle plus sign represents elementwise addition.

to a $B \times 512 \times 8 \times 8$ tensor, where B is the minibatch size. The encoded tensor is then decoded by the same number of decoder blocks, with skip connections between blocks at the same hierarchical level as proposed in U-Net [RFB15]. It has been noted that such deep architectures may wash away some necessary information from the input [HSL16, SGS15], so we apply spatially-adaptive normalization (SPADE) [PLW19] to address this problem. Motivated by their architecture, we replace the modulation parameters of batch normalization layers [IS15] in each decoder block with parameters learned from downsampled polarization images using simple, two-layer convolutional sub-networks. The details of our adaptations to the SPADE module are depicted in Fig. 3.3. Lastly, we normalize the output estimated normal vectors to unit length, and apply the cosine similarity loss function:

$$L_{cosine} = \frac{1}{W \times H} \sum_{i}^{W} \sum_{j}^{H} (1 - \langle \hat{\boldsymbol{N}}_{ij}, \boldsymbol{N}_{ij} \rangle), \tag{3.7}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, $\hat{\boldsymbol{N}}_{ij}$ is the estimated surface normal at pixel location (i, j), and $\boldsymbol{N}_{ij}$ is the corresponding ground truth surface normal. This loss is minimized when $\hat{\boldsymbol{N}}_{ij}$ and $\boldsymbol{N}_{ij}$ have identical orientation.

3.4 Dataset and Implementation Details

In what follows, we describe the dataset capture and organization as well as software implementation details. This is the first real-world dataset of its kind in the SfP domain, containing polarization images and corresponding ground truth surface normals for a variety of objects, under multiple different lighting conditions. The Deep Shape from Polarization dataset can thus provide a baseline for future attempts at applying learning to the SfP problem.

3.4.1 Dataset

A polarization camera [Luc18] with a layer of polarizers above the photodiodes (as described in Sec. 3.3.2) is used to capture four polarization images at angles $0^\circ, 45^\circ, 90^\circ$ and 135° in a single shot. Then a structured light based 3D scanner [SHI18] (with single shot accuracy no more than 0.1 mm, point distance from 0.17 mm to 0.2 mm, and a synchronized turntable for automatically registering scanning from multiple viewpoints) is used to obtain high-quality 3D shapes. Our real data capture setup is shown in Fig. 3.4. The scanned 3D shapes are aligned from the scanner's coordinate system to the image coordinate system of the polarization camera by using the shape-to-image alignment method adopted in [SMW19]. Finally, we compute the surface normals of the aligned shapes by using the Mitsuba renderer [Jak10]. Our introduced dataset consists of 25 different objects, each object with 4 different orientations for a total of 100 object-orientation combinations. For each object-orientation combination, we capture images in 3 lighting conditions: indoors, outdoors on an overcast day, and outdoors on a sunny day. In total, we capture 300 images for this dataset, each with 4 polarization angles.[2]

[2]The dataset is available at: `https://visual.ee.ucla.edu/deepsfp.htm`.

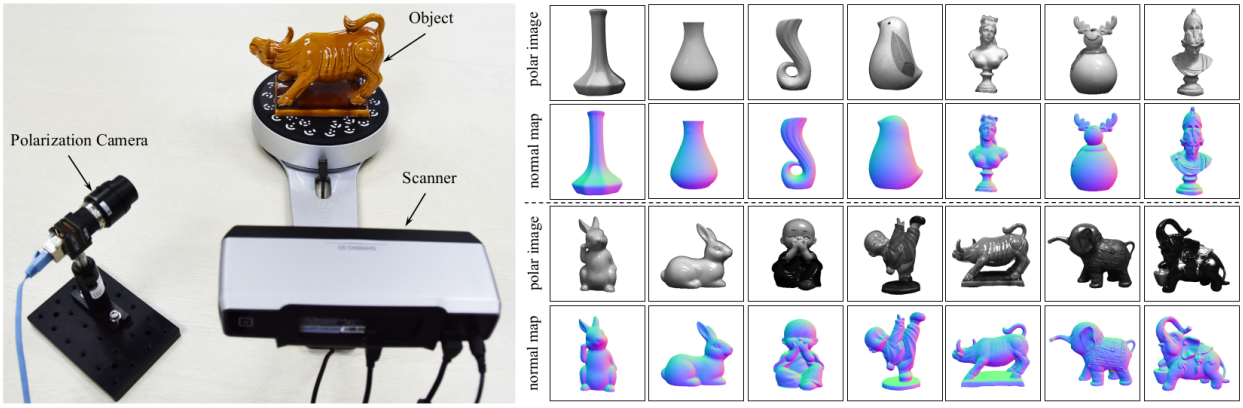


Figure 3.4: **This is the first dataset of its kind for the SfP problem.** The capture setup and several example objects are shown above. We use a polarization camera to capture four gray-scale images of an object with four polarization angles in a single shot. The scanner is put next to the camera for obtaining the 3D shape of the object. The polarization images shown have a polarizer angle of 0 degrees. The corresponding normal maps are aligned below. For each object, the capture process was repeated for 4 different orientations (front, back, left, right) and under 3 different lighting conditions (indoor lighting, outdoor overcast, and outdoor sunlight).

3.4.2 Software Implementation

Our model was implemented in PyTorch [PGC17] and trained for 500 epochs with a batch size of 4. It took around 8 hours for the network to converge with a single NVIDIA GeForce RTX 2070. We used the Adam optimizer [KB14] with default parameters with a base learning rate of 0.01. We train our model on randomly cropped 256×256 image patches, which is relatively common in shape estimation tasks [XCB14, MSL18] as a form of data augmentation.

3.5 Experimental Results

In this section, we evaluate our model with the presented challenging real-world scene benchmark and compare it against three physics-only methods for SfP. All neural networks were

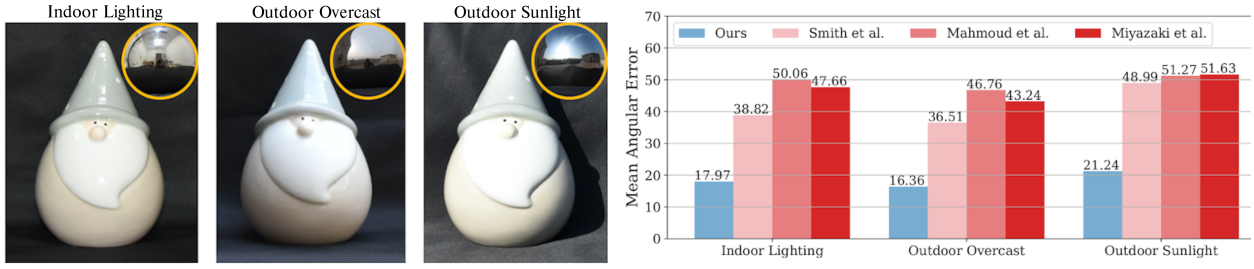


Figure 3.5: **The proposed method handles objects under varied lighting conditions**. Note that our method has very similar mean angular error among all test objects across the three lighting conditions (bottom row).

trained on the same training data as discussed in Sec. 3.4.1. To quantify shape accuracy, we compute the widely used mean angular error (MAE) score on the surface normals.

3.5.1 Comparisons to Physics-based SfP

We used a test dataset consisting of scenes that include BALL, HORSE, VASE, CHRISTMAS, FLAMINGO, DRAGON. On this test set, we implement three physics-based methods for SfP as a baseline: **1.** Smith *et al.* [SRT18]. **2.** Mahmoud *et al.* [MEF12]. **3.** Miyazaki *et al.* [MTH03]. The first method recovers the depth map directly, and we only use the diffuse model due to the lack of specular reflection masks. The surface normals are obtained from the estimated depth with bicubic fit. Both the first and the second methods require lighting input, and we use the estimated lighting from the first method during comparison. The second method also requires known albedo maps, and following convention, we assume a uniform albedo of 1. Note the method proposed in [MTH03] is the same as that presented in [AH06]. We omit comparison with Tozza *et al.* [TSZ17], as it requires two unpolarized intensity images, with two different light source directions. To motivate a fair comparison, we obtained the comparison codes directly from Smith *et al.* [SRT18]. [3]

[3] `https://github.com/waps101/depth-from-polarisation`

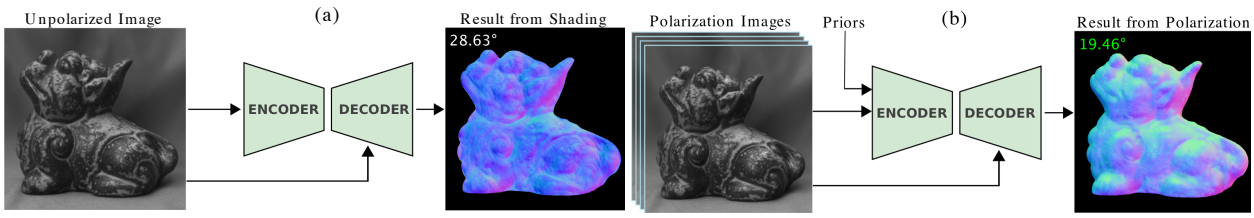


Figure 3.6: **Our network is learning from polarization cues, not just shading cues.** An ablation study conducted on the DRAGON scene. In (a) the network does not have access to polarization inputs. In (b) the network can learn from polarization inputs and polarization physics. Please refer to Fig. 3.8, row c, for the ground truth shape of the DRAGON.

3.5.2 Robustness to Lighting Variations

Figure 3.5 shows the robustness of the method to various lighting conditions. Our dataset includes lighting in three broad categories: (a) indoor lighting; (b) outdoor overcast; and (c) outdoor sunlight. Our method has the lowest MAE, over the three lighting conditions. Furthermore, our method is consistent across conditions, with only slight differences in MAE for each object between lightings.

3.5.3 Importance of Polarization

An interesting question is how much of the shape information is learned from polarization cues as compared to shading cues. Figure 3.6 explores the benefit of polarization by ablating network inputs. We compare two cases. Figure 3.6(a) shows the resulting shape reconstruction when using a network architecture optimized for unpolarized image input. The shape has texture copy artifacts and a high MAE of 28.63 degrees. In contrast, Fig. 3.6(b) shows shape reconstruction from our proposed method of learning from four polarization images and a model of polarization physics. We observe that shape reconstruction using polarization cues is more robust to texture copy artifacts, and has a lower MAE of only 19.46 degrees. Although only one image is used in the shading network (as is typical for shape from shading), this image is computed using an average of the four polarization images. Thus the

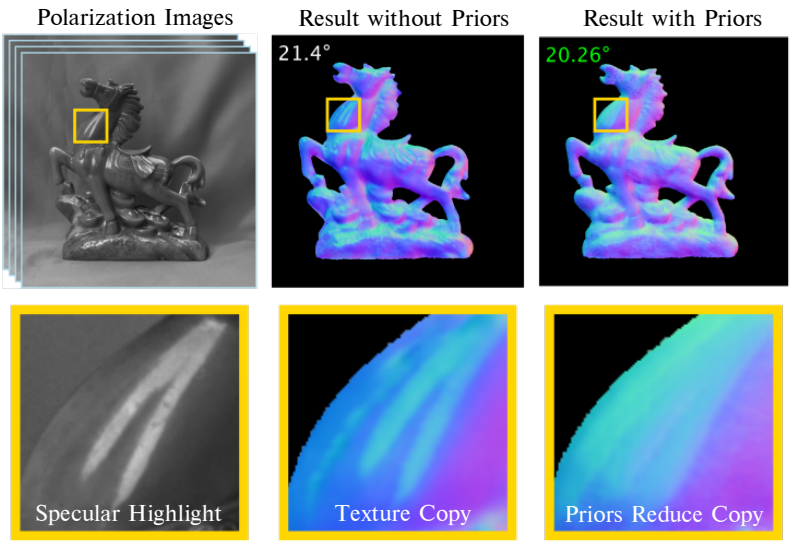


Figure 3.7: **Ablation test shows that the physics-based prior reduces texture copy artifacts.** We see that the specular highlight in the input polarization image is directly copied into the normal reconstruction without priors. Note that our prior-based method shows stronger suppression of the texture copy artifacts.

distinction between the two cases in Fig. 3.6(a) and Fig. 3.6(b) is the polarization diversity, rather than improvements in photon noise.

3.5.4 Importance of Physics Revealed by Ablating Priors

Figure 3.7 highlights the importance of physics-based learning, as compared to traditional machine learning. Here, we refer to "traditional machine learning" as learning shape using only the polarization images as input. These results are shown in the middle column of Fig. 3.7. Shape reconstructions based on traditional machine learning exhibit image-based artifacts, because the polarization images contain brightness variations that are not due to geometry, but due to specular highlights (e.g., the HORSE is shiny). Learning from just the polarization images alone causes these image-based variations to masquerade as shape variations, as shown in the zoomed inset of Fig. 3.7. A term used for this is *texture copy*,

Scene	**Proposed**	Smith [SRT18]	Mahmoud [MEF12]	Miyazaki [MTH03]
Box	**23.31°**	**31.00°**	41.51°	45.47°
Dragon	**21.55°**	**49.16°**	70.72°	57.72°
Father Christmas	**13.50°**	39.68°	**39.20°**	41.50°
Flamingo	**20.19°**	**36.05°**	47.98°	45.58°
Horse	**22.27°**	55.87°	**50.55°**	51.34°
Vase	**10.32°**	**36.88°**	44.23°	43.47°
Whole Set	**18.52°**	**41.44°**	49.03°	47.51°

Table 3.2: **Our method outperforms previous methods for each object in the test set.** Numbers represent the MAE averaged across the three lighting conditions for each object. The best model is marked in magenta and the second-best is in blue.

where image texture is undesirably copied onto the geometry [KTS15]. In contrast, the proposed results with physics priors are shown in the rightmost inset of Fig. 3.7, showing less dependence on image-based texture (because we also input the geometry-based physics model).

3.5.5 Quantitative Evaluation

We use MAE[4] to make a quantitative comparison between our method and the previous physics-based approaches. Table 3.2 shows that the proposed method has the lowest MAE on each object, as well as the overall test set. The two most challenging scenes in the test set are the horse and the dragon. The former has intricate detail and specularities, while the latter has a mixed material surface. The physics-based methods struggle on these challenging scenes as all scenes have over 49 degrees of mean angular error. The method from Smith *et al.* [SRT18] has the second-lowest error on the dragon scene, but the method from

[4]MAE is the most commonly reported measure for surface normal reconstruction, but in many cases, it is a deceptive metric. We find that a few outliers in high-frequency regions can skew the MAE for entire reconstructions. Accordingly, we emphasize the qualitative comparisons of the proposed method to its physics-based counterparts.

Miyazaki *et al.* [MTH03] has the second-lowest error on the HORSE scene. On the overall test set, the physics-based methods are all clustered between 41.4 and 49.0 degrees, while the physics-based deep learning approach we propose achieves over a two-fold reduction in error to 18.5 degrees.

The reader may wonder why the physics-based methods perform poorly on tested scenes. The result from Smith *et al.* [SRT18] assumes a reflection model and combinatorial lighting estimation, which do not appear to scale to unconstrained, real-world environments, resulting in a normal map with a larger error. Mahmoud *et al.* [MEF12] uses shading constraints that assume a distant light source, which is not the case for some of the tested scenes, especially the indoor ones. Finally, the large region-wise anomalies on many of the results from Miyazaki *et al.* [MTH03] are due to the sensitive nature of their histogram normalization method.

3.5.6 Qualitative Evaluation

Figure 3.8 shows qualitative and quantitative data for various objects in our test set. The RGB images in (row a) are not used as input but are shown in the top row of the figure for context about material properties. The input to all the methods shown is four polarization images, shown in (row b) of Fig. 3.8. The ground truth shape is shown in (row c), and corresponding shape reconstructions for the proposed method are shown in (row d). Comparison methods are shown in (row e) through (row g). It is worth noting that the physics-based methods particularly struggle with *texture copy* artifacts, where color variations masquerade as geometric variations. This can be seen in Fig. 3.8, (row f), where the physics-based reconstruction of Mahmoud [MEF12] confuses the color variation in the beak of the FLAMINGO with a geometric variation. In contrast, our proposed method, shown in (row d), recovers the beak more accurately. Beyond texture copy, another limitation of physics-based methods lies in the difficulty of solving the *ambiguity problem*, discussed earlier in this chapter. In row g, the physics-based approach from Miyazaki *et al.* [MTH03] has significant ambiguity errors. This can be seen as the fixed variations in color of normal maps, which are not due to

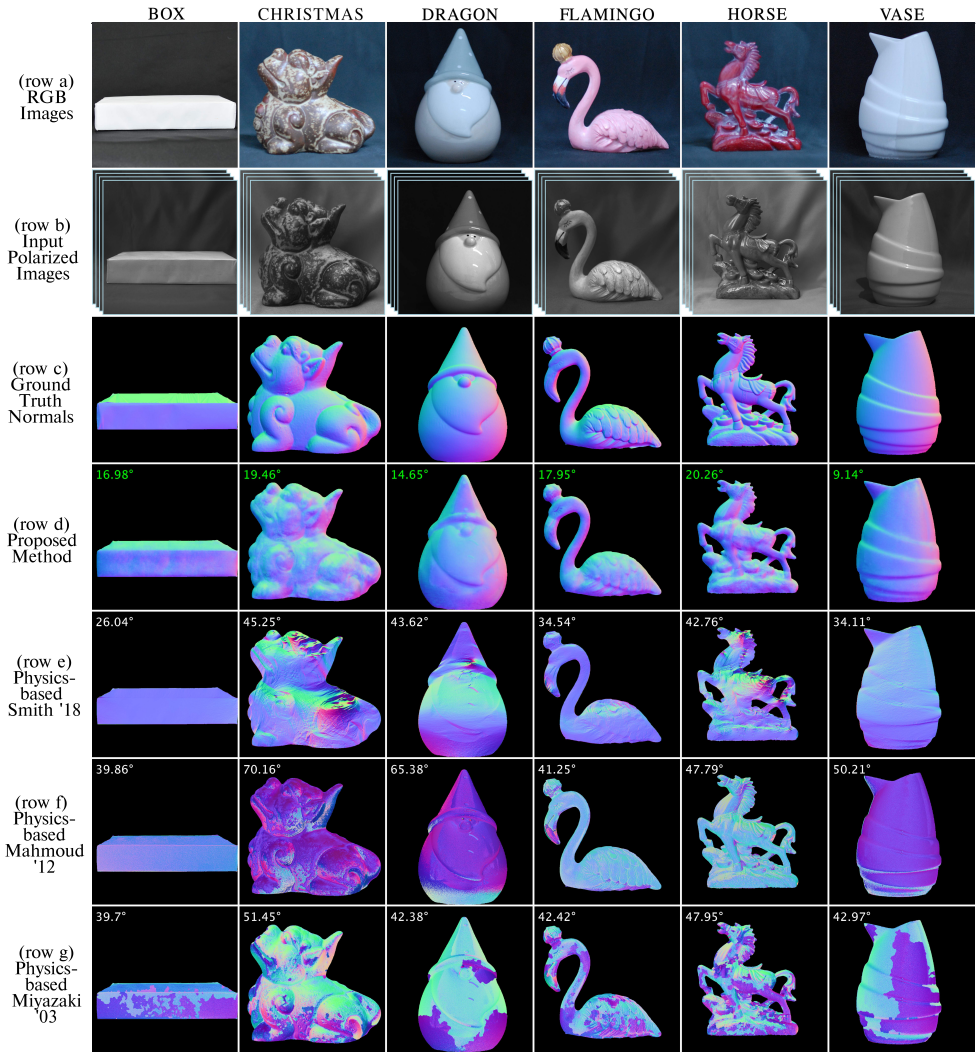


Figure 3.8: **The proposed method shows qualitative and quantitative improvements in shape recovery in our test dataset.** (row a) The RGB scene photographs for context—these are not used as the input to any of the methods. (row b) The input to all methods is a stack of four polarization photographs at angles of 0°, 45°, 90°, and 135°(row c). The ground truth normals obtained experimentally. (row d) The proposed approach for shape recovery. (row e-g) We compare with physics-based SfP methods by Smith *et al.* [SRT16], Mahmoud *et al.* [MEF12] and Miyazaki *et al.* [MTH03]. (We omit the results from Atkinson *et al.* [AH06], which uses a similar method as [MTH03]).

random noise. Although less drastic, the physics-based method of Smith *et al.* [SRT18] also shows such fixed pattern artifacts, due to the underdetermined nature of the problem. Our proposed method is fairly robust to fixed pattern error, and our deviation from ground truth is largely in areas with high-frequency detail. Although the focus of Fig. 3.8 is to highlight qualitative comparisons, it is worth noting that the MAE in of the proposed method is the lowest for all these scenes (lowest MAE is highlighted in green font).

3.6 Discussion

In summary, we presented a first attempt at re-examining SfP through the lens of deep learning, and specifically, physics-based deep learning. Table 3.2 shows that our network achieves over a two-fold reduction in shape error, from 41.4 degrees [SRT18] to 18.5 degrees. An ablation test verifies the importance of using the physics-based prior in the deep learning model. In experiments, the proposed model performs well under varied lighting conditions, while previous physics-based approaches have either higher error or variation across lighting.

Future Work: The framerate of our technique is limited both by the feed-forward pass, as well as the time required to calculate the physical prior (about 1 second per frame). Future work could explore parallelizing the physics-based calculations or using approximations for more efficient computing. As discussed in Sec. 3.5.5, the high MAE is largely due to a few regions with extremely fine detail. Finding ways to effectively weigh these areas more heavily or add a refinement stage focused on these challenging regions, are promising avenues for future exploration. Moreover, identifying a metric better able to capture the quality of reconstructions than MAE would be valuable for the continued study of learning-based SfP.

Conclusion: We hope the results of this study encourage future explorations at the seamline of deep learning and polarization as well as the broader field of fusion of data-driven and physics-driven techniques.

CHAPTER 4

Style Transfer with Bio-realistic Appearance Manipulation for Skin-tone Inclusive rPPG

4.1 Introduction

During the pandemic, telehealth consults have increased more than 50-fold for certain groups (e.g., those with chronic diseases) [WBH21] due to the concerns that the congregation of people may increase the risk of contraction. Although contact sensors (e.g., electrocardiograms, oximeters) provide a gold-standard measurement of human body functions, these contact devices are not widely available, which makes a non-contact way of detecting vital signs crucial for telehealth settings [BDO20, SLW20, Bok21]. Non-contact health sensing can also benefit applications in clinical settings, such as neonatal intensive care unit (ICU) sensing [VCJ19], as the contact sensors may cause infection for these vulnerable groups. For non-contact health sensing systems to be deployed at scale in society, it is important to ensure their performance consistency across a broad range of ethnic groups [Kad21]. In this chapter, we use remote photoplethysmography (rPPG) as an example to explore how to push Pareto frontier by promoting both accuracy and fairness in heart rate estimation with synthetic augmentation as shown in Fig. 4.1. We select camera-based rPPG [VSN08, PMP10a] since it provides a solution to the above scenarios given that web cameras are more ubiquitously available, contactless, and low-cost. In the meantime, the existing rPPG datasets are usually overwhelmed by subjects of light skin tones, which makes it problematic to deploy rPPG for various demographic groups.

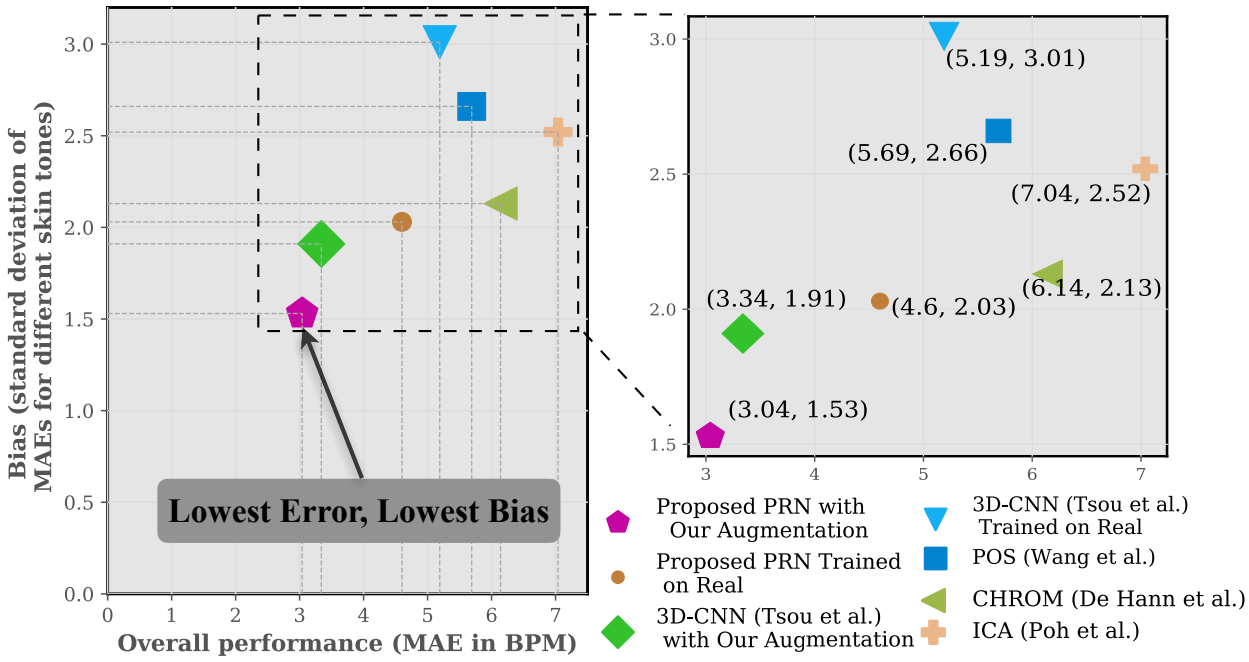


Figure 4.1: **Our proposed augmentation method pushes the Pareto frontier toward both axes: accuracy and equity for rPPG.** We use the mean absolute error (MAE) of the heart rate (HR) estimation for all skin tones as the overall performance metric and the standard deviation of MAEs across different skin-tone groups as the bias metric. Our proposed augmentation method has the lowest estimation error with minimized bias as compared with the existing solutions. HR MAE is measured in the unit of beats per minute (BPM) in the plot.

Camera-based rPPG uses subtle skin color variations on the face to obtain physiological signals. When the light hits the face, the amount of light reflected or absorbed is determined by the physiological processes, and the color change corresponding to the Blood Volume Pulse (BVP) is synchronized with the heart rate (HR), which provides the feasibility to extract HR from facial videos. While data-driven neural networks have exhibited remarkable estimation accuracy for non-contact camera-based sensing [McD18, YPL19, RIS19, NSH19], there exist several practical constraints towards collecting large-scale data from patients for these deep learning models: (1) demographic biases in society that translate to data (e.g., innovation

happening in some countries/regions may not have access to a diverse dataset); (2) the requirement of medical-grade sensors and necessity of intrusive/semi-intrusive traditional methods for data collection; and (3) patient privacy concerns (e.g., OBF dataset [LAS18] is not publicly available due to the license issue).

Recent studies have shown that computer vision algorithms have been disadvantaging the underrepresented groups in some applications, such as face recognition [BG18]. Non-contact rPPG estimation is not an exception given the unbalanced and relatively small datasets in the field [NMV20b]. There are very rare subjects with dark skin tones in the existing benchmark datasets. More specifically, MMSE-HR [ZGW16], AFRL [EBM14], and UBFC-RPPG [BMB19] only contain roughly 10%, 0%, and 5% dark-skinned subjects respectively. With the training sets heavily biased towards subjects of light skin tones, the state-of-the-art data-driven rPPG models usually fail to generalize their performance to the underrepresented groups [NMV20b]. This prohibits the clinical deployment of these algorithms since it is critical for rPPG algorithms to have consistent performance across different demographic groups in clinical settings.

Realizing the difficulty of recruiting patients to collect large-scale rPPG datasets in the university setting, synthetic augmentation of facial videos has become an active research topic recently. McDuff *et al.* [MHW20] use synthetic avatars with ray tracing to reflect the blood volume changes under various configurations. However, as the authors point out, that infrastructure is labor-intensive and requires a significant amount of rendering time for each frame (approximately 20 seconds per frame), which impedes their scalability. Pulse signals can also be incorporated to make the synthetic avatars more lifelike, yet it is difficult for avatar-based methods to generate a balanced dataset due to the lack of dark-skinned avatars [MN21]. Tsou *et al.* [TLH20] augment source rPPG videos with other specified pulse signals, however, their framework is restricted to the face appearance in the original source videos and fails to produce novel videos with dark skin tones.

In contrast to these prior arts, we do a first attempt to directly augment the existing

rPPG dataset by translating videos of light-skinned subjects to dark skin tones. This is difficult because the color variations due to blood volume changes are subtle, and the generation network has to be carefully designed to reflect these subtle changes while conducting skin tone translation without accessing real rPPG videos of dark-skinned subjects. However, this technique is rewarding, since it is capable of producing both photo-realistic and physiologically accurate synthetic videos in a fast manner (approximately 0.005 seconds per frame on average for our model) and can assist the development of algorithms and techniques for remote diagnostics and healthcare. In the experiment, our proposed method can reduce around 31% HR estimation error for the dark-skinned group and show 46% improvement in bias mitigation for all the groups, as compared with the existing architecture trained with just real samples.

Yucer *et al.* [YAA20] introduce a race translation model across various racial domains with a CycleGAN-based architecture [ZPI17]. However, their work is not designed to incorporate pulsatile signals. As illustrated in Fig. 4.2, this vanilla skin tone translation network [YAA20] merely focuses on the visual appearance, and the pulsatile signals are not preserved. To address this issue, we propose a learning framework that can augment realistic rPPG videos with dark skin tones that are of high fidelity. The framework consists of two interconnected components: (1) a generator to translate light skin tones to dark skin tones and (2) an rPPG estimator named PhysResNet (PRN) to encourage pulsatile signals within the generated videos. The generator is trained to learn both the visual appearance and the subtle color variations with respect to the underlying blood volume variations, and the rPPG network can simultaneously benefit from the generator to generalize its performance in diverse groups. We also demonstrate that our generated synthetic videos can be directly utilized to improve the performance of the state-of-the-art data-driven rPPG method with reduced bias across different skin color groups.

To summarize, the contributions of our work include:

- We introduce a first attempt to translate facial videos of light-skinned subjects to dark

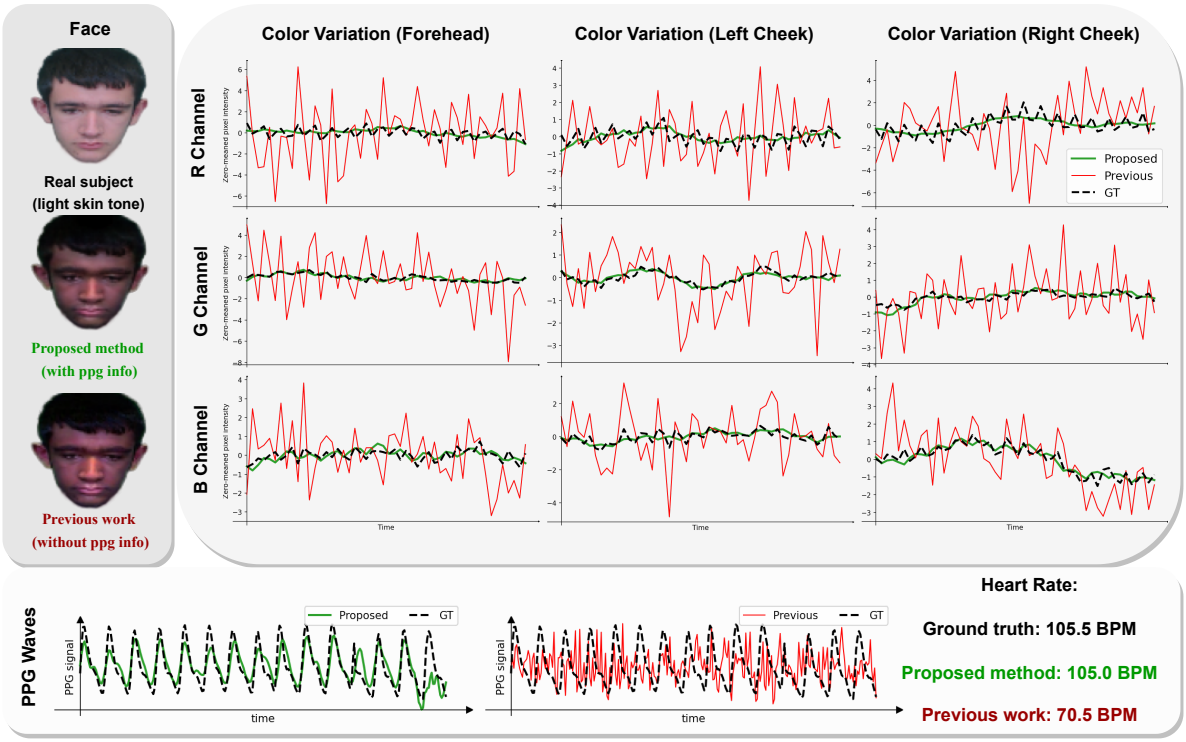


Figure 4.2: **The proposed method successfully incorporates pulsatile signals into the generated videos, while the existing work [YAA20] only focuses on the visual appearance.** For different facial regions, frames generated by the proposed method exhibit similar pixel intensity variations as compared with frames from real videos, while the prior work shows unrealistic RGB variations. As a result, pulsatile signals can be well preserved in our method as opposed to the vanilla skin tone translation.

tones while preserving the underlying blood volume variations;

- We demonstrate that our synthetic videos can be directly utilized to improve the performance of the state-of-the-art deep rPPG methods with mitigated bias across different demographic groups;

- We propose a simple yet efficient rPPG estimation model based on 3D convolution operations and show that the proposed model can achieve state-of-the-art performance on various facial videos.

4.2 Related Work

4.2.1 Imaging Photoplethysmography

Imaging PPG methods aim to recover the pulsatile signal from the subtle color changes in the face videos. Algorithms of detecting non-contact PPG signal can roughly be divided into three categories: Signal decomposition [LRK11, PMP10a, PMP10b, TAR16, WSD15], model-based methods [DV14, DJ13, WBS16, KVS15], and deep learning methods [McD18, YPL19, RIS19, NSH19]. Signal decomposition techniques based on Blind Source Separation (BSS) techniques decompose/demix the face videos into different sources utilizing PCA [LRK11] or ICA [PMP10a]. However, these methods do not exploit skin reflectance properties that are specific to rPPG problems.

Model-based methods, such as CHROM [DJ13], apply color space transforms to linearly combine the chrominance signals to obtain the final PPG signals. The Pulse Blood Vector [DV14] method uses characteristic blood volume changes to weigh different color channels. This method can be further improved by first projecting the temporally-normalized skin tone onto the plane which is orthogonal to the intensity variation term and then linearly combining the projected signals [WBS16]. These methods use all the face skin pixels for the rPPG measurement, which may achieve sub-optimal results as each pixel may have a rather different contribution to the pulse signals.

More recently, data-driven methods have gained more attention [CM18, YPL19, SCC21, LHZ21, NMV21a, TLH20]. More specifically, DeepPhys [CM18] proposes a Convolutional Attention Network (CAN) which uses appearance information to guide motion estimation to recover physiological signals. PhysNet [YPL19] captures the temporal correlation of the pulse signals in the rPPG face videos using a 3D spatial-temporal Convolutional Neural Network (3D-CNN) or a Recurrent Neural Network (RNN). While these methods exhibit remarkable performance improvement as compared with model-based solutions, their generalization capability is highly affected by the diversity of the training samples.

4.2.2 Synthetic Augmentation in Healthcare

Medical images have been widely used in clinics and played a critical role in various clinical applications. Due to the significant cost of collecting high-quality medical images, most datasets are very limited in size, and this has impeded the scientific progress. Traditional data augmentation schemes, such as horizontal/vertical flipping, rotation, and translation, are used and have become a standard procedure for training deep neural networks in computer vision applications [KSH12]. However, the diversity of the dataset can not be improved significantly by such schemes. Medical image synthesis can be of great benefit to address this problem [NTL18], such as synthetic skin lesion images [QLZ20] and synthetic Magnetic Resonance (MR) images for brain tumors [FE20].

In the rPPG field, McDuff *et al.* [MHW20] use synthetic avatars with blood volume changes to generate rPPG face videos under various settings. The infrastructure for their pipeline is expensive and labor-intensive, which makes it difficult to scale up their generation process. Tsou *et al.* [TLH20] propose to augment the source rPPG videos with a specified rPPG signal present in another video and show improvement on the heart rate estimation task with the augmented dataset. Their model cannot augment the original dataset with different face appearance, such as skin tones. In contrast, we use a generator to synthesize bio-realistic videos with dark skin tones to reflect the underlying subtle PPG signal variations in a scalable way and show that it is beneficial to improving the measurement of heart rate for remote clinical use.

4.2.3 Neural Style Transfer for Medicine

Neural style translation has been applied to various medical applications, such as digital histopathology since the images of the same tissue recorded from different labs and hospitals usually exhibit a large variation in terms of their colors [MNM09, BLT15, LP15, BRC15]. Color translation frameworks based on neural networks [XDV19, CLC17, LPL20]

have been proposed to learn not only the certain color distribution but also the corresponding histopathological patterns. The performance of tissue segmentation and classification is improved with the color-augmented histopathological datasets. Inspired by these successful applications, our work provides a first attempt to bridge the gap between neural style transfer and rPPG for bio-realistic skin tone augmentation.

4.3 Method

Our bio-realistic skin translation framework is designed to adhere to the light transport analysis of human skin. In Sec. 4.3.1, we briefly review the existing skin reflection theory that models pulsatile blood variations. In Sec. 4.3.2, we detail our pipeline to translate videos of real subjects with light skin tones to synthetic dark skin tones. The implementation details are provided in Sec. 4.3.3.

4.3.1 Optical Model for Pulsatile Blood Variations

Under the assumption of a light source with a constant spectral composition and varying intensity, RGB channels $\mathbf{C}_k(t)$ at the kth skin pixel measured by a remote color camera can be described by the dichromatic reflection model as a time-varying function [WBS16] as illustrated in Fig. 4.3:

$$\mathbf{C}_k(t) = I(t) \cdot \big(\mathbf{v}_s(t) + \mathbf{v}_d(t)\big) + \mathbf{v}_n(t), \tag{4.1}$$

where $I(t)$ is the luminance intensity level, $\mathbf{v}_s(t)$ and $\mathbf{v}_d(t)$ are the time-varying specular and diffuse reflections respectively, and $\mathbf{v}_n(t)$ is quantization noise. Specular component $\mathbf{v}_s(t)$ in Eq. (4.1) is a result of the mirror-like reflection from the skin surface, which is usually considered to be BVP independent. We can write $\mathbf{v}_s(t)$ as the following equation [WBS16]:

$$\mathbf{v}_s(t) = \mathbf{u}_s \cdot \big(s_0 + s(t)\big), \tag{4.2}$$

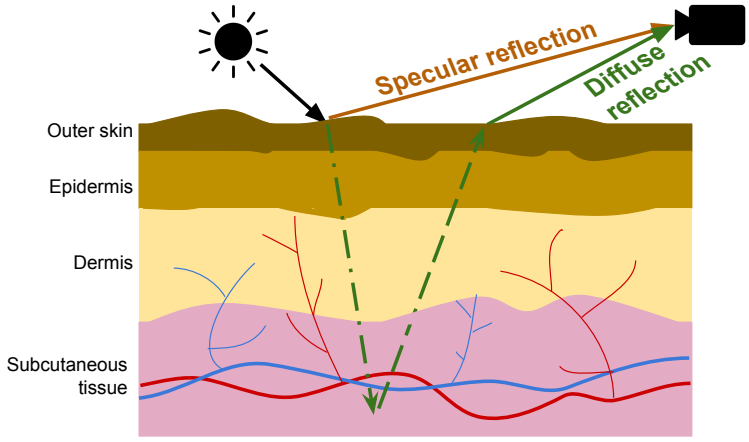


Figure 4.3: **Illustration of the dichromatic skin model.** The specular component is due to the reflection from the skin surface, and the diffuse component is related to the absorption and scattering properties of the skin tissues. Our bio-realistic skin tone translation model aims to conduct skin tone translation while preserving the relative variations between BVP and skin appearance.

where $\mathbf{u}_s$ is the unit color vector of the incident light, s_0 is the stationary part of the specular reflection, and $s(t)$ is the varying part of the specular reflection induced by motion. Diffuse reflection $\mathbf{v}_d(t)$ in Eq. (4.1) is related to the absorption and scattering properties of the skin tissues, and its varying component is identified as a key indicator to the blood volume changes [WBS16]:

$$\mathbf{v}_d(t) = \mathbf{u}_d \cdot d_0 + \mathbf{u}_p \cdot p(t), \tag{4.3}$$

where $\mathbf{u}_d$ is the unit color vector of the skin, d_0 is the stationary reflection strength, $\mathbf{u}_p$ is the relative pulsatile strengths in RGB channels, and $p(t)$ is the pulse signal. Substituting Eq. (4.2) and Eq. (4.3) into Eq. (4.1), we can write $\mathbf{C}_k(t)$ as follows:

$$\mathbf{C}_k(t) = I(t) \cdot \Big(\mathbf{u}_s \cdot \big(s_0 + s(t)\big) + \mathbf{u}_d \cdot d_0 + \mathbf{u}_p \cdot p(t)\Big) + \mathbf{v}_n(t). \tag{4.4}$$

The stationary parts of the specular and diffuse components can be combined into a single skin stationary term:

$$\mathbf{u}_c \cdot c_0 = \mathbf{u}_s \cdot s_0 + \mathbf{u}_d \cdot d_0, \tag{4.5}$$

where $\mathbf{u}_c$ is the unit color vector of the skin reflection, and c_0 denotes the reflection strength. This further simplifies Eq. (4.4) as:

$$\mathbf{C}_k(t) = I_0 \cdot \big(1 + i(t)\big) \cdot \big(\mathbf{u}_c \cdot c_0 + \mathbf{u}_s \cdot s(t) + \mathbf{u}_p \cdot p(t)\big) + \mathbf{v}_n(t), \tag{4.6}$$

where $I(t)$ is expressed as the sum of a stationary part I_0 and a time-varying motion-induced part $I_0 \cdot i(t)$. Video-based PPG measurement algorithms aim to estimate the pulse signal $p(t)$ from the pixel intensity $\mathbf{C}_k(t)$ by separating the physiological and non-physiological variations, while the primary focus of our work is to establish an inverse mapping between $p(t)$ and $\mathbf{C}_k(t)$ for dark-skin realistic human faces in a data-driven manner.

4.3.2 Bio-realistic Skin Tone Translation

In order to translate real subjects with light skin tones to synthetic subjects with dark skin tones, we utilize two interconnected networks: a video generator G and an rPPG estimator E, as illustrated in Fig. 4.4. We next describe the proposed 3D convolutional video generator, the rPPG estimation network, and our joint optimization scheme.

4.3.2.1 3D Convolutional Video Generator

The goal of our video generator G is to translate frame sequences of real light-skinned subjects to synthetic dark-skinned subjects. We propose a novel 3D convolutional neural network to accomplish this goal. The model consists of an encoder (several convolutional layers), a transformer (6 ResNet Blocks), and finally a decoder (several convolutional layers).

The generator takes 256 consecutive frames $\mathbf{I}_{light}$ at size 80×80 as the input and generates the corresponding translated frames in the same dimension. Since the paired ground-truth translated frames do not exist, we use a race transfer model [YAA20] pretrained on VGGFace2 [CSX18] to generate the pseudo target frames $\mathbf{I}_{dark}$. More specifically, the generator *Caucasian-to-African* in [YAA20] is utilized to translate videos of light-skinned subjects in the existing rPPG dataset to dark skin tones.

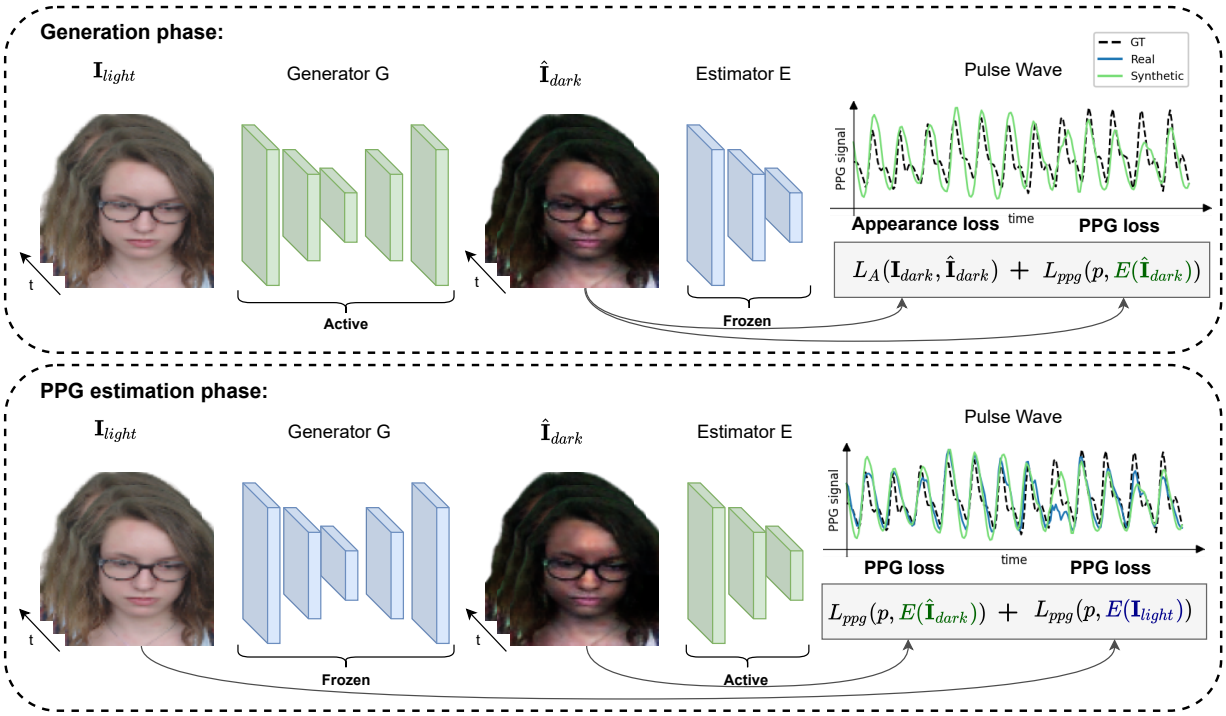


Figure 4.4: **Illustration of the proposed joint optimization framework.** Our framework is capable of translating light-skinned facial videos to dark skin tones while maintaining the original pulsatile signals. With a two-phase weight updating scheme, the rPPG estimation network can benefit from the synthetic dark-skinned videos and gradually learn to conduct inference on dark-skinned subjects without accessing real facial videos with dark skin tones.

The generator is first supervised by the L1 distance between the pseudo target frames $\mathbf{I}_{dark}$ and the generated frames $\hat{\mathbf{I}}_{dark} = G(\mathbf{I}_{light})$ to learn the visual appearance of the synthetic dark-skinned subjects. At this stage, the output frames $\hat{\mathbf{I}}_{dark}$ do not contain pulsatile signal, since the target frames $\mathbf{I}_{dark}$ from [YAA20] are generated in a frame-by-frame manner without temporal pulse correspondence along the time dimension. In the joint optimization part, we describe how to further incorporate the pulsatile signals presented in the original videos $\mathbf{I}_{light}$ into the generated frames.

4.3.2.2 PRN: rPPG Estimator with Residual Connections

The rPPG estimator is designed to model the BVP temporal information from a sequence of facial frames. Similarly, it takes 256 consecutive frames at size 80×80 as the input, and its output is the corresponding BVP value for each input frame. We build our novel rPPG estimator based on 3D convolution operations. It consists of three consecutive 3D convolutional blocks with residual connections, and an average pooling is performed after each block for the downsampling purpose.

To supervise the network, we use a negative Pearson correlation loss between the estimated pulse signals $\hat{p} \in \mathbb{R}^T$ and the ground-truth pulse signals $p \in \mathbb{R}^T$:

$$L_{ppg}(p, \hat{p}) = 1 - \frac{T \sum_i p_i \hat{p}_i - \sum_i p_i \sum_i \hat{p}_i}{\sqrt{\left(T \sum_i p_i^2 - \left(\sum_i p_i\right)^2\right)\left(T \sum_i \left(\hat{p}_i\right)^2 - \left(\sum_i \hat{p}_i\right)^2\right)}}, \tag{4.7}$$

where the summation $\sum_i$ is over the frame length T. This negative Pearson correlation loss has shown to be more effective as compared with the point-wise mean squared error (MSE) loss in the previous work [YPL19]. We first train PRN with only real subjects, and this simple yet efficient architecture can already achieve state-of-the-art performance on the existing rPPG datasets. In the next part, we detail how to further incorporate the synthetic subjects into the training process.

4.3.2.3 Joint Optimization

The generator trained with L1 loss in the previous part fails to produce synthetic dark-skinned subjects with desired pulsatile information, and the rPPG estimator trained with only real light-skinned subjects exhibits poor generalization capability on unseen data or data that rarely appears in the training set (i.e., the underrepresented group with dark skin tones). To make use of these two models, we design a joint optimization mechanism to incorporate pulsatile signals into the synthetic videos and improve the generalizability of the

rPPG estimator simultaneously.

We use a two-phase weight updating scheme to train the video generator and the rPPG estimator simultaneously. These two phases are alternated within each mini-batch as illustrated in Fig. 4.4. In the generation phase, we freeze the weight of the rPPG estimator E, and the generator G is supervised by the following loss function to maintain both the visual appearance and the pulsatile information:

$$L_G(\mathbf{I}_{light}, p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + \lambda * L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}), \tag{4.8}$$

$$L_A(\mathbf{I}_{dark}, \hat{\mathbf{I}}_{dark}) = \frac{1}{\sum_i z_i} \sum_i z_i |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}|, \tag{4.9}$$

$$z_i = \begin{cases} 0 & \text{if } |\mathbf{I}_{dark_i} - \hat{\mathbf{I}}_{dark_i}| < \epsilon \\ 1 & \text{otherwise} \end{cases}, \tag{4.10}$$

where $\hat{\mathbf{I}}_{dark} = G(\mathbf{I}_{light})$ is the generated frame sequence from synthetic dark-skinned subjects, λ is the balance factor, $L_A(\cdot)$ is the visual appearance loss designed based on a threshold L1 loss, and ϵ is the selected threshold. The weighting factor λ is chosen to be 1.0. Directly enforcing an L1 loss between $\mathbf{I}_{dark}$ and $\hat{\mathbf{I}}_{dark}$ causes the generator to struggle between the visual appearance and the pulse information, since the pseudo ground truth $\mathbf{I}_{dark_i}$ from [YAA20] does not contain the desired BVP variations. Therefore, we relax the appearance loss $L_A(\cdot)$ by a threshold ϵ. The relaxation is based on the observation that the color changes due to BVP variations are subtle in the RGB domain. In our implementation, we select $\epsilon = 0.1$ based on an empirical analysis of the color variations in real videos.

In the rPPG estimation phase, we freeze the weight of the generator G and train the rPPG estimator E with both real and synthetically augmented frame sequences:

$$L_E(\mathbf{I}_{light}, \hat{\mathbf{I}}_{dark}), p) = L_{ppg}(p, E(\hat{\mathbf{I}}_{dark})) + L_{ppg}(p, E(\mathbf{I}_{light})). \tag{4.11}$$

Both real and synthetic subjects are utilized to supervise the rPPG network E while updating its weights. This arrangement allows E to gradually adapt to the synthetic dark-skinned subjects without losing estimation accuracy on real subjects. With this two-phase updating

rule, both the generator and the rPPG estimator benefit from each other in an alternate manner. At convergence, the generator G can successfully translate frame sequences from real light-skinned subjects to dark skin tones while maintaining the original BVP variations, and the estimator E can generalize its performance to dark skin tones without using actual real videos from dark-skinned subjects.

4.3.3 Implementation Details

The facial bounding box for each video is estimated by applying a face detector based on Multitask Cascaded Convolutional Neural Networks (MTCNN) [ZZL16] to its first frame, and a square region with 160% width and height of the detected bounding box is cropped and resized to 80×80 using linear interpolation. The learning rate for the generator and the rPPG network are 0.0001 and 0.0003 respectively. The learning rates are modified based on a cosine annealing schedule during training [LH17]. The networks are initialized with Kaiming initialization [HZR15] with a batch size of two and ReLU activation. We use Adam [KB14] solver with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The network architectures are implemented with batch normalization [IS15] in PyTorch [PGM19a], and the experiments are conducted on a single NVIDIA Tesla V100 GPU.

4.4 Experiments

To demonstrate the effectiveness of the proposed method, we conduct a comprehensive evaluation on several commonly used rPPG datasets. We describe the datasets for our experiment in Sec. 4.4.1, the comparison methods in Sec. 4.4.2, and the evaluation metrics in Sec. 4.4.3. Some illustration of the generated synthetic videos is provided in Sec. 4.4.4. The performance of different comparison models and the proposed solutions are listed in Sec. 4.4.5 and Sec. 4.4.6. The bias mitigation analysis is shown in Sec. 4.4.7.

4.4.1 Datasets

4.4.1.1 UBFC-RPPG [BMB19]:

UBFC-RPPG database contains 42 front facial videos from 42 subjects, and the corresponding ground-truth PPG signals are collected from a fingertip pulse oximeter. The videos are recorded at 30 frames per second with a resolution of 640x480 in the uncompressed 8-bit AVI format. Each video is roughly one minute long.

4.4.1.2 VITAL Dataset [CKK20]:

Facial videos are recorded at 1920x1080 pixel resolution and 30 frames per second for 60 subjects at room lighting in the highly compressed MP4 format. Each video is roughly 2 minutes long. A Philips IntelliVue MX800 patient monitor is utilized for ground-truth vital sign monitoring. The subject wears a blood pressure cuff, 5-ECG leads, and a finger pulse oximeter, which is connected to the MX800 unit. Diverse skin tones and varied demographic groups are represented in the dataset. We use 58 subjects in the VITAL dataset (subject 26 and subject 40 are left out due to data errors in the collecting process). For the skin types quantified by Fitzpatrick scales [Fit88], there are 5, 16, 14, 11, 5, and 7 subjects respectively from I (lightest) to VI (darkest).

4.4.2 Comparison Methods

We compare our model with three conventional methods: POS [WBS16], CHROM [DJ13], and ICA [PMP10a]. These rPPG baseline methods are implemented based on the publicly available MATLAB toolbox [MB19], and we follow the procedures in the toolbox to obtain facial pixels of interest, i.e., converting facial frames from RGB to YC_RC_B and identifying skin pixels based on a predefined threshold. We also compare with a data-driven state-of-the-art rPPG algorithm 3D-CNN [TLH20]. It is implemented based on the architecture

description as detailed in the original publication.

4.4.3 Evaluation Metrics

After obtaining the estimated pulse waves from each model, we apply a Butterworth filter to the output signals with cut-off frequencies of 0.7 and 2.5 Hz for heart rate estimation. The filtered waves are divided with sliding windows of 30-second length and 1-second stride, and a heart rate is estimated based on the position of the peak frequency for each window. For each subject, four error metrics are calculated and averaged over all windows. The four metrics include mean absolute error (MAE), root mean square error (RMSE), Pearson's correlation coefficient (PCC) between the estimated heart rate and the ground-truth heart rate, and signal-to-noise ratio (SNR) of the estimated PPG waves. The ground-truth HR for UBFC-RPPG is obtained by applying the same procedures as described above to the ground-truth pulse waves, and the ground-truth HR for the VITAL dataset is obtained from the MX800 patient monitor through ECG signals. Details of these metrics are provided as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{N} \left| \text{HR}_i - \hat{\text{HR}}_i \right|}{N}, \tag{4.12}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(\text{HR}_i - \hat{\text{HR}}_i \right)^2}{N}}, \tag{4.13}$$

$$\text{PCC} = \frac{T \sum_i p_i \hat{p}_i - \sum_i p_i \sum_i \hat{p}_i}{\sqrt{\left(T \sum_i p_i^2 - \left(\sum_i p_i \right)^2 \right) \left(T \sum_i \left(\hat{p}_i \right)^2 - \left(\sum_i \hat{p}_i \right)^2 \right)}}, \tag{4.14}$$

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{f=0.75}^{2.5} \left(U_t(f) \hat{S}(f) \right)^2}{\sum_{f=0.75}^{2.5} \left(\left(1 - U_t(f) \right) \hat{S}(f) \right)^2} \right), \tag{4.15}$$

where HR is the ground-truth heart rate, $\hat{\text{HR}}$ is the estimated heart rate, N is the total number of windows, p is the ground-truth pulse wave, $\hat{p}$ is the estimated pulse signal, $\hat{S}$ is the power spectrum of the estimated pulse signal, f is the frequency in Hz, and $U_t(\cdot)$ is a binary mask. For the heart frequency region from f_{HR} - 0.1 Hz to f_{HR} + 0.1 Hz and its first

harmonic region from 2 * f_{HR}- 0.1 Hz to 2 * f_{HR} + 0.1 Hz, $U_t(\cdot)$ is set to be one. For other regions, $U_t(\cdot)$ is set to be zero.

4.4.4 Generating Synthetic Dark-skinned Subjects

We demonstrate the superiority of our proposed method with empirical results on UBFC-RPPG [BMB19] and VITAL [CKK20] for HR estimation using the above four metrics. The synthetic videos generated by our model can also further improve the performance of the existing data-driven PPG estimation model with reduced bias across different skin tones.

The UBFC-RPPG dataset is randomly split into a training set (32 subjects) and a validation set (10 subjects). The training set is used to jointly optimize the generator G and the rPPG estimator E. Models with minimum validation loss are selected for a cross-dataset evaluation on the VITAL videos. Some generated frames in the UBFC-RPPG validation set are illustrated in Fig. 4.5. Our generator G can successfully produce photo-realistic videos that reflect the associated underlying blood volume changes. Estimated pulse waves from the real videos and the synthetic videos are both closely aligned with the ground truth. In the frequency domain, the power spectrum of the PPG waves is also preserved with a clear peak near the gold-standard HR value.

4.4.5 Performance on UBFC-RPPG

Performance metrics of different models in the UBFC-RPPG validation set are listed in Tab. 4.1. We list the HR estimation accuracy of PRN trained with the proposed joint optimization pipeline (referred to as PRN augmented), real samples (referred to as PRN w/ Real), and synthetic samples (referred to as PRN w/ Synth). The synthetic samples are generated by our generator G through translating the real samples in the UBFC-RPPG training set when the joint optimization converges. As a comparison, we also include the performance of a state-of-the-art deep learning model 3D-CNN [TLH20] that is trained with both real and

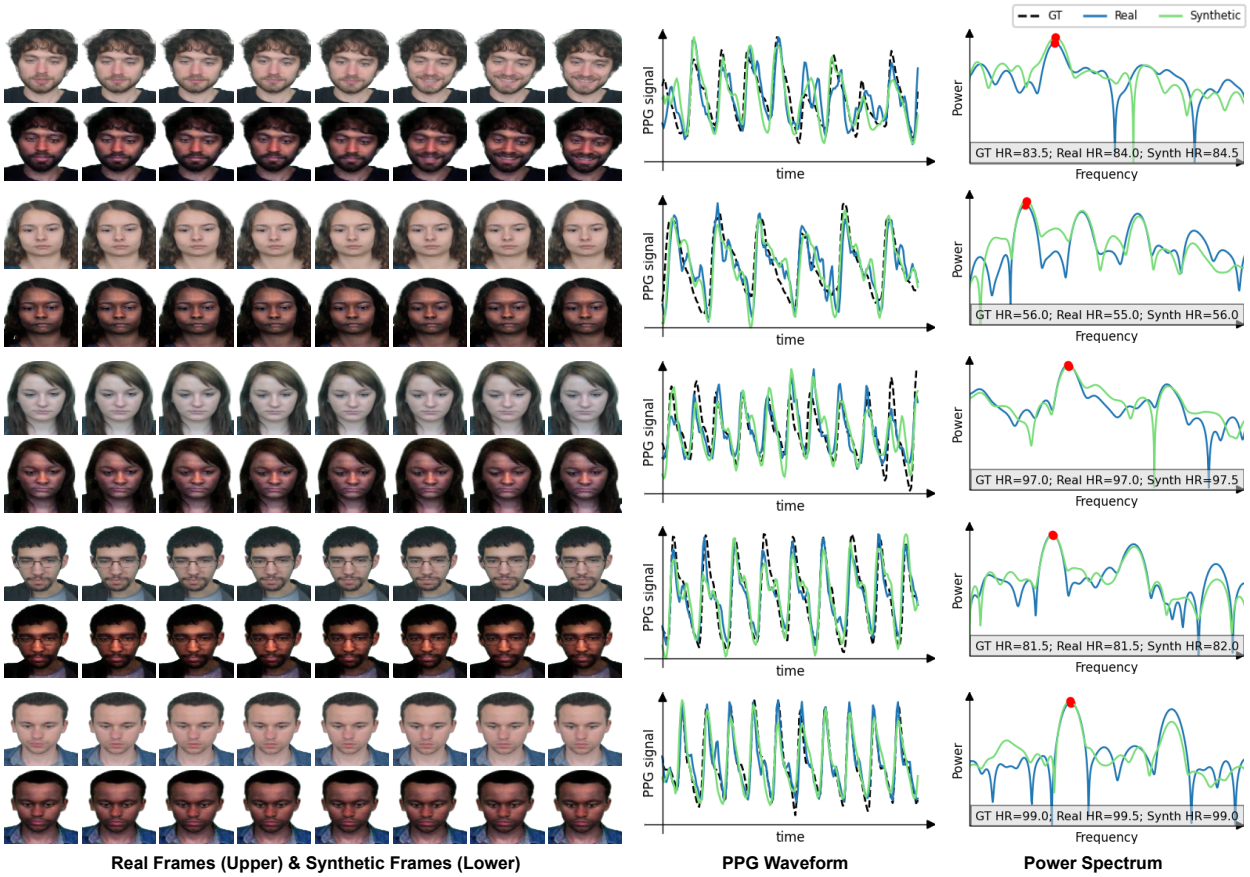


Figure 4.5: **Illustration of real frames and the corresponding synthetic frames in the UBFC-RPPG dataset.** Our proposed framework has successfully incorporated pulsatile signals when translating skin color. The estimated pulse waves from PRN exhibit a high correlation to the ground-truth waves, and the heart rates are preserved in the frequency domain.

synthetic samples (referred to as 3D-CNN w/ Real&Synth), just real samples (referred as 3D-CNN w/ Real), and just synthetic samples (referred as 3D-CNN w/ Synth). Performance of three traditional methods (POS [WBS16], CHROM [DJ13] and ICA [PMP10a]) are also provided in the table.

Notably, the proposed PRN architecture has already outperformed other rPPG estimation methods even without synthetic skin color augmentation. More specifically, the pro-

Method	MAE↓	RMSE↓	PCC↑	SNR↑
PRN augmented	**0.68**	**1.31**	0.86	5.76
PRN w/ Real	0.75	1.64	0.83	**7.91**
PRN w/ Synth	4.32	6.56	0.54	-1.93
3D-CNN [TLH20] w/ Real&Synth	0.89	1.66	**0.88**	7.74
3D-CNN [TLH20] w/ Real	1.09	1.91	0.84	7.80
3D-CNN [TLH20] w/ Synth	0.95	1.80	0.82	3.48
POS [WBS16]	3.69	5.31	0.75	3.07
CHROM [DJ13]	1.84	3.40	0.77	4.84
ICA [PMP10a]	8.28	9.82	0.55	1.45

Table 4.1: **Performance of HR estimation on UBFC-RPPG.** Boldface font represents the preferred results.

posed PRN has around 31% improvement on MAE and around 14% improvement on RMSE over the state-of-the-art 3D-CNN using real training samples. With synthetic augmentation, the performance of PRN can be further improved. PRN trained with augmentation achieves 9% improvement on MAE (from 0.75 BPM to 0.68 BPM) as compared with PRN trained with just real samples. This suggests that even for the UBFC-RPPG dataset which is overwhelmed by subjects with light skin tones, increasing the diversity of training samples is still able to enhance the performance. This finding is consistent with the recent research [LNP20, CBA22] that demonstrates the benefit of a diverse dataset.

The jointly optimized generator G can be beneficial to other data-driven models as well. We train 3D-CNN with both real and corresponding synthetic samples from G. As compared with the 3D-CNN model trained with just real samples, the 3D-CNN model trained with both real and synthetic samples exhibits 18% improvement on MAE and 13% improvement on RMSE. This further indicates that our generator has successfully learned to produce both

visually satisfying and BVP-informative facial videos, and these synthetic videos can facilitate the learning progress of the existing data-driven rPPG estimation algorithm without conducting the joint optimization process again to adapt to another new network architecture.

4.4.6 Cross-dataset Performance on VITAL

In real-world applications, it is common that the test subjects are in a different environment (e.g., illumination conditions) in contrast to the training samples. Therefore, we conduct a cross-dataset evaluation on the VITAL dataset using the models trained on the UBFC-RPPG videos. The VITAL dataset contains different subjects and is captured in an entirely different environment as compared to the UBFC-RPPG dataset. This type of cross-dataset verification can provide more visibility on the generalization capability of the models.

Similarly, we report MAE, RMSE, PCC, and SNR of various models trained with real and synthetic samples in Tab. 4.2. Since the VITAL dataset contains testing subjects of diverse skin tones with the associated Fitzpatrick scale labels (F1-6), we group the subjects into three categories, i.e., F1-2 (light skin color), F3-4 (medium skin color), and F5-6 (dark skin color), to measure the performance across different demographic groups.

PRN trained with the joint optimization pipeline exhibits significant improvement across these metrics as compared with PRN trained with just real samples. More precisely, there is 1.01 BPM reduction on MAE and 1.33 BPM reduction on RMSE for the light skin color group, 1.72 BPM reduction on MAE and 2.01 BPM reduction on RMSE for the medium skin color group, and 2.22 BPM reduction on MAE and 2.5 BPM reduction on RMSE for the dark skin color group. For all the methods, it is observed that the error of the light skin tone group is generally lower than other groups. This is probably due to that the melanin concentration of the light-skinned subjects is the least, and more light can be reflected to the camera. However, it should also be noted that models trained by both real and synthetic data have a relatively smaller performance difference among the three groups. For the dark skin

Method	F1-2		F3-4		F5-6		Overall	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
PRN augmented	2.37	3.13	**2.95**	**3.82**	**4.39**	**5.98**	**3.04**	**4.01**
PRN w/ Real	3.38	4.46	4.67	5.83	6.61	8.48	4.60	5.88
PRN w/ Synth	4.27	6.01	4.52	6.18	5.64	8.33	4.66	6.57
3D-CNN [TLH20] w/ Real&Synth	**2.32**	**3.11**	3.18	4.09	5.45	7.07	3.34	4.35
3D-CNN [TLH20] w/ Real	3.31	4.64	5.86	6.78	7.07	8.89	5.19	6.44
3D-CNN [TLH20] w/ Synth	3.88	5.23	4.68	6.07	7.81	9.88	5.04	6.56
POS [WBS16]	4.97	6.28	5.36	6.86	7.25	9.74	5.69	7.25
CHROM [DJ13]	6.51	8.92	5.01	6.38	7.83	14.56	6.14	8.99
ICA [PMP10a]	7.65	9.66	7.14	8.40	5.75	7.31	7.04	8.63
	F1-2		F3-4		F5-6		Overall	
	PCC↑	SNR↑	PCC↑	SNR↑	PCC↑	SNR↑	PCC↑	SNR↑
PRN augmented	0.40	3.45	0.63	**5.73**	**0.30**	**-3.38**	**0.48**	**3.02**
PRN (w/ Real)	0.36	0.32	0.50	0.03	0.08	-7.00	0.36	-1.32
PRN (w/ Synth)	0.29	-0.64	0.42	-0.44	0.11	-6.35	0.31	-1.74
3D-CNN [TLH20] (w/ Real&Synth)	**0.42**	**3.96**	**0.65**	5.21	0.17	-4.84	0.47	2.68
3D-CNN [TLH20] (w/ Real)	0.30	-0.61	0.48	-1.26	0.11	-8.26	0.34	-2.47
3D-CNN [TLH20] (w/ Synth)	0.07	-2.04	0.38	-1.34	0.10	-6.38	0.21	-2.64
POS [WBS16]	0.26	-2.22	0.42	-1.04	0.27	-5.59	0.33	-2.41
CHROM [DJ13]	0.15	-2.14	0.46	-1.11	-0.10	-5.53	0.23	-2.40
ICA [PMP10a]	0.24	-2.06	0.32	-1.73	0.06	-5.04	0.23	-2.53

Table 4.2: **The proposed method shows an improved HR estimation accuracy on the VITAL dataset.** Boldface font denotes the preferred results.

color groups, PRN trained with synthetic data shows lower estimation errors as compared with real data, and the errors are reversed for the light skin color group. This validates the fact that data-driven rPPG estimation models are heavily impacted by the skin color distribution of training samples, and it is critical to create a diverse and balanced training set for generalizability and real-world deployment of rPPG algorithms.

To assess the cross-dataset generalization capability of synthetic videos, we also evaluate 3D-CNN trained on real and synthetic samples from UBFC-RPPG on the VITAL dataset. Similar improvement can be observed in the 3D-CNN model, where 3D-CNN trained with both real and synthetic samples outperforms the model trained on only real or only synthetic samples. This supports that our generator can generate synthetic videos that can accurately reflect subtle color variations due to blood volume changes, instead of simply overfitting the UBFC-RPPG training samples. Our synthetic data can therefore serve as a bio-realistic augmentation to the real samples.

POS [WBS16], CHROM [DJ13] and ICA [PMP10a] show relatively large HR estimation errors as compared with the data-driven models, where their MAEs on the light skin color group is usually larger than 4 BPM. Their MAEs are even higher for other groups. Unlike the end-to-end rPPG estimation networks, these conventional methods usually require preprocessing steps which may diminish the subtle color changes on the face and degrade the performance. Besides, these models need to average the pixel intensities over the skin region, and this might be a sub-optimal solution since skin pixels at different facial regions can contribute differently to the pulse signals.

The cross-dataset experiment indicates that the improvement of our proposed framework is more substantial as compared with the intra-dataset evaluation where all the samples are obtained within the same environment. This suggests that synthetic videos can provide more significant benefits by diversifying the training samples when there exist some data distribution shifts between real training and testing videos. This finding is also consistent with the observation for ray-tracing based augmentation method [MHW20]. Synthetic augmentation techniques thus become particularly effective for cross-domain learning and can improve the generalization capability of HR estimation for real-world applications.

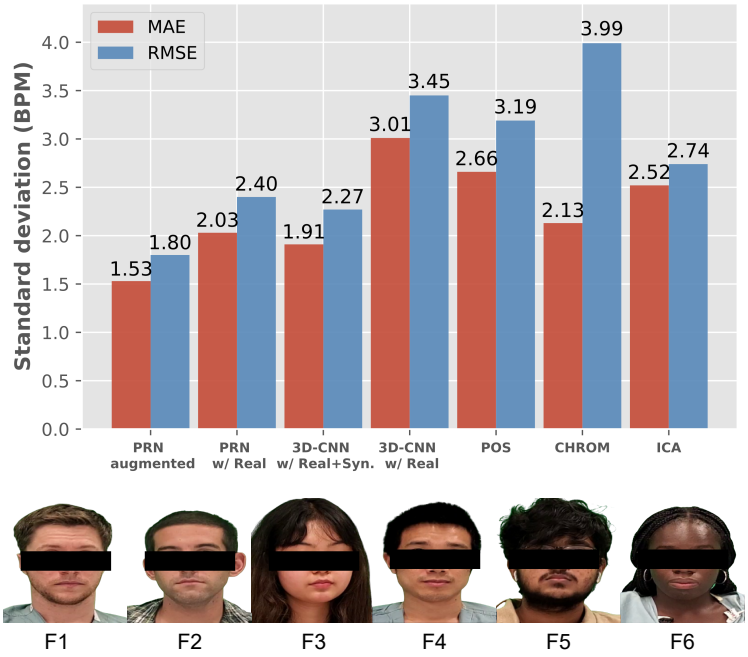


Figure 4.6: **Synthetic dark-skinned videos can help to reduce bias in HR estimation.** The augmented PRN and the 3D-CNN [TLH20] trained on both real and synthetic videos show a reduced standard deviation on MAE and RMSE across Fitzpatrick scales F1-6 in the VITAL dataset.

4.4.7 Bias Mitigation

It is critical for an algorithm to have consistent performance across different demographic groups in real-world medical deployment. To quantify the performance gap for each group, we use the standard deviation of MAE and RMSE for each Fitzpatrick scale as the measurement. This measurement has also been used in some prior work [MHW20, YAA20]. The standard deviation for each method in the VITAL dataset is illustrated in Fig. 4.6, together with a sample portrait for each skin scale from F1 to F6. The conventional POS method exhibits large variation (MAE: 2.66 BPM, RMSE: 3.19 BPM) across different Fitzpatrick scales, while the jointly optimized PRN shows the lowest bias (MAE: 1.53 BPM, RMSE: 1.80 BPM) as compared with all the conventional methods. In contrast to PRN trained with just

real samples (MAE: 2.03 BPM), the augmented training offers a 25% improvement of bias mitigation among different groups while simultaneously improving the overall performance of all the groups. This suggests our joint training framework can provide a more desirable trade-off between performance and bias. For 3D-CNN, the standard deviations for MAE and RMSE are also reduced by adding the synthetic samples into the training set. We attribute this improvement to the more diverse and balanced dataset augmented by our generator.

4.5 Discussion and Limitations

Our work has made an attempt to tackle bias in rPPG. The lack of dark-skinned subjects in existing rPPG datasets (MMSE-HR, AFRL, and UBFC-RPPG have roughly 10%, 0%, and 5% dark-skinned subjects) has produced unwanted bias against some underrepresented groups, and there exist several practical constraints towards collecting a large-scale balanced dataset for rPPG. To address this issue, an attempt is proposed to translate facial frames from light-skinned subjects to dark skin tones while preserving the subtle color variations corresponding to the pulsatile signals. The jointly optimized rPPG estimator can outperform the existing state-of-the-art methods with reduced estimation bias across different demographic groups. More specifically, PRN trained with augmentation has around 38% reduction in MAE for the dark-skinned group along with 49% improvement on bias mitigation in the VITAL dataset, as compared with 3D-CNN [TLH20] trained with just real samples. Our generated synthetic videos maintain both photo-realistic and bio-realistic features and can be directly used to improve the performance of the existing deep learning rPPG estimation model.

Video synthesis, such as DeepFake, has raised public concerns in the community [ML21]. Over half a decade, these 'fake' videos generated by deep learning have been used for face manipulation, and the malicious usage has drawn a lot of social attention. We demonstrate a positive example that these bio-realistic 'fake' videos can also be utilized for the purpose of

social good. Our synthetic videos are capable of reducing both HR estimation error and bias for rPPG models and further facilitate the development of remote healthcare. We hope our framework can act as a tool to address some social issues in the existing medical applications.

We now discuss a few limitations of our approach. Our current pipeline is an initial attempt that focuses on skin color translation, and all the remaining factors (e.g., pulse signals, body motion, and other facial attributes) are directly copied from the original videos. To maximize the benefit of synthetic augmentation, it is also important to extend the generation framework to incorporate arbitrary facial attributes and pulse waves. We hope the method presented in our work could inspire following work on synthetic generation for a more diverse dataset. Besides, it should also be noted that the generated frames are limited by a fixed resolution at 80×80. Future work may produce solutions to generate frames at arbitrary pixel resolution to fit the requirements of various subsequent rPPG estimation models without frame size interpolation. The primary goal of our work is to overcome the shortness of real dark-skinned subjects by synthetic generation. Therefore, the current framework is designed based on Caucasian-to-African translation. Future work may extend this to other appropriate racial group(s) to further diversify the training data. Our framework relies on a generator designed based on 3D convolutions, where its output is not directly supervised by videos from real dark-skinned subjects. While the improved heart rate estimation results support the effectiveness of the proposed solution, inductively generalizing claims in our work of reducing bias need to be validated in much larger-scale clinical trials than what is possible in an academic work introducing a new method.

In our work, we used existing metrics to evaluate rPPG quality, such as standard waveform measures of MAE and RMSE. These metrics were carefully chosen so they are regressable against previous rPPG papers. It could be that the metrics could themselves be biased (e.g., if the rPPG waveform has a unique shape amongst demographics and/or if the synthetic data has an unusual shape). Ultimately, we felt more comfortable using the same error metrics used in previous works, to aid in comparisons. Identifying biases in a metric

and/or proposing solutions requires thought and experiment, particularly when the context involves fairness. An option for future work is to evaluate if there is possibly a better metric for the rPPG problem.

4.6 Conclusion

To conclude, we perform appearance transfer while retaining the subtle transient characteristics of realistic blood flow. During training, we demonstrate that heart rate estimation can be improved in *both* performance and equity. Other than heart rate estimation, we hope that future work can apply physiologically-sound appearance transfer to other vital signs, such as blood pressure, blood oxygen saturation, and respiration rate.

Ethics Statement: We envision positive benefits of bio-realistic avatars, as a way to expand training datasets for medical instruments, like remote vital sign monitors. We condemn the use of this technique to fool DeepFake catchers.

CHAPTER 5

Synthetic Generation of Face Videos with Plethysmograph Physiology

5.1 Introduction

Photoplethysmography (PPG) is an optical technique that measures vital signs such as Blood Volume Pulse (BVP) by detecting the light reflected or transmitted through the skin. Remote Photoplethysmography (rPPG) based on camera videos has several advantages over the conventional PPG methods. It is non-contact thus allowing for a wide range of applications in neonatal monitoring [KPA21, VCJ19]. It causes no skin irritation and prevents the risk of developing into infection for those whose skins are fragile and sensitive to the adhesive sensing electrodes. As cameras are ubiquitous in electronic devices nowadays (such as smartphones and laptops), rPPG can be applied for telemedicine with patients at home and no equipment setup is needed [APM21]. Camera-based rPPG techniques have also been used in other applications such as driver monitoring [NMM18] and face anti-spoofing [LYZ16].

Traditional rPPG methods either use Blind Source Separation (BSS) [PMP10a, PMP10b, LRK11] or models based on skin reflectance [WBS16, DJ13, KVS15] 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 their expressiveness and flexibility [CM18, YPL19, LFP20, NSH19, NYH20, LHZ21]. CNNs learn the mapping between the pulse signal and the color variations with end-to-end supervised training on the labeled

Dataset	# Subjects	# Videos	Demo. diversity	Orig. Videos Free Avail.
AFRL [EBM14]	25	300	✗	✓
MMSE-HR [ZGW16]	40	102	✗	✗
UBFC-rPPG [BMB19]	42	42	✗	✓
UBFC-Phys [MBD21]	56	168	✗	✓
VIPL-HR [NHS18]	107	3130	✗	✓
Dasari *et al.* [DPJ21]	140	140	✗	✗
Our synthetic method	480	480	High	✓

Table 5.1: **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.

dataset, thus achieving state-of-the-art performance on vital sign detection. However, the performance of data-driven rPPG networks hinges on the quality of the dataset [NMV20b].

There are some efforts (as shown in Tab. 5.1) on collecting a large rPPG dataset for better physiological measurement. Nonetheless, there exist 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 [BWK22], a diverse rPPG dataset may not be accessible for some countries/regions due to the 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 a pressing need for the concept of 'digital patients': physiologically accurate graphical renders that may assist the development of algorithms and techniques to improve 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

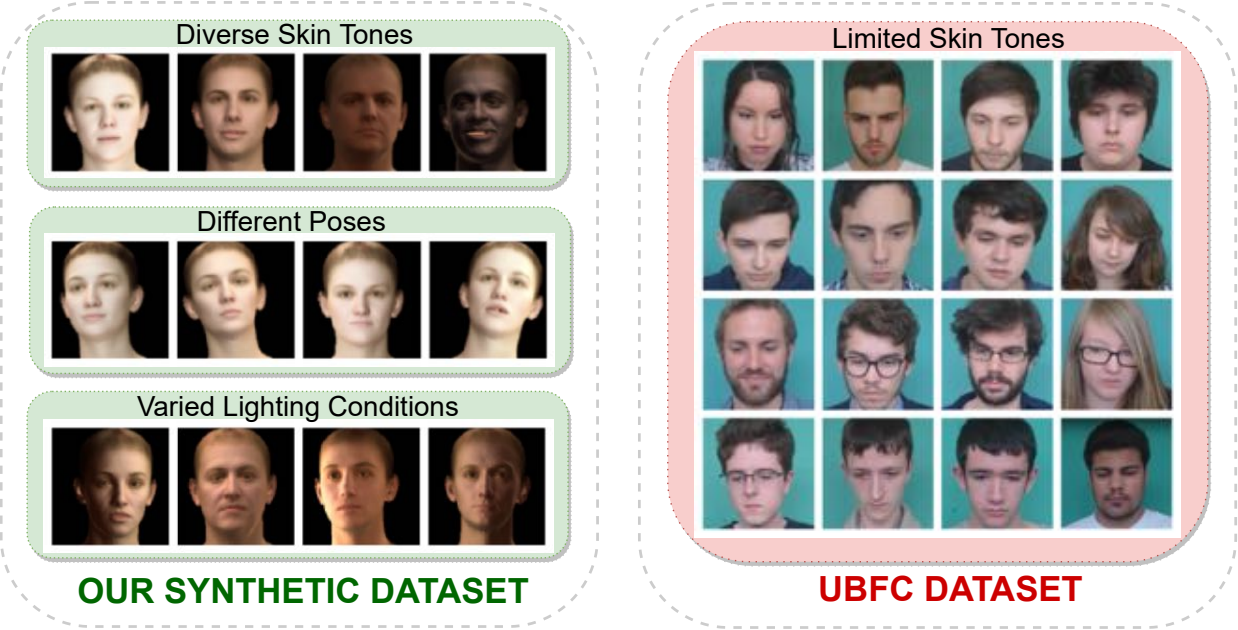


Figure 5.1: **Our proposed scalable model can generate synthetic rPPG videos with diverse attributes, such as poses, skin tones, and lighting conditions.** In contrast, existing real datasets (e.g., UBFC) only contain limited races.

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) [FFB21]. Our model can generate rPPG videos with large variations of various attributes, such as facial appearance and expression, head motions, and environmental lighting as shown in Fig. 5.1.

We summarize our contributions as follows:

- 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.

5.2 Related Work

5.2.1 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 [LRK11] 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) [PMP10a, MGP14] 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) [DJ13] is proposed to extract the blood volume pulse by assuming a standardized skin color to white-balance the image and then linearly combine the chrominance signals. Plane Orthogonal to Skin-tone (POS) [WBS16] 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 [CM18, YPL19, LFP20, NSH19, NYH20, LHZ21, BWK22]. 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 [YPL19] and PRN [BWK22] in our experiments to demonstrate the performance of the rPPG models on both real and synthetic datasets.

5.2.2 Real rPPG Datasets

There are many efforts on collecting real datasets for more accurate physiological sensing [EBM14, ZGW16, BMB19, MBD21, NHS18, DPJ21]. However, these datasets are usually very limited in the number of subject participants and also biased toward certain demographic groups. Some work includes subjects with darker skin types, but the number is still very limited [ZGW16]. Making machine learning methods equitable is of increasing interest in the medical domain [ZS21, Kad21]. 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.* [DPJ21] proposed a dataset that only contains dark skin tones. However, only the color space values of skin regions of interest are shared instead of the actual videos. The current best-performing deep learning algorithms require sizeable input data. The rPPG model trained on such a biased 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 society would not cause biases against certain groups that are underrepresented. Our real dataset represents a first step towards filling this gap.

5.2.3 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 [MHW20, TLH20,

BWK22, NMV21b]. Mcduff *et al.* [MHW20] 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 the 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 [BWK22] 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 [TLH20], 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.

5.3 Methods

In this section, we propose a scalable method that can generate a synthetic dataset with any given reference image and target rPPG signal in Sec. 5.3.1. The generated videos can be used to train the state-of-the-art rPPG networks, which we introduce in Sec. 5.3.2.

5.3.1 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. 5.2 for an illustration of the entire synthetic generation pipeline.

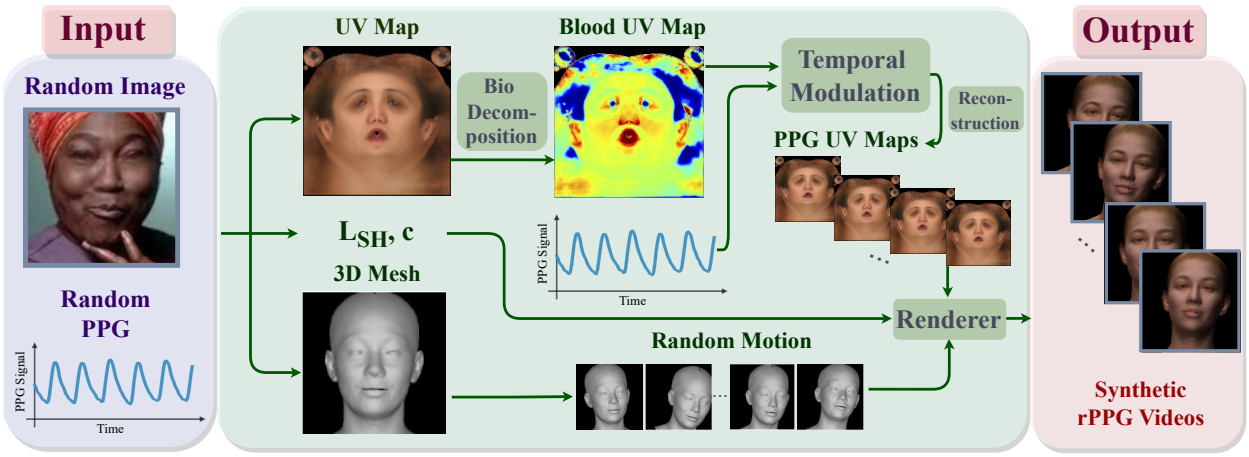


Figure 5.2: **Pipeline of our cross-modal synthetic generation model that can generate rPPG face videos given any face image and target rPPG signal as input.** The input image is encoded into UV albedo map, 3D mesh, illumination model L_{SH} and camera model c. We then decompose the UV albedo map into a 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.

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 [FFB21] to predict subject-specific albedo, shape, pose, and lighting parameters from an image. In detail, it uses a statistical 3D head model FLAME [LBB17] 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) [PKA09] 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 the illumination model as the Spherical Harmonics (SH) [RH01], the shaded face

image can be represented as the following equation:

$$B\left(\boldsymbol{\alpha}, \mathbf{l}, N_{uv}\right)_{i,j} = A(\boldsymbol{\alpha})_{i,j} \odot \sum_{k=1}^{9} \mathbf{l}_k H_k\left(N_{i,j}\right), \tag{5.1}$$

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 $\mathcal{R}(\cdot)$:

$$I_r = \mathcal{R}(M, B, \mathbf{c}). \tag{5.2}$$

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 show how this biophysically interpretable manipulation is achieved in the following sections.

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}. Following 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 [AS19a]:

$$A_c = \int_{\lambda} E(\lambda) R(f_{\text{mel}}, f_{\text{blood}}, \lambda) S_c(\lambda) d\lambda. \tag{5.3}$$

An optical skin reflectance model [AS19b] 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}:

$$R\left(f_{\text{mel}}, f_{\text{blood}}, \lambda\right) = T_{\text{epidermis}}\ \left(f_{\text{mel}}, \lambda\right)^2 R_{\text{dermis}}\ \left(f_{\text{blood}}, \lambda\right). \tag{5.4}$$

The transmittance in the epidermis is modeled by Lambert-Beer law [PC04] as light not absorbed by the melanin in this layer is propagated to the dermis [AS17]:

$$T_{\text{epidermis}}(f_{\text{mel}}, \lambda) = e^{-\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda)}, \tag{5.5}$$

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

$$\mu_{a.\text{epidermis}}(f_{\text{mel}}, \lambda) = f_{\text{mel}}\mu_{a.\text{mel}}(\lambda) + (1 - f_{\text{mel}})\mu_{\text{skinbaseline}}(\lambda), \tag{5.6}$$

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

The reflectance in the dermis can be modeled using the Kubelka-Munk theory [INN07], and the proportion of light remitted from a layer is given by [AS17]:

$$R_{\text{dermis}}\,(f_{\text{blood}}, \lambda) = \frac{(1 - \beta^2)\left(e^{Kd_{\text{pd}}} - e^{-Kd_{\text{pd}}}\right)}{(1 + \beta^2)\,e^{Kd_{\text{pd}}} - (1 - \beta)^2 e^{-Kd_{\text{pd}}}}, \tag{5.7}$$

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. (5.7).

Biophysical decomposition and variation of UV albedo map: With the light transport theory of the skin, we follow a physics-based learning framework (BioFaceNet [AS19a]) 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 the 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. (5.3).

Different from the previous work [AS19a], 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:

$$\mathcal{L} = w_1 \mathcal{L}_{\text{appearance}} + w_2 \mathcal{L}_{\text{CameraPrior}}, \tag{5.8}$$

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 the 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 the UV map. The final image is rendered using the UV map combined with illumination and camera model according to Eq. (5.2).

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 [WDH19]. It is categorized according to ethnicity (i.e., Caucasian, Indian, Asian, and African). We use these images as reference images for generating the synthetic videos as shown in Fig. 5.2.

PPG recordings: To synthesize videos of a given input PPG signal, we use PPG waveforms recordings from BIDMC PPG and Respiration Dataset [PJC16]. It contains 53 8-minute contact PPG recordings with a 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.

5.3.2 Physiological Measurement Networks

We use two state-of-the-art deep rPPG networks PhysNet [YPL19] and PRN [BWK22] 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 connections 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 $\hat{p}$:

$$L_{ppg}(p,\hat{p}) = 1- \frac{T\sum_i p_i\hat{p}_i - \sum_i p_i \sum_i \hat{p}_i}{\sqrt{\left(T\sum_i p_i^2 - \left(\sum_i p_i\right)^2\right)\left(T\sum_i \hat{p}_i^2 - \left(\sum_i \hat{p}_i\right)^2\right)}}, \tag{5.9}$$

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 papers, 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 [PGM19b] and trained on Nvidia V100 GPU.

5.4 Experiments

In this section, we introduce the datasets we use for the experiments and evaluation protocol in Sec. 5.4.1. We report and analyze the experimental results for our real dataset in Sec. 5.4.2 and UBFC-rPPG dataset in Sec. 5.4.3.

5.4.1 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 groups. The Fitzpatrick (FP) skin type scale [Fit88] 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 503 videos in total. All the videos in our dataset are uncompressed and synchronized with the ground truth heart rate.

Figure 5.3 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 their finger and look into the camera. Both the camera and the oximeter are connected to a laptop to get synchronous data.

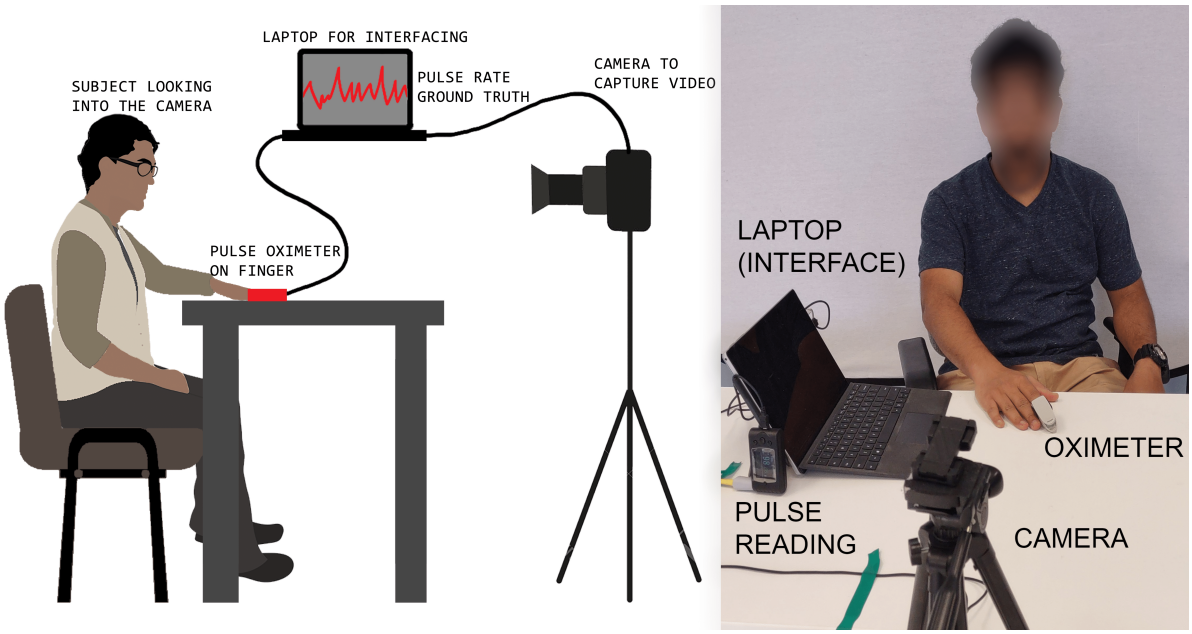


Figure 5.3: **Experimental setup of data collection.** The subject wears an oximeter on their finger and sits looking directly into the camera. The camera and the oximeter are connected to a laptop to get synchronous video and ground-truth pulse reading. Face blurred to preserve anonymity.

UBFC-rPPG [BMB19]: 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), Pearson's Correlation Coefficient (PCC) and Signal-to-Noise Ratio (SNR). Pearson's Correlation Coefficient (PCC) and Signal-to-Noise Ratio (SNR) is defined as in [NMV20a].

For traditional baseline methods POS, CHROM, and ICA, we use the iPhys toolbox [MB19]

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 of 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.

5.4.2 Performance on UCLA-rPPG

For the study of our work, we split the subjects into three skin tone groups based on the Fitzpatrick skin type [Fit88]. 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 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. 5.2. In general, models trained with both real and synthetic data perform consistently better than using real data alone on all skin tones across 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. 5.3.1. 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 take the average of all the pixels and ignore the inhomogeneous spatial contribution of the pixels to pulsatile signals.

Method	F1-2		F3-4		F5-6		Overall	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
PhysNet [YPL19] w/ Real&Synth	**0.54**	0.84	0.38	0.70	**1.55**	**2.17**	**0.71**	**1.10**
PhysNet [YPL19] w/ Real	0.81	1.21	0.43	0.77	2.61	3.34	1.06	1.51
PhysNet [YPL19] w/ Synth	1.06	1.52	1.16	1.66	4.96	6.20	2.06	2.73
PRN [BWK22] w/ Real&Synth	**0.54**	**0.79**	**0.36**	**0.65**	3.41	4.09	1.15	1.53
PRN [BWK22] w/ Real	0.65	1.02	0.40	0.71	4.35	5.26	1.43	1.90
PRN [BWK22] w/ Synth	1.47	2.00	0.63	1.07	8.89	9.88	2.87	3.47
POS [WBS16]	3.40	4.34	3.03	3.98	8.07	10.23	4.27	5.49
CHROM [DJ13]	4.06	5.11	3.99	5.25	7.45	9.74	4.79	6.22
ICA [PMP10a]	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 [YPL19] w/ Real&Synth	**0.84**	**14.40**	**0.80**	**17.11**	**0.60**	**9.19**	**0.76**	**14.45**
PhysNet [YPL19] w/ Real	0.81	13.13	0.77	15.83	0.59	6.54	0.74	12.84
PhysNet [YPL19] w/ Synth	0.74	7.19	0.64	6.11	0.23	-3.33	0.57	4.10
PRN [BWK22] w/ Real&Synth	0.81	12.24	0.79	14.61	0.57	4.84	0.74	11.59
PRN [BWK22] w/ Real	0.77	10.73	0.77	13.22	0.48	2.38	0.70	9.91
PRN [BWK22] w/ Synth	0.69	5.14	0.67	5.27	0.21	-5.81	0.56	2.53
POS [WBS16]	0.50	-0.30	0.42	-0.09	0.27	-5.38	0.41	-1.34
CHROM [DJ13]	0.41	-1.81	0.31	-1.60	0.26	-5.31	0.33	-2.49
ICA [PMP10a]	0.45	-0.60	0.38	-0.19	0.27	-5.24	0.37	-1.44

Table 5.2: **Heart rate estimation results on our real dataset UCLA-rPPG show that both PhysNet and PRN trained with real and synthetic datasets perform consistently better than the models trained with only real data.** The improved performance shows the benefit of the synthetic video dataset we generate.

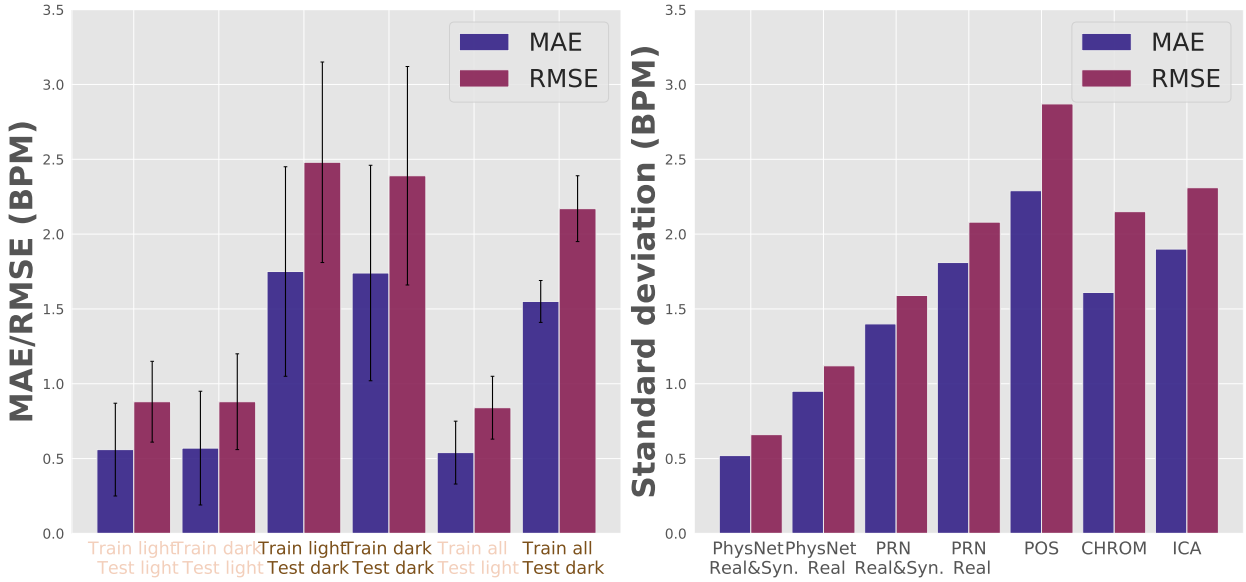


Figure 5.4: **Left: Ablation study.** The model pretrained with all synthetic dataset outperforms these pretrained on either light or dark skin tones alone. **Right: Bias mitigation.** The standard deviation of MAE and RMSE of the deep rPPG models trained with real and synthetic datasets are smaller than real data alone and the traditional models.

Bias mitigation: To evaluate the bias 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. 5.4, we can see the standard deviation of PhysNet with both real and synthetic datasets is the smallest, and the MAE disparity among all the three groups is 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 pretrain 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. 5.4, we can see the model with the

Method	MAE ↓	RMSE ↓	PCC ↑	SNR ↑
PhysNet [YPL19] w/ Real&Synth	0.90	1.80	**0.84**	6.28
PhysNet [YPL19] w/ Real	1.42	2.74	0.78	5.64
PhysNet [YPL19] w/ Synth	**0.84**	**1.76**	0.83	**6.70**
PRN [BWK22] w/ Real&Synth	1.15	2.38	0.82	5.36
PRN [BWK22] w/ Real	2.36	4.21	0.66	-1.24
PRN [BWK22] w/ Synth	1.09	1.99	0.83	3.00
POS [WBS16]	3.69	5.31	0.75	3.07
CHROM [DJ13]	1.84	3.40	0.77	4.84
ICA [PMP10a]	8.28	9.82	0.55	1.45

Table 5.3: **Performance of HR estimation on UBFC-rPPG shows the superiority of the synthetic datasets.** Boldface font represents the preferred results.

pretrained rPPG network on diverse races is consistently better than these on a single race. The improvement is more obvious on the dark skin tone test set. This demonstrates the benefits of a diverse synthetic dataset.

5.4.3 Performance on UBFC-rPPG

We use the model with the best performance on our real dataset to test them on UBFC-rPPG dataset [BMB19] 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 UBFC-rPPG. All the results with four evaluation metrics are reported in Tab. 5.3. 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 the 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

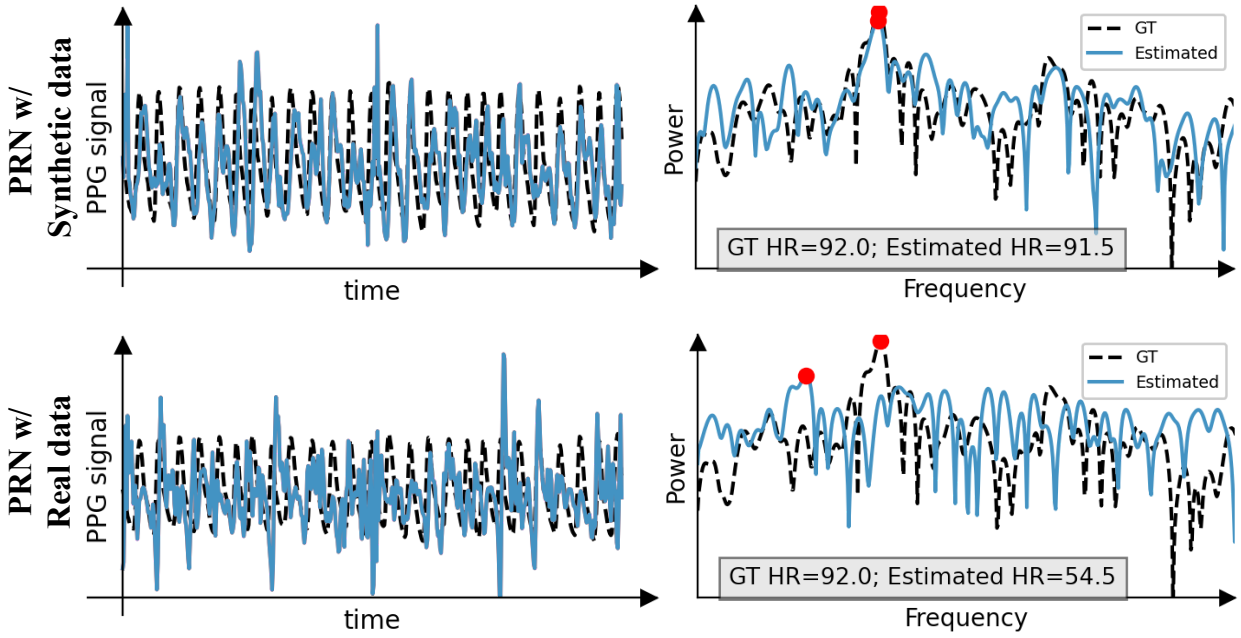


Figure 5.5: **The example shows that PRN [BWK22] trained with synthetic data (above) generalizes better than PRN trained with real data (bottom) on UBFC-rPPG dataset.** The waves are more aligned with the ground-truth PPG wave (dashed black line) and the power spectrum plot is also more consistent with the ground truth for the PRN trained with synthetic data.

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 settings, such as lighting. We also give a qualitative study in Fig. 5.5 that shows that the rPPG waves extracted using our synthetic dataset resemble more closely to the ground truth than that using real dataset. As a result, it gives more accurate heart rate estimation.

5.4.4 Visualization

As shown in Fig. 5.6, our model can successfully produce synthetic avatar videos that reflect the associated underlying blood volume changes. Estimated pulse waves from the synthetic

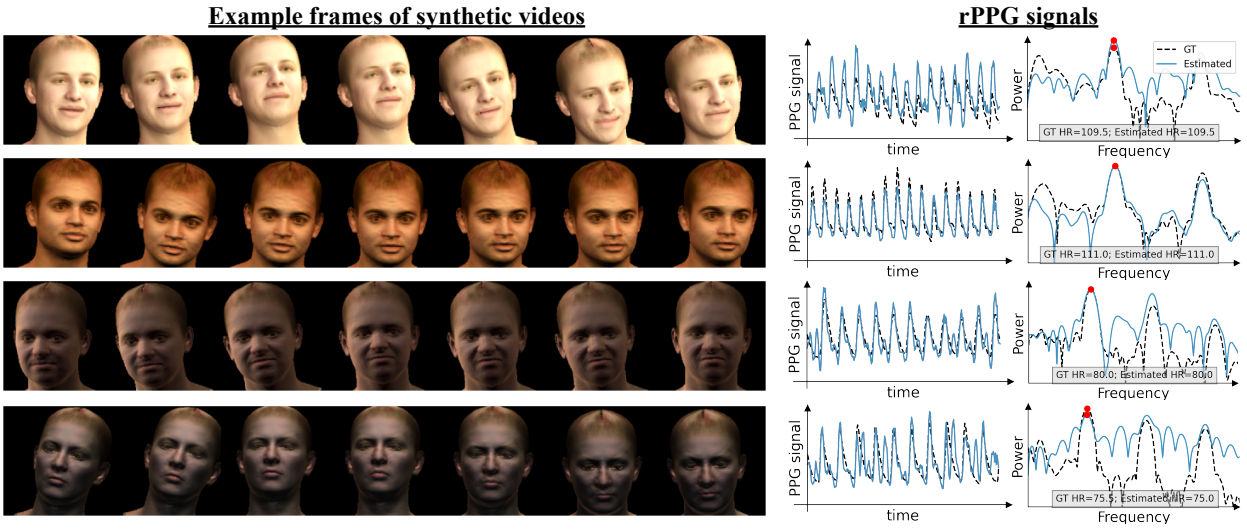


Figure 5.6: **Illustration of example frames of our generated synthetic videos.** Our proposed framework has successfully incorporated PPG signals into the reference image. The estimated pulse waves from PRN for generated synthetic videos are highly correlated to the ground-truth waves, and the heart rates are preserved as shown in the power spectrum plot.

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.

5.5 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 our work. However, it is found in [NMV20a] that the background can be utilized for better pulsatile signals extraction. In addition, 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: The novelty of our work is to generate synthetic face videos that are physiologically consistent with the heartbeat, and we hope it can be a tool to address some social issues, such as biases 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 the UCLA-rPPG dataset we collect that contains diverse skin tones and the UBFC-rPPG dataset. The experimental results show that the synthetic dataset can improve the performance on both datasets and help reduce the bias among different demographic groups.

CHAPTER 6

Not Just Streaks: Towards Ground Truth for Single Image Deraining

6.1 Introduction

Single-image deraining aims to remove degradations induced by rain from images. Restoring rainy images not only improves their aesthetic properties but also supports the reuse of abundant publicly available pretrained models across computer vision tasks. Top performing methods use deep networks, but suffer from a common issue: it is not possible to obtain ideal real ground-truth pairs of rain and clean images. The same scene, in the same space and time, cannot be observed both with and without rain. To overcome this, deep learning based rain removal relies on synthetic data.

The use of synthetic data in deraining is prevalent [FHZ17, HFZ19, LCT19, LTG16, YTF17, ZP18, ZSP19]. However, current rain simulators cannot model all the complex effects of rain, which leads to unwanted artifacts when applying models trained on them to real-world rainy scenes. For instance, a number of synthetic methods add *rain streaks* to clean images to generate the pair [FHZ17, LTG16, YTF17, ZP18, ZSP19], but rain does not only manifest as streaks: If raindrops are further away, the streaks meld together, creating *rain accumulation*, or *veiling* effects, which are exceedingly difficult to simulate. A further challenge with synthetic data is that results on real test data can only be evaluated qualitatively, for no real paired ground truth exists.

Realizing these limitations of synthetic data, we tackle the problem from another angle by

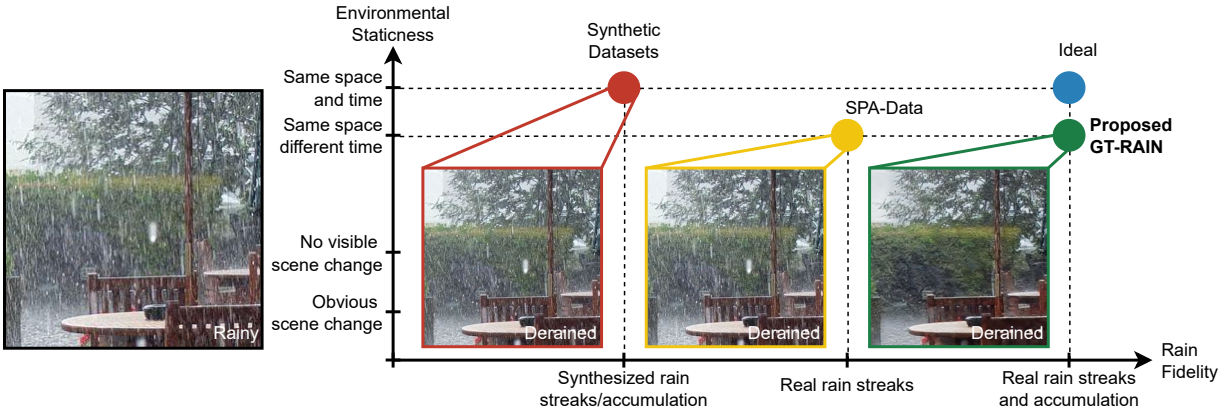


Figure 6.1: **The points above depict datasets and their corresponding outputs from models trained on them.** These outputs come from a real rain image from the Internet. Our opinion* is that GT-RAIN can be the right dataset for the deraining community to use because it has a smaller domain gap to the ideal ground truth. * Why an asterisk? The asterisk emphasizes that this is an "opinion". It is impossible to quantify the domain gap because collecting true real data is infeasible. To date, deraining is largely a viewer's imagination of what the derained scene should look like. Therefore, we present the derained images above and leave it to the viewer to judge the gap. Additionally, GT-RAIN can be used in complement with the litany of synthetic datasets [FHZ17, HFZ19, LCT19, LTG16, YTF17, ZP18, ZSP19], as illustrated in Tab. 6.4.

relaxing the concept of ideal ground truth to a sufficiently short time window (see Fig. 6.1). We decide to conduct the experiment of obtaining short time interval paired data, particularly in light of the timely growth and diversity of landscape YouTube live streams. We strictly filter such videos with objective criteria on illumination shifts, camera motions, and motion artifacts. Further correction algorithms are applied for subtle variations, such as slight movements of foliage. We call this dataset GT-RAIN, as it is a first attempt to provide real paired data for deraining. Although our dataset relies on streamers, YouTube's fair use policy allows its release to the academic community.

Defining "real, paired ground truth": Clearly, obtaining real, paired ground truth data by capturing a rain and rain-free image pair at the exact same space and time is not feasible. However, the dehazing community has accepted several test sets [AAS19, AAT20, AAT18a, AAT18b] following these guidelines as a satisfactory replacement for evaluation purposes:

- A pair of degraded and clean images is captured as real photos at two different timestamps;
- Illumination shifts are limited by capturing data on cloudy days;
- The camera configuration remains identical while capturing the degraded and clean images.

We produce the static pairs in GT-RAIN by following the above criterion set forth by the dehazing community while enforcing a stricter set of rules on the sky and local motion. More importantly, as a step closer towards obtaining real ground truth pairs, we capture natural weather effects instead, which address problems of scale and variability that inherently come with simulating weather through man-made methods. In the results of the proposed method, we not only see quantitative and qualitative improvements but also showcase a unique ability to handle diverse rain physics that was not previously handled by synthetic data.

Contributions: In summary, we make the following contributions:

- We propose a real-world paired dataset: GT-RAIN. The dataset captures *real* rain phenomena, from rain streaks to accumulation under various rainfall conditions, to bridge the domain gap that is too complex to be modeled by synthetic [FHZ17, HFZ19, LCT19, LTG16, YTF17, ZP18, ZSP19] and semi-real [WYX19] datasets.
- We introduce an avenue for the deraining community to now have standardized quantitative and qualitative evaluations. Previous evaluations were quantifiable only with respect to simulations.

- We propose a framework to reconstruct the underlying scene by learning representations robust to the rain phenomena via a rain-robust loss function. Our approach outperforms the state of the art [ZAK21] by 12.1% PSNR on average for deraining real images.

6.2 Related Work

6.2.1 Rain Physics

Raindrops exhibit diverse physical properties while falling, and many experimental studies have been conducted to investigate them, e.g, equilibrium shape [BC87], size [MP48], terminal velocity [FD69, GK49], spatial distribution [Man93], and temporal distribution [ZLQ06]. A mixture of these distinct properties transforms the photometry of a raindrop into a complex mapping of the environmental radiance which considers refraction, specular reflection, and internal reflection [GN07]:

$$L(\hat{n}) = L_r(\hat{n}) + L_s(\hat{n}) + L_p(\hat{n}), \tag{6.1}$$

where $L(\hat{n})$ is the radiance at a point on the raindrop surface with normal $\hat{n}$, $L_r(\cdot)$ is the radiance of the refracted ray, $L_s(\cdot)$ is the radiance of the specularly reflected ray, and $L_p(\cdot)$ is the radiance of the internally reflected ray. In real images, the appearance of rain streaks is also affected by motion blur and background intensities. Moreover, the dense rain accumulation results in sophisticated veiling effects. Interactions of these complex phenomena make it challenging to simulate realistic rain effects. Until GT-RAIN, previous works [GSJ21, HZW21, JWY20, LCT19, WXZ20, WYX19, ZAK21] have relied heavily on simulated rain and are limited by the sim2real gap.

6.2.2 Deraining Datasets

Most data-driven deraining models require paired rainy and clean, rain-free ground-truth images for training. Due to the difficulty of collecting real paired samples, previous works focus on synthetic datasets, such as Rain12 [LTG16], Rain100L [YTF17], Rain100H [YTF17], Rain800 [ZSP19], Rain12000 [ZP18], Rain14000 [FHZ17], NYU-Rain [LCT19], Outdoor-Rain [LCT19], and RainCityscapes [HFZ19]. Even though synthetic images from these datasets incorporate some physical characteristics of real rain, significant gaps still exist between synthetic and real data [YTW20]. More recently, a "paired" dataset with real rainy images (SPA-Data) was proposed in [WYX19]. However, their "ground-truth" images are in fact a product of a video-based deraining method – synthesized based on the temporal motions of raindrops which may introduce artifacts and blurriness; moreover, the associated rain accumulation and veiling effects are not considered. In contrast, we collect pairs of real-world rainy and clean ground-truth images by enforcing rigorous selection criteria to minimize environmental variations. To the best of our knowledge, our dataset is the first large-scale dataset with real paired data. Please refer to Tab. 6.1 for a detailed comparison of the deraining datasets.

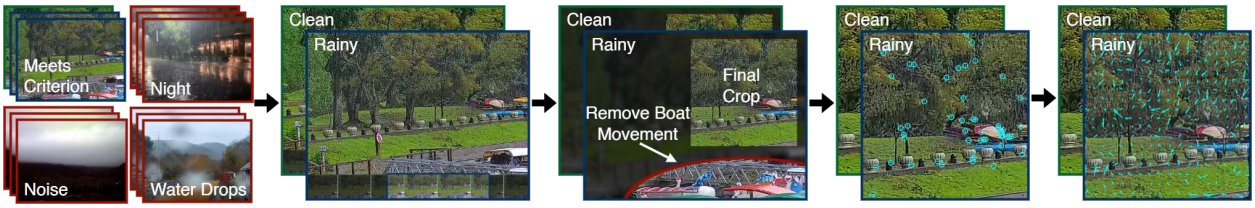
6.2.3 Single-image Deraining

Previous methods used model-based solutions to derain [CH13, JHZ18, LTG16, LXJ15]. More recently, deep learning based methods have seen increasing popularity and progress [FHD17, GSJ21, HZW21, JWY20, LCT19, PLS18, RSZ20, WXZ20, WYX19, YTF17, ZAK21, ZP18]. The multi-scale progressive fusion network (MSPFN) [JWY20] characterizes and reconstructs rain streaks at multiple scales. The rain convolutional dictionary network (RCD-Net) [WXZ20] encodes the rain shape using the intrinsic convolutional dictionary learning mechanism. The multi-stage progressive image restoration network (MPRNet) [ZAK21] splits the image into different sections in various stages to learn contextualized features at

Dataset	Type	Rain Effects	Size
Rain12 [LTG16]	Simulated	Synth. streaks only	12
Rain100L [YTF17]	Simulated	Synth. streaks only	300
Rain800 [ZSP19]	Simulated	Synth. streaks only	800
Rain100H [YTF17]	Simulated	Synth. streaks only	1.9K
Outdoor-Rain [LCT19]	Simulated	Synth. streaks & Synth. accumulation	10.5K
RainCityscapes [HFZ19]	Simulated	Synth. streaks & Synth. accumulation	10.62K
Rain12000 [ZP18]	Simulated	Synth. streaks only	13.2K
Rain14000 [FHZ17]	Simulated	Synth. streaks only	14K
NYU-Rain [LCT19]	Simulated	Synth. streaks & Synth. accumulation	16.2K
SPA-Data [WYX19]	Semi-real	Real streaks only	29.5K
Proposed	Real	Real streaks & Real accumulation	31.5K

Table 6.1: **Our proposed large-scale dataset enables paired training and quantitative evaluation for real-world deraining.** We consider SPA-Data [WYX19] as a semi-real dataset since it only contains real rainy images, where the pseudo ground-truth images are synthesized from a rain streak removal algorithm.

different scales. The spatial attentive network (SPANet) [WYX19] learns physical properties of rain streaks in a local neighborhood and reconstructs the clean background using non-local information. EfficientDeRain (EDR) [GSJ21] aims to derain efficiently in real time by using pixel-wise dilation filtering. Other than rain streak removal, the heavy rain restorer (HRR) [LCT19] and the depth-guided non-local network (DGNL-Net) [HZW21] have also attempted to address rain accumulation effects. All of these prior methods use synthetic or semi-real datasets and show limited generalizability to real images. In contrast, we propose a derainer that learns a rain-robust representation directly.



(1) Scene selection (2) Time control (3) Large motion removal (4) Camera motion removal (5) Local motion removal

Figure 6.2: **We collect a real paired deraining dataset by rigorously controlling the environmental variations.** First, we remove heavily degraded videos such as scenes without proper exposure, noise, or water droplets on the lens. Next, we carefully choose the rainy and clean frames as close as possible in time to mitigate illumination shifts before cropping to remove large movements. Lastly, we correct for small camera motion (due to strong wind) using SIFT [Low04] and RANSAC [FB81] and perform elastic image registration [Thi98, VPP09] by estimating the displacement field when necessary.

6.3 Dataset

We now describe our method to control variations in a real dataset of paired images taken at two different timestamps, as illustrated in Fig. 6.2.

6.3.1 Data Collection

We collect rain and clean ground-truth videos using a Python program based on FFmpeg to download videos from YouTube live streams across the world. For each live stream, we record the location in order to determine whether there is rain according to the OpenWeatherMap API [Ltd]. We also determine the time of day to filter out nighttime videos. After the rain stops, we continue downloading in order to collect clean ground-truth frames. Note: while our dataset is formatted for single-image deraining, it can be re-purposed for video deraining as well by considering the timestamps of the frames collected.

6.3.2 Collection Criteria

To minimize variations between rainy and clean frames, videos are filtered based on a strict set of collection criteria. Note that we perform realignment for camera and local motion only when necessary – with manual oversight to filter out cases where motion still exists after realignment.

- **Heavily degraded scenes** that contain excessive noise, webcam artifacts, poor resolution, or poor camera exposure are filtered out as the underlying scene cannot be inferred from the images.
- **Water droplets** on the surface of the lens occlude large portions of the scene and also distort the image. Images containing this type of degradation are filtered out as it is out of the scope of our work – we focus on rain streaks and rain accumulation phenomena.
- **Illumination shifts** are mitigated by minimizing the time difference between rainy and clean frames. Our dataset has an average time difference of 25 minutes, which drastically limits large changes in global illumination due to sun position, clouds, etc.
- **Background changes** containing large discrepancies (e.g., cars, people, swaying foliage, water surfaces) are cropped from the frame to ensure that clean and rainy images are aligned. By limiting the average time difference between scenes, we also minimize these discrepancies before filtering. All sky regions are cropped out as well to ensure proper background texture.
- **Camera motion.** Adverse weather conditions, e.g., heavy wind, can cause camera movements between the rainy and clean frames. To address this, we use the Scale Invariant Feature Transform (SIFT) [Low04] and Random Sample Consensus (RANSAC) [FB81] to compute the homography to realign the frames.
- **Local motion.** Despite controlling for motion whenever possible, certain scenes still

contain small local movements that are unavoidable, especially in areas of foliage. To correct for this, we perform elastic image registration when necessary by estimating the displacement field [Thi98, VPP09].

6.3.3 Dataset Statistics

Our large-scale dataset includes a total of 31,524 rainy and clean frame pairs, which is split into 26,124 training frames, 3,300 validation frames, and 2,100 testing frames. These frames are taken from 101 videos, covering a large variety of background scenes from urban locations (e.g., buildings, streets, cityscapes) to natural scenery (e.g., forests, plains, hills). We span a wide range of geographic locations (e.g., North America, Europe, Oceania, and Asia) to ensure that we capture diverse scenes and rainfall conditions. The scenes also include varying degrees of illumination from different times of day and rain of varying densities, streak lengths, shapes, and sizes. The webcams cover a wide array of resolutions, noise levels, intrinsic parameters (focal length, distortion), etc. As a result, our dataset captures diverse rain effects that cannot be accurately reproduced by SPA-Data [WYX19] or synthetic datasets [FHZ17, HFZ19, LCT19, LTG16, YTF17, ZP18, ZSP19]. See Fig. 6.3 for representative image pairs in GT-RAIN.

6.4 Learning to Derain Real Images

To handle greater diversity of rain streak appearance, we propose to learn a representation (illustrated in Fig. 6.4) that is robust to rain for real image deraining.

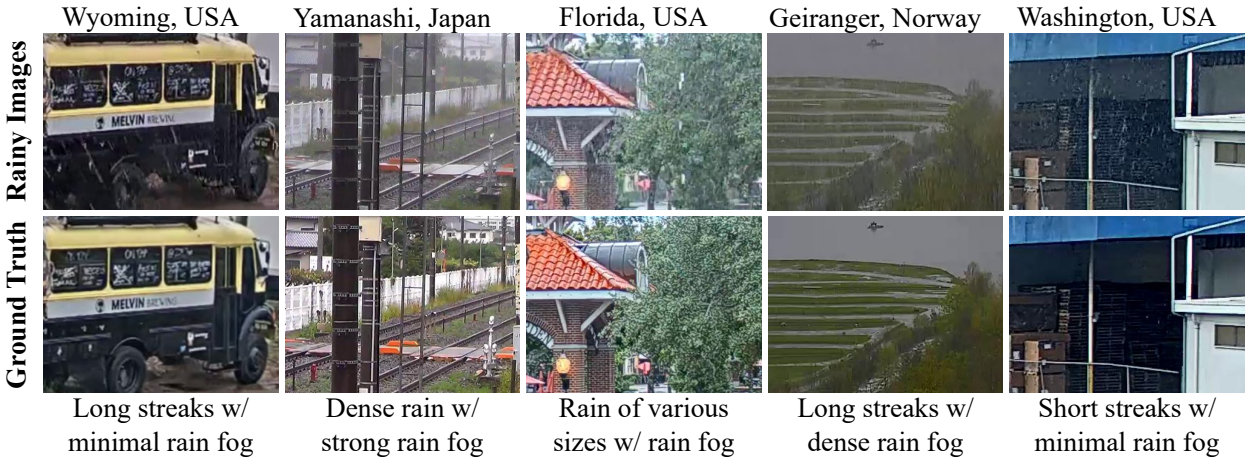


Figure 6.3: **Our proposed dataset contains diverse rainy images collected across the world.** We illustrate several representative image pairs with various rain streak appearances and rain accumulation strengths at different geographic locations.

6.4.1 Problem Formulation

Most prior works emphasize on rain streak removal and rely on the following equation to model rain [DHZ18, FHZ17, LHZ18, LTG16, WXZ20, WYX19, YP19, ZP18, ZFL17]:

$$\mathbf{I} = \mathbf{J} + \sum_{i}^{n} \mathbf{S}_i, \tag{6.2}$$

where $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ is the observed rainy image, $\mathbf{J} \in \mathbb{R}^{3\times H\times W}$ is the rain-free or "clean" image, and $\mathbf{S}_i$ is the i-th rain layer. However, real-world rain can be more complicated due to the dense rain accumulation and the rain veiling effect [LCT19, LTC20, YTF19]. These additional effects, which are visually similar to fog and mist, may cause severe degradation, and thus their removal should also be considered for single-image deraining. With GT-RAIN, it now becomes possible to study and conduct optically challenging, real-world rainy image restoration.

Given an image $\mathbf{I}$ of a scene captured during rain, we propose to learn a function $\mathcal{F}(\cdot, \theta)$ parameterized by θ to remove degradation induced by the rain phenomena. This function is realized as a neural network (see Fig. 6.4) that takes as input a rainy image $\mathbf{I}$ and outputs

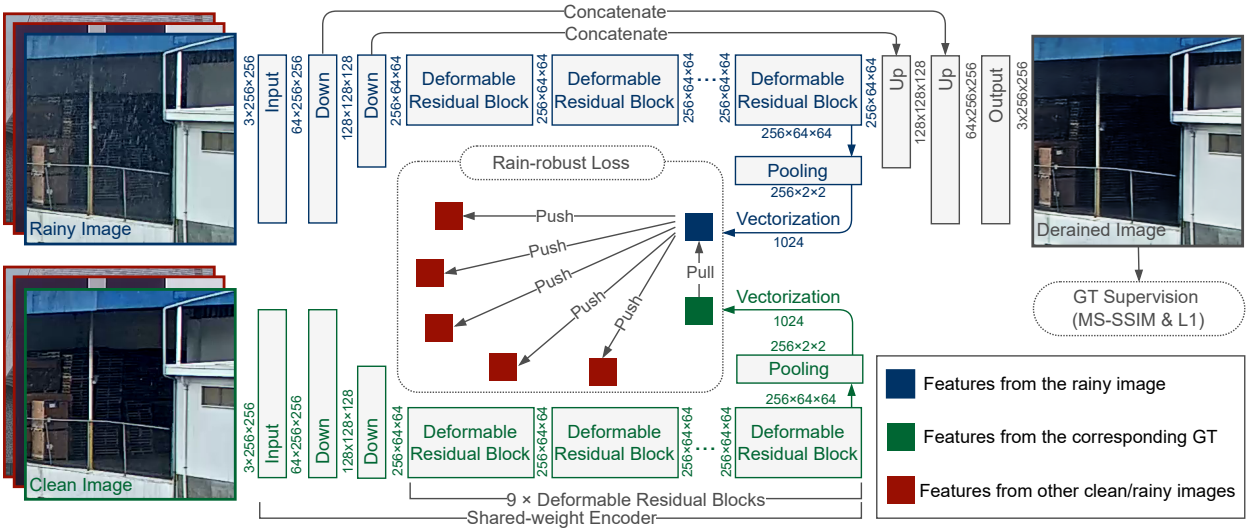


Figure 6.4: **By minimizing a rain-robust objective, our model learns robust features for reconstruction.** When training, a shared-weight encoder is used to extract features from rainy and ground-truth images. These features are then evaluated with the rain-robust loss, where features from a rainy image and its ground truth are encouraged to be similar. Learned features from the rainy images are also fed into a decoder to reconstruct the ground-truth images with MS-SSIM and $\ell 1$ loss functions.

a "clean" image $\hat{\mathbf{J}} = \mathcal{F}(\mathbf{I}, \theta) \in \mathbb{R}^{3\times H\times W}$, where undesirable characteristics, i.e., rain streaks and rain accumulation, are removed from the image to reconstruct the underlying scene $\mathbf{J}$.

6.4.2 Rain-robust Loss

To derain an image $\mathbf{I}$, one may directly learn a map from $\mathbf{I}$ to $\hat{\mathbf{J}}$ simply by minimizing the discrepancies between $\hat{\mathbf{J}}$ and the ground truth $\mathbf{J}$, i.e., an image reconstruction loss—such is the case for existing methods. Under this formulation, the model must explore a large hypothesis space, e.g., any region obfuscated by rain streaks is inherently ambiguous, making learning difficult.

Unlike previous works, we constrain the learned representation such that it is robust to

rain phenomena. To "learn away" the rain, we propose to map both the rainy and clean images of the same scene to an embedding space where they are close to each other by optimizing a similarity metric. Additionally, we minimize a reconstruction objective to ensure that the learned representation is sufficient to recover the underlying scene. Our approach is inspired by the recent advances in contrastive learning [CKN20], and we aim to distill rain-robust representations of real-world scenes by directly comparing the rainy and clean images in the feature space. But unlike [CKN20], we do not define a positive pair as an augmentation to the same image, but rather any rainy image and its corresponding clean image from the same scene.

When training, we first randomly sample a mini-batch of N rainy images with the associated clean images to form an augmented batch $\{(\mathbf{I}_i, \mathbf{J}_i)\}_{i=1}^{N}$, where $\mathbf{I}_i$ is the i-th rainy image, and $\mathbf{J}_i$ is its corresponding ground-truth image. This augmented batch is fed into a shared-weight feature extractor $\mathcal{F}_E(\cdot, \theta_E)$ with weights θ_E to obtain a feature set $\{(\mathbf{z}_{\mathbf{I}_i}, \mathbf{z}_{\mathbf{J}_i})\}_{i=1}^{N}$, where $\mathbf{z}_{\mathbf{I}_i} = \mathcal{F}_E(\mathbf{I}_i, \theta_E)$ and $\mathbf{z}_{\mathbf{J}_i} = \mathcal{F}_E(\mathbf{J}_i, \theta_E)$. We consider every $(\mathbf{z}_{\mathbf{I}_i}, \mathbf{z}_{\mathbf{J}_i})$ as the positive pairs. This is so that the learned features from the same scene should be close to each other regardless of the rainy conditions. We treat the other $2(N-1)$ samples from the same batch as negative samples. Based on the noise-contrastive estimation (NCE) [GH10], we adopt the following InfoNCE [OLV18] criterion to measure the rain-robust loss for a positive pair $(\mathbf{z}_{\mathbf{J}_i}, \mathbf{z}_{\mathbf{I}_i})$:

$$\ell_{\mathbf{z}_{\mathbf{J}_i}, \mathbf{z}_{\mathbf{I}_i}} = -\log \frac{\exp\left(\text{sim}_{\cos}(\mathbf{z}_{\mathbf{I}_i}, \mathbf{z}_{\mathbf{J}_i})/\tau\right)}{\sum_{\mathbf{k}\in\mathcal{K}} \exp\left(\text{sim}_{\cos}(\mathbf{z}_{\mathbf{J}_i}, \mathbf{k})/\tau\right)}, \tag{6.3}$$

where $\mathcal{K} = \{\mathbf{z}_{\mathbf{I}_j}, \mathbf{z}_{\mathbf{J}_j}\}_{j=1, j\neq i}^{N}$ is a set that contains the features extracted from other rainy and ground-truth images in the selected mini-batch, $\text{sim}_{\cos}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\intercal\mathbf{v}/\left\|\mathbf{u}\right\|\left\|\mathbf{v}\right\|$ is the cosine similarity between two feature vectors $\mathbf{u}$ and $\mathbf{v}$, and τ is the temperature parameter [WXY18]. We set τ as 0.25, and this loss is calculated across all positive pairs within the mini-batch for both $(\mathbf{z}_{\mathbf{I}_i}, \mathbf{z}_{\mathbf{J}_i})$ and $(\mathbf{z}_{\mathbf{J}_i}, \mathbf{z}_{\mathbf{I}_i})$.

6.4.3 Full Objective

While minimizing Eq. (6.3) maps features of clean and rainy images to the same subspace, we also need to ensure that the representation is sufficient to reconstruct the scene. Hence, we additionally minimize a Multi-Scale Structural Similarity Index (MS-SSIM) [WSB03] loss and a $\ell 1$ image reconstruction loss to prevent the model from discarding useful information for the reconstruction task. Our full objective $\mathcal{L}_{\text{full}}$ is as follows:

$$\mathcal{L}_{\text{full}}(\hat{\mathbf{J}}, \mathbf{J}) = \mathcal{L}_{\text{MS-SSIM}}(\hat{\mathbf{J}}, \mathbf{J}) + \lambda_{\ell 1}\mathcal{L}_{\ell 1}(\hat{\mathbf{J}}, \mathbf{J}) + \lambda_{\text{robust}}\mathcal{L}_{\text{robust}}(\mathbf{z}_{\mathbf{J}}, \mathbf{z}_{\mathbf{I}}), \tag{6.4}$$

where $\mathcal{L}_{\text{MS-SSIM}}(\cdot)$ is the MS-SSIM loss that is commonly used for image restoration [ZGF16], $\mathcal{L}_{\ell 1}(\cdot)$ is the $\ell 1$ distance between the estimated clean images $\hat{\mathbf{J}}$ and the ground-truth images $\mathbf{J}$, $\mathcal{L}_{\text{robust}}(\cdot)$ is the rain-robust loss in Eq. (6.3), and $\lambda_{\ell 1}$ and λ_{robust} are two hyperparameters to control the relative importance of different loss terms. In our experiments, we set both $\lambda_{\ell 1}$ and λ_{robust} as 0.1.

6.4.4 Network Architecture and Implementation Details

We design our model based on the architecture introduced in [JAF16, ZPI17]. As illustrated in Fig. 6.4, our network includes an encoder of one input convolutional block, two downsampling blocks, and nine residual blocks [HZR16] to yield latent features $\mathbf{z}$. This is followed by a decoder of two upsampling blocks and one output layer to map the features to $\mathbf{J}$. We fuse skip connections into the decoder using 3×3 up-convolution blocks to retain information lost in the bottleneck. Note: normal convolution layers are replaced by deformable convolution [ZHL19] in our residual blocks – in doing so, we enable our model to propagate non-local spatial information to reconstruct local degradations caused by rain effects. Latent features $\mathbf{z}$ are used for the rain-robust loss described in Eq. (6.3). Since these features are high dimensional ($256 \times 64 \times 64$), we use an average pooling layer to condense the feature map of each channel to 2×2. The condensed features are flattened into a vector of length 1024 for the rain-robust loss. It is worth noting that our rain-robust loss does not require

Data Split	Metrics	Rainy Images	SPANet [WYX19] (CVPR'19)	HRR [LCT19] (CVPR'19)	MSPFN [JWY20] (CVPR'20)	RCDNet [WXZ20] (CVPR'20)	DGNL-Net [HZW21] (IEEE TIP'21)	EDR [GSJ21] (AAAI'21)	MPRNet [ZAK21] (CVPR'21)	Ours
Dense Rain Streaks	PSNR↑	18.46	18.87	17.86	19.58	19.50	17.33	18.86	19.12	**20.84**
	SSIM↑	0.6284	0.6314	0.5872	0.6342	0.6218	0.5947	0.6296	0.6375	**0.6573**
Dense Rain Accumulation	PSNR↑	20.87	21.42	14.82	21.13	21.27	20.75	21.07	21.38	**24.78**
	SSIM↑	0.7706	0.7696	0.4675	0.7735	0.7765	0.7429	0.7766	0.7808	**0.8279**
Overall	PSNR↑	19.49	19.96	16.55	20.24	20.26	18.80	19.81	20.09	**22.53**
	SSIM↑	0.6893	0.6906	0.5359	0.6939	0.6881	0.6582	0.6926	0.6989	**0.7304**

Table 6.2: **Quantitative comparison on GT-RAIN.** Our method outperforms the existing state-of-the-art derainers. The preferred results are marked in **bold**.

additional modifications on the model architectures.

Our deraining model is trained on 256×256 patches and a mini-batch size $N = 8$ for 20 epochs. We use the Adam optimizer [KB14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 2×10^{-4}, and it is steadily modified to 1×10^{-6} based on a cosine annealing schedule [LH17]. We also use a linear warm-up policy for the first 4 epochs. For data augmentation, we use random cropping, random rotation, random horizontal and vertical flips, and RainMix augmentation [GSJ21].

6.5 Experiments

We compare to state-of-the-art methods both quantitatively and qualitatively on GT-RAIN, and qualitatively Internet rainy images [WMZ19]. To quantify the difference between the derained results and ground truth, we adopt peak signal-to-noise ratio (PSNR) [HG08] and structure similarity (SSIM) [WBS04].

6.5.1 Quantitative Evaluation on GT-RAIN

To quantify the sim2real gap of the existing datasets, we test seven representative existing state-of-the-art methods [GSJ21, HZW21, JWY20, LCT19, WXZ20, WYX19, ZAK21] on our GT-RAIN test set. Since there exist numerous synthetic datasets proposed by previous works [FHZ17, HFZ19, LCT19, LTG16, YTF17, ZP18, ZSP19], we found it intractable to train our method on each one; whereas, it is more feasible to take the best derainers for each respective dataset and test on our proposed dataset as a proxy (Tab. 6.2). This follows the conventions of previous deraining dataset papers [FHD17, HZW21, LTG16, WYX19, YTW20, ZP18, ZSP19] to compare with top performing methods from each existing dataset.

SPANet [WYX19] is trained on SPA-Data [WYX19]. HRR [LCT19] utilizes both NYU-Rain [LCT19] and Outdoor-Rain [LCT19]. MSPFN [JWY20] and MPRNet [ZAK21] are trained on a combination of multiple synthetic datasets [FHZ17, LTG16, YTF17, ZSP19]. DGNL-Net [HZW21] is trained on RainCityscapes [HFZ19]. For RCDNet [WXZ20] and EDR [GSJ21], multiple weights from different training sets are provided. We choose RCDNet trained on SPA-Data and EDR V4 trained on Rain14000 [FHZ17] due to superior performance.

Compared to training on GT-RAIN (ours), methods trained on other data perform worse, with the largest domain gap being in NYU-Rain and Outdoor-Rain (HRR) and RainCityscapes (DGNL). Two trends do hold: training on (1) more synthetic data gives better results (MSPFN, MPRNet), and (2) semi-real data also helps (SPANet). However, even when multiple synthetic [FHZ17, LTG16, YTF17, ZSP19] or semi-real [WYX19] datasets are used, their performance on real data is still around 2dB lower than training on GT-RAIN (ours).

Figure 6.5 illustrates some representative derained images across scenarios with various rain appearance and rain accumulation densities. Training on GT-RAIN enables the network to remove most rain streaks and rain accumulation; whereas, training on synthetic/semi-real

data tends to leave visible rain streaks. We note that HRR [LCT19] and DGNL [HZW21] may seem like they remove rain accumulation, but they in fact introduce undesirable artifacts, e.g., dark spots on the back of the traffic sign, tree, and sky. The strength of having ground-truth paired data is demonstrated by our 2.44 dB gain compared to the state of the art [ZAK21]. On test images with dense rain accumulation, the boost improves to 3.40 dB.

6.5.2 Qualitative Evaluation on Other Real Images

Other than the models described in the above section, we also include EDR V4 [GSJ21] trained on SPA-Data [WYX19] for the qualitative comparison, since it shows more robust rain streak removal results as compared to the version trained on Rain14000 [FHZ17]. The derained results on Internet rainy images are illustrated in Fig. 6.6. The model trained on the proposed GT-RAIN (i.e., ours) deals with large rain streaks of various shapes and sizes as well as the associated rain accumulation effects while preserving the features present in the scene. In contrast, we observe that models [HZW21, LCT19] trained on data with synthetic rain accumulation introduce unwanted color shifts and residual rain streaks in their results. Moreover, the state-of-the-art methods [JWY20, WXZ20, ZAK21] are unable to remove the majority of rain streaks in general as highlighted in the red zoom boxes. This demonstrates the gap between top methods on synthetic versus one that can be applied to real data.

6.5.3 Retraining Other Methods on GT-RAIN

We additionally train several state-of-the-art derainers [GSJ21, WXZ20, ZAK21] on the GT-RAIN training set to demonstrate that our real dataset leads to more robust real-world deraining and benefits all models. We have selected the most recent derainers for this retraining study.[1] All the models are trained from scratch, and the corresponding PSNR

[1]Both DGNL-Net [HZW21] and HRR [LCT19] cannot be retrained on our real dataset, as both require additional supervision, such as transmission maps and depth maps.

Figure 6.5: **Our model simultaneously removes rain streaks and rain accumulation, while the existing models fail to generalize to real-world data.** The red arrows highlight the difference between the proposed and existing methods on the GT-RAIN test set (zoom for details, PSNR and SSIM scores are listed below the images).

Figure 6.6: **Our model can generalize across real rainy images with robust performance.** We select representative real rainy images with various rain patterns and backgrounds for comparison (zoom for details). EDR V4 (S) [GSJ21] denotes EDR trained on SPA-Data [WYX19], and EDR V4 (R) [GSJ21] denotes EDR trained on Rain14000 [FHZ17].

and SSIM scores on the GT-RAIN test set are provided in Tab. 6.3. For all the retrained models, we can observe a PSNR and SSIM gain by using the proposed GT-RAIN dataset. In addition, with all models trained on the same dataset, our model still outperforms others in all categories.

Data Split	Metrics	Rainy Images	RCDNet [WXZ20] (Original)	RCDNet [WXZ20] (GT-RAIN)	EDR [GSJ21] (Original)	EDR [GSJ21] (GT-RAIN)	MPRNet [ZAK21] (Original)	MPRNet [ZAK21] (GT-RAIN)	Ours
Dense Rain Streaks	PSNR↑	18.46	19.50	19.60	18.86	19.95	19.12	20.19	**20.84**
	SSIM↑	0.6284	0.6218	0.6492	0.6296	0.6436	0.6375	0.6542	**0.6573**
Dense Rain Accumulation	PSNR↑	20.87	21.27	22.74	21.07	23.42	21.38	23.38	**24.78**
	SSIM↑	0.7706	0.7765	0.7891	0.7766	0.7994	0.7808	0.8009	**0.8279**
Overall	PSNR↑	19.49	20.26	20.94	19.81	21.44	20.09	21.56	**22.53**
	SSIM↑	0.6893	0.6881	0.7091	0.6926	0.7104	0.6989	0.7171	**0.7304**

Table 6.3: **Retraining comparison methods on GT-RAIN.** The improvement of these derainers further demonstrates the effectiveness of real paired data.

6.5.4 Fine-tuning Other Methods on GT-RAIN

To demonstrate of the effectiveness of combining real and synthetic datasets, we also fine-tune several more recent derainers [GSJ21, WXZ20, ZAK21] that are previously trained on synthetic datasets with the proposed GT-RAIN dataset. We fine-tune from the official weights as described in the above quantitative evaluation section, and the fine-tuning learning rate is 20% of the original learning rate for each method. For the proposed method, we pretrain the model on the synthetic dataset used by MSPFN [JWY20] and MPRNet [ZAK21]. The corresponding PSNR and SSIM scores on the GT-RAIN test set are listed in Tab. 6.4. In the table, we can observe a further boost as compared with training the models from scratch with just real or synthetic data.

6.5.5 Ablation Study

We validate the effectiveness of the rain-robust loss with two variants of the proposed method: (1) the proposed network with the full objective as described in Sec. 6.4; and (2) the proposed network with just MS-SSIM loss and ℓ_1 loss. The rest of the training configurations and hyperparameters remain identical. The quantitative metrics for these two variants on the

Data Split	Metrics	Rainy Images	RCDNet [WXZ20] (O)	RCDNet [WXZ20] (F)	EDR [GSJ21] (O)	EDR [GSJ21] (F)	MPRNet [ZAK21] (O)	MPRNet [ZAK21] (F)	Ours (O)	Ours (F)
Dense Rain Streaks	PSNR↑	18.46	19.50	19.33	18.86	20.03	19.12	20.65	**20.84**	20.79
	SSIM↑	0.6284	0.6218	0.6463	0.6296	0.6433	0.6375	0.6561	0.6573	**0.6655**
Dense Rain Accumulation	PSNR↑	20.87	21.27	22.50	21.07	23.57	21.38	24.37	24.78	**25.20**
	SSIM↑	0.7706	0.7765	0.7893	0.7766	0.8016	0.7808	0.8250	0.8279	**0.8318**
Overall	PSNR↑	19.49	20.26	20.69	19.81	21.55	20.09	22.24	22.53	**22.68**
	SSIM↑	0.6893	0.6881	0.7076	0.6926	0.7111	0.6989	0.7285	0.7304	**0.7368**

Table 6.4: **Fine-tuning comparison methods on GT-RAIN.** (F) denotes the fine-tuned models, and (O) denotes the original models trained on synthetic/real data.

Metrics	Rainy Images	Ours w/o $\mathcal{L}_{\text{robust}}$	Ours w/ $\mathcal{L}_{\text{robust}}$
PSNR↑	19.49	21.82	**22.53**
SSIM↑	0.6893	0.7148	**0.7304**

Table 6.5: **Ablation study.** Our rain-robust loss improves both PSNR and SSIM.

proposed GT-RAIN test set are listed in Tab. 6.5. Our model trained with the proposed rain-robust loss produces a normalized correlation between rainy and clean latent vectors of .95 ± .03; whereas it is .85 ± .10 for the one without. These rain-robust features help the model to show improved performance in both PSNR and SSIM.

6.5.6 Clean Images as Input

An ideal derainer should be able to preserve the image appearance when clean images are used as its input. Some typical output images are illustrated in Fig. 6.7 by directly feeding images without any rain effect into our model. We can observe that the majority of the model output remains closely identical to the corresponding clean input, which validates

Clean Input | Model Output | Clean Input | Model Output

Figure 6.7: **Our proposed model is capable of preserving image appearance when using clean images as its input.** Two typical scenes with different backgrounds are selected for illustration.

Rainy | EDR V4 (R) [GSJ21] | MPRNet [ZAK21] | Ours | Ground Truth

Figure 6.8: **Deraining is still an open problem.** Both the proposed method and the existing work have difficulty in generalizing the performance to some challenging scenes.

the effectiveness of our proposed model on clean image preservation. However, it should also be noted that our model still introduces some artifacts (e.g., blurriness) in its output. A potential future direction of the work can focus on removing these artifacts, such as encouraging the existence of high-frequency details in the output space.

6.5.7 Failure Cases

Apart from the successful cases illustrated in Fig. 6.5, we also provide some of the failure cases in the GT-RAIN test set in Fig. 6.8. Deraining is still an open problem, and we hope future work can take advantage of both real and synthetic samples to make derainers more robust in diverse environments.

6.6 Conclusion

Many of us in the deraining community probably wish for the existence of parallel universes, where we could capture the exact same scene with and without weather effects at the exact same time. Unfortunately, however, we are stuck with our singular universe, in which we are left with two choices: (1) synthetic data at the same timestamp with simulated weather effects or (2) real data at different timestamps with real weather effects. Though it is up to opinion, it is our belief that the results of our method in Fig. 6.6 reduce the visual domain gap more than those trained with synthetic datasets. Additionally, we hope the introduction of a real dataset opens up exciting new pathways for future work, such as the blending of synthetic and real data or setting goalposts to guide the continued development of existing rain simulators [HLC19, NCY21, WYX21, YCZ21, YXZ21].

CHAPTER 7

Conclusion

In this dissertation, we investigate the general paradigm of integrating physical priors into neural networks as well as how physical knowledge can benefit the learning process with specific applications in the field of computational imaging and computer vision. Applications in this dissertation primarily focus on incorporating physical knowledge in a single-modality setting (e.g., RGB domain for skin-tone inclusive rPPG), while complicated real-world inference usually requires a collaboration of multiple sensor modalities. One potential future work direction is to extend the idea of learning with physical priors and inductive biases to broader applications with multiple sensor modalities and advance the technique from low-level vision to higher-level machine intelligence, such as 3D estimation and scene understanding for autonomous agents in complex or extreme environments.

In addition, contriving an automated way to blend suitable prior information from a list of potential candidates can be another line of future work. In this dissertation, we mainly focus on finding a suitable way to blend a specific type of physical knowledge for a specific task, however, it is common that there exist multiple types of prior knowledge designed for the same task. Each type of prior knowledge may contribute differently when meeting with real data. Establishing an efficient mechanism to identify and combine suitable prior knowledge from all the potential candidates can be an important step toward more robust and reliable inference in real-world scenarios.

REFERENCES

[AAA19] Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, John Barrett, Dan Bintley, et al. "First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole." *The Astrophysical Journal Letters*, **875**(1):L4, 2019.

[AAS19] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. "Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images." In *2019 IEEE international conference on image processing (ICIP)*, pp. 1014–1018. IEEE, 2019.

[AAT18a] Codruta O Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 754–762, 2018.

[AAT18b] Cosmin Ancuti, Codruta O Ancuti, Radu Timofte, and Christophe De Vleeschouwer. "I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images." In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 620–631. Springer, 2018.

[AAT20] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. "NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 444–445, 2020.

[ABW19] Anurag Ajay, Maria Bauza, Jiajun Wu, Nima Fazeli, Joshua B Tenenbaum, Alberto Rodriguez, and Leslie P Kaelbling. "Combining Physical Simulators and Object-Based Networks for Control." *arXiv preprint arXiv:1904.06580*, 2019.

[AE18] Gary A. Atkinson and Jürgen D. Ernst. "High-sensitivity analysis of polarization by surface reflection." *Machine Vision and Applications*, 2018.

[AFE18] Stefano Amalfitano, Stefano Fazi, Elisabet Ejarque, Anna Freixa, Anna M Romaní, and Andrea Butturini. "Deconvolution model to resolve cytometric microbial community patterns in flowing waters." *Cytometry Part A*, **93**(2):194–200, 2018.

[AH05] Gary A. Atkinson and Edwin R. Hancock. "Multi-view surface reconstruction using polarization." *ICCV*, 2005.

[AH06] Gary A Atkinson and Edwin R Hancock. "Recovery of surface orientation from diffuse polarization." *IEEE TIP*, 2006.

[APM21] Edem Allado, Mathias Poussel, Anthony Moussu, Véronique Saunier, Yohann Bernard, Eliane Albuisson, and Bruno Chenuel. "Innovative measurement of routine physiological variables (heart rate, respiratory rate and oxygen saturation) using a remote photoplethysmography imaging system: A prospective comparative trial protocol." *BMJ open*, **11**(8):e047896, 2021.

[AS17] Sarah Alotaibi and William AP Smith. "A biophysical 3D morphable model of face appearance." In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 824–832, 2017.

[AS19a] Sarah Alotaibi and William Smith. "BioFaceNet: Deep Biophysical Face Image Interpretation." In *British Machine Vision Conference (BMVC)*, 2019.

[AS19b] Sarah Alotaibi and William AP Smith. "Decomposing multispectral face images into diffuse and specular shading and biophysical parameters." In *IEEE International Conference on Image Processing (ICIP)*, pp. 3138–3142. IEEE, 2019.

[ASA18] Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, Josh Tenenbaum, and J Zico Kolter. "End-to-end differentiable physics for learning and control." In *Advances in Neural Information Processing Systems*, pp. 7178–7189, 2018.

[Atk17] Gary A. Atkinson. "Polarisation photometric stereo." *Computer Vision and Image Understanding*, 2017.

[BC87] Kenneth V. Beard and Catherine Chuang. "A new model for the equilibrium shape of raindrops." *Journal of Atmospheric Sciences*, **44**(11):1509–1524, 1987.

[BDO20] Bastiaan R Bloem, E Ray Dorsey, and Michael S Okun. "The coronavirus disease 2019 crisis as catalyst for telemedicine for chronic neurological disorders." *JAMA neurology*, **77**(8):927–928, 2020.

[BG18] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

[BGN16] Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. "Designing neural network architectures using reinforcement learning." *arXiv preprint arXiv:1611.02167*, 2016.

[BGW20] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. "Deep shape from polarization." In *European Conference on Computer Vision*, pp. 554–571. Springer, 2020.

[BJT18] Seung-Hwan Baek, Daniel S Jeon, Xin Tong, and Min H Kim. "Simultaneous acquisition of polarimetric SVBRDF and normals." *ACM SIGGRAPH (TOG)*, 2018.

[BJZ16] Katherine L Bouman, Michael D Johnson, Daniel Zoran, Vincent L Fish, Sheperd S Doeleman, and William T Freeman. "Computational imaging for vlbi image reconstruction." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 913–922, 2016.

[BLT15] Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM van der Laak. "Stain specific standardization of whole-slide histopathological images." *IEEE transactions on medical imaging*, **35**(2):404–415, 2015.

[BMB19] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. "Unsupervised skin tissue segmentation for remote photoplethysmography." *Pattern Recognition Letters*, **124**:82–90, 2019.

[Bok21] Anthony Jnr Bokolo. "Exploring the adoption of telemedicine and virtual software for care of outpatients during and after COVID-19 pandemic." *Irish Journal of Medical Science (1971-)*, **190**(1):1–10, 2021.

[BOS19] George Barbastathis, Aydogan Ozcan, and Guohai Situ. "On the use of deep learning for computational imaging." *Optica*, **6**(8):921–943, 2019.

[BRC15] Aldo Badano, Craig Revie, Andrew Casertano, Wei-Chung Cheng, Phil Green, Tom Kimpe, Elizabeth Krupinski, Christye Sisson, Stein Skrøvseth, Darren Treanor, et al. "Consistency and standardization of color in medical imaging: a consensus report." *Journal of digital imaging*, **28**(1):41–52, 2015.

[BSP02] Kiran S Bhat, Steven M Seitz, Jovan Popović, and Pradeep K Khosla. "Computing the physical parameters of rigid-body motion from video." In *European Conference on Computer Vision*, pp. 551–565. Springer, 2002.

[BVM17] Kai Berger, Randolph Voorhies, and Larry H. Matthies. "Depth from stereo polarization in specular scenes for urban robotics." *ICRA*, 2017.

[BWK22] Yunhao Ba, Zhen Wang, Kerim Doruk Karinca, Oyku Deniz Bozkurt, and Achuta Kadambi. "Style Transfer with Bio-realistic Appearance Manipulation for Skin-tone Inclusive rPPG." In *2022 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–12. IEEE, 2022.

[BZK19] Yunhao Ba, Guangyuan Zhao, and Achuta Kadambi. "Blending diverse physical priors with neural networks." *arXiv preprint arXiv:1910.00201*, 2019.

[BZY22] Yunhao Ba, Howard Zhang, Ethan Yang, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Celso M de Melo, Suya You, Stefano Soatto, Alex Wong, et al. "Not Just Streaks: Towards Ground Truth for Single Image Deraining." In *European Conference on Computer Vision*, pp. 723–740. Springer, 2022.

[CBA22] Pradyumna Chari, Yunhao Ba, Shreeram Athreya, and Achuta Kadambi. "MIME: Minority Inclusion for Majority Group Enhancement of AI Performance." In *European Conference on Computer Vision*, pp. 326–343. Springer, 2022.

[CBD15] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. "Binaryconnect: Training deep neural networks with binary weights during propagations." In *Advances in neural information processing systems*, pp. 3123–3131, 2015.

[CGG18] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. "Reblur2deblur: Deblurring videos via self-supervised learning." *ICCP*, 2018.

[CGS17] Zhaopeng Cui, Jinwu Gu, Boxin Shi, Ping Tan, and Jan Kautz. "Polarimetric multi-view stereo." *CVPR*, 2017.

[CH13] Yi-Lei Chen and Chiou-Ting Hsu. "A generalized low-rank appearance model for spatio-temporally correlated rain streaks." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1968–1975, 2013.

[Cha16] Ayan Chakrabarti. "Learning sensor multiplexing design through back-propagation." In *Advances in Neural Information Processing Systems*, pp. 3081–3089, 2016.

[CHW18] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. "PS-FCN: A flexible learning framework for photometric stereo." *ECCV*, 2018.

[CKK20] Pradyumna Chari, Krish Kabra, Doruk Karinca, Soumyarup Lahiri, Diplav Srivastava, Kimaya Kulkarni, Tianyuan Chen, Maxime Cannesson, Laleh Jalilian, and Achuta Kadambi. "Diverse R-PPG: Camera-Based Heart Rate Estimation for Diverse Subject Skin-Tones and Scenes." *arXiv preprint arXiv:2010.12769*, 2020.

[CKN20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[CLC17] Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. "Neural stain-style transfer learning using gan for histopathological images." *arXiv preprint arXiv:1710.08543*, 2017.

[CLZ18] Chengqian Che, Fujun Luan, Shuang Zhao, Kavita Bala, and Ioannis Gkioulekas. "Inverse transport networks." *arXiv preprint arXiv:1809.10820*, 2018.

[CM18] Weixuan Chen and Daniel McDuff. "Deepphys: Video-based physiological measurement using convolutional attention networks." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–365, 2018.

[Csa01] Balázs Csanád Csáji. "Approximation with artificial neural networks." *Faculty of Sciences, Etvs Lornd University, Hungary*, **24**:48, 2001.

[CSX18] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. "Vggface2: A dataset for recognising faces across pose and age." In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

[CTB19] Pradyumna Chari, Chinmay Talegaonkar, Yunhao Ba, and Achuta Kadambi. "Visual physics: Discovering physical laws from videos." *arXiv preprint arXiv:1911.11893*, 2019.

[CZH18] Han Cai, Ligeng Zhu, and Song Han. "Proxylessnas: Direct neural architecture search on target task and hardware." *arXiv preprint arXiv:1812.00332*, 2018.

[CZS18] Lixiong Chen, Yinqiang Zheng, Art Subpa-asa, and Imari Sato. "Polarimetric three-view geometry." *ECCV*, 2018.

[DAD18] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. "Single-image SVBRDF capture with a rendering-aware deep network." *ACM SIGGRAPH (TOG)*, 2018.

[DHZ18] Liang-Jian Deng, Ting-Zhu Huang, Xi-Le Zhao, and Tai-Xiang Jiang. "A directional global sparse model for single image rain removal." *Applied Mathematical Modelling*, **59**:662–679, 2018.

[DJ13] Gerard De Haan and Vincent Jeanne. "Robust pulse rate from chrominance-based rPPG." *IEEE Transactions on Biomedical Engineering*, **60**(10):2878–2886, 2013.

[DPJ21] Ananyananda Dasari, Sakthi Kumar Arul Prakash, László A Jeni, and Conrad S Tucker. "Evaluation of biases in remote photoplethysmography methods." *NPJ digital medicine*, **4**(1):1–13, 2021.

[DS01] Ondrej Drbohlav and Radim Sara. "Unambiguous determination of shape from photometric stereo with unknown light sources." *ICCV*, 2001.

[DSH17] Steven Diamond, Vincent Sitzmann, Felix Heide, and Gordon Wetzstein. "Unrolled optimization with deep priors." *arXiv preprint arXiv:1705.08041*, 2017.

[DV14] Gerard De Haan and Arno Van Leest. "Improved motion robustness of remote-PPG by using the blood volume pulse signature." *Physiological measurement*, **35**(9):1913, 2014.

[EBM14] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography." In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1462–1469. IEEE, 2014.

[ESS15] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. "Multimodal deep learning for robust RGB-D object recognition." In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 681–687. IEEE, 2015.

[FAL15] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. "Learning visual predictive models of physics for playing billiards." *arXiv preprint arXiv:1511.07404*, 2015.

[FB81] Martin A. Fischler and Robert C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography." *Communications of the ACM*, **24**(6):381–395, 1981.

[FD69] G. B. Foote and P. S. Du Toit. "Terminal velocity of raindrops aloft." *Journal of Applied Meteorology*, **8**(2):249–253, 1969.

[FE20] Mehdi Foroozandeh and Anders Eklund. "Synthesizing brain tumor images and annotations by combining progressive growing GAN and SPADE." *arXiv preprint arXiv:2009.05946*, 2020.

[FFB21] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. "Learning an animatable detailed 3D face model from in-the-wild images." *ACM Transactions on Graphics (TOG)*, **40**(4):1–13, 2021.

[FHD17] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. "Clearing the skies: A deep network architecture for single-image rain removal." *IEEE Transactions on Image Processing*, **26**(6):2944–2956, 2017.

[FHZ17] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. "Removing rain from single images via a deep detail network." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3855–3863, 2017.

[Fit88] Thomas B Fitzpatrick. "The validity and practicality of sun-reactive skin types I through VI." *Archives of dermatology*, **124**(6):869–871, 1988.

[FWS19] Xiaohan Fei, Alex Wong, and Stefano Soatto. "Geo-supervised visual depth prediction." *IEEE Robotics and Automation Letters*, 2019.

[GAL18] Alexandre Goy, Kwabena Arthur, Shuai Li, and George Barbastathis. "Low photon count phase retrieval using deep learning." *APS PRL*, 2018.

[GCP10] Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A Wilson, and Paul Debevec. "Circularly polarized spherical illumination reflectometry." *ACM SIGGRAPH (TOG)*, 2010.

[GFT11] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. "Multiview face capture using polarized spherical gradient illumination." *ACM SIGGRAPH (TOG)*, 2011.

[GH10] Michael Gutmann and Aapo Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models." In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

[GK49] Ross Gunn and Gilbert D. Kinzer. "The terminal velocity of fall for water droplets in stagnant air." *Journal of Atmospheric Sciences*, **6**(4):243–248, 1949.

[GL10] Karol Gregor and Yann LeCun. "Learning fast approximations of sparse coding." *ICML*, 2010.

[GN07] Kshitiz Garg and Shree K. Nayar. "Vision and rain." *International Journal of Computer Vision*, **75**(1):3–27, 2007.

[GPD12] Giuseppe Claudio Guarnera, Pieter Peers, Paul Debevec, and Abhijeet Ghosh. "Estimating surface normals from spherical stokes reflectance fields." *ECCV*, 2012.

[GSJ21] Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, Yang Liu, and Jianjun Zhao. "EfficientDeRain: Learning Pixel-wise Dilation Filtering for High-Efficiency Single-Image Deraining." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1487–1495, 2021.

[GTH00] Radek Grzeszczuk, Demetri Terzopoulos, and Geoffrey Hinton. *NeuroAnimator: fast neural network emulation and control of physics-based models.* University of Toronto, 2000.

[HCS19] Zhuo Hui, Ayan Chakrabarti, Kalyan Sunkavalli, and Aswin C Sankaranarayanan. "Learning to Separate Multiple Illuminants in a Single Image." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3780–3789, 2019.

[HFZ19] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. "Depth-attentional features for single-image rain removal." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8022–8031, 2019.

[HG08] Quan Huynh-Thu and Mohammed Ghanbari. "Scope of validity of PSNR in image/video quality assessment." *Electronics letters*, **44**(13):800–801, 2008.

[HKK18] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. "Learning a variational network for reconstruction of accelerated MRI data." *Magnetic resonance in medicine*, **79**(6):3055–3071, 2018.

[HLC19] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. "Physics-based rendering for improving robustness to rain." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10203–10212, 2019.

[HRH10] Cong Phuoc Huynh, A. Robles-Kelly, and Edwin R. Hancock. "Shape and refractive index recovery from single-view polarisation images." *CVPR*, 2010.

[HRH13] Cong Phuoc Huynh, A. Robles-Kelly, and Edwin R. Hancock. "Shape and refractive index from single-view spectro-polarimetric images." *IJCV*, 2013.

[HSL16] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. "Deep Networks with Stochastic Depth." *CoRR*, 2016.

[HZR15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[HZW21] Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. "Single-Image Real-Time Rain Removal Based on Depth-Guided Non-Local Features." *IEEE Transactions on Image Processing*, **30**:1759–1770, 2021.

[Ike18] Satoshi Ikehata. "CNN-PS: CNN-based photometric stereo for general non-convex surfaces." *ECCV*, 2018.

[INN07] Takanori Igarashi, Ko Nishino, and Shree K Nayar. *The appearance of human skin: A survey.* Now Publishers Inc, 2007.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

[JAF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *European Conference on Computer Vision*, 2016.

[Jak10] Wenzel Jakob. "Mitsuba renderer.", 2010. http://www.mitsuba-renderer.org.

[JHZ18] Tai-Xiang Jiang, Ting-Zhu Huang, Xi-Le Zhao, Liang-Jian Deng, and Yao Wang. "Fastderain: A novel video rain streak removal method using directional gradient priors." *IEEE Transactions on Image Processing*, **28**(4):2089–2102, 2018.

[JMF17] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. "Deep convolutional neural network for inverse problems in imaging." *IEEE Transactions on Image Processing*, 2017.

[JWY20] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. "Multi-scale progressive fusion network for single image deraining." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8346–8355, 2020.

[Kad21] Achuta Kadambi. "Achieving fairness in medical devices." *Science*, **372**(6537):30–31, 2021.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[KBR19] Michael Kellman, Emrah Bostan, Nicole Repina, and Laura Waller. "Physics-based learned design: Optimized coded-illumination for quantitative phase imaging." *IEEE Transactions on Computational Imaging*, 2019.

[KMY17] Eunhee Kang, Junhong Min, and Jong Chul Ye. "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction." *Medical Physics*, 2017.

[KPA21] Fatema-Tuz-Zohra Khanam, Asanka G Perera, Ali Al-Naji, Kim Gibson, Javaan Chahl, et al. "Non-Contact Automatic Vital Signs Monitoring of Infants in a Neonatal Intensive Care Unit Based on Neural Networks." *Journal of Imaging*, **7**(8):122, 2021.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, **25**:1097–1105, 2012.

[KTS15] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. "Polarized 3D: High-quality depth sensing with polarization cues." *ICCV*, 2015.

[KTS17] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. "Depth sensing using geometrically constrained polarization normals." *IJCV*, 2017.

[KVS15] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. "DistancePPG: Robust non-contact vital signs monitoring using a camera." *Biomedical optics express*, **6**(5):1565–1588, 2015.

[KWB13] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar. "Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles." *ACM Transactions on Graphics (ToG)*, **32**(6):1–10, 2013.

[KWR17a] Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. "Physics-guided neural networks (pgnn): An application in lake temperature modeling." *arXiv preprint arXiv:1710.11431*, 2017.

[KWR17b] Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. "Physics-guided Neural Networks (PGNN): An application in lake temperature modeling." *CoRR*, 2017.

[LAS18] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. "The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection." In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 242–249. IEEE, 2018.

[LBB17] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. "Learning a model of facial shape and expression from 4D scans." *ACM Trans. Graph.*, **36**(6):194–1, 2017.

[LCL19] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. "Reflection separation using a pair of unpolarized and polarized images." In *Advances in Neural Information Processing Systems*, pp. 14559–14569, 2019.

[LCT19] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. "Heavy Rain Image Restoration: Integrating Physics Model and Conditional Adversarial Learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1633–1642, 2019.

[LDP17] Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. "Modeling surface appearance from a single photograph using self-augmented convolutional neural networks." *ACM SIGGRAPH (TOG)*, 2017.

[LFP20] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. "Multi-task temporal shift attention networks for on-device contactless vitals measurement." *arXiv preprint arXiv:2006.03790*, 2020.

[LH17] Ilya Loshchilov and Frank Hutter. "SGDR: Stochastic Gradient Descent with Warm Restarts." In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[LHZ18] Guanbin Li, Xiang He, Wei Zhang, Huiyou Chang, Le Dong, and Liang Lin. "Non-locally enhanced encoder-decoder network for single image de-raining." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1056–1064, 2018.

[LHZ21] Hao Lu, Hu Han, and S Kevin 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*, pp. 12404–12413, 2021.

[LLM17] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. "PDE-net: Learning PDEs from data." *arXiv preprint arXiv:1710.09668*, 2017.

[LNP20] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis." *Proceedings of the National Academy of Sciences*, **117**(23):12592–12594, 2020.

[Low04] David Lowe. "Sift-the scale invariant feature transform." *Int. J*, **2**(91-110):2, 2004.

[LOW18] David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. "Single-photon 3D imaging with deep sensor fusion." *ACM SIGGRAPH (TOG)*, 2018.

[LP15] Xingyu Li and Konstantinos N Plataniotis. "A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics." *IEEE Transactions on Biomedical Engineering*, **62**(7):1862–1873, 2015.

[LPL20] Hanwen Liang, Konstantinos N Plataniotis, and Xingyu Li. "Stain Style Transfer of Histopathology Images Via Structure-Preserved Generative Learning." *arXiv preprint arXiv:2007.12578*, 2020.

[LRK11] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity." In *2011 federated conference on computer science and information systems (FedCSIS)*, pp. 405–410. IEEE, 2011.

[LSC18] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. "Materials for masses: SVBRDF acquisition with a single mobile phone image." *ECCV*, 2018.

[LSY18] Hanxiao Liu, Karen Simonyan, and Yiming Yang. "Darts: Differentiable architecture search." *arXiv preprint arXiv:1806.09055*, 2018.

[LTC20] Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. "All in one bad weather removal using architectural search." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3175–3185, 2020.

[Ltd] OpenWeather Ltd. "OpenWeatherMap API." `https://openweathermap.org/`. Accessed: 2021-11-05.

[LTG16] Yu Li, Robby T. Tan, Xiaojie Guo, Jiangbo Lu, and Michael S. Brown. "Rain streak removal using layer priors." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2736–2744, 2016.

[Luc18] Lucid Vision Phoenix polarization camera. "`https://thinklucid.com/product/phoenix-5-0-mp-polarized-model/`." 2018.

[LXJ15] Yu Luo, Yong Xu, and Hui Ji. "Removing rain from a single image via discriminative sparse coding." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3397–3405, 2015.

[LXR18] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. "Learning to reconstruct shape and spatially-varying reflectance from a single image." *ACM SIGGRAPH Asia (TOG)*, 2018.

[LYZ16] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. "3D mask face anti-spoofing with remote photoplethysmography." In *Proceedings of the European Conference on Computer Vision*, pp. 85–100. Springer, 2016.

[Man93] Robert M. Manning. *Stochastic Electromagnetic Image Propagation*. McGraw-Hill Companies, 1993.

[MB19] Daniel McDuff and Ethan Blackford. "iphys: An open non-contact imaging-based physiological measurement toolbox." In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6521–6524. IEEE, 2019.

[MBD21] Rita Meziatisabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. "UBFC-Phys: A Multimodal Database For Psychophysiological Studies Of Social Stress." *IEEE Transactions on Affective Computing*, 2021.

[McD18] Daniel McDuff. "Deep super resolution for recovering physiological information from videos." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1367–1374, 2018.

[MEF12] Ali H Mahmoud, Moumen T El-Melegy, and Aly A Farag. "Direct method for shape recovery from polarization and shading." *ICIP*, 2012.

[MGP14] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. "Improvements in remote cardiopulmonary measurement using a five band digital camera." *IEEE Transactions on Biomedical Engineering*, **61**(10):2593–2601, 2014.

[MHM17] Julio Marco, Quercus Hernandez, Adolfo Munoz, Yue Dong, Adrian Jarabo, Min H Kim, Xin Tong, and Diego Gutierrez. "DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging." *ACM SIGGRAPH (TOG)*, 2017.

[MHP07] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination." *Eurographics Conference on Rendering Techniques*, 2007.

[MHW20] Daniel McDuff, Javier Hernandez, Erroll Wood, Xin Liu, and Tadas Baltrusaitis. "Advancing Non-Contact Vital Sign Measurement using Synthetic Avatars." *arXiv preprint arXiv:2010.12949*, 2020.

[MKI04] Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. "Transparent surface modeling from a pair of polarization images." *PAMI*, 2004.

[MKS18] Tomohiro Maeda, Achuta Kadambi, Yoav Y Schechner, and Ramesh Raskar. "Dynamic heterodyne interferometry." *ICCP*, 2018.

[ML21] Yisroel Mirsky and Wenke Lee. "The creation and detection of deepfakes: A survey." *ACM Computing Surveys (CSUR)*, **54**(1):1–41, 2021.

[MN21] Daniel McDuff and Ewa Nowara. ""Warm Bodies": A Post-Processing Technique for Animating Dynamic Blood Flow on Photos and Avatars." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 2021.

[MNM09] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. "A method for normalizing histology slides for quantitative analysis." In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1107–1110. IEEE, 2009.

[MP48] JS Marshall and W. McK. Palmer. "THE DISTRIBUTION OF RAINDROPS WITH SIZE." *Journal of Meteorology*, **5**(4):165–166, 1948.

[MSB16] Daisuke Miyazaki, Takuya Shigetomi, Masashi Baba, Ryo Furukawa, Shinsaku Hiura, and Naoki Asada. "Surface normal estimation of black specular objects from multiview polarization images." *International Society for Optics and Photonics, Optical Engineering*, 2016.

[MSL18] Zhipeng Mo, Boxin Shi, Feng Lu, Sai-Kit Yeung, and Yasuyuki Matsushita. "Uncalibrated photometric stereo under natural illumination." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2936–2945, 2018.

[MSV18] Christopher A Metzler, Philip Schniter, Ashok Veeraraghavan, and Richard G Baraniuk. "prDeep: Robust phase retrieval with a flexible deep network." *arXiv preprint arXiv:1803.00212*, 2018.

[MTH03] Daisuke Miyazaki, Robby T Tan, Kenji Hara, and Katsushi Ikeuchi. "Polarization-based inverse rendering from a single view." *ICCV*, 2003.

[MYK19] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. "Learned reconstructions for practical mask-based lensless imaging." *Optics Express*, **27**(20):28075–28090, 2019.

[NCY21] Siqi Ni, Xueyun Cao, Tao Yue, and Xuemei Hu. "Controlling the rain: From removal to rendering." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6328–6337, 2021.

[NHS18] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video." In *Asian Conference on Computer Vision*, pp. 562–576. Springer, 2018.

[NMM18] Ewa Magdalena Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan. "SparsePPG: Towards Driver Monitoring Using Camera-Based Vital Signs Estimation in Near-Infrared." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1353–135309, 2018.

[NMV20a] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. "The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention." *arXiv preprint arXiv:2010.07770*, 2020.

[NMV20b] Ewa M Nowara, Daniel McDuff, and Ashok 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*, pp. 284–285, 2020.

[NMV21a] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. "The benefit of distraction: Denoising camera-based physiological measurements using inverse attention." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4955–4964, 2021.

[NMV21b] Ewa M. Nowara, Daniel McDuff, and Ashok 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*, pp. 3810–3819, 2021.

[NNT15] Trung Thanh Ngo, Hajime Nagahara, and Rin-ichiro Taniguchi. "Shape and light directions from shading and polarization." *CVPR*, 2015.

[NSH19] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation." *IEEE Transactions on Image Processing*, **29**:2409–2423, 2019.

[NTL18] Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. "Medical image synthesis with deep convolutional adversarial networks." *IEEE Transactions on Biomedical Engineering*, **65**(12):2720–2730, 2018.

[NWM18] Elias Nehme, Lucien E Weiss, Tomer Michaeli, and Yoav Shechtman. "Deep-STORM: super-resolution single-molecule microscopy by deep learning." *Optica*, **5**(4):458–464, 2018.

[NXL18] Thanh Nguyen, Yujia Xue, Yunzhe Li, Lei Tian, and George Nehmetallah. "Deep learning approach for Fourier ptychography microscopy." *Optics express*, **26**(20):26470–26484, 2018.

[NYH20] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. "Video-based remote physiological measurement via cross-verified feature disentangling." In *Proceedings of the European Conference on Computer Vision*, pp. 295–310. Springer, 2020.

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748*, 2018.

[PC04] Stephen J Preece and Ela Claridge. "Spectral filter optimization for the recovery of parameters which describe human skin." *IEEE transactions on pattern analysis and machine intelligence*, **26**(7):913–922, 2004.

[PGC17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in pytorch." *NIPS-W*, 2017.

[PGM19a] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

[PGM19b] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library." *Advances in neural information processing systems*, **32**:8026–8037, 2019.

[PJC16] Marco AF Pimentel, Alistair EW Johnson, Peter H Charlton, Drew Birrenkott, Peter J Watkinson, Lionel Tarassenko, and David A Clifton. "Toward a robust estimation of respiratory rate from pulse oximeters." *IEEE Transactions on Biomedical Engineering*, **64**(8):1914–1923, 2016.

[PKA09] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. "A 3D face model for pose and illumination invariant face recognition." In *IEEE international conference on advanced video and signal based surveillance*, pp. 296–301, 2009.

[PLD18] Jinshan Pan, Yang Liu, Jiangxin Dong, Jiawei Zhang, Jimmy Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang. "Physics-based generative adversarial models for image restoration and beyond." *arXiv preprint arXiv:1808.00605*, 2018.

[PLS18] Jinshan Pan, Sifei Liu, Deqing Sun, Jiawei Zhang, Yang Liu, Jimmy Ren, Zechao Li, Jinhui Tang, Huchuan Lu, Yu-Wing Tai, et al. "Learning dual convolutional neural networks for low-level vision." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3070–3079, 2018.

[PLW19] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. "Semantic image synthesis with spatially-adaptive normalization." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

[PMP10a] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. "Advancements in noncontact, multiparameter physiological measurements using a webcam." *IEEE transactions on biomedical engineering*, **58**(1):7–11, 2010.

[PMP10b] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Optics express*, **18**(10):10762–10774, 2010.

[Pol17] PolarM polarization camera. "`http://www.4dtechnology.com/products/polarimeters/polarcam/`." 2017.

[QLZ20] Zhiwei Qin, Zhao Liu, Ping Zhu, and Yongbo Xue. "A GAN-based image synthesis method for skin lesion classification." *Computer Methods and Programs in Biomedicine*, p. 105568, 2020.

[Rai18] Maziar Raissi. "Deep hidden physics models: Deep learning of nonlinear partial differential equations." *The Journal of Machine Learning Research*, **19**(1):932–955, 2018.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *MICCAI*, 2015.

[RH01] Ravi Ramamoorthi and Pat Hanrahan. "An efficient representation for irradiance environment maps." In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 497–500, 2001.

[RIS19] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. "Deep ppg: Large-scale heart rate estimation with convolutional neural networks." *Sensors*, **19**(14):3079, 2019.

[RPK17] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations." *arXiv preprint arXiv:1711.10561*, 2017.

[RRF17] Jérémy Riviere, Ilya Reshetouski, Luka Filipi, and Abhijeet Ghosh. "Polarization imaging reflectometry in the wild." *ACM SIGGRAPH (TOG)*, 2017.

[RSZ20] Dongwei Ren, Wei Shang, Pengfei Zhu, Qinghua Hu, Deyu Meng, and Wangmeng Zuo. "Single image deraining using bilateral recurrent network." *IEEE Transactions on Image Processing*, **29**:6852–6863, 2020.

[RWO19] Yair Rivenson, Yichen Wu, and Aydogan Ozcan. "Deep learning in holography and coherent imaging." *Light: Science & Applications*, **8**(1):1–8, 2019.

[SCC21] Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. "PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography." *IEEE Journal of Biomedical and Health Informatics*, **25**(5):1373–1384, 2021.

[Sch15] Yoav Y Schechner. "Self-calibrating imaging polarimetry." *ICCP*, 2015.

[SDP18] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging." *ACM Transactions on Graphics (TOG)*, **37**(4):114, 2018.

[SE17] Russell Stewart and Stefano Ermon. "Label-free supervision of neural networks with physics and domain knowledge." *AAAI*, 2017.

[SF19] Shuran Song and Thomas Funkhouser. "Neural Illumination: Lighting Prediction for Indoor Environments." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6918–6926, 2019.

[SGS15] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Highway Networks." *CoRR*, 2015.

[SHI18] SHINING 3D scanner. "`https://www.einscan.com/einscan-se-sp`." 2018.

[SHW18] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. "Deep end-to-end time-of-flight imaging." *CVPR*, 2018.

[SKC18] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. "SfSNet: Learning shape, reflectance and illuminance of faces in the wild." *CVPR*, 2018.

[SLK19] Yu Sun, Jiaming Liu, and Ulugbek S Kamilov. "Block Coordinate Regularization by Denoising." *arXiv preprint arXiv:1905.05113*, 2019.

[SLL17] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. "Lensless computational imaging through deep learning." *Optica*, **4**(9):1117–1125, 2017.

[SLW20] Xuan Song, Xinyan Liu, and Chunting Wang. "The role of telemedicine during the COVID-19 epidemic in China—experience from Shandong province.", 2020.

[SMW19] Boxin Shi, Zhipeng Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. "A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo." *PAMI*, 2019.

[SRT16] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. "Linear depth estimation from an uncalibrated, monocular polarisation image." *ECCV*, 2016.

[SRT18] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. "Height-from-polarisation with unknown Lighting or albedo." *PAMI*, 2018.

[SSO19] Guanya Shi, Xichen Shi, Michael O'Connell, Rose Yu, Kamyar Azizzadenesheli, Animashree Anandkumar, Yisong Yue, and Soon-Jo Chung. "Neural lander: Stable drone landing control using learned dynamics." *ICRA*, 2019.

[SSS17] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. "Deep photometric stereo network." *ICCV Workshops*, 2017.

[STG17] Guy Satat, Matthew Tancik, Otkrist Gupta, Barmak Heshmat, and Ramesh Raskar. "Object classification through scattering media with deep learning on time resolved measurement." *OSA Optics Express*, 2017.

[TAR16] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2396–2404, 2016.

[Thi98] J-P Thirion. "Image matching as a diffusion process: an analogy with Maxwell's demons." *Medical image analysis*, **2**(3):243–260, 1998.

[TLH20] Yun-Yun Tsou, Yi-An Lee, and Chiou-Ting Hsu. "Multi-Task Learning for Simultaneous Video Generation and Remote Photoplethysmography Estimation." In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[TM18] Tatsunori Taniai and Takanori Maehara. "Neural inverse rendering for general reflectance photometric stereo." *ICML*, 2018.

[TSR18] Matthew Tancik, Guy Satat, and Ramesh Raskar. "Flash photography for data-driven hidden scene recovery." *arXiv preprint arXiv:1810.11710*, 2018.

[TSS18] Matthew Tancik, Tristan Swedish, Guy Satat, and Ramesh Raskar. "Data-driven non-line-of-sight imaging with a traditional camera." *OSA Imaging and Applied Optics*, 2018.

[TSZ17] Silvia Tozza, William A. P. Smith, Dizhong Zhu, Ravi Ramamoorthi, and Edwin R. Hancock. "Linear differential constraints for photo-polarimetric height estimation." *ICCV*, 2017.

[TSZ18] Daniel Teo, Boxin Shi, Yinqiang Zheng, and Sai-Kit Yeung. "Self-calibrating polarising radiometric calibration." *CVPR*, 2018.

[VCJ19] Mauricio Villarroel, Sitthichok Chaichulee, João Jorge, Sara Davis, Gabrielle Green, Carlos Arteta, Andrew Zisserman, Kenny McCormick, Peter Watkinson, and Lionel Tarassenko. "Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit." *NPJ digital medicine*, **2**(1):1–18, 2019.

[VPP09] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. "Diffeomorphic demons: Efficient non-parametric image registration." *NeuroImage*, **45**(1):S61–S72, 2009.

[VSN08] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. "Remote plethysmographic imaging using ambient light." *Optics express*, **16**(26):21434–21445, 2008.

[VWG12] Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G Bawendi, and Ramesh Raskar. "Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging." *Nature Communications*, 2012.

[WBC22] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. "Synthetic Generation of Face Videos With Plethysmograph Physiology." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20587–20596, 2022.

[WBH21] Jonathan P Weiner, Stephen Bandeian, Elham Hatef, Daniel Lans, Angela Liu, and Klaus W Lemke. "In-Person and Telehealth Ambulatory Contacts and Costs in a Large US Insured Cohort Before and During the COVID-19 Pandemic." *JAMA network open*, **4**(3):e212618–e212618, 2021.

[WBS04] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing*, **13**(4):600–612, 2004.

[WBS16] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. "Algorithmic principles of remote PPG." *IEEE Transactions on Biomedical Engineering*, **64**(7):1479–1491, 2016.

[WDH19] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai 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*, pp. 692–702, 2019.

[WJX22] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. "Integrating scientific knowledge with machine learning for engineering and environmental systems." *ACM Computing Surveys*, **55**(4):1–37, 2022.

[WMZ19] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. "Semi-supervised transfer learning for image rain removal." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3877–3886, 2019.

[Wol97] Lawrence B. Wolff. "Polarization vision: A new sensory approach to image understanding." *Image Vision Computing*, 1997.

[WSB03] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

[WSD15] Wenjin Wang, Sander Stuijk, and Gerard De Haan. "A novel algorithm for remote photoplethysmography: Spatial subspace rotation." *IEEE transactions on biomedical engineering*, **63**(9):1974–1984, 2015.

[WXY18] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. "Unsupervised feature learning via non-parametric instance discrimination." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

[WXZ20] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. "A Model-Driven Deep Neural Network for Single Image Rain Removal." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[WYX19] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. "Spatial attentive single-image deraining with a high quality real rain dataset." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12270–12279, 2019.

[WYX21] Hong Wang, Zongsheng Yue, Qi Xie, Qian Zhao, Yefeng Zheng, and Deyu Meng. "From rain generation to rain removal." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14791–14801, 2021.

[XAJ18] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. "Pointfusion: Deep sensor fusion for 3d bounding box estimation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 244–253, 2018.

[XCB14] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J Gortler, David W Jacobs, and Todd Zickler. "From shading to local shape." *IEEE transactions on pattern analysis and machine intelligence*, **37**(1):67–79, 2014.

[XDV19] Yang Xiao, Etienne Decencière, Santiago Velasco-Forero, Hélène Burdin, Thomas Bornschlögl, Françoise Bernerd, Emilie Warrick, and Thérèse Baldeweck. "A new color augmentation method for deep learning segmentation of histological images." In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 886–890. IEEE, 2019.

[YAA20] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–19, 2020.

[YCZ21] Yuntong Ye, Yi Chang, Hanyu Zhou, and Luxin Yan. "Closing the loop: Joint rain generation and removal via disentangled image translation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2053–2062, 2021.

[Yil01] Öz Yilmaz. *Seismic data analysis: Processing, inversion, and interpretation of seismic data*. Society of exploration geophysicists, 2001.

[YLD18] Wenjie Ye, Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. "Single image surface appearance modeling with self-augmented CNNs and inexact supervision." *Wiley Online Library Computer Graphics Forum*, 2018.

[YP19] Rajeev Yasarla and Vishal M. Patel. "Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8405–8414, 2019.

[YPL19] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. "Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 151–160, 2019.

[YTF17] Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. "Deep joint rain detection and removal from a single image." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1357–1366, 2017.

[YTF19] Wenhan Yang, Robby T. Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. "Joint rain detection and removal from a single image with contextualized deep networks." *IEEE transactions on pattern analysis and machine intelligence*, **42**(6):1377–1393, 2019.

[YTL18] Luwei Yang, Feitong Tan, Ao Li, Zhaopeng Cui, Yasutaka Furukawa, and Ping Tan. "Polarimetric dense monocular SLAM." *CVPR*, 2018.

[YTW20] Wenhan Yang, Robby T. Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. "Single image deraining: From model-based to data-driven and beyond." *IEEE Transactions on pattern analysis and machine intelligence*, 2020.

[YXZ21] Zongsheng Yue, Jianwen Xie, Qian Zhao, and Deyu Meng. "Semi-supervised video deraining with dynamical rain generator." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 642–652, 2021.

[ZAK21] Syed W. Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad S. Khan, Ming-Hsuan Yang, and Ling Shao. "Multi-stage progressive image restoration." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14821–14831, 2021.

[ZFL17] Lei Zhu, Chi-Wing Fu, Dani Lischinski, and Pheng-Ann Heng. "Joint bi-layer optimization for single-image rain streak removal." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2526–2534, 2017.

[ZGF16] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. "Loss functions for image restoration with neural networks." *IEEE Transactions on computational imaging*, **3**(1):47–57, 2016.

[ZGW16] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. "Multimodal spontaneous emotion corpus for human behavior analysis." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3438–3446, 2016.

[ZHL19] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. "Deformable convnets v2: More deformable, better results." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9308–9316, 2019.

[ZL16] Barret Zoph and Quoc V Le. "Neural architecture search with reinforcement learning." *arXiv preprint arXiv:1611.01578*, 2016.

[ZLQ06] Xiaopeng Zhang, Hao Li, Yingyi Qi, Wee Kheng Leow, and Teck Khim Ng. "Rain removal in video by combining temporal and chromatic properties." In *2006 IEEE international conference on multimedia and expo*, pp. 461–464. IEEE, 2006.

[ZP18] He Zhang and Vishal M Patel. "Density-aware single image de-raining using a multi-stream dense network." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 695–704, 2018.

[ZPI17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[ZS19] Dizhong Zhu and William A. P. Smith. "Depth from a polarisation + RGB stereo pair." *CVPR*, 2019.

[ZS21] James Zou and Londa Schiebinger. "Ensuring that biomedical AI benefits diverse populations." *EBioMedicine*, p. 103358, 2021.

[ZSL19] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. "TossingBot: Learning to Throw Arbitrary Objects with Residual Physics." *arXiv preprint arXiv:1903.11239*, 2019.

[ZSP19] He Zhang, Vishwanath Sindagi, and Vishal M Patel. "Image de-raining using a conditional generative adversarial network." *IEEE transactions on circuits and systems for video technology*, **30**(11):3943–3956, 2019.

[ZZL16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks." *IEEE Signal Processing Letters*, **23**(10):1499–1503, 2016.