Motion Matters: Neural Motion Transfer for Better Camera Physiological Sensing

Anonymous ICCV submission

Paper ID 1080

Abstract

Machine learning models for camera-based physiological measurement can have weak generalization due to a lack of representative training data. Body motion is one of the most significant sources of noise when attempting to recover the subtle cardiac pulse from a video. We explore motion transfer as a form of data augmentation to introduce motion variation while preserving physiological changes. We adapt a neural video synthesis approach to augment videos for the task of remote photoplethysmography (PPG) and study the effects of motion augmentation with respect to 1) the magnitude and 2) the type of motion. After training on motion-augmented versions of publicly available datasets, the presented inter-dataset results on five benchmark datasets show improvements of up to 75% over existing state-of-the-art results. Our findings illustrate the utility of motion transfer as a data augmentation technique for improving the generalization of models for camera-based physiological sensing. We release our code and pre-trained models for using motion transfer as a data augmentation technique.

1. Introduction

Scalable health sensors enable frequent, opportunistic, and more equitable access to vital information about the body's internal state. Cameras are some of the most versatile and widely available sensors. Videos capture spatial, temporal, and ultimately frequency-specific information making them suitable for imaging dynamic processes, even below the surface of the skin [26]. Camera-based measurement of cardiac signals is one such application [19], in which cameras are used to measure the pulse via light reflected from the body, a principle known as photoplethysmography (PPG) [2, 39]. The PPG signals can be used to derive respiration [28], heart rate variability [28], arrhythmia [29], and blood pressure [12]. As a result this technology has the potential to turn webcams and smartphones into meaningful health sensors.

However, unlike traditional medical sensors, extracting physiological signals from a video requires more than filtering and simple signal processing. The state-of-the-art (SOTA) algorithms are supervised neural models [4, 34, 47, 16, 48]. Despite the prowess of these models, they are inherently limited by the diversity of the data used to train them. Public datasets (e.g., UBFC-rPPG [3], PURE [35]) serve as an extremely valuable resource for the research community, containing videos and synchronized physiological gold-standard measurements making them suitable for training and testing models. Building datasets such as these is challenging for two reasons: (1) collecting videos with gold-standard signals from a medical-grade sensor is time consuming and labor intensive, (2) it requires storing and distributing privacy sensitive biometric data. Therefore,

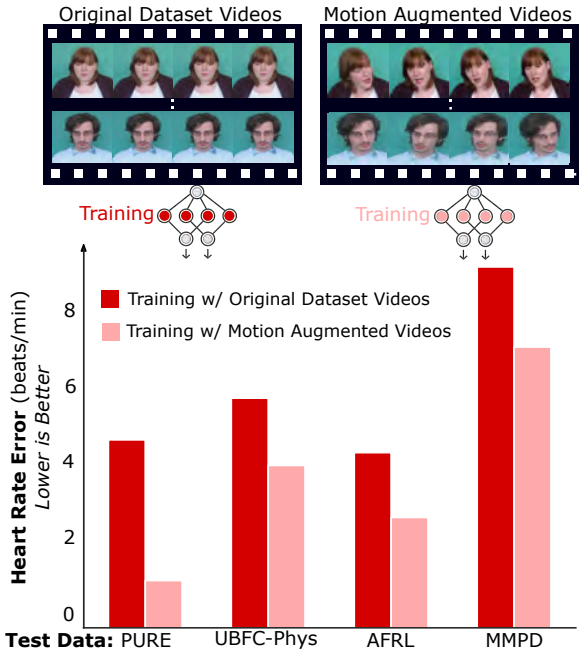


Figure 1: **Motion augmentation improves rPPG.** We present the first neural motion augmentation pipeline for the task of remote PPG estimation and empirically show it reduces error in heart rate estimation by up to 75%.

more data efficient methods for training rPPG sensing models would be desirable.

Synthetic data are a powerful resource in machine learning. The two main sources of synthetic data are (1) parametric computer graphics engines and (2) statistically-based generative machine learning models. Data created using these approaches have been used successfully for many computer vision tasks, including face detection, landmark localization, face parsing and face recognition [20, 44, 22], body pose estimation [31] and eye tracking [45, 36].

However, creating synthetic data that preserve the subtle and nuanced peripheral pulse in a video is non-trivial. McDuff et al. [23] released a large dataset (2,800 videos) of avatars and cardiac signals; however, their computer graphics pipeline had an extremely high overhead. Creating a pipeline for generating videos of avatars required years worth of investment, procuring assets (e.g., 3D facial scans and environments), building a parameterized rendering pipeline, and then generating each video frame-by-frame. Wang et al. [43] used a learning based method to generate synthetic videos given a reference image and target PPG signal. Their creative approach successfully incorporated PPG signals producing videos that benefited training. However, the videos created lacked the visual fidelity of other synthetics or real video datasets, and their pipeline involved several relatively complex components.

We question whether existing motion transfer algorithms can be used effectively for augmenting rPPG video data and explore what steps need to be taken to achieve optimal results. Our main contributions are as follows: (1) We perform a systematic investigation of the impact of motion augmentation on the physiological information within the video and in-turn the corresponding labels, (2) through experimentation, we provide quantitative, empirical evidence that training with certain kinds of motion-augmented data is effective for camera-based physiological measurement algorithms, and (3) we demonstrate, through inter-dataset results, the usefulness of motion augmentation for improving the generalization of models for camera-based physiological sensing. We achieve state-of-the-art results on multiple public benchmark datasets, including those with significant motion. We summarize the key findings of this paper about the effectiveness of motion transfer as a data augmentation tool in Sec. 5. We provide our code for augmenting datasets, training using these data, and pre-trained models trained on motion-augmented data (all assets are released with responsible use licenses [5]).

2. Background

Generative Synthetics for Training Models: Statistical generative models [9, 14, 13, 33, 10, 6] capture a probabilistic representation of a dataset from which samples can be drawn. These models are typically trained to mimic the distribution of the training set and can be trained without the need for labels, allowing large sets of data to be used. Facial video generation using generative models has advanced rapidly over recent years [15, 30]. Numerous image-driven works have accomplished the ability to separate identity and pose in source and driving images used for high quality, robust video generation using generative adversarial networks (GANs) [49, 32, 41, 11]. Image-driven facial video generation methods attempt to preserve the identity of a given source image while manipulating the pose based on a driving video to generate a new video. The identity from the driving video is excluded with the help of a keypoint-based motion transfer approach, where keypoints are predicted for both a source image and a driving image in order to model local motion using shifts in the corresponding keypoints [32, 41, 11]. Face video generation that is achieved by using keypoints that take pose and expression into account can be successful for the task of head video generation, but can at times have a loss in source image identity and unwanted temporal artifacts [32, 49, 11]. FaceVid2Vid [41] utilizes canonical keypoints in addition to source and driving image keypoints in order to capture a target person's geometry signature, which includes the shape of the target's face, nose, and eyes. This allows for improved head video generation that minimizes source identity loss while effectively transferring motion from a driving video.

rPPG Models: The principle that photoplethysmography could be performed with a camera and without contact with the body was established by Blazek et al. [2] and replicated in a series of following experiments [37, 39]. The application of more advanced signal processing methods helped make measurement somewhat more robust under real-world conditions [28, 42], as did leveraging knowledge of physiological and physical properties [42]. Yet, these models were still very sensitive to body motions. Neural data-driven models currently achieve state-of-the-art results in most cases [4, 47, 16, 48, 17], but are a function of the data used to train them. While intra-dataset performance is generally strong, inter-dataset performance is often substantively worse. In order to alleviate the dependency on labeled data, several researchers have proposed unsupervised learning procedures [8, 40, 46]. However, most require fine-tuning on a labeled set and also reveal that supervised learning still holds some additional benefit. As an alternative or a complement, generative methods have been suggested to "create" data [21, 43].

rPPG Datasets: As with many health applications, those working in camera physiological measurement face challenges associated with collecting and managing data. Public datasets (such as UBFC-rPPG [3], PURE [35], VIPL-HR [25]) are valuable resources. However, given the challenging nature of the rPPG task researchers have collected and released data under heavily constrained condi-

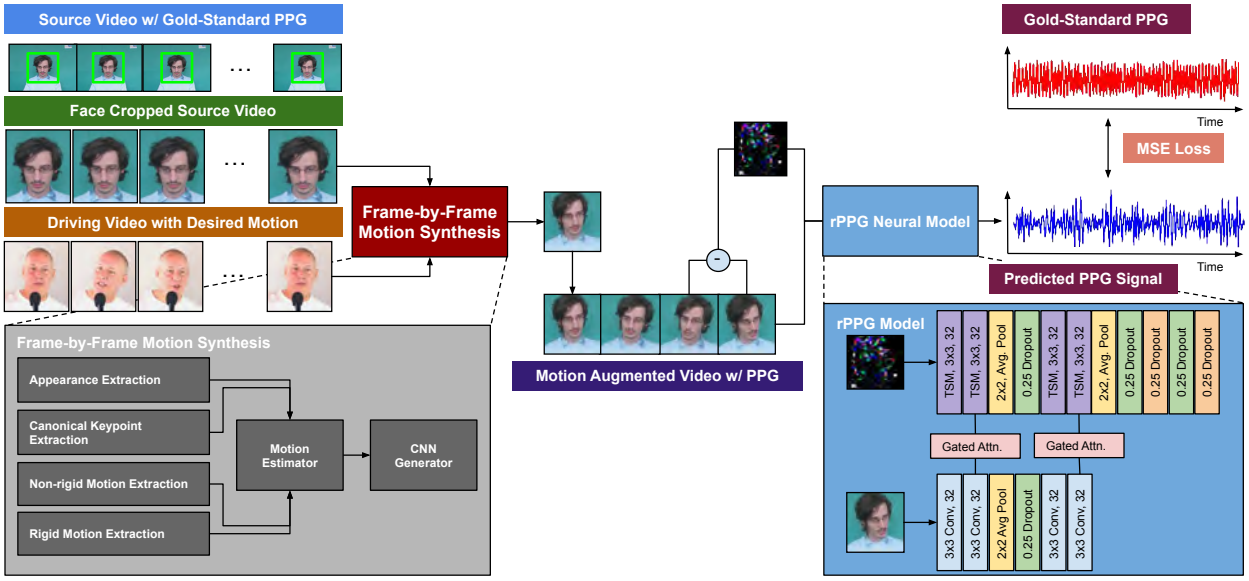


Figure 2: **Motion augmentation and training pipeline.** We augment each frames of a source video with corresponding frames of a randomly selected driving video to create an augmented video with the identity of the source video and motion of the target video. We then train a PPG estimation network on the augmented video with Mean Squared Error Loss.

tions with very little physical motion. More recent datasets (such as UBFC-PHYS [24] and MMPD [38]) contain larger and more natural motions. However, the baseline results on these datasets are not very strong.

3. Motion Augmented rPPG Video Pipeline

We propose neural motion transfer as a data augmentation technique to train machine learning models for predicting physiological measurements, specifically Photoplethysmography (PPG) signal, from facial videos. First, we describe our proposed pipeline to augment facial videos with naturalistic human head motion and expression in section 3.1. Neural motion transfer algorithms often use generative models to synthesize new videos of a person by transferring the rigid head motion and non-rigid facial expressions from a driving video of another person. Since these models generate image pixels from scratch, it is possible that images generated by neural motion transfer algorithms can destroy the underlying physiological signal. Thus, in section 3.2, we provide qualitative evidence to prove that neural motion transfer algorithms do not destroy the original PPG signal, and the original heart rate is preserved. This allows us to effectively use neural motion transfer as a data augmentation technique for training rPPG networks.

3.1. Motion Augmentation Pipeline

In a camera-based physiological sensing (e.g., rPPG) task, a machine learning model is trained on facial videos with time-aligned physiological labels. These may take the form of continuous waveforms (e.g., a gold-standard PPG or a respiration wave) or vital statistics (e.g., heart or breathing rates). In this project, we consider video labels in the form of a PPG signal. The goal of designing a data augmentation strategy is to apply more naturalistic motion to the facial videos without changing the PPG labels.

To apply naturalistic motion to these facial videos, we consider neural talking-head video synthesis models that transfer more naturalistic motion from a *driving* video of a person to the *source* video with PPG signal labels. Our goal is to find a neural motion transfer algorithm that can: (a) inject a large variety of rigid and non-rigid head motions into the source video, (b) not introduce any artifacts that significantly degrade the generated video quality, and (c) maintain the key properties of the underlying PPG signal in terms of frequency information indicating physiological signals like heart rate.

Our pipeline takes in a source video with PPG signal labels from the training data, **S**, and a driving video, **D**, randomly selected from a curated driving video set as inputs for motion augmentation. Both **S** and **D** can be represented as a sequence of frames, respectively $\{s_1, s_2, ..., s_n\}$ and $\{d_1, d_2, ..., d_n\}$. Motion is transferred from driving video **D** to source video **S** on a frame-by-frame basis, such that an output video **Y** represents the motion-augmented sequence of frames $\{y_1, y_2, ..., y_n\}$. Thus we search for a motion transfer algorithm $M(\cdot; \theta)$, such that $y_t = \mathbf{M}(s_t, d_t; \theta)$.

We choose Face-Vid2Vid [41], a neural talking-head synthesis model for transferring motion from a driving video to a source video. The original Face-Vid2Vid paper was intended for teleconferencing applications where a motion-augmented video is generated from a single source image using a driving video. In contrast, we redesign

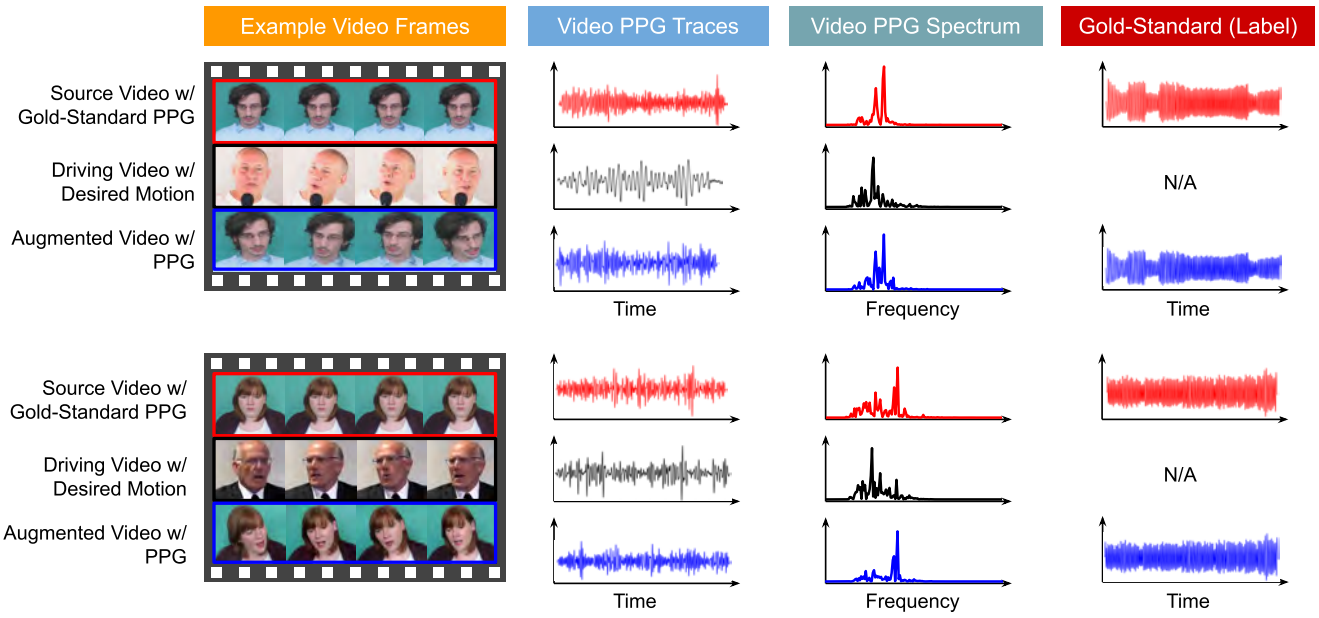


Figure 3: **Preserving physiological signals in motion augmented videos.** We show that applying neural motion transfer preserve the physiological signal corresponding to the heart-rate present in the peak of the frequency spectrum of the source and the augmented video.

and reimplement this algorithm such that each frame of the source video is augmented with motion from the corresponding frame of the driving video. The motion-augmented video **Y**, along with the original PPG signal label, is ultimately used as training data for various deep learning-based camera physiological measurements. This pipeline is shown in Figure 2.

Source Video Datasets: We utilize the UBFC-rPPG [3] and PURE [35] rPPG video datasets as source videos. The UBFC-rPPG dataset contains videos with a very minimal amount of both rigid motion and non-rigid motion, making them ideal for motion augmentation. The PURE dataset contains videos of various tasks with a variety of constrained rigid and non-rigid motion.

Driving Video Datasets: The driving video datasets used include a self-captured, constrained driving video set (CDVS) and the TalkingHead-1KH [41] dataset. The CDVS contains 90 self-captured videos by 5 subjects with heavily constrained, unnatural motion used only for ablation studies to understand the impact of augmenting data with various degrees of rigid and non-rigid motion. The CDVS will be released in the future for research purposes. Talkinghead-1KH is a publicly available, large-scale talking-head video dataset used as a benchmark for Face-Vid2Vid [41] and entirely sourced from YouTube videos. It contains 180K unconstrained videos of people speaking in a variety of real-world contexts, leading to a rich diversity in both rigid and non-rigid motion.

Deep Networks for estimating PPG signal: For our experiments, we focus on using TS-CAN [16] to predict the 1st-order derivative of the PPG signal after training on videos augmented with motion. We also use DeepPhys [4] and EfficientPhys [17] to highlight the consistent benefits of motion augmentation across different neural models.

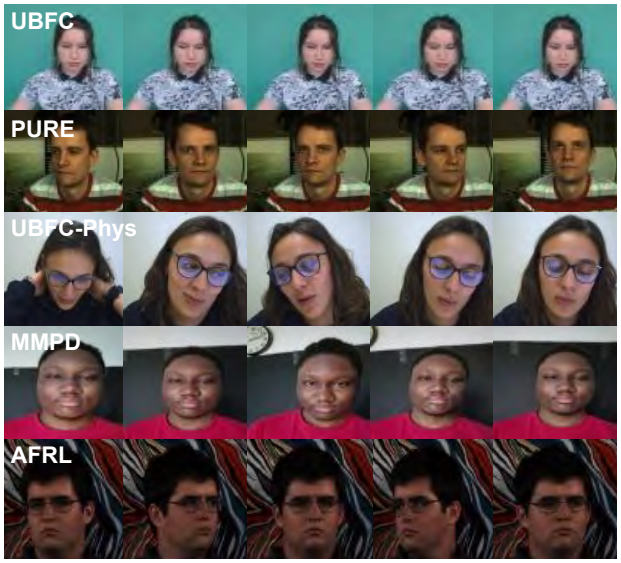
3.2. The Effect of Motion Transfer on PPG

Neural Motion Transfer algorithms are based on generative models where every pixel of the generated image is synthesized by a neural network. While these algorithms succeed in producing photorealistic facial images that are indistinguishable from real images, it is not obvious if the synthesized videos can preserve the underlying PPG signal.

In an ideal world, a motion transfer algorithm is expected to perturb the PPG signal since head motion will induce certain changes in raw pixel intensities. However, the frequency domain analysis of the PPG signal should preserve the peaks related to the heart rate of the patient. It is highly unlikely that the peak frequency of head motion and heart rate will be exactly the same.

Thus, our goal is to first analyze if the motion transfer algorithm of Face-Vid2Vid [41] can preserve the peak heart rate indicated in the frequency domain analysis of the PPG signal extracted from the source video and the synthesized video. In Figure 3, we qualitatively analyze the time-domain and frequency domain PPG signals extracted from the source and the synthesized (augmented) video. We choose a simple unsupervised algorithm, POS [42], for extracting the PPG signal from all the facial videos to focus more on the original signal contents in the videos. We observe that the most prominent frequency peak, corresponding to the heart rate, is the same for the source video and the augmented video. This appears to also hold true across

Table 1: A Summary of the rPPG Benchmark Datasets.



Dataset	Subjects / Videos	Motion Tasks
UBFC-rPPG	42 / 42	Stationary
PURE	10 / 59	Stationary, Talking, Rotation, Translation
UBFC-Phys	56 / 168	Stationary, Talking, Head Rotation
MMPD	33 / 660	Stationary, Talking, Walking, Head Rotation
AFRL	25 / 300	Stationary, Head Rotation

different appearances and motion conditions, both in the source videos and the driving videos. We present additional results in the supplementary materials. Thus, we can effectively claim that motion transfer algorithms like FaceVid2Vid [41] do preserve the underlying physiological signal, like heart rate, and they can be a very effective tool for large scale augmentation of training videos for PPG estimation tasks. Our quantitative experimental results show that deep neural networks for camera physiological measurement can take advantage of this to significantly improve model performance by training on motion-augmented data.

4. Experiments

We consider five datasets for training and evaluation, **UBFC-rPPG** [3], **PURE** [35], **UBFC-PHYS** [24], **AFRL** [7], and **MMPD** [38] (see Table 1). They consist of facial videos and corresponding gold-standard PPG signal labels. We use some of these datasets for augmentation with neural motion transfer and training the rPPG models, and use the rest to evaluate different aspects of the effectiveness of neural motion augmentation. To our knowledge, we perform the most extensive inter-dataset evaluation of PPG estimation to date, testing on five independent test datasets.

Implementation Details: For all of the results reported in this section, the predicted PPG signals were filtered using a band-pass signal with cut-offs 0.75 Hz and 2.5 Hz. The heart rate was calculated based on the PPG signal using the Fast Fourier Transform (FFT), with a measurement window of the video length for UBFC-rPPG, PURE, UBFC-PHYS, and MMPD. To evaluate the AFRL dataset, a measurement window of 30 seconds was utilized for heart rate calculations. All networks were trained using an NVIDIA RTX A4500 and PyTorch [27] implementations in the publicly available rPPG-Toolbox [18]. A cyclic learning rate scheduler was utilized with 30 epochs, a learning rate of 0.009, and a batch size of 4 for both training and inference.

4.1. Training with Motion Augmented Data

In Table 2, we compare the performance of a supervised PPG estimation network, TS-CAN [16], trained on existing video datasets and motion-augmented versions of those datasets. We also show the performance of unsupervised methods for comparison. For the sake of space and clarity, the following tables only show the mean absolute error (MAE) in heart rate estimation for each video (and the corresponding mean absolute percentage error (MAPE)). Equivalent tables with root mean squared error (RMSE) and Pearson correlation metrics can be found in our supplementary material. The driving videos used for augmentation in Table 2 contain significant amounts of unconstrained motion – both rigid and non-rigid.

We observe that training TS-CAN on augmented videos produces state-of-the-art (SOTA) performance in most cases. Additionally, we observe that in most cases, the augmented versions outperform the non-augmented versions, with a gain in performance up to 75% and an average gain of 26%. However, when comparing the performance of MA-PURE versus PURE when tested on UBFC-PHYS, we note a minor drop in performance rather than an improvement due to the difficulty in effectively augmenting the PURE dataset. This is because the PURE dataset already contains significant amounts of rigid motion, and when augmented, it may provide training data with artifacts that make the learned rPPG task less useful in the face of a highly unconstrained dataset with natural rigid and non-rigid motion.

Details: We utilize all downloadable videos from the TalkingHead-1KH [41] dataset as our driving videos for augmenting various PPG estimation datasets with motion. We analyze the videos using OpenFace [1] to obtain the intensity (0 to 5) of 17 Facial Action Units (AUs) and the head pose rotations R_x, R_y, and R_z in radians (rad). To generate MAUBFC-rPPG, we choose driving videos from a pool of 60 driving videos with a range of mean standard deviation in head pose rotations from 0.10 to 0.14 rad to augment as much rigid motion as possible into a source video dataset that has very little of both rigid and non-rigid motion. We do not constrain for non-rigid motion in this case, so we observe a wide range of mean standard deviation in facial AUs from 0.15 to 0.5 intensity. To generate MAPURE, we choose driving videos with a range of mean standard deviation in facial AUs from 0.45 to 0.55 intensity to augment as much non-rigid motion as possible into a source

video dataset that has very little non-rigid motion. We do not constrain for rigid motion in this case, so we observe a wide range of mean standard deviation in head pose rotations from 0.03 to 0.14 rad.

4.2. Effect of Motion Types

A key question in designing a motion augmentation strategy is deciding what type of motion should be applied to obtain the best performance on a certain evaluation dataset. To answer this question, we separately analyze two types of motion: rigid and non-rigid, by augmenting training data with different magnitudes of motion. Rigid motion refers to head pose rotation, while having minimal change in facial action units or expressions. Non-rigid motion refers to changes in facial expression, i.e. motion in facial action units for various tasks like talking, while having minimal head pose rotation.

Rigid Motion: For rigid motion, we consider UBFC-rPPG as training data, which has very little head motion and AFRL as test data which has large variations in rigid head motion. We classify videos in the AFRL dataset into different rigid head motion categories: 'very small motion', 'small motion' (10 deg rotation per sec), and 'large motion' (30 deg rotation per sec). Based on this categorization, we also select driving videos from our captured CDVS to have 'small motion' and 'large motion' using the mean standard deviation in estimated head pose rotations across all the frames of a video. Specifically, for 'small motion' we used mean standard deviation between 0.03 to 0.07 rad and for 'large motion' between 0.10 to 0.14 rad. These parameters are chosen to roughly match the distribution of head pose rotation in 'small motion' and 'large motion' categories of AFRL. We then use these videos from the CDVS dataset to augment the source videos of UBFC-rPPG to create 3 separate categories of augmented videos for 'very small motion' (which is the original UBFC-rPPG dataset), 'small motion', and 'large motion' respectively. We then train TS-CAN on augmented data in each category and test on the same categories of the AFRL dataset. We present these results in Table 3.

We observe that when the test data of AFRL has 'very small motion' or 'small motion', augmenting UBFC-rPPG with small motion performs the best. In fact, augmenting with large motion worsens the result by 19% in this case. However, when testing on the 'large motion' split of AFRL, UBFC-rPPG augmented with 'large motion' outperforms 'small motion' by 13.5% and 'very small motion' by 52%.

Non-rigid Motion: For non-rigid motion, we also consider UBFC-rPPG as training data since it has very little motion, and the speech task of the PURE dataset [35] as the test data which has significant non-rigid head motion. We also augment the UBFC-rPPG dataset with non-rigid head motion from our captured CDVS with 'small' and 'large' non-rigid motions and minimal rigid motion. For this experiment, we define small non-rigid motion to have a range of mean standard deviation in facial action units from 0.15 to 0.25 intensity and large non-rigid motion to have a range of mean standard deviation in facial action units from 0.45 to 0.55 intensity. We train TS-CAN on 'small' and 'large' motion augmented versions of UBFC-rPPG and test it on the speech task of PURE, in which recorded participants are asked to talk while avoiding head movements as much as possible. We observe that augmenting UBFC-rPPG with 'large' non-rigid motion improves over 'very small motion' (original UBFC-rPPG) by 89.2% and over 'small' non-rigid motion by 37%.

4.3. Effect of Multiple Augmentations

We consider whether it is plausible to augment the same source video with multiple driving videos using neural motion transfer. Thus, the newly augmented dataset has the same number of identities as the original dataset but a significantly larger variation in motions. Our goal is to analyze how many times one can augment a single source video before the performance starts to saturate or drop. We consider UBFC-rPPG as training data that we augment with randomly sampled driving videos from the TalkingHead-1K dataset to produce MAUBFC-rPPG. We augment the same source video from 1 to 4 times with different driving videos and evaluate on the PURE [35] dataset and the UBFC-PHYS [24] dataset and report the results in Table 5. We notice that the results saturate pretty quickly and can start to decline after augmenting more than 2 times.

4.4. Synthetic vs Naturalistic Head Motion

In order to further evaluate the impact of motion transfer as a data augmentation technique, we explore whether data augmented with natural head motion using a neural motion transfer algorithm is better than augmenting data with synthetically generated motion using parametric motion animation, as used in the SCAMPS dataset [23]. The SCAMPS dataset consists of synthetic human heads that can be rigged to induce parametric motion. We consider 200 such samples from the SCAMPS dataset that consist of significant synthetically generated rigid and non-rigid head motion (ID 1801 to 2000) as SCAMPS-200 (Motion). We then take instances from the SCAMPS dataset with no head motion (ID 1 to 200) and augment them with naturalistic head motion using our motion synthesis pipeline and a subset of driving videos from the TalkingHead-1KH dataset to produce MASCAMPS-200. We choose driving videos with a range of mean standard deviation in AUs from 0.35 to 0.40 intensity and a range of mean standard deviation in head pose rotations from 0.05 to 0.125 rad. Note that both SCAMPS-200 (Motion) and MASCAMPS-200 consist of synthetics with the same number of identities, with the only difference

Table 2: **Evaluation across all datasets.** We motion-augment two training datasets, UBFC-rPPG and PURE, to create MAUBFC-rPPG and MAPURE, respectively. We observe that the motion-augmented versions produce significant improvements (shown in bold).

		Testing Set									
		UBFC-rPPG		PURE		UBFC-PHYS		AFRL		MMPD	
Training Set	**Method**	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓	MAE↓	MAPE↓
Unsupervised	Green	19.82	18.78	10.09	10.28	13.45	16.00	7.01	9.24	16.27	20.09
	ICA	14.70	14.34	4.77	4.47	8.00	9.48	6.77	8.96	13.10	16.33
	CHROM	3.98	3.78	5.77	11.52	4.68	6.20	5.41	7.95	8.85	11.93
	POS	4.00	3.86	3.67	7.25	4.62	6.29	6.93	10.00	8.18	11.12
UBFC-rPPG	TS-CAN	-	-	4.55	4.67	5.56	7.25	4.24	5.84	8.74	10.51
MAUBFC-rPPG	TS-CAN	-	-	**1.14**	**1.30**	**3.93**	**5.24**	2.67	3.65	**6.80**	**7.97**
PURE	TS-CAN	1.34	1.55	-	-	4.43	5.89	2.63	3.51	8.96	10.33
MAPURE	TS-CAN	**1.03**	**1.17**	-	-	4.39	5.90	**2.37**	**3.26**	8.08	9.54
MAUBFC-RPPG VS. UBFC-RPPG		-	-	+74.95%	+72.16%	+29.32%	+27.72%	+37.03%	+37.50%	+22.20%	+24.17%
MAPURE VS. PURE		+23.13%	+24.52%	-	-	+0.90%	-0.17%	+9.89%	+7.12%	+9.82%	+7.65%

MAE = Mean Absolute Error in HR estimation (Beats/Min), MAPE = Mean Absolute Percentage Error in HR estimation

Table 3: **Effect of Motion Types – Rigid.** We augment UBFC-rPPG with various types of rigid head motions and test on AFRL [7].

		Testing Set			
		No Motion	Small Motion	Large Motion	All Motion
Training Set	**Rigid Motion**	MAE↓	MAE↓	MAE↓	MAE↓
UBFC-rPPG	Very Small	1.00	2.28	7.59	4.72
MAUBFC-rPPG	Small	**0.84**	**1.44**	4.21	**3.19**
MAUBFC-rPPG	Large	1.00	1.78	**3.64**	3.39
OURS VS. BASELINE		+16.0%	+36.8%	+52.0%	+32.4%

Table 4: **Effect of Motion Types – Non-rigid.** We augment UBFC-rPPG with various types of non-rigid motions (expressions) and test on the speech task, in PURE [35].

		Testing Set Non-rigid Motion Task	
Training Set	**Non-Rigid Motion**	MAE↓	MAPE↓
UBFC-rPPG	Very Small	10.84	11.40
MAUBFC-rPPG	Small	1.86	2.94
MAUBFC-rPPG	Large	**1.17**	**1.55**
OURS VS. BASELINE		+89.2%	+86.4%

Table 5: **Effect of Multiple Augmentations.** Augmenting each source video of UBFC-rPPG 1x, 2x, 3x, and 4x, we test on PURE and UBFC-PHYS datasets. The best results are shown in bold.

			Testing Set			
			PURE		UBFC-PHYS	
Training Set	**Size**	**Subjects**	MAE↓	MAPE↓	MAE↓	MAPE↓
UBFC-rPPG	42	42	4.55	4.67	5.56	7.25
MAUBFC-rPPG	42	42	1.14	1.30	3.93	5.24
MAUBFC-rPPG 2x	84	42	1.12	1.29	**3.90**	**5.22**
MAUBFC-rPPG 3x	126	42	**1.11**	**1.27**	3.97	5.31
MAUBFC-rPPG 4x	168	42	1.19	1.33	4.10	5.40
OURS VS. BASELINE			+2.63%	+2.31%	+0.76%	+0.38%

Table 6: **Synthetic vs Naturalistic Head Motion.** We compare the effect of adding head motions to SCAMPS and UBFC-rPPG and contrast this with using motion data in SCAMPS. Average time for augmenting each frame of a sequence is presented. The best results are shown in bold.

	Testing Set				
	PURE		AFRL		
Training Set	MAE↓	MAPE↓	MAE↓	MAPE↓	Synth. Time
SCAMPS-200 (No motion)	10.29	11.09	7.75	10.54	37.00s
SCAMPS-200 (Motion)	5.38	5.42	7.25	10.20	37.00s
UBFC-rPPG	4.55	4.67	4.72	6.59	-
MASCAMPS-200	4.67	4.22	5.00	6.69	1.20s
MAUBFC-rPPG	**1.14**	**1.30**	**3.24**	**4.37**	2.39s
MASCAMPS VS. SCAMPS	+13.2%	+22.1%	+31.1%	+34.4%	+96.8%
MAUBFC VS. UBFC-RPPG	+74.4%	+72.2%	+31.4%	+33.7%	-

Avg. Synth. Time = time (in seconds) to synthesize a frame

being synthetic and naturalistic head motion, respectively.

We train TS-CAN on both SCAMPS-200 (Motion) and MASCAMPS-200, and evaluated its performance on PURE and AFRL, as shown in Table 6. We observed that adding naturalistic motion improved performance by 13.2% on PURE and 31.1% on AFRL compared to synthetically generated motion. It is worth noting that the average time taken to add synthetic motion to each frame of a sequence is 37 seconds, compared to only 1.2 seconds for adding naturalistic motion using the neural motion transfer algorithm. For comparison, we also included real-world training data, UBFC-rPPG, which showed that having real images significantly improved performance over synthetic images. Furthermore, the only way to augment real images is to use the neural motion transfer algorithm, as parametric rigged head motion cannot be applied to real data.

4.5. Effect of PPG Estimation Models

It is important to decouple any data augmentation technique from additional factors that affect its usefulness for a given set of training data. One such factor is the neural network model used for training and evaluation. Thus,

Table 7: **Effect of PPG Estimation Models.** We train different PPG estimation networks on UBFC-rPPG and MAUBFC-rPPG and evaluate on PURE. The best results are shown in bold.

		Testing Set PURE	
Training Set	**Method**	MAE↓	MAPE↓
UBFC-rPPG	DeepPhys [4]	5.14	4.90
MAUBFC-rPPG	DeepPhys	1.24	1.56
UBFC-rPPG	EfficientPhys [17]	4.95	4.56
MAUBFC-rPPG	EfficientPhys	1.45	1.76
UBFC-rPPG	TS-CAN [16]	4.55	4.67
MAUBFC-rPPG	TS-CAN	**1.14**	**1.30**

in addition to TS-CAN, we evaluate two more rPPG models - DeepPhys and EfficientPhys - in Table 7. We utilize MAUBFC-rPPG as training data and evaluate on PURE. We observe that the results are reasonably consistent across neural rPPG models.

5. Discussion

Can motion augmented videos achieve SOTA results? In Section 3.2, we demonstrated that motion transfer algorithms can be used to create motion-augmented videos while still preserving the variations in skin appearance from the cardiac pulse. This means we can augment the training data to create novel samples with greater variance than those in the original set. We conducted a set of systematic empirical validation studies that show that these videos can be used to effectively train rPPG models that generalize to independent benchmark datasets (see Table 2). Cross-dataset experiments show a 23.1% reduction in HR MAE on UBFC-rPPG when using the motion-augmented PURE datasets for training and a 74.95% reduction in HR MAE on PURE when using the motion-augmented UBFC-rPPG dataset for training. Other than PURE, the largest gains were observed training on MAUBFC-rPPG and testing on videos with large rigid and/or non-rigid head motions (UBFC-PHYS: 29.32%, AFRL: 37.03% and MMPD: 22.20% reduction in HR MAE).

What type of motion is best to augment? In learning tasks, designing training data that matches the distribution of the testing data is advantageous. Does augmenting motion in the training set that is similar to that in a testing set lead to optimal results? Our experiments show that this is the case for both rigid (see Table 3) and non-rigid (see Table 4) head motions. Beyond the type of motion, if the motions have a larger magnitude, then including larger magnitude motions in the training set seems to have a benefit.

Does the effect of motion augmentation saturate? While source videos with gold-standard PPG data might be limited, it is possible to augment each source video with many different motions by leveraging large video datasets like TalkingHead-1KH [41]. We looked at whether augmenting the same source videos multiple times with different motions improved results. We observed that most of the improvements were obtained by augmenting the data once. Incremental improvements were obtained by augmenting a second or third set of videos, but the results quickly saturated (see Table 5). This is presumably due to the fact that we were not augmenting other aspects of the subjects' appearance (e.g., skin tone, identity, etc.).

Is natural motion augmentation best? Finally, there are different methods for synthesizing motion in video data. State-of-the-art synthetic datasets are generated using parametric computer graphics, but they require a large amount of computational resources. As a result, if the motions present in those datasets are sub-optimal, it is costly to remedy. Can motion augmentation add motions to these datasets "cheaply" and still obtain the performance benefits of graphics approaches? Our results in Table 6 suggest that the motion in the SCAMPS dataset is sub-optimal when tested on PURE and AFRL. We were able to obtain a performance gain by using our simple motion augmentation.

What are the limitations of our method? There are several limitations that we would like to highlight. First, detecting artifacts in augmented videos is not always trivial, and we used motion driving videos without extreme motions to mitigate the chance of augmented videos with unnatural artifacts. We did not conduct an extensive investigation to determine if other physiological changes (e.g., respiration) that might be correlated with the PPG signal are preserved in the augmented videos. However, empirically we have shown that these data can be used to effectively train *heart rate* estimation models. We did not thoroughly test whether the waveform dynamics, beyond the dominant frequency, were faithfully preserved in the augmented videos. For tasks such as blood pressure estimation from PPG waveforms, morphological information is important. Our method does not address diversity across other dimensions, particularly identity diversity. The augmented datasets we produced, while contributing to significant improvements over the baselines, only contain examples from the same number of subjects as the original dataset. Other synthetic generation techniques [43] could help in these regards alongside more generic neural rendering approaches such as ours.

6. Conclusion

Motion artifacts are a significant challenge in camera physiological measurement. The PPG signal presents only very subtle changes in diffuse light reflections from the skin, whereas motion of the head causes large changes in specular reflections. We have shown that neural motion augmentation can be used to create training data with more motion, while still preserving the pulse signal. Motion augmented data leads to up to 75% reduction in error in cross-dataset experiments compared to training with unaugmented data.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[2] Vladimir Blazek, Ting Wu, and Dominik Hoelscher. Near-infrared ccd imaging: Possibilities for noninvasive and contactless 2d mapping of dermal venous hemodynamics. In *Optical Diagnostics of Biological Fluids V*, volume 3923, pages 2–9. International Society for Optics and Photonics, 2000.

[3] 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.

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

[5] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 778–788, 2022.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[7] 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 *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 1462–1469. IEEE, 2014.

[8] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3995–4004, 2021.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[11] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3387–3396, 2022.

[12] In Cheol Jeong and Joseph Finkelstein. Introducing contactless blood pressure assessment using a high speed video camera. *Journal of medical systems*, 40(4):77, 2016.

[13] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 109:839–862, 2020.

[16] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *NeurIPS*, 2020.

[17] Xin Liu, Brian L Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement. *arXiv preprint arXiv:2110.04447*, 2021.

[18] Xin Liu, Xiaoyu Zhang, Girish Narayanswamy, Yuzhe Zhang, Yuntao Wang, Shwetak Patel, and Daniel McDuff. Deep physiological sensing toolbox. *arXiv preprint arXiv:2210.00716*, 2022.

[19] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys (CSUR)*, 2021.

[20] Daniel McDuff, Roger Cheng, and Ashish Kapoor. Identifying bias in ai using simulation. 2018.

[21] Daniel McDuff, Xin Liu, Javier Hernandez, Erroll Wood, and Tadas Baltrusaitis. Synthetic data for multi-parameter camera-based physiological sensing. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2021.

[22] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. *Advances in Neural Information Processing Systems*, 32:5403–5414, 2019.

[23] Daniel McDuff, Miah Wander, Xin Liu, Brian L Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. Scamps: Synthetics for camera measurement of physiological signals. *arXiv preprint arXiv:2206.04197*, 2022.

[24] 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.

[25] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. *arXiv preprint arXiv:1810.04927*, 2018.

[26] Ewa Nowara, Daniel Mcduff, Ashutosh Sabharwal, and Ashok Veeraraghavan. Seeing beneath the skin with computational photography. *Communications of the ACM*, 65(12):90–100, 2022.

[27] 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, 2019.

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

[29] Ming-Zher Poh, Yukkee Cheung Poh, Pak-Hei Chan, Chun-Ka Wong, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Daniel Wai-Sing Chu, and

Chung-Wah Siu. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart*, 104(23):1921–1928, 2018.

[30] Tong Sha, Wei Zhang, Tong Shen, Zhoujun Li, and Tao Mei. Deep person generation: A survey from the perspective of face, pose and cloth synthesis. *ACM Computing Surveys*, 2021.

[31] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

[32] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[34] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6, 2018.

[35] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014.

[36] Lech Świrski and Neil Dodgson. Rendering synthetic ground truth images for eye tracker evaluation. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 219–222, 2014.

[37] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007.

[38] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: Multi-domain mobile video physiology dataset, 2023.

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

[40] Hao Wang, Euijoon Ahn, and Jinman Kim. Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2431–2439, 2022.

[41] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10034–10044, 2020.

[42] 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, 2017.

[43] 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*, pages 20587–20596, 2022.

[44] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3681–3691, 2021.

[45] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3756–3764, 2015.

[46] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets. *arXiv preprint arXiv:2210.03115*, 2022.

[47] 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*, pages 151–160, 2019.

[48] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. *arXiv preprint arXiv:2111.12082*, 2021.

[49] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.