Motion Matters: Neural Motion Transfer for Better Camera Physiological Sensing Supplementary Material

Akshay Paruchuri[1], Xin Liu[2], Yulu Pan[1], Shwetak Patel[2], Daniel McDuff[2,*], Soumyadip Sengupta[1,*]
[1]UNC Chapel Hill [2]University of Washington
{akshay, ronisen, yulupan}@cs.unc.edu, {xliu0, shwetak, dmcduff}@cs.washington.edu

A. Overview of Appendices

Our appendices contain the following additional details and results:

- Tables 1, 2, 3, 4, and 5 in Section B include additional metrics, RMSE and the Pearson correlation coefficient, for experimental results included in the main paper. We also provide Scatter and Bland-Altman Plots in Figures 1 and 2 that correspond to the overall results shown in Table 3. Section B.1 contains additional details regarding our experimental process. Section B.2 contains an additional experiment toward the effect of scaling a dataset using motion augmentation.
- Section C, and the corresponding Table 6, describe and show intra-dataset results using the PURE [15] dataset.
- Section D briefly describes additional materials that we provide for research purposes, including our motion augmentation pipeline code, pre-trained models, and motion analysis scripts. Additionally, Section D.1 shows more qualitative examples of the effects of motion augmentation on the underlying PPG signal, as well as briefly addresses prior works involving analysis of physiological signals in deep fake videos.
- Sections E.1, E.2, and E.3, provide further details on source, driving, and evaluation datasets used in the main paper.
- Section F is our broader impact statement.

B. Experimental Results

The following section contains tables that include additional metrics, RMSE and the Pearson correlation coefficient, for experimental results already included in the main paper. We also provide scatter and Bland-Altman plots in Figures 1 and 2 that correspond to results shown in Table 3.

Table 1. **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 [15]. The best results are shown in bold.

		Testing Set Non-rigid Motion Task			
Training Set	**Non-Rigid Motion**	MAE↓	RMSE↓	MAPE↓	ρ↑
UBFC-rPPG	Very Small	10.84	24.64	11.40	0.46
MAUBFC-rPPG	Small	1.86	2.79	2.94	**0.99**
MAUBFC-rPPG	Large	**1.17**	**1.90**	**1.55**	**0.99**
OURS VS. BASELINE		**+89.21%**	**+92.29%**	**+86.40%**	+0.00%

MAE = Mean Absolute Error in HR estimation (Beats/Min), RMSE = Root Mean Square Error in HR estimation (Beats/Min), MAPE = Mean Absolute Percentage Error in HR estimation, ρ = Pearson Correlation in HR estimation

B.1. Experimental Details

The predicted PPG signals were filtered using a band-pass filter with cut-offs 0.75 Hz and 2.5 Hz. The heart rate was calculated based on the predicted 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 [14] implementations in the publicly available rPPG-Toolbox [8]. All pre-processing steps and evaluation was also done in a reproducible fashion using the toolbox. The AdamW [9] optimizer, a mean squared error (MSE) loss, and 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.

For both the UBFC-rPPG dataset and PURE datasets, all subjects were augmented with motion. For our experiments, we elect to use all of the subjects in our training and train to the very last epoch. A variety of appropriately titled pre-trained models corresponding to results in the main paper and the appendices are included alongside our code in the *pretrained_models* folder.

Table 2. **Effect of Motion Types – Rigid.** We augment UBFC-rPPG with various types of rigid head motions and test on AFRL [5]. The best results are shown in bold.

		Testing Set															
		No Motion				Small Motion				Large Motion				All			
Training Set	**Rigid Motion**	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑
UBFC-rPPG	Very Small	1.00	3.86	1.48	0.95	2.28	6.36	3.44	0.85	7.59	12.91	10.99	0.49	4.72	10.01	6.59	0.67
MAUBFC-rPPG	Small	**0.84**	**3.25**	**1.18**	**0.96**	**1.44**	**4.44**	**2.03**	**0.93**	4.21	9.11	5.96	0.74	**3.19**	**7.96**	**4.36**	**0.79**
MAUBFC-rPPG	Large	1.00	3.61	1.37	0.96	1.78	5.23	2.49	0.90	**3.64**	**8.14**	**5.12**	**0.78**	3.39	8.26	4.58	0.77
OURS VS. BASELINE		**+16.00%**	**+15.80%**	**+20.27%**	**+1.05%**	**+36.84%**	**+30.19%**	**+40.99%**	**+9.41%**	**+52.04%**	**+36.95%**	**+53.41%**	**+59.18%**	**+32.42%**	**+20.48%**	**+33.84%**	**+17.91%**

MAE = Mean Absolute Error in HR estimation (Beats/Min), RMSE = Root Mean Square Error in HR estimation (Beats/Min), MAPE = Mean Absolute Percentage Error in HR estimation, ρ = Pearson Correlation in HR estimation

Table 3. **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↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑
Unsupervised	Green	19.82	31.49	18.78	0.37	10.09	23.85	10.28	0.34	13.45	19.11	16.00	0.31	7.01	12.52	9.24	0.52	16.27	21.74	20.09	-0.04
	ICA	14.70	23.71	14.34	0.53	4.77	16.70	4.47	0.72	8.00	13.51	9.48	0.48	6.77	12.25	8.96	0.51	13.10	17.84	16.33	0.03
	CHROM	3.98	8.72	3.78	0.89	5.77	14.93	11.52	0.81	4.68	8.09	6.20	0.77	5.41	10.71	7.95	0.60	8.85	**12.77**	11.93	0.29
	POS	4.00	7.58	3.86	0.92	3.67	11.82	7.25	0.88	4.62	8.02	6.29	0.78	6.93	11.89	10.00	0.49	8.18	13.04	11.12	0.31
UBFC-rPPG	TS-CAN	-	-	-	-	4.55	14.47	4.67	0.80	5.56	9.88	7.25	0.68	4.24	8.72	5.84	0.75	8.74	15.55	10.51	0.25
MAUBFC-rPPG	TS-CAN	-	-	-	-	**0.96**	**4.17**	**1.13**	**0.97**	**3.93**	**7.50**	**5.24**	**0.81**	2.67	6.55	3.65	0.85	**6.80**	14.20	**7.97**	**0.29**
PURE	TS-CAN	1.34	3.01	1.55	**0.99**	-	-	-	-	4.43	8.12	5.89	0.78	2.63	7.35	3.51	0.82	8.96	16.59	10.33	0.15
MAPURE	TS-CAN	**1.03**	**2.70**	**1.17**	**0.99**	-	-	-	-	4.39	8.10	5.90	0.78	**2.37**	**6.28**	**3.26**	**0.87**	8.08	15.38	9.54	0.18
MAUBFC-RPPG VS. UBFC-RPPG		-	-	-	-	**+78.9%**	**+71.18%**	**+75.8%**	**+21.25%**	**+29.32%**	**+24.09%**	**+27.72%**	**+19.12%**	**+37.03%**	**+24.89%**	**+37.50%**	**+13.33%**	**+22.20%**	**+8.68%**	**+24.17%**	**+16.00%**
MAPURE VS. PURE		**+23.13%**	**+10.30%**	**+24.52%**	+0.00%	-	-	-	-	**+0.90%**	**+0.25%**	**-0.17%**	+0.00%	**+9.89%**	**+14.56%**	**+7.12%**	**+6.10%**	**+9.82%**	**+7.29%**	**+7.65%**	**+20.00%**

MAE = Mean Absolute Error in HR estimation (Beats/Min), RMSE = Root Mean Square Error in HR estimation (Beats/Min), MAPE = Mean Absolute Percentage Error in HR estimation, ρ = Pearson Correlation in HR estimation

Table 4. **Naturalistic vs Synthetic 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				Per Frame
Training Set	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑	Synthesis Time
SCAMPS-200 (No motion)	10.29	23.81	11.09	0.35	7.75	13.08	10.54	0.48	37s
SCAMPS-200 (Motion)	5.38	16.98	5.42	0.72	7.25	12.85	10.20	0.48	37s
Wang et al. [18]	7.40	22.45	6.13	0.44	-	-	-	-	N/A
UBFC-rPPG	4.55	14.47	4.67	0.80	4.72	10.01	6.59	0.67	-
MASCAMPS-200	4.67	16.35	4.22	0.75	5.00	10.10	6.69	0.67	**1.20s**
MAUBFC-rPPG	**0.96**	**4.17**	**1.13**	**0.97**	**3.24**	**7.89**	**4.37**	**0.79**	2.39s
MASCAMPS VS. SCAMPS BASELINE	**+13.20%**	**+3.71%**	**+22.14%**	**+4.17%**	**+31.03%**	**+21.40%**	**+34.41%**	**+39.58%**	**+96.76%**
MAUBFC VS. UBFC-RPPG BASELINE	**+78.9%**	**+71.18%**	**+75.8%**	**+21.25%**	**+31.36%**	**+21.18%**	**+33.69%**	**+17.91%**	-

MAE = Mean Absolute Error in HR estimation (Beats/Min), RMSE = Root Mean Square Error in HR estimation (Beats/Min), MAPE = Mean Absolute Percentage Error in HR estimation, ρ = Pearson Correlation in HR estimation, Synthesis Time = the amount of time (in seconds) it takes to synthesize a single frame, when relevant

Table 5. **Effect of rPPG 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↓	RMSE↓	MAPE↓	ρ↑
UBFC-rPPG	DeepPhys	5.14	17.20	4.90	0.72
MAUBFC-rPPG	DeepPhys	1.24	6.01	1.56	**0.97**
UBFC-rPPG	PhysNet	8.06	19.71	13.67	0.61
MAUBFC-rPPG	PhysNet	2.38	11.29	2.44	0.88
UBFC-rPPG	TS-CAN	4.55	14.47	4.67	0.80
MAUBFC-rPPG	TS-CAN	**0.96**	**4.17**	**1.13**	**0.97**

MAE = Mean Absolute Error in HR estimation (Beats/Min), RMSE = Root Mean Square Error in HR estimation (Beats/Min), MAPE = Mean Absolute Percentage Error in HR estimation, ρ = Pearson Correlation in HR estimation

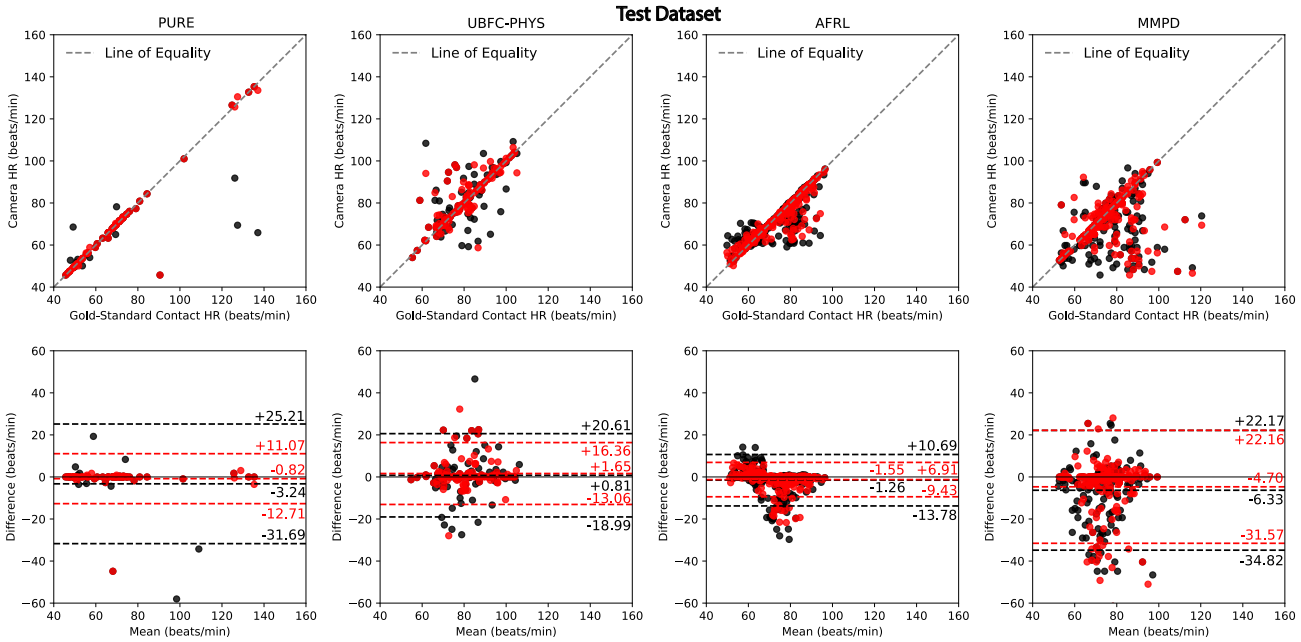


Figure 1. **Scatter and Bland-Altman Plots.** Scatter (top row) and Bland-Altman (bottom row) plots for models trained on UBFC-rPPG (black) and MAUBFC-rPPG (red) and tested on (from left to right), PURE, UBFC-PHYS, AFRL, and MMPD.

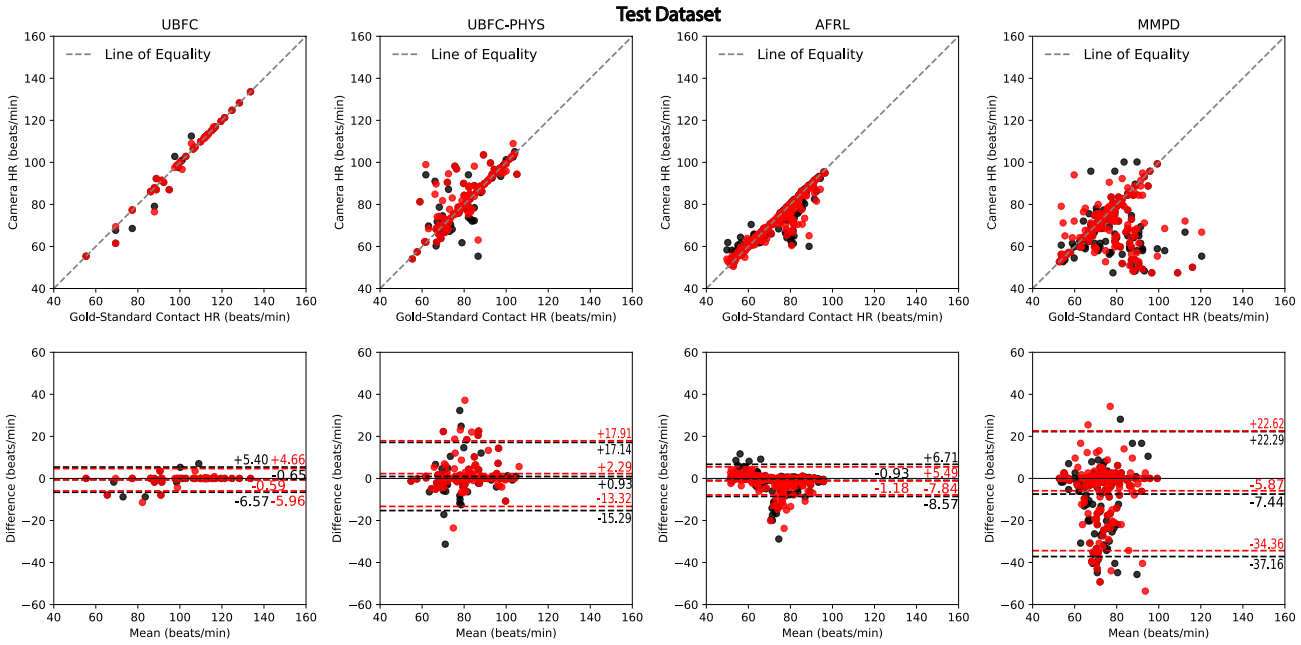


Figure 2. **Scatter and Bland-Altman Plots.** Scatter (top row) and Bland-Altman (bottom row) plots for models trained on PURE (black) and MAPURE (red) and tested on (from left to right), UBFC-rPPG, UBFC-PHYS, AFRL, and MMPD.

For Table 4, we consider 200 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 being synthetic and naturalistic head motion, respectively.

B.2. 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 [15] dataset and the UBFC-PHYS [13] dataset and report the results in Table 7. We notice that the results saturate pretty quickly and can start to decline after augmenting more than 2 times. This is presumably due to the fact that we were not augmenting other aspects of the subjects' appearance (e.g., skin tone, identity, etc.).

C. Intra-dataset Results

We include intra-dataset results not included in the main paper here for reference. We utilize all of the tasks from the PURE dataset. We train on subjects 1, 2, 3, 4, and 5 and then test on subjects 6, 7, 8, 9, and 10. We then train on subjects 6, 7, 8, 9, and 10 and then test on subjects 1, 2, 3, 4, and 5. We average the results from these two experiments and repeat the aforementioned process for the motion-augmented version of PURE. We find that motion augmentation helps as an intra-dataset augmentation technique.

Table 6. **PURE Intra-dataset Results.** We use motion augmentation to augment half of the subjects in the PURE dataset at a time, while testing on the corresponding other half. The averaged results are shown below, with the best result in each column bolded.

	Testing Set PURE			
Training Set	MAE↓	RMSE↓	MAPE↓	ρ ↑
PURE	2.52	8.92	2.55	0.92
MAPURE	**1.61**	**5.50**	**1.77**	**0.97**
OURS VS. BASELINE	**+36.1%**	**+38.34%**	**+30.59%**	**+5.43%**

MAE = Mean Absolute Error in HR estimation (Beats/Min), RMSE = Root Mean Square Error in HR estimation (Beats/Min), MAPE = Mean Absolute Percentage Error in HR estimation, ρ = Pearson Correlation in HR estimation

Table 7. **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		Testing Set 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	0.96	1.13	3.93	5.24
MAUBFC-rPPG 2x	84	42	0.94	1.10	**3.90**	**5.22**
MAUBFC-rPPG 3x	126	42	**0.92**	**1.09**	3.97	5.31
MAUBFC-rPPG 4x	168	42	1.02	1.25	4.10	5.40
OURS VS. BASELINE			**+2.63%**	**+2.31%**	**+0.76%**	**+0.38%**

D. Motion Augmented rPPG Videos

We provide code for augmenting various camera physiology datasets and pre-trained models trained on motion-augmented data. Additionally, we provide various files to easily train on baselines and motion augmented data using the publicly available rPPG-Toolbox [8]. Pre-trained models using the baseline UBFC-rPPG or PURE datasets can be found in the rPPG-Toolbox. We also include motion analysis scripts that utilize OpenFace [1] to analyze both rigid and non-rigid motion in videos and generate plots. All of these additional materials be found through our project page: https://motion-matters.github.io/.

D.1. The Effect of Motion Transfer on PPG

In Figure 3, we provide additional qualitative examples of the effect of motion transfer on the underlying PPG signal in rPPG videos. All examples include a plot of the gold-standard PPG signal, the predicted PPG signal from the unaugmented source rPPG video by a TS-CAN model, and the predicted PPG signal from the motion-augmented source rPPG video by a TS-CAN model. The TS-CAN models utilized are trained on a larger superset of the shown examples (e.g., UBFC, MAUBFC) with the same experimental settings mentioned in Section B.1. As shown by

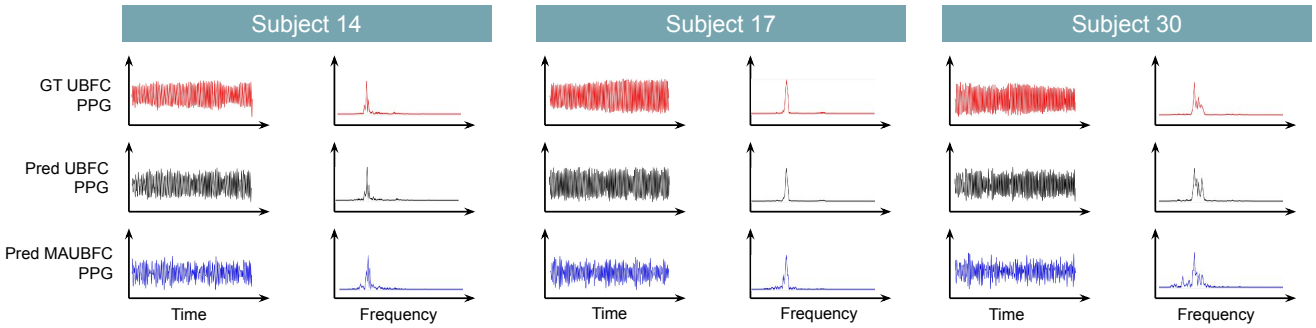


Figure 3. **Signal Prediction on UBFC-rPPG and MAUBFC-rPPG.** We provide three, subject-wise examples of signal prediction using a neural method, TS-CAN, on unaugmented and motion-augmented videos from the UBFC-rPPG dataset.

these qualitative examples, a neural method such as TS-CAN is capable of recovering the underlying PPG signal despite the application of motion augmentation.

Our observation that the underlying physiological signal is reasonably preserved after motion augmentation and, therefore, useful as training data may seem contradictory to prior works (e.g., [3]) that analyze the loss of physiological information as a means to identify deep fake videos. The methods utilized in such prior works typically animate a single frame or swap faces, which we would respectively expect to not have and not preserve a useful rPPG signal. When a method that does animate a whole video without destroying the video subject identity is mentioned, the source videos in question can be highly compressed videos rather than raw, uncompressed videos from rPPG datasets that we used and are expected to have a useful rPPG signal to begin with. It is well-known that various degrees of compression in videos can severely degrade the underlying rPPG signal [12].

E. Datasets

E.1. Source Videos for Motion Synthesis

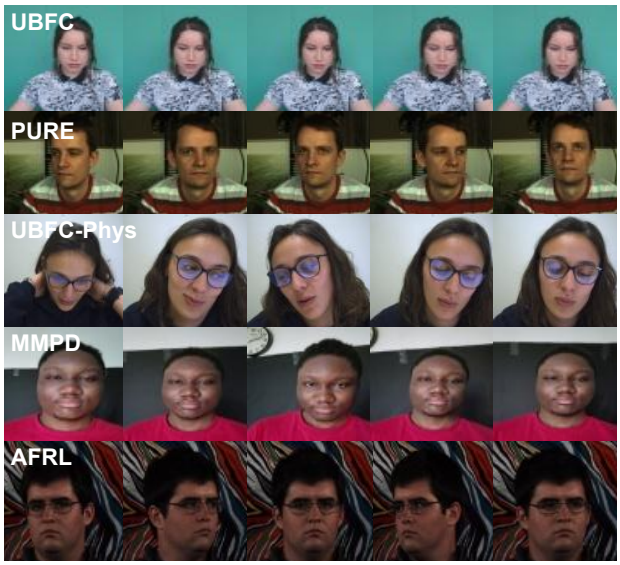
We use the following state-of-the-art datasets for source videos used in our motion synthesis pipeline:

UBFC-rPPG [2]: The UBFC-rPPG video dataset contains 42 RGB videos, one per subject, at 30 Hz. The videos were collected with a Logitech C920 HD Pro with a resolution of 640x480 and a CMS50E transmissive pulse oximeter was utilized in order to record gold-standard PPG signals. The UBFC-rPPG dataset contains minimal motion, with subjects being asked to simply sit one meter away from the camera in an environment with both artificial and natural lighting. When utilized as source videos, we utilized all videos from the UBFC-rPPG dataset. When utilized for evaluation, we also utilized all videos from the UBFC-rPPG dataset.

PURE [15]: The PURE dataset contains 59 videos, each corresponding to a unique task, per a subject. The six tasks involve staying steady, talking, slow head translation, fast head translation, small head rotation, and medium head rotation. There are 10 subjects total with subject 6's talking task video being excluded. All of the videos were captured with an RGB eco274CVGE camera (SVS-Vistek GmbH) at a resolution of 640x480 and 60 Hz. During all tasks, the subject was asked to be seated in front of the camera at an average distance of 1.1 meters and lit from the front with ambient natural light through a window. A gold-standard measure of PPG was collected with a pulse oximeter, CMS50E, attached to the finger. When utilized as source videos, we utilized all videos from the PURE dataset. When utilized for evaluation, we also utilized all videos from the PURE dataset.

SCAMPS [11]: The SCAMPS dataset contains 2,800 synthetic videos that were generated using a blendshape-based rig with 7,667 vertices and 7,414 polygons, with

Table 8. 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

distinct identities being learned from a set of high-quality facial scans. Blood flow, and subsequently the underlying physiological signals, are simulated using the modification of physically-based shading materials. The SCAMPS dataset contains a variety of rigid and non-rigid head motions, with varying intensities. The dataset also contains a variety of lighting conditions and background conditions. Each SCAMPS video is 20 seconds in length, with 600 frames at a sampling rate of 30 Hz. We only utilize portions of the SCAMPS dataset as source videos in our ablation study regarding synthetic versus naturalistic head motion.

E.2. Driving Videos for Motion Synthesis

We use the following datasets for driving videos used in our motion synthesis pipeline:

TalkingHead-1KH [17]: The TalkingHead-1KH dataset is a publicly available, large-scale talking-head video dataset used as a benchmark for Face-Vid2Vid [17] 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. The videos are of varied resolutions, but there is an emphasis on collecting high quality, high resolution videos which compose a significant portion of the dataset (with a resolution of at least 512x512). We elect to filter the dataset for head pose such that any videos where the head pose, on average, is outside +/- 20 degrees are removed. This prevents damaging motion augmentation artifacts due to impractical differences in the head pose in the source video and the head pose in the driving videos, but comes at the cost of reduced head pose variations. We also filter by facial action units (AUs) (0 to 5, in units of intensity) such that any videos below a mean standard deviation in facial AUs of 0.15 is filtered out. This prevents driving videos that are not suitable for our application from being used - for example, a driving video that is effectively a slide show and doesn't have any naturalistic motion upon qualitative inspection.

CDVS: 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. Subjects self-capture the videos in a variety of settings with artificial lighting of the face in an indoors setting. When capturing a video to show one of the two types of motion we study, subjects are asked to constrain the other motion type as much as possible. The CDVS will be released in the future for research purposes.

E.3. Additional Datasets for Evaluation

In addition to using UBFC-rPPG [2] and PURE [15] as both source video datasets in the motion synthesis pipeline and datasets for evaluation, we use three additional state-of-the-art datasets for evaluation:

UBFC-PHYS [13]: The UBFC-PHYS dataset is a multimodal dataset with 168 RGB videos, with 56 subjects (46 women and 10 men) per a task. There are three tasks with significant amounts of both rigid and non-rigid motion - a rest task, a speech task, and an arithmetic task. Gold-standard BVP and electrodermal activity (EDA) measurements were collected via the Empatica E4 wristband. The videos were recorded at a resolution of 1024x1024 and 35Hz with a EO-23121C RGB digital camera. We utilized all of the tasks and the same subject sub-selection list provided by the authors of the dataset in the second supplementary material of Sabour et al. [13] for evaluation. This means we eliminated 14 subjects (s3, s8, s9, s26, s28, s30, s31, s32, s33, s40, s52, s53, s54, s56) for the rest task, 30 subjects (s1, s4, s6, s8, s9, s11, s12, s13, s14, s19, s21, s22, s25, s26, s27, s28, s31, s32, s33, s35, s38, s39, s41, s42, s45, s47, s48, s52, s53, s55) for the speech task, and 23 subjects (s5, s8, s9, s10, s13, s14, s17, s22, s25, s26, s28, s30, s32, s33, s35, s37, s40, s47, s48, s49, s50, s52, s53) for the arithmetic task.

AFRL [5]: The AFRL dataset contains 300 videos of 25 participants (17 males, 8 females) recorded at 658x492 resolution and 120 FPS. Gold-standard physiological signals were measured using the fingertip reflectance PPG method. Participants were asked to perform a series of tasks, resulting in 12 tasks total. With a black background behind the participant, the tasks entailed sitting still with a chin-rest, sitting still without a chin rest, rotating the head with an angular velocity of 10 degrees/second, 20 degrees/second, and 30 degrees/second, and finally randomly orienting their head once per a second to a predefined location. This resulted in six recordings, which were repeated once with a colorful background, resulting in 12 videos per a participant. As a part of our pre-processing steps for AFRL, we down-sampled the videos to 30 FPS. We utilized all of the videos for evaluation.

MMPD [16]: The Multi-domain Mobile Video Physiology Dataset (MMPD) dataset contains 11 hours of recordings from mobile phones of 33 subjects. Gold-standard PPG signals were simultaneously recorded using an HKG-07C+ oximeter. The dataset was designed to capture variations in skin tone, body motion, and lighting conditions in videos useful for the rPPG task. Videos were collected under three artificial light sources: i) low LED light (100 lumens on the face region), ii) mid-level incandescent light (200 lumens on the face region), and iii) high LED light (300 lumens on the face region). Videos were also collected under natural light, which varied from 300-800 lumens intensity on the face region. Videos were recorded following an experimental procedure in which participants performed a variety of tasks in different lighting conditions - a stationary task, a head rotation task, a talking task, and a walking task. We

evaluated on videos with artificial lighting, Fitzpatrick scale skin tone type 3, and any of the four tasks (stationary, head rotation, talking, and walking) that correspond to varying degrees of rigid and non-rigid motion.

F. Broader Impact Statement

While generating synthetic videos that are indistinguishable from those of real people has concerning use cases, there are positive applications of this technology can enabled. In the medical domain simulators are increasingly being tested within specific applications [6, 7]. It is important that the limitations of generative models are understood as these may impact the performance of the resulting models trained using simulated data. It is possible for generative approaches to compound harmful biases [10] and motion augmentation algorithms can be used for troubling negative applications. To mitigate negative outcomes, we license our source code using responsible behavioral use licenses used across a large number of publicly released machine learned models [4].

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. 4

[2] 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. 5, 6

[3] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

[4] 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. 7

[5] 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. 2, 6

[6] Waldemar Hahn, Katharina Schütte, Kristian Schultz, Olaf Wolkenhauer, Martin Sedlmayr, Ulrich Schuler, Martin Eichler, Saptarshi Bej, and Markus Wolfien. Contribution of synthetic data generation towards an improved patient stratification in palliative care. *Journal of Personalized Medicine*, 12(8):1278, 2022. 7

[7] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 2022. 7

[8] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Yuntao Wang, Soumyadip Sengupta, Shwetak Patel, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. *arXiv preprint arXiv:2210.00716*, 2022. 1, 4

[9] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 1

[10] Vongani H. Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A. Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race, 2022. 7

[11] 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. 5

[12] Daniel J. McDuff, Ethan B. Blackford, and Justin R. Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, pages 63–70, 2017. 5

[13] 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. 4, 6

[14] 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. 1

[15] 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. 1, 4, 5, 6

[16] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: Multi-domain mobile video physiology dataset, 2023. 6

[17] 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. 6

[18] 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. 2