egoPPG: Heart Rate Estimation from Eye-Tracking Cameras in Egocentric Systems to Benefit Downstream Vision Tasks

Björn Braun, Rayan Armani, Manuel Meier, Max Moebus, and Christian Holz
Department of Computer Science, ETH Zürich, Switzerland
https://siplab.org/projects/egoPPG

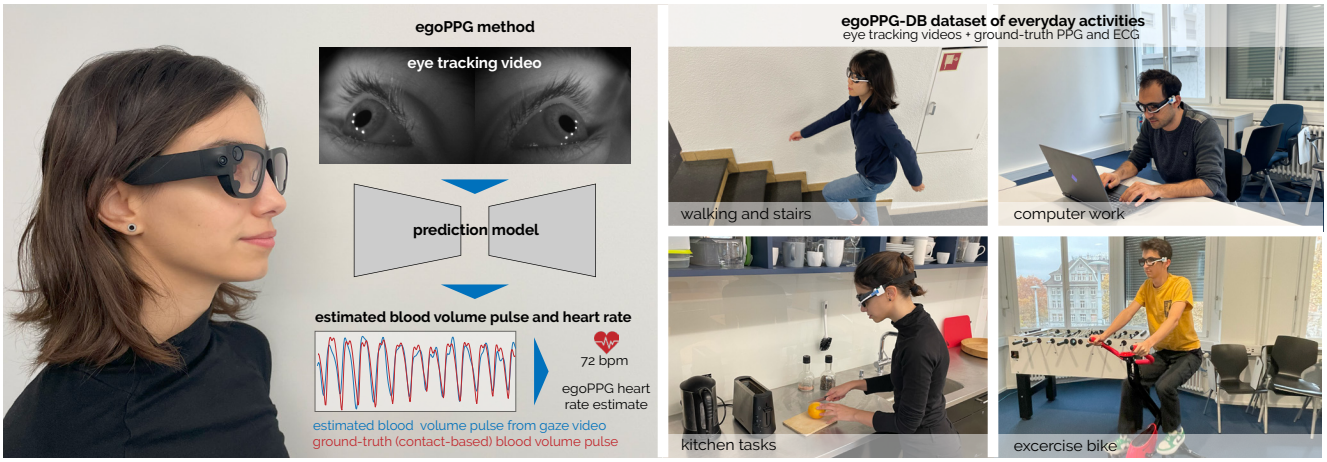

Figure 1. We propose *egoPPG* as a novel computer vision task: tracking a person's heart rate (HR) on unmodified egocentric vision headsets. Taking eye tracking videos as input, our method *PulseFormer* estimates the photoplethysmogram (PPG) from areas around the eyes to derive HR values. For training and validation, we introduce *egoPPG-DB*, a dataset of eye tracking videos while participants performed everyday activities with synchronized ground-truth PPG (via nose-based contact sensor) and HR values (via ECG chest strap).

Abstract

Egocentric vision systems aim to understand the spatial surroundings and the wearer's behavior inside it, including motions, activities, and interactions. We argue that egocentric systems must additionally detect physiological states to capture a person's attention and situational responses, which are critical for context-aware behavior modeling. In this paper, we propose egoPPG, a novel vision task for egocentric systems to recover a person's cardiac activity to aid downstream vision tasks. We introduce PulseFormer, a method to extract heart rate as a key indicator of physiological state from the eye tracking cameras on unmodified egocentric vision systems. PulseFormer continuously estimates the photoplethysmogram (PPG) from areas around the eyes and fuses motion cues from the headset's inertial measurement unit to track HR values. We demonstrate egoPPG's downstream benefit for a key task on EgoExo4D, an existing egocentric dataset for which we find PulseFormer's estimates of HR to improve proficiency estimation by 14%. To train and validate PulseFormer, we collected a dataset of 13+ hours of eye tracking videos from Project Aria and contact-based PPG signals as well as an electrocardiogram (ECG) for ground-truth HR values. Similar to EgoExo4D, 25 participants performed diverse everyday activities such as office work, cooking, dancing, and exercising, which induced significant natural motion and HR variation (44–164 bpm). Our model robustly estimates HR (MAE=7.67 bpm) and captures patterns (r=0.85). Our results show how egocentric systems may unify environmental and physiological tracking to better understand users and that egoPPG as a complementary task provides meaningful augmentations for existing datasets and tasks. We release our code, dataset, and HR augmentations for EgoExo4D to inspire research on physiology-aware egocentric tasks.

1. Introduction

Egocentric vision systems, such as Mixed Reality (MR) glasses by Meta [56], Magic Leap [43], and others have emerged as powerful devices for capturing and analyzing a person's behavior and their environment from a first-person

perspective. The wider availability of promising wearable capture platforms has sparked a large amount of research on egocentric vision tasks for environment understanding and navigation [18], including localization [39, 71, 74], and simultaneous localization and mapping [15, 33, 68]. Since egocentric systems simultaneously capture parts of the wearer's behavior in addition to their environment, prior work has investigated egocentric action recognition [51, 87, 89, 96] and hand-object interaction [26, 73, 97] to understand user behavior. Several large-scale datasets now accelerate data-driven research in this domain with multimodal data for training and evaluation (e.g., Ego4D [26], Nymeria [50], EgoExo4D [27]).

Most recently, Meta's Project Aria 2 introduced a contact-based heart rate (HR) sensor, with which egocentric systems can gauge the wearer's cognitive performance, attention, and situational responses [13, 20, 52, 76, 83]. Numerous additional conditions manifest in a person's HR, such as emotions, stress and fatigue [1, 9, 59, 65, 80]—capturing these dynamics can thus benefit models of human behavior to enable a richer understanding of user behavior.

In this paper, we introduce a method to make such HR estimates available to many *existing* egocentric systems and *already recorded* large datasets, such as EgoExo4D [27] or Nymeria [50]. Our method *PulseFormer* accurately recovers a person's HR from the eye tracking videos in egocentric headsets. *PulseFormer* first estimates the person's photoplethysmogram (PPG) from the subtle fluctuations in skin intensity due to pulsatile artery expansion beneath the surface following a blood volume pulse (BVP), in particular deriving it from regions around the wearer's eye for robust tracking. Our spatial attention module ensures that PPG is estimated from robust regions around the eye, while our cross-attention fusion with the system's inertial measurement unit (IMU) learns a motion-informed temporal attention to optimally weight the eye tracking images for more accurate PPG estimates in scenarios with heavy motion.

We validate our method's efficacy on a novel dataset that we collected to capture some of the activities included in large-scale egocentric datasets alongside physiological reference recordings. Our dataset *egoPPG-DB* contains 13 hours of recordings from 25 participants, who wore Project Aria glasses and performed six real-world tasks with varying motion and intensity, causing their HR values to reach levels between 44–164 bpm.

Downstream benefits for egocentric vision tasks

A key contribution of our paper is that we demonstrate that knowing a person's continuous HR values benefits egocentric vision tasks downstream. We augment an existing architecture with *PulseFormer*'s HR estimates and show its impact on EgoExo4D's proficiency estimation benchmark, which improves accuracy on this task by 14.1%.

Contributions

We summarize our key contributions as follows:

1. *egoPPG* as a novel task and *PulseFormer* as an HR estimation method for egocentric systems that operates on eye tracking videos. Our method robustly predicts continuous HR across a series of activities and interactions (MAE=7.67 bpm), with a 23.8% lower error than current state-of-the-art rPPG models [10, 45, 90, 92, 93].
2. *egoPPG-DB*, a dataset of eye tracking videos and synchronized BVP (contact-based) and ECG recordings (chest strap-based) to verify all physiological signals. We captured these across diverse everyday activities that were inspired by those included in existing large-scale egocentric datasets, such as EgoExo4D [26, 27, 50].
3. a validation of *egoPPG's* downstream benefits for egocentric vision tasks. We demonstrate the implications of our method *PulseFormer* on the proficiency estimation benchmark of the EgoExo4D dataset, which increases the accuracy by 14.1% when augmenting EgoExo4D with our continuously predicted HR values.

2. Related work

Egocentric vision. In recent years, research in egocentric vision has surged, driven by advances in AR/VR glasses [4, 18, 31, 43, 56, 57], which provide new ways for understanding user interaction from a first-person perspective. Much of this work has focused on tasks such as action recognition [41, 51, 87, 89, 96] and anticipation [14, 25, 87], full-body pose estimation [36, 37, 75], responding to user needs [67, 69, 91], and social behavior analysis [21, 26, 35]. Additionally, tracking vital signs in AR/VR settings and for affective computing applications [1, 9, 59, 65, 80] has become an important tool for understanding users' physiological states [53], their behavior, attention, and intent [3, 58, 88].

Physiological measurements. Wearable sensors have had a tremendous impact on health monitoring in recent years, enabling continuous measurement of key physiological metrics, such as heart rate (HR), oxygen saturation, and activity levels [11, 17, 55, 61, 62]. HR, in particular, is a key measure for assessing an individual's health and performance [19, 23, 40, 48, 72]. In parallel to wearable sensors such as smartwatches, recent research has extensively explored using cameras as an unobtrusive, non-contact alternative for measuring HR, generally called remote photoplethysmography (rPPG) [34, 66, 84]. rPPG measures HR based on the BVP via subtle color changes of the skin. Generally, rPPG methods can be broadly divided into traditional signal processing techniques [7, 16, 34, 66, 85] and deep learning-based approaches [8, 10, 45, 78, 90, 92]. So far, rPPG has been mostly applied to facial videos with the camera and user being stationary, such as while sitting in front of a laptop, as it requires a continuous video feed

of the same skin region. This limitation is shown in current rPPG datasets, which primarily capture individuals in seated positions with either a stationary camera directed at their face [6, 29, 70, 79] or requiring users to hold a smartphone steadily in front of their face [82]. As a result, rPPG is not feasible to be deployed in more dynamic settings.

Eye tracking cameras. Eye tracking in egocentric vision systems is mostly done using inward-facing cameras directed at the eyes [2]. Even during motion, eye tracking in VR devices demonstrated accurate performance showcasing that the cameras remain almost stationary *relative to* the user's eyes [12]. Furthermore, IR illumination makes them robust to lighting variations and low-light conditions [49]. To the best of our knowledge, videos from eye tracking cameras have not yet been explored for HR estimation.

3. Overview

Our aim is to enable egocentric vision systems i) to model a person's physiological state via continuously estimated HR and ii) to integrate these HR estimates into downstream tasks that benefit from knowledge of the user's state. Sec. 4 first describes our dataset of synchronized eye tracking videos and ground-truth HR measurements. Sec. 5 introduces our method *PulseFormer* for recovering continuous HR from eye tracking videos. Sec. 6 outlines the downstream benefits of our novel task, using HR as input for modeling user proficiency on the EgoExo4D dataset.

4. egoPPG-DB

The *egoPPG-DB* dataset was developed to support HR estimation from eye tracking videos under real-world conditions and contains significant motion and HR fluctuations. By including diverse everyday activities, we provide a challenging benchmark for egocentric HR estimation models.

4.1. Recruiting and recording

We recruited $N = 25$ participants (12 female, 13 male, ages 19–32, $\mu = 25.1$ and $\sigma = 3.3$) on a voluntary basis, resulting in over 13 hours of video recordings. Based on the Fitzpatrick scale [22], 9 participants had skin type II, 8 had skin type III, 3 had skin type IV, and 5 had skin type V. All participants signed a consent form before the data collection, agreeing with using and sharing their data for academic and non-commercial purposes. The data collection was approved by the ETH Zurich Ethics Commission (no. 2023-N-08). In terms of duration, *egoPPG-DB* is among the longest rPPG datasets as listed in Tab. 10 (Supplementary). Participants of *egoPPG-DB* are not included in EgoExo4D.

4.2. Apparatus

Fig. 2 illustrates our experimental setup. We used Project Aria glasses [18] with Profile 21 to record eye tracking videos at 30 fps with a resolution of 320×240 pixels per eye. To capture ground truth PPG measurements, with which we train our model, we developed a custom sensor that records PPG data offline at 128 Hz. The sensor consists of a main board, mounted on the left side of the frame, featuring a DA14695 system-on-chip interfacing with a MAX86141ENP+ PPG sensor. The LEDs and photodiodes used by the PPG sensor are embedded in the left nose pad and connected to the main board using a flat flexible cable. For each participant, we individually adjusted the nose pad position to ensure the sensor aligned with their left angular artery [30]. To validate our custom PPG sensor, we also recorded gold-standard ECG data using a movisens ECGMove 4 chest belt sampling at 1024 Hz. We synchronized all devices at the start and end of each recording with a synchronization pattern, using their built-in IMUs.

Figure 2. Apparatus used to record the *egoPPG-DB* dataset.

4.3. Capture protocol

The average recording lasted 32 minutes. The capture protocol comprised 5 activities (Tab. 1): watching a video, office work, kitchen work, dancing, and exercising on an indoor bike (Fig. 1). We included these activities for three purposes: (1) Incorporate everyday activities including the corresponding HR changes and motion artifacts; (2) cover a wide range of HR values (low HR when watching a video vs. high HR when exercising), and (3) resemble activities that were captured in large-scale egocentric vision datasets, such as EgoExo4D [27]. In Tab. 11 (Supplementary), we give a detailed description of all activities. Tab. 3 shows mean HR values for each activity. Exercising on the bike produced the highest mean HR values (113 bpm), whereas watching the video resulted in the lowest (71 bpm).

4.4. Dataset and signal quality verification

To ensure the contact PPG sensor, whose signal we later use as the target for model training, produces accurate HR

Activity	Actions	Minutes
Watch video	Watch a documentary	5
Office work	Work on a computer	4
	Write on a paper	2
	Talk to the experimenter	2
Walking	Walk to the kitchen	1
Kitchen work	Cut vegetables Prepare a sandwich Wash the dishes	5
Walking	Walk to the dancing room	1.5
Dancing	Follow random dance video	5
Exercise bike	Ride an exercise bike	5
Walking	Walk back to the start	1.5

Table 1. Capture protocol for recording the *egoPPG-DB* dataset.

values, we evaluate it against the gold-standard ECG. We calculated the MAE and Pearson correlation between HR estimates from the ECG and PPG signals for each participant using a 30-second sliding window. For activity labeling, we manually annotated the start and end times of each task (see Tab. 1) for each participant using the Point of View (POV) RGB videos recorded by the Project Aria glasses. To ensure that the signal quality of the contact PPG is sufficient for model training, we excluded all tasks with an MAE over 3.0 bpm between the PPG and ECG (e.g. when the PPG sensor moved). This applied to 20 of 150 tasks (13%, see Tab. 6 in Supplementary). During the remaining tasks, our custom-built PPG nose sensor achieved very high accuracy, with an MAE of 1.3 bpm and a correlation of 0.94 compared to the ECG signal, showing its suitability as ground truth.

5. PulseFormer method

5.1. Problem definition

Our objective is to estimate BVP and HR from periodic changes in pixel intensity in eye tracking video frames $\boldsymbol{F} \in \mathbb{R}^{w \times h}$. Physically, this means extracting physiological signals from the information in the light reflected by the arteries and arterioles that carry blood beneath the skin. This light reflection can be modeled as a combination of diffuse and specular reflections. Wang *et al*. [85] model the reflected light intensity $C(t)$ as:

$$\boldsymbol{C(t)} = I(t)(\boldsymbol{v_s}(t) + \boldsymbol{v_d}(t)) + \boldsymbol{v_n}(t) \tag{1}$$

where $I(t)$ is the luminance intensity, $\boldsymbol{v_s}(t)$ the specular reflection, $\boldsymbol{v_d}(t)$ the diffuse reflection, and $\boldsymbol{v_n(t)}$ the sensor noise. While the specular reflection $\boldsymbol{v_s}(t)$ lacks pulsatile information, the diffuse reflection $\boldsymbol{v_d}(t)$ contains information about the absorption and scattering of the light in skin tissue [85]. Thus, $\boldsymbol{v_d}(t)$ can be further decomposed as:

$$\boldsymbol{v_d}(t) = \boldsymbol{u_d} d_0 + \boldsymbol{u_p} p(t) \tag{2}$$

where $\boldsymbol{u_d}$ is the unit color vector of the skin, d_0 the stationary reflection strength, $\boldsymbol{u_p}$ the relative absorption, and $p(t)$ the signals of interest. $p(t)$ is in our case the BVP, which our model aims to learn from the camera recordings.

5.2. Deep learning model

Our architecture is built upon a 3D CNN backbone (PhysNet) [92] with a temporal input length of $T = 128$ frames (corresponding to 4.3 seconds) downsampled to $(h = 48) \times (w = 128)$ pixels, resulting in an input of dimensions (T, C, w, h). The channel is $C = 1$ in our case, as our input is from monochrome videos. The input in our network is the consecutive standardized frame differences (per participant frame-wise differences divided by the STD of the frames) of the eye tracking videos to help the network focus on the changes between frames [10]. As labels, we use the standardized consecutive differences of the PPG signals. Since eye tracking videos offer additional challenges compared to facial videos, usually used for rPPG tasks, we have designed our model to address these challenges (see Fig. 3).

Motion-informed temporal attention (MITA). Egocentric glasses are body-worn and subject to considerable motion artifacts when the user moves. Therefore, we propose to leverage the IMU within the glasses to obtain a motion-informed temporal attention. We employ a cross-attention module to integrate the IMU data with the video input, allowing our model to weigh each frame differently along the temporal dimension based on the motion intensity encoded by the IMU. This allows the model, *e.g*., to give less emphasis to frames heavily affected by motion. Given the input feature map $\boldsymbol{F_{in}} \in \mathbb{R}^{T \times 1 \times w \times h}$, we use ResNet18 [28] and a linear layer to obtain the image embeddings $\boldsymbol{F_e} \in \mathbb{R}^{T \times D}$, where $D = 128$ is the embedding dimension. Given IMU measurements $\mathbf{I_{in}} \in \mathbb{R}^{T \times 1}$, we use two 1D convolutional layers to obtain the IMU embeddings $\mathbf{I_e} \in \mathbb{R}^{T \times D}$. We then calculate the cross-attention $\mathbf{A} \in \mathbb{R}^{T \times D}$ as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}}\right)\mathbf{V} \tag{3}$$

where $\mathbf{I_e}$ serve as queries Q and $\mathbf{F_e}$ as keys K and values V. Using a linear layer, we obtain the motion-informed temporal attention $\mathbf{T} \in \mathbb{R}^{T \times 1 \times 1 \times 1}$, which we multiply with $\boldsymbol{F_{in}}$.

Spatial attention (SA). While the bulbar conjunctiva (white of the eyes) contains many blood vessels from which the BVP could theoretically be estimated, eyes typically move strongly during everyday situations and are closed while blinking (see participant 2 in Fig. 4). Consequently, extracting the BVP from the eye regions would introduce substantial motion artifacts and reduce the signal-to-noise

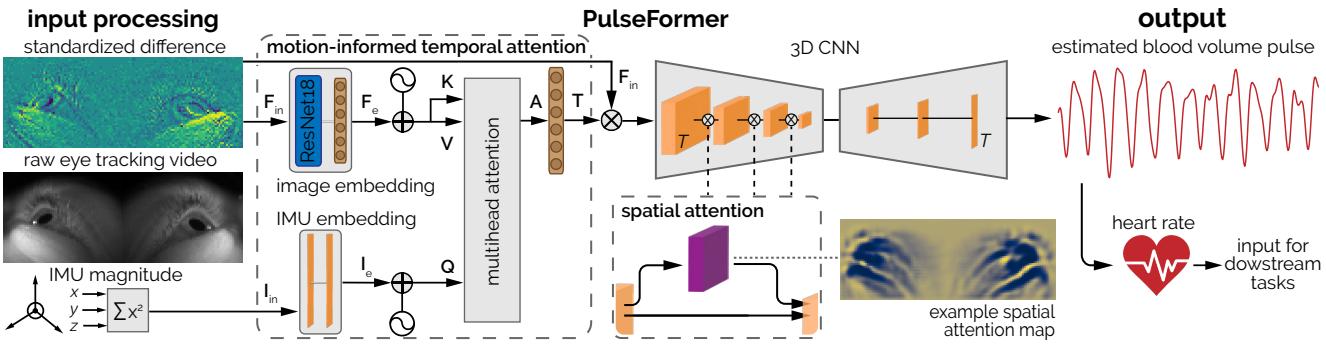

Figure 3. Architecture of our model for continuous BVP estimation from eye tracking videos and consecutive HR computation.

ratio (SNR). In contrast, when qualitatively analyzing eye tracking images, we see that the skin around the eyes exhibits considerably less motion than the eyes themselves and could thus provide a more stable source of BVP information. To address this, we introduce spatial attention modules [32, 64, 86] before each pooling (see Fig. 3) to allow our network to focus on high-SNR regions, such as the skin, and reduce the influence of low-SNR regions with frequent motion, like the eyes. Given some feature map $\boldsymbol{F} \in \mathbb{R}^{T\times C\times w\times h}$, the spatial attention modules infer a spatial attention map $\boldsymbol{M_s} \in \mathbb{R}^{T\times 1\times w\times h}$ as:

$$\boldsymbol{M_s}(\boldsymbol{F}) = \sigma * (f^{7\times 7}([\boldsymbol{F_{avg}}; \boldsymbol{F_{max}}])) \quad (4)$$

where σ is the sigmoid function, $f^{7\times 7}$ a 7×7 convolution operation and $\boldsymbol{F_{avg}} \in \mathbb{R}^{T\times 1\times w\times h}$ and $\boldsymbol{F_{max}} \in \mathbb{R}^{T\times 1\times w\times h}$ are the average-pooled and max-pooled feature maps respectively. The final output $\boldsymbol{F_{out}} \in \mathbb{R}^{T\times C\times w\times h}$ of each attention process is then the product of $\boldsymbol{M_s}$ and $\boldsymbol{F}$.

Data augmentation. Furthermore, individual variations in the fit of the glasses result in different parts of the skin around the eyes being visible. For some individuals, the eye tracking cameras capture only the areas above the eyes, for others, only below, and in some cases, the glasses sit at an incline (see Fig. 4). To account for such variations, we apply three targeted data augmentations during training that reflect these specific differences in camera angles and coverage: (1) random rotation between -20 and +20 degrees to account for slight inclinations; (2) random horizontal cropping to help the network distinguish between high and low SNR regions across various skin areas and camera positions; and (3) horizontal and vertical flipping to further increase robustness to differences in skin region visibility. Our model requires approximately 399 GFLOPS per batch and has about 12M parameters. The frame rate is 2.9k fps on an RTX 4090 and 180 fps on an AMD EPYC CPU.

5.3. Experiments setup

Training. We trained all models using five-fold cross-validation split by participants to ensure a strict separation between training, validation, and test sets. We iteratively held out the data from five participants (20%) as the test set, two as validation, and trained on the remaining with a batch size of 4 for 100 epochs, a learning rate of 0.0009, and mean squared error (MSE) as loss. In addition to our model, we used ten state-of-the-art rPPG baseline networks to compare the performance of our proposed model to these established models. Our model was trained on a GeForce RTX 4090, with a total runtime of about 20 hours for all folds.

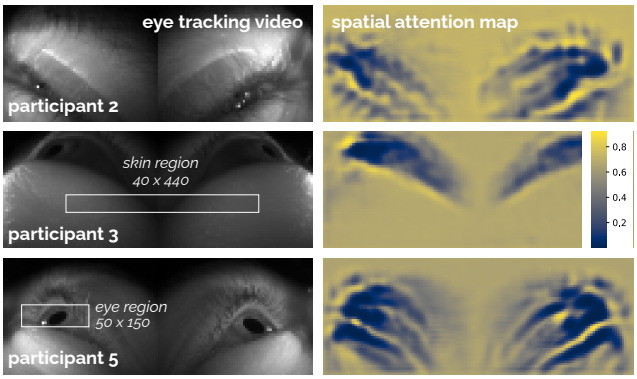

Figure 4. Left: Head geometry determines the regions that the eye tracker captures. Right: Learned spatial attention maps show that eye regions are excluded and *PulseFormer* instead extracts BVP from the surrounding skin regions, which moves less than the eyes.

Evaluation. To calculate the HR, we filter the predicted BVP with a Butterworth filter (0.5–2.8 Hz) and then detect peaks. To assess model accuracy, we use the mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and Pearson correlation (r) using a non-overlapping 60-second window [16, 45, 46].

Video sampling rate. While we recorded the eye tracking videos with 30 fps, large-scale datasets such as EgoExo4D [27] or Nymeria [50] used only 10 fps. To assess the impact of reduced fps, we evaluated model performance when (1) downsampling our videos to 10 fps by retaining only every third frame and (2) downsampling to 10 fps, then

linearly interpolating between frames to upsample to 30 fps.

6. Downstream use for proficiency estimation

To demonstrate the utility of predicting a user's physiological state for egocentric vision applications, we use the user proficiency estimation benchmark from the EgoExo4D dataset, which contains over 5000 videos from 740 participants performing skilled human activities [27]. This benchmark aims to classify the proficiency of a user (novice, early expert, intermediate expert, late expert) using only egocentric videos (*Ego*), only exocentric videos (*Exo*), or all videos together (*Ego + Exo*). Our goal was to assess if we can improve the performance of the current baseline model (TimeSFormer [5]) when integrating our predicted HR data into the network. This results in three additional configurations: using egocentric/exocentric videos and HR (*Ego + HR*/*Exo + HR*) and using all videos and HR (*Ego + Exo + HR*). To predict the continuous HRs for all EgoExo4D videos, we use *PulseFormer*, pre-trained on *egoPPG-DB*.

We implement the TimeSFormer model in exactly the same configuration as for the benchmark results [27] with a clip size of 16 frames and a sampling rate of 16, trained for 15 epochs. We use all videos of the EgoExo4D dataset, for which the proficiency estimation labels are available (using the official benchmark training/validation sets) and which have at least 16 frames at a sampling rate of 16, resulting in 2044 videos. From the official training set, we use 10% as validation, and the held-out official validation set for testing. We summarize our predicted HR data by calculating five features (mean, STD, minimum and maximum HR, and mean HR change) for the corresponding videos. We integrate these features via normalization and a 50-parameter linear layer whose output we concatenate with the output of TimeSFormer's backbone before feeding it into the classification head. We train all models from random initialization and evaluate using top-1 accuracy per EgoExo4D protocol.

7. Experiments

7.1. Heart rate estimation

7.1.1. Signal-processing baseline

We employed signal processing to verify that the BVP signal is present in the eye tracking videos, to determine in which regions the SNR is highest, and to establish a baseline (see Tab. 2). Since the glasses remain mostly stable throughout the recording, we manually define two spatial cropping regions per participant. One region that includes mostly skin, and one region that includes mainly eyes (see Fig. 4). We calculate the mean pixel intensity for both regions, remove motion artifacts by discarding any changes outside the interquartile range and finally filter the signal with a 4^{th} order Butterworth bandpass filter (0.5 to 2.8 Hz) to obtain the BVP (see Fig. 5 in Supplementary).

7.1.2. PulseFormer method

Using our proposed network *PulseFormer*, we obtain an MAE of 7.67 bpm and a correlation of 0.85 between our predicted HR and the ground truth HR (see Tab. 2). This is an improvement of 2.40 bpm (23.8%) of the MAE and 0.13 for the correlation compared to the current SOTA FactorizePhys [38]. Split by activity, we obtain the lowest MAE while the participants are watching a video (MAE=5.52 bpm) and the highest MAE during exercising on a bike (MAE=12.91 bpm), which is the task with the highest mean HR (113.1 bpm) and the second highest motion magnitude. In addition, the MAE decreases for all activities when adding MITA, with the greatest performance improvement for dancing. We define the motion magnitude as the root-mean-squared sum of the absolute differences across the 3-axis IMU recorded by the Aria glasses and normalize it between zero and one across all activities to get a measure of motion of each activity. See Fig. 6 (Supplementary) for a boxplot of the MAEs of *PulseFormer*'s predictions. Using signal processing, we obtain an MAE of 12.40 bpm when using the skin region around the eyes and an MAE of 14.60 bpm using the eye regions as input. This is also reflected in the spatial attention maps that our model implicitly learns (see Fig. 4), which exclude the eyes to predict the HR. To qualitatively cross-check these results, Fig. 5 (Supplementary) shows an example plot of the raw mean intensity values (before filtering) of the skin region compared to the eye region, with the BVP clearly visible for the skin region. Tab. 4 shows the results when downsampling our videos to 10 fps. MAE increases to 11.13 bpm and the correlation decreases to 0.7 when training and testing using 10 fps. When upsampling the videos again to 30 fps, the MAE decreases to 10.20 bpm and the correlation increases to 0.77. In Sec. 16 (Supplementary), we show that *PulseFormer* also outperforms the baselines in a cross-dataset evaluation.

7.2. Downstream task: proficiency estimation

Tab. 5 summarizes the results of our experiments to evaluate the value of HR estimation for the proficiency estimation benchmark on EgoExo4D. We see that integrating our predicted HRs into the TimeSFormer model [5] improved accuracy for all scenarios but one (Soccer). We also achieved the highest accuracy for each of these individual scenarios with our HR integration. When combining the egocentric videos with our predicted HRs, we achieved an overall accuracy of 45.29%, a 14.1% increase compared to using egocentric videos alone. The largest gains appeared in the cooking and dancing tasks, where accuracy rose from 20.00% to 40.00% and from 43.44% to 53.27%. Also, when using the egocentric and exocentric videos, and our predicted HRs together, the accuracy increased by 12.67% (4.94 percentage points) from 39.00% to 43.94% compared

Model	MAE	RMSE	MAPE	r
Yue et al. [94]	29.63	32.99	37.86	0.1
DeepPhys [10]	28.26	31.97	36.68	0.08
TS-CAN [45]	26.32	32.39	29.13	0.11
ContrastPhys+ [81]	19.12	24.13	22.57	0.21
RhythmMamba [99]	15.05	19.78	17.46	-0.16
Baseline eyes	14.60	18.18	18.37	0.20
PhysMamba [90]	13.94	16.86	17.76	0.61
RhythmFormer [98]	13.13	17.43	14.73	0.51
Baseline skin	12.40	15.54	15.29	0.50
PhysNet [92]	12.09	15.43	15.14	0.66
PhysFormer [93]	10.71	13.97	12.69	0.72
PulseFormer w/o SA	10.49	13.62	12.83	0.73
FactorizePhys [38]	10.07	13.43	12.36	0.67
PulseFormer w/o MITA	8.82	12.03	10.82	0.81
***PulseFormer* (ours)**	**7.67**	**10.69**	**9.45**	**0.85**
Improvement over second-best method	**−2.40**	**−2.74**	**−2.91**	**+0.13**

Table 2. Results for HR prediction from eye tracking videos using different models (*PulseFormer*, *PulseFormer* without SA, *PulseFormer* without MITA and established rPPG baselines).

Activity	μ HR	Motion magnitude	*PulseFormer*	*PulseFormer* w/o MITA
Video	71.5	0	**5.52**	**5.97**
Office	75.7	0.45	7.50	8.22
Kitchen	85.3	0.54	7.22	8.89
Dancing	89.1	**1.00**	7.85	10.54
Bike	**113.1**	0.77	12.91	14.62
Walking	93.7	0.30	8.23	8.29

Table 3. Results for HR prediction (MAE) split by activity using *PulseFormer* and *PulseFormer* without MITA.

Input video	MAE	RMSE	MAPE	r
10 fps (other datasets)	11.13	15.18	12.28	0.70
Upsampled to 30 fps	**10.18**	**13.07**	**12.48**	**0.77**

Table 4. Results for HR prediction with different frame rates. In the first row, we downsample our videos to a frame rate of 10 fps, commonly used by large-scale datasets such as EgoExo4D [27]. In the second row, we first downsample our videos to 10 fps and then upsample them to 30 fps by linearly interpolating between frames.

to using only the egocentric and exocentric videos.

8. Discussion

8.1. Heart rate estimation

Evaluating *PulseFormer* on *egoPPG-DB*, we showed that HR can be reliably predicted from eye tracking videos and IMU signals from unmodified egocentric vision headsets. While SOTA rPPG models (e.g., PhysFormer) achieve MAEs as low as 0.50 bpm [93, 98] on datasets such as UBFC-RPPG [6] or OBF [44], these datasets captured participants while *calmly sitting* at a table looking at the camera (with very little motion or HR changes). However, during even just *light* motion (e.g., on MMPD [82] or VIPL-HR [63]), their MAE increases to 5.0–12.0 bpm [93, 98].

On *egoPPG-DB*, which contains much stronger motion (dancing, exercise bike) and HR fluctuations (between 44–164 bpm), *PulseFormer's* MAE is 7.67 bpm and outperforms the rPPG SOTA FactorizePhys (MAE=10.07 bpm). Given the strong motion and diverse everyday activities in *egoPPG-DB*, we believe that our results demonstrate the robustness of *PulseFormer* in dynamic, everyday conditions. For context, even HR estimates from contact sensors tightened to the body (e.g., Apple Watch) yield an MAE of 3.0 bpm during rest and an MAE of 4.6 bpm on a bike [24].

8.1.1. Performance depending on method

We introduced MITA and leverage SA modules to improve the performance of our model. The performance of *PulseFormer* decreases from 7.67 bpm to 8.82 bpm when removing the MITA and to 10.49 bpm when removing the SA modules (see Tab. 2). When qualitatively analyzing the learned SA maps, we see that our model implicitly learned to exclude the eyes for estimating BVP from the eye tracking videos (see Fig. 4). This aligns with our results using signal processing, obtaining better performance for the skin region compared to the eyes (see Tab. 2).

8.1.2. Performance depending on activity

Analyzing our results split by activity (see Tab. 3), we obtain the highest MAE when exercising on a bike (MAE=12.91 bpm) and the lowest MAE when watching a video (MAE=5.52 bpm). While watching a video yields the lowest MAE, it is higher than MAEs typically reported for rPPG datasets, such as UBFC-RPPG [6], despite similar levels of motion. We attribute this to two factors: first, the higher variability in HR across *egoPPG-DB*, requiring improved generalization, and second, the inherent motion artifacts in eye tracking videos from blinking and natural eye movements, even during static tasks like watching videos. Such inherent motion artifacts and, e.g., slipping glasses can make capturing rPPG more difficult in this manner. Furthermore, although dancing has the highest motion magnitude, its MAE (7.85 bpm) is comparable to that of lower-motion tasks such as office and kitchen activities. When comparing performance with and without our MITA module, we

Scenario	Majority	Ego	Ego + HR (ours)	Exo	Exo + HR (ours)	Ego + Exo	Ego + Exo + HR (ours)
Basketball	38.00	45.45	47.47	48.48	48.48	49.49	**50.50**
Cooking	0.00	20.00	**40.00**	35.00	**40.00**	25.00	**40.00**
Dancing	24.59	43.44	53.27	42.62	48.36	50.82	**59.84**
Music	57.89	78.94	**81.58**	57.89	57.89	57.89	60.53
Bouldering	15.29	24.50	**27.81**	8.61	12.58	15.89	21.19
Soccer	62.50	50.00	56.25	**81.25**	75.00	75.00	62.50
Overall	27.80	39.69	**45.29**	34.75	37.67	39.00	43.94

Table 5. Results for proficiency estimation benchmark on EgoExo4D dataset. Note that for all scenarios except Soccer, the accuracy increases when integrating *PulseFormer*'s heart rate estimate into the existing and otherwise unmodified baseline model.

observe an improvement of 2.6 bpm for dancing, indicating that MITA effectively addresses motion-induced artifacts.

8.1.3. Performance depending on camera fps

Using eye tracking videos recorded at only 10 fps considerably decreases performance (see Tab. 4). However, upsampling the frame rate to 30 fps through linear interpolation between frames substantially improves the performance again. This is especially important as many large-scale datasets, such as EgoExo4D [27] or Nymeria [50], for which predicting a user's physiological state could help for further downstream tasks, are recorded at only 10 fps.

8.2. Benefits for proficiency estimation downstream

We found that incorporating HR data into the baseline model of the proficiency estimation task substantially improved accuracy across all three configurations. The egocentric videos combined with the HR achieved the highest overall accuracy at 45.29%, marking a 14.1% increase over using only egocentric videos (39.69%). Adding HR especially improved accuracy for cooking (from 20% to 40%) and dancing (from 43.44% to 53.27%), which had the lowest accuracies besides bouldering when using only egocentric videos, demonstrating the value of HR in enhancing model performance. Combining egocentric videos, exocentric videos, and HR provided further accuracy gains for some scenarios, achieving the best results for basketball, cooking, and dancing. Results using exocentric views alone were lower overall, which is consistent with benchmark results [27]. Soccer was the only scenario, for which the performance decreased for *Exo+HR* and *Ego+Exo+HR*. We see two reasons for that. 1) Of EgoExo4D's 2044 official train/test videos, only 77 are soccer, making it the scenario with the least training/test data by far. 2) Our HR estimates may be less accurate for "Stop-and-Go" sports, which are not captured in *egoPPG-DB* right now. In Tab. 9 (Supplementary), we show that we obtain the best downstream performance when using the HR features calculated with *PulseFormer* compared to the baselines. For training and testing, we used the available subset of EgoExo4D videos for which proficiency labels are available, following the official training and test splits. While our used data shows slight variations from the official release in majority class distributions and accuracy scores, the observed trends align well with the established benchmark results.

8.3. Broader impacts

Beyond health applications, such as predicting stress and fatigue [9, 60, 65], cardiac measurements could also help models better understand user behavior to, *e.g.*, improve personalized assistance [42]. Furthermore, we believe that our approach requires the user's explicit consent, regardless of application. Mechanisms must make users aware of measurements and require consent, *e.g.*, on the Aria platform.

9. Conclusion

egoPPG is a novel task for egocentric vision systems to extract the wearer's heart rate for integrating their physiological state into egocentric vision tasks downstream. We have introduced *PulseFormer*, a method that processes input from the eye tracking cameras on unmodified egocentric vision systems and fuses them with motion cues from the headset's IMU to robustly estimate the person's HR in various everyday scenarios. We validate *PulseFormer's* robustness on our dataset *egoPPG-DB* and demonstrate significant improvements over existing rPPG models. With HR estimations from *PulseFormer*, we also significantly improve the proficiency estimation benchmark on the large-scale EgoExo4D dataset. Our results emphasize the potential of physiological insights obtained via *egoPPG* methods for further egocentric vision applications. By making our dataset available to the community, we aim to support physiological state estimation via HR in future research and new downstream tasks for egocentric vision systems. Given our promising results, we believe that future work could now focus on collecting more participants with a broader demographic background across different age groups, skin types, and ethnicities, and also extend data collection to outdoor settings to assess the impact of varying lighting conditions.

References

[1] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015. 2

[2] Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye tracking in virtual reality: a broad review of applications and challenges. *Virtual Reality*, 27(2):1481–1505, 2023. 3

[3] Henny Admoni and Siddhartha Srinivasa. Predicting user intent through eye gaze for shared autonomy. In *2016 AAAI fall symposium series*, 2016. 2

[4] Apple. Apple vision pro. https://www.apple.com/apple-vision-pro/, 2024. Accessed: 2024.11.13. 2

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 6

[6] 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. 3, 7, 2

[7] Björn Braun, Daniel McDuff, Tadas Baltrusaitis, and Christian Holz. Video-based sympathetic arousal assessment via peripheral blood flow estimation. *Biomedical Optics Express*, 14(12):6607–6628, 2023. 2

[8] Björn Braun, Daniel McDuff, and Christian Holz. How suboptimal is training rppg models with videos and targets from different body sites? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 410–418, 2024. 2

[9] Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. *The Oxford handbook of affective computing*. Oxford University Press, USA, 2015. 2, 8

[10] 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. 2, 4, 7

[11] Hsueh-Wen Chow, Chao-Ching Yang, et al. Accuracy of optical heart rate sensing technology in wearable fitness trackers for young and older adults: validation and comparison study. *JMIR mHealth and uHealth*, 8(4):e14707, 2020. 2

[12] Viviane Clay, Peter König, and Sabine Koenig. Eye tracking in virtual reality. *Journal of eye movement research*, 12(1), 2019. 3

[13] Stephen A Coombes, Torrie Higgins, Kelly M Gamble, James H Cauraugh, and Christopher M Janelle. Attentional control theory: Anxiety, emotion, and motor planning. *Journal of anxiety disorders*, 23(8):1072–1079, 2009. 2

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[15] Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1403–1410. IEEE, 2003. 2

[16] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013. 2, 5

[17] Jessilyn Dunn, Ryan Runge, and Michael Snyder. Wearables and the medical revolution. *Personalized medicine*, 15(5): 429–448, 2018. 2

[18] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2, 3

[19] Harun Evrengul, Halil Tanriverdi, Sedat Kose, Basri Amasyali, Ayhan Kilic, Turgay Celik, and Hasan Turhan. The relationship between heart rate recovery and heart rate variability in coronary artery disease. *Annals of Noninvasive Electrocardiology*, 11(2):154–162, 2006. 2

[20] Michael W Eysenck, Nazanin Derakshan, Rita Santos, and Manuel G Calvo. Anxiety and cognitive performance: attentional control theory. *Emotion*, 7(2):336, 2007. 2

[21] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 2

[22] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988. 3

[23] Kim Fox, Jeffrey S Borer, A John Camm, Nicolas Danchin, Roberto Ferrari, Jose L Lopez Sendon, Philippe Gabriel Steg, Jean-Claude Tardif, Luigi Tavazzi, Michal Tendera, et al. Resting heart rate in cardiovascular disease. *Journal of the American College of Cardiology*, 50(9):823–830, 2007. 2

[24] Stephen Gillinov, Muhammad Etiwy, Robert Wang, Gordon Blackburn, Dermot Phelan, A Marc Gillinov, Penny Houghtaling, Hoda Javadikasgari, and Milind Y Desai. Variable accuracy of wearable heart rate monitors during aerobic exercise. *Medicine & Science in Sports & Exercise*, 2017. 7

[25] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021. 2

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[27] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 3, 5, 6, 7, 8

[28] 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*, pages 770–778, 2016. 4

[29] Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017. 3

[30] Christian Holz and Edward J Wang. Glabella: Continuously sensing blood pressure behavior using an unobtrusive wearable device. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–23, 2017. 3

[31] HTC. Htc vive. https://vive.com/, 2024. Accessed: 2024.11.13. 2

[32] Min Hu, Fei Qian, Xiaohua Wang, Lei He, Dong Guo, and Fuji Ren. Robust heart rate estimation with spatial–temporal attention network from facial videos. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):639–647, 2021. 5

[33] Patrick Hübner, Kate Clintworth, Qingyi Liu, Martin Weinmann, and Sven Wursthorn. Evaluation of hololens tracking and depth sensing for indoor mapping applications. *Sensors*, 20(4):1021, 2020. 2

[34] Markus Huelsbusch and Vladimir Blazek. Contactless mapping of rhythmical phenomena in tissue perfusion using ppgi. In *Medical Imaging 2002: Physiology and Function from Multidimensional Images*, pages 110–117. International Society for Optics and Photonics, 2002. 2

[35] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022. 2

[36] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 2

[37] Jiaxi Jiang, Paul Streli, Manuel Meier, Andreas Fender, and Christian Holz. Egoposer: Robust real-time ego-body pose estimation in large scenes. *arXiv preprint arXiv:2308.06493*, 3(7), 2023. 2

[38] Jitesh Joshi, Sos S Agaian, and Youngjun Cho. Factorizephys: Matrix factorization for multidimensional attention in remote physiological sensing. *arXiv preprint arXiv:2411.01542*, 2024. 6, 7

[39] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 2

[40] Robert E Kleiger, Phyllis K Stein, and J Thomas Bigger Jr. Heart rate variability: measurement and clinical utility. *Annals of Noninvasive Electrocardiology*, 10(1):88–101, 2005. 2

[41] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16020–16030, 2021. 2

[42] Nikola Kovacevic, Christian Holz, Markus Gross, and Rafael Wampfler. On multimodal emotion recognition for human-chatbot interaction in the wild. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 12–21, 2024. 8

[43] Magic Leap. Magic leap 2. https://www.magicleap.com/magic-leap-2, 2024. Accessed: 2024.11.13. 1, 2

[44] 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)*, pages 242–249. IEEE, 2018. 7, 3

[45] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 2, 5, 7

[46] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[47] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, 2019. 2

[48] Tiffany Luong and Christian Holz. Characterizing physiological responses to fear, frustration, and insight in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3917–3927, 2022. 2

[49] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 3

[50] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024. 2, 5, 8

[51] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 2

[52] Alan C MacPherson, Dave Collins, and Sukhvinder S Obhi. The importance of temporal structure and rhythm for the optimum performance of motor skills: A new focus for practitioners of sport psychology. *Journal of Applied Sport Psychology*, 21(S1):S48–S61, 2009. 2

[53] Javier Marín-Morales, Carmen Llinares, Jaime Guixeres, and Mariano Alcañiz. Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*, 20(18):5163, 2020. 2

[54] Daniel McDuff, Miah Wander, Xin Liu, Brian Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. Scamps: Synthetics for camera measurement of physiological signals. *Advances in Neural Information Processing Systems*, 35: 3744–3757, 2022. 3

[55] Manuel Meier, Berken Utku Demirel, and Christian Holz. Wildppg: a real-world ppg dataset of long continuous recordings. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2025. Curran Associates Inc. 2

[56] Meta. Meta quest. https://www.meta.com/quest/, 2024. Accessed: 2024.11.13. 1, 2

[57] Microsoft. Microsoft hololens. https://learn.microsoft.com/en-us/hololens/, 2024. Accessed: 2024.11.13. 2

[58] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021. 2

[59] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE transactions on affective computing*, 12(2):479–493, 2018. 2

[60] Max Moebus, Shkurta Gashi, Marc Hilty, Pietro Oldrati, and Christian Holz. Meaningful digital biomarkers derived from wearable sensors to predict daily fatigue in multiple sclerosis patients and healthy controls. *Iscience*, 27(2), 2024. 8

[61] Max Moebus, Lars Hauptmann, Nicolas Kopp, Berken Demirel, Björn Braun, and Christian Holz. Nightbeat: Heart rate estimation from a wrist-worn accelerometer during sleep. *IEEE Journal of Biomedical and Health Informatics*, 2024. 2

[62] Subhas Chandra Mukhopadhyay. Wearable sensors for human activity monitoring: A review. *IEEE sensors journal*, 15(3):1321–1330, 2014. 2

[63] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 562–576. Springer, 2019. 7, 3

[64] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 5

[65] Rosalind W Picard. *Affective computing*. MIT press, 2000. 2, 8

[66] Ming-Zher Poh, Daniel 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. 2

[67] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252, 2021. 2

[68] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3437–3444. IEEE, 2023. 2

[69] Michael S Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 295–302, 2015. 2

[70] Rita Meziati Sabour, 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. 3

[71] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *2011 International Conference on Computer Vision*, pages 667–674. IEEE, 2011. 2

[72] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5: 258, 2017. 2

[73] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2

[74] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2733–2742, 2021. 2

[75] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins. Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*, pages 1–10, 2011. 2

[76] Isabelle M Shuggi, Hyuk Oh, Helena Wu, Maria J Ayoub, Arianna Moreno, Emma P Shaw, Patricia A Shewokis, and Rodolphe J Gentili. Motor performance, mental workload and self-efficacy dynamics during learning of reaching movements throughout multiple practice sessions. *Neuroscience*, 423:232–248, 2019. 2

[77] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011. 3

[78] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474, 2023. 2

[79] 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, 2014. 3, 2

[80] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9 (2):147–160, 2016. 2

[81] Zhaodong Sun and Xiaobai Li. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 7

[82] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: multi-domain mobile video physiology dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5. IEEE, 2023. 3, 7, 2

[83] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 8:235933, 2017. 2

[84] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2

[85] 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. 2, 4

[86] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[87] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2

[88] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Advances in Image and Video Technology: 5th Pacific Rim Symposium, PSIVT 2011, Gwangju, South Korea, November 20-23, 2011, Proceedings, Part I 5*, pages 277–288. Springer, 2012. 2

[89] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022. 2

[90] Zhixin Yan, Yan Zhong, Wenjun Zhang, Lin Shu, Hongbin Xu, and Wenxiong Kang. Physmamba: Leveraging dual-stream cross-attention ssd for remote physiological measurement. *arXiv preprint arXiv:2408.01077*, 2024. 2, 7

[91] Yu Yao, Mingze Xu, Chiho Choi, David J Crandall, Ella M Atkins, and Behzad Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9711–9717. IEEE, 2019. 2

[92] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 2, 4, 7

[93] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022. 2, 7

[94] Zijie Yue, Miaojing Shi, and Shuai Ding. Facial video-based remote physiological measurement via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13844–13859, 2023. 7

[95] 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*, pages 3438–3446, 2016. 3

[96] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 2

[97] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20110–20120, 2023. 2

[98] Bochao Zou, Zizheng Guo, Jiansheng Chen, and Huimin Ma. Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer. *arXiv preprint arXiv:2402.12788*, 2024. 7

[99] Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. Rhythmmamba: Fast remote physiological measurement with arbitrary length videos. *arXiv preprint arXiv:2404.06483*, 2024. 7

egoPPG: Heart Rate Estimation from Eye-Tracking Cameras in Egocentric Systems to Benefit Downstream Vision Tasks

Supplementary Material

10. Related datasets

Tab. 10 gives a comparison of the dataset size and activities of some related remote photoplethysmography (rPPG) datasets. In terms of hours of recordings and recorded frames, *egoPPG-DB* is among the largest dataset. Furthermore, we see that all comparable rPPG datasets only include activities with very little motion and heart rate (HR) changes such as watching videos, head rotations or talking. In contrast, *egoPPG-DB* features a wide variety of challenging everyday activities, such as kitchen work, dancing and riding an exercise bike, which induce significant motion artifacts and HR changes.

11. Excluded tasks

For all participants and activities, we checked the mean absolute error (MAE) between the predicted HR from our custom contact PPG sensor on the nose and the gold standard ECG from the chest belt. We excluded all tasks with an MAE over 3.0 beats per minute (bpm), which can happen, for example, when the PPG sensor loses alignment with the angular artery due to movement. In this way, we ensured that the photoplethysmography (PPG) signal from the nose, which we used as the target signal to train our model, is highly accurate. As a result, we had to exclude 20 out of the 150 tasks (13%), which we list in Tab. 6. We can see that this applied only to tasks with more motion (dancing, exercise bike, and walking). Since the participants had to walk multiple stairs throughout the data recording, this mostly happened during walking.

Activity	Excluded participants
Watch video	—
Office work	—
Kitchen work	—
Dancing	012, 015
Exercise bike	009, 012, 014, 015, 016, 023
Walking	004, 012, 013, 014, 018, 021, 022

Table 6. Detailed table of all excluded tasks.

12. Detailed description of activities

Tab. 11 gives a comprehensive description of the actions for each activity during our recording. Generally, participants were free to talk during the entire duration of the recording and conduct the tasks as they would do it normally. For example, during the kitchen work, the participants were completely free to prepare the sandwich and if they would like to eat or drink while doing it.

13. Data recording

In Fig. 8, we show a variety of different images and people of our data recording from a third person view to visualize the apparatus and capture protocol. All participants visible in these images explicitly agreed to be visualized.

14. Initial signal verification

In Fig. 5, we show the raw mean intensity values after spatial cropping of the skin region and the eye region (see Fig. 4) compared to the ground truth contact PPG signal from the nose. We can clearly see that the blood volume pulse is present both in the eyes and skin region with the skin region having a higher signal-to-noise ratio (SNR) compared to the eyes.

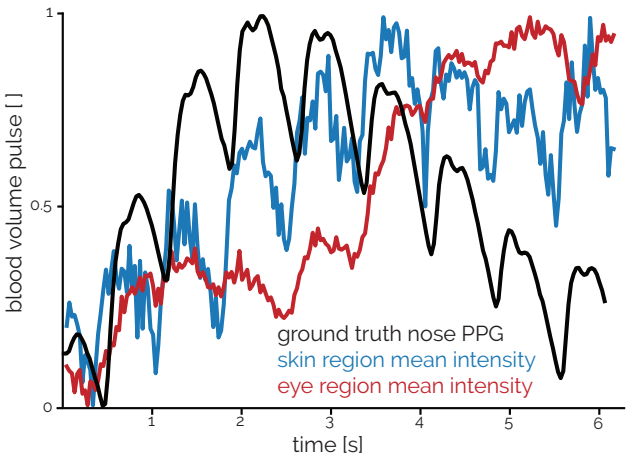

Figure 5. Example raw mean intensity of the skin and eye region, showing the higher SNR for the skin region around the eyes compared to the eyes.

15. Variance of results

In Fig. 6 we show the boxplot of the MAEs of the predictions of *PulseFormer* on *egoPPG-DB* by split. The interquartile range across all splits is between 1.7 and 10.5 bpm.

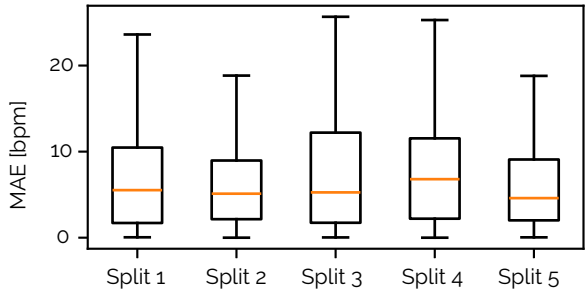

Figure 6. Boxplot of the MAEs of the predictions of *PulseFormer*.

16. Cross-dataset evaluation

We evaluated *PulseFormer* and the two strongest baselines when training on three conventional rPPG datasets (MMPD [82], UBFC-rPPG [6], and PURE [79]) and testing on *egoPPG-DB* (Tab. 7), and vice versa (Tab. 8). For the rPPG datasets, we extracted the eye region using MediaPipe [47], resized to 48×128, and converted to grayscale. *PulseFormer* consistently outperforms the baselines across all scenarios and datasets (except one case), showing strong generalization to unseen data. Please note that we can only evaluate *PulseFormer* w/o MITA as conventional rPPG datasets do not contain IMU data from the participants' heads.

Train Set	Model	MAE	MAPE
MMPD	PhysFormer	20.56	27.06
	FactorizePhys	Not converging	
	***PulseFormer* w/o MITA**	**13.66**	**16.64**
UBFC-rPPG	PhysFormer	18.32	23.63
	FactorizePhys	18.58	24.46
	***PulseFormer* w/o MITA**	**14.83**	**18.57**
PURE	PhysFormer	24.39	24.94
	FactorizePhys	13.20	15.44
	***PulseFormer* w/o MITA**	**12.99**	**13.46**

Table 7. Results (MAE) when training on conventional rPPG datasets and testing on *egoPPG-DB*.

Model	MMPD		UBFC-rPPG		PURE	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
PhysFormer	11.76	14.57	16.80	16.46	23.89	37.50
FactorizePhys	12.06	15.11	**14.28**	**14.98**	26.10	40.62
***PulseFormer* (ours)**	**11.48**	**15.08**	15.09	15.81	**23.56**	**36.71**

Table 8. Results (MAE) when training on *egoPPG-DB* and testing on conventional rPPG datasets.

17. HR distribution

egoPPG-DB exhibits the widest HR range (44–164 bpm, see Fig. 7) and significantly more motion (e.g., dancing, exercise bike) than other evaluated rPPG datasets, where participants typically sit calmly at a table.

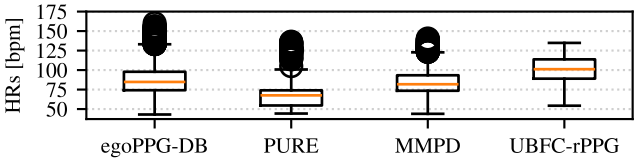

Figure 7. Boxplot of HRs of *egoPPG-DB* and three rPPG datasets.

18. Downstream performance comparison

HR features from the other evaluated baselines perform progressively worse than those from *PulseFormer* when used for proficiency estimation on EgoExo4D, highlighting the importance of accurate HR estimation for downstream tasks (see Tab. 9).

Model	Ego+HR	Exo+HR	Ego+Exo+HR
FactorizePhys	44.62	36.72	40.13
PhysFormer	44.39	36.66	43.07
***PulseFormer* (ours)**	**45.29**	**37.67**	**43.94**

Table 9. Downstream performance (accuracy) on EgoExo4D using the HR predictions from the three best baseline models.

Dataset	Part.	Frames	Hours	Tasks
PURE [79]	10	110 K	1	Resting, talking, small head movements
MAHNOB-HCI [77]	27	2.6 M	12	Watching videos
MMPD [82]	33	1.2 M	11	Resting, head rotation, selfie videos
MMSE-HR [95]	40	310 K	2	Talking, watching videos, experiencing different emotions
UBFC-rPPG [6]	43	150 K	1.5	Gaming on a computer
UBFC-PHYS [70]	56	2.4 M	19	Resting, Trier Social Stress Test
OBF [44]	106	3.8 M	18	Resting with varying HR levels
VIPL-HR [63]	107	**4.3 M**	**20**	Resting, talking, head rotation, different lighting conditions
SCAMPS (synthetic) [54]	**2800**	1.7 M	16	Different facial actions
egoPPG-DB (ours)	25	1.4 M	13	Watching videos, office and kitchen work, dancing, biking, walking

Table 10. Summary of existing datasets for rPPG.

Activity	Actions	Description
Watch video	Watch a documentary	Watch a relaxing documentary on a computer.
Office work	Work on a computer	Randomly browse through websites and type text from a PDF into Word.
	Write on a paper	Write a text from a PDF on a computer onto a piece of paper.
	Talk to the experimenter	Have a free, unscripted conversation with the experimenter.
Walking	Walk to the kitchen	Walk along a hallway, down the stairs into the kitchen.
Kitchen work	Get ingredients	Get all ingredients for a sandwich from the fridge.
	Cut vegetables	Get a cutting board, knife and a plate and cut vegetables.
	Prepare a sandwich	Put the bread into the toaster and afterward freely prepare sandwich.
	Eat sandwich/drink	Participants are free to eat the sandwich or drink during the recording.
	Wash the dishes	Wash everything used while preparing the sandwich.
Walking	Walk to the dancing room	Walking along a hallway into a new room for dancing and biking.
Dancing	Follow random dance video	Choose a dance video and afterward follow it.
Exercise bike	Ride an exercise bike	Ride an exercise bike with moderate to high intensity.
Walking	Walk back to the physical location of the start	Walk back to the physical location of the start either up the stairs or using the elevator.

Table 11. Detailed capture protocol and action descriptions of the *egoPPG-DB* dataset.

Figure 8. Additional images of the data recording showing the variety of everyday activities our dataset includes.