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

Björn Braun Rayan Armani Manuel Meier Max Moebus Christian Holz
Department of Computer Science, ETH Zurich
{bjoern.braun, rayan.armani, max.moebus, christian.holz}@inf.ethz.ch

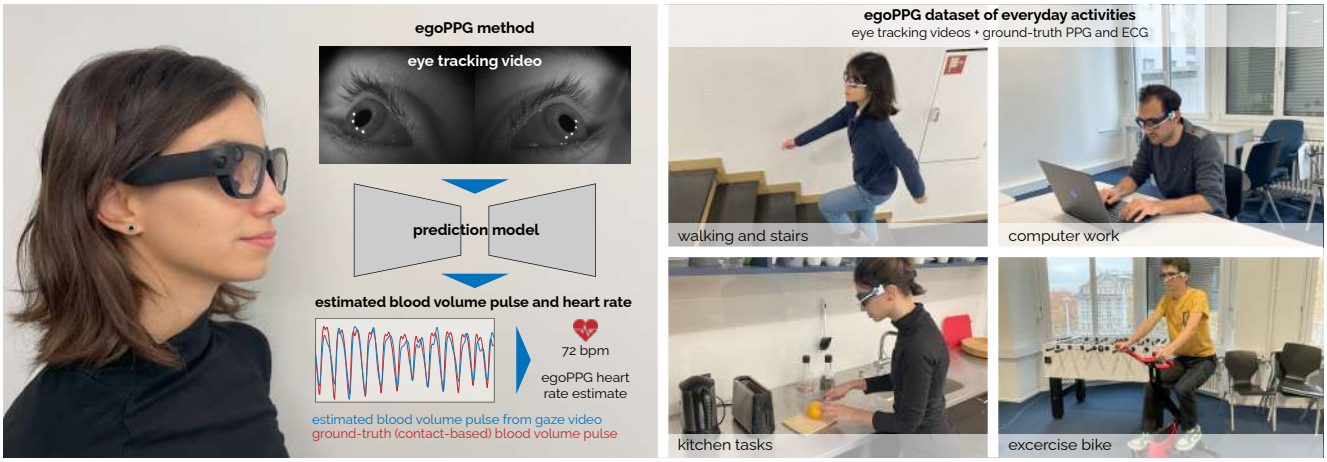

Figure 1. Implementing a novel vision task egoPPG, *EgoPulseFormer* tracks a person's heart rate (HR) from eye-tracking videos captured by unmodified egocentric vision headsets. Our method estimates the person's photoplethysmogram (PPG) of the blood volume pulse from areas around the eyes to extract HR values. For training and validation, we collected egoPPG-DB, a dataset of participants' eye-tracking videos during everyday activities with synchronized ground-truth PPG signals (white contact sensor) and HR values (ECG chest strap).

Abstract

Egocentric vision systems aim to understand the spatial surroundings and the wearer's behavior inside it, including motions, activities, and interaction with objects. Since a person's attention and situational responses are influenced by their physiological state, egocentric systems must also detect this state for better context awareness. In this paper, we propose egoPPG, a novel task for egocentric vision systems to extract a person's heart rate (HR) as a key indicator of the wearer's physiological state from the system's built-in sensors (e.g., eye tracking videos). We then propose EgoPulseFormer, a method that solely takes eye-tracking video as input to estimate a person's photoplethysmogram (PPG) from areas around the eyes to track HR values—without requiring additional or dedicated hardware. We demonstrate the downstream benefit of EgoPulseFormer on EgoExo4D, where we find that augmenting existing models with tracked HR values improves proficiency estimation by 14%. To train and validate EgoPulseFormer, we collected a dataset of 13+ hours of eye-tracking videos from Project Aria and contact-based blood volume pulse signals as well as an electrocardiogram (ECG) for ground-truth HR values. 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=8.82 bpm) and captures patterns (r=0.81). Our results show how egocentric systems may unify environmental and physiological tracking to better understand user actions and internal states.

1. Introduction

Egocentric vision systems, such as Mixed Reality (MR) glasses by Meta [56], Magic Leap [46], and others have emerged as powerful devices for capturing and analyzing a person's behavior as well as their surrounding environment from a first-person perspective. The wider availability of promising wearable capture platforms (e.g., Project

Aria glasses [23]) has sparked a large amount of research on egocentric vision tasks for environment understanding and navigation, including localization [41, 71, 74], and simultaneous localization and mapping (SLAM) [17, 38, 68]. Since egocentric systems can simultaneously capture parts of the wearer's behavior, many efforts have investigated egocentric action recognition [44, 52, 86, 89, 94] and hand-object interaction [22, 32, 73, 95] to understand user behavior. Large-scale datasets have also been introduced in recent years to accelerate data-driven research in this domain, such as Ego4D [32], Nymeria [51], and EgoExo4D [33], offering rich, multimodal data to enable training and evaluation for these tasks.

In addition to spatial awareness, understanding the user's behavior, their attention, and intent is equally important for egocentric systems [3, 58, 88]. For instance, anticipating the user's next action is crucial for applications in navigation, personalized feedback, and autonomous assistance [20, 84, 91]. Objects of interest are commonly estimated from analyzed gaze patterns [27, 37, 45, 47] to support behavioral analysis and social understanding [26, 32, 33, 40].

However, holistically modeling a person's behavior and intent requires knowledge of their physiological state, which influences cognitive performance, attention, and situational responses [15, 25, 53, 76, 81]. Key components of physiological state include cardiovascular indicators such as heart rate (HR) and electrodermal activity [1, 9, 59, 79], which reflect emotions, stress, fatigue, and alertness [1, 11, 59, 65, 79]. Capturing these dynamics can thus benefit models of human behavior to enable a richer understanding of user actions for adaptive systems.

In this paper, we introduce *egoPPG*, a novel task for egocentric vision systems to accurately extract a person's HR measurements from the system's built-in sensors, specifically the eye-tracking cameras in unmodified headsets. We then propose *EgoPulseFormer*, a novel method that implements egoPPG on Project Aria glasses to demonstrate the benefit of egoPPG for existing downstream vision tasks on large egocentric datasets. Our learning-based method EgoPulseFormer is designed to recover the person's photoplethysmogram (PPG) from the subtle fluctuations in skin intensity due to light absorption in pulsatile arteries beneath the skin surface following a blood volume pulse (BVP), in particular deriving it from regions around the wearer's eye for robust tracking. Leveraging the infrared (IR) illuminant of the eye tracker, it work across light conditions [50]. In addition, we demonstrate the benefits of estimated HR values in egocentric recordings for a key vision task downstream: We augment an existing architecture with *EgoPulseFormer* and show its impact on EgoExo4D's proficiency estimation benchmark, whose accuracy improves by 14.1% with *EgoPulseFormer*'s HR estimates.

The sensing configuration of our task can be considered a hybrid between typical contact-based BVP sensors (e.g., those in smartwatches [5, 13, 21, 31, 61, 63]) and rPPG methods that aim to extract HR from a person's face using a camera [39, 66, 82]. While egocentric glasses are body-worn much like wrist watches, they couple more loosely to the body and are subject to considerable motion artifacts. Unlike contact sensors, eye trackers observe the wearer's eye regions from a short distance and capture eye motions and blinks, leading to ambiguity and noise for capturing fluctuations in skin intensity. Unlike rPPG configurations, egocentric capture systems move with the wearer's body and head, and eye trackers use controlled illumination.

Therefore, we designed *EgoPulseFormer* to extract BVP from the region with the least motion—around the wearer's eyes—by incorporating spatial attention within our model's backbone. We validate *EgoPulseFormer*'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 five real-world tasks with varying motion and intensity, causing their HR values to reach levels between 44–164 bpm.

We summarize our key contributions as follows:

1. egoPPG as a novel task and EgoPulseFormer 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=8.82 bpm), with a 27% lower error than current state-of-the-art rPPG models [12, 48, 90, 92].
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 [32, 33, 51].
3. a validation of egoPPG's downstream benefits for proficiency estimation. Augmenting EgoExo4D with continuous HR values and additionally feeding them into an existing architecture, we demonstrate the implications of our method on the proficiency estimation benchmark with an increase in accuracy by 14.1%.

2. Related work

Egocentric vision. In recent years, research in egocentric vision has surged, driven by advances in AR/VR glasses [4, 23, 36, 46, 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 [44, 52, 86, 89, 94] and anticipation [16, 30, 60, 86], full-body pose estimation [75, 87], responding to user needs [67, 69, 91], and social behavior analysis [26, 32, 40]. Ad-

ditionally, tracking vital signs in AR/VR settings and for affective computing applications [1, 11, 59, 65, 79] has become an important tool for understanding users' physiological states [54], further aiding in understanding users' behavior, their 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 [13, 21, 62, 63]. Heart rate (HR), in particular, is a key measure for assessing an individual's health and performance [24, 29, 42, 72]. While wearable sensors, such as wrist-worn smartwatches, provide accurate HR measurements and also challenging scenarios (e.g., exercising), they are intrusive and can cause discomfort [43]. Recent research has, thus, extensively explored using cameras as an unobtrusive, non-contact alternative for measuring HR, generally called remote photoplethysmography (rPPG) [39, 66, 82]. rPPG measures HR based on subtle color changes in the skin caused by the BVP. Generally, rPPG methods can be broadly divided into traditional signal processing techniques [8, 18, 19, 39, 66, 82, 83] and deep learning-based approaches [10, 12, 48, 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 [7, 34, 64, 70, 78] or requiring users to hold a smartphone steadily in front of their face [80]. As a result, rPPG is not feasible to be deployed in more dynamic settings, such as during exercise.

Eye tracking cameras. Eye tracking in egocentric vision systems is mostly done using inward-facing cameras direct 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 [14]. Furthermore, IR illumination makes them robust to lighting variations and low-light conditions [50]. To the best of our knowledge, videos from eye tracking cameras have not yet been explored for HR estimation using the BVP despite their promise to enable unobtrusive HR measurements during everyday life.

3. Overview

Our aim is to enable egocentric vision systems i) to track 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 during a series of everyday activities. Sec. 5 then outlines our method *EgoPulseFormer* that continuously estimates a person's BVP from eye-tracking videos and derives HR values from it. Fig. 3 illustrates our approach. To demonstrate *EgoPulseFormer*'s benefit for downstream applications, we leverage HR estimation for the user proficiency benchmark of the EgoExo4D dataset, described in Sec. 6. Finally, Sec. 7 provides all results from our evaluations and Sec. 8 discusses our findings.

4. egoPPG-DB

The *egoPPG-DB* dataset was developed to support HR estimation from eye-tracking videos under real-world conditions, with a protocol designed to elicit significant motion and fluctuations in HR. 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 [28], 9 participants had skin type II, 10 had skin type III, 2 had skin type IV, and 4 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. Participants were instructed to avoid wearing makeup prior to the recording. The data collection was approved by the *anonymized* Ethics Commission (no. *#anonymized*). In terms of dataset size by duration, *egoPPG-DB* is third amongst the longest rPPG datasets [7, 34, 64, 70, 78, 80] as listed in Tab. 6.

4.2. Apparatus

Fig. 2 illustrates our experimental setup. We used Project Aria glasses [23] 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 [35]. 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.

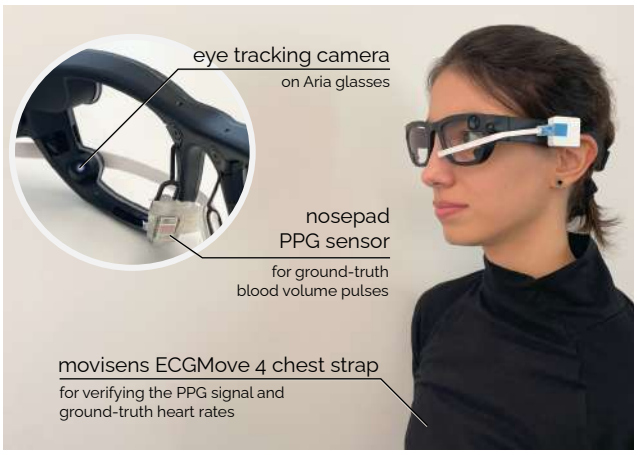

Figure 2. Apparatus used to record the *egoPPG-DB* dataset.

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.

4.3. Capture protocol

The average duration of the recording of participant was 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 [33] or Nymeria [51]. In Tab. 8, we give a detailed description of each activity, and in Tab. 3 we show the mean HR values for each activity. Exercising on the bike produced the highest HR values (113 bpm), whereas watching the video resulted in the lowest mean HR (71 bpm).

4.4. Dataset and signal quality verification

To evaluate that the contact PPG sensor, whose signal we later use as the target for model training, produces accurate HR values, we evaluate it against the gold-standard ECG. We assessed the performance by calculating 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, which can happen when the PPG sensor occasionally loses alignment with the angular artery due to movement. This applied to 20 of 150 tasks (13%, see Tab. 7). During the remaining tasks, our custom-built PPG nose sensor achieved very high accuracy, with an average MAE of 1.3 bpm and a mean correlation of 0.94 compared to the ECG signal, showing its suitability as ground truth.

5. EgoPulseFormer: a first method for egoPPG

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 a 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. [83] 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[83]. 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 [92] with a temporal input length of $T = 128$ frames (corresponding to 4.3 seconds) downsampled to $(h = 48) \times (w = 128)$

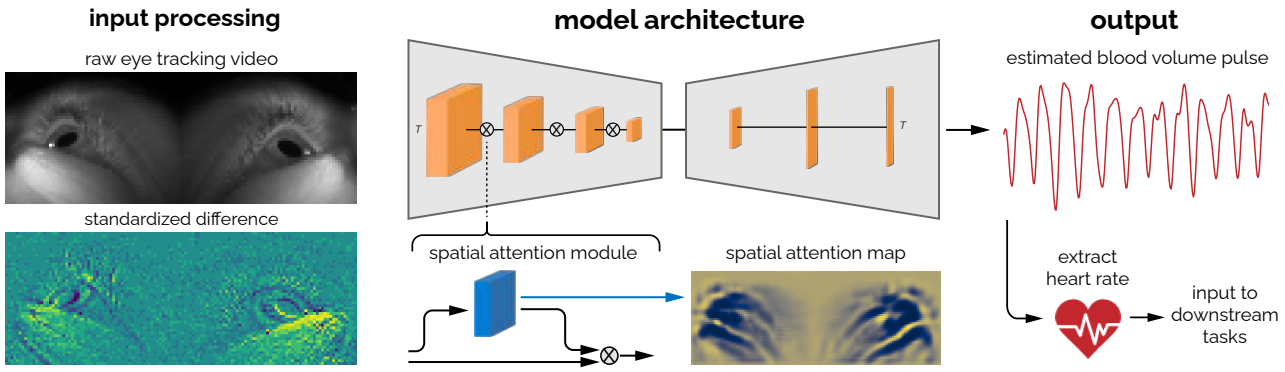

Figure 3. Architecture of our model for continuous BVP estimation from eye-tracking videos and consecutive HR computation.

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. Compared to the video of a person's face, which is usually used for rPPG tasks, eye tracking videos offer additional challenges. 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 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 [85] 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 an input feature map $\boldsymbol{F} \in \mathbb{R}^{T \times C \times w \times h}$, the spatial attention modules infers a spatial attention map $\boldsymbol{M_s} \in \mathbb{R}^{T \times 1 \times w \times h}$ as:

$$\boldsymbol{M_s}(\boldsymbol{F}) = \sigma * (f^{7\times7}([\boldsymbol{F_{avg}}; \boldsymbol{F_{max}}])) \tag{3}$$

where σ is the sigmoid function, $f^{7\times7}$ 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 $\mathbf{F_{out}}$ of the attention process is then:

$$\mathbf{F_{out}} = \mathbf{M_s} \otimes \mathbf{F}, \tag{4}$$

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 in the glasses' positioning; (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 individual differences in skin region visibility. Finally, to help the network focus on the changes between the frames caused by the periodic BVP, we use the standardized consecutive frame differences of the eye tracking videos as input into our network. We standardize each frame by subtracting the mean pixel intensity and dividing it by the standard deviation of the pixel intensities values [12]. We use the standardized consecutive differences of the PPG signals from the nose as the labels for our model. Fig. 3 shows our architecture with the input preprocessing demonstrated on the left side, and an example learned spatial attention map on the right side. For one batch, the total number of FLOPS for our model is about 164 GigaFLOPS. The total number of parameters is about 770k, where the spatial attention modules added only 400 parameters.

5.3. Experiments setup

5.3.1. Training

We trained all implemented 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 used the remaining participants as the training set. The training was conducted with a batch size of 4 for 100 epochs, a learning rate of 0.0009, and mean squared error (MSE) as the loss function. In addition to our model, we used four baseline networks (one state-space and three CNNs), which have state-of-the-art performance for rPPG, to compare the performance of our proposed model to the performance of these established models. Our model was trained on a GeForce RTX 4090, with a total runtime

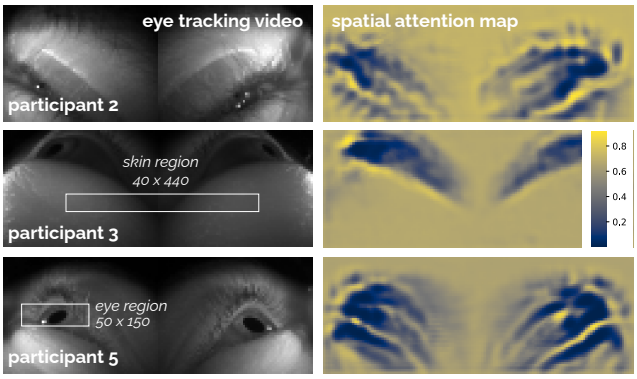

Figure 4. Left: Head geometry determines the regions that the eye tracker captures. Right: Learned spatial attention maps show that eye regions are excluded and *EgoPulseFormer* instead extracts BVP from the surrounding skin regions, which moves less than the eyes.

of about 15 hours for all folds.

5.3.2. Metrics

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 sliding window, as commonly used for rPPG [18, 48, 49, 55, 92].

5.3.3. Video sampling rate

While we recorded the eye tracking videos with 30 fps, large-scale datasets such as EgoExo4D [33] or Nymeria [51] used only 10 fps. To assess the impact of reduced frame rates, we evaluated i) model performance when downsampling our videos to 10 fps by retaining only every third frame and ii) model performance when downsampling to 10 fps, then linearly interpolating between frames to upsample to 30 fps again. For both scenarios, we train from random initialization.

5.3.4. Hyperparameter optimization

We also provide insights into optimal hyperparameter configurations by analyzing the impact of image size, input length, loss function, and choice of label (see Tab. 9).

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 [33]. The user proficiency estimation benchmark aims to classify the proficiency of a user (novice, early expert, intermediate expert, late expert) using only egocentric video clips (*Ego*), only exocentric video clips (*Exo*), or all video clips together (*Ego + Exo*). Our goal was to assess if we can improve the performance of the current baseline model (TimeSFormer [6]) when integrating our predicted HR data into the network. This results in two additional configurations: using egocentric videos and HR together (*Ego + HR*), and using all video clips and HR together (*Ego + Exo + HR*). To predict the continuous HRs for all EgoExo4D videos, we use our proposed method, pre-trained on *egoPPG-DB*.

We implement the TimeSFormer model in exactly the same configuration as for the current benchmark results [33] with a clip size of 16 frames and a sampling rate of 16, trained for 15 epochs on four GeForce RTX 4090. We use all videos of the EgoExo4D dataset, for which the proficiency estimation labels are available (using the official benchmark training and validation sets) and which have at least 16 frames at a sampling rate of 16, resulting in totally 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, standard deviation, minimum and maximum HR, and mean HR change) for the corresponding videos. To integrate our HR predictions into the TimeSFormer architecture, we simply feed our calculated features into a 50-parameter linear layer and concatenate the output with the backbone's output before feeding it into the network head. We train all models from random initialization and evaluate proficiency estimation using top-1 accuracy per the EgoExo4D protocol.

7. Experiments

7.1. Heart rate estimation

7.1.1. 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 for comparison. 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). After spatially cropping the images, we calculate the mean pixel intensity values and remove motion artifacts by discarding any changes in the signal that are outside the first and third quantiles of the signal's changes. Finally, we filter the signal with a 4th order Butterworth bandpass filter between 0.6 and 3.0 Hz (corresponding to 36 to 180 bpm) to obtain the BVP (see Fig. 6).

7.1.2. EgoPulseFormer: an egoPPG method

Using our proposed network, we obtain an MAE of 8.82 bpm and a correlation of 0.81 between our predicted HR and the ground truth HR (see Tab. 2). This is an improvement of 3.27 bpm (27.0%) of the MAE and 0.15 for

Model	MAE	RMSE	MAPE	r
DeepPhys [12]	28.26	31.97	36.68	0.08
TS-CAN [48]	26.32	32.39	29.13	0.11
Mean intensity eyes	14.60	18.18	18.37	0.20
PhysMamba [90]	13.94	16.86	17.76	0.61
Mean intensity skin	12.40	15.54	15.29	0.50
PhysNet [92]	12.09	15.43	15.14	0.66
EgoPulseFormer (ours)	**8.82**	**12.03**	**10.82**	**0.81**
Improvement over second-best method	**−3.27**	**−3.4**	**−4.32**	**+0.15**

Table 2. Results for HR prediction from eye tracking videos using different models (*EgoPulseFormer* and established rPPG baselines).

Activity	μ HR	Motion magnitude	MAE	RMSE	MAPE
Video	71.45	0	**6.25**	**8.50**	**9.88**
Office	75.65	0.45	8.90	11.97	12.60
Kitchen	85.31	0.54	8.75	11.11	10.42
Dancing	89.07	**1.00**	10.02	12.79	11.22
Bike	**113.06**	0.77	13.51	16.52	11.04
Walking	93.71	0.30	7.02	9.31	6.70

Table 3. Results for HR prediction split by activity.

Input video	MAE	RMSE	MAPE	r
10 fps (other datasets)	12.40	16.80	14.10	0.52
Upsampled to 30 fps	**10.20**	**13.56**	**12.78**	**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 [33]. In the second row, we first downsample our videos to 10 fps and then upsample them to 30 fps by linearly interpolating between frames.

the correlation compared to the second-best network (PhysNet [92]). Split by activity, we obtain the lowest MAE while the participants are watching a video (MAE of 6.25 bpm) and the highest MAE during exercising on a bike (MAE of 13.51 bpm) and while dancing (MAE of 10.02 bpm), which are both also the tasks with the highest normalized motion magnitude. 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. We, then, normalize it between zero and one across all activities to get a measure of the amount of motion of each activity. Furthermore, using simple signal processing and spatial cropping, we obtain an MAE of 12.40 and a correlation of 0.50 when using the skin region around the eyes. When using the eyes regions as input, the MAE increases to 14.60, and the correlation drops to 0.20. This is also reflected in spatial attention maps that our model implicitly learns (see Fig. 4), which exclude the eyes for predicting the HR. To qualitatively cross-check these results, Fig. 6 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. The MAE increases to 12.40 bpm and the correlation decreases to 0.52 when training and testing using 10 fps. When upsampling the videos again to 30 fps using linear interpolation, the MAE decreases to 10.20 bpm and the correlation increases to 0.77.

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 the EgoExo4D dataset. We see that integrating our predicted HRs into the TimeSFormer model [6] improved accuracy for all scenarios but one. Additionally, we also achieved the highest accuracy for each individual scenario with our HR integration for the same scenarios. When combining the egocentric videos with our predicted HRs, we achieved an overall accuracy of 45.20%, 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%, respectively. Also, when using the egocentric videos, exocentric videos, and our predicted HRs together, the accuracy increased by 8.64% from 39.00% to 42.37% compared to using only the egocentric and exocentric videos. Using only exocentric videos yielded the lowest overall accuracy at 35.93%.

8. Discussion

8.1. Heart rate estimation

When evaluating our proposed method on *egoPPG-DB*, we showed that HR can reliably be predicted from eye tracking videos of unmodified egocentric vision headsets. Compared to the performances achieved on popular rPPG datasets such as PURE [78], UBFC-RPPG [7], and MMPD [80], we achieved very competitive performance in challenging settings [49]. While the lowest reported MAE for UBFC-RPPG in the rPPG-toolbox [49] is 1.21 bpm, its participants sit almost motion-free with their eyes closed. Already on MMPD, a dataset with varied lighting and little motion, the lowest reported MAE increases to 10.23 bpm [49], as it incorporates more realistic challenges due to recordings on mobile devices with light motion (head rotation, talking, and taking selfies). In comparison, using *EgoPulseFormer*, we achieved a lower MAE of 8.82 bpm while even having

Scenario	Majority	Ego	Ego + HR (ours)	Exo	Exo + HR (ours)	Ego + Exo	Ego + Exo + HR (ours)
Basketball	38.00	45.45	47.47	49.24	49.24	49.49	**52.52**
Cooking	0.00	20.00	**40.00**	33.75	40.00	25.00	**40.00**
Dancing	24.59	43.44	53.27	45.08	48.36	50.82	**55.73**
Music	57.89	78.94	**81.58**	57.89	57.89	57.89	60.53
Bouldering	15.29	24.50	**27.81**	10.26	12.58	15.89	18.54
Soccer	62.50	50.00	56.25	**76.56**	75.00	75.0	62.50
Overall	27.80	39.69	**45.29**	35.93	35.93	39.00	42.37

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

tasks with heavy motion (dancing, walking stairs) and very high HR changes (44–164 bpm).

8.1.1. Performance depending on activity

Analyzing our results split by activity (see Tab. 3), we see that the tasks with the highest mean HR (bike, 113 bpm) and motion magnitude (dancing) also have the highest MAE. Given the substantially bigger motion artifacts and HR variability (HRs between 44 and 164 bpm) in our dataset — captured during diverse everyday activities and physical exercises — we believe that our results demonstrate the robustness of *EgoPulseFormer* in more dynamic, everyday conditions.

8.1.2. Performance depending on method

For our model, we leveraged spatial attention maps to improve performance. When qualitatively analyzing the learned spatial attention maps, we see that our model implicitly learned to exclude the eyes for estimating the BVP from the eye tracking videos (see Fig. 4). This aligns with our obtained qualitative (see Fig. 6) and quantitative (see Tab. 2) results using simple signal processing, which show a higher SNR for the skin region compared to the eyes. When analyzing the performance of other state-of-the-art rPPG models on our dataset, we found that only PhysNet [92] and PhysMamba [90] achieved reasonable performances, though their results did not reach our performance.

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 [33] or Nymeria [51], 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 both 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 43.4% to 53.3%), 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 [33]. These findings suggest that adding HR data as an auxiliary signal enhances performance for the proficiency estimation benchmark on the EgoExo4D dataset. For training and testing, we used the available subset of EgoExo4D videos for which proficiency labels are available, following the official training and validation 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. Limitations and future work

We observed the highest MAE in tasks with elevated HRs and motion, such as dancing and biking. We believe that promising approaches to address these limitations could be to record more tasks with high HRs and integrating the IMU data from the glasses. Furthermore, future studies could extend data collection to outdoor settings to assess the impact of varying lighting conditions. While we ensured a balanced gender ratio (13 male, 12 female), our sample size of 25 participants restricts broader demographic conclusions. In future work, we aim to expand the dataset to investigate performance variations across different age groups, skin types, and ethnicities. Finally, we believe that

it is a highly interesting problem to explore further downstream applications of a user's physiological state, such as personalized feedback, autonomous assistance, as well as health-related applications.

9. Conclusion

We introduced *egoPPG*, a novel task for egocentric vision systems to extract the wearer's heart rate for integrating their physiological state into egocentric vision tasks downstream. Our method *EgoPulseFormer* processes input from the eye tracking cameras on unmodified egocentric vision systems to robustly estimate the person's HR in various everyday scenarios. We validate *EgoPulseFormer's* robustness on our dataset *egoPPG-DB* and demonstrate significant improvements over existing rPPG models. With HR estimations from *EgoPulseFormer* we 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.

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, 3

[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, 3

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

[5] Apple. Apple watch. https://www.apple.com/watch/, 2024. Accessed: 2024.11.13. 2

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 6, 7

[7] 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

[8] 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. 3

[9] Björn Braun, Daniel McDuff, Tadas Baltrusaitis, Paul Streli, Max Moebus, and Christian Holz. Sympcam: Remote optical measurement of sympathetic arousal. *arXiv preprint arXiv:2410.20552*, 2024. 2

[10] Björn Braun, Daniel McDuff, and Christian Holz. How suboptimal is training rppg models with videos and targets from different body sites? *arXiv preprint arXiv:2403.10582*, 2024. 3

[11] Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. *The Oxford handbook of affective computing*. Oxford University Press, USA, 2015. 2, 3

[12] 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, 3, 5, 7

[13] 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, 3

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

[15] 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

[16] 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

[17] 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

[18] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013. 3, 6

[19] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 3

[20] Benjamin J Dixon, Michael J Daly, Harley Chan, Allan D Vescan, Ian J Witterick, and Jonathan C Irish. Surgeons blinded by enhanced navigation: the effect of augmented reality on attention. *Surgical endoscopy*, 27:454–461, 2013. 2

[21] Jessilyn Dunn, Ryan Runge, and Michael Snyder. Wearables and the medical revolution. *Personalized medicine*, 15(5): 429–448, 2018. 2, 3

[22] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Visual object tracking in first person vision. *International Journal of Computer Vision*, 131(1):259–283, 2023. 2

[23] 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

[24] 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. 3

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

[26] 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

[27] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 314–327. Springer, 2012. 2

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

[29] 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. 3

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

[31] Google. Google pixel watch 3. https://store.google.com/product/pixel_watch_3?hl=de&pli=1, 2024. Accessed: 2024.11.13. 2

[32] 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

[33] 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, 4, 6, 7, 8

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

[35] 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

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

[37] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 2

[38] 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

[39] 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, 3

[40] 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

[41] 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

[42] 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. 3

[43] James F. Knight, Daniel Deen-Williams, Theodoros N. Arvanitis, Chris Baber, Sofoklis Sotiriou, Stamatina Anastopoulou, and Michael Gargalakos. Assessing the wearability of wearable computers. In *2006 10th IEEE International Symposium on Wearable Computers*, pages 75–82, 2006. 3

[44] 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

[45] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *arXiv preprint arXiv:2208.04464*, 2022. 2

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

[47] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 2

[48] 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, 3, 6, 7

[49] 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. 6, 7

[50] 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. 2, 3

[51] 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, 4, 6, 8

[52] 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

[53] 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

[54] 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. 3

[55] Daniel McDuff, Javier Hernandez, Erroll Wood, Xin Liu, and Tadas Baltrusaitis. Advancing non-contact vital sign measurement using synthetic avatars. *arXiv preprint arXiv:2010.12949*, 2020. 6

[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, 3

[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, 3

[60] Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18580–18590, 2024. 2

[61] Max Moebus and Christian Holz. Personalized interpretable prediction of perceived sleep quality: Models with meaningful cardiovascular and behavioral features. *Plos one*, 19(7): e0305258, 2024. 2

[62] 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. 3

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

[64] 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. 3, 2

[65] Rosalind W Picard. *Affective computing*. MIT press, 2000. 2, 3

[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, 3

[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, 2

[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. 3

[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. 2

[78] 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 Sym-*

posium on Robot and Human Interactive Communication, pages 1056–1062, 2014. 3, 7, 2

[79] 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, 3

[80] 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

[81] 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

[82] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2, 3

[83] 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. 3, 4

[84] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023. 2

[85] 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

[86] 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

[87] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. 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, 3

[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, 3, 7, 8

[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, 3, 4, 6, 7, 8

[93] 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. 2

[94] 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

[95] 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

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

Supplementary Material

10. Related datasets

Tab. 6 gives a comprehensive comparison of the dataset size and activities of related remote photoplethysmography (rPPG) datasets. In terms of hours of recordings and recorded frames, *egoPPG-DB* is the third 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 riding and 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 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. 7. 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.

12. Detailed description of activities

Tab. 8 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. 5, 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. 6, 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.

15. Hyperparameter optimization

In Tab. 9, we show the results of *EgoPulseFormer* when training our model with different hyperparameters such as different image sizes, input window sizes, or target signals.

Dataset	Part.	Frames	Hours	Tasks
PURE [78]	10	110 K	1	Resting, talking, small head movements
MAHNOB-HCI [77]	27	2.6 M	12	Watching videos
MMPD [80]	33	1.2 M	11	Resting, head rotation, selfie videos
MMSE-HR [93]	40	310 K	2	Talking, watching videos, experiencing different emotions
UBFC-rPPG [7]	43	150 K	1.5	Gaming on a computer
UBFC-PHYS [70]	56	2.4 M	19	Resting, Trier Social Stress Test
VIPL-HR [64]	**107**	**4.3 M**	**20**	Resting, talking, head rotation, different lighting conditions
egoPPG-DB (ours)	25	1.4 M	13	Watching videos, office and kitchen work, dancing, biking, walking

Table 6. Summary of existing datasets for rPPG.

Figure 5. Additional images of the data recording showing the variety of everyday activities our dataset includes.

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 7. Detailed table of all excluded tasks.

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 start	Walk back to the start either up the stairs or using the elevator.

Table 8. Detailed capture protocol and action descriptions of the *egoPPG-DB* dataset.

Hyperparameter	Configuration	MAE	RMSE	MAPE	r
Image preprocessing	Raw	16.33	20.04	20.93	0.31
	Standardized	14.44	18.21	18.66	0.46
	Difference	8.99	12.09	10.96	0.80
	Standardized difference	8.82	12.03	10.82	0.81
Image size	24×64	9.73	12.69	12.12	0.77
	$\mathbf{48 \times 128}$	8.82	12.03	10.82	0.81
	96×256	10.27	13.10	12.46	0.79
Window size	64	9.95	12.82	11.93	0.80
	128	8.82	12.03	10.82	0.81
	256	9.71	13.02	11.41	0.78
Loss function	Negative Pearson [92]	10.62	13.57	13.16	0.81
	MAE	9.71	13.02	11.41	0.78
	MSE	8.82	12.03	10.82	0.81
Target signal	Heart rate	10.99	14.03	13.49	0.77
	Nose PPG	8.82	12.03	10.82	0.81
Augmentation	Without	10.57	14.19	12.78	0.70
	With	8.82	12.03	10.82	0.81

Table 9. Results of our model on the *egoPPG-DB* dataset when training with different hyperparameters.

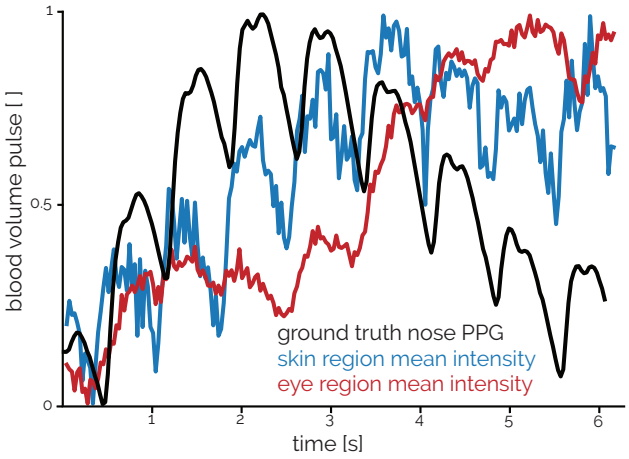

Figure 6. Example raw mean intensity of the skin and eye region, showing the higher SNR for the skin region around the eyes compared to the eyes.