Orientation-conditioned Facial Texture Mapping for Video-based Facial Remote Photoplethysmography Estimation

Sam Cantrill[1,2*] David Ahmedt-Aristizabal[2]
Lars Petersson[2] Hanna Suominen[1,2,3] Mohammad Ali Armin[2]
[1]Australian National University, Canberra, Australia
[2]Data61, Commonwealth and Scientific Industrial Research Organization, Canberra, Australia
[3]University of Turku, Turku, Finland
{sam.cantrill, hanna.suominen}@anu.edu.au
{ali.armin, david.ahmedtaristizabal, lars.petersson}@data61.csiro.au

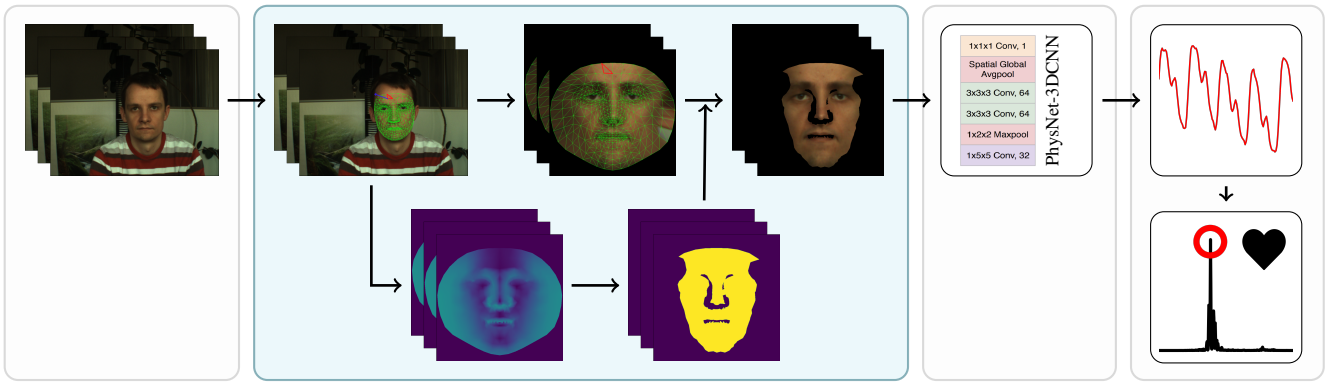

Figure 1. Proposed methodology for constructing the orientation-conditioned facial texture video using UV-coordinate texture mapping to enhance the motion robustness of camera-based remote photoplethysmography (rPPG) and downstream pulse rate (PR) estimation.

Abstract

Camera-based remote photoplethysmography (rPPG) enables contactless measurement of important physiological signals such as pulse rate (PR). However, dynamic and unconstrained subject motion introduces significant variability into the facial appearance in video, confounding the ability of video-based methods to accurately extract the rPPG signal. In this study, we leverage the 3D facial surface to construct a novel orientation-conditioned facial texture video representation which improves the motion robustness of existing video-based facial rPPG estimation methods. Our proposed method achieves a significant 18.2% performance improvement in cross-dataset testing on MMPD over our baseline using the PhysNet model trained on PURE, highlighting the efficacy and generalization benefits of our designed video representation. We demonstrate significant performance improvements of up to 29.6% in all tested motion scenarios in cross-dataset testing on MMPD, even in the presence of dynamic and unconstrained subject motion. Emphasizing the benefits the benefits of disentangling motion through modeling the 3D facial surface for motion robust facial rPPG estimation. We validate the efficacy of our design decisions and the impact of different video processing steps through an ablation study. Our findings illustrate the potential strengths of exploiting the 3D facial surface as a general strategy for addressing dynamic and unconstrained subject motion in videos. The code is available at https://samcantrill.github.io/orientation-uv-rppg/.

1. Introduction

In camera-based remote photoplethysmography (rPPG) we estimate the rPPG signal using video obtained from consumer-grade cameras. The rPPG signal contains valuable and meaningful physiological information including pulse rate (PR), respiration rate (RR), and pulse rate variability (PRV) [20]. Contactless approaches offer distinct advantages over traditional contact-based methods, conse-

quently finding applications in telehealth [13], in-patient monitoring [1], and various affective computing tasks [7, 12, 41, 43]. Despite significant advancement in facial rPPG estimation, existing methods often display performance degradation in challenging real-world scenarios [6, 13, 34], particularly in handling unconstrained and dynamic subject motion, thus limiting the potential of this technology.

Facial rPPG estimation methods primarily rely on detecting subtle color changes on the face surface caused by the sub-surface blood volume pulse (BVP) to estimate downstream physiological signals such as pulse rate (PR) [4, 30, 37]. However, unconstrained and dynamic subject motion introduces significant variability into the observed appearance of the face in video, with large pixel-level variations potentially overshadowing the subtle changes associated with the rPPG signal.

Continued advancement in facial rPPG estimation has solidified the strengths of deep-learning based methods [3, 13, 21, 32] in capturing the intricate and non-linear relationships between spatio-temporal input features in video and the target rPPG signal. Video-based approaches [14, 39, 42] operate directly on video-formatted data, demonstrating robust spatio-temporal modeling capabilities. However, they are easily influenced by real-world conditions [40] such as subject motion. Spatio-temporal map (STMap)-based methods [16, 17, 21, 23] leverage prior knowledge about the characteristics of the subtle rPPG signal to design hand-crafted 2D spatio-temporal input representations. Consequently, STMap-based methods often exhibit superior performance and robustness over video-based approaches. However, their coarse spatial representation limits the extraction of rich and contextual spatio-temporal features.

Modeling and localizing the spatial region of the face in a video through facial detection represents a strong inductive bias for facial rPPG estimation. It is frequently used to enhance performance, rendering it a standard component of the input processing pipeline [15] across state-of-the-art video-based [42] and STMap-based [16] approaches. A limited number of methods [2, 31] have leveraged facial structure modeling to extract dynamic surface regions instead of fixed spatial regions from video, aiming to improve robustness against subject translation, rotation, and facial expressions. Several studies have noted the importance of considering the 3D facial structure [19, 38] for improving both performance and motion robustness. However, within video-based methods, there is a notable absence of works that exploit facial structure for enhancing the motion robustness of rPPG estimation.

In this work, we investigate how modeling the 3D facial structure can be used as a fundamental strategy for disentangling rigid and non-rigid subject motion from video. UV coordinate texture maps are commonly used to represent the texture of a 3D surface using an image, providing a way to represent the 3D surface of the face in a form compatible with existing video-based facial rPPG estimation methods, as illustrated in Figure 1. Our methodology uses existing 3D landmark detection techniques to model the 3D surface of the face as a mesh. Subsequently, we apply UV coordinate texture mapping to warp the observed facial surface within video frames into a facial texture representation. Given the introduction of geometric distortion during the transformation, we mask the UV coordinate facial texture using the facial surface orientation to remove regions with re-projected and distorted appearance. We demonstrate a significant improvement in generalization performance across all motion scenarios through cross-dataset testing, employing a well-understood baseline method.

Our contributions are summarized as follows:

- We propose a novel UV-coordinate facial texture video representation conditioned on facial surface orientation designed to be compatible with video-based methods which significantly enhances the generalization performance and motion-robustness of facial rPPG estimation.

2. Related Work

Signal Processing Methods: Early efforts in camera-based facial rPPG estimation primarily focused on extracting temporal signals with domain-specific knowledge such as blind source separation (BSS) methods and maximum periodicity criteria [29, 30]. These approaches are restricted to limited motion conditions, under which these assumptions remain valid. Subsequent works sought to improve the robustness of the rPPG signal extraction against sources of noise such as subject motion and illumination by exploiting prior domain knowledge. For instance, leveraging knowledge about skin optical properties [4] or the blood volume pulse dynamics [5] helped isolate a more robust color vector subspace with a higher signal-to-noise ratio. Further works like S2R [36] and POS [37] adopted data-driven approaches to dynamically extract such a sub-space.

Despite the strong priors these methods employ, they suffer from significant performance degradation in the presence of unconstrained and dynamic subject motion, as the underlying assumptions may not hold. As a result, deep-learning methods gained prominence due to their ability to learn the complex and non-linear relationship between the spatio-temporal video features and the rPPG signal.

STMap-based Deep Learning Methods: Spatial-temporal map (STMap)-based methods in facial rPPG estimation rely on hand-crafted spatial-temporal representations of video to exploit prior knowledge about the rPPG signal, aiming to improve the signal-to-noise ratio. Similarly to signal processing methods, they leverage information about the facial position and surface within video alongside prior domain knowledge as strong inductive biases.

Construction of the STMap or multi-scale STMap

(MSTMap) commonly involves extracting a static spatial region of interest from video containing the face [16, 17, 21, 25], dividing it into grid-cells, and processing them into STMap-pixels for each time-step. While this method aids the performance by increasing the signal-to-noise ratio of the rPPG signal in video, it remains susceptible to both subject translation, head rotation and facial expressions. Other approaches extract dynamic 2D facial surface regions of interest [31], providing improved robustness to subject translation and subject rotation, as each STMap pixel represents a temporally consistent region on the face surface.

Despite the advantages of disentangling subject motion through the input representation, it has not been explored whether using the dynamic 3D surface of the face can be used to further improve the robustness to motion. Furthermore, the coarse nature of STMap pixel-regions limits the ability to extract the rich spatio-temporal information present in video [26, 41].

Video-based Deep Learning Methods: Video-based methods differ from STMap-based methods by directly operating on video-formatted data, sequences of spatial frames, employing models with strong spatio-temporal modeling capabilities to learn to extract subtle changes in skin color due to the rPPG signal. The performance and robustness of video-based methods implicitly rely on regularization in the learning process. Despite the strengths of video-based methods they are often outperformed by STMap-based methods which employ more explicit inductive biases for facial rPPG estimation [40].

Given the relevance of the facial surface as the object of analysis for facial rPPG estimation, numerous video-based methods explore different network architectures and functional forms to refine spatial features and improve motion robustness. Some approaches incorporate explicit spatial masking in feature space [3, 13, 26, 28] to guide the model's attention to salient spatial regions. Other works utilize spatial feature refinement modules [9, 10, 24] to allow the model to implicitly learn more informative spatial features. Despite performance improvements, such approaches are still sensitive to various motion scenarios due to a lack of an explicit inductive bias for motion robustness.

Video input representation can also be used to exploit facial and motion features, enhancing motion robustness. DeepPhys [3] proposed the normalized frame-difference as a motion representation to better guide motion estimation, a technique employed in further works [13, 15]. Optical flow representations have also been used to provide motion context to the model, guiding the spatial feature alignment over time to improve the robustness to motion [11]. Furthermore, leveraging facial position through static facial detection is commonly employed in state-of-the-art approaches [42] to increase the signal-to-noise ratio of the rPPG signal in the video.

Using different motion-based and frame-based video representations can provide a strong and explicit inductive bias for motion robustness. However, leveraging the 3D surface of the face to disentangle rigid and non-rigid motion has not been explored as a general mechanism to enhance the robustness of facial rPPG estimation.

3. Method

UV coordinate texture maps are used to represent the texture of a 3D surface using an image as shown in Figure 2. UV coordinate texture mapping can therefore be used to disentangle both rigid and non-rigid facial motion from the observed regions of the face within a video frame. This UV frame representation can be leveraged to reduce motion-related feature variability within a video, thus enhancing the motion robustness of video-based facial rPPG estimation method. Figure 3 outlines the pipeline for constructing the orientation-conditioned facial texture video representation.

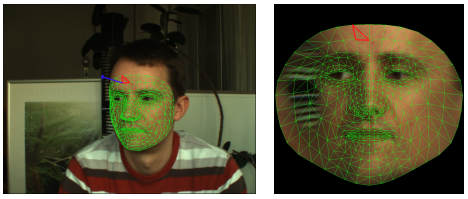

Figure 2. Example from PURE [33] of a XY coordinate image-space frame and the computed UV coordinate texture-space frame with overlaid 3D facial meshes.

3.1. UV Facial Texture Representation

We begin by modeling the 3D geometry of the face within the image-plane using the MediaPipe FaceMesh [18] which detects 468 3D facial landmarks in a non-metrical geometry-space per frame. In this geometry-space, the camera normal is aligned with the negative z-axis, $\vec{n}_{cam} = [0, 0, -1]^T$. These landmarks are represented as a mesh using a pre-defined triangular tessellation scheme [18].

Computing the UV Frame: To compute the facial texture representation, we start by projecting the XYZ coordinate geometry-space facial landmarks onto the XY-plane, yielding the XY coordinate image-space facial landmarks. To perform UV coordinate texture mapping, we compute a series of affine transformations based on the mesh triangles in the XY coordinate image-space, aligning them with the pre-defined and corresponding triangles in UV coordinate texture-space [18]. These affine transformations are then applied to the video frame pixels, converting them from XY coordinate image-space to UV coordinate texture-space. We utilize bi-linear interpolation to compute missing values between transformed points, ensuring smooth transitions. We fill undefined areas with 0, as no points exist

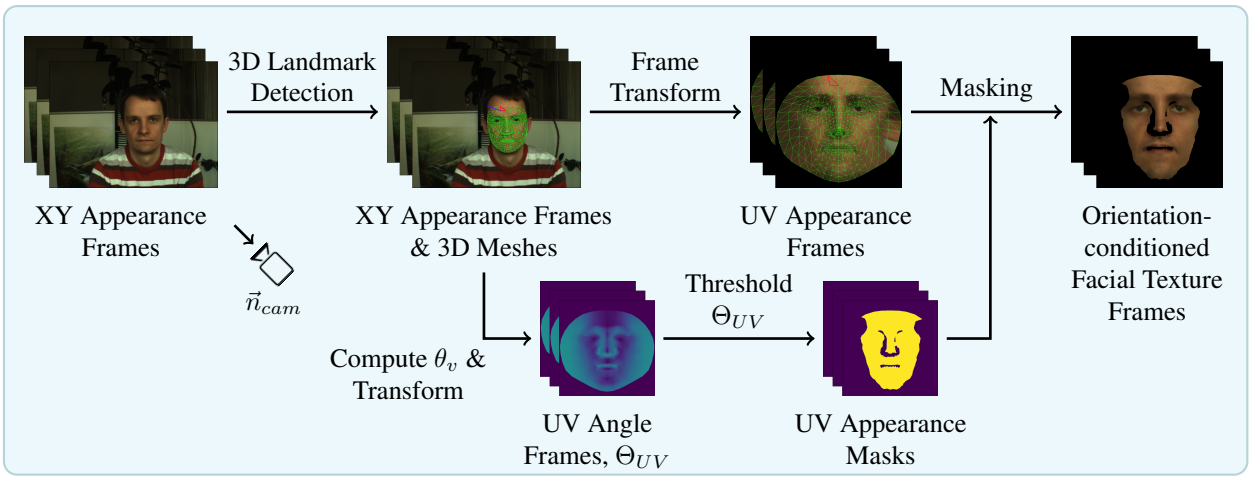

Figure 3. Pipeline for constructing orientation-conditioned facial texture video from input video frames. It leverages a temporally coherent 3D facial mesh [18] to warp the observed XY coordinate facial surface into a pre-defined UV coordinate texture-space [18], followed by masking based on orientation, Θ_{UV}, between the camera and the facial surface to reduce appearance distortion.

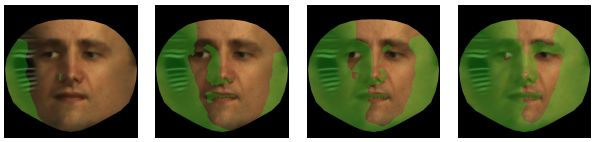

Figure 4. Example of a UV texture-space frame from PURE [33] with facial surface highlighted green based on the relative angle between the surface and the camera of $\Theta_{UV} \geq 90°$, $60°$, $45°$, and $30°$ respectively, to highlight regions with re-projected and/or distorted appearance.

outside of the facial convex hull in UV coordinate texture-space this naturally provides facial segmentation. Thus we obtain the facial texture representation, an example of this is shown in Figure 2. This method of computing the facial texture representation through inherently provides both dynamic localization and extraction of the facial surface.

However, we note two issues that arise due to the facial texture mapping transformation. Firstly, facial surface regions with a normal vector obtuse to $\vec{n}_{cam}$ will be re-projected onto the XY plane and contain duplicate appearance information. Secondly, facial surface regions with a normal vector close to perpendicular to $\vec{n}_{cam}$ will experience significant distortion of the appearance when mapping to UV-space. These issues are highlighted in Figure 4.

Since the aforementioned issues stem from the projection, transformation and interpolation of the UV appearance frame, masking can be applied based on the relative angle between the facial surface and the camera. Hence, we propose conditioning the UV appearance frame based on the relative orientation of the facial surface to the camera, to remove regions with re-projected and distorted appearance.

Masking based on Relative Surface Orientation We begin by computing the relative angles for each mesh vertex, $\theta_v \in [-180°, 180°)$, based on the cosine formula for the dot product between $\vec{n}_v$ and $\vec{n}_{cam}$. We compute $\vec{n}_v$ as the average of the triangle normal's which share that vertex, the triangle normal's are computed from the cross-product of two edge vectors.

Next, we apply the previously defined piece-wise affine transformations to transform the vertices from XY image-space into UV texture-space. Subsequently, spatial bi-linear interpolation of θ_v in the UV texture-space is performed to estimate the relative angle across the frame. This results in the representation of the relative surface angle frame in UV texture-space, denoted Θ_{UV}, as shown in Figure 3.

We mask the UV facial frame based on Θ_{UV} to address the issues of re-projected and distorted appearance. Specifically, regions where $\Theta_{UV} \geq 90°$ are masked to eliminate re-projected appearance. However, masking based on Θ_{UV} can remove frame information. Considering that appearance distortion is greater for $\Theta_{UV} \approx 90°$, based on cross-dataset experimentation we apply masking where $\Theta_{UV} \geq 45°$ to balance between removing information and distortion whilst eliminating re-projection, as illustrated in Figure 5.

Figure 5. Example of a computed UV angle frame Θ_{UV}, the subsequent UV appearance mask for $\Theta_{UV} < 45°$ and resultant masked UV appearance frame to be provided to the video-based model from PURE [33].

4. Experiments

4.1. Experimental Setup

Experiments for rPPG-based estimation of pulse rate (PR) [20] are conducted on two publicly available datasets containing diverse motion scenarios: PURE [33] and MMPD [34]. **PURE** [33] is a small-scale dataset for facial rPPG estimation. It contains instantaneous PPG alongside 60 RGB videos for 10 subjects recorded with diverse head movement scenarios: steady, talking, slow/fast translation and small/medium rotation. **MMPD** [34] is a dataset for facial rPPG estimation under diverse conditions with comprehensive metadata. It contains instantaneous PPG at 30 Hz alongside 660 RGB videos at 30 FPS and 320 $\times$ 240 resolution for 33 subjects recorded with different head/body motions: stationary, rotation, talking and walking.

4.2. Implementation Details

Video-based Deep Learning Model: We adopt PhysNet [39] as our baseline video-based rPPG estimation model to evaluate the efficacy of our designed orientation-conditioned facial texture video representation. Following the implementation and training details outlined in [15], we trained PhysNet [39] with a batch size of 4 for 30 epochs using the Adam optimizer with a OneCycleLR scheduler using a maximum learning rate of 9e-3. We retained the model from the epoch with the lowest validation error for subsequent testing.

Data Preparation: We perform cubic spline interpolation to adjust the sampling rate of the ground-truth rPPG signal to match the corresponding video sequence frame rate. Following the approach in [3, 15], we compute the first-order normalized signal difference as our ground-truth signal. We employ MediaPipe FaceMesh [18] for 3D facial landmark detection with a confidence threshold of 0.45. We linearly temporally interpolate the landmarks for up to three consecutive frames with missing landmarks. We follow our proposed pipeline and compute the orientation-conditioned facial texture video representation with a frame size of 128 $\times$ 128 pixels. We then compute the first-order normalized frame difference [3, 15] followed by pixel outlier-clipping and standardization [3] to serve as our input. We employ standardization per extracted window instead of per video as in [15]. Following [15], we resize the frames to 72 $\times$ 72 pixels to facilitate equitable comparison with previously reported results given the known impact of frame size on model performance from [39]. We refer to PhysNet trained on this video representation as *PhysNet-UV*. We do not employ any data augmentation techniques such as [27], to isolate the focus of our study on the impact of the proposed orientation-conditioned facial texture video representation.

Baseline Data Preparation: To enable comparison with standardized video preparation pipelines, we replace our orientation-conditioned facial texture video pipeline with static facial detection pipeline based on [15] as a baseline. We leverage the previously obtained MediaPipe FaceMesh [18] to derive a static bounding box based on the minimum and maximum bounds of the convex hull of the landmarks projected onto the XY-plane from the 0-th frame of a video sample. We scale the bounding box ($\times$1.5) with a fixed center and crop the video onto the resultant bounding box. We resize the frames to 72 $\times$ 72 pixels and subsequently apply the first-order normalized frame difference, pixel outlier-clipping, and standardization operations as previously outlined. We refer to PhysNet trained on this video representation as *PhysNet-XY*.

4.3. Evaluation

Pulse Rate Estimation: To facilitate comparison with existing methods, we evaluate trained models using the downstream task of pulse rate (PR) estimation through a signal process of the estimated rPPG signal. We begin by estimating the rPPG signal for the entire video, we apply the video-based model across all frames with a video using a sliding non-overlapping window and concatenate the results. This approach maximizes the number of samples and thus the frequency domain resolution from the Fast Fourier Transform (FFT), providing a higher resolution estimate of the PR. Following the standardized procedure outlined in [13, 15] for estimating PR, we detrend the aggregated signal using [35] and apply a 2nd-order Butterworth filter to the rPPG signal, with cut-off frequencies set to [0.75, 2.50] Hz to ensure equitable comparison with previously reported results [15]. This filtering step helps to remove noise and unwanted frequencies, enhancing the quality of the signal. Then, we compute the estimated PR by identifying the dominant frequency within the power spectrum, which is computed using the FFT of the processed rPPG signal. This process is applied to obtain both the predicted and ground-truth rPPG signals per video.

Performance Metrics: Consistent with prior research [15], we report commonly used performance metrics to evaluate model performance on PR estimation. These metrics include the mean absolute error (MAE) measured in Beats Per Minute (BPM), root mean square error (RMSE) (BPM) and Pearson's correlation coefficient (r) of the estimated PR per video across all videos within the test set to provide insight into the error, variance and correlation of the estimated PR. We compute the signal-to-noise ratio (SNR) [4] in decibels (dB) to provide insight into the frequency domain characteristics of the signals. We additionally compute the standard error (SE) to provide a measure of the statistical accuracy of the different estimates in the full results. The reporting for all performance metrics alongside standard errors are provided in the supplementary material in Section 7.2 and Section 7.3. We refer readers to [15] for further details

Method	MAE ↓ (BPM)	RMSE ↓ (BPM)	r ↑
CHROM [15] ★	2.07	9.92	0.99
CHROM [15] ★	5.77	14.93	0.81
2SR [36] ★	2.44	3.06	0.98
POS [37] ★	3.14	10.57	0.95
POS [15] ★	3.67	11.82	0.88
HR-CNN [32] ♦	1.84	2.37	0.98
PhysNet [39] ♦	2.1	2.6	0.99
PhysNet+TFA+PFE [11] ♦	1.44	2.50	-
ETA-rPPGNet [8] ♦	**0.34**	0.77	0.99
Dual-GAN [17] ■	0.82	1.31	0.99
Dual-TL [31] ■	0.36	**0.68**	0.99
rPPG-MAE [16] ■	0.40	0.92	0.99
PhysNet-XY (Excl. S7-T) ♦	0.341	1.108	**0.999**
PhysNet-UV (Excl. S7-T) ♦	0.500	1.397	0.998
PhysNet-XY (Ours) ♦	1.318	7.632	0.945
PhysNet-UV (Ours) ♦	1.639	8.919	0.924
	+24.3%	**+28.7%**	**-2.2%**

★ Signal Processing; ♦ Video-based Deep Learning; ■ STMap-based Deep Learning;

Table 1. Intra-dataset subject-independent performance on PURE [33]. Best results are marked in **bold** and second best in underline.

on these performance metrics. Note we have used publicly available results, codes, and experimental settings for these error metrics and evaluation pipelines.

4.4. Intra-dataset Testing

PR Estimation on PURE: PhysNet-UV demonstrates a significant increase in the error compared to PhysNet-XY in intra-dataset testing on PURE [33]. We strongly emphasize that intra-dataset testing on PURE [33] does not provide meaningful performance differentiation as is elaborated in Section 5, this highlights the importance of cross-dataset testing for meaningful evaluation. We adopt a subject-exclusive 5-fold cross-validation training protocol for PURE [33], the testing results are averaged across the folds to obtain the subject-independent performance. We observe that samples from video *Subject 7 - Talking* (S7-T) exhibit a large PR estimation error (≈58 BPM) due to confounding frequency domain characteristics of the rPPG signal which the PR estimation pipeline fails to handle. We report our results alongside existing works in Table 1 and refer readers to Section 7.2 in the supplementary material for the full results including the standard error, and results excluding *Subject 7 - Talking*.

4.5. Cross-dataset Testing

PR Estimation on MMPD: PhysNet-UV demonstrates a significant 18.2% reduction in the MAE (BPM) compared to PhysNet-XY in cross-dataset testing on MMPD [34], validating the efficacy of our proposed orientation-conditioned facial texture representation for enhancing the performance of facial rPPG estimation methods. We evaluate the generalization performance of the PhysNet-XY and PhysNet-UV models trained on PURE [4] by conducting cross-dataset testing on MMPD [34] across all folds from Section 4.4. We average the results across the folds to obtain the subject-independent performance. We report our results in Table 2 alongside existing results on MMPD [34] reported by [15]. PhysNet-UV outperforms both the previous state-of-the-art and all deep-learning based approaches in terms of both MAE (BPM) and r as compared to other methods also trained on PURE [33]. We refer reads to Section 7.3 in the supplementary material for the full results.

Method	MAE ↓ (BPM)	RMSE ↓ (BPM)	r ↑
CHROM [15] ★	13.66	18.76	0.08
POS [15] ★	12.36	**17.71**	0.18
DeepPhys [15] ♦	16.92	24.61	0.05
PhysNet [15] ♦	13.93	20.29	0.17
TS-CAN [15] ♦	13.93	21.61	0.20
PhysFormer [15] ♦	14.57	20.71	0.15
EfficientPhys-C [15] ♦	14.03	21.62	0.17
PhysNet-XY (Ours) ♦	14.905	22.542	0.155
PhysNet-UV (Ours) ♦	**12.187**	19.849	**0.294**
	-18.2%	**-11.9%**	**+89.8%**

★ Signal Processing; ♦ Video-based Deep Learning; ■ STMap-based Deep Learning;

Table 2. Cross-dataset subject-independent performance on MMPD [34] trained on PURE [33]. Best results are marked in **bold** and second best in underline.

Motion Analysis on MMPD: PhysNet-UV displays significant performance improvements across all tested motion scenarios compared to PhysNet-XY, validating the efficacy of our proposed video representation for enhancing the motion robustness of existing video-based facial rPPG estimation methods. We follow the same process described in Section 4.4 to obtain the subject-independent performance for motion analysis on MMPD [34]. We report the comparison between motion scenarios in Table 3. We demonstrate significant improvements in scenarios with rigid subject motion. We observe a lower performance improvement for *Talking*, which contains a combination of subtle subject rigid and non-rigid head movement. Despite significant improvements, we still observe poor performance for *Walking* which contains significant relative motion between the subject and the camera, highlighting the difficulty of robust PR estimation in the presence of dynamic and unconstrained subject motion.

4.6. Ablation Study

Given the limited insights gained from intra-dataset testing on PURE [33] and the relevance of generalization performance in real-world contexts, we employ cross-dataset test-

Scenario	PhysNet	MAE ↓ (BPM)	RMSE ↓ (BPM)	r ↑
Stationary	XY	7.501	14.364	0.394
	UV	5.887	11.931	0.553
		-21.5%	**-16.9%**	**+40.6%**
Stationary (after exercise)	XY	17.281	26.447	0.113
	UV	15.522	24.298	0.247
		-10.2%	**-8.1%**	**+118.9%**
Rotation	XY	12.251	17.579	0.173
	UV	8.627	14.659	0.344
		-29.6%	**-16.6%**	**+98.0%**
Talking	XY	8.979	14.898	0.381
	UV	8.308	14.519	0.377
		-7.5%	**-2.5%**	**-1.2%**
Walking	XY	28.490	33.202	0.036
	UV	22.565	28.463	0.070
		-20.8%	**-14.3%**	**+96.7%**

Table 3. Cross-dataset subject-independent performance on MMPD [34] of PhysNet-XY and PhysNet-UV trained on PURE [33] across different motion scenarios.

ing to evaluate our ablations. We train on PURE [4] and test on MMPD [34] using the same protocol described in Section 4.5. We denote the sequence of first-order normalized frame difference, pixel outlier clipping, and standardization operations described in Section 4.2 as F_D for brevity. We also denote the UV transformation operation as T_{UV}.

Impact of Θ_{UV}: Masking the facial texture using the surface orientation provides an effective mechanism to mitigate the negative impact of re-projected and distorted information in the UV coordinate frame as a result of the UV transformation process. In this ablation, we vary Θ_{UV} to remove increasing amounts of re-projected and distorted information. We report the results in Table 4 and highlight PhysNet-XY and PhysNet-UV for easier comparison. We demonstrate that balancing the removal of distortion with retaining training information by masking regions with $\Theta_{UV} \geq 45°$ is necessary to optimize the performance of the proposed facial texture video representation.

Impact of Video Processing: The proposed facial texture representation provides improved performance over both dynamic facial localization and/or segmentation, highlighting the efficacy of the UV coordinate mapping process. In this ablation, we vary the video processing pipeline to evaluate the impact of both static and dynamic facial detection and/or segmentation. We derive the bounding box and segmentation mask from the projected XY landmarks to ensure consistency across experiments. We report the results in Table 4 and highlight PhysNet-XY for easier comparison. We refer readers to Sec. 7.3 in the supplementary material for the full results, including additional ablations.

We demonstrate that employing dynamic segmentation degrades the performance over the baseline, in-line with [26]. We also observe that employing dynamic facial detection degrades the performance, we suspect that dynamic facial detection results in jitter in the cropped region, resulting in large pixel-level changes in F_D. However, dynamic facial cropping with segmentation may remove the large pixel-level variations in the background, improving the performance over the baseline. We note the performance impact that subtle changes to steps and operations in the video processing pipeline can have.

5. Discussion

We have proposed to model the 3D surface of the face as a strategy to disentangle rigid and non-rigid subject motion from video, reducing the spatio-temporal feature variability in video due to subject motion. We leveraged the 3D to 2D correspondence of UV coordinate texture mapping to construct video frames which enhance the performance and motion robustness of existing video-based facial rPPG estimation methods. Our method achieves a 18.2% cross-dataset performance improvement using the proposed orientation-conditioned facial texture video representation as demonstrated in Table 2 over our baseline, which represents a commonly employed video processing pipeline. We demonstrated significant generalization performance improvements of up to 29.6% across a diverse range of motion scenarios in Table 3, further validating the efficacy of our proposed video representation to improve motion-robustness. We highlighted the importance of mitigating the effects of re-projected and distorted facial texture in Table 4 through leveraging the surface orientation. We demonstrated the advantages of UV coordinate mapping over both dynamic facial detection and segmentation, and the impact of subtle changes in the video processing pipeline. Our proposed orientation-conditioned facial texture video representation provides a robust and explicit inductive bias for enhancing the motion robustness of existing video-based rPPG methods.

Limitations: Our proposed representation inherently introduces distortion through the UV coordinate texture mapping process. We mitigated this by masking the appearance frame based on the facial surface orientation. We observed performance trade-offs when masking a significant amount of training information, resulting in degraded performance as shown in Table 4. Furthermore, our proposed method explicitly relies on accurate and consistent 3D facial landmark detection to compute the orientation-conditioned facial texture video representation. Within this work we did not explore the impact of different 3D facial reconstruction methods or the effects of noisy landmark detection.

We emphasize the importance of increasing the frequency domain resolution of the FFT for fine-grained PR estimation, this is critical in scenarios where we expect high performance from our models. Our evaluation process provides a frequency domain resolution of ≈0.88 BPM in

Video Processing Pipeline	MAE ↓ (BPM)	RMSE ↓ (BPM)	r ↑	SNR ↑ (dB)
Crop_{Static} (×1.5-Box) + Resize + F_D (**PhysNet-XY**)	14.905	22.542	0.155	-6.882
Crop_{Static} (×1.5-Box) + Segment + Resize + F_D	15.237	23.524	0.120	**-6.053**
$\text{Crop}_{Dynamic}$ (×1.5-Box) + Pad_{Square} + Resize + F_D	17.988	25.183	0.033	<u>-6.263</u>
$\text{Crop}_{Dynamic}$ (×1.5-Box) + Pad_{Square} + Segment + Resize + F_D	14.683	22.563	0.138	-6.553
T_{UV} + F_D + Resize	<u>12.687</u>	<u>20.454</u>	0.248	-6.679
T_{UV} + Mask ($\Theta_{UV} \geq 90°$) + F_D + Resize	13.038	20.900	0.216	-6.473
T_{UV} + Mask ($\Theta_{UV} \geq 60°$) + F_D + Resize	12.890	20.629	0.256	-6.284
T_{UV} + Mask ($\Theta_{UV} \geq 45°$) + F_D + Resize (**PhysNet-UV**)	**12.187**	**19.849**	**0.294**	-6.265
T_{UV} + Mask ($\Theta_{UV} \geq 30°$) + F_D + Resize	13.300	20.834	<u>0.277</u>	-6.496

Table 4. Ablative results with different video processing pipelines, obtained from cross-dataset testing on MMPD [34] of PhysNet trained on PURE [4]. Best results are marked in **bold** and second best in <u>underline</u>.

PURE [33] , indicating the accuracy of the trained model is below the resolution of the evaluation task. Furthermore we observe significant variability in the PR estimation for certain subjects despite similar frequency domain characteristics, highlighting the difficulty in designing robust PR evaluation pipelines. This concern is also highlighted by the cut-off frequencies of the band-pass filtering which correspond to 45-150 BPM, in high-intensity exercise we may reasonably except PR to exceed 150 BPM. These issues significantly limit the meaningful insights and performance differentiation we can obtain.

Future Work: Future directions include exploring alternative ways to exploit the 3D facial structure for more performant and robust facial rPPG estimation. Existing STMap-based [16] approaches commonly extract image regions, we expect these approaches to benefit from leveraging the 3D facial structure to extract temporally consistent surface regions for use in the STMap. Exploring the use of our orientation-conditioned facial texture representation with more recent and powerful video-based architectures such as PhysFormer [42] in this context may similarly provide performance benefits.

Further experimental work leveraging large-scale datasets such as VIPL-HR [22] is necessary to meaningfully evaluate the intra-dataset performance, given the limitations of training expressive deep-learning models on small-scale datasets such as PURE [33]. Further evaluation of rPPG estimation methods in real-world scenarios is also necessary to identify and establish the performance limitations to be addressed moving forwards.

Impact Statement: Subject motion represents a significant challenge in camera-based facial rPPG estimation. Addressing this challenge is crucial to unlocking the potential of camera-based remote physiological measurement, particularly in vital applications like telehealth. However, it is important to acknowledge the risks of these technologies being used in an unethical manner, such as for covert measurement. Therefore, responsible and ethical use and deployment in validated and regulated scenarios is critical for realizing the benefits of this impactful technology.

6. Conclusion

In this paper, we have demonstrated that our proposed orientation-conditioned facial texture video representation improves the performance and motion robustness of existing video-based facial rPPG estimation methods. Using the proposed video representation, PhysNet [39] achieves a substantial 18.2% overall performance improvement compared to our baseline, and demonstrates significant improvements of up to 29.6% in all tested motion scenarios. Despite the limitations associated with the proposed video representation, our results and further investigations Section 4.6 underscore the strength of disentangling subject motion through UV coordinate mapping. More generally, this represents an interesting direction for future research to explore explicitly leveraging the facial structure as a strong inductive bias for more robust facial rPPG estimation in challenging rigid and non-rigid subject motion scenarios. We hope the work and insights demonstrated in this work will contribute towards ensuring reliable and performant facial rPPG estimation in real-world scenarios.

Acknowledgments: This work was supported by the MRFF Rapid Applied Research Translation grant (RARUR000158), CSIRO AI4M Minimising Antimicrobial Resistance Mission, and Australian Government Training Research Program (AGRTP) Scholarship.

Compliance with Ethical Standards: This study was performed in line with the principles of the Declaration of Helsinki. The experimental procedures involving human subjects described in this paper were approved by CSIRO Health and Medical Human Research Ethics Committee (CHMHREC) [ethics protocol 2022_016_LR] and the Australian National University Human Research Ethics Committee (ANU HREC) [ethics protocols 2023/403 and 2023/483].

References

[1] Lonneke AM Aarts, Vincent Jeanne, John P Cleary, C Lieber, J Stuart Nelson, Sidarto Bambang Oetomo, and Wim Verkruysse. Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—a pilot study. *Early Human Development*, 89(12):943–948, 2013. 2

[2] Constantino Alvarez Casado and Miguel Bordallo López. Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces. *IEEE Journal of Biomedical and Health Informatics*, 2023. 2

[3] Weixuan Chen and Daniel McDuff. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In *IEEE/CVF European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 2, 3, 5

[4] 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, 6, 7, 8, 1

[5] 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. 2

[6] Amogh Gudi, Marian Bittner, and Jan van Gemert. Real-Time Webcam Heart-Rate and Variability Estimation with Clean Ground Truth for Evaluation. *Applied Sciences*, 10 (23):8630, 2020. 2

[7] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation. *arXiv preprint arXiv:2010.00400*, 2020. 2

[8] Min Hu, Fei Qian, Dong Guo, Xiaohua Wang, Lei He, and Fuji Ren. ETA-rPPGNet: Effective Time-domain Attention Network for Remote Heart Rate Measurement. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. 6

[9] 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. 3

[10] Bin Li, Panpan Zhang, Jinye Peng, and Hong Fu. Non-contact PPG signal and heart rate estimation with multi-hierarchical convolutional network. *Pattern Recognition*, 139:109421, 2023. 3

[11] Jianwei Li, Zitong Yu, and Jingang Shi. Learning Motion-Robust Remote Photoplethysmography through Arbitrary Resolution Videos. In *AAAI Conference on Artificial Intelligence*, pages 1334–1342, 2023. 3, 6

[12] Bofan Lin, Xiaobai Li, Zitong Yu, and Guoying Zhao. Face Liveness Detection by rPPG Features and Contextual Patch-Based CNN. In *International Conference on Biometric Engineering and Applications*, pages 61–68. Association for Computing Machinery, 2019. 2

[13] 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, 5

[14] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Cardiac Measurement. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4997–5006. IEEE, 2023. 2

[15] 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. 2, 3, 5, 6, 1

[16] Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. rPPG-MAE: Self-supervised Pre-training with Masked Autoencoders for Remote Physiological Measurement. *IEEE Transactions on Multimedia*, 2024. 2, 3, 6, 8

[17] Hao Lu, Hu Han, and S. Kevin Zhou. Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12399–12408, 2021. 2, 3, 6

[18] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3, 4, 5

[19] Yuichiro Maki, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Remote Heart Rate Estimation Based on 3D Facial Landmarks. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2634–2637, 2020. 2

[20] Daniel McDuff. Camera Measurement of Physiological Vital Signs. *ACM Computing Surveys*, 55(9):1–40, 2023. 1, 5

[21] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3580–3585, 2018. 2, 3

[22] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. VIPL-HR: A Multi-modal Database for Pulse Estimation from Less-Constrained Face Video. In *IEEE/CVF Asian Conference on Computer Vision (ACCV)*, pages 562–576. Springer International Publishing, 2019. 8

[23] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. RhythmNet: End-to-end Heart Rate Estimation from Face via Spatial-temporal Representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 2

[24] 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, 2019. 3

[25] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based Remote Physiological Measurement via Cross-verified Feature Disentangling. In *European ConECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 295–310. Springer, 2020. 3

[26] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The Benefit of Distraction: Denoising Camera-Based Physiological Measurements Using Inverse Attention. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4955–4964, 2021. 3, 7

[27] Akshay Paruchuri, Xin Liu, Yulu Pan, Shwetak Patel, Daniel McDuff, and Soumyadip Sengupta. Motion matters: Neural motion transfer for better camera physiological measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5933–5942, 2024. 5

[28] Olga Perepelkina, Mikhail Artemyev, Marina Churikova, and Mikhail Grinenko. HeartTrack: Convolutional Neural Network for Remote Video-Based Heart Rate Monitoring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1163–1171, 2020. 3

[29] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2010. 2

[30] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010. 2

[31] Wei Qian, Dan Guo, Kun Li, Xilan Tian, and Meng Wang. Dual-path TokenLearner for Remote Photoplethysmography-based Physiological Measurement with Facial Videos. *arXiv preprint arXiv:2308.07771*, 2023. 2, 3, 6

[32] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual Heart Rate Estimation with Convolutional Neural Network. In *British Machine Vision Conference (BMVC)*, pages 3–6, 2018. 2, 6

[33] 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, 4, 5, 6, 7, 8, 1, 2

[34] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. MMPD: Multi-Domain Mobile Video Physiology Dataset. *arXiv preprint arXiv:2302.03840*, 2023. 2, 5, 6, 7, 8, 1

[35] Mika P Tarvainen, Perttu O Ranta-Aho, and Pasi A Karjalainen. An advanced detrending method with application to HRV analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002. 5

[36] Wenjin Wang, Sander Stuijk, and Gerard de Haan. A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation. *IEEE Transactions on Biomedical Engineering*, 63(9):1974–1984, 2015. 2, 6

[37] 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, 6

[38] Kwan Long Wong, Jing Wei Chin, Tsz Tai Chan, Ismoil Odinaev, Kristian Suhartono, Kang Tianqu, and Richard H. Y. So. Optimising rPPG Signal Extraction by Exploiting Facial Surface Orientation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2164–2170, 2022. 2

[39] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. In *British Machine Visison Conference (BMVC)*, 2019. 2, 5, 6, 8

[40] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. AutoHR: A Strong End-to-end Baseline for Remote Heart Rate Measurement with Neural Searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020. 2, 3

[41] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. TransRPPG: Remote Photoplethysmography Transformer for 3d Mask Face Presentation Attack Detection. *IEEE Signal Processing Letters*, 28:1290–1294, 2021. 2, 3

[42] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4186–4196, 2022. 2, 3, 8

[43] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, 2016. 2