International Journal of Science and Engineering Applications
Volume 15-Issue 10, 11 – 20, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1510.1003

GridPulseFormer for Remote Heart Rate Estimation with Regional Temporal Consensus and Robust Waveform Reconstruction

Xun Zhang
School of Electronic Information and Electrical Engineering
Yangtze University, Hubei, Jingzhou, 434000, China

Abstract: Remote photoplethysmography estimates pulse signals from subtle color variations in facial videos, but remains sensitive to motion, illumination, and partial occlusion. We propose GridPulseFormer, a video Transformer that combines regional temporal consensus with robust waveform reconstruction. A three-dimensional convolutional stem converts video into a spatial grid that retains the arrangement of facial regions. A divided spatiotemporal Transformer adjusts spatial attention according to the agreement between latent regional trajectories. The output stage combines regional amplitude and first-order difference estimates through differentiable iterative reweighting to reconstruct the pulse waveform. We evaluate the model on PURE, UBFC-rPPG, and MMPD using within-dataset and cross-dataset comparisons, module ablations, and visualizations. In the within-dataset PURE experiment, the full model achieves a mean absolute error of 0.2250 bpm and a root mean square error of 0.3486 bpm. The ablations favor the joint use of the two modules over either module alone. Spectral and spatial visualizations reveal differences in harmonic peak selection and regional weight distributions. GridPulseFormer retains spatial information throughout feature modeling and imposes explicit waveform constraints for remote heart rate estimation.

Keywords: remote photoplethysmography; heart rate estimation; video Transformer; regional temporal consensus; robust waveform reconstruction

1. INTRODUCTION

Heart rate is a basic measure of cardiac activity. Remote photoplethysmography (rPPG) uses a conventional camera to record periodic changes in light reflected from the skin, reducing the need for contact sensors. Its performance, however, depends on illumination, motion, and recording distance [1]. Verkruysse et al. demonstrated the feasibility of measuring pulse signals from video under ambient light [2]. Because pulse-related color variations are much weaker than static skin appearance and motion components, rPPG requires both weak-signal extraction and interference suppression.

Early approaches relied primarily on signal separation and color-space projection. Poh et al. used blind source separation to extract the pulse component [3]. CHROM combines chrominance signals to reduce motion sensitivity, whereas POS derives color projections from a skin reflection model [4][5]. These methods are computationally simple, but their fixed processing rules can be inadequate when several sources of interference occur together.

Deep learning combines video feature extraction with waveform recovery. DeepPhys uses convolutional attention, PhysNet employs a spatiotemporal network, and EfficientPhys focuses on simple, efficient camera-based physiological measurement [6][7][8]. These methods improve learned representations, while leaving room to investigate how regional features should be fused and how the output waveform should be constrained.

Transformers model dependencies over extended feature sequences. The Vision Transformer represents images as token sequences, and TimeSformer separates temporal and spatial attention for video modeling [9][10]. PhysFormer uses a temporal difference Transformer to model pulse features [11]. RhythmFormer introduces periodic sparse attention, while PhysMamba and RhythmMamba explore state space modeling, providing alternative approaches to temporal modeling in rPPG [12][13][14].

Shadows, motion at facial boundaries, and occlusion affect facial regions differently. Collapsing the spatial dimension too early limits the network's ability to distinguish regional contributions, whereas global attention at the original image resolution is computationally expensive. We retain the regional layout on a low-resolution spatial grid and use agreement between latent regional trajectories to guide information exchange. Consensus acts as a soft bias because shared motion and global illumination changes can also produce coherent responses.

Output layers often generate waveforms directly through convolutional or linear mappings, leaving the relationship between regional amplitudes and local temporal changes implicit in the network parameters. Inspired by robust estimation, which adjusts the contribution of each observation according to its residual [15], we first predict regional amplitudes, slopes, and weights, then reconstruct the final waveform explicitly.

GridPulseFormer therefore introduces two components: spatial attention guided by regional temporal consensus and a robust reconstruction head with joint amplitude and slope constraints. The attention component controls information exchange across regions, while the reconstruction head combines regional evidence at the output. We examine their effects through structural ablations, a loss-function comparison, and waveform and spatial visualizations. Three-dimensional convolution, Transformers, and iterative reweighting serve as established building blocks.

2. MATERIALS AND METHODS

2.1 Datasets

We evaluate GridPulseFormer on three public datasets, PURE, UBFC-rPPG, and MMPD, in both within-dataset and cross-

International Journal of Science and Engineering Applications
Volume 15-Issue 10, 11 – 20, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1510.1003

dataset settings. Their differences in recording devices, environmental conditions, and activities allow us to examine heart rate estimation under varied acquisition conditions.

PURE contains videos and reference pulse signals from 10 participants performing six activities. Videos are recorded at 30 Hz, and the activities include remaining still, speaking, translation, and head rotation [16]. For within-dataset experiments, we use participants 1 to 7 for training, participant 8 for validation, and participants 9 to 10 for testing. The three sets are subject disjoint.

We retain a fixed file list after preprocessing and use the same training, validation, and test splits for all ablations.

UBFC-rPPG is a widely used dataset for camera-based pulse measurement; the associated work considers both skin-region selection and pulse extraction [17]. We use it for within-dataset evaluation and bidirectional cross-dataset testing with PURE.

MMPD contains smartphone videos of 33 participants with varied skin tones, motion, and illumination conditions [18]. Our within-dataset experiments use a subset that excludes walking activities, and the reported results apply to this subset.

We use subject-disjoint training, validation, and test sets for within-dataset experiments. In cross-dataset experiments, model weights are selected using the source-domain validation loss. Target-domain labels are not used for model selection.

2.2 Data Preprocessing

We detect and crop the face, enlarge the bounding box by a factor of 1.5, and resize the crop to 128 × 128 pixels. The facial region is fixed throughout each recording, with dynamic detection disabled. This keeps grid locations relatively stable, although it cannot fully compensate for large head movements.

Let I_t and I_{t+1} denote consecutive RGB frames. The difference-normalized input is computed using Eq. (1), where ε_i prevents division by zero. We then normalize the frame differences by their standard deviation over the video and append a zero frame to preserve the temporal length.

$$R_t = \frac{(I_{t+1} - I_t)}{(I_{t+1} + I_t + \varepsilon_i)} \quad (1)$$

The reference blood volume pulse (BVP) is resampled to the video length and standardized. In Eq. (2), b denotes the original label, μ_β and σ_β are its mean and standard deviation, and $\tilde{b}$ is the standardized label. During training, we check that labels are finite and have nonzero variation. The input is difference-normalized, while the target remains a waveform consistent with the output of the reconstruction decoder.

$$\tilde{b}_t = (b_t - \mu_\beta) / \sigma_\beta \quad (2)$$

Each video is divided into nonoverlapping clips of 160 frames, corresponding to approximately 5.33 s at 30 Hz. At test time, predicted and reference waveforms are ordered by recording and clip index, concatenated, and used to estimate heart rate for each recording. We save the preprocessing parameters and file lists with each experiment so that all configurations use the same labels and test samples.

2.3 Model Architecture

2.3.1 Overall Architecture

Figure 1 shows GridPulseFormer: a three-dimensional convolutional stem and grid pooling retain regional features; three divided Transformer blocks model spatiotemporal dependencies; an evidence head predicts regional amplitudes, slopes, and weights; and three reconstruction iterations produce the pulse waveform. The input is $X \in \mathbb{R}^{BxTx3xHxW}$, where B is batch size, T is temporal length, and H and W are spatial dimensions. Omitting the batch dimension, the grid tokens have shape T × 16 × 64.

The network reduces the spatial resolution to a 4 × 4 grid with 16 regional tokens. Each grid cell represents a local feature region rather than a fixed anatomical area. Some cells may include background as a result of cropping or motion.

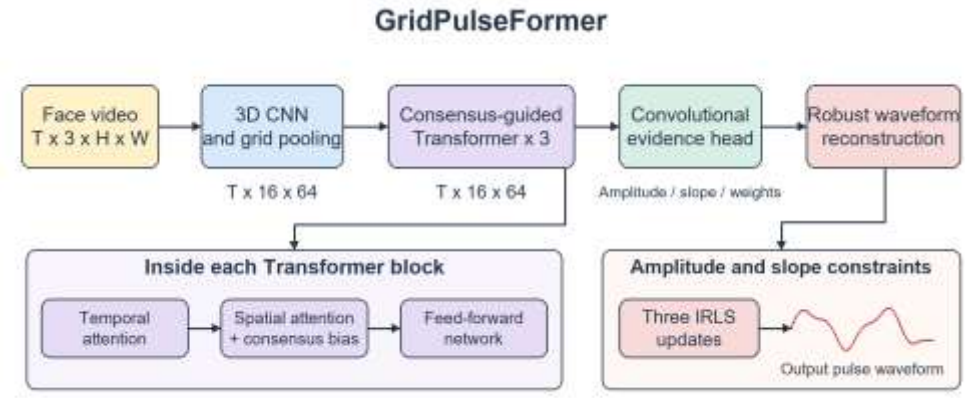

Figure 1. GridPulseFormer architecture: spatial grid feature extraction, spatiotemporal modeling, evidence prediction, and robust waveform reconstruction.

International Journal of Science and Engineering Applications
Volume 15-Issue 10, 11 – 20, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1510.1003

2.3.2 Convolutional Feature Extraction with a Spatial Grid

Figure 2 shows three blocks of three-dimensional convolution, group normalization, and GELU. Kernel sizes are 3 × 5 × 5, 3 × 3 × 3, and 3 × 3 × 3, with spatial stride 2 and temporal stride 1. Channels increase from 3 to 16, 32, and 64; spatial resolution decreases from 128 × 128 to 64 × 64, 32 × 32, and 16 × 16. Adaptive pooling produces a 4 × 4 grid while retaining all 160 frames.

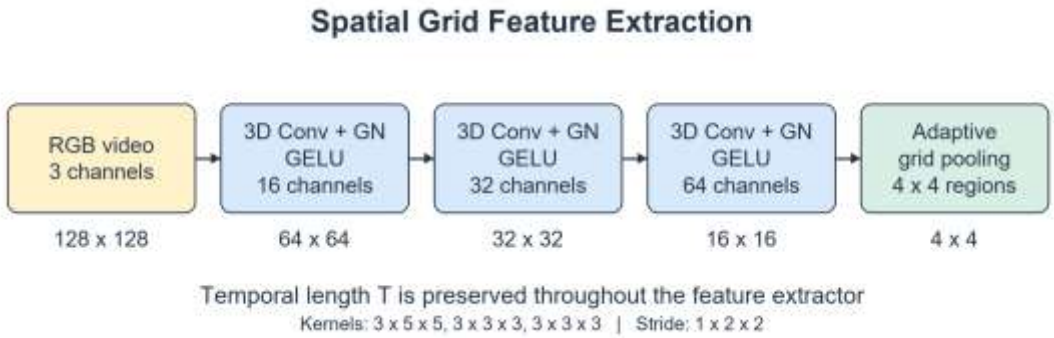

Figure 2. Spatial grid feature extraction. Spatial resolution is reduced while temporal length is preserved. GN denotes group normalization.

Spatial-only downsampling reduces computation while preserving temporal resolution. Group normalization stabilizes feature scales for small training batches.

Adaptive average pooling produces a T × 4 × 4 grid. With N = G² regions and C channels, the features are rearranged into tokens $Z \in \mathbb{R}^{BxTxNxc}$. We add a learnable spatial positional embedding E_s as shown in Eq. (3). Sinusoidal temporal encodings are added only to the queries and keys of temporal attention, leaving the values unchanged.

$$Z_0 = GridPool(Stem(X)) + E_s \quad (3)$$

2.3.3 Divided Spatiotemporal Transformer

Figure 3 shows the divided Transformer block. Temporal attention processes T samples within each region, followed by spatial attention across N regions and a feed-forward network. Each stage uses pre-normalization and a residual connection. Temporal positional encodings enter only the queries and keys; regional consensus biases the spatial keys. We extend divided video attention [10] with consensus-guided regional interaction.

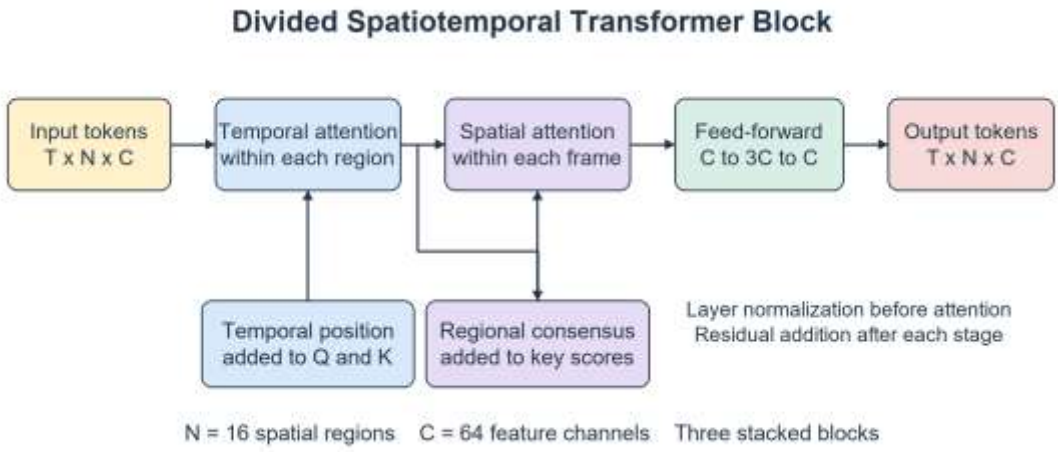

Figure 3. Divided spatiotemporal Transformer with temporal positional encoding, regional consensus bias, and residual connections.

Equation (4) computes temporal attention from normalized Z with positional embedding P added to the queries and keys. We use 4 heads, 64-dimensional features, and 3 Transformer blocks.

$$U = Z + Attention_t(LN(Z) + P, LN(Z) + P, LN(Z)) \quad (4)$$

Spatial attention is then computed across regions at each time step with the consensus bias. The feed-forward network expands the channel dimension to 3C and projects it back to C, using GELU, dropout, and a residual connection. T and N remain unchanged throughout the block. The attention interaction cost is approximately $O(NT^2 + TN^2)$, compared with $O(T^2N^2)$ for joint spatiotemporal attention.

International Journal of Science and Engineering Applications
Volume 15-Issue 10, 11 – 20, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1510.1003

2.3.4 Spatial Attention Guided by Regional Temporal Consensus

Figure 4 shows the consensus computation. A linear probe projects each C-dimensional regional feature in U onto a latent trajectory r_n. Temporal centering and L2 normalization reduce scale differences before pairwise comparison. These trajectories are learned indirectly through the training objective, without regional BVP supervision. A learnable positive coefficient scales the resulting consensus scores before they enter spatial attention.

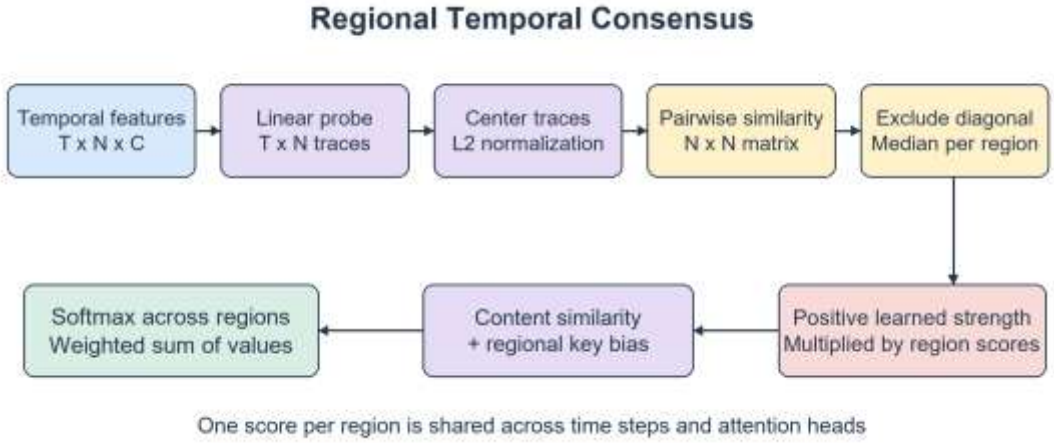

Figure 4. Regional temporal consensus. Median pairwise trajectory similarities produce regional key biases shared across time and attention heads.

$$r_{tn} = w_r^{T} U_{tn}$$

$$q_n = (r_n - mean(r_n)) \,/\, max(\|r_n - mean(r_n)\|_2, \varepsilon) \quad (5)$$

Inner products between regional trajectories yield the agreement scores A_{nm}. After excluding self-similarity, we define the consensus score s_n as the median similarity to the other regions, as shown in Eq. (6). This reduces the influence of a small number of atypical pairwise relationships.

$$A_{nm} = q_n^{T} q_m, \; s_n = median\{A_{nm} : m \neq n\} \quad (6)$$

A nonnegative learnable coefficient α scales the consensus scores to form a bias for the keys in spatial attention. In Eq. (7), i and j index the query and key regions, respectively. The bias is shared across time steps and attention heads and adjusts regional contributions alongside content similarity.

$$P_{ij} = softmax_j(Q_i K_j^{T}/\sqrt{d} + \alpha s_j)$$

$$\alpha = softplus(\theta) \quad (7)$$

The score measures agreement between latent trajectories, not physiological signal quality itself. Shared motion can produce misleading consensus, and phase differences can reduce similarity. We therefore assess the module through ablations and analysis of its outputs.

2.3.5 Regional Amplitude and Slope Evidence

The Transformer output is rearranged into a C × T × G × G grid. A convolution with a temporal kernel size of 3 is followed by a 1 × 1 × 1 convolution that predicts regional amplitudes a, temporal changes u, and confidence scores c. These quantities provide evidence for reconstruction rather than directly defining the final waveform.

Applying softmax to c along the regional dimension produces nonnegative weights w that sum to 1. The initial waveform $y^{(0)}$ is a weighted sum of the regional amplitudes. The slope sequence v aggregates u using the regional weights at the next time step and has length T−1. These operations are defined in Eqs. (8) and (9), respectively.

$$w_{tn} = exp(c_{tn}) \,/\, \Sigma_m exp(c_{tm})$$

$$y_t^{(0)} = \Sigma_n w_{tn} a_{tn} \quad (8)$$

$$v_t = \Sigma_n w_{t+1,n} u_{t+1,n}, \; t = 1, \ldots, T-1 \quad (9)$$

When reconstruction is disabled, the model outputs the weighted waveform in Eq. (8) directly. Registered parameters of inactive modules may remain in the ablation models, so equal total parameter counts do not imply equal computational costs.

2.3.6 Robust Waveform Reconstruction with Amplitude and Slope Constraints

The reconstruction stage seeks a waveform y consistent with both regional amplitude estimates and local slopes. Let D denote the first-order difference matrix and λ a learnable regularization strength. The objective is given in Eq. (10). The amplitude term uses the pseudo-Huber penalty in Eq. (11) to reduce the influence of evidence with large residuals.

$$min_y \Sigma_t \Sigma_n w_{tn} \rho\delta(y_t - a_{tn}) + (\lambda/2)\|Dy - v\|_2^{\,2} \quad (10)$$

$$\rho\delta(e) = \delta^2(\sqrt{(1 + (e/\delta)^2)} - 1) \quad (11)$$

At iteration k, robust coefficients r are computed from the residuals and multiplied by the regional weights. We fix δ at 0.5 and impose a coefficient floor of 0.02, as shown in Eqs. (12) and (13). The objective is solved approximately using a finite number of reweighting steps with a stabilization term.

$$r_{tn}^{(k)} = max\{0.02, [1 + ((y_t^{(k)} - a_{tn})/\delta)^2]\text{^}(-1/2)\} \quad (12)$$

$$\hat{w}_{tn}^{(k)} = w_{tn} r_{tn}^{(k)}$$

$$d_t^{(k)} = \Sigma_n \hat{w}_{tn}^{(k)}$$

$$b_t^{(k)} = \Sigma_n \hat{w}_{tn}^{(k)} a_{tn} \quad (13)$$

With the effective weights fixed, we solve the linear system in Eq. (14). The diagonal stabilization coefficient is $\eta = 10^{-4}$, and softplus ensures that λ is positive. D uses natural endpoint conditions without connecting the beginning and end of a clip periodically. The model performs 3 updates in single precision. Waveform-loss gradients pass through the differentiable solver to the evidence branches.

$$[diag(d^{(k)}) + \lambda D^{T} D + \eta I] y^{(k+1)} = b^{(k)} + \lambda D^{T} v \quad (14)$$

The reconstruction head retains regional amplitudes, weights, slopes, and the initial waveform for inspection of the aggregation process. First-order differences provide a general signal representation; we do not impose ECG-specific QRS detection rules or fixed waveform templates.

3. EXPERIMENTAL SETUP

3.1 Implementation Details

We implement the model in rPPG-Toolbox [19] using Python 3.10.21, PyTorch 2.3.1, and CUDA 12.1 on a single CUDA device. The model has 256,039 parameters, approximately 0.256 M.

We train for 30 epochs with AdamW, a batch size of 4, an initial learning rate of 3×10^{-4}, and weight decay of 0.01. A cosine schedule reduces the learning rate each epoch. After backpropagation, gradients are clipped to a norm of 1.0 to limit unusually large updates. The architecture uses 64-dimensional tokens, 4 attention heads, 3 Transformer blocks, a 4×4 spatial grid, and a dropout rate of 0.1. These settings are held constant across the controlled ablations.

The random seed is fixed at 100. We compute validation loss after each epoch and test the checkpoint with the lowest validation loss. For PURE, the full model and the model with additional losses use checkpoints from zero-indexed epochs 9 and 5, respectively. Configurations, data lists, training logs, and predicted waveforms are saved for each experiment.

3.2 Loss Function

The main model is supervised by a combination of correlation and waveform-shape losses. We first center and normalize the prediction and reference separately using Eq. (15), with $\varepsilon = 10^{-6}$ for numerical stability. Let p and g denote the normalized sequences. Their correlation loss is defined in Eq. (16).

$$N\varepsilon(x) = (x - mean(x)) \,/\, \sqrt{(mean[(x - mean(x))^2] + \varepsilon)} \quad (15)$$

$$Lcorr = 1 - mean(p \odot g) \quad (16)$$

The shape term uses the Smooth L1 loss, which is quadratic for small residuals and linear for large residuals, as defined in Eq. (17). The main objective combines the correlation loss with the shape term weighted by 0.2, as shown in Eq. (18).

$$l(e) = 0.5e^2, \ |e| < 1$$
$$l(e) = |e| - 0.5, \ otherwise \quad (17)$$

$$Lbase = Lcorr + 0.2mean[l(p - g)] \quad (18)$$

For the additional-loss comparison, we introduce autocorrelation and slope supervision. The autocorrelation term compares correlations at matching lags over approximately 0.25 to 1.5 s, subject to the clip length, as defined in Eq. (19).

$$R_x(k) = (T - k)^{-1} \Sigma_{t=1}^{T-k} x_t x_{t+k}$$
$$Lacf = mean_k |R_p(k) - Rg(k)| \quad (19)$$

For the slope term, v is normalized by s_γ, the root mean square amplitude of the centered predicted waveform, and compared with the first-order label differences Dg using Smooth L1 loss. Equation (20) defines the extended objective. Even without the additional slope loss, the main model learns the slope branch through the reconstruction constraint Dy≈v.

$$Lslope = mean[l(v/s_\gamma - Dg)]$$

$$Lextended = Lbase + 0.2Lacf + 0.1Lslope \quad (20)$$

3.3 Evaluation Protocol

Test clips are concatenated in their original order without independently normalizing each clip for the primary evaluation. The concatenated signals undergo smoothness-prior detrending and first-order Butterworth bandpass filtering from 0.75 to 2.5 Hz. Predictions and references receive identical processing.

Heart rate is estimated from the largest peak in the Welch power spectrum within 45 to 150 bpm [20]. Each Welch segment contains at most 256 samples, and the FFT length is 3333, giving a spectral sampling interval of approximately 0.5401 bpm. Zero-padding increases the density of spectral samples but does not improve the independent frequency resolution determined by the observation duration.

Let N be the number of test recordings, and let HR_i^{pred} and HR_i^{gt} denote the predicted and reference heart rates. Mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) are defined in Eqs. (21) to (23). MAE and RMSE are expressed in bpm, and MAPE is expressed as a percentage.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |HR_i^{pred} - HR_i^{gt}| \quad (21)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(HR_i^{pred} - HR_i^{gt})^2} \quad (22)$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N} \left|\frac{HR_i^{pred} - HR_i^{gt}}{HR_i^{gt}}\right| \quad (23)$$

The Pearson correlation coefficient (PCC) is calculated using Eq. (24), where an overbar denotes the mean across recordings. We also assess absolute agreement using Bland–Altman analysis [21]. Differences are defined as prediction minus reference, and the 95% limits of agreement are the mean difference ± 1.96 times the sample standard deviation of the differences.

$$PCC = \frac{\sum_{i=1}^{N}(HR_i^{pred} - \overline{HR}^{pred})(HR_i^{gt} - \overline{HR}^{gt})}{\sqrt{\sum(HR_i^{pred} - \overline{HR}^{pred})^2}\sqrt{\sum(HR_i^{gt} - \overline{HR}^{gt})^2}} \quad (24)$$

SNR is calculated in dB using $10\log_{10}$ of the power ratio in Eq. (25). Within the heart rate evaluation band, power near the reference fundamental frequency and its second harmonic contributes to Psignal, and the remaining power contributes to Pnoise. Because the signal term includes the second harmonic, a high SNR does not guarantee correct heart rate peak selection.

$$SNR = 10log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \quad (25)$$

We evaluate every recording in each predefined test set without excluding recordings according to prediction error. All metrics in Table 4 are computed on the same test set.

4. RESULTS AND DISCUSSION

4.1 Within-dataset Evaluation

Tables 1 and 2 compare within-dataset performance against convolutional networks, Transformers, and state space models using heart rate error, correlation, and SNR. Rankings refer to the displayed values, with equal rounded values treated as ties.

International Journal of Science and Engineering Applications
Volume 15-Issue 10, 11 – 20, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1510.1003

Table 1. Performance on UBFC-rPPG and PURE. Bold indicates the best result for each metric; underlining indicates the second-best result.

Method	UBFC-rPPG			PURE		
	MAE	RMSE	PCC	MAE	RMSE	PCC
DeepPhys [6]	6.27	10.82	0.65	0.83	1.54	0.99
EfficientPhys [8]	1.14	1.81	0.99	1.67	3.28	0.95
PhysFormer [11]	**0.50**	**0.71**	0.99	1.10	1.75	0.99
PhysNet [7]	2.95	3.67	0.97	2.10	2.60	0.99
RhythmFormer [12]	**0.50**	0.78	0.99	0.27	0.47	0.99
RhythmMamba [14]	**0.50**	0.75	0.99	0.23	**0.34**	0.99
PhysMamba [13]	0.54	0.76	0.99	0.25	0.40	0.99
GridPulseFormer	0.54	0.74	0.99	**0.22**	**0.34**	0.99

On UBFC-rPPG, GridPulseFormer achieves an MAE of 0.54 bpm, an RMSE of 0.74 bpm, and a PCC of 0.99 (Table 1). Its RMSE is lower than RhythmMamba's 0.75 bpm but higher than PhysFormer's 0.71 bpm, and its MAE exceeds the best reported value of 0.50 bpm. On PURE, our model achieves an MAE of 0.22 bpm, compared with 0.23 bpm for RhythmMamba, and both obtain an RMSE of 0.34 bpm. These results indicate low heart rate error on PURE and performance close to the leading methods on UBFC-rPPG.

Table 2. Performance on MMPD. Bold indicates the best result for each metric; underlining indicates the second-best result. A dash denotes an unavailable or unreported result.

Method	MMPD				
	MAE	RMSE	MAPE	PCC	SNR
DeepPhys [6]	22.27	28.92	23.90	-0.03	-13.44
EfficientPhys [8]	13.47	21.32	14.19	0.21	-9.25
PhysFormer [11]	11.99	18.41	12.50	0.14	-9.51
PhysNet [7]	4.80	11.80	4.74	0.60	1.77
RhythmFormer [12]	3.07	6.81	3.24	0.86	**5.46**
RhythmMamba [14]	3.16	7.27	—	0.84	—
PhysMamba [13]	—	—	—	—	—
GridPulseFormer	**2.16**	**5.67**	**2.32**	**0.89**	2.27

On MMPD, GridPulseFormer achieves an MAE of 2.16 bpm, an RMSE of 5.67 bpm, a MAPE of 2.32%, and a PCC of 0.89, outperforming the other reported methods on these metrics (Table 2). Relative to RhythmFormer, MAE decreases from 3.07 to 2.16 bpm, a reduction of approximately 29.6%, and RMSE decreases from 6.81 to 5.67 bpm, a reduction of approximately 16.7%. Its SNR of 2.27 dB is lower than RhythmFormer's 5.46 dB. Lower heart rate error therefore does not imply better performance on every spectral-quality metric.

Across the within-dataset comparisons, the improvement varies by dataset. Regional interaction and waveform constraints yield low heart rate errors on PURE and MMPD, although our method does not achieve the best result on every dataset and metric. On MMPD, RMSE remains substantially higher than MAE, indicating large errors in the tail of the distribution that warrant further analysis by motion condition.

4.2 Cross-dataset Evaluation

We evaluate both training on PURE and testing on UBFC-rPPG, and training on UBFC-rPPG and testing on PURE. Checkpoints are selected using source-domain validation data without using target-domain labels. Because the two directions involve different target sets, their error magnitudes should not be used to rank dataset difficulty directly.

Table 3. Training on PURE and testing on UBFC-rPPG. Bold indicates the best result for each metric; underlining indicates the second-best result. A dash denotes an unavailable or unreported result.

Method	UBFC-rPPG				
	MAE	RMSE	MAPE	PCC	SNR
DeepPhys [6]	1.21	2.90	1.42	0.99	1.74
EfficientPhys [8]	2.07	6.32	2.10	0.94	-0.12
PhysFormer [11]	1.44	3.77	1.66	0.98	0.18
PhysNet [7]	0.98	2.48	1.12	0.99	1.49
RhythmFormer [12]	0.89	<u>1.83</u>	1.97	0.99	<u>6.05</u>
RhythmMamba [14]	**0.95**	<u>1.83</u>	**1.04**	0.99	**6.35**
PhysMamba [13]	<u>0.97</u>	1.93	—	0.99	—
GridPulseFormer	0.99	**1.82**	<u>1.08</u>	0.99	2.43

When trained on PURE and tested on UBFC-rPPG, our method achieves an MAE of 0.99 bpm and an RMSE of 1.82 bpm (Table 3). MAE is slightly higher than RhythmFormer's 0.89 bpm. RMSE is the lowest in the table, although the difference from 1.83 bpm is small. MAPE, PCC, and SNR are 1.08%, 0.99, and 2.43 dB, respectively. Our method maintains low heart rate error in this transfer direction, while its spectral SNR remains below those of RhythmFormer and RhythmMamba.

Table 4. Training on UBFC-rPPG and testing on PURE. Bold indicates the best result for each metric; underlining indicates the second-best result. A dash denotes an unavailable or unreported result.

Method	PURE				
	MAE	RMSE	MAPE	PCC	SNR
DeepPhys [6]	5.54	18.51	5.32	0.66	4.40
EfficientPhys [8]	5.47	17.04	5.40	0.71	4.07
PhysFormer [11]	12.92	24.36	23.92	0.47	2.16
PhysNet [7]	8.06	19.71	13.67	0.61	6.68
RhythmFormer [12]	**0.97**	**3.36**	**1.60**	**0.99**	**12.01**
RhythmMamba [14]	1.98	6.51	<u>3.59</u>	0.96	<u>8.74</u>
PhysMamba [13]	<u>1.20</u>	<u>5.99</u>	—	<u>0.97</u>	—
GridPulseFormer	8.16	18.50	16.47	0.72	3.74

When trained on UBFC-rPPG and tested on PURE, GridPulseFormer achieves an MAE of 8.16 bpm, an RMSE of 18.50 bpm, a MAPE of 16.47%, a PCC of 0.72, and an SNR of 3.74 dB (Table 4). The MAE and RMSE are substantially higher than RhythmFormer's 0.97 and 3.36 bpm. The improvements observed within datasets therefore do not transfer fully to this cross-dataset setting.

We examine the error distribution to understand this gap. Of the 59 test recordings, 9 have absolute errors above 35 bpm. Calculations from the aggregate metrics indicate that these recordings have an MAE of approximately 45.78 bpm and account for approximately 93.7% of the total squared error. The remaining recordings have an MAE of 1.38 bpm and an RMSE of 5.05 bpm. A small number of large-error recordings therefore explain most of the difference between the full-set results and the lower-error subset. Squaring the errors further increases their influence on RMSE. We retain all test recordings and assess cross-dataset performance using the full-set metrics.

Generalization differs between the two transfer directions. Differences between source-domain training conditions and target-domain activities, illumination, and facial motion may alter the relationship between learned regional features and the pulse signal. When motion or illumination affects multiple regions simultaneously, consensus may reinforce the interference. Robust reconstruction can reduce inconsistent evidence, but may not correct a systematic bias shared across regions. These mechanisms are plausible explanations based on the architecture; aggregate metrics alone cannot establish the cause. We will examine waveforms, power spectra, and activity conditions for the large-error recordings to distinguish domain shift, interference peaks, and harmonic peak selection.

International Journal of Science and Engineering Applications
Volume 15-Issue 10, 11 – 20, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1510.1003

4.3 Ablation Study with Training and Testing on PURE

The ablation study compares the baseline, consensus-only, reconstruction-only, full, and additional-loss models. The first four use the correlation and shape losses. The fifth adds autocorrelation function (ACF) and slope losses to the full architecture. Disabling reconstruction leaves weighted regional amplitude aggregation as the output, while disabling consensus removes its bias from spatial attention. All other settings are held constant.

The full model enables both modules; the additional-loss variant also uses ACF and slope losses. All variants use the same PURE split.

Table 5. Ablation study on PURE. Consensus and reconstruction denote the two structural modules. Extra losses combine ACF and slope supervision. MAE and RMSE are expressed in bpm. All results use the full test set. Bold indicates the best result. The first four rows are taken from the experiment results workbook, and the last row is taken from the saved test log.

Method	Consensus	Reconstr.	Extra losses	MAE	RMSE	PCC
Baseline	No	No	No	4.0504	13.1012	0.7489
Consensus only	Yes	No	No	4.0504	13.1012	0.7489
Reconstruction only	No	Yes	No	4.0054	13.1003	0.7471
Full model	**Yes**	**Yes**	**No**	**0.2250**	**0.3486**	**0.9998**
Full model + extra losses	Yes	Yes	Yes	4.0054	13.1003	0.7471

The baseline and consensus-only models yield identical heart rate errors, while reconstruction alone slightly reduces MAE. Combining both modules reduces full-set MAE and RMSE to 0.2250 and 0.3486 bpm, respectively, and reduces the number of recordings with errors above 35 bpm from 1 to 0. This supports a complementary effect under the same experimental settings. Equal heart rate metrics do not require identical waveforms because the evaluation selects discrete spectral peaks.

Adding ACF and slope losses increases MAE and RMSE to 4.0054 and 13.1003 bpm. Most of this difference arises from one recording with a harmonic-selection error. Its reference heart rate is 45.3645 bpm, which the full model also predicts, whereas the additional-loss model predicts 90.7291 bpm. We therefore use the simpler main objective. This joint comparison does not isolate the individual effects of the two additional losses.

4.4 Visualization

4.4.1 Waveform and Spectral Comparison

Figure 5 compares the full model and the additional-loss model on the recording with the largest error. The upper panel shows the first 10 s of the waveforms; the lower panel shows Welch spectra computed over the complete 64 s recording. All three waveforms undergo the same detrending and bandpass filtering and are standardized separately, without additional phase alignment. We use this recording to diagnose the error rather than to represent the entire test set.

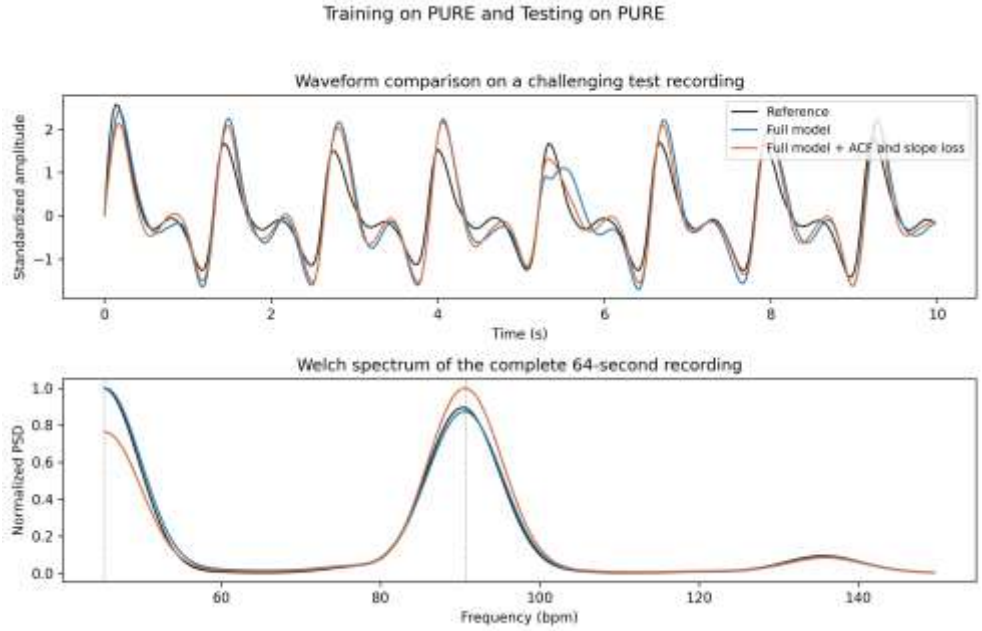

Figure 5. Waveforms and spectra for a harmonic-error case. The upper panel shows the first 10 s; the lower panel uses the full recording. Dashed lines indicate approximately 45.36 and 90.73 bpm.

Both outputs are periodic, but the lower-frequency peak dominates in the full model and reference signal. The additional-loss model has a stronger second-harmonic peak near 90.73 bpm, resulting in a doubled heart rate estimate. The reference peak at 45.36 bpm lies close to the 45 bpm lower search boundary, so this interpretation is specific to the evaluation band. Because SNR includes power near the second harmonic in its signal term, error metrics must be considered alongside spectral peak locations.

4.4.2 Heart Rate Agreement

Figure 6 presents the heart rate scatter plot and Bland–Altman plot for all 12 test recordings evaluated with the full model. The mean prediction-minus-reference difference is 0.0450 bpm, the 95% limits of agreement range from −0.6627 to 0.7527 bpm, and PCC is 0.9998. Original coordinates are retained without jittering overlapping points.

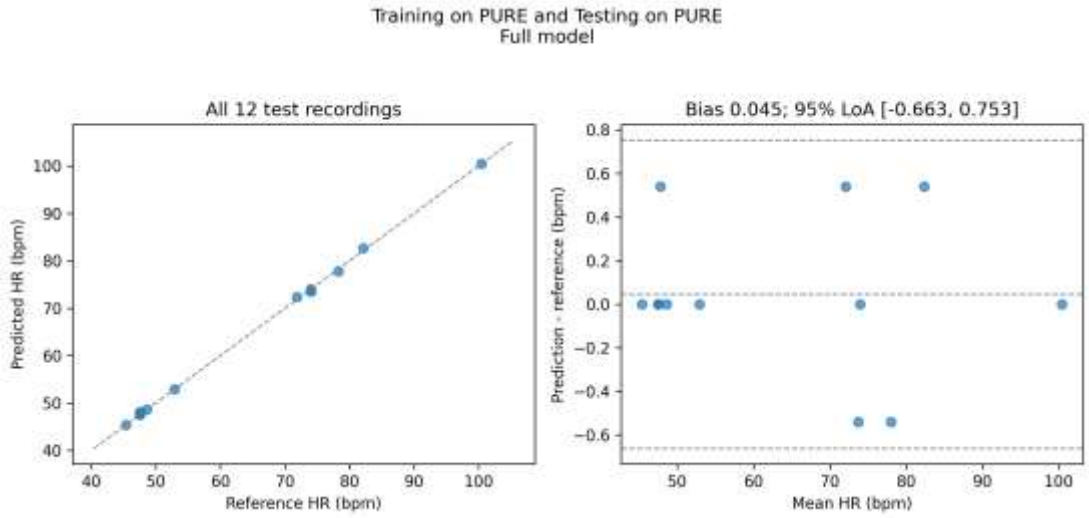

Figure 6. Heart-rate agreement for the full model. The left panel shows predicted versus reference heart rates; the right panel shows Bland–Altman differences, mean bias, and 95% limits of agreement. Overlapping points are retained.

The small mean difference indicates no pronounced overall overestimation or underestimation in this test. However, the recordings come from two participants and cannot establish agreement across broader populations. PCC measures the relationship between recording-level heart rates, not sample-by-sample waveform correlation.

4.4.3 Spatial Grids and Regional Weights

Figure 7 compares spatial representations from the first exported clip shared by the two models. Each clip contains 160 frames and uses a 4 × 4 grid. The three columns show feature energy, temporally averaged regional weights, and final-layer consensus scores. Corresponding maps use the same color scale. This clip differs from the recording in Figure 5: the two visualizations examine regional representations and spectral peak selection, respectively.

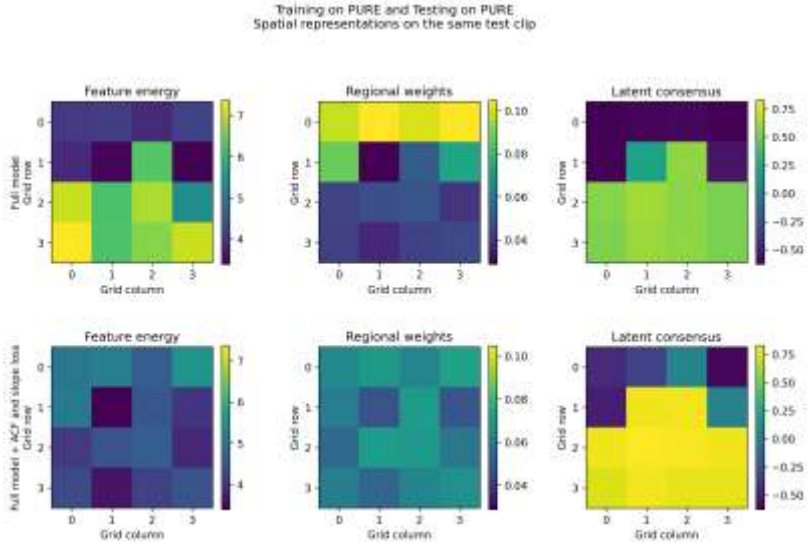

Figure 7. Spatial grid visualization. The upper and lower rows show the full model and the model with extra losses. Columns show feature energy, temporal mean regional weights, and final-layer consensus scores. Corresponding maps share color scales.

Nonuniform grid responses show that the model retains spatial differences. Consensus scores regulate information exchange within the Transformer, whereas regional weights control output evidence aggregation, so their distributions need not coincide. The maps depict latent activations and weights. They are not skin segmentation masks or ground-truth measures of physiological signal quality, and they do not establish a causal contribution for any region on their own.

5. LIMITATIONS

Our ablations use a fixed random seed and a fixed PURE split. Repeated experiments are needed to assess sensitivity to initialization and data partitioning. Large tail errors remain when training on UBFC-rPPG and testing on the full PURE test set, indicating limited tolerance to interference across domains.

The evaluation band is restricted to 45 to 150 bpm and does not cover irregular rhythms or a wider heart rate range. Shared motion or illumination changes can contaminate regional consensus, and the robust threshold in finite-step reconstruction depends on the scale of internal amplitude estimates. The spatial maps are low-resolution latent representations. Further work should include independent datasets, targeted analysis of difficult clips, and measurements of practical runtime efficiency.

6. CONCLUSIONS

We present GridPulseFormer, which retains a spatial grid through three-dimensional convolution, introduces regional temporal consensus into a divided video Transformer, and reconstructs pulse waveforms under joint amplitude and slope constraints. The model achieves low heart rate errors in within-dataset comparisons, and the PURE ablations support using the two modules together. Cross-dataset testing shows low error when training on PURE and testing on UBFC-rPPG, but large-error recordings remain a limitation in the reverse direction. Waveform, spectral, and spatial visualizations provide insight into harmonic peak selection and regional information allocation. Future work will address interference in difficult cross-dataset cases and improve reconstruction efficiency.

7. REFERENCES

[1] Sun, Y., and Thakor, N. Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging. IEEE Transactions on Biomedical Engineering, 63, 463–477, 2016. https://doi.org/10.1109/TBME.2015.2476337

[2] Verkruysse, W., Svaasand, L. O., and Nelson, J. S. Remote plethysmographic imaging using ambient light. Optics Express, 16, 21434–21445, 2008. https://doi.org/10.1364/OE.16.021434

[3] Poh, M.-Z., McDuff, D. J., and Picard, R. W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics Express, 18, 10762–10774, 2010. https://doi.org/10.1364/OE.18.010762

[4] de Haan, G., and Jeanne, V. Robust Pulse Rate From Chrominance-Based rPPG. IEEE Transactions on Biomedical Engineering, 60, 2878–2886, 2013. https://doi.org/10.1109/TBME.2013.2266196

[5] Wang, W., den Brinker, A. C., Stuijk, S., and de Haan, G. Algorithmic Principles of Remote PPG. IEEE Transactions on Biomedical Engineering, 64, 1479–1491, 2017. https://doi.org/10.1109/TBME.2016.2609282

[6] Chen, W., and McDuff, D. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. European Conference on Computer Vision, 2018. https://doi.org/10.1007/978-3-030-01216-8_22

[7] Yu, Z., Li, X., and Zhao, G. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. British Machine Vision Conference, 2019. https://arxiv.org/abs/1905.02419

[8] Liu, X., Hill, B., Jiang, Z., Patel, S., and McDuff, D. EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Cardiac Measurement. IEEE/CVF Winter Conference on Applications of Computer Vision, 2023. https://doi.org/10.1109/WACV56688.2023.00498

[9] Dosovitskiy, A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations, 2021. https://arxiv.org/abs/2010.11929

[10] Bertasius, G., Wang, H., and Torresani, L. Is Space-Time Attention All You Need for Video Understanding? International Conference on Machine Learning, 139, 813–824, 2021. https://proceedings.mlr.press/v139/bertasius21a.html

[11] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. Torr, G. Zhao, PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer, in: 2022 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR, IEEE, New Orleans, LA, USA, 2022: pp. 4176–4186. https://doi.org/10.1109/CVPR52688.2022.00415.

[12] B. Zou, Z. Guo, J. Chen, J. Zhuo, W. Huang, H. Ma, RhythmFormer: Extracting Patterned rPPG Signals based on Periodic Sparse Attention, (2025). https://doi.org/10.48550/arXiv.2402.12788.

[13] C. Luo, Y. Xie, Z. Yu, PhysMamba: Efficient Remote Physiological Measurement with SlowFast Temporal Difference Mamba, (2024). https://doi.org/10.48550/arXiv.2409.12031.

[14] B. Zou, Z. Guo, X. Hu, H. Ma, RhythmMamba: Fast, Lightweight, and Accurate Remote Physiological Measurement, in: 2025: pp. 11077–11085. https://doi.org/10.1609/aaai.v39i10.33204.

[15] Huber, P. J. Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35, 73–101, 1964. https://doi.org/10.1214/aoms/1177703732

[16] Stricker, R., Müller, S., and Gross, H.-M. Non-contact Video-based Pulse Rate Measurement on a Mobile Service Robot. IEEE International Symposium on Robot and Human Interactive Communication, 1056–1062, 2014. https://doi.org/10.1109/ROMAN.2014.6926392

[17] Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., and Dubois, J. Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters, 124, 82–90, 2019. https://doi.org/10.1016/j.patrec.2017.10.017

[18] Tang, J., Chen, K., Wang, Y., Shi, Y., Patel, S., McDuff, D., and Liu, X. MMPD: Multi-Domain Mobile Video Physiology Dataset. IEEE Engineering in Medicine and Biology Society Conference, 2023. https://arxiv.org/abs/2302.03840

[19] Liu, X., et al. rPPG-Toolbox: Deep Remote PPG Toolbox. Advances in Neural Information Processing Systems, Datasets and Benchmarks Track, 2023. https://arxiv.org/abs/2210.00716

[20] Welch, P. D. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. IEEE Transactions on Audio and Electroacoustics, 15, 70–73, 1967. https://doi.org/10.1109/TAU.1967.1161901

[21] Bland, J. M., and Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327, 307–310, 1986.