BEAT: Bottleneck-Enhanced Adaptive Temporal Learning for Heart Rate Estimation from Egocentric NIR Eye Videos

Juha Park[1] and Sang Jun Lee[1] *

Division of Electronic Engineering, Jeonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju 54896, Republic of Korea
juhapark@jbnu.ac.kr, sj.lee@jbnu.ac.kr

Abstract. Remote photoplethysmography (rPPG) from wearable near-infrared (NIR) eye-camera videos enables contact-free heart rate (HR) monitoring. Reliable estimation remains challenging because these inward-facing videos capture only a small periocular skin region and are strongly affected by blinks, gaze shifts, and headset motion. Existing methods mainly rely on global or frame-wise motion cues, leaving spatially varying temporal corruption insufficiently addressed. We present Bottleneck-Enhanced Adaptive Temporal (BEAT) Learning, which combines location-wise temporal refinement with pulse-domain phase-progression supervision. A lightweight residual depthwise temporal convolutional network (TCN) refines each bottleneck trajectory before global spatial pooling while adding only 18,560 parameters. We further introduce Multi-Lag Pulse-Phase Progression Loss (MPPL), which aligns local phase increments across multiple lags in a band-limited pulse representation and downweights unreliable low-amplitude intervals. On egoPPG-DB, BEAT achieves the best reported performance, with 7.20 BPM MAE, 9.68 BPM RMSE, 8.65 % MAPE, and $r = 0.88$.

Keywords: Remote photoplethysmography · Heart rate estimation · Near-infrared imaging · Wearable eye cameras · Phase supervision

1 Introduction

Remote photoplethysmography (rPPG) estimates cardiac activity from subtle visual changes caused by blood-volume variation [6, 18]. It enables contact-free heart rate (HR) and pulse waveform estimation using ordinary video sensors. Recent CNN- and transformer-based methods have improved rPPG estimation by learning spatial and temporal patterns from facial videos [28, 30, 32]. Most of these methods, however, assume that a broad facial skin region is visible under a relatively stable viewing condition. Such assumptions do not hold for inward-facing near-infrared (NIR) eye cameras, which capture only a small periocular region. This gap motivates rPPG methods designed specifically for restricted and highly dynamic wearable views.

* Corresponding author.

Wearable NIR eye cameras provide a practical sensing configuration but introduce several challenges for pulse estimation. The visible periocular region contains substantially less skin than a frontal facial image. Blinks, gaze shifts, eyelashes, and local occlusions can produce image changes that are much stronger than pulse-related variations. Headset motion further introduces blur and frame-to-frame displacement throughout the recording. These disturbances are not uniformly distributed across the image because some regions may remain stable while others are strongly corrupted. Reliable wearable rPPG therefore requires preserving weak pulse periodicity while handling spatially varying temporal corruption.

Synchronized inertial measurements provide useful information about headset motion. Prior egoPPG research combines NIR eye-camera video with headset IMU signals to estimate continuous pulse waveforms and HR under wearable conditions [1]. Frame-wise motion conditioning can adapt visual features according to the motion observed at each time step. However, a single frame-wise coefficient applies the same modulation to all spatial locations within that frame. Spatial attention can emphasize informative regions, but the temporal trajectories of different locations are eventually merged through global spatial pooling. Once stable and corrupted features are pooled into a shared representation, their distinct temporal behavior becomes difficult to recover. This observation motivates refining location-specific temporal features before global spatial pooling.

A second challenge arises from the difference between the model output and the signal representation used for HR estimation. The model directly predicts differential PPG, whereas HR is computed from a reconstructed pulse waveform [19]. Waveform supervision aligns the temporal shape of the differential signal. Spectral supervision further encourages the prediction to contain a physiologically valid periodic structure. Nevertheless, these objectives do not directly constrain how the pulse phase progresses over short time intervals. Signals with similar magnitude spectra can still have different pulse timing, inter-beat intervals, and accumulated phase drift. This limitation motivates supervision in the reconstructed pulse domain, where local pulse progression can be measured more directly.

To address these challenges, we present Bottleneck-Enhanced Adaptive Temporal (BEAT) Learning, which combines location-wise temporal refinement with multi-lag phase-progression supervision. At the feature level, the shared temporal convolutional network (TCN) adapter refines each bottleneck spatial location before global spatial pooling. The adapter preserves the original spatial correspondence and does not introduce additional interaction across spatial locations. At the signal level, the proposed MPPL reconstructs a band-limited pulse representation and compares relative phase increments over multiple temporal lags. Target-envelope confidence reduces the influence of intervals in which analytic phase estimation is unreliable. Together, the two components address location-dependent temporal corruption in the feature space and local pulse-timing errors in the reconstructed signal space.

Our contributions are summarized as follows:

- We introduce a TCN adapter that independently refines the temporal trajectory of each bottleneck spatial location before global spatial pooling. This design preserves spatial correspondence while adding only 18,560 parameters.
- We introduce MPPL to align local phase increments over multiple temporal lags in the reconstructed pulse domain. The proposed objective uses target-envelope confidence to reduce unstable phase supervision in low-amplitude intervals.
- BEAT achieves the best reported performance on egoPPG-DB, with an MAE of 7.20 BPM, an RMSE of 9.68 BPM, a MAPE of 8.65 %, and $r = 0.88$, obtaining better reported values than PulseFormer across all evaluation metrics.

2 Related Work

2.1 Motion-Robust Temporal Modeling for Wearable rPPG

Conventional rPPG methods estimate cardiac signals by modeling subtle spatiotemporal variations in facial videos. Early approaches improved robustness to realistic motion and illumination changes through adaptive filtering, local patch selection, and dynamic facial-region selection [10, 12, 25]. DeepPhys and TS-CAN subsequently learned appearance and motion representations using convolutional attention mechanisms [3, 15]. PhysNet instead estimates continuous pulse waveforms using spatiotemporal convolutions [28]. Transformer-based methods such as PhysFormer and RhythmFormer model longer-range temporal dependencies and periodic physiological structure [30, 32]. Other studies address practical deployment issues, including video compression and computational efficiency [16, 29]. Meta-learning and domain-generalization approaches further improve adaptation to unseen recording conditions [11, 17].

However, these methods are mainly designed for frontal facial videos with broad skin visibility and relatively stable viewing conditions. Wearable NIR eye cameras instead capture only a small periocular region and are strongly affected by blinks, gaze shifts, local occlusions, and headset motion. PulseFormer introduced the egoPPG task and combines NIR eye-camera video with headset IMU cues to estimate continuous pulse waveforms and HR under wearable conditions [1]. Its frame-wise motion conditioning captures global motion context, while spatial attention emphasizes informative image regions. Nevertheless, temporal features from different spatial locations are eventually merged through global spatial pooling. As a result, stable and corrupted spatial trajectories become entangled before their distinct temporal behavior is fully modeled. In contrast, BEAT independently refines each bottleneck spatial trajectory before global spatial pooling, preserving location-specific temporal information until spatial aggregation.

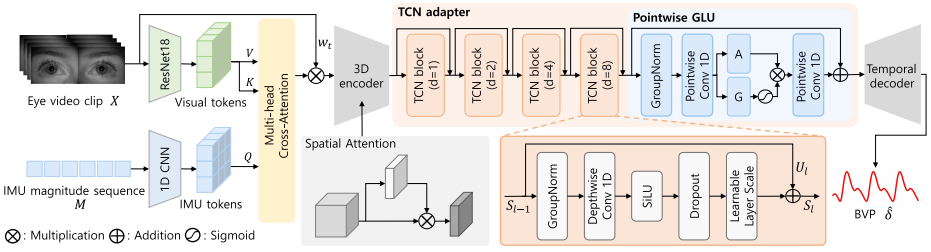

Fig. 1: Overview of the proposed BEAT framework. Given an NIR eye-camera sequence and synchronized IMU magnitudes, the encoder produces a spatially structured bottleneck representation. The TCN adapter independently refines each spatial trajectory before global spatial pooling, and the decoder predicts a differential PPG waveform. During training, MPPL aligns multi-lag phase progression in the reconstructed band-limited pulse representation.

2.2 Physiological Supervision for rPPG

rPPG models commonly combine time-domain waveform losses with frequency-domain objectives. Waveform losses preserve temporal correspondence between the prediction and target, whereas spectral losses encourage physiologically plausible periodicity. However, magnitude-spectrum supervision does not preserve temporal order and therefore cannot directly constrain local pulse timing. Contrast-Phys learns physiological representations from spatiotemporal agreement without synchronized contact-sensor labels [23]. Non-contrastive unsupervised learning instead exploits spectral sparsity and variation across samples to discover periodic physiological signals [20]. Pseudo-label and self-supervised strategies further combine signal-processing priors with learned representations to reduce sensitivity to irrelevant periodic patterns [13]. These approaches primarily constrain global periodicity or consistency across complete signal segments.

Several studies instead address uncertainty in temporal alignment between facial videos and physiological labels. TALOS compensates for temporal offsets during training to reduce sensitivity to global label shifts [4]. BYHE uses periodic self-similarity representations to reduce the need for explicit waveform alignment [21]. DOHA-HPS similarly transforms physiological signals into temporal self-similarity maps to alleviate phase-delay conflicts across instances and datasets [22]. Although these approaches reduce the influence of uncertain global or instance-level delays, they do not directly supervise time-varying local phase progression. This distinction is important because signals with similar global alignment or spectral content can still exhibit different local pulse timing and accumulated phase drift. In contrast, MPPL reconstructs a band-limited pulse representation and compares relative analytic-phase increments over multiple short temporal lags. It therefore complements waveform and spectral supervision by constraining local pulse-phase progression rather than global alignment or all-pairs temporal similarity.

3 Method

The overall architecture is illustrated in Figure 1. Given a single-channel NIR eye-camera sequence and synchronized headset IMU magnitudes, BEAT predicts a 128-sample differential PPG waveform $\hat{\delta}$. We retain the IMU-conditioned visual encoder and temporal decoder of PulseFormer [1] and insert the proposed adapter at the spatially structured bottleneck. The IMU–visual cross-attention module first generates frame-wise modulation coefficients that condition the visual features on headset motion. A 3D CNN backbone with spatial attention modules then emphasizes informative regions and encodes the modulated sequence into a compact feature grid while preserving its spatial structure. Before global spatial pooling, the proposed TCN adapter independently refines the temporal features at each spatial location. The refined representation is subsequently decoded into the differential PPG waveform $\hat{\delta}$. The proposed MPPL further reconstructs a band-limited pulse representation and supervises its local phase progression.

3.1 IMU-Conditioned Visual Feature Extraction

Let $X \in \mathbb{R}^{B\times 1\times T\times H\times W}$ denote a single-channel NIR eye-camera sequence and $M \in \mathbb{R}^{B\times T}$ denote the synchronized IMU magnitudes. Here, B, T, H, and W denote the batch size, temporal length, image height, and image width, respectively. The visual and IMU encoders transform the two modalities into visual tokens and IMU tokens. We perform cross-attention using IMU tokens as the query and visual tokens as the key and value. The attended representation is projected into frame-wise modulation coefficients and applied to the visual stream. We denote the resulting motion-conditioned visual sequence by X'.

The modulated sequence is processed by a PulseFormer-derived 3D CNN backbone equipped with spatial attention modules. The spatial attention modules reweight spatial responses, while the 3D convolutions capture local spatiotemporal patterns. The encoder E_{SA} reduces the temporal resolution from 128 to 32 and produces

$$F_0 = E_{\mathrm{SA}}(X') \in \mathbb{R}^{B\times 64\times 32\times 3\times 8}. \tag{1}$$

The bottleneck feature F_0 is passed to the proposed TCN adapter before global spatial pooling.

3.2 Location-Wise Temporal Refinement

The bottleneck representation $F_0 \in \mathbb{R}^{B\times 64\times 32\times 3\times 8}$ retains a spatial grid of 3×8 locations. We fold the spatial dimensions into the batch dimension and treat each location as an independent temporal sequence:

$$F_0 \longrightarrow S_0 \in \mathbb{R}^{24B\times 64\times 32}. \tag{2}$$

A shared TCN adapter independently refines all 24 spatial trajectories before global spatial pooling.

The adapter consists of four non-causal residual depthwise temporal convolution blocks:

$$\begin{aligned} Z_\ell &= \mathrm{GN}(S_{\ell-1}), \\ U_\ell &= \mathrm{Dropout}\left[\mathrm{SiLU}\left(\mathcal{D}_{d_\ell}(Z_\ell)\right)\right], \\ S_\ell &= S_{\ell-1} + \gamma_\ell \odot U_\ell, \qquad \ell = 1, \ldots, 4. \end{aligned} \tag{3}$$

Here, GN denotes group normalization [26]. And $\mathcal{D}_{d_\ell}$ denotes a depthwise temporal convolution with kernel size 3 and dilation d_ℓ, γ_ℓ is a learnable channel-wise layer-scale parameter, and $\odot$ denotes element-wise multiplication. We use $d_\ell \in \{1, 2, 4, 8\}$, giving the temporal receptive field

$$R = 1 + \sum_{\ell=1}^{4} 2d_\ell = 31. \tag{4}$$

After the final TCN block, we apply a pointwise gated linear unit (GLU) [5] to adaptively modulate the refined channels at each temporal position:

$$\begin{aligned} (V, G) &= \mathrm{split}\left[\Phi_{\mathrm{in}}\left(\mathrm{GN}(S_4)\right)\right], \\ Q &= \mathrm{GLU}_{\mathrm{pt}}(S_4) = \Phi_{\mathrm{glu}}\left[V \odot \mathrm{sigmoid}(G)\right], \\ S &= S_0 + \Phi_{\mathrm{out}}(Q). \end{aligned} \tag{5}$$

Here, split divides the projected channels into the feature and gate tensors V and G, respectively. The operators Φ_{in}, Φ_{glu}, and Φ_{out} are pointwise linear projections applied independently at each temporal position. The output projection is initialized to zero, making the adapter an identity mapping at initialization. Finally, S is restored to $F \in \mathbb{R}^{B\times 64\times 32\times 3\times 8}$ according to the original spatial ordering.

3.3 Loss Function

Let $\hat{\delta}, \delta \in \mathbb{R}^{B\times T}$ denote the predicted and target differential PPG waveforms. For notational convenience, we define $\delta^{\mathrm{pred}} = \hat{\delta}, \delta^{\mathrm{gt}} = \delta$. The complete objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{wave}} + \lambda_{\mathrm{spec}}\mathcal{L}_{\mathrm{spec}} + \lambda_{\mathrm{MPPL}}\mathcal{L}_{\mathrm{MPPL}}. \tag{6}$$

We set $\lambda_{\mathrm{spec}} = 0.1$ whenever $\mathcal{L}_{\mathrm{spec}}$ is enabled. The waveform and spectral terms establish waveform correspondence and physiological periodicity, while MPPL supervises local pulse-phase progression.

Waveform Loss. Following PulseFormer [1], we use the same batch-global z-normalized MSE waveform loss, with the standard-deviation denominator lower-bounded by ε_0 for numerical stability. The waveform loss preserves temporal

correspondence between the predicted and target differential PPG waveforms. For $u \in \{\text{pred}, \text{gt}\}$, we define the batch-global normalization

$$\mathcal{N}_{BT}(\delta^u) = \frac{\delta^u - \mu_{BT}(\delta^u)}{\max\left(\sigma_{BT}(\delta^u), \varepsilon_0\right)}, \tag{7}$$

where $\mu_{BT}(\delta^u)$ and $\sigma_{BT}(\delta^u)$ denote the mean and standard deviation over all batch and temporal elements, respectively, and $\varepsilon_0 > 0$ is a small constant.

The waveform loss is

$$\mathcal{L}_{\text{wave}} = \frac{1}{BT} \left\| \mathcal{N}_{BT}(\delta^{\text{pred}}) - \mathcal{N}_{BT}(\delta^{\text{gt}}) \right\|_F^2 . \tag{8}$$

Physiological Spectral Loss. The spectral loss encourages the predicted waveform to preserve physiologically plausible periodicity. For each sample b and u, we define the one-sided power spectrum

$$P_b^u(n) = \left| \text{FT}_t \left[\left(\delta_b^u - \bar{\delta}_b^u \right) \odot h \right]_n \right|^2 , \tag{9}$$

where $\bar{\delta}_b^u$ is the temporal mean, $h \in \mathbb{R}^T$ is a Hann window [8], and FT_t denotes the temporal Fourier transform.

We consider the physiological frequency-bin set

$$\Omega_f = \left\{ n \;\middle|\; f_{\min} \le \frac{n f_s}{T} \le f_{\max} \right\}, \tag{10}$$

where f_s is the sampling frequency and $f_{\min} = 0.7\,\text{Hz}$ and $f_{\max} = 2.8\,\text{Hz}$. Within this range, the normalized spectral distribution is

$$\pi_b^u(n) = \frac{P_b^u(n) + \varepsilon_0}{\sum_{m \in \Omega_f} \left(P_b^u(m) + \varepsilon_0 \right)}, \qquad n \in \Omega_f. \tag{11}$$

The physiological spectral loss is

$$\mathcal{L}_{\text{spec}} = \frac{1}{B} \sum_{b=1}^{B} \text{JS}\left(\pi_b^{\text{pred}}, \pi_b^{\text{gt}} \right), \tag{12}$$

where JS denotes the Jensen–Shannon divergence [14]. This objective aligns the relative spectral-energy distribution within the physiological frequency range.

Multi-Lag Pulse-Phase Progression Loss. Magnitude-spectrum supervision does not preserve temporal order and therefore does not directly constrain local pulse timing. We introduce MPPL, which constrains local pulse timing by aligning relative analytic-phase increments over multiple short temporal lags in a reconstructed, band-limited pulse representation.

For $u \in \{\mathrm{pred}, \mathrm{gt}\}$, we reconstruct a pulse representation by taking the negative cumulative sum of δ^u, removing its linear trend, and applying a fourth-order zero-phase Butterworth band-pass filter [2] with cutoff frequencies of 0.7 and 2.8 Hz, yielding $y^u \in \mathbb{R}^{B\times T}$. The corresponding analytic signal is

$$\zeta^u = \mathcal{C}_{L_c}\left[y^u + \mathrm{i}\,\mathrm{Hilb}(y^u)\right] \in \mathbb{C}^{B\times T_v}, \tag{13}$$

where $\mathrm{i}^2 = -1$, Hilb denotes the Hilbert transform [7], and $\mathcal{C}_{L_c}$ removes L_c samples from each temporal boundary, giving $T_v = T - 2L_c$.

For each sample b, we define the target-derived reference scale β_b and, for $t = 1, \ldots, T_v$, the stabilized analytic-phase phasor $\varphi^u_{b,t}$ as

$$\begin{aligned} \beta_b &= \sqrt{\frac{1}{T_v}\sum_{\tau=1}^{T_v}\left|\zeta^{\mathrm{gt}}_{b,\tau}\right|^2 + \varepsilon_0}, \\ \varphi^u_{b,t} &= \frac{\zeta^u_{b,t}}{\sqrt{\left|\zeta^u_{b,t}\right|^2 + \varepsilon_0\beta_b^2}}, \qquad u \in \{\mathrm{pred}, \mathrm{gt}\}. \end{aligned} \tag{14}$$

We use the lag set $\mathcal{K} = \{1, 2, 4\}$ and the valid index set $\mathcal{I}_\kappa = \{\kappa+1, \ldots, T_v\}$. Here, $|\mathcal{K}|$ and $|\mathcal{I}_\kappa|$ denote the corresponding set cardinalities. For $u \in \{\mathrm{pred}, \mathrm{gt}\}$, $\kappa \in \mathcal{K}$, and $t \in \mathcal{I}_\kappa$, the relative phase-progression phasor is

$$\omega^{u,(\kappa)}_{b,t} = \frac{\varphi^u_{b,t}\left(\varphi^u_{b,t-\kappa}\right)^*}{\sqrt{\left|\varphi^u_{b,t}\left(\varphi^u_{b,t-\kappa}\right)^*\right|^2 + \varepsilon_0}}, \tag{15}$$

where $(\cdot)^*$ denotes complex conjugation.

Because analytic phase is less reliable at weak target envelopes, we define the target-derived confidence

$$\begin{aligned} \chi^{(\kappa)}_{b,t} &= \sqrt{\frac{\left|\zeta^{\mathrm{gt}}_{b,t}\right|\left|\zeta^{\mathrm{gt}}_{b,t-\kappa}\right|}{\beta_b^2} + \varepsilon_0}, \\ \alpha^{(\kappa)}_{b,t} &= \min\left\{2,\ \frac{\chi^{(\kappa)}_{b,t}}{\dfrac{1}{|\mathcal{I}_\kappa|}\displaystyle\sum_{\tau\in\mathcal{I}_\kappa}\chi^{(\kappa)}_{b,\tau} + \varepsilon_0}\right\}. \end{aligned} \tag{16}$$

The normalized confidence $\alpha^{(\kappa)}_{b,t}$ is clipped at 2 to limit the influence of individual intervals.

The confidence-weighted MPPL objective is

$$\mathcal{L}_{\mathrm{MPPL}} = \frac{1}{B|\mathcal{K}|}\sum_{b=1}^{B}\sum_{\kappa\in\mathcal{K}}\frac{\displaystyle\sum_{t\in\mathcal{I}_\kappa}\alpha^{(\kappa)}_{b,t}\frac{1}{2}\left|\omega^{\mathrm{pred},(\kappa)}_{b,t} - \omega^{\mathrm{gt},(\kappa)}_{b,t}\right|^2}{\displaystyle\sum_{t\in\mathcal{I}_\kappa}\alpha^{(\kappa)}_{b,t} + \varepsilon_0}. \tag{17}$$

By comparing relative rather than absolute phase, MPPL is invariant to a constant global phase offset and a global polarity reversal, while penalizing discrepancies in local phase progression. The loss weight and numerical constants are specified in the experimental setup.

4 Experiments

4.1 Experimental Setup

Dataset. We use egoPPG-DB, which contains more than 13 hours of synchronized NIR eye-camera videos and physiological signals collected from 25 participants during daily activities [1]. We use the video, office, kitchen, dancing, bike, and walking activities. The model takes single-channel NIR eye-camera clips and synchronized headset IMU magnitudes as input. Each clip has a spatial resolution of 48×128. Frame differencing followed by standardization is applied to the video input. Each clip contains $T = 128$ frames sampled at the native 30 Hz without input-frame subsampling, corresponding to approximately 4.27 seconds.

Settings. We train the model for 100 epochs with a batch size of 4 using Adam and OneCycleLR, with a maximum learning rate of 9×10^{-4}, on a single NVIDIA B200 GPU. We use participant-disjoint five-fold cross-validation and repeat each fold with three random seeds. All reported learning-based results, including the ablations, are averaged over the resulting 15 runs. We additionally reproduce PulseFormer using its original architecture and batch-global z-normalized MSE waveform objective under the same participant splits, preprocessing, and HR evaluation protocol as BEAT. For MPPL, we use $\mathcal{K} = \{1, 2, 4\}$, $L_c = 27$, $\varepsilon_0 = 10^{-8}$. The confidence weights are clipped to $[0, 2]$ and we set $\lambda_{\mathrm{MPPL}} = 0.5$.

Evaluation Metrics. We evaluate HR estimation using mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and Pearson correlation coefficient (r). MAE and RMSE are reported in BPM, while MAPE is reported in percentage. Lower MAE, RMSE, and MAPE and higher r indicate better performance. For computational efficiency, we report parameters, multiply–accumulate operations (MACs), latency, and kFPS. One multiplication followed by one accumulation is counted as one MAC for a single clip. Latency denotes the time required for one forward pass on one clip. kFPS denotes video-frame throughput in thousands of frames per second.

HR Computation. We concatenate the predicted and target differential PPG clips in participant-specific temporal order and divide each sequence into 60-second windows. For $u \in \{\mathrm{pred}, \mathrm{gt}\}$, let $x^u \in \mathbb{R}^{N_w}$ denote a differential PPG waveform within one window, where $N_w = 60 f_s = 1800$ at $f_s = 30$ Hz. We reconstruct and band-limit x^u using the same temporal integration, linear detrending, and Butterworth filtering operations defined above.

Table 1: Comparison with existing rPPG methods on egoPPG-DB. Published baseline results are taken from egoPPG [1]. PulseFormer (our reproduction) denotes our reimplementation evaluated under the protocol described in Section 4.1. Results for our PulseFormer reproduction and BEAT are averaged over five participant-disjoint folds and three random seeds per fold. Lower MAE, RMSE, and MAPE and higher r indicate better performance.

Model	MAE ↓	RMSE ↓	MAPE (%) ↓	r ↑
Yue et al. [31]	29.63	32.99	37.86	0.10
DeepPhys [3]	28.26	31.97	36.68	0.08
TS-CAN [15]	26.32	32.39	29.13	0.11
ContrastPhys+ [24]	19.12	24.13	22.57	0.21
RhythmMamba [33]	15.05	19.78	17.46	-0.16
Signal processing (eyes) [1]	14.60	18.18	18.37	0.20
PhysMamba [27]	13.94	16.86	17.76	0.61
RhythmFormer [32]	13.13	17.43	14.73	0.51
Signal processing (skin) [1]	12.40	15.54	15.29	0.50
PhysNet [28]	12.09	15.43	15.14	0.66
PhysFormer [30]	10.71	13.97	12.69	0.72
FactorizePhys [9]	10.07	13.43	12.36	0.67
PulseFormer [1]	7.67	10.69	9.45	0.85
PulseFormer (our reproduction)	9.12	12.15	11.19	0.79
BEAT	**7.20**	**9.68**	**8.65**	**0.88**

Let $\nu_1^u < \cdots < \nu_{J_u}^u$ denote the sample indices of the J_u detected pulse peaks. For $J_u \geq 2$, the heart rate is computed from the mean inter-peak interval as

$$\mathrm{HR}^u = 60 f_s \left[\frac{1}{J_u - 1} \sum_{j=1}^{J_u - 1} \left(\nu_{j+1}^u - \nu_j^u \right) \right]^{-1}. \tag{18}$$

The same evaluation procedure is applied independently to the prediction and target. Unlike MPPL, HR evaluation operates on 60-second windows and does not remove boundary samples.

4.2 Main Results

Table 1 compares BEAT with existing rPPG methods on egoPPG-DB. BEAT achieves the best results across all metrics, with an MAE of 7.20 BPM, an RMSE of 9.68 BPM, a MAPE of 8.65 %, and $r = 0.88$. Our PulseFormer reproduction obtains an MAE of 9.12 BPM, compared with the 7.67 BPM reported in egoPPG [1]. We report the published result as the literature benchmark and include our reproduction for transparency. Relative to the PulseFormer result reported in egoPPG [1], BEAT achieves a 0.47 BPM lower MAE, a 1.01 BPM lower RMSE, and a 0.80 percentage-point lower MAPE. The correlation is also higher, at 0.88 compared with 0.85. The lower errors and higher correlation indicate

Table 2: Computational comparison between our PulseFormer implementation and BEAT on an NVIDIA B200 GPU. All measurements use FP32 inference, a batch size of 4, and 128-frame clips. Parameters and MACs quantify model complexity, while latency and kFPS report inference efficiency. Latency is reported as the median over seven runs.

Model	T	Params (M)	MACs (G)	Latency (ms)	kFPS
PulseFormer	128	12.08	49.69	4.01	31.94
BEAT	128	12.10	49.70	4.61	27.77

Table 3: Individual and combined effects of the TCN adapter and MPPL. All configurations use $\mathcal{L}_{\text{wave}} + 0.1\mathcal{L}_{\text{spec}}$; $0.5\mathcal{L}_{\text{MPPL}}$ is added only where indicated. The first row retains the PulseFormer architecture under this common waveform–spectral objective. The joint configuration yields the best HR estimation performance across all metrics.

TCN adapter	MPPL	MAE	RMSE	MAPE (%)	r
		8.83	12.63	11.65	0.83
✓		7.84	10.62	9.41	0.87
	✓	7.82	11.07	9.96	0.86
✓	✓	**7.20**	**9.68**	**8.65**	**0.88**

better absolute HR accuracy and agreement with the reference HR values. The substantially larger errors of methods designed mainly for frontal facial videos further show the difficulty of recovering pulse cues from the restricted periocular region.

4.3 Efficiency Analysis

Table 2 compares the computational efficiency of BEAT and our PulseFormer implementation under the same hardware and input settings. BEAT uses 12.10M parameters and 49.70G MACs, compared with 12.08M parameters and 49.69G MACs for PulseFormer. The proposed adapter therefore changes both the parameter count and MACs by less than 0.2 %. The measured latency increases from 4.01 to 4.61 ms per clip, while frame throughput decreases from 31.94 to 27.77 kFPS. The practical runtime difference is larger than the difference in nominal MACs. Nevertheless, BEAT processes a 128-frame clip in 4.61 ms while preserving nearly unchanged model complexity.

4.4 Ablation Studies

Adapter and MPPL. Table 3 evaluates the individual and combined effects of the TCN adapter and MPPL. The configuration without either proposed component retains the PulseFormer architecture and uses $\mathcal{L}_{\text{wave}}+0.1\mathcal{L}_{\text{spec}}$, obtaining an MAE of 8.83 and a correlation of 0.83. Using only the adapter reduces the

Table 4: Ablation of temporal dilation schedules and their corresponding receptive fields. The receptive field is computed for four non-causal residual depthwise TCN blocks with a kernel size of 3. The $[1, 2, 4, 8]$ schedule provides the best overall error performance. All other model and loss settings are fixed, and only the dilation schedule is varied.

Dilations	Receptive field	MAE	RMSE	MAPE (%)	r
$[1, 1, 1, 1]$	9	9.45	12.03	12.17	**0.88**
$[1, 2, 2, 2]$	15	8.59	11.66	10.57	0.86
$[1, 2, 4, 4]$	23	8.80	11.73	11.33	0.87
$[1, 2, 4, 8]$	31	**7.20**	**9.68**	**8.65**	**0.88**

Table 5: Individual and combined effects of the waveform, spectral, and MPPL terms. All configurations use the final TCN architecture and vary only the indicated loss terms. Combining all three objectives achieves the best performance across all metrics.

Waveform Loss	Spectral Loss	MPPL	MAE	RMSE	MAPE (%)	r
✓			8.64	11.77	9.64	0.82
✓	✓		7.84	10.62	9.41	0.87
✓		✓	9.05	12.25	10.72	0.83
✓	✓	✓	**7.20**	**9.68**	**8.65**	**0.88**

MAE to 7.84 and improves the correlation to 0.87. Using only MPPL gives a similar MAE of 7.82 and a correlation of 0.86. The adapter gives better RMSE and correlation, whereas MPPL gives a slightly lower MAE. Their combination achieves the best result across all metrics, with an MAE of 7.20, an RMSE of 9.68, a MAPE of 8.65 %, and a correlation of 0.88. These results indicate that bottleneck temporal refinement and multi-lag phase-progression supervision provide complementary improvements.

Dilation Schedule. Table 4 evaluates different dilation schedules and temporal receptive fields. Increasing the receptive field from 9 to 15 reduces the MAE from 9.45 to 8.59, but the receptive field of 23 slightly degrades the error metrics. This non-monotonic trend indicates that performance depends on both receptive-field size and the distribution of dilation factors. The exponential schedule $[1, 2, 4, 8]$ achieves the lowest MAE, RMSE, and MAPE of 7.20, 9.68, and 8.65 %, respectively. Its receptive field of 31 provides broad temporal context while progressively covering short- and long-range patterns. We therefore adopt $[1, 2, 4, 8]$ as the final dilation schedule.

Loss Composition. Table 5 analyzes the relationship among the waveform, spectral, and MPPL terms. Adding the spectral loss to the waveform loss reduces MAE from 8.64 to 7.84 and increases the correlation from 0.82 to 0.87. This improvement shows that spectral supervision helps establish more reliable physiological periodicity. In contrast, combining MPPL only with the waveform

Table 6: Controlled comparison of location-wise and spatially pooled temporal refinement at the bottleneck. Both variants operate at $T = 32$ and use the same TCN architecture, dilation schedule, parameter count, and training objective. The pooled control broadcasts a residual correction derived from spatially averaged features, whereas ours independently refines the 24 spatial trajectories using shared parameters.

Temporal refinement	MAE	RMSE	MAPE (%)	r
Spatially pooled control	7.86	10.76	9.50	0.86
Location-wise (ours)	**7.20**	**9.68**	**8.65**	**0.88**

loss increases MAE to 9.05, although the correlation slightly improves to 0.83. This result suggests that multi-lag phase-progression supervision alone is insufficient when the prediction lacks stable physiological periodicity. When MPPL is combined with both waveform and spectral supervision, the model achieves the best result across all metrics. Compared with the waveform–spectral setting, MPPL further reduces MAE from 7.84 to 7.20 and RMSE from 10.62 to 9.68, while increasing the correlation from 0.87 to 0.88. The results support a connected objective in which waveform supervision aligns the basic signal, spectral supervision stabilizes its periodicity, and MPPL refines local analytic-phase progression.

Location-Wise Refinement. Table 6 more directly tests whether the adapter gain arises from temporal processing alone or from preserving location-specific temporal trajectories. Both variants operate on the same $B \times 64 \times 32 \times 3 \times 8$ bottleneck representation and use identical 64-channel, four-block depthwise TCN parameterization with a kernel size of 3, dilations $[1, 2, 4, 8]$, a receptive field of 31, and 18,560 parameters. They also use the same total parameter count, training objective, and 100-epoch training schedule.

The spatially pooled control first averages the 3×8 spatial grid, applies the TCN to the resulting $B \times 64 \times 32$ temporal sequence, and broadcasts the resulting residual correction to all spatial locations. In contrast, the location-wise variant applies the shared TCN independently to each of the 24 spatial trajectories. Location-wise refinement reduces MAE from 7.86 to 7.20 BPM, RMSE from 10.76 to 9.68 BPM, and MAPE from 9.50 % to 8.65 %, while increasing the correlation from 0.86 to 0.88. These correspond to relative reductions of 8.42 %, 10.00 %, and 8.96 % in MAE, RMSE, and MAPE, respectively. Because the two variants are matched in bottleneck resolution, temporal configuration, parameter count, and training objective, these results support preserving location-specific temporal trajectories during temporal refinement rather than deriving a single correction from their spatial mean.

5 Conclusion

We presented BEAT, a framework for HR estimation from NIR videos captured by wearable eye cameras. BEAT combines a TCN adapter that refines

each bottleneck spatial location before global spatial pooling with MPPL, which supervises analytic phase progression over multiple temporal lags in the reconstructed pulse domain. The adapter preserves and refines location-specific temporal features, while MPPL complements waveform and spectral supervision by constraining local pulse timing through multi-lag phase progression. Experiments on egoPPG-DB show that BEAT achieves improved HR estimation accuracy with only a small increase in parameters and MACs. A matched bottleneck control further shows that location-wise temporal refinement outperforms a spatially pooled correction under the same temporal configuration and parameter count. The remaining ablations show that the adapter and MPPL work best together and that MPPL is most effective after spectral supervision establishes stable physiological periodicity.

We evaluate participant-disjoint, fixed-length clips from egoPPG-DB. The current model uses both past and future frames within each clip. In addition, MPPL requires a stable band-limited pulse representation. Its benefit may therefore diminish under very low pulse amplitudes or severe motion corruption. Future work will evaluate the method across different wearable eye-camera settings and investigate causal temporal modeling for online HR estimation.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT)(IITP-2026-RS-2024-00439292).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Braun, B., Armani, R., Meier, M., Moebus, M., Holz, C.: egoppg: Heart rate estimation from eye-tracking cameras in egocentric systems to benefit downstream vision tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5579–5590 (2025)
2. Butterworth, S.: On the theory of filter amplifiers. Wireless Engineer **7**(6), 536–541 (1930)
3. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the european conference on computer vision (ECCV). pp. 349–365 (2018)
4. Comas, J., Ruiz, A., Sukno, F.: Efficient remote photoplethysmography with temporal derivative modules and time-shift invariant loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2182–2191 (2022)

5. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: International conference on machine learning. pp. 933–941. PMLR (2017)
6. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE transactions on biomedical engineering **60**(10), 2878–2886 (2013)
7. Gabor, D.: Theory of communication. part 1: The analysis of information. Journal of the Institution of Electrical Engineers-part III: radio and communication engineering **93**(26), 429–441 (1946)
8. Harris, F.J.: On the use of windows for harmonic analysis with the discrete fourier transform. Proceedings of the IEEE **66**(1), 51–83 (1978)
9. Joshi, J., Agaian, S.S., Cho, Y.: Factorizephys: Matrix factorization for multidimensional attention in remote physiological sensing. arXiv preprint arXiv:2411.01542 (2024)
10. Lam, A., Kuno, Y.: Robust heart rate measurement from video using select random patches. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3640–3648 (2015)
11. Lee, E., Chen, E., Lee, C.Y.: Meta-rPPG: Remote heart rate estimation using a transductive meta-learner. In: Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol. 12372, pp. 392–409. Springer (2020). https://doi.org/10.1007/978-3-030-58583-9_24
12. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4264–4271 (2014)
13. Li, Z., Yin, L.: Contactless pulse estimation leveraging pseudo labels and self-supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20588–20597 (2023)
14. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (1999)
15. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. Advances in Neural Information Processing Systems **33**, 19400–19411 (2020)
16. Liu, X., Hill, B., Jiang, Z., Patel, S., McDuff, D.: EfficientPhys: Enabling simple, fast and accurate camera-based cardiac measurement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5008–5017 (2023)
17. Lu, H., Yu, Z., Niu, X., Chen, Y.C.: Neuron structure modeling for generalizable remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18589–18599 (2023)
18. Niu, X., Han, H., Shan, S.: Remote photoplethysmography-based physiological measurement: a survey. Journal of Image and Graphics **25**(11), 2321–2336 (2020)
19. Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., Chen, X.: Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE Journal of Biomedical and Health Informatics **25**(5), 1373–1384 (2021)
20. Speth, J., Vance, N., Flynn, P., Czajka, A.: Non-contrastive unsupervised learning of physiological signals from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14464–14474 (2023)
21. Sun, W., Zhang, X., Chen, Y., Ge, Y., Ji, C., Huang, X.: BYHE: A simple framework for boosting end-to-end video-based heart rate measurement network. arXiv preprint arXiv:2207.01697 (2022)

22. Sun, W., Zhang, X., Lu, H., Chen, Y., Ge, Y., Huang, X., Yuan, J., Chen, Y.: Resolve domain conflicts for generalizable remote physiological measurement. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8214–8224 (2023)
23. Sun, Z., Li, X.: Contrast-Phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In: Computer Vision – ECCV 2022. Lecture Notes in Computer Science, vol. 13672, pp. 492–510. Springer (2022). https://doi.org/10.1007/978-3-031-19775-8_29
24. Sun, Z., Li, X.: Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5835–5851 (2024)
25. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2396–2404 (2016)
26. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
27. Yan, Z., Zhong, Y., Zhang, W., Shu, L., Xu, H., Kang, W.: Physmamba: Leveraging dual-stream cross-attention ssd for remote physiological measurement. arXiv preprint arXiv:2408.01077 (2024)
28. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. arXiv preprint arXiv:1905.02419 (2019)
29. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 151–160 (2019)
30. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P.H., Zhao, G.: Physformer: Facial video-based physiological measurement with temporal difference transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4186–4196 (2022)
31. Yue, Z., Shi, M., Ding, S.: Facial video-based remote physiological measurement via self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13844–13859 (2023)
32. Zou, B., Guo, Z., Chen, J., Zhuo, J., Huang, W., Ma, H.: Rhythmformer: Extracting patterned rppg signals based on periodic sparse attention. Pattern Recognition **164**, 111511 (2025)
33. Zou, B., Guo, Z., Hu, X., Ma, H.: Rhythmmamba: Fast, lightweight, and accurate remote physiological measurement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 11077–11085 (2025)