SKIA: Spatial-Knowledge Informed Attention for Robust Remote Photoplethysmography under Lower-Face Motion Artifacts

Van-Truong Le[a,b], Viet-Tham Huynh[a,b], and Minh-Triet Tran[a,b]

[a]University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[b]Vietnam National University, Ho Chi Minh City, Vietnam

ABSTRACT

Remote photoplethysmography (rPPG) enables contactless physiological monitoring from facial videos, but non-rigid facial motion can overwhelm subtle hemodynamic variations. Existing approaches generally compensate for global motion or learn spatial attention implicitly, leaving the effect of local lower-face activity insufficiently characterized. We investigate this effect and present SKIA, a lightweight mask-guided framework for robust blood volume pulse reconstruction. Experiments on UBFC-rPPG show that the mouth and jaw regions contribute substantial motion artifacts while providing less stable physiological information than the forehead and upper cheeks. Based on this observation, we introduce a static spatial-knowledge mask that attenuates the lower face while preserving informative skin regions. We further propose SKIA-Net, a compact spatiotemporal architecture that combines raw appearance, masked representations, and temporal residuals. Static masking improves CHROM Pearson correlation from 0.3507 to 0.4166 ($p < 0.001$), whereas SKIA-Net achieves a correlation of 0.7522 and a heart-rate mean absolute error of 3.73 bpm. Ablations show that the stable anatomical prior is more effective than dynamic optical-flow masking. These results demonstrate that explicit spatial knowledge provides an efficient and interpretable approach to motion-robust rPPG. Code is available at github.com/lee-vtruong/SKIA-rPPG.

Keywords: remote photoplethysmography, rPPG, physiological computing, motion artifact suppression, facial video analysis, heart rate estimation, spatiotemporal learning

1. INTRODUCTION

Remote photoplethysmography (rPPG) estimates blood volume pulse (BVP) and heart rate from subtle skin-color variations in facial videos, enabling unobtrusive physiological monitoring without contact sensors.[1] However, the weak hemodynamic signal is easily corrupted by illumination changes and facial motion. Non-rigid mouth and jaw movements are particularly challenging because they alter local appearance and color aggregation throughout a video.

Classical methods such as CHROM and POS suppress some illumination and motion variation through chrominance projections, while deep models learn spatiotemporal representations directly from video. Nevertheless, most methods process the face globally or learn motion invariance implicitly, without identifying which regions introduce the strongest artifacts. We therefore examine a simple spatial hypothesis: the lower face contributes disproportionate task-induced motion but less stable physiological information than the forehead and upper cheeks.

This distinction is important because region selection changes both the strength of the physiological signal and the amount of motion contamination. Completely discarding moving pixels may remove useful skin information, whereas rapidly changing attention weights can introduce artificial temporal variation. A useful spatial prior must therefore suppress unstable regions while remaining consistent across frames.

Accordingly, we study two related questions: whether a fixed lower-face prior can improve a conventional color-based estimator, and whether the same prior remains useful when fused with learned appearance and motion representations. Evaluating both settings separates the contribution of spatial knowledge from that of network capacity.

Experiments on UBFC-rPPG show that static lower-face suppression improves CHROM waveform reconstruction for most subjects, whereas frame-varying optical-flow masks degrade performance by changing the sampled skin area over time. Motivated by this finding, we propose SKIA (Spatial-Knowledge Informed Attention) and its compact reconstruction network, SKIA-Net. The framework uses a stable anatomical mask and jointly models raw appearance, masked appearance, and temporal residuals, retaining complementary information instead of discarding motion cues entirely.

The contributions of this work are summarized as follows:

- We analyze lower-face motion artifacts and show that stable suppression of motion-dominant regions improves rPPG reconstruction.
- We introduce a static spatial-knowledge mask and SKIA-Net, which combines raw, masked, and temporal representations in a compact architecture.
- We demonstrate statistically significant improvements over CHROM and single-stream variants, and show that static masking outperforms dynamic optical-flow masking.

2. RELATED WORK

Early rPPG methods recovered pulse signals by spatially averaging facial color traces and separating physiological components from noise. Poh *et al.* used independent component analysis for webcam-based pulse measurement,[2] but such source-separation approaches remained sensitive to motion and illumination. Chrominance methods subsequently improved robustness without costly learning. CHROM projects normalized RGB traces into a chrominance subspace designed to reduce motion-induced color distortion,[3] while POS constructs a plane orthogonal to the skin-tone direction.[4] These efficient methods remain common baselines, yet their performance depends strongly on region-of-interest selection because all included pixels contribute to the aggregated color traces. Motion-contaminated areas can therefore dominate the weak physiological component.

Deep learning enables direct estimation of physiological signals from spatiotemporal video representations. DeepPhys introduced appearance-guided attention for camera-based heart-rate and respiration measurement,[5] and PhysNet used a 3D convolutional encoder–decoder to reconstruct BVP waveforms.[6] Later work explored lightweight temporal modeling in EfficientPhys[7] and frequency-domain supervision.[8] Transformer[9] and self-supervised approaches[10] further improved temporal representation learning. Although these architectures improve accuracy, robustness to unseen motion remains difficult because motion invariance is commonly learned implicitly from the training distribution. SKIA-Net instead supplies an explicit anatomical prior while retaining raw and temporal information, allowing the network to learn from both complete and motion-suppressed observations.

Motion-robust rPPG research has considered face stabilization, skin segmentation, adaptive ROI tracking, optical flow, temporal filtering, and spatial attention. The forehead and cheeks are comparatively stable and informative, whereas deformable regions can reduce signal quality.[11] Realistic motion nevertheless remains challenging for learned rPPG systems.[12] Unlike methods that dynamically change spatial weights, SKIA uses a fixed, parameter-free lower-face prior with both CHROM and SKIA-Net. Experiments use UBFC-rPPG, which provides synchronized facial videos and contact signals under a mathematical task;[13] it contains natural facial activity but is not a dedicated speaking benchmark.

3. METHOD

3.1 Problem Formulation

Given a facial clip $X \in \mathbb{R}^{T\times H\times W\times 3}$, the goal is to reconstruct its BVP waveform $y \in \mathbb{R}^T$. SKIA-Net learns a mapping $f_\theta : X \mapsto \hat{y}$ while reducing sensitivity to local appearance changes that are unrelated to blood-volume variation. Because BVP amplitude depends on acquisition conditions, reconstruction quality is measured primarily by temporal waveform agreement. The Pearson correlation is

$$\rho(\hat{y}, y) = \frac{\sum_t (\hat{y}_t - \bar{\hat{y}})(y_t - \bar{y})}{\sqrt{\sum_t (\hat{y}_t - \bar{\hat{y}})^2}\sqrt{\sum_t (y_t - \bar{y})^2}}, \tag{1}$$

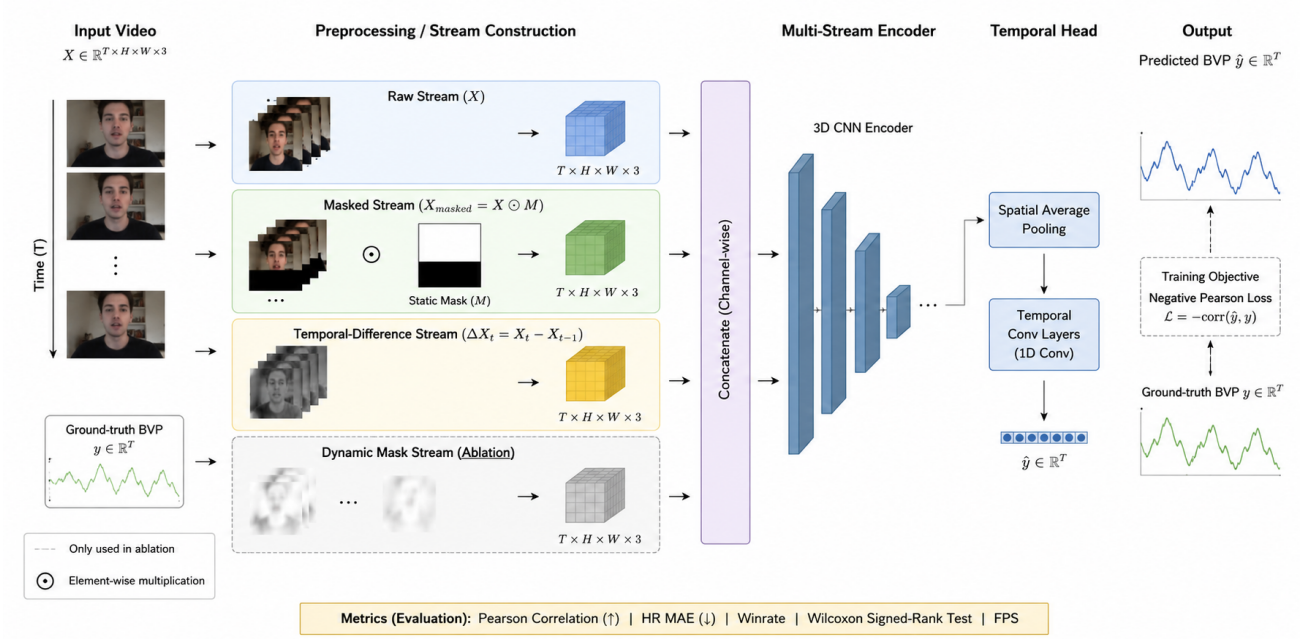


Figure 1: Overview of the proposed SKIA framework. The input facial video is processed into three complementary representations: raw frames, static lower-face suppressed frames, and temporal-difference frames. These representations are jointly encoded by a lightweight spatiotemporal network to reconstruct the target BVP waveform.

where higher values indicate better morphological and temporal consistency. Heart rate is obtained from the dominant FFT peak and evaluated by mean absolute error following standard rPPG protocols.[14]

3.2 Static Lower-Face Suppression and Masked rPPG

Lower-face articulation produces non-rigid deformation and intensity changes that can dominate pulse-induced color variation. We therefore suppress the mouth and jaw while retaining the forehead and upper cheeks using a static mask $M \in \mathbb{R}^{H\times W}$:

$$M(i,j) = \begin{cases} 1, & i < \alpha H, \\ 0, & \text{otherwise}, \end{cases}$$

where $\alpha = 0.7$ retains the upper 70% of the face and $X_{\text{masked}} = X \odot M$. Unlike frame-varying motion masks, this anatomical prior maintains the same sampled region throughout a clip. The retained ratio controls a trade-off between preserving skin area and removing motion: aggressive suppression discards physiological information, whereas a large ratio retains part of the moving mouth. The ratio ablation in Section 4.5 identifies $\alpha = 0.7$ as the best balance. The mask is defined relative to the cropped facial height, so it can be applied consistently after resizing without requiring subject-specific landmarks. It is intended as a lightweight prior rather than an exact anatomical segmentation.

Masked CHROM formulation. To test the spatial prior independently of deep learning, masked CHROM computes each color trace only over retained pixels, for example

$$R_t = \frac{\sum_{i,j} M_{i,j} X^R_{t,i,j}}{\sum_{i,j} M_{i,j}},$$

with analogous definitions for G_t and B_t. Within each temporal window $\mathcal{W}$, the traces are normalized by their window means, $\widetilde{C}_t = C_t/\mu_C - 1$ for $C \in \{R, G, B\}$. The CHROM components are

$$X_{chr} = 0.77\widetilde{R} - 0.51\widetilde{G}, \qquad Y_{chr} = 0.77\widetilde{R} + 0.51\widetilde{G} - 0.77\widetilde{B}.$$

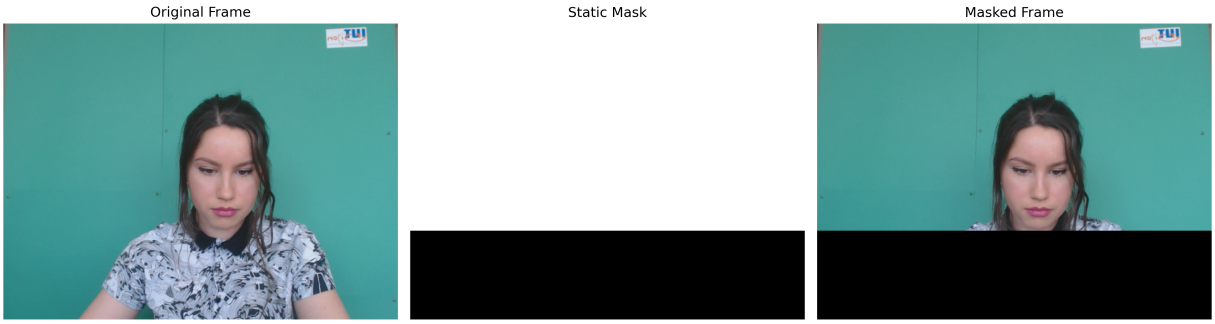


Figure 2: Static lower-face suppression preserving the forehead and upper cheeks.

After bandpass filtering both components within the heart-rate range, the window-level pulse estimate is $S = X_{chr} - [\sigma(X_{chr})/\sigma(Y_{chr})]Y_{chr}$, where $\sigma(\cdot)$ is computed within $\mathcal{W}$. Window estimates are combined into the complete rPPG signal. Thus, masking changes only spatial aggregation; the chrominance projection and temporal processing remain identical to the baseline. This controlled comparison isolates the contribution of lower-face suppression from changes in the rPPG algorithm.

Dynamic mask attempt. We additionally generate motion-dependent soft masks from dense optical flow between consecutive frames,[15] assigning lower weights to regions with larger motion magnitude. Although this strategy adapts to observed movement, its frame-wise variation changes the sampled color distribution and degrades reconstruction, as quantified in the ablation study.

3.3 SKIA-Net Architecture and Training

Building upon the effectiveness of static lower-face suppression, we introduce SKIA-Net, a lightweight spatiotemporal architecture that jointly exploits three complementary input representations: the raw appearance stream X, the static masked stream $X_{\text{masked}} = X \odot M$, and the temporal-difference stream $\Delta X_t = X_t - X_{t-1}$, which explicitly captures short-term motion variations and physiological dynamics between adjacent frames. These representations are concatenated along the channel dimension and processed by a lightweight 3D convolutional encoder. The encoder extracts joint spatiotemporal physiological features while preserving temporal coherence. A spatial average pooling layer followed by a temporal convolutional prediction head then reconstructs the target BVP waveform. Figure 1 presents the overall architecture.

All three representations retain the same temporal length and spatial resolution, ensuring frame-level alignment before fusion. The 3D convolutions therefore observe appearance and motion evidence within the same local temporal neighborhood. Spatial pooling aggregates distributed skin responses and reduces sensitivity to their exact image coordinates, while the temporal head preserves the input sequence length and produces one BVP estimate per frame.

The streams play distinct roles: raw frames preserve all skin-color evidence, masked frames emphasize stable upper-face regions, and temporal differences highlight short-term variation. Early channel-wise fusion allows the encoder to learn when motion cues support pulse reconstruction and when they should be attenuated by the anatomical prior.

To systematically analyze the contribution of each representation, we evaluate four variants summarized in Table 1. Experimental results demonstrate that combining complementary representations substantially improves physiological reconstruction robustness compared with single-stream variants.

Training objective. The network is trained using Negative Pearson Correlation Loss, directly optimizing

Table 1: SKIA-Net variants and their input configurations.

Variant	Raw	Masked	Temporal Diff.
SKIA Raw	✓		
SKIA Masked		✓	
SKIA Raw+Masked	✓	✓	
SKIA Full	✓	✓	✓

waveform similarity between predicted and ground-truth physiological signals:

$$\mathcal{L} = -\frac{\sum_t (\hat{y}_t - \bar{\hat{y}})(y_t - \bar{y})}{\sqrt{\sum_t (\hat{y}_t - \bar{\hat{y}})^2}\sqrt{\sum_t (y_t - \bar{y})^2}}.$$

Compared with point-wise regression losses such as Mean Squared Error (MSE), Pearson-based optimization focuses on waveform consistency and temporal alignment, which are more suitable for physiological signal reconstruction. It is also invariant to affine changes in waveform amplitude, which is useful because video-derived pulse magnitude can vary with illumination, skin exposure, and camera response even when the underlying cardiac timing is unchanged.

4. EXPERIMENTS AND RESULTS

4.1 Dataset

UBFC-rPPG contains RGB facial videos, contact-sensor BVP, and heart-rate signals from 42 subjects performing a time-sensitive mathematical task.[13] Because CHROM requires no training, the classical CHROM comparison and all masking-ratio and dynamic-mask ablations are evaluated subject-wise on all 42 participants. For SKIA-Net, we use a fixed split of 32 training and 10 test subjects. The test subjects are `43--49`, `5`, `8`, and `9`; all others are used for training, with no participant shared across sets. The task includes natural head and lower-face activity but is not treated as a dedicated speaking corpus.

4.2 Implementation Details

Models were implemented in PyTorch[16] and trained on an NVIDIA A100 using Adam,[17] a learning rate of 10^{-4}, batch size 4, and Negative Pearson loss. Clips contain 64 or 128 frames resized to 96×96 or 128×128, and each variant is trained for 8–20 epochs. Clip stride is either equal to half or all of the window length, depending on the configuration. Predicted and reference BVP signals are bandpass filtered at 0.7–2.5 Hz (42–150 bpm) before evaluation; the same post-processing is applied to all compared neural variants.

4.3 Evaluation Metrics

We evaluate physiological reconstruction quality using multiple complementary metrics: the Pearson correlation coefficient to measure waveform similarity between predicted and ground-truth BVP signals (higher values indicate better temporal fidelity); Mean Absolute Error (MAE) of heart rate estimated via FFT peak analysis, expressed in beats per minute (bpm); winrate, which reports the percentage of subjects for which a method outperforms the baseline; paired Wilcoxon signed-rank tests over the 42 subject-level CHROM results between baseline and masked conditions[18] to assess statistical significance; and runtime throughput measured in frames-per-second (FPS) to evaluate inference efficiency.

4.4 Physiological Reconstruction Results

We first evaluate the effect of static lower-face suppression using the classical CHROM rPPG pipeline. Table 2 presents waveform reconstruction and heart-rate estimation performance. Static lower-face suppression improves CHROM waveform correlation by +0.0659 Pearson while reducing heart-rate MAE by 10.06 bpm. The improvement is highly consistent across subjects, with a winrate of 88.1% and a Wilcoxon signed-rank p-value of 2.0×10^{-7}, confirming statistical significance and supporting our hypothesis that lower-face motion substantially contaminates chrominance-based physiological estimation.

Table 2: CHROM results over all 42 subjects.

Method	Pearson ↑	HR MAE ↓
CHROM	0.3507	33.50
CHROM + Static Mask	**0.4166**	**23.44**

Next, we evaluate the proposed SKIA-Net variants under different input configurations. Table 3 summarizes the reconstruction performance.

The masked-only variant has slightly lower correlation but better HR MAE than the raw stream, indicating that suppression improves rate stability while discarding some useful appearance information. Combining raw and masked inputs improves both metrics, confirming that the mask is a complementary prior rather than a replacement for the original signal. Adding temporal differences gives the best overall performance, showing that aligned appearance, anatomical, and short-term motion cues provide the most robust representation.

Table 3: SKIA-Net variants on the 10-subject test set.

Method	Pearson ↑	HR MAE ↓
PhysNet-lite[6]	0.2350	–
SKIA Raw	0.6320	8.99
SKIA Masked	0.6246	7.55
SKIA Raw+Masked	0.6844	4.97
SKIA Full	**0.7522**	**3.73**

4.5 Ablation Studies

To analyze the effect of lower-face suppression strength, we evaluate multiple masking ratios. Table 4 summarizes the results. Excessively aggressive suppression (ratio 0.5) removes useful physiological skin regions, degrading performance, while weak suppression (ratio 0.8) leaves motion-contaminated lower-face regions largely intact. The best performance is achieved at ratio 0.7, which provides an effective balance between motion suppression and physiological information preservation.

Table 4: Mask-ratio ablation over all 42 subjects.

Ratio	Mean Δ Pearson	Winrate
0.5	-0.0210	40.5%
0.6	+0.0592	81.0%
0.7	**+0.0659**	**88.1%**
0.8	+0.0410	85.7%

We further investigate optical-flow-based dynamic masking strategies where motion-dependent soft masks were generated from dense optical flow between consecutive frames. Table 5 compares this dynamic approach with static masking. Static masking improves correlation to 0.4166 (88.1% winrate), whereas the dynamic mask obtains 0.1534 (4.8%). Its varying weights alter frame-wise color sampling and introduce non-physiological fluctuations, showing that motion detection alone does not ensure robust rPPG.

4.6 Efficiency Analysis

We also evaluate computational efficiency for both classical and deep-learning methods. Table 6 reports runtime throughput. Deep model throughput was measured on preloaded clips using GPU inference only (NVIDIA A100), excluding preprocessing and data loading. Although these values do not reflect full end-to-end deployment speed, they demonstrate that lightweight spatiotemporal reconstruction can achieve very high inference efficiency under optimized GPU conditions.

Table 5: Static and dynamic masking over all 42 subjects.

Method	Pearson ↑	Δ vs CHROM	Winrate
CHROM Baseline	0.3507	–	–
Static Mask	**0.4166**	**+0.0659**	**88.1%**
Dynamic Soft Mask	0.1534	-0.1973	4.8%

Table 6: Runtime throughput comparison.

Method	FPS
CHROM	568.73
CHROM + Static Mask	305.63
SKIA Raw	4868.11
SKIA Raw+Masked	17523.12
SKIA Full	10626.73

5. DISCUSSION

The results indicate that lower-face activity introduces structured variation rather than random noise. Static suppression benefits both CHROM and learned reconstruction, suggesting that the mouth and jaw can contribute more motion contamination than stable hemodynamic information under the evaluated task. A fixed mask preserves the forehead and upper cheeks while avoiding changes in the sampled skin area. By contrast, optical-flow masking varies from frame to frame; even when it detects motion correctly, these changing weights can create color fluctuations that are interpreted as physiological activity.

SKIA-Net performs best when raw, masked, and temporal-difference streams are combined. The raw stream preserves complete appearance information, the masked stream emphasizes anatomically stable regions, and temporal differences provide short-term dynamics. Their complementarity explains why the full model improves both waveform fidelity and heart-rate estimation over single-stream variants. More broadly, the findings show that a simple, temporally stable spatial prior can be preferable to a more complex but unstable adaptive mask.

This study is limited to UBFC-rPPG, so cross-dataset generalization remains unverified. The dataset is not designed specifically for speech; consequently, our interpretation concerns lower-face motion rather than validated speaking behavior. The static mask is heuristic and does not adapt to facial geometry, and the dynamic baseline uses a basic optical-flow formulation. In addition, reported neural-network throughput excludes preprocessing and data loading. Future work will evaluate landmark-aware masks, stronger motion-adaptive baselines, and cross-dataset performance in speaking-intensive settings. Such evaluation should also examine whether the optimal retained ratio changes with camera resolution, head pose, skin coverage, and the type or intensity of facial activity.

6. CONCLUSION

We presented SKIA, a spatial-knowledge framework for reducing lower-face motion artifacts in rPPG. Static masking improved CHROM correlation from 0.3507 to 0.4166, while the full SKIA-Net achieved a correlation of 0.7522 and an HR MAE of 3.73 bpm on UBFC-rPPG. The results show that combining raw, masked, and temporal representations is more effective than single-stream modeling, and that stable anatomical suppression outperforms naive dynamic masking. SKIA therefore provides a compact and interpretable basis for motion-robust physiological sensing.

REFERENCES

[1] Verkruysse, W., Svaasand, L. O., and Nelson, J. S., "Remote plethysmographic imaging using ambient light.," *Optics express* **16**(26), 21434–21445 (2008).

[2] 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**(10), 10762–10774 (2010).

[3] De Haan, G. and Jeanne, V., "Robust pulse rate from chrominance-based rppg," *IEEE transactions on biomedical engineering* **60**(10), 2878–2886 (2013).

[4] Wang, W., Den Brinker, A. C., Stuijk, S., and De Haan, G., "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering* **64**(7), 1479–1491 (2016).

[5] Chen, W. and McDuff, D., "Deepphys: Video-based physiological measurement using convolutional attention networks," in [*Proceedings of the european conference on computer vision (ECCV)*], 349–365 (2018).

[6] Yu, Z., Li, X., and Zhao, G., "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," *arXiv preprint arXiv:1905.02419* (2019).

[7] Liu, X., Hill, B., Jiang, Z., Patel, S., and 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*], 5008–5017 (2023).

[8] Špetlík, R., Franc, V., and Matas, J., "Visual heart rate estimation with convolutional neural network," in [*Proceedings of the british machine vision conference, Newcastle, UK*], 3–6 (2018).

[9] Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P. H., and 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*], 4186–4196 (2022).

[10] Gideon, J. and Stent, S., "Estimating heart rate from unlabelled video," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 2743–2749 (2021).

[11] Lempe, G., Zaunseder, S., Wirthgen, T., Zipser, S., and Malberg, H., "Roi selection for remote photoplethysmography," in [*Bildverarbeitung für die Medizin 2013: Algorithmen-Systeme-Anwendungen. Proceedings des Workshops vom 3. bis 5. März 2013 in Heidelberg*], 99–103, Springer (2013).

[12] Chen, W., Yi, Z., Lim, L. J. R., Lim, R. Q. R., Zhang, A., Qian, Z., Huang, J., He, J., and Liu, B., "Deep learning and remote photoplethysmography powered advancements in contactless physiological measurement," *Frontiers in bioengineering and biotechnology* **12**, 1420100 (2024).

[13] 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).

[14] Liu, X., Narayanswamy, G., Paruchuri, A., Zhang, X., Tang, J., Zhang, Y., Sengupta, R., Patel, S., Wang, Y., and McDuff, D., "rppg-toolbox: Deep remote ppg toolbox," *Advances in Neural Information Processing Systems* **36**, 68485–68510 (2023).

[15] Farnebäck, G., "Two-frame motion estimation based on polynomial expansion," in [*Scandinavian conference on Image analysis*], 363–370, Springer (2003).

[16] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems* **32** (2019).

[17] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[18] Wilcoxon, F., "Individual comparisons by ranking methods," *Biometrics bulletin* **1**(6), 80–83 (1945).