Fusion-E2Pulse: A Multimodal Event-RGB Fusion Network for Non-contact Pulse Wave Reconstruction

Qian Feng[1], Hao Guo[1], Yan Niu[1], Zhenhuan Xu[2]✉, and Yidi Li[1]✉

[1] College of Computer Science and Technology, Taiyuan University of Technology, China
[2] College of Artificial Intelligence, Taiyuan University of Technology, China
{xuzhenhuan, liyidi}@tyut.edu.cn

Abstract. Non-contact pulse wave reconstruction hinges on the precise recovery of waveform morphology, including the dicrotic notch. Conventional RGB-based methods, which extract physiological signals from recorded facial videos, are constrained by the integral imaging mechanism of standard cameras, where the exposure process induces a smoothing effect that attenuates subtle vascular pulsation details. Conversely, neuromorphic event cameras, while offering exceptional sensitivity to intensity fluctuations, are inherently susceptible to noise and artifacts induced by minor motion. To exploit the synergy between frame-based integration and event-based differential sensing, we propose a novel multimodal network named Fusion-E2Pulse. This framework utilizes filtered RGB signals as structural priors to suppress motion artifacts, while leveraging the high-sensitivity of event streams to recover fine-grained morphological details. Experimental results demonstrate that Fusion-E2Pulse achieves compelling performance, effectively balancing noise suppression and morphological fidelity; under the subject-independent split (w/ ICA prior), it achieves a mean absolute error of 0.73 bpm for heart rate estimation, a waveform correlation of up to 0.919, and a diastolic phase duration error of 22.18 ms, supporting its ability to preserve fine-grained physiological waveform morphology.

Keywords: Pulse Wave Reconstruction · Event Camera · Non-Contact Physiological Measurement.

1 Introduction

Cardiovascular disease remains a leading cause of global mortality, requiring effective screening and long-term monitoring to mitigate its impact [21,25]. As a vital physiological window, pulse wave morphology reflects the dynamic interplay between the heart and the peripheral vascular system. Through pulse wave analysis, clinicians can obtain fundamental indicators such as Heart Rate (HR). More importantly, the precise extraction of morphological features—notably the dicrotic notch—provides critical pathological information regarding vascular resistance and arterial stiffness, which are key predictors of early-stage cardiovascular deterioration [2,25]. While traditional contact-based sensors offer high precision, they are often impractical for sensitive scenarios such as neonatal care or burn patient monitoring [19,7,28]. Consequently, remote photoplethysmography (rPPG) has emerged as a promising non-contact alternative due to its low hardware cost and accessibility [17,24]. This field has evolved significantly, ranging from classical signal processing methods such as CHROM [5], ICA [20], and POS [26] to advanced deep learning architectures like PhysNet [27] and RhythmFormer [29].

Despite these advancements, standard Red-Green-Blue (RGB)-based methods face an inherent physical bottleneck. These methods rely on capturing Photoplethysmography (PPG) signals—a measure of blood volume changes via skin color variations—from time-series pixel values in facial video sequences. The integral imaging mechanism of standard cameras acts as a physical low-pass filter by accumulating photons over fixed exposure windows, which inevitably smooths out the subtle, rapid pulsations essential for capturing subtle morphological structures [5]. The intrinsic advantages of neuromorphic event cameras precisely compensate for these deficiencies. Characterized by acute sensitivity to intensity fluctuations and high dynamic range, event cameras asynchronously encode pixel-level intensity changes to acutely capture the subtle variations caused by skin micro-vibrations [8]. Recent studies have confirmed the feasibility of non-contact HR monitoring using event cameras [11,18,6], highlighting their potential to record subtle physiological dynamics. However, the differential sensing of event cameras [16] is a double-edged sword; while providing high sensitivity, it is highly susceptible to non-physiological clutter triggered by minute motions, compromising HR estimation robustness.

In this study, we propose an innovative multimodal network named **Fusion-E2Pulse** to achieve robust noise suppression and high-fidelity waveform reconstruction. By establishing a complete framework that synergizes RGB structural stability with event dynamic sensitivity, we bridge the gap between global stability and local precision. The principal contributions are summarized as follows: (1) we propose a novel fusion paradigm that integrates RGB structural priors with event streams to effectively suppress motion-induced artifacts; (2) we establish a time-frequency adversarial supervision framework with synergistic optimization objectives, integrating spectral and morphological constraints to ensure high-fidelity physiological recovery; and (3) we perform extensive validation on the Event-based Multimodal Physiological Dataset (EMPD) [6], where Fusion-E2Pulse establishes a strong benchmark in HR estimation and enables precise reconstruction of subtle physiological landmarks.

2 Method

To implement the aforementioned synergy between RGB structural stability and event dynamic sensitivity, the Fusion-E2Pulse framework is organized into a two-

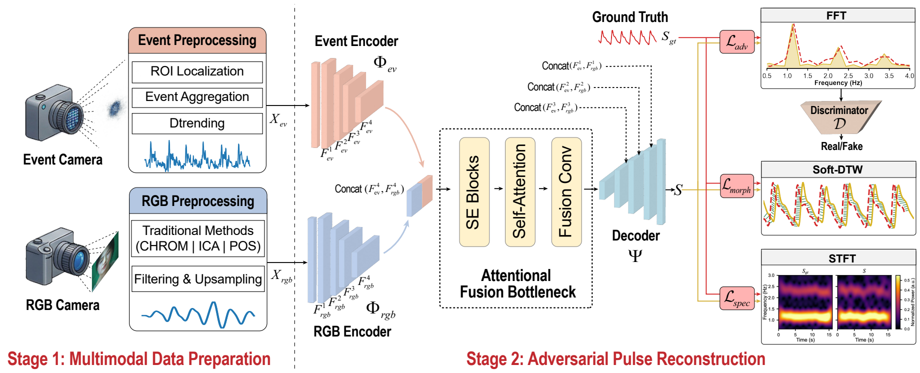

Fig. 1. The architecture of the proposed Fusion-E2Pulse framework.

stage pipeline consisting of multimodal data preparation and adversarial pulse reconstruction, as illustrated in Fig. 1. We formulate pulse wave reconstruction as a multimodal sequence-to-sequence generation task, where the objective is to recover the underlying PPG signal (S_{gt})—serving as the ground truth recorded via a fingertip pulse oximeter—from heterogeneous sensor data. By bridging the gap between global stability and local precision, the framework transforms raw sensor captures into high-fidelity physiological waveforms.

2.1 Multimodal Data Preparation

The raw input modalities comprise asynchronous event streams captured from the radial artery of the wrist by an event camera and recorded facial RGB video sequences. To handle modal heterogeneity and prepare the data for the generative network, we apply distinct preprocessing pipelines to derive the final processed signals $X_{ev} \in \mathbb{R}^{1\times T}$ and rPPG signals $X_{rgb} \in \mathbb{R}^{1\times T}$, where T denotes the time window length. Unlike traditional contact-based sensors, our framework extracts physiological information from the spatio-temporal pixel variations within video sequences and sparse event spikes, leveraging their unique sensing mechanisms to ensure robust reconstruction.

For the event stream, we first apply Region of Interest (ROI) localization followed by temporal aggregation to convert asynchronous events into a synchronous continuous signal matching the sampling frequency of the ground-truth PPG (60Hz), with detrending applied to remove motion-induced drifts. Simultaneously, for the RGB stream, rPPG waveforms are extracted from video via traditional rPPG algorithms (CHROM [5], ICA [20], and POS [26]), which leverage the specific absorption characteristics of oxygenated hemoglobin in the visible spectrum. These signals are band-pass filtered and linearly interpolated for temporal alignment. As visualized in Fig. 2, the resulting processed signals X_{ev} and X_{rgb} exhibit significant morphological complementarity: X_{ev} displays rich fine-grained dynamics that acutely capture subtle vascular pulsations, whereas

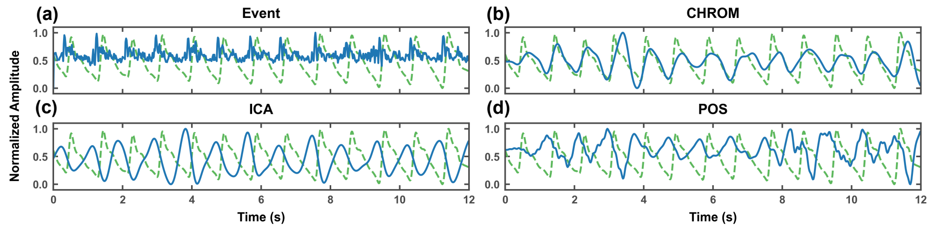

Fig. 2. Visualization of synchronized multimodal inputs. Inputs are overlaid with the ground-truth PPG signal (green dashed). (a) Event signal. (b)–(d) RGB signals extracted via CHROM, ICA, and POS.

X_{rgb} maintains a robust low-frequency periodic structure but is overly smoothed due to the integral imaging effect. Finally, both signals are Z-score standardized [15,1] to facilitate stable fusion.

2.2 Adversarial Pulse Reconstruction

Dual-Stream Encoder. To extract complementary features, the encoder employs a dual-stream architecture. The event branch (Φ_{ev}) extracts fine-grained local features F_{ev}^i via cascaded 1D convolutions, where i is the layer index. Complementarily, the RGB branch (Φ_{rgb}) extracts low-frequency periodic structural features F_{rgb}^i. Both branches output semantic features F_{ev}^L and F_{rgb}^L at the deep bottleneck (L being the total layers) to avoid early fusion interference. This late-fusion design follows a common strategy in multimodal perception, where modality-specific encoders are used to preserve heterogeneous cues before adaptive fusion [13,14].
Attentional Fusion Bottleneck. To address modal heterogeneity, we introduce Squeeze-and-Excitation (SE) [10] and Self-Attention (SA) [23] mechanisms. Features are first concatenated as $F_{cat} = \text{Concat}(F_{ev}^L, F_{rgb}^L)$. An SE module then dynamically calibrates channel weights to produce F_{se}, allowing adaptive weighting of event and RGB features based on signal quality. Subsequently, SA captures global temporal dependencies to yield F_{sa}. Finally, a fusion convolution (1×1 Conv) reduces the channel dimensionality to produce the final fused feature F_{fused} for the decoder.
Decoder. The decoder (Ψ) upsamples the fused feature F_{fused} via transposed convolutions to restore temporal resolution. Crucially, skip connections link the shallow features (F_{ev}^i, F_{rgb}^i) from the encoders to the corresponding decoder layers. This fuses texture details with semantic information, finally generating the reconstructed pulse waveform $S = \Psi(F_{fused})$.
Discriminator. Unlike standard GANs that operate directly on time-domain signals, our discriminator $\mathcal{D}$ transforms the waveform into the frequency domain via Fast Fourier Transform (FFT). It analyzes the magnitude spectrum of the signal, which allows the network to explicitly distinguish the fundamental HR frequency and harmonic structures from noise. This spectral discrimination strat-

egy guides the generator to synthesize realistic physiological periodicity without over-constraining the temporal phase.

Optimization Objectives. To ensure consistency in both time-domain topology and frequency-domain distribution, we train our network with a composite loss: $\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{morph}\mathcal{L}_{morph} + \lambda_{spec}\mathcal{L}_{spec}$. We employ an adversarial loss ($\mathcal{L}_{adv}$) based on Wasserstein GAN [9] with gradient penalty to constrain the distribution of the generated pulse waves. To ensure morphological fidelity and robustness against phase shifts, we utilize the Soft-DTW loss ($\mathcal{L}_{morph}$) [4], which provides a measure of temporal alignment. Finally, a time-frequency spectral loss ($\mathcal{L}_{spec}$) based on Short-Time Fourier Transform (STFT) [3] is applied to suppress noise and enforce consistency in the spectrogram domain.

3 Experiment and Results

3.1 Datasets and Implementation Details

To comprehensively evaluate Fusion-E2Pulse, we utilize the EMPD [6] dataset, comprising 193 multimodal recordings (68s each) from 83 subjects, including synchronized facial RGB, wrist-pulse events, and ground-truth PPG. We adopt two evaluation protocols: (1) time split, partitioning each 68s record into training (48s), validation (8s), and testing (12s) segments to assess temporal continuity;

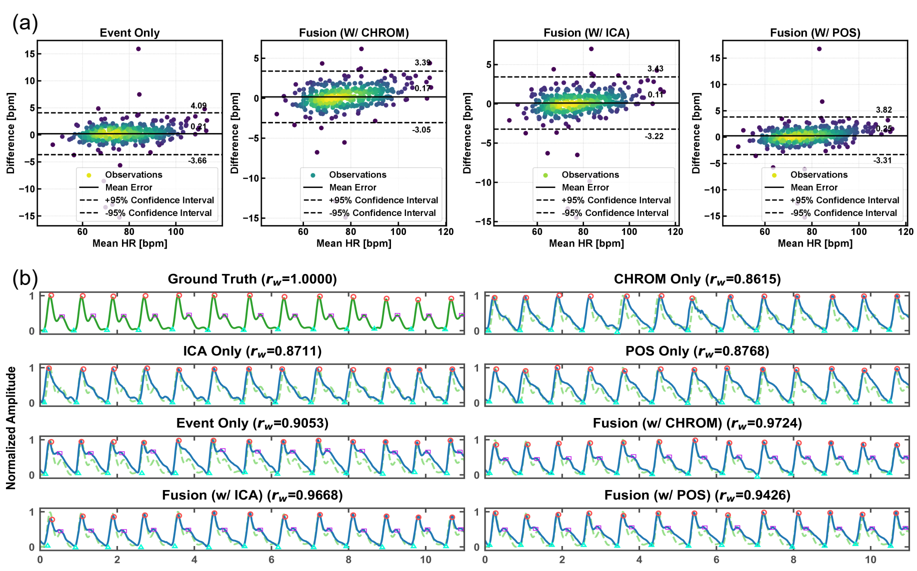

Fig. 3. **(a)** Bland-Altman plots showing the agreement between the estimated and ground-truth HR. The plots display the difference (y-axis) against the mean (x-axis), with dashed lines indicating the 95% Limits of Agreement (LoA); **(b)** Waveform morphology analysis.

Table 1. Quantitative comparison of the proposed Fusion-E2Pulse against its single-modal counterparts.

Method	HR				Morphology				
	MAE	**RMSE**	**MAPE**	r_h	**SPD**	**DPD**	**PWD**	**SNR**	r_w
Time Split									
Event	0.87	2.68	1.18	0.977	22.50	29.22	17.93	11.46	0.883
CHROM	5.10	9.50	6.16	0.713	28.49	60.04	64.50	6.41	0.543
ICA	6.63	11.36	8.31	0.543	22.67	69.47	73.41	6.15	0.498
POS	5.74	10.27	6.98	0.656	26.26	58.84	65.64	6.40	0.498
w/ CHROM	0.82	**2.51**	1.14	**0.980**	**16.07**	**19.24**	**15.98**	**11.93**	0.894
w/ ICA	0.89	2.65	1.26	0.978	18.52	21.85	18.82	11.81	**0.895**
w/ POS	**0.78**	**2.51**	**1.05**	**0.980**	16.74	21.85	17.93	11.89	0.891
Subject-Independent Split									
Event	0.86	2.76	1.25	0.976	21.11	28.69	17.91	11.97	**0.919**
CHROM	5.11	8.59	7.14	0.729	24.56	59.12	64.47	5.55	0.484
ICA	7.38	11.53	10.34	0.517	22.52	77.91	80.76	5.42	0.428
POS	5.39	9.23	7.33	0.688	28.04	66.08	67.38	5.74	0.464
w/ CHROM	0.76	2.50	1.10	0.978	18.28	22.85	**16.71**	**12.84**	0.918
w/ ICA	**0.73**	**2.43**	**1.02**	**0.980**	17.56	**22.18**	17.41	12.76	**0.919**
w/ POS	0.76	2.44	1.07	**0.980**	**17.44**	23.48	**17.06**	12.41	0.913

and (2) subject-independent split, partitioning the 83 subjects (193 records) into training (58 subjects, 136 recordings, 70.5%), validation (8 subjects, 20 recordings, 10.4%), and testing (17 subjects, 37 recordings, 19.2%) sets to strictly prevent data leakage and evaluate generalization across unseen individuals. Experiments are implemented in PyTorch on an NVIDIA RTX 4060 GPU using the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) with an initial learning rate of 2×10^{-4}. Loss weights are set to $\lambda_{morph} = 1.0$ and $\lambda_{spec} = 10.0$ to balance constraints. Evaluation metrics for heart rate estimation accuracy include Mean Absolute Error (MAE, in *bpm*), Root Mean Square Error (RMSE, in *bpm*), Mean Absolute Percentage Error (MAPE, in %), and the Pearson coefficient (r_h). For morphological fidelity, we evaluate the waveform Pearson coefficient (r_w), Signal-to-Noise Ratio (SNR, in *dB*), and the Systolic/Diastolic Phase Duration (SPD/DPD) and Pulse Width (PWD) measured in *ms* [22,12].

3.2 Comparative Analysis with Single-Modal Counterparts

We benchmark Fusion-E2Pulse against various configurations to validate the efficacy of multimodal integration. These counterparts include: (1) event-only, a single-modal network variant (denoted as Event in our experimental tables) utilizing only wrist-pulse event streams as input without RGB guidance; and (2) RGB-only: pulse signals extracted solely from facial video sequences via classic rPPG algorithms, including CHROM [5], ICA [20], and POS [26]. The proposed fusion variants, denoted as w/ CHROM, w/ ICA, and w/ POS, represent the full Fusion-E2Pulse architecture integrating event micro-dynamics with the corresponding structural priors derived from these RGB-based methods. This setup allows for a direct assessment of how different structural references influence the reconstruction of morphological details.

Performance on Time Split. To evaluate the fundamental reconstruction capability, we benchmark the models using the time split (Table 1). Fusion-E2Pulse (w/ POS) achieves the lowest MAE (0.78 bpm) while maintaining high waveform correlation ($r_w \approx 0.89$), significantly outperforming single-modal counterparts in overall reconstruction quality. Qualitatively, our method reconstructs sharp dicrotic notches (Fig. 3(b), $r_w \approx 0.97$) and demonstrates the tightest Limits of Agreement (LoA) in Bland-Altman analysis (Fig. 3(a)). Critically, the errors remain consistently low across the varying heart rate range shown on the x-axis, indicating robustness against physiological variations.

Generalization on Subject-Independent Split. To assess robustness against unseen subjects and strictly avoid data leakage, we evaluate the models using the subject-independent split (Table 1). RGB-only baselines remain markedly inferior on unseen subjects, whereas fusion variants consistently outperform the event-only network in HR and phase-duration errors. While the event-only model achieves an MAE of 0.86 bpm and an SPD error of 21.11 ms, the fusion variant (w/ ICA) achieves the best generalizability with an MAE of 0.73 bpm and a DPD error of 22.18 ms. This suggests that the RGB structural prior acts as a universal structural constraint, guiding the event branch to adapt to diverse physiological characteristics and ensuring high-fidelity reconstruction even for unseen individuals.

3.3 Ablation Study

Impact of Optimization Objectives. We analyze the contribution of each loss component (Table 2). The configuration trained solely with $\mathcal{L}_{adv}$ fails to recover meaningful pulse morphology, highlighting the difficulty of stabilizing GANs on unconstrained physiological signals. The introduction of $\mathcal{L}_{spec}$ proves critical by anchoring the fundamental frequency. Under the subject-independent split, configuring the model without $\mathcal{L}_{morph}$ (i.e., $\mathcal{L}_{spec} + \mathcal{L}_{adv}$) yields minor numerical improvements in HR error (e.g., Event MAE of 0.72 bpm) by strictly optimizing the spectral distribution. However, the absence of explicit time-domain constraints leads to a clear degradation in waveform morphology, evidenced by lower SNR (6.96 dB) and reduced correlation ($r_w = 0.810$). The full objective, integrating $\mathcal{L}_{morph}$ via Soft-DTW, provides necessary temporal alignment regularization. This integration substantially improves morphological fidelity—restoring Event SNR to 11.97 dB and r_w to 0.919—at a minor cost in HR MAE for the event-only setting (0.86 vs. 0.72 bpm), thereby providing a more comprehensive physiological reconstruction.

Robustness of Fusion Strategy. We further investigate the fusion effectiveness under the subject-independent split (Table 2). Consistent with time-split results, the inclusion of structural priors (CHROM, ICA, POS) generally improves

Table 2. Ablation study of Fusion-E2Pulse loss functions and architecture components. (Top) Impact of different optimization objectives under the time split. (Bottom) Robustness of fusion strategies and component contributions under the subject-independent split.

Setting	Method	HR				Morphology				
		MAE	RMSE	MAPE	r_h	SPD	DPD	PWD	SNR	r_w
Time Split										
$\mathcal{L}_{adv}$	Event	21.35	37.26	24.91	0.294	–	–	–	–	–
	w/ CHROM	29.27	43.97	35.95	0.154	–	–	–	–	–
	w/ ICA	26.14	40.37	32.37	0.072	–	–	–	–	–
	w/ POS	17.71	23.34	20.95	0.071	–	–	–	–	–
$\mathcal{L}_{spec}$ $+\mathcal{L}_{adv}$	Event	1.02	4.02	1.32	0.949	**16.50**	**20.32**	**16.70**	11.18	0.881
	w/ CHROM	**0.97**	**2.70**	1.31	**0.979**	18.90	25.52	21.36	**11.56**	0.883
	w/ ICA	1.09	3.61	1.50	0.958	19.31	23.51	17.90	11.31	**0.884**
	w/ POS	**0.97**	2.84	**1.29**	0.974	17.99	21.54	20.41	11.31	0.880
$\mathcal{L}_{morph}$ $+\mathcal{L}_{adv}$	Event	1.11	3.10	1.49	0.970	20.92	28.97	19.99	11.41	0.880
	w/ CHROM	**1.04**	**2.84**	**1.37**	**0.974**	18.63	25.48	19.55	11.24	0.875
	w/ ICA	1.22	3.63	1.59	0.958	18.53	**24.03**	**19.52**	11.28	0.876
	w/ POS	1.07	3.02	1.41	0.970	**18.50**	24.13	20.36	**11.77**	**0.882**
Full	Event	0.87	2.68	1.18	0.977	22.50	29.22	17.93	11.46	0.883
	w/ CHROM	0.82	**2.51**	1.14	**0.980**	**16.07**	**19.24**	**15.98**	**11.93**	0.894
	w/ ICA	0.89	2.65	1.26	0.978	18.52	21.85	18.82	11.81	**0.895**
	w/ POS	**0.78**	**2.51**	**1.05**	**0.980**	16.74	21.85	17.93	11.89	0.891
Subject-Independent Split										
$\mathcal{L}_{adv}$	Event	27.57	47.00	35.53	0.252	–	–	–	–	–
	w/ CHROM	24.81	39.80	32.55	0.055	–	–	–	–	–
	w/ ICA	16.71	27.24	21.53	0.055	–	–	–	–	–
	w/ POS	21.60	35.57	28.60	0.031	–	–	–	–	–
$\mathcal{L}_{spec}$ $+\mathcal{L}_{adv}$	Event	**0.72**	**2.37**	**1.02**	**0.980**	20.45	29.02	**15.55**	6.96	0.810
	w/ CHROM	0.78	2.53	1.11	0.978	**16.90**	**24.96**	17.96	**8.87**	**0.841**
	w/ ICA	0.78	2.69	1.12	0.975	19.19	27.21	16.81	7.49	0.826
	w/ POS	0.83	2.68	1.15	0.975	18.02	25.13	16.64	7.90	0.822
$\mathcal{L}_{morph}$ $+\mathcal{L}_{adv}$	Event	0.96	2.85	1.31	0.972	21.16	29.06	18.54	11.36	0.898
	w/ CHROM	0.92	2.89	1.24	0.972	**19.15**	25.36	17.56	11.70	0.903
	w/ ICA	**0.90**	2.85	**1.23**	0.972	20.97	28.90	17.70	**12.53**	**0.907**
	w/ POS	0.94	**2.82**	1.28	**0.973**	20.09	**24.64**	**17.37**	11.78	0.901
Full	Event	0.86	2.76	1.25	0.976	21.11	28.69	17.91	11.97	**0.919**
	w/ CHROM	0.76	2.50	1.10	0.978	18.28	22.85	**16.71**	**12.84**	0.918
	w/ ICA	**0.73**	**2.43**	**1.02**	**0.980**	17.56	**22.18**	17.41	12.76	**0.919**
	w/ POS	0.76	2.44	1.07	**0.980**	**17.44**	23.48	**17.06**	12.41	0.913

HR accuracy and morphological timing/SNR metrics over the event-only configuration. Specifically, the fusion variant (w/ ICA) improves the SNR from 11.97 dB (event-only) to 12.76 dB (and w/ CHROM reaches 12.84 dB). Notably, even without the full loss constraints, the structural prior provides essential regularization, preventing the event branch from overfitting to subject-specific noise. This confirms that the attentional fusion bottleneck successfully extracts complementary features—RGB for global stability and event for local dynamic precision.

4 Conclusion

In this work, we propose Fusion-E2Pulse, a novel multimodal framework synergizing neuromorphic event cameras and RGB video for high-fidelity pulse reconstruction. By integrating RGB structural priors with event micro-dynamics under a time-frequency adversarial supervision framework, our approach overcomes single-modal limitations such as exposure smoothing and noise. Comprehensive evaluations on the EMPD dataset under both time and subject-independent splits—with the latter reaching 0.73 bpm MAE, 0.919 waveform correlation, and 22.18 ms DPD error (w/ ICA prior)—demonstrate its capability in balancing accurate HR estimation with the high-fidelity recovery of fine-grained morphological biomarkers. Despite these results, we acknowledge current limitations including the single-dataset dependency and the necessity for further clinical studies to fully validate its diagnostic utility.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62403345), the Shanxi Provincial Department of Science and Technology Basic Research Project (202403021212174, 202403021221074).

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

References

1. Abdi, H.: Z-scores. Encyclopedia of measurement and statistics **3**, 1055–1058 (2007)
2. Adami, B., Karimian, N.: rppg-sysdiagan: Systolic-diastolic feature localization in rppg using generative adversarial network with multi-domain discriminator. In: European Conference on Computer Vision. pp. 193–210. Springer (2024)
3. Allen, J.: Short term spectral analysis, synthesis, and modification by discrete fourier transform. In: IEEE Trans. on Acoust., Speech, and Sig. Proc. vol. 4, pp. 21–24 (1997)
4. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: International conference on machine learning. pp. 894–903. PMLR (2017)
5. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE transactions on biomedical engineering **60**(10), 2878–2886 (2013)
6. Feng, Q., Li, P., Gao, R., Xu, J., Gong, R., Li, Y.: Empd: An event-based multimodal physiological dataset for remote pulse wave detection. arXiv preprint arXiv:2603.26699 (2026)

7. Figl, M., Rueckert, D., Hawkes, D., Casula, R., Hu, M., Pedro, O., Zhang, D.P., Penney, G., Bello, F., Edwards, P.: Registration of a 4d cardiac motion model to endoscopic video for augmented reality image guidance of robotic coronary artery bypass. In: Proceedings of AMI-ARCS workshop in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'08). pp. 122–129 (2008)
8. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020)
9. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
11. Jagtap, A., Saripalli, R.V., Lemley, J., Shariff, W., Smeaton, A.F.: Heart rate detection using an event camera. In: 2023 IEEE International Symposium on Multimedia (ISM). pp. 243–246. IEEE (2023)
12. Korpas, D., Halek, J., Doležal, L.: Parameters describing the pulse wave. Physiological research **58**(4) (2009)
13. Li, Y., Liu, H., Yang, B.: Stnet: Deep audio–visual fusion network for robust speaker tracking. IEEE Transactions on Multimedia **27**, 1835–1847 (2024)
14. Li, Y., Li, Y., Xu, Z., Guo, H., Liu, M., Wan, W.: Avclnet: Multimodal multispeaker tracking network using audio-visual contrastive learning. CAAI Transactions on Intelligence Technology **11**(1), 238–255 (2026)
15. Liu, X., Narayanswamy, G., Paruchuri, A., Zhang, X., Tang, J., Zhang, Y., Sengupta, R., Patel, S., Wang, Y., McDuff, D.: rppg-toolbox: Deep remote ppg toolbox. Advances in Neural Information Processing Systems **36**, 68485–68510 (2023)
16. Lv, Y., Zhou, L., Liu, Z., Zhang, H.: Structural vibration frequency monitoring based on event camera. Measurement Science and Technology **35**(8), 085007 (2024)
17. McDuff, D.: Camera measurement of physiological vital signs. ACM Computing Surveys **55**(9), 1–40 (2023)
18. Nakamura, H., Lin, G., Mitsukura, Y.: A novel approach for pulse signals extraction and heart rate estimation with hilbert space orthogonalization using event-based camera. IEEE Access (2025)
19. Oronti, I.B., Haleem, M.S., Pecchia, L., Popham, T.: Optimizing classification of congestive heart failure using feature weight importance correlation. In: Meets Africa Workshop. pp. 24–31. Springer (2024)
20. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express **18**(10), 10762–10774 (2010)
21. 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)
22. Tong, Y., Huang, Z., Zhang, Z., Yin, M., Shan, G., Wu, J., Qin, F.: Detail-preserving arterial pulse wave measurement based biorthogonal wavelet decomposition from remote rgb observations. Measurement **222**, 113605 (2023)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

24. Wang, J., Shan, C., Liu, L., Hou, Z.: Camera-based physiological measurement: Recent advances and future prospects. Neurocomputing **575**, 127282 (2024)
25. Wang, P., Yang, M., Zhang, X., Wang, J., Wang, C., Jia, H.: Non-contact blood pressure monitoring using radar signals: A dual-stage deep learning network. Bioengineering **12**(3), 252 (2025)
26. Wang, W., Stuijk, S., De Haan, G.: A novel algorithm for remote photoplethysmography: Spatial subspace rotation. IEEE transactions on biomedical engineering **63**(9), 1974–1984 (2015)
27. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. arXiv preprint arXiv:1905.02419 (2019)
28. Zhao, Z.h., Zhou, Y., Zhang, S., Chen, S., Wen, C.b., Xu, Q., Li, W.h.: Pulsenet: Multi-task learning-based non-contact pulse condition diagnosis using multi-scale fusion and transformer. Knowledge-Based Systems **302**, 112333 (2024)
29. Zou, B., Guo, Z., Chen, J., Zhuo, J., Huang, W., Ma, H.: RhythmFormer: Extracting patterned rPPG signals based on periodic sparse attention. Pattern Recognit. **164**, 111511 (Aug 2025)