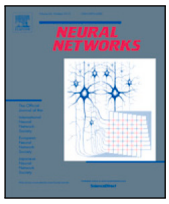
Neural Networks 185 (2025) 107128

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Spiking-PhysFormer: Camera-based remote photoplethysmography with parallel spike-driven transformer

Mingxuan Liu [a,1], Jiankai Tang [a,1], Yongli Chen [b], Haoxiang Li [a], Jiahao Qi [a], Siwei Li [a], Kegang Wang [a], Jie Gan [b], Yuntao Wang [a,c,*], Hong Chen [a,*]

[a] Tsinghua University, Beijing, China
[b] Beijing Smartchip Microelectronics Technology Co., Ltd, Beijing, China
[c] National Key Laboratory of Human Factors Engineering, Beijing, China

ARTICLE INFO

Keywords:
Brain-inspired neural networks
Remote photoplethysmography
Biomedical signal
Transformer

ABSTRACT

Artificial neural networks (ANNs) can help camera-based remote photoplethysmography (rPPG) in measuring cardiac activity and physiological signals from facial videos, such as pulse wave, heart rate and respiration rate with better accuracy. However, most existing ANN-based methods require substantial computing resources, which poses challenges for effective deployment on mobile devices. Spiking neural networks (SNNs), on the other hand, hold immense potential for energy-efficient deep learning owing to their binary and event-driven architecture. To the best of our knowledge, we are the first to introduce SNNs into the realm of rPPG, proposing a hybrid neural network (HNN) model, the Spiking-PhysFormer, aimed at reducing power consumption. Specifically, the proposed Spiking-PhyFormer consists of an ANN-based patch embedding block, SNN-based transformer blocks, and an ANN-based predictor head. First, to simplify the transformer block while preserving its capacity to aggregate local and global spatio-temporal features, we design a parallel spike transformer block to replace sequential sub-blocks. Additionally, we propose a simplified spiking self-attention mechanism that omits the value parameter without compromising the model's performance. Experiments conducted on four datasets—PURE, UBFC-rPPG, UBFC-Phys, and MMPD demonstrate that the proposed model achieves a 10.1% reduction in power consumption compared to PhysFormer. Additionally, the power consumption of the transformer block is reduced by a factor of 12.2, while maintaining decent performance as PhysFormer and other ANN-based models.

1. Introduction

The pulse wave is a vital sign of significant importance in healthcare, used to measure cardiac activity (O'Rourke, Pauca, & Jiang, 2001). However, measuring the pulse wave via contact PPG sensors is not convenient and may not be suitable for the elderly and preterm infants (Balakrishnan, Durand, & Guttag, 2013). Consequently, camera-based remote photoplethysmography (rPPG) has been proposed to predict the pulse wave through light reflected off the face, thereby obtaining heart rate (HR), respiratory rate (RR), and pulse transit time (PTT) (Sinhal, Singh, & Raghuwanshi, 2020). As illustrated in Fig. 1(a), the periodic movement of blood from the heart to the head through the abdominal aorta and carotid artery results in periodic motion of the facial color (Allen, 2007; Poh, McDuff, & Picard, 2010b). The rPPG detects the pulse from the movement without contact measurements, which allows for continuous monitoring outside clinical settings, providing doctors with timely samples, offering long-term trends and statistical analysis (Balakrishnan et al., 2013). In addition, the SARS-CoV-2 (COVID-19) pandemic has increased the demand for remote diagnostics. One issue of current telemedicine systems is that physicians are not able to assess patients' physiological status remotely. With rPPG technology, the problem can be solved (Liu, Narayanswamy, et al., 2023).

Early rPPG methods used traditional signal processing to analyze facial color changes (Balakrishnan et al., 2013; Li, Chen, Zhao, & Pietikainen, 2014; Poh, McDuff, & Picard, 2010a; Poh et al., 2010b; Tulyakov et al., 2016; Verkruysse, Svaasand, & Nelson, 2008; Wu et al., 2012). For example, Balakrishnan et al. (2013) tracked head features and used PCA to break down motions. Wu et al. (2012) introduced Eulerian Video Magnification for spatial decomposition and temporal filtering. However, these methods are limited by body movement and

* Corresponding authors.
E-mail addresses: yuntaowang@tsinghua.edu.cn (Y. Wang), hongchen@tsinghua.edu.cn (H. Chen).
[1] Mingxuan Liu and Jiankai Tang are co-first authors of the article.

https://doi.org/10.1016/j.neunet.2025.107128

Available online 10 January 2025

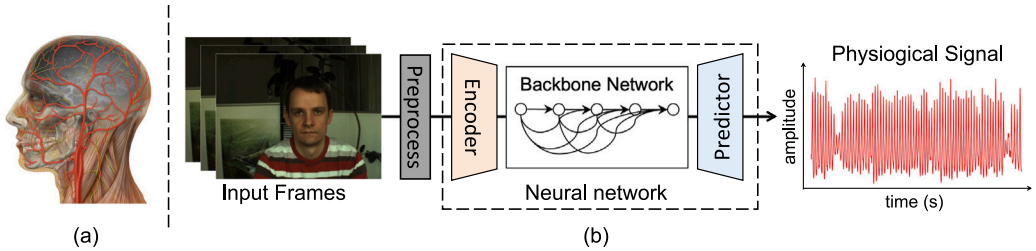


Fig. 1. (a) Human head anatomy with external and internal carotid arteries (Lynch, 2007) (b) rPPG pipeline of neural methods (Liu, Narayanswamy, et al., 2023).

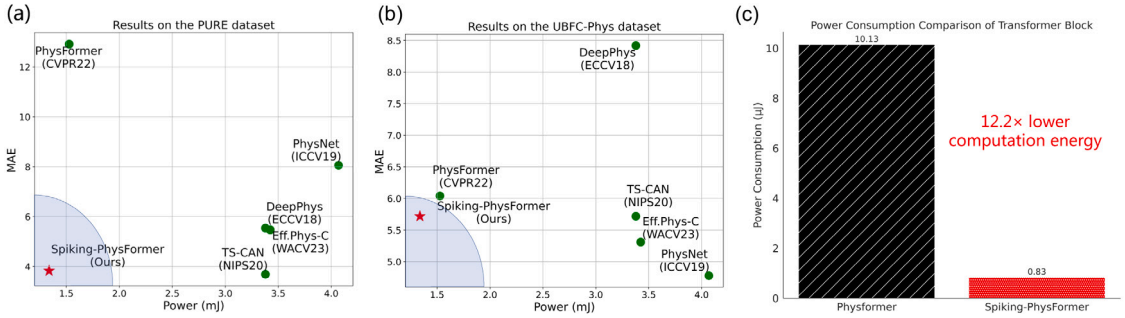


Fig. 2. MAE (lower is better) vs inference energy of different neural methods implemented in 45 nm technology (Horowitz, 2014) with an input frame size of 128 × 128, the shaded blue region shows the preferred region (The FLOPs for ANN-based models are derived from PhysBench (Wang, Wei, et al., 2023) and adjusted based on input size). (a) Results on the PURE dataset (Stricker, Müller, & Gross, 2014) after training on the UBFC-rPPG dataset (Bobbia, Macwan, Benezeth, Mansouri, & Dubois, 2019). (b) Results on the UBFC-Phys dataset (Sabour, Benezeth, De Oliveira, Chappe, & Yang, 2021) after training on the PURE dataset (Stricker et al., 2014) (c) The computational energy required by the transformer block in the Spiking-PhysFormer is 12.2 times lower than that in the PhysFormer (Yu et al., 2022).

lighting conditions, and may be biased in the presence of motion, lighting changes, and noise (McDuff, 2023).

With the advancement of deep learning, supervised learning methods based on artificial neural networks (ANNs) (shown in Fig. 1(b)) have been proposed to address the aforementioned problems (Chen & McDuff, 2018; Li, Yu, & Shi, 2023; Liu, Hill, Jiang, Patel, & McDuff, 2023; Shao, Luo, Chen, Hu, & Yang, 2023; Špetlík, Franc, & Matas, 2018; Yu, Peng, Li, Hong, & Zhao, 2019; Yu, Shen, et al., 2023; Yu et al., 2022; Yue, Shi, & Ding, 2023). These neural methods are mainly based on convolutional networks (CNNs) or transformers. The DeepPhys CNN by Chen and McDuff (2018) and the two-step CNN by Špetlík et al. (2018) are notable examples. Transformers, renowned for their exceptional long-range capabilities, have found applications in various domains such as natural language processing (NLP) and video analysis (Cao, Xu, Sun, Cheng, et al., 2023; Chen, Zhang, Pan, Xu, & Guan, 2023; Huang et al., 2023; Jiang, Cao, & Shen, 2023; Liu, Zhang, & Liu, 2023; Wang, Li, Zhang, Chi, & Dai, 2023; Zhang et al., 2024; Zong et al., 2023). Yu et al. (2022) introduced a video transformer, and Gupta, Kumar, Birla, and Gupta (2023) proposed RADIANT to enhance rPPG features. Despite their accuracy, ANNs, and specifically transformer models, necessitate considerable computational power. This requirement becomes particularly challenging for long-term remote photoplethysmography (rPPG) applications due to increased energy consumption. Additionally, the deployment of such models on edge computing devices is hindered by their limited memory and processing capabilities, making it difficult to support complex modules. Specifically, excessive computational complexity of models may render edge devices inoperable or result in high power consumption, whereas an excessive number of parameters can be constrained by limited memory. Therefore, finding a balance between performance and computational efficiency continues to be a significant challenge in this field (Zhuge, Wang, Xu, & Xu, 2023). In response to the challenge, a new approach emerges with spiking neural networks (SNNs).

As the third generation of neural networks, SNNs are designed with biological plausibility to encode and transmit information in the form of spikes, mimicking the dynamics of neurons in the brain Maass (1997), Wei et al. (2024), Zhang, Qu, Belatreche, Chen, and Yi (2018), Zhang, Wang, et al. (2021). Compared with ANNs, SNNs' event-driven nature enables a significant reduction in energy consumption when running on neuromorphic chips (Zhang, Huo, et al., 2023). For example, using a 28 nm asynchronous SNN accelerator, SNNs achieve an inference power efficiency of 3.97 pJ/SOP and a classification accuracy of 95.7% on the N-MNIST test dataset (Zhang, Liang, Wei, Wei, & Chen, 2021). Previously, SNNs were solely utilized for simple classification tasks (Wade, McDaid, Santos, & Sayers, 2010; Wu, Chua, Zhang, Li, & Tan, 2018; Yan, Zhou, & Wong, 2021), but with the introduction of complex SNN backbones in recent years (Fang et al., 2021; Wei et al., 2023; Yao et al., 2023; Zhou et al., 2022,), SNNs have been employed for intricate tasks such as image generation (Cao, Wang, et al., 2023; Liu, Gan, Wen, Li, Chen, & Chen, 2023; Watanabe, Mukuta, & Harada, 2023), optical flow estimation (Cuadrado, Rançon, Cottereau, Barranco, & Masquelier, 2023), image deblurring (Cao, Fu, Zhu, Sun, & Zha, 2023), and language generation (Zhu, Zhao, Li, & Eshraghian, 2023). However, the potential of SNNs in rPPG has not yet been explored.

In order to reduce the computational energy used by rPPG models, we introduce SNNs into this field for the first time by proposing a new hybrid neural network (HNN) called the Spiking-PhysFormer. As the encoding process required by SNNs to convert feature values into spike sequences, leading to information loss during feature extraction (Gerhards, Weih, Huang, Knobloch, & Mayr, 2023; Liu et al., 2024), we utilize ANNs in the feature extraction step to improve the accuracy and employs SNNs in constructing subsequent transformer blocks. Moreover, we adopt a spike self-attention mechanism to eliminate irrelevant information for pulse wave prediction. Specifically, Spiking-PhysFormer consists of a patch embedding (PE) block, transformer blocks, and a predictor head. To balance performance and energy efficiency, we adopt the PhysFormer design for the PE block and predictor head with ANNs, and specifically design the transformer blocks with SNNs. To enhance the transformer blocks with global spatio-temporal attention based on fine-grained temporal skin color differences, we propose a parallel spike-driven transformer, which combines temporal difference convolution (TDC) with spike-driven

self-attention (SDSA) mechanisms, executing multi-layer perceptron (MLP) and attention submodules in parallel to improve efficiency while minimizing performance degradation. Additionally, we introduce simplified spiking self-attention (S3A), omitting the value parameter, further reducing the complexity of the attention sub-block. The main contributions of this paper are listed:

(1) We introduce the Spiking-PhysFormer, an HNN that integrates SNNs with transformer architecture for efficient global spatio-temporal attention in rPPG models, featuring our innovative parallel spike-driven transformer and the simplified spiking self-attention (S3A) to reduce computational complexity. Spiking-PhysFormer represents the inaugural application of SNNs in rPPG signal analysis.

(2) Experiments on four datasets—PURE (Stricker et al., 2014), UBFC-rPPG (Bobbia et al., 2019), UBFC-Phys (Sabour et al., 2021), and MMPD (Tang et al., 2023)—show that Spiking-PhysFormer cuts energy use by 10.1% compared to PhysFormer. Its transformer block requires 12.2 times less computational energy (Fig. 2), while maintaining performance equivalent to PhysFormer (Yu et al., 2022) and other ANN-based models.

(3) Analysis of the spatio-temporal attention map based on spike firing rate (SFR) highlights Spiking-PhysFormer's capability to effectively capture facial regions in the spatial dimension. Furthermore, it demonstrates the model's ability to identify pulse wave peaks in the temporal dimension, verifying the interpretability of the proposed method.

2. Related work

2.1. Camera-based remote photoplethysmography

Camera-based rPPG is garnering increasing research interest due to its critical importance for telemedicine and remote health monitoring. DeepPhys (Chen & McDuff, 2018) is the first to demonstrate the superiority of deep learning over traditional signal processing algorithms such as POS (Wang, den Brinker, Stuijk, & de Haan, 2017) and ICA (Poh, McDuff, & Picard, 2011), hence current research is primarily focused on designing end-to-end models using ANNs. ANN-based methods are predominantly categorized into two types: CNN-based methods (Chaichulee et al., 2019; Chen & McDuff, 2018; Li, Yu, & Shi, 2023; Liu, Fromm, Patel, & McDuff, 2020; Liu, Wei, Kuang, & Ma, 2022; Yu et al., 2019) and Transformer-based methods (Gupta et al., 2023; Liu, Hill, et al., 2023; Yu, Shen, et al., 2023; Yu et al., 2022; Zhang, Xia, Liu, & Feng, 2023). Among these, CNNs represent the most prevalent form of supervised learning utilized for camera-based physiological measurement (McDuff, 2023). For example, a multi-task temporal shift convolutional attention network (TS-CAN) (Liu et al., 2020) was introduced to predict both PPG and breathing wave signals simultaneously. Another study (Chaichulee et al., 2019) utilized a CNN-based skin segmentation network prior to signal extraction. The CDCA-rPPGNet (Liu et al., 2022) incorporated an attention mechanism to fuse spatial and temporal features. To address the challenges of varying distance and head motion, two plug-and-play modules, namely the physiological signal feature extraction block (PFE) and the temporal face alignment block (TFA), were proposed (Li, Yu, & Shi, 2023). On the other hand, transformer-based rPPG research remains relatively unexplored, with PhysFormer (Yu et al., 2022) pioneering the integration of transformer blocks to amplify quasi-periodic rPPG features and refine spatio-temporal representation. PhysFormer++ (Yu, Shen, et al., 2023) extended this approach with two-pathway SlowFast architecture and additional temporal difference periodic and cross-attention transformers. Furthermore, EfficientPhys (Liu, Hill, et al., 2023) adopted transformer blocks as the network backbone, eliminating the need for preprocessing steps such as face detection, segmentation, normalization, and color space transformation. Indeed, ANN-based rPPG models offer high precision but come with significant computational demands. The multi-head self-attention (MHSA) mechanism in transformer blocks, in particular, involves complex matrix multiplications (Zhuge et al., 2023), posing challenges for deploying these models on edge devices.

2.2. Transformer-based spiking neural networks

Transformer-based ANNs have demonstrated remarkable success across various domains, such as natural language processing (NLP), computer vision, and audio processing (Cao, Xu, Sun, Gao, & Shen, 2023; Lin, Wang, Liu, & Qiu, 2022). However, the use of self-attention (SA) mechanisms in SNNs is still in its early stage. This is primarily due to the incompatibility of traditional SA mechanisms (Vaswani et al., 2017) with the computation characteristics of SNNs, which aim to avoid multiplicative operations and reduce computational overhead (Zhou et al., 2022). Recent research focused on addressing these challenges and exploring the integration of SA in SNNs for advanced deep learning (Mueller, Studenyak, Auge, & Knoll, 2021; Wang, Fang, Cao, Wang, & Xu, 2022; Wang, Fang, et al., 2023; Yao et al., 2021, 2023; Zhou et al., 2023, 2022; Zhu et al., 2023). For instance, Yao et al. (2021) proposed a temporal-wise attention SNN (TA-SNN) model to reduce redundant time steps. Mueller et al. (2021) proposed a method to convert ANN-transformer into SNN, but it still relies on the conventional vanilla self-attention (VSA) mechanism. In the groundbreaking work of Zhou et al. (2022), a novel spiking self-attention (SSA) mechanism was introduced, which models sparse visual features using spike-form Query, Key, and Value without the need for softmax. This marks the first transformer-based SNN. Subsequent enhancements were made to improve energy efficiency, such as the spike-driven self-attention (SDSA) in Yao et al. (2023). SDSA utilizes only mask and addition operations, completely avoiding multiplications and achieving up to 87.2 times lower computation energy compared to VSA. To further reduce energy consumption without sacrificing performance, Wang, Fang, et al. (2023) proposed the masked spiking transformer (MST) by incorporating the random spike masking (RSM) technique. Additionally, Liu, Xiao, Li, and Yu (2023) introduced SparseSpikformer, a co-design framework that aims to achieve sparsity in Spikformer through token and weight pruning techniques. Lastly, Wang, Zhao, Cui, Liu, and Xu (2023) represents the first neural architecture search (NAS) method specifically tailored for transformer-based models. The advancements have greatly enhanced the performance of SNNs, allowing them to be used in complex tasks, such as audio-visual classification, human pose tracking, and language generation (Guo et al., 2023; Li & Liu, 2023; Lv et al., 2023; Wang, Wu, et al., 2023; Zou, Mu, Zuo, Wang, & Li, 2023). In addition, the transformer-based SNNs have set the groundwork for our proposed Spiking-PhysFormer, which can capture long-range spatio-temporal attentional rPPG features from facial videos.

2.3. Hybrid neural networks

In recent years, there has been a growing interest in exploring the potential benefits of integrating ANNs and SNNs to achieve high-performance and low-power hybrid neural networks (HNNs). Various strategies have been investigated for combining these networks across different tasks. Specifically, SpikeGAN, proposed by Rosenfeld, Simeone, and Rajendran (2022), includes a conditional generator using an SNN and a discriminator using a conventional ANN. It addresses the problem of learning to emulate a spatio-temporal distribution and allows for a flexible definition of target outputs leveraging the temporal encoding nature of spiking signals. Another example is Spike-FlowNet, introduced by Lee et al. (2020), which enables energy-efficient optical flow estimation using sparse event camera data. Inspired by the auditory cortex, Gall, Kocanaogullari, Akcakaya, Erdogmus, and Kubendran (2023) proposed a CNN-SNN corticomorphic architecture, achieving auditory attention detection with an accuracy of 91.03% based on EEG data, while maintaining a low latency. For robot place recognition, Yu, Wu, et al. (2023) proposed a multimodal hybrid neural network (MHNN) that effectively encodes and integrates multimodal cues from conventional and neuromorphic sensors. They further deployed this MHNN on the Tianjic hybrid neuromorphic chip (Deng

et al., 2020) and integrated it into a quadruped robot. Furthermore, HNNs have been employed to achieve storage-efficient traffic sign recognition (Zhang, Xu, Huang & Chen, 2023) and efficient object detection in autonomous driving (Seras, Del Ser, & Garcia-Bringas, 2023). However, the application of HNNs in camera-based vitals measurement has received limited attention. To balance performance and energy efficiency in the rPPG task, we introduce the Spiking-PhysFormer. To the best of our knowledge, the proposed Spiking-PhysFormer is the first HNNs-based model in camera-based rPPG that includes a comprehensive evaluation of multiple public datasets.

3. The proposed method

The proposed Spiking-PhysFormer integrates transformer-based SNNs into the neural method of rPPG. We adopt the SNNs learning algorithms in SpikingJelly platform (Fang et al., 2023). The basic computational unit is the Leaky Integrate and Fire (LIF) neuron model (Stein & Hodgkin, 1967), which can emulate the behavior of biological neurons by generating discrete spikes and can be described by:

$$H[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1] - V_{reset})) \tag{1}$$

$$S[t] = \Theta(H[t] - V_{th}) \tag{2}$$

$$V[t] = H[t](1 - S[t]) + V_{reset}S[t] \tag{3}$$

where τ is the membrane time constant influencing the rate of potential change over time, $X[t]$ represents the synaptic input current at time step t signifying the cumulative input from connected synapses, and $H[t]$ is the neuron's membrane potential post charging and pre-spike, derived by integrating the input current. The spike occurrence at time t, denoted by $S[t]$, is determined by the Heaviside step function Θ, which outputs a spike (value of 1) when $H[t]$ surpasses the firing threshold V_{th}, indicating an action potential. The membrane potential after spiking, $V[t]$, is then updated to either remain at $H[t]$ if no spike occurs or reset to V_{reset}, reflecting the neuron's return to a baseline state post firing.

Above definitions and equations capture the dynamics of a LIF neural, where the membrane potential of neurons is updated based on the input current, and spikes are generated when the membrane potential surpasses a certain threshold. Due to the fact that the function $\Theta(x)$ used in Eq. (2) is non-differentiable, the surrogate gradient method is required. Specifically, we use the gradient g' of the arctangent function as a replacement for Θ' in order to facilitate training of SNNs using backpropagation:

$$g'(x) = \frac{\alpha}{2(1 + (\frac{\pi}{2}\alpha x)^2)} \tag{4}$$

3.1. Overall architecture

Fig. 3 illustrates the framework of Spiking-PhysFormer, which consists of three main components: ANN-based patch embedding (PE), parallel spike-driven transformer blocks, and an ANN-based predictor head. The PE block is utilized to extract rich spatio-temporal representations from the input video, while the simplified spiking self-attention (S3A) module in the transformer guides the model's attention towards key features. The final predictor head is responsible for mapping these features to pulse waveform peak signals.

Given an input RGB facial video, denoted as $X \in \mathbb{R}^{3\times T\times H\times W}$, where T, W, and H represent the sequence length, width, and height respectively, we begin with an initial preprocessing (IP) step to obtain preprocessed frames, denoted as $\hat{X} \in \mathbb{R}^{3\times T\times \hat{H}\times \hat{W}}$. The PE block, comprising four 3D convolutional layers and three max pooling layers, downsampling the input frames and partitions them into spatio-temporal tube tokens $X_{tube} \in \mathbb{R}^{D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}}$, where D represents the number of channels. To prepare the input for the parallel spike-driven transformer block, we use direct encoding (DE) (Bodo, Iulia-Alexandra, Yuhuang, Michael, & Shih-Chii, 2017) to replicate X_{tube} T_s times, resulting in the input U_0 of shape $\mathbb{R}^{T_t\times D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}}$. Therefore, the PE block is written as:

$$\hat{X} = \mathrm{IP}(X) \quad X \in \mathbb{R}^{3\times T\times H\times W}, \hat{X} \in \mathbb{R}^{3\times T\times \hat{H}\times \hat{W}} \tag{5}$$

$$X_{tube} = \mathrm{PE}(\hat{X}) \quad X_{tube} \in \mathbb{R}^{D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}} \tag{6}$$

$$U_0 = \mathrm{DE}(X_{tube}) \quad U_0 \in \mathbb{R}^{T_s\times D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}} \tag{7}$$

Subsequently, a LIF spike neuron (SN) is used to convert U_0 into S_0. The spike sequence S_0 is then passed to the parallel spike-driven transformer blocks, which consist of a simplified spiking self-attention (S3A) block and an MLP block. As the main component in Spiking-PhysFormer, S3A offers an efficient method to model the local–global information of videos using spike-form Query (Q), Key (K), and Value (V) without softmax. The outputs of the MLP and S3A blocks are summed together, and the sum is then added to the input again using residual connections. After L transformer blocks, the final output membrane potentials U_L are obtained. To get a concise representation, we compute the average of U_L along the spike temporal dimension, resulting in $U_{mean} \in \mathbb{R}^{D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}}$. Finally, we utilize a Predictor Head (PH) with temporal dimension upsampling blocks to map the features extracted by the transformer-based SNN into a 1D pulse wave $Y \in \mathbb{R}^T$. In summary, the output of S3A, MLP and predictor head blocks can be written as follows:

$$S_l = \mathcal{SN}(U_l) \quad S_l \in \{0,1\}^{T_s\times D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}}, l = 0...L \tag{8}$$

$$U_l = \mathrm{S3A}(S_{l-1}) + \mathrm{MLP}(S_{l-1}) + U_{l-1} \quad U_l \in \mathbb{R}^{T_s\times D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}}, l = 1...L \tag{9}$$

$$U_{mean} = \frac{1}{T_s}\sum_{t=1}^{T_s} U_L[t,:,:,:] \quad U_{mean} \in \mathbb{R}^{D\times \frac{T}{4}\times \frac{\hat{H}}{32}\times \frac{\hat{W}}{32}} \tag{10}$$

$$Y = \mathrm{PH}(U_{mean}) \quad Y \in \mathbb{R}^T \tag{11}$$

3.2. Data initial preprocessing

The proposed Spiking-PhysFormer model relies on the estimated PPG signal derived from subtle variations in facial skin color caused by the cardiac pulse cycle. Therefore, it is critical to preprocess the signal to ensure accurate face detection in each frame. In line with established techniques (Liu, Narayanswamy, et al., 2023), we adopt the straightforward Haar cascade detector (Sharifara, Rahim, & Anisi, 2014). Then, we utilize DiffNormalized, a method that calculates the difference between consecutive frames and labels, and then normalizes them based on their standard deviation. While some approaches, such as MSTMap (Niu et al., 2020), employ regions of interest (ROI) segmentation and landmark detection to extract crucial regions from detected faces and enhance prediction accuracy, our S3A module incorporates a spatio-temporal attention mechanism that automatically focuses on salient regions, rendering the aforementioned steps superfluous.

3.3. Parallel spike-driven transformer

Spike coding and decoding. To bridge the SNN block with the ANN blocks before and after SNN block, it is necessary to perform spike coding at the beginning of the transformer and decoding at the tail. The parallel spike-driven transformer blocks operate over spikes $s \in \{0,1\}$, i.e. temporal binary values, so we perform spike coding on the features extracted by the PE block. Previous approaches employed temporal coding (Zhang, Zhou, Zhi, Du, & Chen, 2019) and Poisson coding, which may bring information loss, particularly when the number of spike time steps decreases, leading to further loss of features. Although direct coding can preserve all feature representations, it brings the loss of sparsity in the computation of Q, K, and V in the first transformer block. To address the problem, we introduce a fixed-parameter LIF neuron before the first S3A sub-block and MLP sub-block. The features

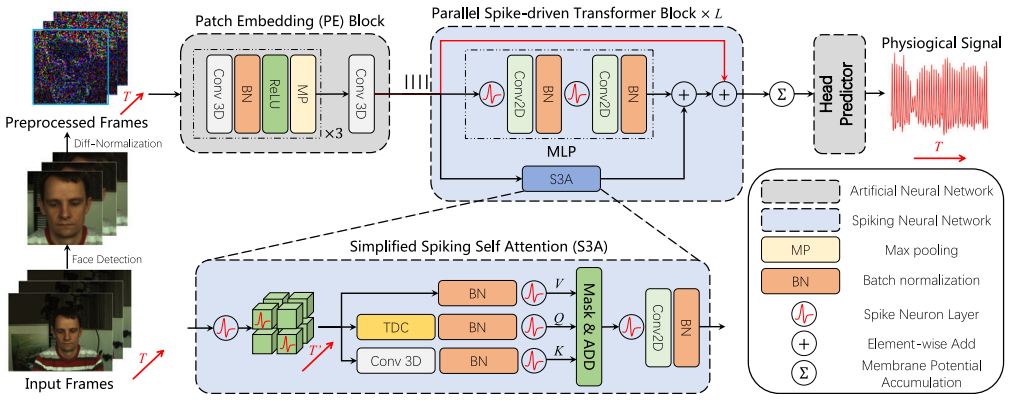


Fig. 3. Framework of the Spiking-PhysFormer. It consists of an ANN-based patch embedding (PE) block, several parallel spike-driven transformer blocks, and an ANN-based predictor head. The icon above the arrow between the PE and the Parallel spike-driven transformer blocks represents direct encoding of the output from the PE block. For the ANN-based components of our model, we follow the network structure in PhysFormer (Yu et al., 2022). Additionally, we initialize our model by pretraining PhysFormer and extracting the weights of the PE block as pre-trained parameters.

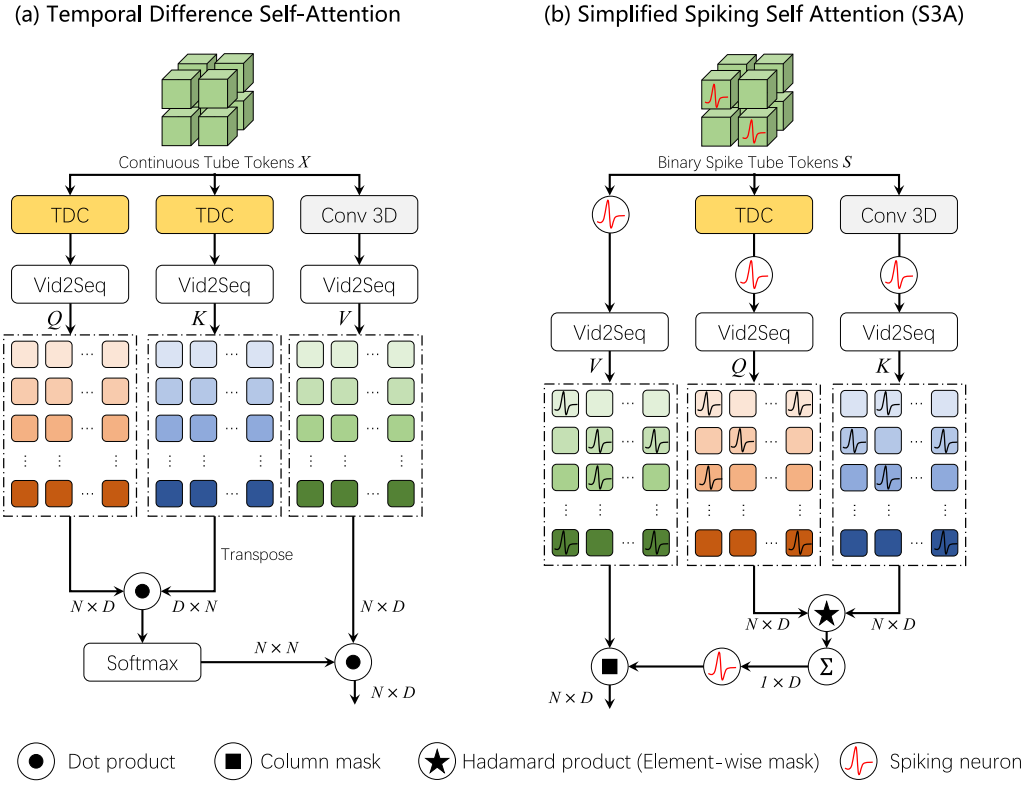


Fig. 4. Comparison temporal difference self-attention (TDSA) used in PhysFormer (Yu et al., 2022) and our simplified spiking self-attention (S3A). (a) In TDSA, Q, K, and V are obtained through linear projections using TDC (Yu, Li, Niu, Shi, & Zhao, 2020) and Conv3D. Since the input X is a floating-point matrix, this involves a significant amount of multiplication operations. Furthermore, the subsequent SA operation involves matrix multiplication, specifically requiring $2N^2D$ multiply-and-accumulate operations, where N is the number of tokens, D is the channel dimensions. (b) Compared with TDSA, S3A utilizes TDC exclusively for query computation. Additionally, since the input S is a binary spike sequence, the linear operation involved here is limited to addition. For SA computation, S3A employs an element-wise mask (Hadamard product), column summation, and column mask. As a result, only fND accumulate operations are required, where f represents the non-zero ratio of the matrix after applying the mask to Q and K. Typically, f is less than 0.06 (Fig. 6).

extracted by the PE block are expanded dimensionally using direct coding and connected to the LIF neuron. This spike coding method interprets the values of the feature vectors inputted at each time step as the current intensity reaches the neuron membrane. For spike decoding, we straightforwardly average the temporal dimension of the output of the final transformer block. As the spike coding and decoding processes described above do not affect the feasibility of backpropagation, the SNN component of Spiking-PhysFormer can be trained using Surrogate Gradient (SG) methods (Nunes, Carvalho, Carneiro, & Cardoso, 2022).

The loss functions of PhysFormer (Yu et al., 2022) is adopt, i.e.:

$$\mathcal{L}_{\text{overall}} = \alpha \cdot \mathcal{L}_{\text{time}} + \beta \cdot (\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{LD}}) \tag{12}$$

$$\beta = \beta_0 \cdot \eta^{\frac{\text{Epoch}_{\text{current}}-1}{\text{Epoch}_{\text{total}}}} \tag{13}$$

where $\mathcal{L}_{\text{time}}$ represents the Negative Pearson loss, which captures temporal constraints, while $\mathcal{L}_{\text{CE}}$ reflects frequency constraints through frequency cross-entropy loss. $\mathcal{L}_{\text{LD}}$ is a label distribution loss used for learning the distribution in HR estimation.

Parallel sub-blocks. Previous Transformer-based SNNs (Liu et al., 2023; Wang, Fang, et al., 2023; Wang, Zhao, et al., 2023; Yao et al., 2023; Zhou et al., 2022; Zhu et al., 2023) adhered to the format of a standard transformer block, in which, the output U_{out} is derived from the input $U_{\text{in}} \in \mathbb{R}^{T_s \times N \times D}$ containing N tokens and dimension D using two sequential sub-blocks (one SA and one MLP) with residual connections:

$$U_{\text{out}} = \alpha_{\text{FF}} \hat{U} + \beta_{\text{FF}} \text{MLP}(\mathcal{SN}(\hat{U})) \tag{14}$$

$$\hat{U} = \alpha_{\text{SA}} U_{\text{in}} + \beta_{\text{SA}} \text{SA}(\mathcal{SN}(U_{\text{in}})) \tag{15}$$

where scalar gain weights α_{FF}, β_{FF}, α_{SA}, β_{SA} fixed to 1 by default. In our work, to simplify the traditional transformer block, we draw inspiration from the use of parallel blocks in GPT-J-6B by Wang and Komatsuzaki (2021) and He and Hofmann (2023), and remove the residual connections in the MLP sub-blocks, obtaining the following output:

$$U_{\text{out}} = \alpha_{comb} U_{\text{in}} + \beta_{\text{FF}} \text{MLP}(\mathcal{SN}(U_{\text{in}})) + \beta_{\text{SA}} \text{SA}(\mathcal{SN}(U_{\text{in}})) \tag{16}$$

with skip gain $\alpha_{comb} = 1$, and residual gains $\beta_{\text{FF}} = \beta_{\text{SA}} = 1$ as default. Moreover, by parallelizing the processing of SA and MLP blocks, our model saves the computation time per layer. Additionally, the attention mechanism based on SFR explicitly provides crucial features (Section 4.5), so that parallel processing does not compromise model performance.

Simplified spiking self-attention (S3A). The S3A sub-block is a crucial module in Spiking-PhysFormer, designed to extract key information from the spatio-temporal features generated by the PE block. A comparison between S3A and the temporal difference self-attention (TDSA) in PhysFormer is illustrated in Fig. 4. To capture subtle differences in local temporal features, both S3A and TDSA utilize temporal difference convolution (TDC) (Yu et al., 2020):

$$\text{TDC}(x) = \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla 3D convolution}} + \theta \cdot \underbrace{(-x(p_0) \cdot \sum_{p_n \in \mathcal{R}'} w(p_n))}_{\text{temporal difference term}} \tag{17}$$

where w represents the learnable weight parameters, the variables p_0, $\mathcal{R}$, and $\mathcal{R}'$ represent the current spatio-temporal location, the sampled local neighborhood of size $(3 \times 3 \times 3)$, and the sampled adjacent neighborhood, respectively. TDC aggregates temporal difference clues within local temporal regions, but the computational cost is high due to multiple convolution operations. Specifically, for an input $S \in \mathbb{R}^{4 \times 96 \times 40 \times 4 \times 4}$, the FLOPs of TDC are approximately 28 times that of a vanilla 3D convolution layer (the result is obtained with the tool FlopCountAnalysis). In addition, we have applied DiffNormalized method (Chen & McDuff, 2018) to the input videos during the first preprocessing stage to extract inter-frame differences. Therefore, we utilize TDC exclusively for calculating the query (Q), and a vanilla 3D convolution layer for computing the key (K), leading to a significant reduction in computational cost.

To further simplify spiking self-attention, we remove the 3D convolution layer (W^V) required for the computation of the value (V). This is because W^V is simply a linear projection of the input sequence representation x, and the additional capacity provided by such a matrix is not particularly substantial. Previous studies (He & Hofmann, 2023) have shown that the value and projection parameters in the attention mechanism contribute minimally to the model's performance. The essential operations of self-attention—calculating relationships between different parts of the input—are primarily handled by the query (Q) and key (K) matrices. By setting W^V to the identity matrix or removing it entirely, the model maintains virtually the same performance while reducing the number of parameters and computational overhead. This leads to shorter training and inference times without sacrificing accuracy. Our ablation experiments (Section 4.6) confirm that removing these additional convolution layers does not significantly impact model performance.

As a result, Q, K, and $V \in \mathbb{R}^{T_s \times N \times D}$ are projected as:

$$Q = \text{Vid2Seq}(\mathcal{SN}(\text{BN}(\text{TDC}(S)))) \tag{18}$$

$$K = \text{Vid2Seq}(\mathcal{SN}(\text{BN}(\text{Conv3D}(S)))) \tag{19}$$

$$V = \text{Vid2Seq}(\mathcal{SN}(\text{BN}(S))) \tag{20}$$

To obtain V, we normalize the spike sequence S and input it into a LIF neuron for a second excitation, which decreases the sparsity of V, leading to less energy needs for SA computation (Section 4.3). Upon receiving Q, K, and V, the SA operation can be expressed as:

$$\text{S3A'}(Q, K, V) = g(Q, K) \otimes V = \mathcal{SN}(\text{SUM}_c(Q \otimes K)) \otimes V \tag{21}$$

$$\text{S3A}(Q, K, V) = \text{Seq2Vid}(\text{BN}(\text{Conv2D}(\mathcal{SN}(\text{S3A'}(Q, K, V))))) \tag{22}$$

where Vid2Seq and Seq2Vid respectively denote the processes of flattening a video's temporal, height, and width dimensions, and subsequently restoring them to their original dimensions. $\otimes$ represents the Hadamard product. The function $g(\cdot)$ is responsible for computing the attention map, and $\text{SUM}_c(\cdot)$ signifies the column-wise summation. Both $g(\cdot)$ and $\text{SUM}_c(\cdot)$ yield D-dimensional row vectors. The Hadamard product applied to spike tensors corresponds to the mask operation. Compared with TDSA, which necessitates $2N^2D$ multiply-and-accumulate operations, the proposed S3A framework requires much less computation because fND accumulation is merely needed when computing the attention map. Here, f denotes the sparsity factor, representing the proportion of non-zero elements in the matrix post-application of a mask to Q and K, N corresponds to the token count, and D indicates the channel dimensions. It is noteworthy that typically, f is less than 0.05, as illustrated in Fig. 6. In summary, the proposed S3A simplifies calculations relative to TDSA by primarily streamlining the computation of Q, K, and V, and leveraging additive operations instead of multiplicative ones in the attention mechanism to reduce computational load. Subsequent ablation experiments have demonstrated that the proposed method effectively balances performance and power consumption.

4. Experimental results

4.1. Datasets and performance metrics

Experiments of rPPG-based physiological measurement for pulse wave are conducted on four benchmark datasets (PURE (Stricker et al., 2014), UBFC-rPPG((Bobbia et al., 2019), UBFC-Phys (Sabour et al., 2021), and MMPD (Tang et al., 2023)). The examples of video frames in the aforementioned dataset are illustrated in Fig. 5.

PURE (Stricker et al., 2014): This dataset comprises vital sign measurements and video recordings from a cohort of 10 subjects, consisting of eight males and two females. Each participant engaged in six distinct recording sessions, encompassing a variety of motion conditions, thereby yielding a comprehensive dataset reflective of diverse physical states. Spatial configuration during these sessions was standardized, with subjects positioned at an approximate distance of 1.1 meters from the recording apparatus. Illumination was achieved through the strategic utilization of ambient natural light, which permeated through a window, ensuring consistent front-facing lighting conditions. Video data acquisition was facilitated through the deployment of an RGB eco274CVGE camera, procured from SVS-Vistek GmbH, featuring operational parameters including a 30 Hz frequency and a resolution specification of 640 × 480 pixels. For the purpose of establishing a robust gold-standard ground truth, photoplethysmography (PPG) and blood oxygen saturation (SpO2) levels were meticulously recorded at a frequency of 60 Hz, employing a CMS50E pulse oximeter affixed to each subject's finger.

UBFC-rPPG (Bobbia et al., 2019): This dataset encompasses a series of RGB video recordings from 42 subjects, derived from a scenario

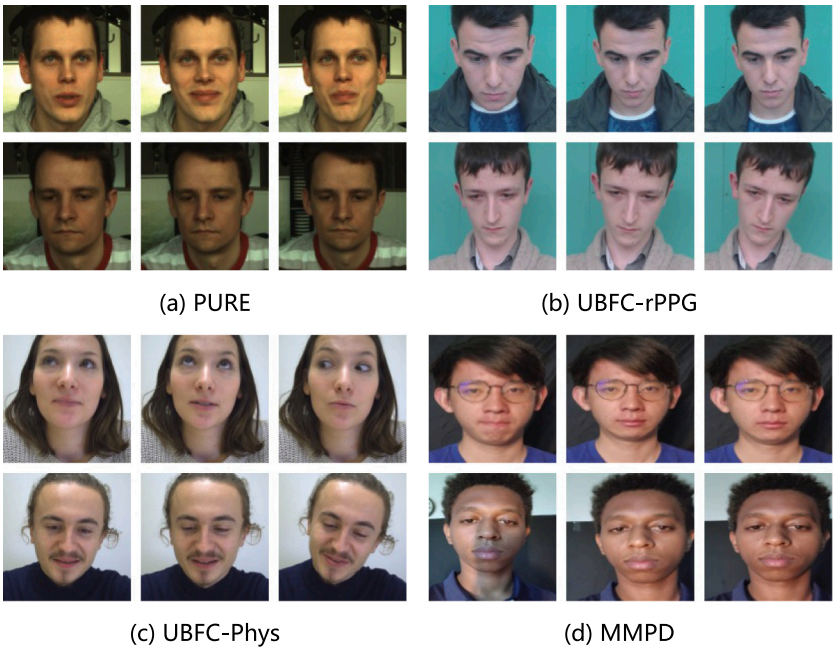


Fig. 5. Example video frames from datasets. (a) PURE (Stricker et al., 2014); (b) UBFC-rPPG (Bobbia et al., 2019); (c) UBFC-Phys (Sabour et al., 2021); and (d) MMPD (Tang et al., 2023).

wherein subjects engaged in a time-limited digital game, simulating typical activities performed in front of a computer. The participants were strategically positioned approximately one meter away from the recording device during these sessions. These sessions were conducted under indoor conditions, utilizing a blend of natural sunlight and artificial illumination to ensure optimal lighting. In terms of technical specifics, the videos were captured using a Logitech C920 HD Pro webcam, operating at a frequency of 30 Hz and delivering a resolution of 640 × 480. These recordings are preserved in an uncompressed 8-bit RGB format. Concurrently, reference photoplethysmography (PPG) data was meticulously acquired using a CMS50E transmissive pulse oximeter, providing a gold-standard dataset for validation purposes.

UBFC-Phys (Sabour et al., 2021): The dataset encompasses recordings from 56 subjects (46 women and 10 men), engaging in three distinct tasks characterized by considerable unconstrained motion under static lighting conditions. These tasks include a rest task, a speech task, and an arithmetic task, each designed to elicit varying physiological responses. The dataset is further enriched with gold-standard blood volume pulse (BVP) and electrodermal activity (EDA) measurements, meticulously captured using the Empatica E4 wristband. In terms of visual data, the recordings were executed using an EO-23121C RGB digital camera, ensuring high-resolution imagery at 1024 × 1024 pixels and a frame rate of 35 Hz. For evaluation purposes, we adhered to the same subject sub-selection list and task framework as outlined in the second supplementary material of Sabour et al. (2021).

MMPD (Tang et al., 2023): This dataset, meticulously assembled for comprehensive analysis, features 660 one-minute videos, showcasing a diverse array of 33 subjects (16 males and 17 females) engaging in four distinct activities (stationary, head rotation, talking, and walking) and additional exercise scenarios. The subjects, representing Fitzpatrick skin types 3–6 from multiple countries, were exposed to four different lighting conditions (LED-low, LED-high, incandescent, natural). For the purpose of validating physiological signals, ground truth photoplethysmogram (PPG) signals were concurrently captured using an HKG-07C+ oximeter at a sampling rate of 200 Hz, later downsampled to 30 Hz. The video data, concurrently recorded using a Samsung Galaxy S22 Ultra mobile phone, was initially captured at 30 frames per second with a resolution of 1280 × 720 pixels, before being compressed to 320 × 240 pixels.

Preprocess: The four datasets are processed using the pipeline described in Section 3.2 to generate video segments, each of which has the size of 128 × 128 × 160, and contains 160 frames. After preprocessing, the number of videos generated for the PURE, UBFC-rPPG, UBFC-Phys, and MMPD datasets are 750, 351, 3939, and 7260, respectively.

Metrics: We benchmark Spiking-PhysFormer against state-of-the-art (SOTA) ANN-based rPPG models: TS-CAN (Liu et al., 2020), PhysNet (Yu et al., 2019), DeepPhys (Chen & McDuff, 2018), Eff.Phys-C (Liu, Hill, et al., 2023), PhysFormer (Yu et al., 2022), iBVPNet (Joshi & Cho, 2024), rFaceNet (Zhu, Zhang, Zeng, Liu, Yang, & Zheng, 2024), DiffPhys (Chen, Wong, Chin, Chan, & So, 2024), PhysNet-XY (Cantrill, Ahmedt-Aristizabal, Petersson, Suominen, & Armin, 2024), and PhysNet-UV (Cantrill et al., 2024). For rFaceNet, DiffPhys, PhysNet-XY, and PhysNet-UV, we use the test results directly from the respective papers, which result in the absence of some experimental metrics. The remaining models are implement within the rPPG-Toolbox framework, ensuring consistent training and testing procedures for a fair comparison.

To evaluate the generalization ability of the rPPG model on out-of-distribution (OoD) data, we conduct cross-dataset testing on other three datasets after training on PURE or UBFC-rPPG. With the same settings of the rPPG-Toolbox (Liu, Narayanswamy, et al., 2023), we choose the UBFC-rPPG and PURE datasets for training because UBFC-Phys and MMPD datasets include lots of human motion and noise, which hinders model's convergence. Specifically, the UBFC-Phys dataset involves considerable unconstrained motion under static lighting conditions, while the MMPD dataset includes data from subjects speaking, walking, and exercising.

To quantitatively assess the predictive accuracy of various models, post-processing is applied to the outputted pulse wave signals. Specifically, the predicted waveforms are first filtered using a second-order Butterworth filter with cutoff frequencies of 0.75 and 2.5 Hz. Subsequently, the Fast Fourier Transform (FFT) is applied to the filtered signals to calculate the heart rates (HRs). The commonly used Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Pearson Correlation (ρ) are adopted as evaluation metrics:

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\left|R_g - R_p\right| \tag{23}$$

$$\text{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{R_g - R_p}{R_g}\right| \tag{24}$$

$$\rho = \frac{\sum_{n=1}^{N}\left(R_{g.n} - \overline{R_g}\right)\left(R_{p.n} - \overline{R_p}\right)}{\sqrt{\left(\sum_{n=1}^{N} R_{g.n} - \overline{R_g}\right)^2\left(\sum_{n=1}^{N} R_{p.n} - \overline{R_p}\right)^2}} \tag{25}$$

Where R_p, R_g, and N denote the predicted signal rate, the ground truth signal rate, and the number of instances, respectively. Additionally, $\overline{R}$ is the average value of R across N samples.

4.2. Implementation details

The proposed Spiking-PhysFormer is implemented with Pytorch and on two open-source platforms, SpikingJelly (Fang et al., 2023, https://github.com/fangwei123456/spikingjelly) and rPPG-Toolbox (Liu, Narayanswamy, et al., 2023, https://github.com/ubicomplab/rPPG-Toolbox), to facilitate fair comparisons. All videos in the datasets are resized to 128 × 128 by pre-preprocessing, and a sequence of 160 consecutive frames is randomly selected as input, denoted as $\hat{X} \in \mathbb{R}^{3\times160\times128\times128}$. Additionally, we set the number of channels, D, in the S3A to 96. We configure the timesteps T_s of the SNN module to 4 and utilize four parallel spike-drive transformer blocks in Spiking-PhysFormer. The LIF neuron model's surrogate gradient function is $g(x) = \frac{1}{\pi}\arctan(\frac{1}{\pi}\alpha x) + \frac{1}{2}$, and its derivative is $g'(x) = \frac{\alpha}{2(1+(\frac{\pi}{2}\alpha x)^2)}$, where α represents the slope parameter. For all neurons, $\alpha = 2$, $V_{reset} = 0$, and $V_{th} = 1$. Our model is trained with Adam optimizer and the initial learning rate and weight decay are 3e-3 and 5e-5, respectively. During the training phase, we first train the PhysFormer (Yu et al., 2022) for 10 epochs using a standard configuration in rPPG-Toolbox (Liu, Narayanswamy, et al., 2023) to obtain a pre-trained ANN-based PE block. Then, we train the Spiking-PhysFormer for 10 epochs, employing the model that performs the best on the validation set for cross-dataset testing. The batch size is set to 4, and the experiments are conducted on GeForce RTX 4090. In practice, we reshape the $Q, K, V \in \mathbb{R}^{T_s\times N\times D}$ (after Vid2Seq Yu et al., 2022) into multi-head form $q_i, k_i, v_i \in \mathbb{R}^{T_s\times N\times d}, i \in [1, \ldots, H]$, where $D = H \times d$. Next, we split Q, K, V into H parts and run H S3A operations, in parallel, which are called multi-head S3A (MHS3A):

$$Q = \left(q_1, q_2, \ldots, q_H\right), K = \left(k_1, k_2, \ldots, k_H\right), V = \left(v_1, v_2, \ldots, v_H\right) \tag{26}$$

$$\text{MHS3A'}(Q, K, V) = \left[\text{S3A'}_1\left(q_1, k_1, v_1\right); \cdots; \text{S3A'}_h\left(q_H, k_H, v_H\right)\right] \tag{27}$$

$$\text{MHS3A}(Q, K, V) = \text{Seq2Vid}(\text{BN}\left(\text{Conv2D}\left(\mathcal{SN}\left(\text{MHS3A'}(Q, K, V)\right)\right)\right)) \tag{28}$$

4.3. Energy consumption analysis

The FLOPs of the ANN module in the proposed Spiking-PhysFormer can be conveniently obtained using the tool THOP (https://github.com/Lyken17/pytorch-OpCounter). Now we need to analyze the energy consumption details of the SNN module. The FLOPs (the number of multiply-and-accumulate (MAC) operations) of ANN-based Conv3D layer (FL_{Conv3D}) and Conv2D layer (FL_{Conv2D}) are:

$$\text{FLOPs}_{Conv3D} = \left(k_n\right)^2 \cdot t_n \cdot h_n \cdot w_n \cdot c_{n-1} \cdot c_n \tag{29}$$

$$\text{FLOPs}_{Conv2D} = \left(k_n\right)^2 \cdot h_n \cdot w_n \cdot c_{n-1} \cdot c_n \tag{30}$$

where k_n is the kernel size, $\left(t_n, h_n, w_n\right)$ is the output feature map size, c_{n-1} and c_n are the input and output channel numbers, respectively. For the SNN-based convolution layer, the computation of theoretical energy consumption begins with the calculation of synaptic operations (SOPs) (Zhou et al., 2022), which is the number of spike-based accumulate (AC) operations:

$$\text{SOPs}_{Conv2D} = fr \cdot T_s \cdot \text{FLOPs}_{Conv2D} \tag{31}$$

$$\text{SOPs}_{Conv3D} = fr \cdot T_s \cdot \text{FLOPs}_{Conv3D} \tag{32}$$

where fr and T_s denote the spike fire rate and timesteps, respectively. The fr is defined as the proportion of non-zero elements within the spike tensor. It is positively correlated with power consumption because when SNNs operate on neuromorphic hardware, sparse computations are triggered only when input spike signals arrive; otherwise, neurons remain quiescent. Practically, we set T_s to 4. Once the FLOPs for the ANN module and the SOPs for the SNN module are determined, we can further compute the energy cost E:

$$E_{\text{FLOPs}} = E_{\text{MAC}} \times \text{FLOPs}, \quad E_{\text{SOPs}} = E_{\text{AC}} \times \text{SOPs} \tag{33}$$

Drawing from previous studies (Yao et al., 2023; Zhou et al., 2022), we assume that the MAC and AC operations are implemented in 32-bit floating point format using 45 nm technology (Horowitz, 2014), where $E_{\text{MAC}} = 4.6pJ$ and $E_{\text{AC}} = 0.9pJ$. When $E_{\text{AC}} \times T_s \times fr < E_{\text{MAC}}$, SNNs are more energy-efficient compared to their ANN counterparts. Additionally, due to T_s is a constant, the energy-saving characteristics of SNNs hinge on the spike firing rate fr.

Fig. 6 illustrates the average spike fire rate, fr, for each tensor within the transformer blocks. We find that the input tensor's spike fire rate (fr) displays an ascending trend but consistently stays below 0.25 (as indicated by the red line in Fig. 6), a consequence of the accumulative impact of inputs across layers, facilitated by residual connections. Conversely, the fr of Q, K, and V show a gradual decrease block by block, consistently staying below 0.06 (as indicated by the blue line in Fig. 6), which indicates that the spatio-temporal attention mechanism guides the network to filter out less significant features as the network depth increases.

In conclusion, as shown in Table 1, the computational complexity comparison reveals that the proposed Spiking-PhysFormer requires merely 1.34mJ to process a single-frame image of size 128 × 128. This represents a 10.1% energy reduction compared to the PhysFormer and is substantially lower than that of other baseline models.

4.4. Cross-dataset testing

Performance metrics, detailed in Table 2 and Fig. 7, are obtained by training on UBFC-rPPG and testing on three other datasets. Similarly, metrics in Table 3 and Fig. 8 are derived from training on PURE and subsequent testing on three other datasets. This cross-dataset testing verifies the model's adaptability to videos with diverse facial features, backgrounds, and illumination.

Results from training on UBFC-rPPG: As shown in Table 2, Spiking PhysFormer achieves performance comparable to the current state-of-the-art ANN-based methods on all three test datasets. Additionally, we can see that Spiking-PhysFormer significantly outperforms PhysFormer on the PURE dataset (MAE: 12.92 → 3.83), and exhibits comparable performance on UBFC-Phys and MMPD. This improvement can be attributed to the enhanced capability of the proposed spike-driven S3A to extract pivotal features, and the long-range spatio-temporal attention mechanism, which mitigates overfitting on in-distribution (ID) data. To further examine the correlations between the predicted HRs and the ground-truth HRs, we present Bland-Altman plots (Kaur & Stoltzfus, 2017) in Fig. 7. As demonstrated in the top two rows, Spiking-PhysFormer shows a strong correlation with the ground-truth HRs across a wide HR range of 40 to 140 bpm. Moreover, as observed from the output examples in the third row, the predictive challenge on the MMPD is considerably higher than on PURE and UBFC-Phys, which is due to the diversity in the MMPD dataset, where subjects engage in various types of physical activities and possess a range of skin tones, complicating the generalization of models pre-trained on UBFC-rPPG to this dataset.

Results from training on PURE: The Spiking PhysFormer demonstrates comparable predictive performance to the PhysFormer on MMPD, ranking third on UBFC Phys and outperforming PhysFormers in

Table 1
Computational complexity comparisons of rPPG methods on frames with size of 128 × 128 (The FLOPs and Params of ANN-based models are derived from PhysBench (Wang, Wei, et al., 2023) and adjusted based on input size). The FLOPs for Spiking-PhysFormer are contributed by the ANN components, while the SOPs are from the SNN components.

Model	Method	Input size	Params	Frame FLOPs	Frame SOPs	Power/mJ
TS-CAN	ANN	128 × 128	532K	717 M	\	3.30
PhysNet			770K	864 M	\	3.97
DeepPhys			532K	717 M	\	3.30
Eff.Phys-C			2.16M	727 M	\	3.34
PhysFormer			7.03M	324 M	\	1.49
iBVPNet			1.43M	868 M	\	3.99
Spiking-PhysFormer	HNN	128 × 128	2.99M	290 M[a]	3.7 M	1.34

[a] The FLOPs of Spiking-PhysFormer are calculated solely for the ANN-based PE block and the predictor head.

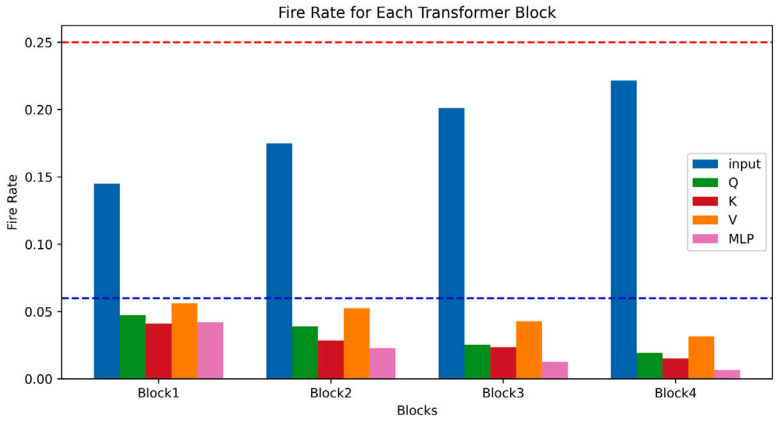


Fig. 6. Fire rate of input, Q, K, V and MLP layer of transformer blocks in Spiking-PhysFormer.

Table 2
Cross-dataset results training with the UBFC-rPPG dataset (Bobbia et al., 2019). MAE = Mean Absolute Error in HR estimation (Beats/Min), MAPE = Mean Percentage Error (%), ρ = Pearson Correlation in HR estimation. Best results are marked in red, second best in **bold**, and third best in underline.

Model	Method	Test set								
		PURE			UBFC-Phys			MMPD		
		MAE ↓	MAPE ↓	ρ ↑	MAE ↓	MAPE ↓	ρ ↑	MAE ↓	MAPE ↓	ρ ↑
TS-CAN	ANN	3.69	3.39	0.82	**5.13**	**6.53**	**0.76**	14.01	15.48	0.24
PhysNet		8.06	13.67	0.61	5.79	7.69	0.70	**9.47**	11.11	**0.31**
DeepPhys		5.54	5.32	0.66	6.62	8.21	0.66	17.50	19.27	0.06
Eff.Phys-C		5.47	5.40	0.71	4.93	6.25	0.79	13.78	15.15	0.09
PhysFormer		12.92	23.92	0.47	6.63	8.91	0.69	12.10	15.41	0.17
rFaceNet		**3.74**	\	0.86	\	\	\	\	\	\
DiffPhys		3.86	**4.07**	**0.84**	\	\	\	\	\	\
iBVPNet		11.74	21.99	0.53	5.82	7.73	0.66	9.23	**11.80**	0.35
Ours	HNN	3.83 ± 0.74	5.70 ± 0.74	0.83 ± 0.06	6.68 ± 0.82	8.33 ± 0.85	0.60 ± 0.08	14.15 ± 0.87	16.22 ± 0.96	0.15 ± 0.06

Table 3
Cross-dataset results training with the PURE dataset (Stricker et al., 2014). MAE = Mean Absolute Error in HR estimation (Beats/Min), MAPE = Mean Percentage Error (%), ρ = Pearson Correlation in HR estimation. Best results are marked in red, second best in **bold**, and third best in underline.

Model	Method	Test set								
		UBFC-rPPG			UBFC-Phys			MMPD		
		MAE ↓	MAPE ↓	ρ ↑	MAE ↓	MAPE ↓	ρ ↑	MAE ↓	MAPE ↓	ρ ↑
TS-CAN	ANN	1.30	1.50	0.99	5.72	7.34	**0.72**	13.94	15.14	**0.20**
PhysNet		0.98	1.12	0.99	**4.78**	6.15	0.73	**13.93**	15.61	0.17
DeepPhys		1.21	**1.42**	0.99	8.42	10.18	0.44	16.92	18.54	0.05
Eff.Phys-C		2.07	2.10	0.94	5.31	**6.61**	0.70	14.03	**15.31**	0.17
PhysFormer		1.44	1.66	0.98	6.04	7.67	0.65	14.57	16.73	0.15
rFaceNet		**1.05**	\	**0.99**	\	\	\	\	\	\
PhysNet-XY		\	\	\	\	\	\	14.91	\	0.15
PhysNet-UV		\	\	\	\	\	\	12.19	\	0.29
iBVPNet		5.21	5.10	0.83	4.51	7.88	5.80	15.78	17.06	0.04
Ours	HNN	2.80 ± 1.69	2.81 ± 1.46	0.95 ± 0.06	5.72 ± 0.75	7.17 ± 0.79	0.68 ± 0.07	14.57 ± 1.52	16.55 ± 1.34	0.14 ± 0.06

terms of performance. However, we find that the test performance on UBFC-rPPG is inferior to that of the ANN-based model, which may be attributed to the limited sample size of UBFC-rPPG, resulting in individual sample test errors having a disproportionate impact on the overall

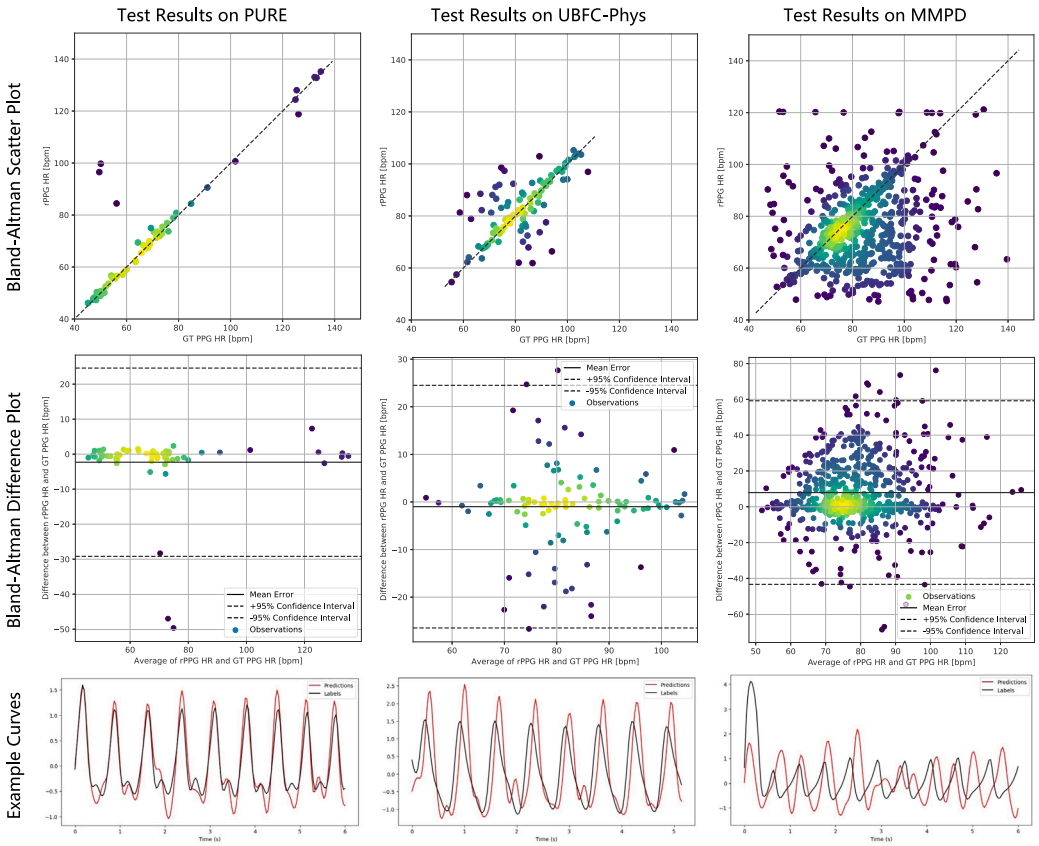


Fig. 7. Bland-Altman plots (Kaur & Stoltzfus, 2017) and output examples of cross-dataset results training with the UBFC-rPPG dataset (Bobbia et al., 2019).

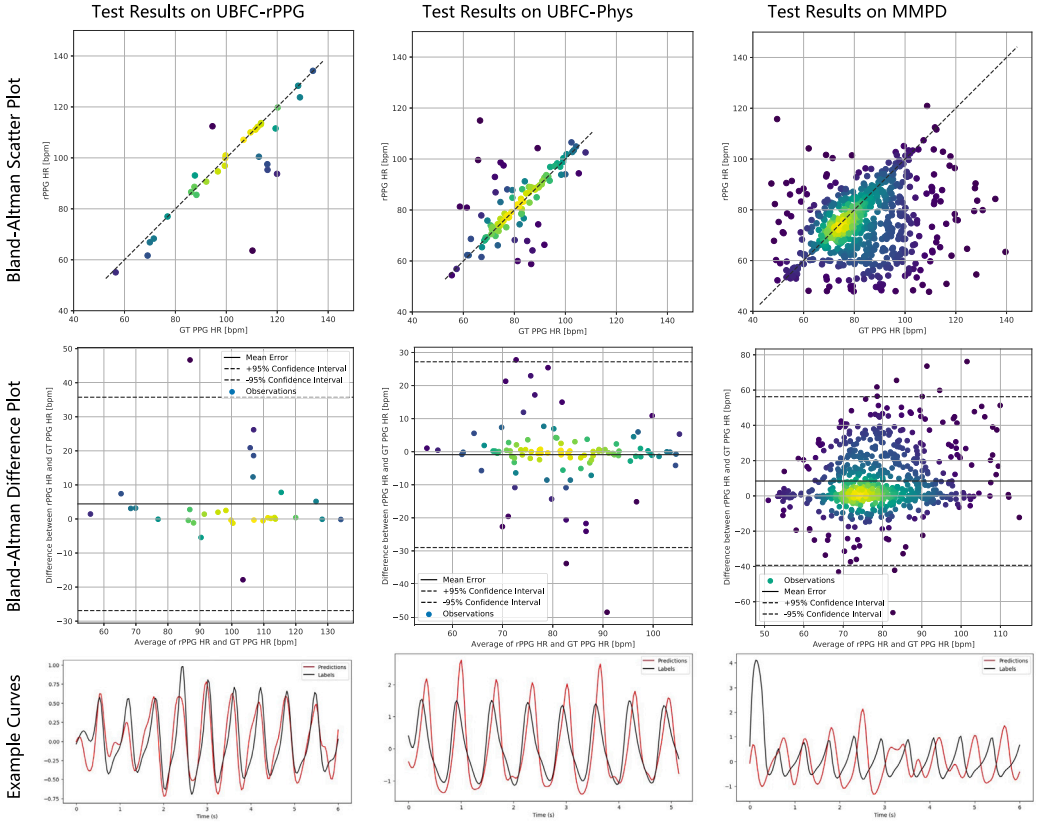


Fig. 8. Bland-Altman plots (Kaur & Stoltzfus, 2017) and output examples of cross-dataset results training with the PURE dataset (Stricker et al., 2014).

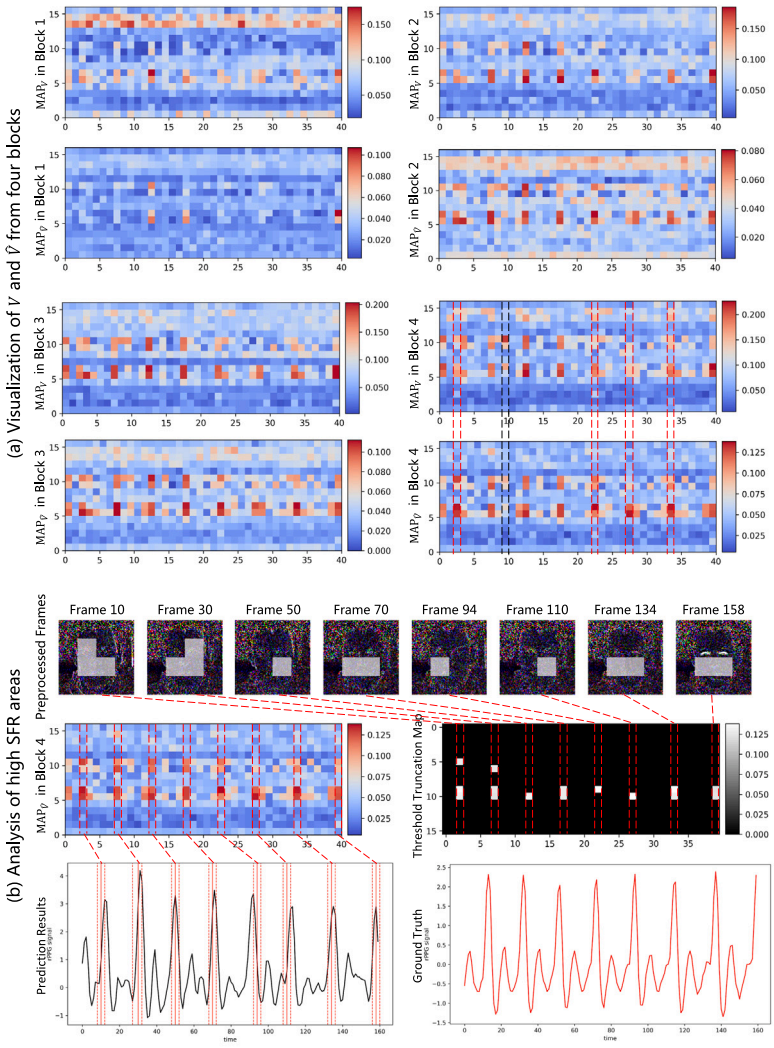


Fig. 9. Spatio-temporal attention map based on spike firing rate (SFR), the redder the higher the SFR, the bluer the smaller the SFR. (a) The visualization of V and $\hat{V}$ from the four parallel spike-driven transformer blocks reveals that irrelevant background signals and noise in V are masked by the hard attention mechanism. Consequently, $\hat{V}$ extracted contains critical spatio-temporal features. Furthermore, from the first block to the fourth, the key features within $\hat{V}$ become progressively more distinct. (b) Analysis of high SFR areas. By overlaying the high SFR areas from the threshold truncation map onto the input image, we find that the S3A mechanism directs the model's focus to the facial regions within the image. Moreover, it is observable that the areas with high SFR correspond to the peaks of the pulse wave in the temporal domain.

metrics. The Bland-Altman plot and output examples for UBFC-rPPG depicted in Fig. 8 reveal that Spiking-PhysFormer's predictions are consistent with the ground-truth HRs across the majority of samples.

To sum up, the proposed Spiking-PhysFormer demonstrates equivalent performance to the state-of-the-art ANN-based rPPG models with less power consumption (Section 4.3), which indicates that the Spiking-PhysFormer can balance efficiency and accuracy.

4.5. Spatio-temporal attention map

As described in Section 3.3, upon obtaining V, Q, and $K \in \mathbb{R}^{T_s \times N \times D}$, the proposed S3A is:

$$\hat{V} = \text{S3A'}(Q, K, V) = g(Q, K) \otimes V = \mathcal{SN}((Q \otimes K)) \otimes V \tag{34}$$

$$\text{S3A}(Q, K, V) = \text{Seq2Vid}(\text{BN}(\text{Conv2D}(\mathcal{SN}(\hat{V})))) \tag{35}$$

We define the output of S3A'(Q, K, V) as $\hat{V} \in \mathbb{R}^{T_s \times N \times D}$. Given that the output from $g(Q, K) = \mathcal{SN}(\text{SUM}_c(Q \otimes K))$ is a D-dimensional row vector composed exclusively of binary spike sequences, the S3A we propose qualifies as a hard attention mechanism (Serra, Suris, Miron, & Karatzoglou, 2018). S3A effectively masks the non-significant channels within V, ensuring that $\hat{V}$ retains only the essential spatio-temporal features, which guide the downstream modules for precise pulse wave prediction.

To demonstrate how the hard attention mechanism within S3A modulates the SFR of V, we present the visualization results in Fig. 9. To obtain the spatio-temporal attention map based on SFR, we first compute the SFR by averaging the spike and channel dimensions of V and $\hat{V}$. Subsequently, we unfold these averages across temporal and spatial dimensions to derive MAP_V and $\text{MAP}_{\hat{V}} \in \mathbb{R}^{\hat{T} \times \hat{H}\hat{W}}$, where $N = \hat{T} \times \hat{H}\hat{W}$.

As depicted in Fig. 9(a), we observe a progressive enhancement of high SFR regions within $\text{MAP}_{\hat{V}}$ from shallower to deeper blocks, indicating a refinement of key features. In the fourth block, highlighted by the red dashed lines, some low SFR areas in MAP_V become relatively high SFR regions in $\text{MAP}_{\hat{V}}$ after masking by the hard attention S3A mechanism. Conversely, the black dashed lines illustrate the opposite effect. Above findings confirm that the attention scores can modulate the spike firing in V.

To further validate the interpretability of the spatio-temporal attention in S3A, we correlated the high SFR regions in $\text{MAP}_{\hat{V}}$ with the input frames and the output pulse wave signals. The results are depicted in

Table 4
Impact of the number of transformer blocks on PURE dataset, PhysFormer and Spiking-PhysFormer are trained on UBFC-rPPG. Best results are marked in red.

Model	Blocks	Test on PURE		
		MAE ↓	MAPE ↓	ρ ↑
PhysFormer (Yu et al., 2022)	4	10.39	22.08	0.49
	6	9.12	15.61	0.60
	8	10.2	18.53	0.60
	10	11.05	18.52	0.49
	12	12.92	23.92	0.47
Spiking-PhysFormer	4	3.32	4.91	0.88
	6	3.72	7.36	0.89
	8	5.79	8.71	0.72
	10	2.90	5.47	0.91
	12	4.16	8.28	0.86

Fig. 9(b), where the threshold truncation map is the result of truncating $MAP_{\hat{V}}$ at a threshold of 0.1, preserving only the high SFR regions. When these regions are superimposed on the corresponding input frames, it is apparent that the high SFR areas are confined to the facial region. This feature enables the model to autonomously focus on changes in facial features without the need of an ROI mask. Additionally, Fig. 9(b) reveals that the high SFR regions correspond temporally with the peak values of the pulse wave, indicating that the attention-score-masked $\hat{V}$ can guide the downstream model to accurately predict the pulse wave and HR.

4.6. Ablation study

To demonstrate the effectiveness of Spiking-PhysFormer, we carry out ablation studies on different factors such as the number of transformer blocks, the impact of parallel sub-blocks, self-attention methods and the ANN components. All experiments were conducted with fixed random seed for training and testing, ensuring reproducibility across single runs.

Impact of block numbers. As illustrated in Table 4, we investigate the impact of the number of transformer blocks on model performance. Given that PhysFormer (Yu et al., 2022) is also a transformer-based rPPG model, we conduct a comparative study. Contrary to the conclusions of most transformer-based research (Yu et al., 2022; Zhuge et al., 2023), we find that for rPPG tasks, both PhysFormer and Spiking-PhysFormer models yield better performance with fewer transformer blocks. Specifically, PhysFormer achieves optimal performance with 6 layers, while performance is at its lowest with 12 layers, contradicting the findings of ablation studies by Yu et al. (2022), who suggested deeper transformer blocks for better performance. The discrepancy arises as we use cross-data testing to assess the models' generalization capabilities on OoD data, where an excessive number of transformer blocks can lead to overfitting. In contrast, Yu et al. (2022) tests their model on ID data. For the Spiking-PhysFormer, the best results are achieved with 10 layers; however, using 4 or 6 layers also yields comparable results (2.90 vs 3.32 or 3.72). Therefore, for rPPG tasks, fewer block layers can alse achieve generalization performance and significantly reduce the number of parameters.

Impact of timesteps of the SNN module. Timesteps play a crucial role in SNNs. Previous research indicates that larger timesteps can enhance the performance of SNNs, as longer spike sequences are capable of encoding more information (Cao, Wang, et al., 2023; Guo et al., 2023; Liu, Gan, et al., 2023; Yao et al., 2023). However, the proposed Spiking-PhysFormer is an HNN, in which the SNN module is engineered to direct the model's attention to crucial spatio-temporal features, thus fewer timesteps actually benefit the hard attention mechanism in filtering out background noise. The ablation study presented in Table 5 shows that Spiking-PhysFormer's performance remains stable with timesteps set to 1, 2, 4, and 8. A notable decrease in performance occurs when the timestep is increased to 16, as excessively long spike sequences introduce redundant information that may hinder the attention mechanism.

Impact of parallel sub-blocks. Parallelizing sub-blocks within the transformer block facilitates faster inference speeds on hardware (He & Hofmann, 2023; Wang & Komatsuzaki, 2021). Additionally, the proposed SFR-based attention map, which provides essential spatio-temporal features, is beneficial for predicting pulse wave signals when directly fed to the subsequent head predictor. As indicated in Table 6, our ablation study investigates the effects of configuring sub-blocks in PhysFormer and Spiking-PhysFormer to operate in parallel. The results demonstrate that for PhysFormer, parallel sub-blocks significantly enhance performance on the PURE dataset (MAE reduced from 12.92 to 9.90). For Spiking-PhysFormer, transitioning to a parallel arrangement improves results on PURE dataset. In summary, for rPPG tasks, parallel sub-blocks offer enhanced model generalizability and faster inference speeds. To further assess how parallelization affects training stability, we plotted the training loss for Spiking-PhysFormer and PhysFormer using both serial and parallel transformer blocks. As shown in Fig. 10, parallel sub-blocks improved the training stability of both models. More concretely, PhysFormer demonstrates faster convergence, while for Spiking-PhysFormer, training on the PURE dataset with serial blocks led to unstable fluctuations in loss values, manifesting as the loss spike (Li, Xu, & Zhang, 2023). In contrast, parallel blocks effectively mitigated this issue. Furthermore, although Spiking-PhysFormer exhibits a higher loss on the training set upon completion of training, it demonstrates comparable or superior performance in cross-dataset testing. This is particularly evident when trained on the UBFC-rPPG dataset and tested on the PURE dataset, indicating that Spiking-PhysFormer effectively mitigates overfitting.

Impact of the attention block. To reduce computational overhead without compromising performance, we simplify the attention block. The results of the ablation study are presented in Table 7. Because of the binary computation and sparsity in SNNs, the Spiking-PhysFormer's transformer block only needs 1.52 μJ of energy, marking a 6.7× decrease compared to the PhysFormer (Yu et al., 2022), even when using the same projection layers, such as TDC for computing Q and K, and a Conv3D layer for V. Besides, we simplify the attention layer by reducing the TDC layers, as the DiffNormalized preprocessing method is used to capture frame differences. The ablation study reveals that eliminating one TDC layer and the Conv3D layer used to compute V results in only a minor performance degradation (MAE increases from 2.49 to 3.32) while reducing power consumption by 45.4%. To sum up, we propose that adopting a TDC layer for computing Q, a Conv3D layer for K, and omitting the projection layer for V represents a viable strategy to balance performance with power efficiency, as delineated in the S3A mechanism.

Impact of the ANN components. To illustrate that the combined use of ANN and SNN yields the best performance, we presented further findings from an ablation study, as detailed in Table 8. When trained on UBFC-rppg, the pure SNN model achieved a MAE of 12.39, markedly lower than the HNN, which builds the transformer block using SNN while relying on ANN for the PE block and predictor.

Table 5
Impact of timesteps of the SNN module on PURE dataset, Spiking-PhysFormer is trained on UBFC-rPPG. Best results are marked in red.

Model	Timesteps	Power (mJ)	Test on PURE		
			MAE ↓	MAPE ↓	ρ ↑
Spiking-PhysFormer	1	1.334	2.56	4.13	0.92
	2	1.335	4.22	6.09	0.82
	4	1.337	3.32	4.91	0.88
	8	1.340	2.72	5.23	0.92
	16	1.347	7.41	14.54	0.76

Table 6
Impact of parallel sub-blocks on PURE dataset, PhysFormer and Spiking-PhysFormer are trained on UBFC-rPPG. Best results are marked in red.

Model	Sub-blocks	Test on PURE		
		MAE ↓	MAPE ↓	ρ ↑
PhysFormer (Yu et al., 2022)	w/o parallel	12.92	23.92	0.47
	parallel	9.90	17.66	0.68
Spiking-PhysFormer	w/o parallel	5.30	10.49	0.83
	parallel	3.32	4.91	0.88

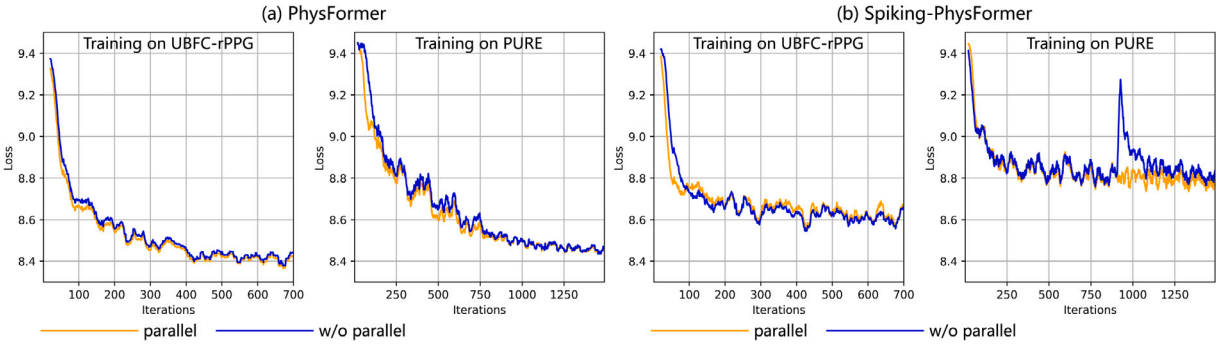


Fig. 10. Impact of parallel sub-blocks on the training convergence of PhysFormer and Spiking-PhysFormer.

Table 7
Impact of the attention block on PURE dataset, PhysFormer and Spiking-PhysFormer are trained on UBFC-rPPG. Best results are marked in red.

Model	Projection layers			Power[a] (μJ)	Test on PURE		
	Q	*K*	*V*		MAE ↓	MAPE ↓	ρ ↑
PhysFormer (Yu et al., 2022)	TDC	TDC	Conv3D	10.13	12.92	23.92	0.47
Spiking-PhysFormer	TDC	TDC	Conv3D	1.52	2.49	4.83	0.92
	TDC	Conv3D	Conv3D	0.85	4.42	8.50	0.87
	TDC	**Conv3D**	**None**	0.83	3.32	4.91	0.88
	Conv3D	Conv3D	None	0.16	4.77	7.15	0.80

[a] The power consumption required for a single transformer block.

5. Discussion

5.1. Limitations of the study

Although the proposed method in this paper balances performance and power consumption and has been comprehensively compared with existing methods to confirm its effectiveness, there are still some limitations. Firstly, our rPPG model is tested on a Linux platform, not in practical scenarios. Secondly, the model needs extensive testing across various conditions such as motion, skin tone, and lighting, which are critical for the algorithm's robustness in diverse environments. Addressing these problems is crucial for the future development of rPPG technology and its successful application in various real-world scenarios.

5.2. Impacts of the study

The advancement of rPPG technology holds promising potential for positive societal impacts, particularly in the context of remote healthcare and continuous physiological monitoring. As the technology matures, it is expected that rPPG can be integrated into endpoint devices, facilitating ongoing inference and health monitoring without interfering with the device's core functionalities. This integration could significantly reduce energy consumption during algorithm deployment, enhancing the feasibility of widespread use in medical and everyday settings. Such developments could revolutionize remote patient monitoring by providing real-time health data, thereby enabling timely medical interventions and promoting overall public health. Despite its potential benefits, the deployment of rPPG technology raises substantial concerns regarding privacy and ethical implications. One of the most pressing issues is the technology's ability to collect physiological data

Table 8
Impact of the ANN components, Spiking-PhysFormer are trained on UBFC-rPPG. Best results are marked in red.

Model	Component			Test on PURE		
	PE block	Transformer block	Predictor head	MAE ↓	MAPE ↓	ρ ↑
Spiking-PhysFormer	SNN	SNN	SNN	12.39	23.05	0.50
	ANN	SNN	SNN	9.56	12.02	0.43
	ANN	**SNN**	**ANN**	3.32	4.91	0.88

through monitoring cameras without the user's consent. This capability could lead to the leakage of sensitive personal health information, posing significant risks. Such data could be exploited by employers to assess employee productivity or even used as a basis for legal scrutiny. The potential for misuse of this sensitive information underscores the need for stringent ethical standards and robust privacy protections to be established as integral components of rPPG technology deployment.

6. Conclusions

In pursuit of a camera-based remote photoplethysmography (rPPG) solution that balances energy efficiency with performance, this paper introduces the Spiking-PhysFormer, a hybrid neural network (HNN) composed of an ANN-based patch embedding block and predictor head, alongside SNN-based parallel transformer blocks. To streamline the transformer block, we design a backbone consisting of parallel sub-blocks and introduce the S3A mechanism for extracting spatio-temporal key features, utilizing a single temporal difference convolution (TDC) layer to reduce the parameter count and computational expenditure. This marks the first incorporation of SNNs into the rPPG domain, injecting new perspectives and potential avenues of exploration. By generating spatio-temporal attention maps based on the spike firing rate (SFR), we validate the interpretability of our proposed model. Experiments across four datasets demonstrate that the Spiking-PhysFormer significantly reduces computational power consumption while maintaining predictive accuracy and generalization capabilities comparable to ANN-based models.

CRediT authorship contribution statement

Mingxuan Liu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Jiankai Tang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Yongli Chen:** Writing – review & editing, Validation, Formal analysis, Data curation. **Haoxiang Li:** Writing – review & editing, Validation, Methodology. **Jiahao Qi:** Validation, Methodology. **Siwei Li:** Writing – review & editing, Methodology. **Kegang Wang:** Visualization, Software, Methodology. **Jie Gan:** Writing – review & editing, Validation, Formal analysis. **Yuntao Wang:** Writing – review & editing, Supervision, Project administration, Data curation, Conceptualization. **Hong Chen:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-4 in order to polish the content. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62334014, 92164110, 62132010, and 62472244), the National Key Research and Development Program of China (Grant No. 2024YFB4505500), the Foundation of National Key Laboratory of Human Factors Engineering (Grant No. HFNKL2024W06), the Beijing Natural Science Foundation (No. QY24248), and the Tsinghua University Initiative Scientific Research Program. This work is partly supported by the Beijing Engineering Research Center (No. BG0149).

Data availability

Data will be made available on request.

References

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), R1.

Balakrishnan, G., Durand, F., & Guttag, J. (2013). Detecting pulse from head motions in video. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3430–3437).

Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., & Dubois, J. (2019). Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, *124*, 82–90.

Bodo, R., Iulia-Alexandra, L., Yuhuang, H., Michael, P., & Shih-Chii, L. (2017). Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, *11*, 682.

Cantrill, S., Ahmedt-Aristizabal, D., Petersson, L., Suominen, H., & Armin, M. A. (2024). Orientation-conditioned facial texture mapping for video-based facial remote photoplethysmography estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 354–363). URL https://api.semanticscholar.org/CorpusID:269148813.

Cao, C., Fu, X., Zhu, Y., Sun, Z., & Zha, Z. J. (2023). Event-driven video restoration with spiking-convolutional architecture. *IEEE Transactions on Neural Networks and Learning Systems*.

Cao, J., Wang, Z., Guo, H., Cheng, H., Zhang, Q., & Xu, R. (2023). Spiking denoising diffusion probabilistic models. arXiv preprint arXiv:2306.17046.

Cao, Y., Xu, X., Sun, C., Cheng, Y., Du, Z., Gao, L., et al. (2023). Segment any anomaly without training via hybrid prompt regularization. arXiv preprint arXiv:2305.10724.

Cao, Y., Xu, X., Sun, C., Gao, L., & Shen, W. (2023). Bias: Incorporating biased knowledge to boost unsupervised image anomaly localization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, http://dx.doi.org/10.1109/TSMC.2023.3344383.

Chaichulee, S., Villarroel, M., Jorge, J., Arteta, C., McCormick, K., Zisserman, A., et al. (2019). Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning. *Physiological Measurement*, *40*, URL https://api.semanticscholar.org/CorpusID:204965464.

Chen, W., & McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European conference on computer vision* (pp. 349–365).

Chen, S., Wong, K. L., Chin, J. W., Chan, T. T., & So, R. H. Y. (2024). DiffPhys: Enhancing signal-to-noise ratio in remote photoplethysmography signal using a diffusion model approach. *Bioengineering*, *11*, URL https://api.semanticscholar.org/CorpusID:271429480.

Chen, J., Zhang, Y., Pan, Y., Xu, P., & Guan, C. (2023). A transformer-based deep neural network model for ssvep classification. *Neural Networks*, *164*, 521–534.

Cuadrado, J., Rançon, U., Cottereau, B. R., Barranco, F., & Masquelier, T. (2023). Optical flow estimation from event-based cameras and spiking neural networks. *Frontiers in Neuroscience*, *17*, Article 1160034.

Deng, L., Wang, G., Li, G., Li, S., Liang, L., Zhu, M., et al. (2020). Tianjic: A unified and scalable chip bridging spike-based and continuous neural computation. *IEEE Journal of Solid-State Circuits*, *55*(8), 2228–2246.

Fang, W., Chen, Y., Ding, J., Yu, Z., Masquelier, T., Chen, D., et al. (2023). SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, *9*(40), eadi1480.

Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., & Tian, Y. (2021). Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, *34*, 21056–21069.

Gall, R., Kocanaogullari, D., Akcakaya, M., Erdogmus, D., & Kubendran, R. (2023). Corticomorphic hybrid CNN-SNN architecture for EEG-based low-footprint low-latency auditory attention detection. arXiv preprint arXiv:2307.08501.

Gerhards, P., Weih, M., Huang, J., Knobloch, K., & Mayr, C. G. (2023). Hybrid spiking and artificial neural networks for radar-based gesture recognition. In *2023 8th international conference on frontiers of signal processing* (pp. 83–87). IEEE.

Guo, L., Gao, Z., Qu, J., Zheng, S., Jiang, R., Lu, Y., et al. (2023). Transformer-based spiking neural networks for multimodal audio-visual classification. *IEEE Transactions on Cognitive and Developmental Systems*, URL https://api.semanticscholar.org/CorpusID:264480751.

Gupta, A. K., Kumar, R., Birla, L., & Gupta, P. (2023). RADIANT: Better rPPG estimation using signal embeddings and transformer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 4976–4986).

He, B., & Hofmann, T. (2023). Simplifying transformer blocks. arXiv preprint arXiv:2311.01906.

Horowitz, M. (2014). 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE international solid-state circuits conference digest of technical papers* (pp. 10–14). IEEE.

Huang, Y., Lin, Z., Liu, X., Gong, Y., Lu, S., Lei, F., et al. (2023). Competition-level problems are effective LLM evaluators. arXiv:2312.02143.

Jiang, Y., Cao, Y., & Shen, W. (2023). A masked reverse knowledge distillation method incorporating global and local information for image anomaly detection. *Knowledge-Based Systems*, *280*, Article 110982.

Joshi, J., & Cho, Y. (2024). Imaging blood volume pulse dataset: RGB-thermal remote photoplethysmography dataset with high-resolution signal-quality labels. *Electronics*, URL https://api.semanticscholar.org/CorpusID:268882929.

Kaur, P., & Stoltzfus, J. C. (2017). Bland–altman plot: A brief overview. *International Journal of Academic Medicine*, *3*(1), 110–111.

Lee, C., Kosta, A. K., Zhu, A. Z., Chaney, K., Daniilidis, K., & Roy, K. (2020). Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks. In *European conference on computer vision* (pp. 366–382). Springer.

Li, X., Chen, J., Zhao, G., & Pietikainen, M. (2014). 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).

Li, L., & Liu, Y. (2023). Multi-dimensional attention spiking transformer for event-based image classification. In *2023 5th International Conference on Communications, Information System and Computer Engineering* (pp. 359–362). URL https://api.semanticscholar.org/CorpusID:259158803.

Li, X., Xu, Z. Q. J., & Zhang, Z. (2023). Loss spike in training neural networks. arXiv preprint arXiv:2305.12133.

Li, J., Yu, Z., & Shi, J. (2023). Learning motion-robust remote photoplethysmography through arbitrary resolution videos. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 37*, (1), (pp. 1334–1342).

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132, URL https://api.semanticscholar.org/CorpusID:235368340.

Liu, X., Fromm, J., Patel, S., & McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, *33*, 19400–19411.

Liu, M., Gan, J., Wen, R., Li, T., Chen, Y., & Chen, H. (2023). Spiking-diffusion: Vector quantized discrete diffusion model with spiking neural networks. arXiv:2308.10187.

Liu, X., Hill, B., Jiang, Z., Patel, S., & McDuff, D. (2023). 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).

Liu, X., Narayanswamy, G., Paruchuri, A., Zhang, X., Tang, J., Zhang, Y., et al. (2023). rPPG-toolbox: Deep remote PPG toolbox. In *Thirty-seventh conference on neural information processing systems datasets and benchmarks track*.

Liu, X., Wei, W., Kuang, H., & Ma, X. (2022). Heart rate measurement based on 3d central difference convolution with attention mechanism. *Sensors*, *22*(2), 688.

Liu, Y., Xiao, S., Li, B., & Yu, Z. (2023). SparseSpikformer: A co-design framework for token and weight pruning in spiking transformer. arXiv:2311.08806.

Liu, X., Zhang, T., & Liu, M. (2023). Joint estimation of pose, depth, and optical flow with a competition-cooperation transformer network. *Neural Networks*.

Liu, F., Zheng, H., Ma, S., Zhang, W., Liu, X., Chua, Y., et al. (2024). Advancing brain-inspired computing with hybrid neural networks. *National Science Review*, nwae066.

Lv, C., Li, T., Xu, J., Gu, C., Ling, Z., Zhang, C., et al. (2023). SpikeBERT: A language spikformer trained with two-stage knowledge distillation from BERT. ArXiv abs/2308.15122, URL https://api.semanticscholar.org/CorpusID:261276843.

Lynch, P. (2007). Human head anatomy with external and internal carotid arteries. http://www.flickr.com/photos/patrlynch/450142019.

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, *10*(9), 1659–1671.

McDuff, D. (2023). Camera measurement of physiological vital signs. *ACM Computing Surveys*, *55*(9), 1–40.

Mueller, E., Studenyak, V., Auge, D., & Knoll, A. (2021). Spiking transformer networks: A rate coded approach for processing sequential data. In *2021 7th international conference on systems and informatics* (pp. 1–5). IEEE.

Niu, X., Yu, Z., Han, H., Li, X., Shan, S., & Zhao, G. (2020). Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, Part II 16* (pp. 295–310). Springer.

Nunes, J. D., Carvalho, M., Carneiro, D., & Cardoso, J. S. (2022). Spiking neural networks: A survey. *IEEE Access*, *10*, 60738–60764. http://dx.doi.org/10.1109/ACCESS.2022.3179968.

O'Rourke, M. F., Pauca, A., & Jiang, X. J. (2001). Pulse wave analysis. *British Journal of Clinical Pharmacology*, *51*(6), 507.

Poh, M. Z., McDuff, D. J., & Picard, R. W. (2010a). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, *58*(1), 7–11.

Poh, M. Z., McDuff, D. J., & Picard, R. W. (2010b). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, *18*(10), 10762–10774.

Poh, M. Z., McDuff, D. J., & Picard, R. W. (2011). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, *58*, 7–11, URL https://api.semanticscholar.org/CorpusID:17635486.

Rosenfeld, B., Simeone, O., & Rajendran, B. (2022). Spiking generative adversarial networks with a neural network discriminator: Local training, bayesian models, and continual meta-learning. *Institute of Electrical and Electronics Engineers. Transactions on Computers*, *71*(11), 2778–2791.

Sabour, R. M., Benezeth, Y., De Oliveira, P., Chappe, J., & Yang, F. (2021). Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*.

Seras, A. M., Del Ser, J., & Garcia-Bringas, P. (2023). Efficient object detection in autonomous driving using spiking neural networks: Performance, energy consumption analysis, and insights into open-set object discovery. arXiv preprint arXiv:2312.07466.

Serra, J., Suris, D., Miron, M., & Karatzoglou, A. (2018). Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning* (pp. 4548–4557). PMLR.

Shao, H., Luo, L., Chen, S., Hu, C., & Yang, J. (2023). Hyperbolic embedding steered spatiotemporal graph convolutional network for video-based remote heart rate estimation. *Engineering Applications of Artificial Intelligence*, *124*, Article 106642.

Sharifara, A., Rahim, M. S. M., & Anisi, Y. (2014). A general review of human face detection including a study of neural networks and haar feature-based cascade classifier in face detection. In *2014 international symposium on biometrics and security technologies* (pp. 73–78). IEEE.

Sinhal, R., Singh, K., & Raghuwanshi, M. (2020). An overview of remote photoplethysmography methods for vital sign monitoring. In *Computer vision and machine intelligence in medical image analysis: international symposium* (pp. 21–31). Springer.

Špetlík, R., Franc, V., & Matas, J. (2018). Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, newcastle, UK* (pp. 3–6).

Stein, R., & Hodgkin, A. L. (1967). The frequency of nerve action potentials generated by applied currents. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *167*(1006), 64–86.

Stricker, R., Müller, S., & Gross, H. M. (2014). Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE international symposium on robot and human interactive communication* (pp. 1056–1062). IEEE.

Tang, J., Chen, K., Wang, Y., Shi, Y., Patel, S. N., McDuff, D. J., et al. (2023). MMPD: Multi-domain mobile video physiology dataset. In *2023 45th Annual international conference of the IEEE engineering in medicine & biology society* (pp. 1–5). URL https://api.semanticscholar.org/CorpusID:256662570.

Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J. F., & Sebe, N. (2016). 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).

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Neural information processing systems*. URL https://api.semanticscholar.org/CorpusID:13756489.

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics Express*, *16*(26), 21434–21445.

Wade, J. J., McDaid, L. J., Santos, J. A., & Sayers, H. M. (2010). SWAT: A spiking neural network training algorithm for classification problems. *IEEE Transactions on Neural Networks*, *21*(11), 1817–1830.

Wang, W., den Brinker, A. C., Stuijk, S., & de Haan, G. (2017). Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, *64*, 1479–1491, URL https://api.semanticscholar.org/CorpusID:6372418.

Wang, Z., Fang, Y., Cao, J., Wang, Z., & Xu, R. (2022). Efficient spiking transformer enabled by partial information. arXiv preprint arXiv:2210.01208.

Wang, Z., Fang, Y., Cao, J., Zhang, Q., Wang, Z., & Xu, R. (2023). Masked spiking transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1761–1771).

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, Q., Li, Z., Zhang, S., Chi, N., & Dai, Q. (2023). A versatile wavelet-enhanced CNN-transformer for improved fluorescence microscopy image restoration. *Neural Networks*.

Wang, K., Wei, Y., Tong, M., Gao, J., Tian, Y., Ma, Y., et al. (2023). PhysBench: A benchmark framework for remote physiological sensing with new dataset and baseline. ArXiv abs/2305.04161, URL https://api.semanticscholar.org/CorpusID:258557743.

Wang, X., Wu, Z. Y., Rong, Y., Zhu, L., Jiang, B., Tang, J., et al. (2023). Sstformer: Bridging spiking neural network and memory support transformer for frame-event based recognition. ArXiv abs/2308.04369, URL https://api.semanticscholar.org/CorpusID:260704582.

Wang, Z., Zhao, Q., Cui, J., Liu, X., & Xu, D. (2023). AutoST: Training-free neural architecture search for spiking transformers. arXiv:2307.00293.

Watanabe, R., Mukuta, Y., & Harada, T. (2023). Fully spiking denoising diffusion implicit models. arXiv preprint arXiv:2312.01742.

Wei, W., Zhang, M., Qu, H., Belatreche, A., Zhang, J., & Chen, H. (2023). Temporal-coded spiking neural networks with dynamic firing threshold: Learning with event-driven backpropagation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10552–10562).

Wei, W., Zhang, M., Zhang, J., Belatreche, A., Wu, J., Xu, Z., et al. (2024). Event-driven learning for spiking neural networks. arXiv preprint arXiv:2403.00270.

Wu, J., Chua, Y., Zhang, M., Li, H., & Tan, K. C. (2018). A spiking neural network framework for robust sound classification. *Frontiers in Neuroscience*, *12*, 836.

Wu, H. Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics*, *31*(4), 1–8.

Yan, Z., Zhou, J., & Wong, W. F. (2021). Energy efficient ECG classification with spiking neural network. *Biomedical Signal Processing and Control*, *63*, Article 102170.

Yao, M., Gao, H., Zhao, G., Wang, D., Lin, Y., Yang, Z., et al. (2021). Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10221–10230).

Yao, M., Hu, J., Zhou, Z., Yuan, L., Tian, Y., Bo, X., et al. (2023). Spike-driven transformer. In *Thirty-seventh conference on neural information processing systems*.

Yu, Z., Li, X., Niu, X., Shi, J., & Zhao, G. (2020). Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, *27*, 1245–1249.

Yu, Z., Peng, W., Li, X., Hong, X., & Zhao, G. (2019). 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).

Yu, Z., Shen, Y., Shi, J., Zhao, H., Cui, Y., Zhang, J., et al. (2023). Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, *131*(6), 1307–1330.

Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P. H., & Zhao, G. (2022). 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).

Yu, F., Wu, Y., Ma, S., Xu, M., Li, H., Qu, H., et al. (2023). Brain-inspired multimodal hybrid neural network for robot place recognition. *Science Robotics*, *8*(78), eabm6996.

Yue, Z., Shi, M., & Ding, S. (2023). Facial video-based remote physiological measurement via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J., Huo, D., Zhang, J., Qian, C., Liu, Q., Pan, L., et al. (2023). 22.6 ANP-I: A 28nm 1.5 pJ/SOP asynchronous spiking neural network processor enabling sub-o. 1 μj/sample on-chip learning for edge-AI applications. In *2023 IEEE international solid-state circuits conference* (pp. 21–23). IEEE.

Zhang, J., Liang, M., Wei, J., Wei, S., & Chen, H. (2021). A 28nm configurable asynchronous SNN accelerator with energy-efficient learning. In *2021 27th IEEE international symposium on asynchronous circuits and systems* (pp. 34–39). IEEE.

Zhang, M., Qu, H., Belatreche, A., Chen, Y., & Yi, Z. (2018). A highly effective and robust membrane potential-driven supervised learning method for spiking neurons. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(1), 123–137.

Zhang, M., Wang, J., Wu, J., Belatreche, A., Amornpaisannon, B., Zhang, Z., et al. (2021). Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(5), 1947–1958.

Zhang, X., Xia, Z., Liu, L., & Feng, X. (2023). Demodulation based transformer for rPPG generation and heart rate estimation. *IEEE Signal Processing Letters*, *30*, 1042–1046, URL https://api.semanticscholar.org/CorpusID:260719975.

Zhang, Y., Xu, H., Huang, L., & Chen, C. (2023). A storage-efficient SNN–CNN hybrid network with RRAM-implemented weights for traffic signs recognition. *Engineering Applications of Artificial Intelligence*, *123*, Article 106232.

Zhang, N., Yu, L., Zhang, D., Wu, W., Tian, S., Kang, X., et al. (2024). CT-net: Asymmetric compound branch transformer for medical image segmentation. *Neural Networks*, *170*, 298–311.

Zhang, L., Zhou, S., Zhi, T., Du, Z., & Chen, Y. (2019). Tdsnn: From deep neural networks to deep spike neural networks with temporal-coding. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 33*, (01), (pp. 1319–1326).

Zhou, C., Yu, L., Zhou, Z., Zhang, H., Ma, Z., Zhou, H., et al. (2023). Spikingformer: Spike-driven residual learning for transformer-based spiking neural network. arXiv preprint arXiv:2304.11954.

Zhou, Z., Zhu, Y., He, C., Wang, Y., Shuicheng, Y., Tian, Y., et al. (2022). Spikformer: When spiking neural network meets transformer. In *The eleventh international conference on learning representations*.

Zhu, D., Zhang, W., Zeng, H., Liu, X., Yang, L., & Zheng, J. (2024). rFaceNet: An end-to-end network for enhanced physiological signal extraction through identity-specific facial contours. URL https://arxiv.org/abs/2403.09034, arXiv:2403.09034.

Zhu, R. J., Zhao, Q., Li, G., & Eshraghian, J. K. (2023). SpikeGPT: Generative pre-trained language model with spiking neural networks. arXiv:2302.13939.

Zhuge, R., Wang, J., Xu, Z., & Xu, Y. (2023). Single image denoising with a feature-enhanced network. *Neural Networks*, *168*, 313–325.

Zong, Z., Yan, H., Sui, H., Li, H., Jiang, P., & Li, Y. (2023). An AI-based simulation and optimization framework for logistic systems. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 5138–5142).

Zou, S., Mu, Y., Zuo, X., Wang, S., & Li, C. (2023). Event-based human pose tracking by spiking spatiotemporal transformer. ArXiv abs/2303.09681, URL https://api.semanticscholar.org/CorpusID:257622694.