This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

FE-SpikeFormer: A Camera-based Facial Expression Recognition method for Hospital Health Monitoring

Zhekang Dong, *Senior Member*, *IEEE*, Liyan Zhu, Shiqi Zhou, Xiaoyue Ji, *Member*, *IEEE*, Chun Sing Lai, *Senior Member*, *IEEE*, Minjiang Chen, and Jiansong Ji

***Abstract*—Facial expression recognition has emerged as a critical research area in health monitoring, enabling healthcare professionals to assess patients' emotional and psychological states for timely intervention and personalized care. However, existing methods often struggle to balance computational accuracy with energy efficiency. To address this challenge, this paper proposes FE-SpikeFormer — a high-accuracy, low-energy, and deployment-friendly Spiking Neural Network (SNN) for facial emotion recognition. The proposed architecture comprises three key components: the initial convolution module, the spiking extraction block, and the spiking integration block. These three modules collectively support detailed and contextual feature extraction, promote spatial feature integration, and strengthen the representational capacity of spiking signals. Meanwhile, a joint verification is conducted in both controlled laboratory settings and real-world hospital scenarios. Experimental results demonstrate that FE-SpikeFormer achieves top-three recognition accuracy among state-of-the-art methods, while utilizing only 6.93 million parameters. Moreover, it exhibits strong robustness against various noise conditions, underscoring its potential for practical deployment in healthcare environments.**

***Index Terms*—Facial Expression Recognition, Hospital Health Monitoring, Dual Attention Mechanism, Spiking Neural Network**

I. Introduction

With the development of deep learning technologies, facial expression recognition has increasingly emerged as a key research direction, particularly in the domain of human health monitoring [1, 2]. Notably, facial expressions are one of the most natural and universal ways for humans to convey their emotional states and behavioral intentions. Accurate and rapid facial expression recognition plays a clinically significant role in healthcare, particularly in the domains of pain assessment, neuropsychiatric monitoring, and patient engagement. It enables clinical interventions and facilitates personalized care [3].

Many studies have sought to enhance the accuracy and real-time performance of facial expression recognition through the use of Artificial Neural Networks (ANNs) [4-13]. Specifically, both [4] and [5] focus on enhancing Convolutional Neural Networks (CNNs) to provide an efficient and computationally lightweight solution for facial expression recognition. In [6], a multimodal-based facial expression recognition method (i.e., IdentiFace) is developed, which requires extensive data and computational resources. In [7], an Adaptive Correlation (Ad-Corre) loss function is designed to enhance facial expression recognition by improving feature discrimination. Then, a fine-tunning VGGNet is proposed to perform facial expression recognition without using additional training data [8]. By simplifying the structure and compressing parameters, a Squeeze-and-Excitation Network (i.e., SENet) is developed to perform facial expression recognition [9]. A Poker Face Vision Transformer (i.e., PF-ViT) that separates emotion-related features from emotion-irrelevant components is proposed for facial expression recognition [10]. [11] proposes an Oriented Attention Enable Network (OAENet) for facial expression recognition, ensuring the sufficient utilization of both global and local features. [12] proposes an end-to-end Facial Expression Recognition model that integrates Expression Synthesis with Representation learning (ESR-FER). In [13], a De-Elements Network (DENet) is proposed to enhance feature discrimination and improve classification performance in facial expression recognition.

Although these above-mentioned methods are able to successfully perform facial expression recognition, they still suffer from some limitations: All these methods rely on different ANN architectures, intensive computing has to execute to guarantee the recognition accuracy, which always makes these networks redundant and hard to deploy. Meanwhile, heavy calculation burden brings high energy consumption. The desire for high-accuracy and energy-efficiency computing paradigm that is compatible with facial expression recognition technology in hospital health monitoring is becoming realistic and attractive [14, 15].

Inspired by human brain, Spiking Neural Networks (SNNs) have emerged as energy-efficient computing paradigms, owing to event-driven mechanisms and binary spike characteristics. So far, SNN has made some progress in image detection and classification task [16-18]. However, most of

This work was supported by National Natural Science Foundation of China under Grant 62401326, China Postdoctoral Science Foundation under Grants 2024T170463 and 2024M751676, Ministry of Science and Technology-Yangtze River Delta Science and Technology Innovation Program under Grant 2023CSJGG1300, and Hangzhou Dianzi University Graduate Research Innovation Fund Project under Grant CXJJ2024044. (Corresponding author: *Xiaoyue Ji*)

Z Dong, L Zhu, and S Zhou are with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou, China, 310018. (e-mail: englishp@hdu.edu.cn; yan@hdu.edu.cn; shiqizhou@hdu.edu.cn).

X. Ji is with the Department of Precision Instrument, Tsinghua University, Beijing 100084, China. (e-mail: jixiaoyue@mail.tsinghua.edu.cn).

C. S. Lai is with Department of Electronic and Computer Engineering, Brunel University London, London, UB8 3PH, UK, (e-mail: chunsing.lai@brunel.ac.uk).

M. Chen and J. Ji are with Key Laboratory of Imaging Diagnosis and Minimally Invasive Intervention Research, they are also with Lishui Central Hospital, Lishui, China, 323000. (e-mail: minjiangchen@wmu.edu.cn; jijiansong@zju.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

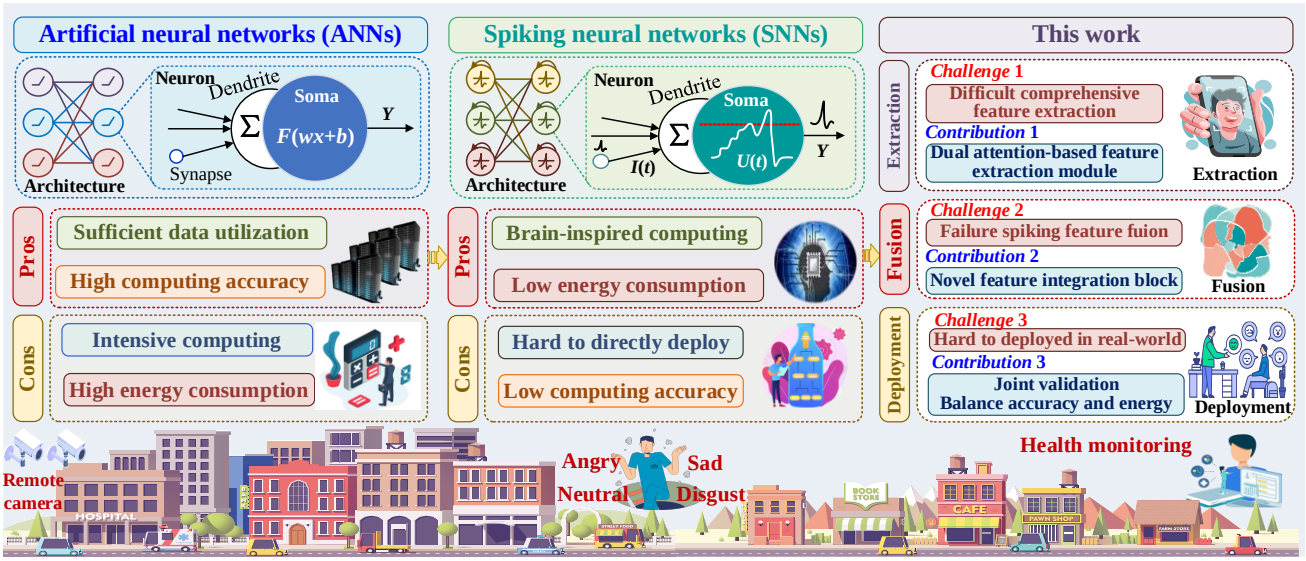

Fig. 1. The systemic comparison of ANNs- and SNNs-based methods for facial expression recognition

existing SNN-based applications are still difficult to achieve ANN-comparable accuracy and hard to directly deploy in the real-world outdoor scenario. For clarity, the systemic comparison of ANNs- and SNNs-based methods for facial expression recognition is illustrated in **Fig. 1**.

From **Fig. 1**, the specific challenges are concluded below:

Challenge 1: In the existing SNNs, different information is processed using sparse computing, which makes sufficient and comprehensive facial feature extraction difficult and challengeable [19]. Hence, Hence, how to design a specific facial feature extraction method suitable for SNNs is important and urgent.

Challenge 2: Although SNNs inherently encode temporal dynamics [20], conventional spiking feature fusion methods fail to effectively integrate multi-scale spatial and temporal cues that are critical for understanding facial expressions. Hence, how to design a specific facial feature fusion module in SNNs is important.

Challenge 3: Almost all existing SNNs have been deployed in controlled laboratory environments [21, 22] to demonstrate their potential for energy-efficient computing. It is important to develop a SNN-based facial expression recognition network in a real-world hospital scenario, enabling to balance the trade-off between accuracy, energy consumption, and running speed.

Based on these challenges, this work proposes FE-SpikeFormer, a high-accuracy, low-energy consumption, and easy-deployable SNN-based facial emotion recognition network designed for hospital health monitoring.

The main contributions of this work can be summarized as follows:

1) To address the challenge of sparse feature extraction, this work proposes a dual attention-based architecture, combining both the local and global attention to capture detailed and contextual features, enhancing the representation of facial expression features in SNNs.

2) To address the challenge of spiking feature fusion, this work proposes a novel feature integration strategy, which facilitates the integration of spatial features in each stage and enhances the representation capability of spiking signals, indicating better overall performance of facial expression recognition.

3) To address deployment issue of SNNs in real-world outdoor environments, the proposed FE-SpikeFormer is evaluated in both a controlled laboratory environment and a real-world hospital setting. The joint validation results demonstrate that the entire scheme is able to balance the tradeoff between accuracy, energy efficiency, and running speed, ensuring reliable performance under diverse and dynamic conditions beyond controlled indoor settings.

The rest of this work is organized as follows: **Section II** elaborates on the design of FE-SpikeFormer. In **Section III**, the dataset and the experimental preparation are described in detail. In **Section IV**, a series of experiments are conducted in both laboratory and hospital environments to demonstrate the superiority of the proposed FE-SpikeFormer in facial expression recognition. **Section V** concludes the entire work, and **Section VI** discusses its limitations and outlines future research directions.

II. Methods

A. Overall Architecture

Fig. 2 demonstrates the overview of FE-SpikeFormer, which primarily consists of three components: Initial-Conv (InConv), Spiking Extraction Block (SEB), and Spiking Integration Block (SIB). As the InConv module has a relatively simple structure [15], the analysis mainly concentrates on the latter two components.

Here, when a 2-Dimensional (2D) image sequence P (as the input of FE-SpikeFormer) is given, the InConv module transforms this image sequence into more compact and informative spike-form patches x. Then, x is sequentially fed into the M-block SEB and N-block SIB to perform local and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

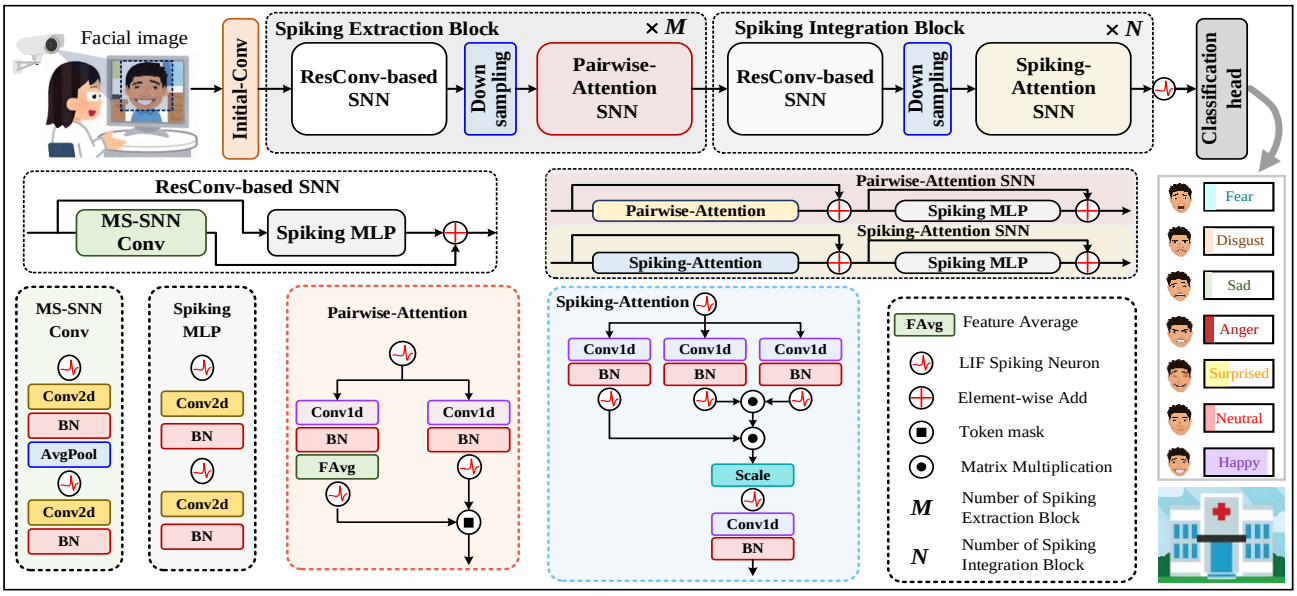

Fig. 2. The overview of FE-SpikeFormer

global attention operations, respectively. The output is then passed to the classification head (*Head*) to produce the final result *I*.

B. Spiking neuron layer

In SNNs, the spike neuron serves as the core unit, integrating incoming currents to accumulate membrane potential. When the potential surpasses a defined threshold, the neuron fires a spike. This study exclusively employs Leaky Integrate-and-Fire (LIF) neurons. The dynamics of LIF model can be expressed 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 represents the membrane time constant, and $X[t]$ denotes the input current at time step t. Notably, a spike $S[t]$ is generated when the membrane potential $H[t]$ surpasses the threshold V_{th}. The Heaviside step function $\Theta(v)$ outputs 1 when $v \geq 0$, and 0 when $v < 0$. After firing, the membrane potential $V[t]$ is reset to V_{reset}; If no spike occurs, it retains the value $H[t]$.

C. Spiking Extraction Block

The SEB consists of a ResConv-based SNN (RCS) and a Pairwise-Attention SNN (PAS). In the RCS module, the model performs convolutional computations based on SNN architecture, while the pairwise-attention computations are performed in the PAS module.

1) ResConv-based SNN

Given an input spiking map $S \in \mathrm{R}^{T\times C\times H\times W}$, the spiking representation is generated in the RCS module. Specifically, the Multi-Stage SNN Conv (MS-SNN Conv) module ($MSConv_1$ and $MSConv_2$) with 3×3 kernel size is used to learn local information. 2D-avg-pooling layer is applied to down-sample the feature map. During this process, the size and channel dimensions of the feature maps change from $\mathrm{R}^{T\times C\times H\times W}$ to $\mathrm{R}^{T\times 2C\times H/2\times W/2}$. To ensure alignment in the shortcut connections and to enhance information representation, the spiking Multi-Layer Perception (MLP) module $SMLP(\cdot)$ with a 1×1 kernel is applied. The operations performed in the RCS module are mathematically expressed by:

$$S_1 = MSConv_1(S) \tag{4}$$

$$S_2 = MSConv_2(S_1) \tag{5}$$

$$S_3 = SMLP(S) \tag{6}$$

$$U = S_2 + S_3 \tag{7}$$

$$MSConv_1(S) = AvgPool(BN(Conv2d(LIF(S)))) \tag{8}$$

$$MSConv_2(S_1) = BN(Conv2d(LIF(S_1))) \tag{9}$$

$$SMLP(S) = BN(Conv2d(LIF(S))) \tag{10}$$

where U denotes the output of RCS, $LIF(\cdot)$ denotes the activation function of the spiking neuron layer, $Conv2d(\cdot)$ represents 2D convolution operation, $BN(\cdot)$ and $AvgPool(\cdot)$ denote batch normalization and average pooling operation. S_1, S_2, and S_3 are intermediate variables in the RCS module.

2) Pairwise-Attention SNN

The feature map $U \in \mathrm{R}^{T\times C\times H\times W}$ is further passed to PAS for attention computation. Similar to RCS, PAS is also composed by two parts: the pairwise-attention (PA) module and the spiking MLP module.

After flattening operation, the feature map U is converted into a sequence of image patches $u \in \mathrm{R}^{T\times C\times N}$, where $N = H\times W$. Then, the stimulus value SV and response value RV with the same size ($\mathrm{R}^{T\times C\times N}$) can be calculated by:

$$SV = LIF_{SV}(BN(u'W_{SV})) \tag{11}$$

$$RV = LIF_{RV}(BN(u'W_{RV})) \tag{12}$$

where $u'=LIF(u)$ is the spiked form of u, W_{SV} and W_{RV} are the learnable linear matrices, $LIF_{SV}(\cdot)$ and $LIF_{RV}(\cdot)$ are the corresponding neuron activation operation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

The spiking-form global features within SV channel dimension $A_{SV} \in \mathrm{R}^{T\times C\times 1}$ can be mathematically calculated by:

$$A_{SV} = LIF(Avg(SV)) \tag{13}$$

where $Avg(\cdot)$ denotes the average computation operation.

Next, the corresponding attention feature map ATT_{Pai} can be obtained and then flattened and transformed into spiking signals. The output of the pairwise-attention (PA) module Y_1 can be written by:

$$ATT_{Pai} = A_{SV} * RV \tag{14}$$

$$Y_1 = BN(Conv1d(LIF(FLA(ATT_{Sti})))) \tag{15}$$

where $FLA(\cdot)$ and $Conv1d(\cdot)$ denote the flattening operation and the one-dimensional convolution (1D convolution) operation, respectively.

Then, the output of the pairwise-attention SNN can be expressed by:

$$Y = PA(U) + U \tag{16}$$

$$OUT_{PAS} = SMLP(Y) + Y \tag{17}$$

where $PA(\cdot)$ denotes the pairwise-attention module, and OUT_{PAS} represents the output of the pairwise-attention SNN.

D. Spiking Integration Block

The SIB module consists of two components: a ResConv-based spiking neural network (RCS) and a spiking-attention-based spiking neural network (SAS). In the SAS module, the spiking-attention computation is performed. Unlike the pairwise-attention module, spiking-attention module focuses more on global modeling across different channels, whereas the pairwise-attention mechanism emphasizes feature modeling within individual channels.

Similarly, the input feature map $Z \in \mathrm{R}^{T\times C\times H\times W}$ can be converted into a sequence of image patches $z \in \mathrm{R}^{T\times C\times N}$ after flattening operation. The query (Q), key (K), and Value (V) are computed through learnable matrices W_Q, W_K, $W_V \in \mathrm{R}^{N\times N}$ firstly. Then they are transformed into spiking sequences Q_S, K_S, and V_S through distinct spiking neuron layers:

$$Q_S = LIF_Q(BN(z'W_Q)) \tag{18}$$

$$K_S = LIF_K(BN(z'W_K)) \tag{19}$$

$$V_S = LIF_V(BN(z'W_V)) \tag{20}$$

where $z'=LIF(z)$ represents the spiked form of z.

Then, a scaling factor b is added to prevent excessively large values during matrix multiplication. And the output of the spiking-attention (SA) module X_1 can be written by:

$$ATT_{Spi} = (Q_S \odot (K_S^T \odot V_S)) * b \tag{21}$$

$$X_1 = BN(Conv1d(FLA(LIF(ATT_{Spi})))) \tag{22}$$

where ATT_{Spi} represents the output of the spiking-attention computation. The overall computation in the spiking-attention SNN can be expressed by:

$$X = SA(Z) + Z \tag{23}$$

$$OUT_{SAS} = SMLP(X) + X \tag{24}$$

where $SA(\cdot)$ denotes the spiking-attention calculation, and OUT_{SAS} represents the output of the spiking-attention SNN.

For clarity and simplicity, the pseudocode of the proposed FE-SpikeFormer is provided in **Table I**.

TABLE I
THE PSEUDOCODE OF THE PROPOSED FE-SPIKEFORMER

Algorithm FE-SpikeFormer

Input: Raw facial images P, Number of blocks M for SEB, Number of blocks N for SIB
Output: Prediction result I
Step 1: Transform images into spike form patches: $x \leftarrow InConv(P)$
Step 2: Extract local features in Spiking Extraction Block:
  for i from 1 to M **do**
    Initial feature extraction: $U \leftarrow RCS(x)$
    Local attention computation: $Y_1 \leftarrow PA(U)$
    Residual connection fusion: $Y \leftarrow Y_1 + U$
    Feature aggregation output: $OUT_{PAS} \leftarrow SMLP(Y) + Y$
  end
Step 3: Integrate global features in Spiking Integration Block:
  for i from 1 to N **do**
    Initial feature extraction: $Z \leftarrow RCS(OUT_{PAS})$
    global attention computation: $X_1 \leftarrow SA(Z)$
    Residual connection fusion: $X \leftarrow X_1 + Z$
    Feature aggregation output: $OUT_{SAS} \leftarrow SMLP(X) + X$
  end
Step 4: Produce final prediction output: $I \leftarrow Head(OUT_{SAS})$
Return: I

III. DATASET AND EXPERIMENTAL PREPARATION

A. Dataset

Currently there is no open-source facial expression recognition dataset specifically for hospital environment, two widely-used public datasets (i.e., FER2013 dataset [23] and AffectNet dataset [24]) are used for laboratory environment experiment. We ensure a rigorous data partitioning between the training, validation, and test sets by strictly adhering to the official dataset splits [23, 24]. Specifically, the FER2013 dataset (36157 images) is further divided into a training set (28709 images), a validation set (3859 images), and a testing set (3589 images) across seven facial expressions (including angry, disgust, fear, happy, sad, surprise, and neutral). The AffectNet dataset (420299 images) is further divided into a training set (416299 images) and a validation set (4000 images) across the same seven facial expressions. The specific distribution of these two datasets is shown in **Table II**.

TABLE II
THE DISTRIBUTION OF FER2013 AND AFFECTNET DATASET.

Dataset	Categories	Training	Testing	Total
FER2013 [23]	Anger	3995	491	4486
	Disgust	436	55	491
	Fear	4097	528	4625
	Happy	7215	879	8094
	Sad	4830	594	5424
	Surprise	3171	416	3587
	Neutral	4965	626	5591
AffectNet [24]	Anger	24882	500	25382
	Disgust	3803	500	4303
	Fear	6378	500	6878
	Happy	134415	500	134915
	Sad	25459	500	25959
	Surprise	14090	500	14590
	Neutral	74874	500	75374

To evaluate the generalization capability of the proposed FE-SpikeFormer in a clinical setting, we conducted a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

TABLE III
DETAILED INFORMATION ON THE DATA COLLECTION IN HOSPITAL SETTING

Category	Description	Note
Participants	40 hospitalized patients	All participants provided written informed consent in accordance with the hospital's ethics committee approval (Approval No.: 2024(1)-089-01)
Age distribution	21–30, 31–40, 41–50, and over 50 years old	Even distribution
Gender	20 males and 20 females	Even distribution
Data collection	Facial expressions during bed rest in a natural hospital setting	Data recording is conducted during daytime only;
Emotion Annotation	3 independent clinical experts	Anonymized labels (e.g., Patient 03: anger)
Ethical compliance	Data anonymization, on-site processing only, de-identified image data (no personal identifiers retained)	Compliant with China's Regulations on Ethical Review of Biomedical Research.

deployment study at Lishui Central Hospital (Zhejiang Province, China). The setup involved a remote camera (Logitech C525) mounted on a short wall bracket with an adjustable viewing angle (0°–180°), and a local edge computing system (HS140) for on-site inference (as shown in **Fig. 3**). This configuration enabled unobtrusive, front-facing facial recording of patients under natural lighting conditions. Detailed information on the data collection in the hospital setting is provided in **Table III**.

As summarized in **Table III**, the study involved 40 long-term hospitalized patients, whose ages are grouped into four categories: 21–30, 31–40, 41–50, and over 50 years. All participants provided informed consent, and the study was approved by the hospital's Ethics Committee (Approval ID: 2024(1)-089-01). A total of 204 facial expression video segments were recorded while patients were resting in bed under natural conditions. Facial expressions were annotated frame by frame by three certified clinicians using a predefined emotion taxonomy comprising the categories: happy, neutral, sad, angry, fear, disgust, and surprise. Two clinicians independently labeled each frame, and a third adjudicated any disagreements to ensure high inter-rater reliability.

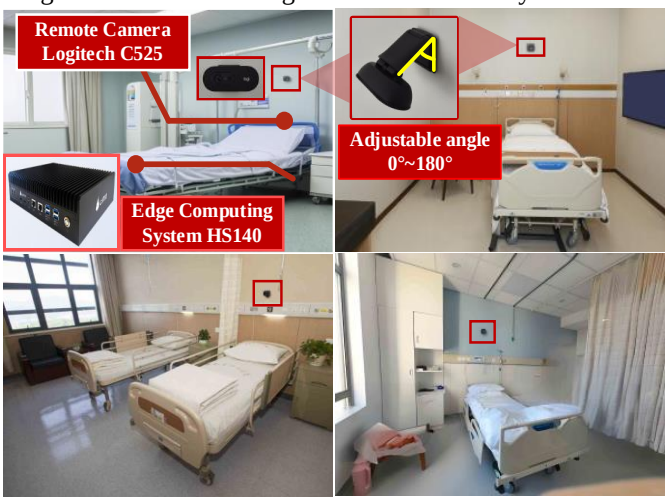

Fig. 3. Necessary sensor & computing devices deployed in a hospital environment

B. Experimental Preparation

1) Experimental Setup

All experiments are conducted on a server equipped with dual NVIDIA GeForce RTX 4090 GPUs, utilizing the PyTorch framework. Each GPU processes a batch size of 48. The AdamW optimizer is employed with a base learning rate of 1×10^{-3}. The actual learning rate is determined by BatchSize/256×10^{-3}. Inspired by [17], [18], [25], and [26], the parameter settings of FE-SpikeFormer are shown in **Table IV**.

TABLE IV
THE PARAMETER SETTING OF FE-SPIKEFORMER

Parameter	Value
Batch size	48
Base learning rate	1×10^{-3}
M	2
N	2
d_{model}	384
T	4
Epoch	300

From **Table IV**, the number of SEB and SIB are both set to 2 (i.e., M=N=2). d_{model} is set to 384 within the attention calculation layer. The time step T is set to 4, and the total number of training epochs is 300. Notably, the configuration of parameters is determined by comprehensive optimization experiments.

2) Evaluation Metrics

In this work, both the confusion matrices [27] and average accuracy [28] are used to evaluate the performance of the proposed FE-SpikeFormer. The former provides detailed information on the model's behavior across various expression categories, while the latter indicates the overall classification accuracy. Additionally, model size and energy consumption [29] are also used to evaluate the computational efficiency.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Comparison Experiments with State-of-the-Art Methods

This part evaluates the performance of FE-SpikeFormer on two datasets and compare it with state-of-the-art (SOTA) methods developed in recent years, including CLCM [4], RTEC [5], IdentiFace [6], Add-Corre [7], VGGNet [8], SENet [9], PF-Vit [10], OAENet [11], ESR-FER [12], DENet [13], ConvSNN [16], Spikformer [17], and Meta-SpikeFormer [18]. The relevant comparison results are presented in **Table V**.

From **Table V**, the proposed FE-SpikeFormer demonstrates competitive performance on both datasets, achieving a notable balance between accuracy and parameter efficiency. For the FER2013 dataset, FE-SpikeFormer achieves the highest accuracy (73.58%) and less energy consumption (7.40 mJ), outperforming all the other SOTA methods. Meanwhile, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

TABLE V
COMPARATIVE RESULTS OF FE-SPIKEFORMER AND OTHER METHODS

Dataset	Method	Year	SNN	Acc (%)	Energy (mJ)	Params (M)	Time Step
FER2013	CLCM [4]	2024	✗	63.01	11.80_3	$\mathbf{2.31_1}$	1
	RTEC [5]	2024	✗	63.21	-	-	1
	IdentiFace [6]	2024	✗	66.13	26.90	35.30	1
	Ad-Corre [7]	2022	✗	72.03_3	30.20	26.15	1
	VGGNet [8]	2021	✗	73.28_2	44.90	138.13	1
	ConvSNN [16]	2024	✓	61.87	21.30	21.50	8
	Spikformer [17]	2022	✓	70.70	7.70_2	9.32_3	4
	Meta-SpikeFormer [18]	2023	✓	71.82	16.70	15.10	4
	FE-SpikeFormer (ours)	2024	✓	$\mathbf{73.58_1}$	$\mathbf{7.40_1}$	6.93_2	4
AffectNet	CLCM [4]	2024	✗	54.11	11.80_3	$\mathbf{2.31_1}$	8
	SENet [9]	2023	✗	56.54	40.50	11.27	1
	PF-Vit [10]	2022	✗	57.99	81.00	86.18	1
	OAENet [11]	2021	✗	58.70	-	-	1
	ESR-FER [12]	2021	✗	60.04_2	98.40	154.10	1
	DENet [13]	2023	✗	$\mathbf{60.94_1}$	39.60	23.33	1
	Spikformer [17]	2022	✓	57.61	7.70_2	9.32_3	4
	Meta-SpikeFormer [18]	2023	✓	58.01	16.70	15.10	4
	FE-SpikeFormer (ours)	2024	✓	59.90_3	$\mathbf{7.40_1}$	6.93_2	4

Note: the subscript 1, 2, and 3 represent the specific ranking results.

proposed FE-SpikeFormer uses only 6.93M parameters (ranking the second place among SOTA methods), which indicates that the proposed FE-SpikeFormer successfully leverages the strengths of SNNs for both computation efficiency and accuracy. For the AffectNet dataset, the proposed FE-SpikeFormer maintains a competitive edge with an accuracy of 59.90%, outperforming almost all the other competitors, except for ESR-FER (60.04%) and DENet (60.94%). But these two methods require more parameters and consume more energy (ESR-FER: 154.1M, 98.4mJ; DENet: 23.33M, 39.60 mJ) compared to the proposed FE-SpikeFormer (only 6.93M, 7.40 mJ). Meanwhile, it is clear that the proposed FE-SpikeFormer is better than Spikformer and Meta-SpikeFormer in terms of accuracy, energy consumption and parameter count.

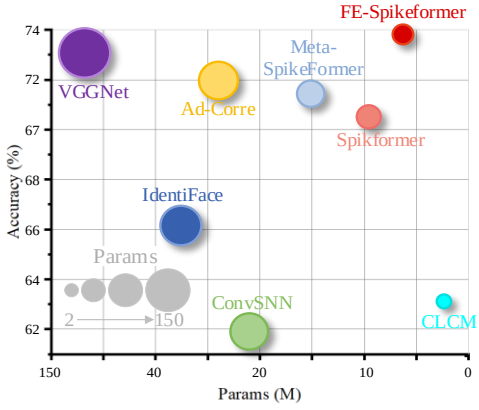

Fig. 4. Comparison of FE-SpikeFormer with SOTA methods in terms of recognition accuracy and parameter size on the FER2013 dataset

To better illustrate that the proposed FE-SpikeFormer is able to balance the trade-off between recognition accuracy and parameter size, an intuitive result is provided in **Fig. 4**. It can be seen that FE-SpikeFormer achieves a superior balance between these two metrics compared to other competitors. Notably, it outperforms VGGNet and Ad-Corre in terms of recognition accuracy, reaching 73.58%, while maintaining a relatively low parameter count. Additionally, compared to traditional models such as ConvSNN and IdentiFace, FE-SpikeFormer delivers higher accuracy with fewer parameters, highlighting its effectiveness in optimizing both performance and efficiency.

Fig. 5 presents attention map examples from the SIB in FE-SpikeFormer, demonstrating its ability to focus on image regions relevant to facial expression semantics. Specifically, the first row displays input images, followed by their corresponding attention maps in the second row. The attention maps highlight key features like the eyes, mouth, and facial expressions, essential for emotion recognition, while irrelevant regions are assigned a value of 0 (black areas). This selective focus enables FE-SpikeFormer to filter out unnecessary information, contributing to event-driven, energy-efficient processing. The results across AffectNet and FER2013 demonstrate the proposed FE-SpikeFormer has the adaptability for diverse inputs (RGB and grayscale images).

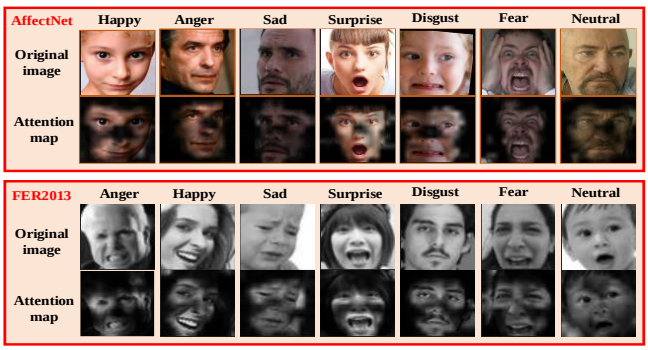

Fig. 5. Attention map visualization of FE-SpikeFormer on AffectNet and FER2013 datasets. The black region is 0.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

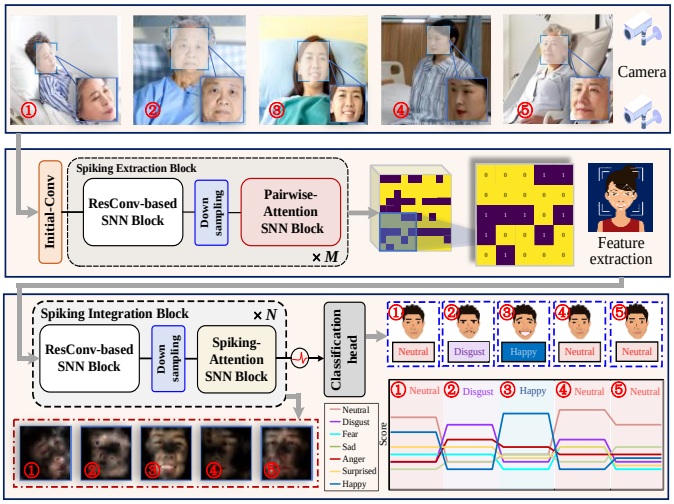

Fig. 6. The workflow of the proposed FE-SpikeFormer model.

Fig. 6 demonstrates the workflow of the proposed FE-SpikeFormer for facial expression recognition. In **Fig. 6**, five facial expression images (labelled as ① ~ ⑤) are input into the well-trained FE-SpikeFormer. After passing through the SEB and SIB, the corresponding attention maps can be obtained. The outputs of the classification head consist of a series of scores, and the final classification results are determined by the highest score, as illustrated in the lower-right subfigure of **Fig. 6**. Notably, signals transmitted between modules are conveyed in spike form. This spike-based computation significantly improves computational efficiency and reduces power consumption, which makes it especially suitable for deployment in hospital settings.

Fig. 7 presents the confusion matrices of FE-SpikeFormer, Meta-SpikeFormer, and Spikformer on two datasets. In FER2013 (**Fig. 7a**), FE-SpikeFormer demonstrates strong performance in recognizing emotions such as happy (0.90), neutral (0.85), and disgust (0.71). These results demonstrate the model is able to effectively extract both local and global features to enhance emotion representation. However, the classification accuracy for fear (0.56) and surprise (0.60) is lower, indicating that these emotions may be more challenging to distinguish due to overlapping facial expressions with other categories, such as anger and sadness. In AffectNet (**Fig. 7b**), the model maintains high accuracy, especially for happy (0.95) and neutral (0.79) emotions, further validating its robustness across datasets. However, it shows some confusion in distinguishing between disgust and anger, with noticeable misclassifications, such as disgust being confused with anger (0.22). This can be attributed to the subtle and ambiguous facial cues present in these emotions, which are inherently challenging for any classifier.

Compared to Meta-SpikeFormer (**Fig. 7c** and **7d**) and Spikformer (**Fig. 7e** and **7f**), FE-SpikeFormer consistently achieves higher accuracy in key emotions across both datasets, especially for happy and neutral, indicating stronger generalization. While Meta-SpikeFormer and Spikformer perform well on certain emotions, they display more variability and a higher tendency to misclassify nuanced emotions like disgust. The dual attention-based feature extraction in proposed FE-SpikeFormer enhances its ability to capture fine-grained features, making it more effective in differentiating subtle expressions. Overall, FE-SpikeFormer demonstrates superior robustness and accuracy in complex facial expression recognition.

B. Ablation Study

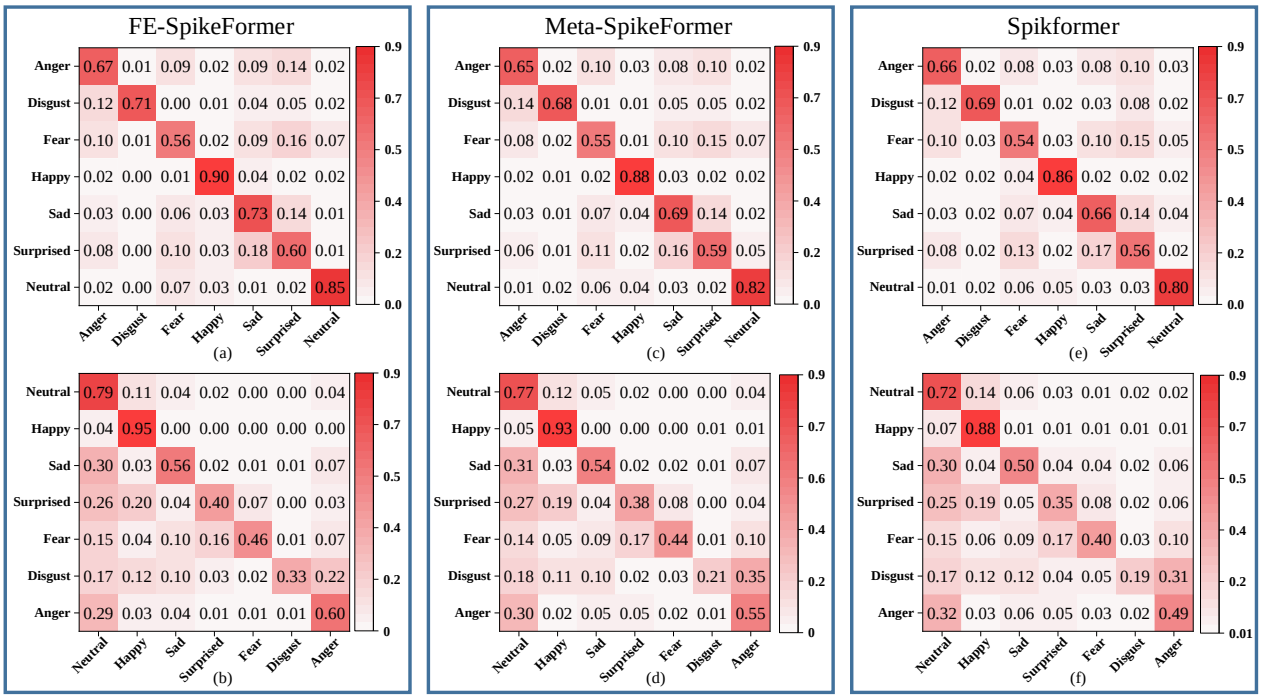

Fig. 7. Confusion Matrices of FE-SpikeFormer, Meta-SpikeFormer, and Spikformer on FER2013 and AffectNet Datasets. (a) FE-SpikeFormer on FER2013. (b) FE-SpikeFormer on AffectNet. (c) Meta-SpikeFormer on FER2013. (d) Meta-SpikeFormer on AffectNet. (e) Spikformer on FER2013. (f) Spikformer on AffectNet

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

In this section, a series of ablation experiments are conducted on FER2013 dataset (including component ablation and block number ablation, and time step ablation).

1) Ablation Study of the Key Components in FE-SpikeFormer

Component ablation experiments are conducted to demonstrate the effectiveness of the proposed MS-SNN Conv and pairwise-attention module. Specifically, for the MS-SNN Conv module, we replace it with a simple convolutional down-sampling module based on an SNN architecture {LIF-Conv-BN}. For the pairwise-attention module, we substitute it with a spiking-attention module, which eliminates the attention computation within individual feature channels and retains only the global attention across channels. Four sets of comparative experiments are conducted, and the relevant results are collected in **Table VI**.

TABLE VI
RESULTS OF COMPARATIVE EXPERIMENTS WITH DIFFERENT MODULE CONFIGURATIONS

MS-SNN Conv	Pairwise-Attention	Acc (%)
✗	✗	70.33
✗	✓	72.74
✓	✗	71.93
✓	✓	**73.58**

Replacing both modules results in the lowest accuracy of 70.33%, indicating their critical role in the overall performance. Retaining only the Pairwise-Attention module improves accuracy to 72.74%, proving the effectiveness of its attention mechanism within feature channels, which better preserves the local semantic features of facial information. As a result, the model can more effectively capture fine details during subsequent global attention computations. When keeping only the MS-SNN Conv module improves accuracy to 71.93%, demonstrating its value for feature extraction and its ability to integrate spatial features during down-sampling operation. The highest accuracy of 73.58% is achieved when both modules are included, showcasing their complementary roles. These results confirm that the combination of MS-SNN Conv and Pairwise-Attention significantly enhances model performance, thereby validating the effectiveness of the proposed design.

Furthermore, we also utilize t-SNE [30] to illustrate the feature distribution. The visualization results for each experimental group are shown in **Fig. 8**. The baseline model (**Fig. 8a**) presents a scattered feature distribution with significant overlap between different emotion categories, resulting in low inter-class separability. Introducing the Pairwise-Attention module (**Fig. 8b**) increases the distance between some categories and improves intra-class compactness, although some categories remain mixed. Using the MS-SNN Conv module alone (**Fig. 8c**) further enhances feature separation, making some category boundaries clearer, but certain overlaps still exist. When combining MS-SNN Conv and Pairwise-Attention modules (**Fig. 8d**), the feature distribution reaches an optimal state, with highly compact intra-class features and distinct inter-class boundaries, significantly improving the facial expression recognition performance. This demonstrates that the combination of MS-SNN Conv and Pairwise-Attention modules provides complementary strengths in feature extraction and category distinction, effectively enhancing model performance in facial expression recognition tasks.

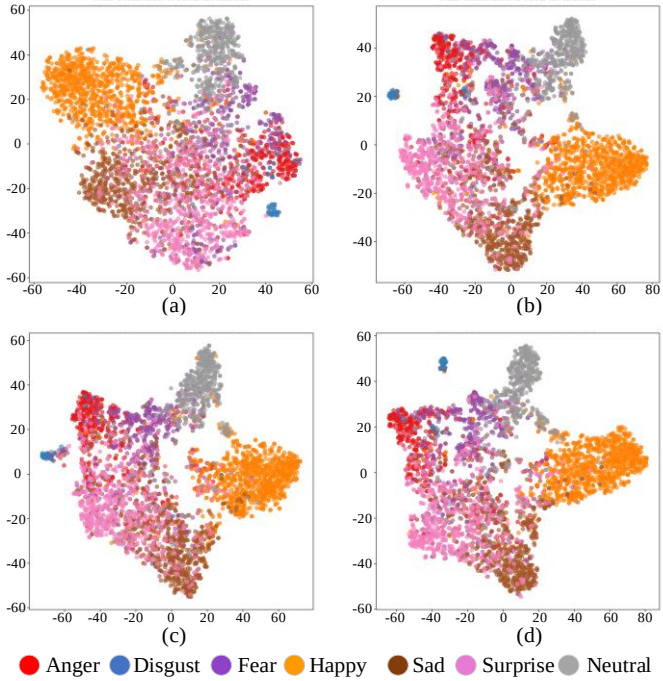

Fig. 8. Visualization of feature distributions for different module configurations (a) Baseline Model (b)Model with pairwise-attention module (c) Model with MS-SNN Conv Module (d)Model with MS-SNN Conv and pairwise-attention modules.

2) Ablation Study on the Number of Blocks

To investigate the effect of module stacking depth on model performance, we conducted a series of controlled experiments on the FER2013 dataset. Inspired by [31-33], a comprehensive exploration for different numbers of SEB (potential values: 1, 2) and SIB (potential values: 2, 4, 6) is undertaken, 6 possible combinations, i.e., (1, 2), (2, 2), (1, 4), (2, 4), (1, 6), and (2, 6), are generated and the corresponding experimental results are provided in **Table VII**.

TABLE VII
IMPACT OF DIFFERENT SEB AND SIB COMBINATIONS ON MODEL PERFORMANCE

Number of blocks			Acc (%)	Params (M)	Acc Gain per M
Spiking Extraction Block	Spiking Integration Block	Total Number			
1	2	3	71.96	5.41	baseline
2	**2**	**4**	**73.58**	**6.93**	**1.07**
1	4	5	73.51	8.22	-0.05
2	4	6	73.72	9.37	0.18
1	6	7	73.99	11.39	0.13
2	6	8	74.06	13.54	0.03

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

From **Table VII,** the results indicate that increasing the total number of blocks improves the model's accuracy. As the number of blocks increases from 3 to 8, the accuracy rises from 71.96% to 74.06%, indicating that deeper architectures enhance the learning capability. However, the performance improvement is accompanied by an increase in model complexity, as the number of parameters grows from 5.41M to 13.54M. Notably, the (2, 2) configuration achieves a balance between accuracy and parameter efficiency, yielding an accuracy of 73.58% with only 6.93 million parameters. yielding the highest accuracy gain per million parameters (1.07) among all configurations. As shown by the diminishing "Acc Gain per M" values with further depth—taking the model with the lowest parameter count as the baseline—the accuracy gain per million parameters decreases as the model's complexity increases. This suggests diminishing returns with increased model complexity, and thus the (2, 2) configuration demonstrates an optimal trade-off, providing a high accuracy with relatively low complexity.

3) Ablation Study on the time steps

The performance of FE-SpikeFormer with different time steps is shown in **Table VIII**. The results indicate that increasing the time step generally improves accuracy, with the accuracy rising from 71.23% at 1 time step to 73.69% at 6-time steps. However, the improvement becomes marginal as the time step increases beyond 4, with only a slight increase from 73.58% (4-time steps) to 73.69% (6-time steps). This suggests that while longer time steps enhance the model's performance by capturing more temporal information, the benefits diminish at higher time steps. Based on this, the selection of the temporal step size is 4.

TABLE VIII
THE RESULTS OF FE-SPIKEFORMER WITH DIFFERENT TIME STEPS

Time step	Acc (%)
1	71.23
2	72.77
4	**73.58**
6	**73.69**

C. Anti-Noise Analysis

In this experiment, we add four types of noise [34, 35] (i.e., Gaussian noise, Pepper noise, Poisson Noise, and Salt Noise) to the images to evaluate the anti-noise ability of the FE-SpikeFormer. Gaussian noise is applied with a variance of 0.001 to simulate random pixel perturbations. Pepper noise is introduced with a density of 0.05, randomly setting 5% of the pixels to 0 (black), while salt noise with the same density replaces some pixels with 1 (white). Additionally, Poisson noise is included, with its intensity proportional to the pixel values. These noise injections provide a comprehensive evaluation of the model's robustness under various noisy conditions. The relevant results are shown in **Table IX**.

TABLE IX
ANTI-NOISE ANALYSIS OF NOISE EFFECT

Noise	Acc (%)
Clean	73.58
Gaussian Noise	71.42
Pepper Noise	71.89
Poisson Noise	67.57
Salt Noise	71.73

Table IX shows the accuracy of the proposed FE-SpikeFormer under different kinds of noise. With Gaussian noise, the accuracy slightly drops to 71.42%, indicating the model's resilience to random perturbations. Under pepper and salt noise, the accuracies remain stable at 71.89% and 71.73%, respectively, showing that the model effectively handles isolated pixel disruptions. Poisson noise, however, causes a more noticeable decline, with the accuracy decreasing to 67.57%. Despite this drop, the model performance keeps in an acceptable range, indicating the solid anti-noise ability of FE-SpikeFormer across various types of noise.

D. Validation in Hospital Environment

To validate the generalization capability of the proposed FE-SpikeFormer in real-world clinical scenarios, we conducted an in-hospital deployment at Lishui Central Hospital. The system configuration and experimental results are presented in **Fig. 9** and **Table X**. A remote camera was mounted at a fixed angle of 35.5° to ensure unobstructed and complete capture of patients' facial regions, while the well-trained model was deployed on a local edge computing device for real-time expression recognition. In accordance with

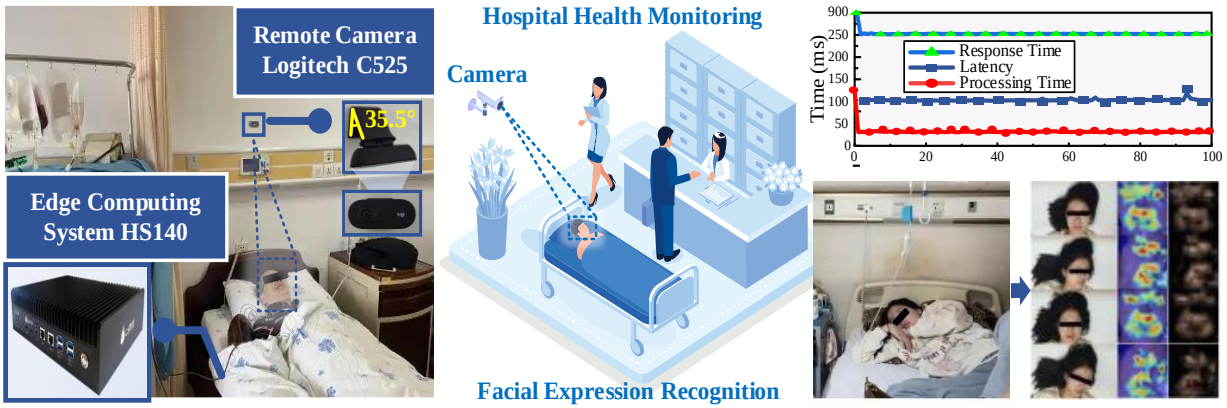

Fig. 9. Validation of the proposed FE-SpikeFormer in hospital scenario

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JBHI.2025.3589267, IEEE Journal of Biomedical and Health Informatics

ethical regulations, all raw video data were processed on-site and the de-identified image data is available for verification purposes. The specific dataset information is provided in Table III. Then, the anonymized classification outputs (e.g., "surprise detected at 14:20") were recorded and shared with clinical personnel. This privacy-preserving deployment demonstrates the model's practical applicability in clinical environments, meeting real-time performance demands under low-power and low-latency constraints.

TABLE X
THE RESULTS OF FE-SPIKEFORMER IN HOSPITAL ENVIRONMENT

Emotion	Acc (%)
Anger 	69.46%
Disgust 	75.29%
Fear 	62.73%
Happy 	92.03%
Sad 	71.96%
Surprise 	59.27%
Neutral 	84.28%
Average	73.57%

From **Fig. 9** and **Table X**, the experimental results demonstrate that the proposed FE-SpikeFormer achieves high recognition accuracy for happy (92.03%) and neutral (84.28%) facial expressions. This high performance is crucial for real-time emotion detection in clinical settings, as it allows healthcare providers to accurately assess a patient's emotional state and formulate appropriate personalized care plans and treatment decisions. Meanwhile, the classification performance for anger, disgust, and sadness also reaches acceptable levels, with recognition accuracy around 70%. These negative emotions are particularly important in clinical contexts, where effective monitoring can support the management of patient stress, anxiety, or pain. Compared to positive emotions, the recognition rates for negative emotions are lower, primarily due to the imbalanced distribution of the data. In addition, the proposed FE-SpikeFormer achieves a response time, latency, and processing time of approximately 0.25s, 0.10s, and 0.03s, respectively. These low-latency characteristics are critical in hospital scenarios, where rapid, real-time emotional feedback is essential for guiding immediate medical interventions. The ability to operate efficiently and with minimal delay ensures the proposed FE-SpikeFormer can be seamlessly integrated into clinical environments, providing valuable emotional insights without compromising the workflow of healthcare professionals.

V. CONCLUSION

This paper focuses on the development of a camera-based facial expression recognition method, i.e., FE-SpikeFormer, for hospital health monitoring. Specifically, a dual attention-based architecture is proposed, which combines both local and global attention mechanisms to capture fine-grained and contextual features, thereby enhancing feature representation in spiking neural networks (SNNs). Then, a feature integration strategy is developed, which facilitates the integration of spatial features in each stage and enhances the representation capability of spiking signals, indicating better overall model performance. For joint validation, the proposed FE-SpikeFormer is applied to realize facial expression recognition task in both the realistic laboratory environment and the real-world hospital scenario. The experimental results and comprehensive analysis (including anti-noise and computational energy analysis) show that the proposed FE-SpikeFormer achieves the good balance between recognition accuracy (FER2013 dataset: 73.58%; AffectNet: 59.90%), model size (6.93 M) and energy consumption (7.40 mJ), outperforming SOTA methods. Additionally, the proposed FE-SpikeFormer is further validated in the hospital environment with an average accuracy of 73.57%, satisfying practical processing requirements.

VI. DISCUSSION

The proposed FE-SpikeFormer advances facial expression recognition for hospital health monitoring by effectively balancing accuracy, energy efficiency, and deployability. Its performance has been validated in both laboratory settings and real-world clinical environments. However, the model achieves higher classification accuracy for positive emotions than for negative emotions due to the imbalanced data distribution. Based on these observations, our future research will explore two key directions: 1) Constructing a facial expression dataset with balanced data distribution in clinical environments; 2) Enhancing the recognition accuracy of negative emotions through more effective feature extraction (potentially by integrating physiological signals)

REFERENCES

[1] M. Chen et al., "Negative information measurement at AI edge: A new perspective for mental health monitoring," *ACM Trans. Internet Technol.*, vol. 22, no. 3, pp. 1-16, Aug. 2022, doi: 10.1145/3471902.

[2] B. Zha et al., "Intelligent wearable photonic sensing system for remote healthcare monitoring using stretchable elastomer optical fiber," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 17317-17329, May 15, 2024, doi: 10.1109/JIOT.2024.3356574.

[3] Z. Dong, X. Ji, C. S. Lai, D. Qi, G. Zhou, and L. L. Lai, "Memristor-based hierarchical attention network for multimodal affective computing in mental health monitoring," *IEEE Consum. Electron. Mag.*, vol. 12, no. 4, pp. 94-106, Jul. 2023, doi: 10.1109/MCE.2022.3159350.

[4] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, and A. Lanata, "Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets," *IEEE Access*, vol. 12, no. 1, pp. 45543-45559, Mar. 2024, doi: 10.1109/ACCESS.2024.3380847.

[5] A. A. O. Díaz, S. C. Tamayo, D. N. De Oliveira, and G. A. Abensur, "Models for real-time emotion classification: FER-2013 dataset," *Intell. Syst. Appl*., K. Arai, Ed., *Lect. Notes Networks Syst.*, vol. 1067, pp. 231-241, Jul. 2024, doi: 10.1007/978-3-031-66431-1_19.

[6] M. Rabea et al., "IdentiFace: A VGG-based multimodal facial biometric system," pp. 1-12, Jan. 2024, doi: 10.48550/arXiv.2401.01227.

[7] A. P. Fard and M. H. Mahoor, "Ad-Corre: Adaptive correlation-based loss for facial expression recognition in the Wild," *IEEE Access*, vol. 10, pp. 26756-26768, Mar. 2022, doi: 10.1109/ACCESS.2022.3156598.

[8] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State-of-the-art performance on FER2013," pp. 1-9, May 2021, doi: 10.48550/arXiv.2105.03588.

[9] Z. Y. Huang et al., "A study on computer vision for facial emotion recognition," *Sci. Rep.*, vol. 13, no. 8425, pp. 1-13, May 2023, doi: 10.1038/s41598-023-35446-4.

[10] J. Li, J. Nie, D. Guo, R. Hong, and M. Wang, "Emotion separation and recognition from a facial expression by generating the poker face with vision transformers," *IEEE Trans. Comput. Soc. Syst.*, Early Access, 2024, doi: 10.1109/TCSS.2024.3478839.

[11] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "OAENet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognit.*, vol. 112, pp. 1-14, Apr. 2021, doi: 10.1016/j.patcog.2020.107694.

[12] X. Zhang, F. Zhang, and C. Xu, "Joint expression synthesis and representation learning for facial expression recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1681-1695, Mar. 2022, doi: 10.1109/TCSVT.2021.3056098.

[13] H. Li, N. Wang, X. Yang, X. Wang, and X. Gao, "Unconstrained facial expression recognition with no-reference de-elements learning," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 173-185, Mar. 2024, doi: 10.1109/TAFFC.2023.3263886.

[14] X. Ji, Z. Dong, Y. Han, C. S. Lai, G. Zhou, and D. Qi, "EMSN: An energy-efficient memristive sequencer network for human emotion classification in mental health monitoring," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 1005-1016, Nov. 2023, doi: 10.1109/TCE.2023.3263672.

[15] X. Ji, Z. Dong, Y. Han, C. S. Lai, and D. Qi, "A Brain-inspired hierarchical interactive in-memory computing system and its application in video sentiment analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7928-7942, Dec. 2023, doi: 10.1109/TCSVT.2023.3275708.

[16] G. Pu and J. Chen, "Facial expression recognition based on convolutional spiking neural network and STDP fine-tune," *Preprints*, pp. 1-18, Jan. 2024, doi: 10.20944/preprints202401.2165.v1.

[17] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," *Proc. 11th Int. Conf. Learn. Represent.*, pp. 1-19, Nov. 2022. doi: 10.48550/arXiv.2209.15425.

[18] M. Yao et al., "Spike-driven Transformer V2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips," pp. 1-24, Feb. 2024, doi: 10.48550/arXiv.2404.03663.

[19] A. Taherkhani, A. Belatreche, Y. Li, G. Cosma, L. P. Maguire, and T. M. McGinnity, "A review of learning in biologically plausible spiking neural networks," *Neural Netw.*, vol. 122, pp. 253-272, Feb. 2020, doi: 10.1016/j.neunet.2019.09.036.

[20] M. Yao et al., "Attention spiking neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9393-9410, Aug. 2023, doi: 10.1109/TPAMI.2023.3241201.

[21] J. Wang, Y. Chen, X. Ji, Z. Dong, M. Gao, and Z. He, "SpikeTOD: A biologically interpretable spike-driven object detection in challenging traffic scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 21297-21314, Dec. 2024, doi: 10.1109/TITS.2024.3468038.

[22] Y. Hu, Q. Zheng, X. Jiang, and G. Pan, "Fast-SNN: Fast spiking neural network by converting quantized ANN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14546-14562, Dec. 2023, doi: 10.1109/TPAMI.2023.3275769.

[23] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," *Neural Inf. Process.*, pp. 145-153, Nov. 2013, doi: 10.1007/978-3-642-37465-8_18.

[24] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18-31, Jan.-Mar. 2019, doi: 10.1109/TAFFC.2017.2740923.

[25] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998-6008, Jun. 2017, doi: 10.48550/arXiv.1706.03762.

[26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. Int. Conf. Learn. Represent.*, pp. 1-22, Jun. 2021. doi: 10.48550/arXiv.2010.11929.

[27] A. Arias-Duart, E. Mariotti, D. Garcia-Gasulla, and J. M. Alonso-Moral, "A confusion matrix for evaluating feature attribution methods," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 17-24, Jun. 2023, doi: 10.1109/CVPRW59228.2023.00380.

[28] X. Ji, Z. Dong, G. Zhou, C. S. Lai, and D. Qi, "MLG-NCS: Multimodal local–global neuromorphic computing system for affective video content analysis," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 54, no. 8, pp. 5137-5149, Aug. 2024, doi: 10.1109/TSMC.2024.3392732.

[29] X. Qiu, R.-J. Zhu, Y. Chou, Z. Wang, L.-J. Deng, and G. Li, "Gated attention coding for training high-performance and efficient spiking neural networks," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 1, pp. 601-610, Mar. 2024, https://doi.org/10.1609/aaai.v38i1.27816.

[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res*., vol. 9, pp. 2579-2605, Nov. 2008.

[31] M. Wen et al., "Multi-agent reinforcement learning is a sequence modeling problem," *Proc. Int. Conf. Neural Inf. Process. Syst.*, pp. 16509-16521, 2022.

[32] C. Yang, J. Xu, S. De Mello, E. J. Crowley, and X. Wang, "GPViT: A high resolution non-hierarchical vision transformer with group propagation," pp. 1-19, Apr. 2023, doi: 10.48550/arXiv.2212.06795.

[33] Z. Tu et al., "MaxViT: Multi-axis vision transformer," *ECCV 2022. Lecture Notes in Computer Science*. pp. 459-479, Nov. 2022, https://doi.org/10.1007/978-3-031-20053-3_27

[34] G. Zhang, C. Li, and X. Xiong, "Analysis and comparison of four signal processing schemes for noise reduction in chaotic communication systems and application of LDPC code," *Chaos Solitons Fractals*, vol. 186, p. 115184, Sep. 2024, doi: 10.1016/j.chaos.2024.115184.

[35] Y. Huang et al., " An accelerated anti-noise adaptive neural network for robotic flexible endoscope with multitype surgical objectives and constraints," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 55, no. 2, pp. 900-1003, Feb. 2025, doi: 10.1109/TSMC.2024.3492324.