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: DOI10.1109/TSMC.2024.3392732

MLG-NCS: Multimodal Local-Global Neuromorphic Computing System for Affective Video Content Analysis

Xiaoyue Ji, *Member, IEEE*, Zhekang Dong, *Senior Member, IEEE*, Guangdong Zhou, Chun Sing Lai, *Senior Member, IEEE*, Donglian Qi, *Senior Member, IEEE*

***Abstract*—Despite neuromorphic computing (NC) technologies offer tremendous potential in executing computationally intensive tasks with high efficiency and low latency, most of existing methods are still difficult to achieve software-comparable accuracy. To address this challenge, we develop a multimodal local-global neuromorphic computing system (MLG-NCS) that can capture local characteristics and exchange global cross-modal information sufficiently. Specifically, a high-density memristor crossbar array is prepared to perform efficient parallel in-memory operations, serving as the fundamental component of the proposed MLG-NCS. To facilitate understanding of the proposed MLG-NCS design, the local feature representation module, the global cross-modal interaction module, and the output module are designed. The experimental results show that the proposed system has advantages in classification accuracy (ranked top three), time consumption (approximately10 times speed up), and latency (about 1.2~15.3 times faster), enabling good inter-related trade-offs between latency, efficiency, and accuracy. This study is expected to promote the revolution and development of next-generation computing system, which takes a firm step toward artificial general intelligence (AGI).**

***Index Terms*—Circuit design, multimodal learning, neuromorphic computing system, affective video content analysis.**

I. Introduction

The long-term goal for artificial intelligence (AI) is to mimic the human level cognitive activities when dealing with complex multimodal information [1-3]. Emotion is an important aspect of human intelligence and is shown to play a significant role in the human learning, memory, decision-making, and communication [4-6]. Recently, with the rapid development of AI, various deep learning algorithms have achieved tremendous successes in affective video content analysis, which can recognize emotion accurately and automatically [7-20]. However, these von Neumann architecture-based deep learning algorithms are usually computationally intensive and suffer from a lack of real-time processing capability with relatively low computational efficiency. In addition, the physical separation of processing and memory units in the von Neumann computing system leads to large latency in data shuffling between different units.

Recently, there has been an increasing number of researches involving neuromorphic computing systems (NCSs) with different applications [21-26], which are capable to perform parallel in-memory operations, enabling greatly improved energy efficiency and computing speed. Although existing NCSs show remarkable advantages in low latency and ultralow-power hardware implementation, the accuracy is inferior to von Neumann architecture-based deep learning algorithms. That is, the inter-related trade-offs between latency, efficiency, and accuracy are hard to balance. At the device level, the device variations may cause low precision in neuromorphic computing system because of the non-uniformity of the switching function layers and electrodes. At the system level, the lack of an efficient remedy for the robust hardware implementation of general feature representation and cross-modal interaction. Specifically, current studies always use different feature representation modules to capture the local characteristics of different modalities, while the intra-modal interactions within each modality are not considered. Meanwhile, almost existing NCSs focus on single-mode information processing and bi-modality information interactions, while global cross-modality information interactions are insufficient. At the algorithm level, the existing ex-situ training methods always suffer from various non-ideal circuit factors (e.g., noise influence), and the in-situ training methods are inevitably faced with the problem of hardware loss.

To fully exploit the potential of NCSs in practical scenarios, this paper aims to investigate a multimodal local-global neuromorphic computing system (MLG-NCS). For clarity, we

Manuscript received July 20, 2023, revised January 15, 2024.

This work was supported in part by the National Postdoctoral Researcher Support Program under Grant GZB20230356, the Shuimu Tsinghua Scholar program under Grant 2023SM035, the National Natural Science Foundation of China under Grant 62206062, and the Fundamental Research Funds for the Provincial University of Zhejiang under Grant GK229909299001-06. (*Corresponding authors: Zhekang Dong and Chun Sing Lai*).

X. Ji is with the Center for Brain-Inspired Computing Research (CBICR), Beijing Innovation Center for Future Chip, Optical Memory National Engineering Research Center, Department of Precision Instrument, Tsinghua University, Beijing 100084, China. (e-mail: jixiaoyue@mail.tsinghua.edu.cn).

Z. Dong is with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: englishp@126.com).

G. Zhou is with the College of Artificial Intelligence, Southwest University, Chongqing 400715, China (e-mail: zhougd@swu.edu.cn).

C. S. Lai is with the Department of Electronic and Electrical Engineering, Brunel University London, London, UB8 3PH, UK and also with the School of Automation, Guangdong University of Technology, Guangzhou, China 510006 (email: chunsing.lai@brunel.ac.uk).

D. Qi is with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: qidl@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: DOI10.1109/TSMC.2024.3392732

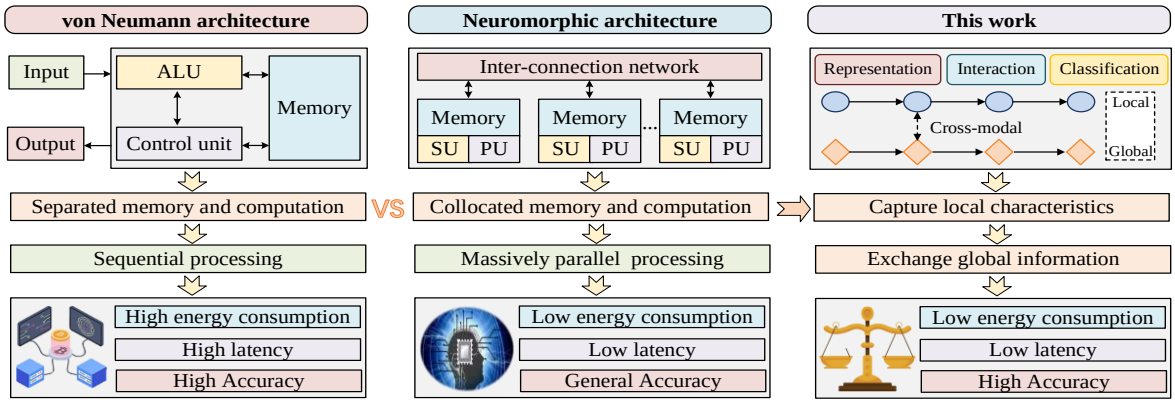


Fig. 1. Comparison of the existing computing architecture.

systemic compare the existing computing architecture, as shown in Fig. 1. The main contributions of this paper are concluded as follows:

1) Different from the existing NCSs, the MLG-NCS is developed that can capture local characteristics and exchange global cross-modal information sufficiently.

2) As the fundamental component, a high-density memristor crossbar array is constructed after fabrication of a good endurance, long retention time, and high stability Au/Fe_2O_3/Fe_2O_3/FTO memristor, enabling high precision parallel computing in the proposed MLG-NCS.

3) Combining the advantages of both in-situ training and ex-situ training, a hybrid training strategy is applied in the proposed MLG-NCS, which achieves good inter-related trade-offs between latency, efficiency, and accuracy in affective video content analysis.

The rest of this paper is organized as follows. Section II describes the overall architecture of the proposed MLG-NCS. Section III describes the fabrication of high-density memristor crossbar array. Section IV presents the specific circuit design of local feature representation module, global cross-modal interaction module, and output module. In Section V, the proposed MLG-NCS is applied to perform affective video content analysis. Finally, the entire work is summarized in Section VI.

II. Multimodal Local-Global Neuromorphic Computing System Architecture

In this work, we propose a novel NCS for affective video content analysis, which can efficiently sense and process the multimodal information from complex environments, as shown in Fig. 2. To facilitate understanding of the MLG-NCS design, we describe it using the following three modules.

Local Feature Representation Module: Inspired by [27], we propose a local feature representation module with cascade configuration to capture the unique local characteristics from multimodal information. The local feature representation module mainly consists of the convolution unit, the bidirectional gated recurrent (Bi-GRU) unit, and the self-attention unit. Specifically, the convolution unit is used to extract local information and unify the dimensions of multimodal information, which is presented as X_{ic}=Conv(X_i) (X_i is the multimodal information, i=t, a, v is the index of the modality). Then, the feature information X_{ic} generated from convolution unit is injected to the Bi-GRU unit, which can capture the high-level sequential feature information. The output of Bi-GRU unit is symbolized by X_{iH}=Bi-GRU(X_{ic}). Furthermore, the self-attention unit is utilized to capture abundant contextual information, which is presented as X_{iself}=Self-attention(X_{iH}). Finally, the local feature representations X_m, m=α, β, γ can be obtained by feed-forward and element-wise additive operations.

Global Cross-modal Interaction Module: The global cross-modal interaction module is designed to perform full multimodal information interaction. Firstly, the average pooling operation is performed on the local feature representations X_m to acquire the expected average feature representations $\overline{X}_m$. For a better understanding, we take audio average feature $\overline{X}_\alpha$ as an example. The cross-attention mechanism [28] is used, taking audio feature representation X_α as values, audio average feature representation $\overline{X}_\alpha$ as keys and the cartesian product $\overline{X}_\beta \otimes \overline{X}_\gamma$ as queries. The attention scores and weights of audio modality S_α can be obtained directly. Similarly, the attention scores and weights of visual modality S_β and text S_γ can be acquired by the same cross-attention operations, which is important for multimodal information fusion. Finally, the three parallel feed forward units are adopted to process the channel-wise information for different modalities, and the outputs of global cross-modal interaction module are present as $X_{m,out.}$, m=α, β, γ.

Output Module: In human brain, the most crucial characteristics in multimodal information are usually given priority attention. Inspired by this processing mechanism, the output module is proposed to extract the key features in the integrated multimodal representations $\sum X_{m,out}$, which mainly consists of the attention unit, the fully connected unit, and the softmax unit. The attention mechanism is employed to capture the weight distribution in the integrated multimodal representations $\sum X_{m,out}$, and generate the global representations of the multimodal information. Then, the global representations are entered into the fully connected layer and softmax layer to generate the final classification X_{out}.

III. High-Density Memristor Crossbar Array

Considering the proposed MLG-NCS needs to execute heavy storage and computing operations, a high-density memristor

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: DOI10.1109/TSMC.2024.3392732

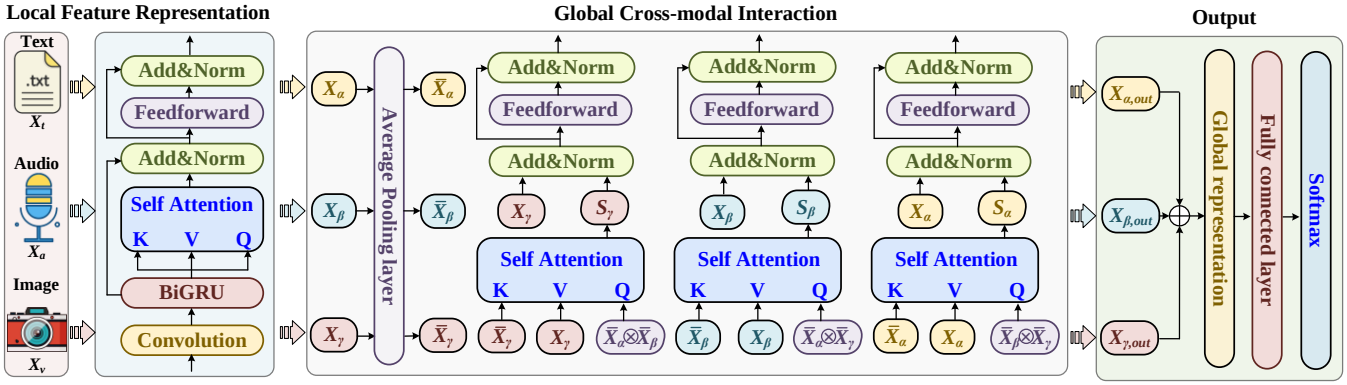


Fig. 2. Schematic of multimodal local-global neuromorphic computing system.

crossbar array is fabricated to perform parallel multiply-accumulate (MAC) operations.

A. High Performance Memristor Fabrication

In this work, the hydrothermal method and magnetron sputtering method [29] are employed to fabricate a highly stable Fe_2O_3-based memristor. The hydrothermal method is used to generate the functional layer, and the magnetron sputtering method is employed to deposit Au top electrode. The specific fabrication process can be provided as follows (Fig. 3).

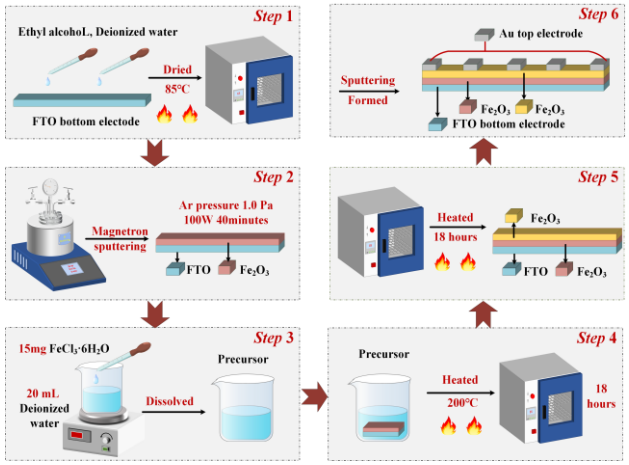


Fig. 3. Flow chart for the fabrication of $Au/Fe_2O_3/Fe_2O_3/FTO$

Step 1: The ethyl alcohol and deionized water are applied to clean possible contaminates on the F-doped-SnO_2 (FTO) bottom electrode, and the FTO bottom electrode is placed to an oven and dried at 85℃.

Step 2: The Fe_2O_3 film is deposited on the FTO bottom electrode to form Fe_2O_3/FTO sample in an ambient atmosphere by magnetron sputtering method.

Step 3: 15 mg $FeCl_3 \cdot H_2O$ is dissolved into 20 ml deionized water to prepare precursor solution.

Step 4: The Fe_2O_3/FTO sample is transferred into precursor solution, then heated in a muffle furnace at 200°C for 18 hours.

Step 5: The FeO_x homojunction is generated from the heated solution by hydrothermal reaction.

Step 6: The Au top electrode is deposited on the FeO_x homojunction by magnetron sputtering method, further developing the $Au/Fe_2O_3/Fe_2O_3/FTO$ memristor.

The electrical characteristics of the $Au/Fe_2O_3/Fe_2O_3/FTO$ memristor are measured within the scanning voltage range of [−2V, 2V], as shown in Fig. 4.

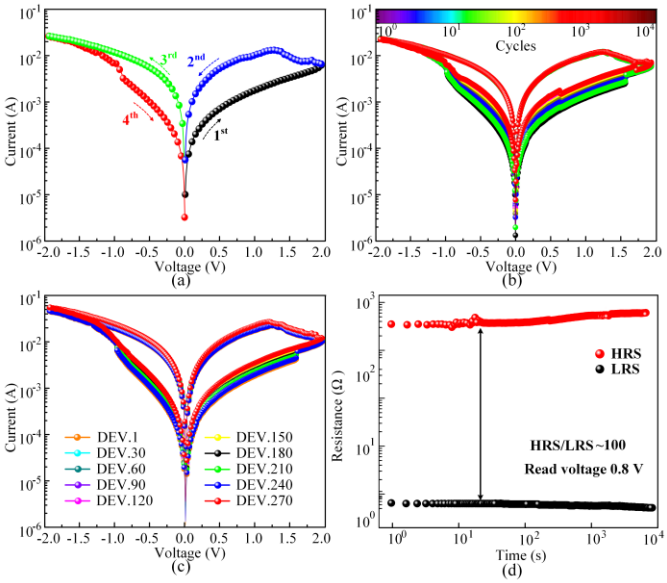


Fig. 4. (a) I-V curve of $Au/Fe_2O_3/Fe_2O_3/FTO$ memristor; (b) C2C analysis; (c) D2D analysis; (d) The stability of HRS and LRS of the prepared $Au/Fe_2O_3/Fe_2O_3/FTO$ memristor.

The current-voltage (I-V) curves demonstrate that the fabricated $Au/Fe_2O_3/Fe_2O_3/FTO$ memristor exhibits obvious self-selective analogue resistance switching (RS) memory behavior, as shown in Fig. 4(a). Specifically, In the first phase, the current gradually increases during the scanning voltage from 0 V to 2 V. In the second phase, the current gradually decreases to a very low value during the scanning voltage from 2 V to 0 V. In the third phase, the current gradually increases to a relatively high value during the scanning voltage sweeps from 0 V to −2 V. In fourth phase, the reverse scanning voltage (−2 V→0 V) is applied to memristor, the current naturally decreases to an ultralow value. It is noted that the fabricated memristor is in the high resistance state (HRS) within the scanning voltage range of [−0.5 V, 0 V] and [0 V, 0.5 V], and the memristor is in low resistance state (LRS) under a high scanning voltage region [−2 V, −0.5 V] and [0.5 V, 2 V]. In order to study the stability of the $Au/Fe_2O_3/Fe_2O_3/FTO$ memristor, over 10000^{th} I–V curves are measured on the same memristor, as shown in Fig. 4(b). The self-selective analogue RS memory behavior can be maintained well, which illustrates high cycle-to-cycle (C2C) stability of the fabricated $Au/Fe_2O_3/Fe_2O_3/FTO$ memristor. Fig. 4(c) shows extensive overlap I–V curves measured by the 270 randomly chosen

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: DOI10.1109/TSMC.2024.3392732

memristors. The results demonstrated that the fabricated Au/Fe_2O_3/Fe_2O_3/FTO memristors exhibit good device-to-device (D2D) stability. Furthermore, a stable resistance ratio (about 100) between the HRS and LRS is well maintained for 10^4 seconds at 0.8 V read voltage, as shown in Fig. 4(d).

B. Memristor Crossbar Array

In this paper, a high-density memristor crossbar array is mainly used to conduct fast MAC operation for neuromorphic computing (NC). As shown in Fig. 5, each 1-transistor-1-memristor (1T1M) cell in high-density memristor crossbar array is implemented by the prepared Au/Fe_2O_3/Fe_2O_3/FTO memristor. The complementary metal oxide semiconductors (CMOS) peripheral circuits such as bit-line (BL), word-line (WL), and source-line (SL) drivers connect at one end of the memristor crossbar array. When the input voltages are applied to the BLs through the BL drivers, the weight can be expressed by the conductance of memristor. Meanwhile, the output currents achieved by Ohm's and Kirchhoff's laws is send to SL registers through corresponding SL switch.

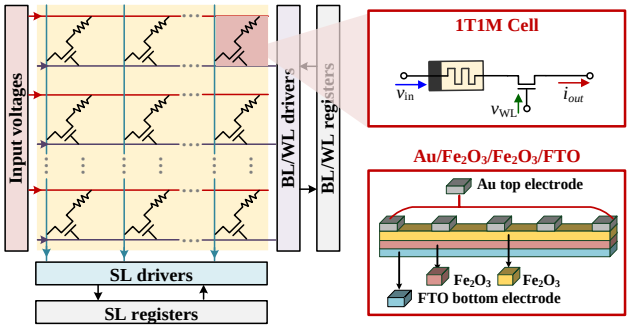


Fig. 5. High-density memristor crossbar array.

We experimentally implemented the 32 × 32 memristor crossbar array, as shown in Fig. 6(a).

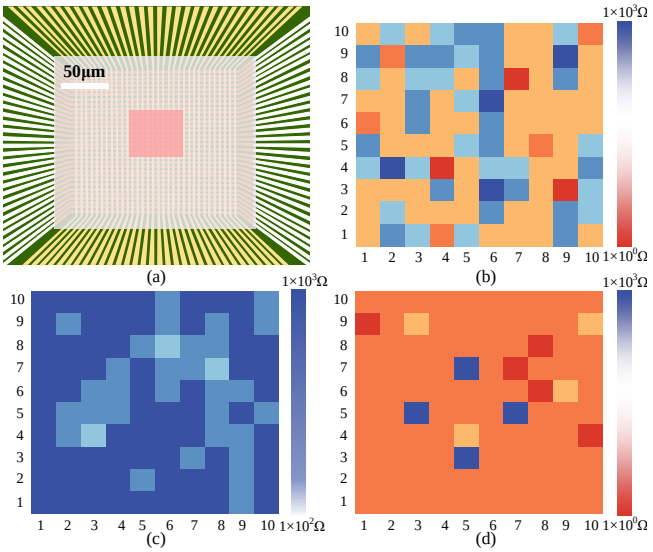


Fig. 6. (a) The optic image of the 32 × 32 memristor crossbar array; (b) The stable resistance ratio between the HRS and LRS of the selected region; (c) The resistances of the selected memristors are operated to the HRS; (d) The resistances of all the selected memristors except the four target memristors are modulated to the LRS.

The selected 10 × 10 memristor crossbar array labeled by pink region, and the resistance of each memristor is operated to an intermediate value between LRS to HRS, as shown in Fig. 6(b). Fig. 6(c) and Fig. 6(d) demonstrate the resistance response of the selected region during the crossbar array modulation operation. From Fig. 6(c), the resistances of the selected memristors are operated to the HRS after 50 cycles. After 50 cycles, the resistances of all the selected memristors except the four target memristors are modulated to the LRS, as shown in Fig. 6(d).

IV. Circuit Design of Multimodal Local-Global Neuromorphic Computing

NC is the potential candidate to break von Neumann bottleneck and provide a new way towards AGI [30]. Our motivation is to design a novel NCS, aimed at realizing high computational accuracy with low computational overhead.

A. Local Feature Representation Module

The local feature representation module is proposed to capture unique characteristics capture features V_m (m=α, β, γ) from the text, audio, and visual modalities V_i (i=t, a, v). The specific circuit architecture of the proposed local feature representation module is shown in Fig. 7.

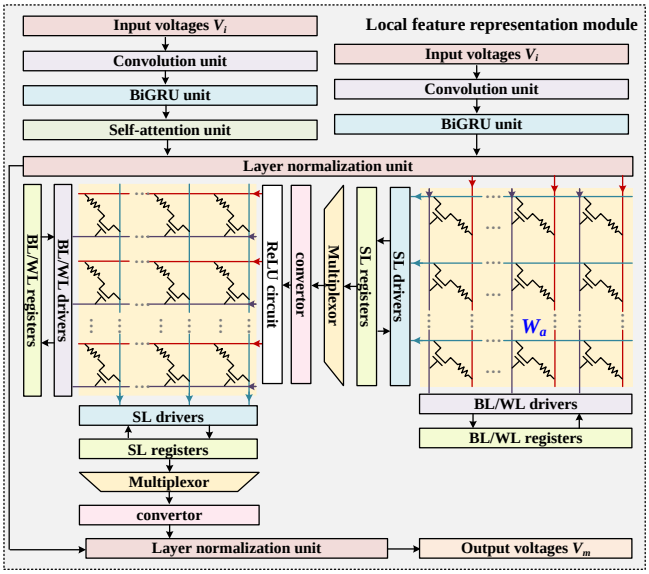


Fig. 7. The circuit architecture of the local feature representation module.

Notably, considering the circuit design of the layer normalization and the ReLU circuits are proposed in our previous work [26], this paper mainly focuses on the following circuits design.

1) Circuit design of convolution unit

The convolution unit is composed by several convolution kernels, and the feature voltages with unified dimensions can be obtained by the convolution computing of kernels and input voltage (containing text, audio, and visual information). In this paper, the convolution unit is constructed using the prepared high-density memristor crossbar array and some peripheral circuits, as shown in Fig. 8.

From, Fig. 8, V_i is the input voltage of the convolution unit (i is the index of the modality. i= t, a, v), V_{ic}=$Conv$(V_i) is 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: DOI10.1109/TSMC.2024.3392732

output voltage of the convolution unit. The prepared 1T1M crossbar array is used to perform highly efficient convolutional operation, in which the number of columns *M* and *N* are equal to the number of input and output channels, respectively.

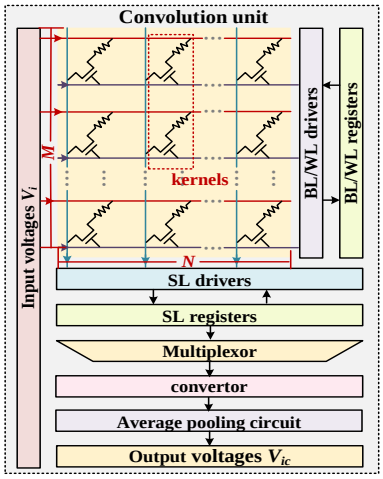


Fig. 8. Circuit design of the convolution unit.

2) Circuit design of Bi-GRU unit

Bi-GRU unit is employed to capture the high-level sequential feature information from feature voltage generated from convolution unit. According to [27], the Bi-GRU unit is comprised by two opposite-direction GRU unit, and the specific circuit architecture of the proposed GRU unit is shown in Fig. 9.

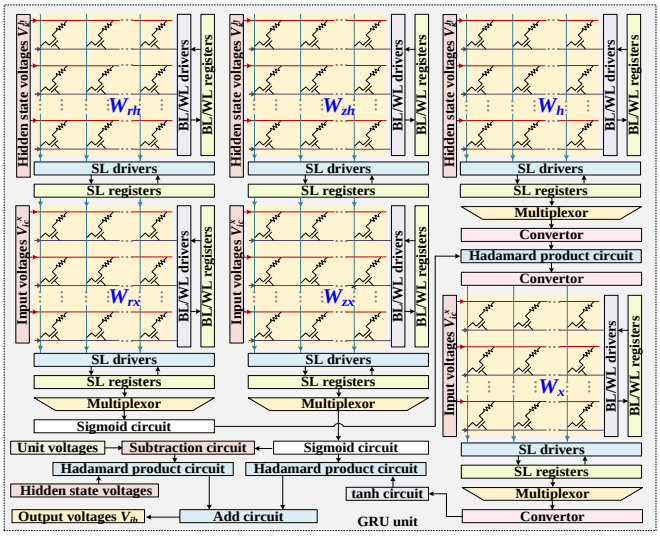


Fig. 9. Circuit design of the GRU unit.

From Fig. 9, the GRU unit mainly consists of the fabricated 1T1M crossbar arrays, sigmoid circuit, subtraction circuit, add circuit, tanh circuit, current-to-voltage convertor, and Hadamard product circuit. Specifically, the difference in conductance between memristors represents the weight of the GRU unit. The 1T1M crossbar arrays are used to storage and compute the learnable parameters of the weight matrixes ***W*** (W_{rx}, W_{zx}, W_h, W_x). The output currents from the 1T1M crossbar arrays are injected into the sigmoid circuit and the tanh circuit, respectively. Then, the output voltages $V_{ir}(t)$, $V_{iz}(t)$ of the reset gate and update gate can be obtained. After several operations of activation, multiplication, and summation, the temporary output state voltage $V'_{ih}(t)$ and the output voltage $V_{ih}(t)$can be produced. It is noted that the above-mentioned sub-circuit module have been proposed in [26]. The specific input and output of the GRU unit are mathematically expressed by:

$$V_{ir}(t)=\sigma\left(\begin{array}{l}\sum_{n=1}^{N}\left(G_{rx,n}^{+}-G_{rx,n}^{-}\right)V_{ic}^{x}(t)\\+\sum_{m=1}^{M}\left(G_{rh,m}^{+}-G_{rh,m}^{-}\right)V_{ic}^{h}(t-1)\end{array}\right) \tag{1}$$

$$V_{iz}(t)=\sigma\left(\begin{array}{l}\sum_{n=1}^{N}\left(G_{zx,n}^{+}-G_{zx,n}^{-}\right)V_{ic}^{x}(t)\\+\sum_{m=1}^{M}\left(G_{zh,m}^{+}-G_{zh,m}^{-}\right)V_{ic}^{h}(t-1)\end{array}\right) \tag{2}$$

$$V_{ih}^{'}(t)=\tanh\left(\begin{array}{l}\sum_{n=1}^{N}\left(G_{x,n}^{+}-G_{x,n}^{-}\right)V_{ic}^{x}(t)\\+V_r(t)\odot\sum_{m=1}^{M}\left(G_{h,m}^{+}-G_{h,m}^{-}\right)V_{ic}^{h}(t-1)\end{array}\right) \tag{3}$$

$$V_{ih}(t)=\left(1-V_z(t)\right)\odot V_h(t-1)+V_h^{'}(t)\odot V_z(t) \tag{4}$$

where $V_{ic}{}^{x}(t)$ and $V_{ic}{}^{h}(t\text{-}1)$ are the input voltage and the hidden state voltage, respectively. G_{rx}, G_{rh}, G_{zx}, G_{zh}, G_x, G_h are the conductance of memristors in corresponding 1T1M crossbar array. *N* and *M* are the row of the 1T1M crossbar array.

On this basis, the circuit design of Bi-GRU unit can be obtained, and the output voltage V_{iH} of Bi-GRU unit is mathematically expressed by:

$$V_{iH}=\left[\overrightarrow{V_{ih}(t)},\overleftarrow{V_{ih}(t)}\right] \tag{5}$$

where $\overrightarrow{V_{ih}(t)}$ and $\overleftarrow{V_{ih}(t)}$ are output voltages of forward GRU unit and backward GRU unit, respectively.

3) Circuit design of self-attention unit

The more abundant characteristics from input voltages V_{iH} are extracted by the self-attention unit. The circuit design of the proposed self-attention unit is illustrated in Fig. 10. In Fig. 10, the prepared high-density memristor crossbar arrays are mainly used to store and compute the attention weight matrixes $W_n{}^{Q}$, $W_n{}^{K}$, and $W_n{}^{V}$. Following the attention weight matrixes, the input voltage V_{iH} can be converted to corresponding current vectors representing the attention key *Q*, attention query *K*, and attention value *V*, respectively. The output current vectors are converted to the voltages via current-to-voltage convertor, and then injected to the Hadamard product circuit. The softmax circuit is used to convert the input voltages V_{isoft} from Hadamard product circuit to a set of voltages expressing a probability distribution.

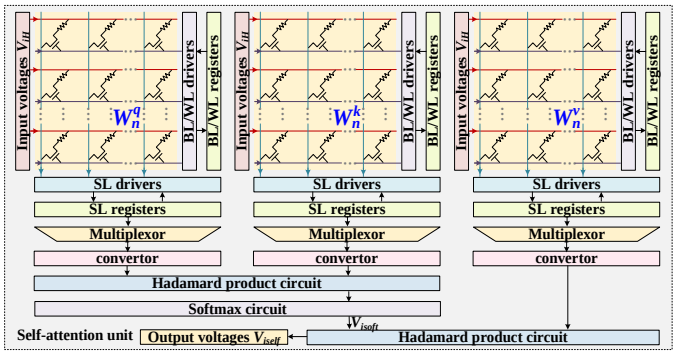


Fig. 10. Circuit design of the self-attention unit.

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: DOI10.1109/TSMC.2024.3392732

After multiplication operation, the output voltage V_{isoft} of self-attention unit can be obtained. The output voltage V_{iself} is mathematically described by:

$$V_{isoft} = soft\max\left(\frac{\text{convertor}\left(W_n^Q V_{iH}\right)^T \cdot \text{convertor}\left(W_n^K V_{iH}\right)^T}{\sqrt{d}}\right) \quad (6)$$

$$V_{iself} = V_{isoft} \cdot \text{convertor}\left(W_n^V V_{iH}\right)^T \quad (7)$$

where T and d are the transpose operation and the dimension of the self-attention unit, respectively.

B. Global Cross-modal Interaction Module

The captured features V_m (m=α, β, γ) are injected into the proposed global cross-modal interaction module, which can sufficiently exchange information between different modalities. The structure of the proposed global cross-modal interaction module is illustrated in Fig. 11.

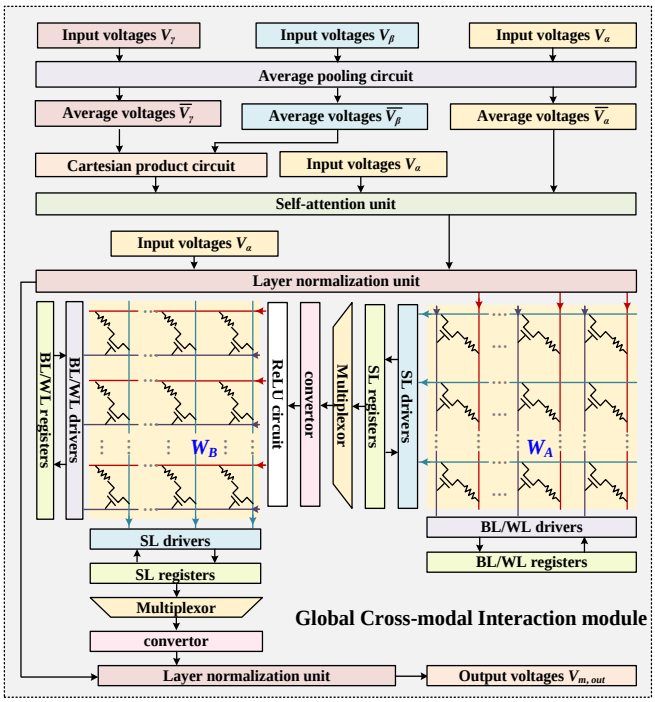


Fig. 11. Circuit design of the global cross-modal interaction module.

Specifically, the input feature signals V_m is fed to the average pooling circuit to generate the expected average feature signals $\overline{V}_m$. Then, the cross-attention mechanism is used to obtain the attention weights and scores. For better comprehending, the audio modality is taken as an example, the attention matrixes are firstly created as:

$$K = W_\alpha^K \cdot \overline{V}_\alpha \quad (8)$$

$$V = W_\alpha^V \cdot V_\alpha \quad (9)$$

$$Q = W_\alpha^Q \cdot \left(\overline{V}_\beta \otimes \overline{V}_\gamma\right) \quad (10)$$

where K, V, and Q are the attention key, value, and query respectively. The attention weight matrixes W_α^K, W_α^V, and W_α^Q are implemented by the prepared 1T1M memristor crossbar arrays. $\otimes$ denotes the Cartesian product operation. V_α and $\overline{V}_\alpha$ are the audio feature signals and average audio feature signals, respectively. $\overline{V}_\beta$ and $\overline{V}_\gamma$ denote the average text and visual feature signals, respectively. Thus, the attention weights of audio modality $V_{S\alpha}$ is symbolized by:

$$V_{S\alpha} = soft\max\left(\frac{W_\alpha^Q \cdot \left(\overline{V}_\beta \otimes \overline{V}_\gamma\right) \cdot \left(W_\alpha^K \cdot \overline{V}_\alpha\right)^T}{d}\right) \cdot W_\alpha^V \cdot V_\alpha \quad (11)$$

Similarly, the attention weights of text and audio modality $V_{S\beta}$ and $V_{S\gamma}$ can be obtained by the same structure:

$$V_{S\beta} = soft\max\left(\frac{W_\beta^Q \cdot \left(\overline{V}_\alpha \otimes \overline{V}_\gamma\right) \cdot \left(W_\beta^K \cdot \overline{V}_\beta\right)^T}{d}\right) \cdot W_\beta^V \cdot V_\beta \quad (12)$$

$$V_{S\gamma} = soft\max\left(\frac{W_\gamma^Q \cdot \left(\overline{V}_\alpha \otimes \overline{V}_\beta\right) \cdot \left(W_\gamma^K \cdot \overline{V}_\gamma\right)^T}{d}\right) \cdot W_\gamma^V \cdot V_\gamma \quad (13)$$

where W_β^K, W_β^V, W_β^Q, W_γ^K, W_γ^V, and W_γ^Q are the attention weight matrixes implemented by the prepared 1T1M memristor crossbar arrays.

Furthermore, the three feed forward units with the parallel configuration are employed to process multi-channel signals V_{Sm} (m=α, β, γ). Finally, the outputs $V_{m,out}$ (m=α, β, γ) of global cross-modal interaction module are mathematically described by:

$$V_{L1} = LN\left(V_m + V_{Sm}\right) \quad (14)$$

$$V_{L2} = \max\left(0, V_{L1} \cdot W_A\right) \cdot W_B \quad (15)$$

$$V_{m,out} = LN\left(V_{L1} + V_{L2}\right) \quad (16)$$

where W_A and W_B are the weight matrixes implemented by 1T1M crossbar array. V_{L1} and V_{L2} are the intermediate results.

C. Output Module

The output module is proposed to extract the key information in the cross-modal interaction signals, which stimulates the processing mechanism of multimodal fusion information in human brain. The specific circuit architecture of the proposed output module is shown in Fig. 12.

From Fig. 12, the output module is mainly comprised by the global representation unit, fully connected circuit, and softmax circuit. The global representation unit is designed to emphasize the features that are crucial for predicting task, which is composed of the prepared 1T1M crossbar array, tanh circuit, Hadamard product circuit, and softmax circuit. The output of the global representation unit is applied to the fully connect circuit and softmax circuit in sequence for multimodal information processing. The specifically input/output of the output module is mathematically expressed by:

$$V_{in} = V_{\alpha,out} + V_{\beta,out} + V_{\gamma,out} \quad (17)$$

$$V_G = soft\max\left(w_l^T \tanh\left(W_L V_{in} + I_B\right)\right) \cdot V_{in} \quad (18)$$

$$V_{out} = soft\max\left(\sum V_G\right) \quad (19)$$

where V_{in} is the input voltage. W_G and w_L denote the weigh matrix and the parameter vector of global representation unit, respectively. I_B is bias current of the global representation unit. V_G and V_{out} are the output voltages of global representation unit and output module, respectively.

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: DOI10.1109/TSMC.2024.3392732

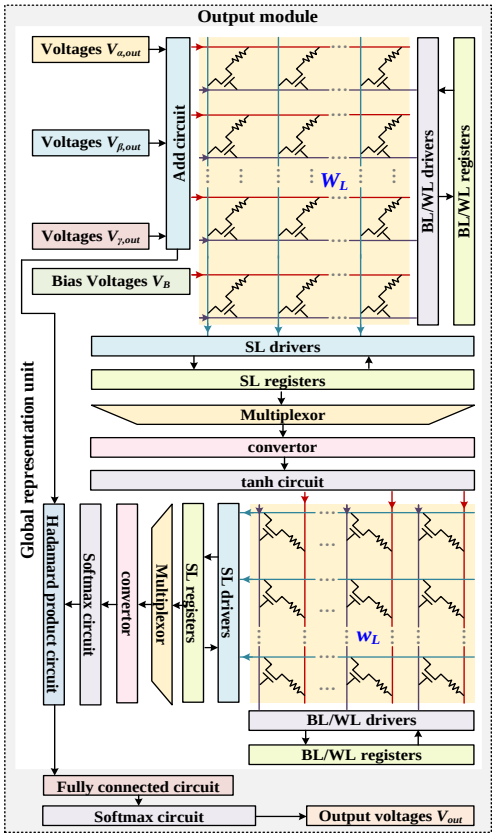


Fig. 12. Circuit design of the output module.

V. Application in Affective Video Content Analysis

The proposed MLG-NCS is further applied to realize affective video content analysis. The parameters (including circuit and neural network parameters) used for the proposed MLG-NCS are provided below:

TABLE I
List of the parameters used for MLG-NCS

	Device	Parameter	
Circuit parameters	1T1M memristor crossbar array	HRS	~$10^3\Omega$
		LRS	~$10^1\Omega$
		Read voltage	0.8V
		SET voltage	2.0V
		RESET voltage	1.5V
		Scan voltage	0.5V/s
	Transistor	RESET gate voltage	5.0V
		Gate width/length ratio	4.3
		Gate voltage	1.1V
		Access resistance	15 KΩ
	ADC	Precision	6 bits
Neural network parameters		Learning rate	10^{-2}
		Momentum	0
		Decay	0.9
		Maximum error	10^{-4}

A. Datasets and Evaluation Metrics

Four benchmark datasets are employed to verify the effectiveness and feasibility of the proposed MLG-NCS, i.e., the LIRIS-ACCEDE dataset, the mediaeval 2015 dataset, the mediaeval 2016 dataset, and the DEAP dataset [10].

The LIRIS-ACCEDE dataset, the largest dataset for affective video content analysis, contains 9800 video excerpts extracted from 160 short movies. All video excerpts in the LIRIS-ACCEDE dataset have discrete levels of valence and arousal. The valence represents the degree of pleasant and unpleasant, while the arousal represents the degree of excitement and calm.

The mediaeval 2015 dataset is proposed for classification task on the LIRIS-ACCEDE dataset. The mediaeval 2015 dataset is an extension of the LIRIS-ACCEDE dataset, which consists of 10900 video excerpts extracted from 199 short movies. In this dataset, 6144 elements are distributed in training dataset and the remaining 4756 elements are distributed in testing dataset.

The mediaeval 2016 dataset is proposed for regression task on the LIRIS-ACCEDE dataset. The mediaeval 2016 consists of 11000 video excerpts extracted from short movies, in which 9800 data elements are distributed in training dataset and the remaining 1200 data elements are distributed in testing dataset. Each video in the mediaeval 2015 and the mediaeval 2016 datasets is labeled in three categories: negative (within the range of [-1, -0.15]), neutral (within the range of [-0.15, 0.15]), and positive (within the range of [0.15, 1]) for valence; clam (within the range of [-1, -0.15]), neutral (within the range of [-0.15, 0.15]), and excited (within the range of [0.15, 1]) for arousal.

The DEAP dataset consists of 120 music video excerpts watched by 32 volunteers. In this work, only 68 music video excerpts are used in our classification task because the length of the remaining excerpts is not enough. We divide the DEAP dataset into two categories: negative (within the range of [-1, 0]) and positive (within the range of [0, 1]) for valence; clam (within the range of [-1, 0]) and excited (within the range of [0, 1]) for arousal.

In addition, the performance of the proposed MLG-NCS is also evaluated on the CMU-MOSI dataset that is composed by 2198 utterances collected form the Internet [16]. In CMU-MOSI dataset, 1283 utterances are included in training dataset, 229 utterances are included in validation dataset, and the remaining utterances are included in testing dataset. Each utterance is uniformly transferred to a [-3, 3] range, indicating the strength of positive and negative emotions.

Then, several performance metrics are used to evaluate the overall performance following the previous works. For the classification task, we report F1-score (F1) and classification accuracy (Acc) as measurements [31, 32]. For the regression task, we report mean absolute error (MAE) and Pearson correlation (Corr) as measurements [33].

B. Hybrid Training Method

The hybrid training method combines the advantages of both in-situ training and ex-situ training, which can reduce hardware loss while ensuring system accuracy. The hybrid training method contains ex-situ training, weights mapping, forward calculation, and weights correction, as illustrated in Fig. 13.

Firstly, the proposed multimodal local-global model is trained ex-situ in PyTorch platform on training datasets. The software performance metrics are taken as the baseline. Then, the well-trained weights are mapping to the 1T1M crossbar array using incremental-pulse write–verify method [22]. Specifically, the acceptance range [G_{targetL}, G_{targetH}] based on

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: DOI10.1109/TSMC.2024.3392732

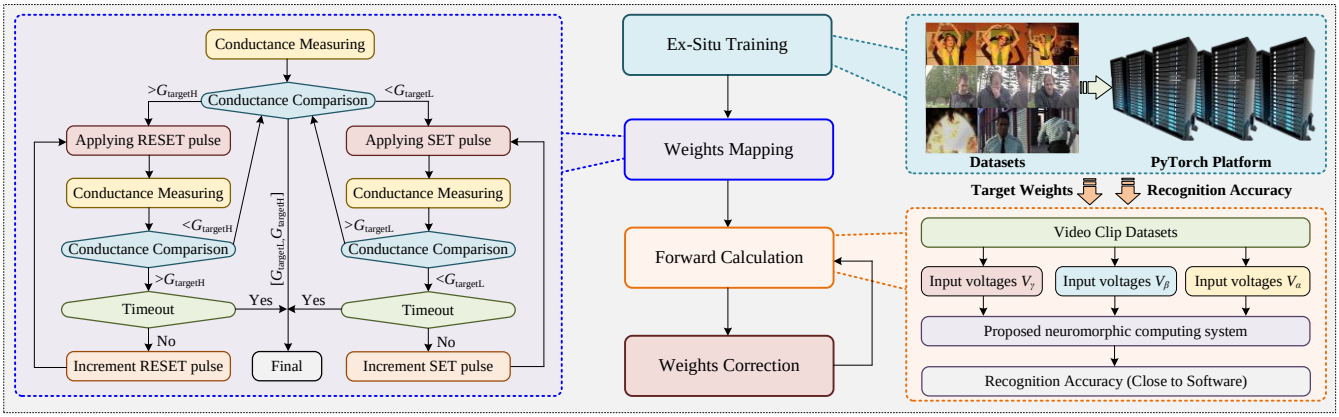


Fig. 13. The flow chart of hybrid training method.

target weight G_{target} is set, and the initial conductance of the memristor is measured. If the conductance is below the lower target weight G_{targetL}, a SET pulse is applied to the corresponding memristor by tuning the gate, aiming at increasing the conductance. After that, the conductance of this memristor is measured again. If the conductance is still below the lower target, a SET pulse with amplitude increased is applied. We repeat set-measure operations until the conductance of memristor is with an acceptance range or over to the higher target weight G_{targetH}. In other case, a reset pulse is applied to the memristor with the same procedure as set operation. When the target weights are transferred to the 1T1M crossbar array, the forward calculation is implemented by the proposed MLG-NCS. In next step, only the output module is trained in-situ to correct the conductance of memristor, which ensures a high performance (closed to the software baseline) on affective video content analysis task.

C. Classification Results and Analysis

In this paper, the well-trained MLG-NCS is used to perform affective video content analysis. The circuit results of the proposed MLG-NCS are illustrated in Fig. 14 to Fig. 17. Specifically, the multimodal information from the benchmark datasets (i.e., the mediaeval 2015 dataset, the mediaeval 2016 dataset, the DEAP dataset, and the CMU-MOSI dataset) can be converted to the voltage signals within the range of [-2, 2] based on digital to analog converter. Notably, image and text information in the benchmark datasets is encoded as voltage maps based on pixel value. The voltage maps are enrolled and fed to the proposed MLG-NCS. The audio information is encoded as the voltage sequence based on amplitude. Fig. 14 ~ Fig. 17 exhibit the selected 10 × 10 image and 5 × 5 text voltage maps V_v and V_t in each period, respectively. The audio voltage signals V_a are labeled as blue solid line in Fig. 14 ~ Fig. 17.

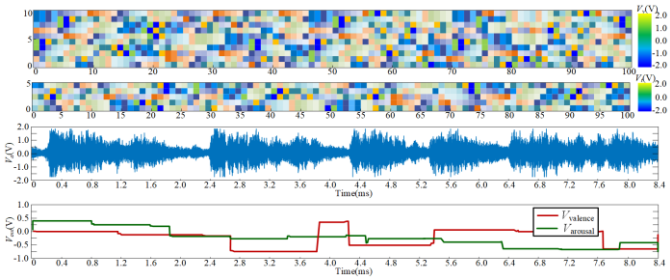


Fig. 14. Circuit results obtained by MLG-NCS on mediaeval 2015 dataset.

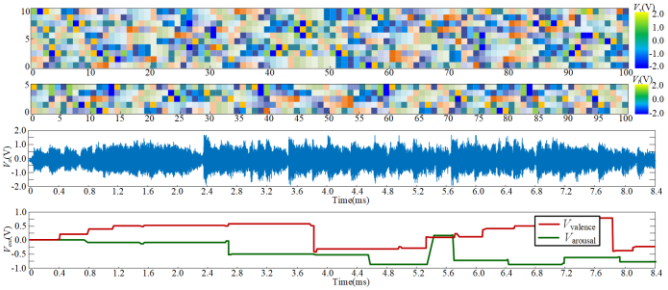


Fig. 15. Circuit results obtained by MLG-NCS on mediaeval 2016 dataset.

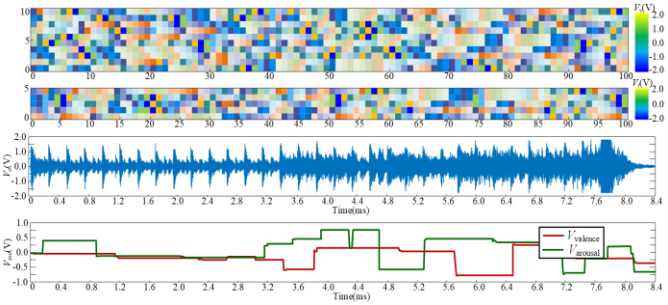


Fig. 16. Circuit results obtained by MLG-NCS on DEAP dataset.

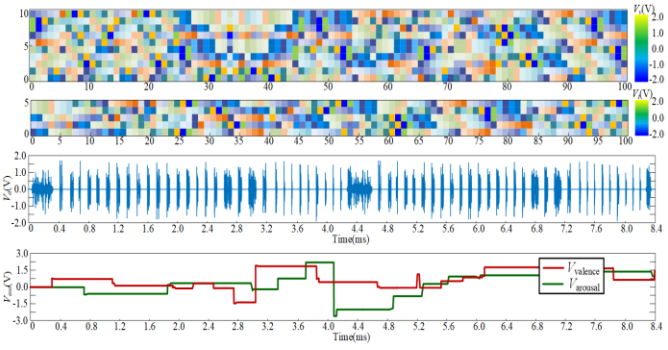


Fig. 17. Circuit results obtained by MLG-NCS on CMU-MOSI dataset.

Then, these input voltages are further injected to the proposed MLG-NCS. The output voltages V_{valence} (red solid line) and V_{arousal} (green solid line) in the two-dimensional emotion state space are obtained, representing corresponding emotional behaviors [10]. In the two-dimensional emotion state space, the X-axis and Y-axis denote valence and arousal, respectively. The output voltages V_{valence} and V_{arousal} are consider as vector with polarity. The V_{valence} with positive or negative polarities represents high- or low-level emotion pleasure. And the $V_{arousal}$ with positive or negative polarities represents high- or low-level mental alertness.

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: DOI10.1109/TSMC.2024.3392732

TABLE II
COMPARISON OF DIFFERENT SOTA METHODS FOR AFFECTIVE VIDEO CONTENT ANALYSIS

Method	Mediaeval 2015				Mediaeval 2016				DEAP			
	Valence		Arousal		Valence		Arousal		Valence		Arousal	
	Acc	F1	Acc	F1	MAE	Corr	MAE	Corr	Acc	F1	Acc	F1
[7]	46.2	/	57.4	/	0.20	0.40	1.17	0.45	/	/	/	/
[8]	**48.6$_2$**	/	**58.2$_3$**	/	**0.19$_3$**	**0.47$_3$**	**0.54$_3$**	**0.52$_2$**	/	/	/	/
[9]	46.6	45.6	57.5	34.6	**0.19$_3$**	0.45	1.08	**0.49$_3$**	76.5	76.4	79.4	79.4
[10]	43.7	**52.5$_1$**	**60.9$_1$**	**37.0$_3$**	0.35	0.31	0.81	0.34	**86.8$_2$**	**85.1$_2$**	84.2	**82.5$_2$**
[11]	/	/	/	/	**0.10$_1$**	/	**0.16$_1$**	/	/	/	/	/
[12]	45.0	45.2	56.8	**37.9$_2$**	/	/	/	/	/	/	/	/
MLG-S	**49.2$_1$**	**48.9$_2$**	**59.6$_2$**	**38.7$_1$**	**0.16$_2$**	**0.49$_1$**	**0.52$_2$**	**0.59$_1$**	**88.2$_1$**	**87.3$_1$**	**86.1$_1$**	**83.4$_1$**
MLG-NCS-WM	46.3	43.4	55.1	33.9	0.25	0.40	0.73	0.37	83.4	82.3	82.5	79.9
MLG-NCS	**48.2$_3$**	**45.7$_3$**	57.9	36.2	**0.19$_3$**	**0.46$_2$**	0.67	0.48	**85.1$_3$**	**84.8$_3$**	83.9	**81.7$_3$**

Note: MLG-S denotes the proposed multimodal local-global model trained in PyTorch platform; MLG-NCS denotes the proposed MLG-NSC with complete hybrid training method; MLG-NCS-WM denotes the proposed MLG-NCS without weights correction during the training process; the subscript 1, 2, 3 represent the corresponding ranking results.

The proposed MLG-NCS is compared with the state-of-the-art (SOTA) methods on the mediaeval 2015 dataset, the mediaeval 2016 dataset, the DEAP dataset, and the CMU-MOSI dataset, as shown in Table II and Table III.

TABLE III
COMPARISON OF DIFFERENT SOTA METHODS ON CMU-MOSI DATASET

Method	Acc-2	F1	Acc-7	MAE	Corr
[13]	82.5	82.3	51.8	**0.58$_3$**	0.70
[14]	84.3	84.3	45.0	0.78	**0.77$_3$**
[15]	80.7	79.8	47.9	0.64	0.66
[16]	84.4	84.6	48.1	0.79	**0.79$_1$**
[17]	**85.6$_2$**	**85.5$_2$**	45.5	0.77	**0.77$_3$**
[18]	**85.5$_3$**	**85.5$_2$**	46.2	0.75	**0.77$_3$**
[19]	84.7	84.7	51.9	**0.58$_3$**	0.70
[20]	85.3	85.2	**52.3$_2$**	**0.56$_2$**	0.74
MLG-S	**86.6$_1$**	**86.4$_1$**	**53.5$_1$**	**0.51$_1$**	**0.78$_2$**
MLG-NCS-WM	82.1	81.9	50.7	0.75	0.65
MLG-NCS	84.8	**84.8$_3$**	**52.1$_3$**	0.63	0.70

From Table II, the proposed MLG-S achieves the highest classification accuracy of valence and the highest F1 of arousal on mediaeval 2015 dataset. Meanwhile, the F1 of valence and accuracy of arousal win the second place over SOTA methods on mediaeval 2015 dataset. For regression task, the proposed MLG-S achieves the highest Corr and ranks top two MAE on mediaeval 2016 dataset. For the DEAP dataset, the MLG-S slightly outperforms SOTA methods. It is noted that inevitable weights mapping errors exist in the proposed hardware system that performs weights mapping operation (i.e., MLG-NCS-WM). The proposed hardware system that performs weights correction operation (i.e., MLG-NCS) achieves high classification and regression performance on all benchmark datasets, closed to the MLG-S. For the above-mentioned benchmark datasets, the proposed MLG-NCS achieves the improvements over other competitors [7-9, 12]. Although [10, 11] is slightly superior to the proposed MLG-NCS, while inferior to running time. Similar conclusion can be observed from Table III. For CMU-MOSI dataset, the proposed MLG-S outperforms other competitors in terms of accuracy, F1, and MAE. The proposed MLG-NCS also achieves the improvements on F1 and Acc-7 over other competitors [13-20].

Furthermore, to study the effectiveness and necessity of each module in the proposed MLG-NCS, the ablation experiments on CMU-MOSI dataset is carried out, as shown in Table IV.

Two common metrics Acc-2 and Acc-7 are used to evaluate the influence of each module. The experiments results are concluded below: 1) Compared with the Convolution unit and Bi-GRU unit, the removal of self-attention unit in local feature

TABLE IV
ABLATION EXPERIMENTS ON CMU-MOSI DATASET

Local feature representation			Global cross-modal interaction	Output	Metrics	
Convolution	Bi-GRU	Self-attention		Global representation	Acc-2	Acc-7
×	✓	✓	✓	✓	84.2	51.3
✓	×	✓	✓	✓	84.0	50.7
✓	✓	×	✓	✓	83.5	50.1
✓	✓	✓	×	✓	72.8	44.6
✓	✓	✓	✓	×	79.5	47.4
✓	✓	✓	✓	✓	84.8	52.1

TABLE V
COMPARISON OF DIFFERENT MODALITY INTERACTION S FOR AFFECTIVE VIDEO CONTENT ANALYSIS

Method	Mediaeval 2015				Mediaeval 2016				DEAP			
	Valence		Arousal		Valence		Arousal		Valence		Arousal	
	Acc	F1	Acc	F1	MAE	Corr	MAE	Corr	Acc	F1	Acc	F1
A	44.4	32.6	56.7	32.7	0.32	0.29	1.00	0.40	72.8	72.1	82.3	80.1
V	46.4	40.9	56.2	30.2	0.26	0.38	1.02	0.29	66.3	65.9	74.6	72.6
T	41.2	29.4	55.9	27.8	0.29	0.13	1.11	0.22	69.5	68.4	77.7	75.4
A+V	47.5	43.3	57.1	34.8	0.25	0.40	0.93	0.48	76.0	75.2	79.2	77.5
A+T	47.7	44.3	57.3	35.0	0.26	0.39	0.87	0.49	77.6	77.8	82.3	80.6
V+T	47.7	45.2	57.4	35.8	0.24	0.40	0.88	0.48	77.5	77.2	81.7	80.2
A+V+T	48.2	45.7	57.9	36.2	0.22	0.41	0.79	0.48	84.1	83.8	83.9	81.7

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: DOI10.1109/TSMC.2024.3392732

representation module has a greater impact on the overall classification performance. The self-attention unit is necessary for local feature representation learning, which plays an important role for sequence modelling. 2) There is a significant drop in classification accuracy, when the global cross-modal interaction module is removed. The experiment result can be demonstrated that the cross-modal learning is essential to affective video content analysis. 3) The proposed MLG-NCS achieves the best classification performance, indicating that the global representation unit can capture the key information in the multimodal information better compared with the removal of this unit.

To explore the effect among different modalities, we use different modalities combinations on the benchmark affective video datasets (i.e., the LIRIS-ACCEDE dataset, the mediaeval 2015 dataset, the mediaeval 2016 dataset, and the DEAP dataset) and CMU-MOSI dataset, as shown in Table V and Table VI.

TABLE VI
COMPARISON OF DIFFERENT MODALITY INTERACTION ON CMU-MOSI DATASET

Modality	Acc-2	F1	Acc-7	MAE	Corr
A	68.0	68.0	30.8	0.79	0.48
V	63.2	60.6	17.7	1.00	0.43
T	73.6	73.6	36.1	0.72	0.59
A+V	73.9	74.2	36.1	0.74	0.59
A+T	77.9	77.1	44.3	0.67	0.66
V+T	79.9	75.9	41.7	0.69	0.65
A+V+T	84.8	84.8	52.1	0.63	0.70

From Table V, the experimental results demonstrate that the visual modality achieves significantly better results than other two modalities in affective video content analysis task. When the text or audio modalities are used together with the visual modality, the system performance slightly outperforming using a single visual modality. From Table VI, compared with visual and audio modalities, the text modality achieves best performance in sentiment analysis task. When the visual or audio modalities are used together with the text modality, the system performance is slightly better than using a single text modality. When all modalities are used together, the proposed MLG-NCS achieves optimal performance in affective video content analysis and sentiment analysis tasks. Based on this, we can draw the following conclusions: Firstly, the importance of different modalities varies, depending on specific tasks. Secondly, the auxiliary modalities can provide additional information to the primary modality.

D. Computational Efficiency Analysis

The computational efficiency analysis is carried out, which concludes the time consumption analysis, area breakdown analysis, latency breakdown analysis, and energy breakdown analysis. To explore the time complexity of the proposed MLG-NCS and make a comparison with baselines, the time consumption of forward propagation is analyzed by comparing with other competitors on the four benchmark datasets, as shown in Fig. 18.

From Fig. 18, the experiments results demonstrate that the proposed MLG-NCS has a significant advantage in terms of time consumption and is approximately 10 times faster than the MLG-S and other competitors but not in terms of space usage. The space complexity of proposed MLG-NCS is $O(L^M)$ where L is the sequential length and M is the number of modalities. The reasons maybe that the high-density memristor crossbar arrays are used in the proposed MLG-NCS, which can realize the dense connectivity between modules and perform parallel MAC operations.

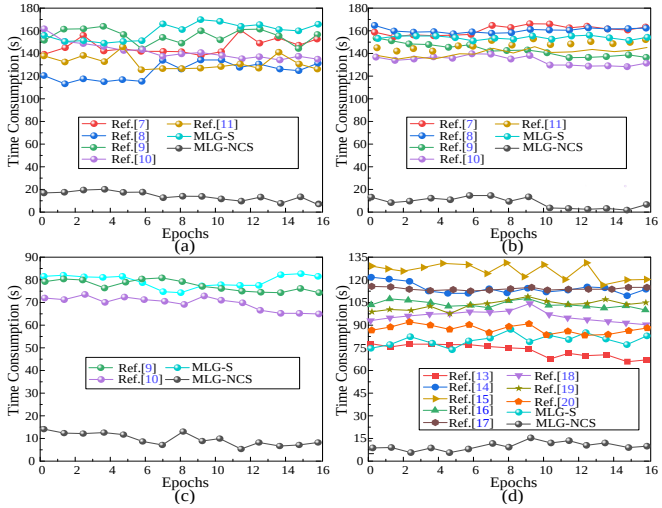


Fig. 18. Time consumption of forward propagation (a) Mediaeval 2015 dataset; (b) Mediaeval 2016 dataset; (c) DEAP dataset; (d) CMU-MOSI dataset.

Then, we measured the latency and power consumption of the proposed multimodal local-global neuromorphic computing system (MLG-NCS) in NeuroSim V3.0 framework [34]. Fig. 19(a) ~ (c) illustrates the area breakdown, latency breakdown, and energy breakdown of the proposed MLG-NCS, respectively.

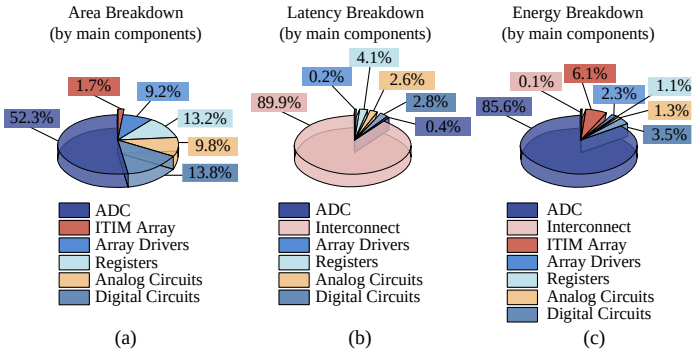


Fig. 19. The breakdown of the proposed MLG-NCS by main component (a) Area; (b) Latency; (c) Energy.

The total area of the proposed MLG-NCS implemented using 28-nm COMS technology is about 185.77μm^2. It is noted that most of the area is occupied by analog to digital converter (ADC) while the 1T1M array occupies 1.7% of the total area. The latency was recorded by capturing the duration of the flag signal on NeuroSim V3.0 framework for inference stage. The total latency of the proposed MLG-NCS is about 1.10μs, and most of the latency is accounted on the interconnect. The total energy consumption for inference mainly depends on the input/output size, the weight precision, the crossbar array size, and the design scheme. From the above measurements and calculations, we obtained that the energy consumption of the proposed MLG-NCS for 1-bit computing with 0.8V, 50ns read voltage is estimated about 3167pJ. Most of the energy consumption is costed by the ADC rather than the 1T1M array.

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: DOI10.1109/TSMC.2024.3392732

TABLE VII
COMPARISON OF DIFFERENT NEUROMORPHIC COMPUTING SYSTEMS

	[21]	[22]	[23]	[24]	[25]	**This work**
Technology	28nm	28nm	130nm	130nm	28nm	28nm
Total area	14.44mm^2	81.83mm^2	0.0704mm^2	159.00mm^2	0.933μm^2	185.77μm^2
Power	0.95W	1.42W	7.44mW	3.00W	224.7μW	62.34mW
Latency	16.8μs	1.30μs	6.69ns	1.40μs	/	1.10μs
Sensing modal	Single	Single	Single	Single	Single	Multimodal
Device stability	High	High	High	High	High	High
Device size	/	/	0.025μm^2	1.69μm^2	/	0.025μm^2
Supply voltage	0.82~0.95V	0.82~0.95V	1.2~1.8V	1.8V	1.0~1.8V	1.0~2.0V
HRS/LRS	/	/	~10^3	/	~10	10^2
Applications	Object detection	Pattern recognition	Pattern recognition	Pattern recognition	Pattern recognition	Affective computing

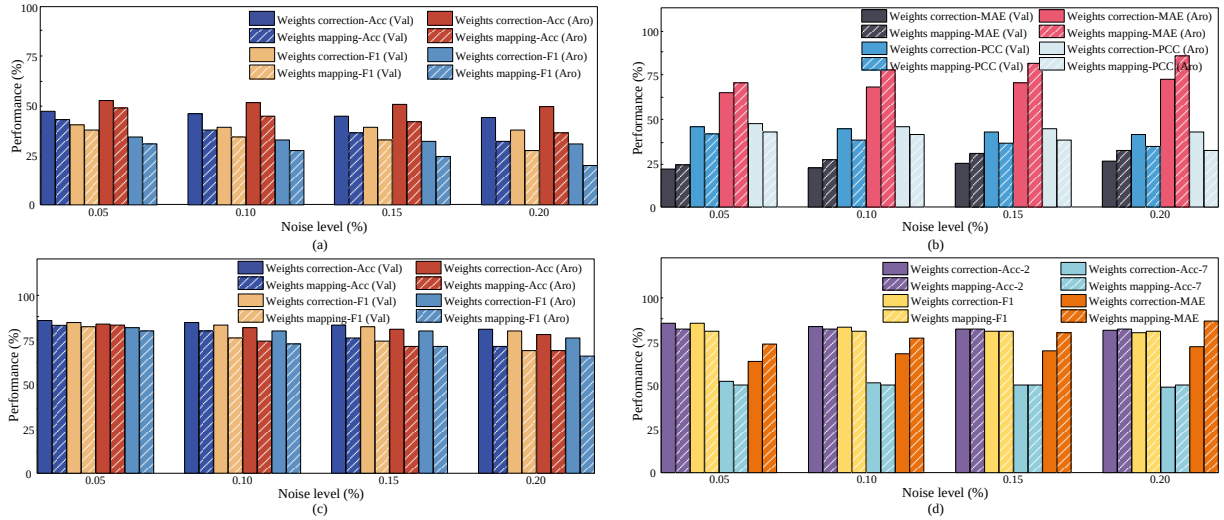


Fig. 20. The noise-resilient analysis of the proposed system (a) Mediaeval 2015 dataset; (b) Mediaeval 2016 dataset; (c) DEAP dataset; (d) CMU-MOSI dataset.

Table VII demonstrates the hardware comparison among the SOTA NCSs. Notably, the hardware metrics of the existing NCSs are excerpted from previous papers.

From Table VII, the proposed MLG-NCS achieves the smallest total area, the lowest power consumption, and the lowest latency among the NCSs at 130-nm COMS technology [23, 24]. At the same 28-nm COMS technology, the proposed MLG-NCS outperforms other competitors [21, 22] in terms of total area, power consumption, and latency, while inferior to [25]. The main reason may be that: 1) At the network structure level: compared to [25] with simple structure (i.e., binary neural network), the proposed MLG-NCS with cascade configuration can capture local characteristics and exchange global cross-modal information sufficiently; 2) At the sensory modality level: the NCS proposed in [25] focuses on single-mode information processing for relatively simple task (i.e., handwritten recognition), and the proposed MLG-NCS has a capability to process multimodal information in the complex fine-grained task. These experiment results show that the proposed MLG-NCS has good performance in latency (about 1.2~15.3 times faster), which balances computational efficiency and computing accuracy to promote versatility.

E. Noise-Resilient Analysis

The noise-resilient analysis is carried out in this section to evaluate the robustness of the proposed MLG-NCS. We inject noise into the proposed MLG-NCS during the weight mapping stage and weight correction stage. We test the proposed MLG-NCS with different levels of noise injection from 0% to 20% on the four benchmark datasets (i.e., the mediaeval 2015 dataset, the mediaeval 2016 dataset, the DEAP dataset, and the CMU-MOSI dataset). The performance metrics (i.e., F1, Acc, MAE, and Corr are used to evaluate the noise influence on the proposed MLG-NCS, as shown in Fig. 20.

The noise experiments show that the proposed MLG-NCS after weights correction operation achieves good robustness and mitigates the interference of noise on the four benchmark datasets. However, the system performance after weights mapping operation is affected as the noise level increases. The results demonstrate that the weights correction operation can help to improve the system noise tolerance capability.

VI. CONCLUSION

This paper investigates a novel NCS for affective video content analysis. Specifically, a high-density memristor crossbar array is prepared using highly stable Fe_2O_3-based memristor, which achieves the dense connectivity between modules and performs parallel-in memory operations. Then, the proposed MLG-NCS mainly consisting of local feature representation module, global cross-modal interaction module, and output module is designed. Through the local feature representation module, the unique local characteristics from multimodal information can be adequately captured. Through the global cross-modal interaction module, the global

cross-modal information can exchange sufficiently. Through output module, the key features can be extracted and reliable output can be obtained effectively. Furthermore, the proposed MLG-NCS with hybrid training method is performed on benchmark datasets, and the experimental results demonstrate that the proposed MLG-NCS achieves high classification and regression performance, closed to the software baseline. In addition, the necessary computational efficiency analysis and noise-resilient analysis are carried out, indicating the high computational efficiency and reliability of proposed system.

VII. References

[1] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, H. Sun, and J. R. Wen, "Towards artificial general intelligence via a multimodal foundation model," *Nat. Commun.,* vol. 13, no. 1, pp. 3094, Dec. 2022, doi:10.1038/S41467-022-30761-2.

[2] B. Zhang, J. Zhu, and H. Su, "Toward the third generation artificial intelligence," *Sci. China Inf. Sci.,* vol. 66, no. 2, pp. 121101, Jan. 2023, doi:10.1007/S11432-021-3449-X.

[3] Z. Dong, X. Ji, C. S. Lai, and D. Qi, "Design and implementation of a flexible neuromorphic computing system for affective communication via memristive circuits," *IEEE Commun. Mag.,* vol. 61, no. 1, pp. 74-80, Jan. 2023, doi:10.1109/MCOM.001.2200272.

[4] S. Grossberg, "Toward autonomous adaptive intelligence: Building upon neural models of how brains make minds," *IEEE Trans. Syst. Man, Cybern. Syst.,* vol. 51, no. 1, pp. 51-75, Jan. 2021, doi:10.1109/TSMC.2020.3041476.

[5] X. Sun, Z. Pei, C. Zhang, G. Li, and J. Tao, "Design and analysis of a human–machine interaction system for researching human's dynamic emotion," *IEEE Trans. Syst. Man, Cybern. Syst.,* vol. 51, no. 10, pp. 6111-6121, Dec. 2021, doi:10.1109/TSMC.2019.2958094.

[6] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: emotion models, databases, and recent advances," *Inf. Fusion,* vol. 83-84, pp. 19-52, July 2022, doi:10.1016/J.INFFUS.2022.03.009.

[7] Y. Yi, and H. Wang, "Multi-modal learning for affective content analysis in movies," *Multimed. Tools Appl.,* vol. 78, no. 10, pp. 13331-13350, Jan. 2019, doi:10.1007/s11042-018-5662-9.

[8] Y. Yi, H. Wang, and Q. Li, "Affective video content analysis with adaptive fusion recurrent network," *IEEE Trans. Multimed.,* vol. 22, no. 9, pp. 2454-2466, Nov. 2020, doi:10.1109/TMM.2019.2955300.

[9] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 31, no. 5, pp. 1901-1914, Aug. 2021, doi:10.1109/TCSVT.2020.3014889.

[10] S. Wang, C. Wang, T. Chen, Y. Wang, Y. Shu, and Q. Ji, "Video affective content analysis by exploring domain knowledge," *IEEE Trans. Affect. Comput.,* vol. 12, no. 4, pp. 1002-1017, Apr. 2021, doi:10.1109/TAFFC.2019.2912377.

[11] S. Zhang, Y. Pan, and J. Z. Wang, "Learning emotion representations from verbal and nonverbal communication," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 18-22 June, 2023, pp. 18993-19004.

[12] Y. Zhu, Z. Chen, and F. Wu, "Affective video content analysis via multimodal deep quality embedding network," *IEEE Trans. Affect. Comput.,* vol. 13, no. 3, pp. 1401-1415, June 2022, doi:10.1109/TAFFC.2020.3004114.

[13] Y. H. Tsai, S. Bai, P. Pu Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *Proc. Conf. Assoc. Comput. Linguist Meet.,* vol. 2019, pp. 6558-6569, July 2019, doi:10.18653/v1/p19-1656.

[14] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, " in *Proceedings of the 2021 International Conference on Multimodal Interaction*, Montreal, 18 Oct., 2021, pp. 6-15.

[15] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," *Inf. Fusion,* vol. 65, pp. 58-71, Jan. 2021, doi:10.1016/j.inffus.2020.08.006.

[16] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, and X. Huang, "TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis," *Knowledge-Based Syst.,* vol. 269, pp. 110502, June 2023, doi:10.1016/j.knosys.2023.110502.

[17] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation," in *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa, 10 Oct., 2022, pp. 3722-3729.

[18] H. Sun, Y. W. Chen, and L. Lin, "TensorFormer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection," *IEEE Trans. Affect. Comput.* Dec. 2022, doi:10.1109/TAFFC.2022.3233070.

[19] S. Zhang, C. Yin, and Z. Yin, "Multimodal sentiment recognition with multi-task learning," *IEEE Trans. Emerg. Top. Comput. Intell.,* vol. 7, no. 1, pp. 200-209, Dec. 2023, doi:10.1109/TETCI.2022.3224929.

[20] H. Deng, Z. Yang, T. Hao, Q. Li, and W. Liu, "Multimodal affective computing with dense fusion transformer for inter- and intra-modality interactions," *IEEE Trans. Multimed.* Sep. 2022, doi:10.1109/TMM.2022.3211197.

[21] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie, and L. Shi, "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature,* vol. 572, no. 7767, pp. 106-111, July 2019, doi:10.1038/s41586-019-1424-8.

[22] S. Ma, J. Pei, W. Zhang, G. Wang, D. Feng, F. Yu, C. Song, H. Qu, C. Ma, M. Lu, F. Liu, W. Zhou, Y. Wu, Y. Lin, H. Li, T. Wang, J. Song, X. Liu, G. Li, R. Zhao, and L. Shi, "Neuromorphic computing chip with spatiotemporal elasticity for multi-intelligent-tasking robots," *Sci. Robot.,* vol. 7, no. 67, pp. abk2948, June 2022, doi:10.1126/scirobotics.abk2948.

[23] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature,* vol. 577, no. 7792, pp. 641-646, Jan. 2020, doi:10.1038/s41586-020-1942-4.

[24] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H. S. P. Wong, and G. Cauwenberghs, "A compute-in-memory chip based on resistive random-access memory," *Nature,* vol. 608, no. 7923, pp. 504-512, Aug. 2022, doi:10.1038/s41586-022-04992-8.

[25] S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S. W. Kwon, Y. Ju, M. Kim, W. Yi, S. Han, B. Kwon, B. Seo, K. Lee, G. H. Koh, K. Lee, Y. Song, C. Choi, D. Ham, and S. J. Kim, "A crossbar array of magnetoresistive memory devices for in-memory computing," *Nature,* vol. 601, no. 7892, pp. 211-216, Jan. 2022, doi:10.1038/s41586-021-04196-6.

[26] 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.* May 2023, doi:10.1109/TCSVT.2023.3275708.

[27] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, 18-24 June, 2022, pp. 815-825.

[28] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 13-19 June, 2020, pp. 10938-10947.

[29] C. Liao, X. Hu, X. Liu, B. Sun, and G. Zhou, "Self-selective analogue FeOx-based memristor induced by the electron transport in the defect energy level," *Appl. Phys. Lett.,* vol. 121, no. 12, pp. 123505, Sep. 2022, doi:10.1063/5.0102076.

[30] S. Wen, R. Hu, Y. Yang, T. Huang, Z. Zeng, and Y. D. Song, "Memristor-based echo state network with online least mean square," *IEEE Trans. Syst. Man, Cybern. Syst.,* vol. 49, no. 9, pp. 1787-1796, Sep. 2019, doi:10.1109/TSMC.2018.2825021.

[31] P. Dhal and C. Azad, "A lightweight filter-based feature selection approach for multi-label text classification," *J. Ambient Intell. Humaniz. Comput.,* vol. 14, no. 9, pp. 12345-12357, Jul. 2022, doi: 10.1007/s12652-022-04335-5.

[32] P. Dhal and C. Azad, "A multi-objective feature selection method using newton's law based pso with gwo," *Appl. Soft Comput.*, vol. 107, pp. 107394, Apr. 2021, doi: 10.1016/j.asoc.2021.107394.

[33] P. Dhal and C. Azad, "Hybrid momentum accelerated bat algorithm with GWO based optimization approach for spam classification," *Multimed. Tools. Appl.*, pp. 1-41, Sep. 2023, doi: 10.1007/s11042-023-16448-w.

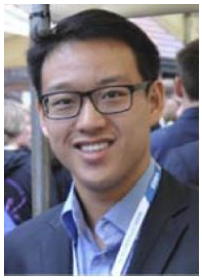
[34] Y. Luo, X. Peng, and S. Yu, "MLP+ NeuroSimV3. 0: Improving on-chip learning performance with device to algorithm optimizations," in *Proceedings of the International Conference on Neuromorphic Systems*, no. 1, pp. 1-7, Jul. 2019, doi: 10.1145/3354265.3354266.

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: DOI10.1109/TSMC.2024.3392732

Xiaoyue Ji (Member, IEEE) received the B.E. degree in electronics and information engineering from the Harbin Engineering University, China, in 2016, the M.S. degree in control science and engineering from the National University of Defense Technology, China, in 2019, and the Ph.D. degree in control science and engineering from the Zhejiang University, China, in 2023. She is currently working toward postdoctoral research at Tsinghua University, China. Her research interests cover memristor and memristive system, artificial neural network, neuromorphic computing.

Zhekang Dong (Senior Member, IEEE) received the B.E. and M.E. degrees in electronics and information engineering in 2012 and 2015, respectively, from Southwest University, Chongqing, China. He received the Ph.D. degree from the School of Electrical Engineering, Zhejiang University, China, in 2019. Currently, he is an associate professor in Hangzhou Dianzi University, Hangzhou, China. He is also a Research Assistant (Joint-Supervision) at The Hong Kong Polytechnic University. His research interests cover memristor and memristive system, artificial neural network, the design and analysis of nonlinear systems based on memristor and computer simulation.

Guangdong Zhou has received his PhD in Faculty of Materials and Energy from Southwest University (P. R. China) in June 2018. He is conducting his postdoctoral research in the Southwest University during 2018.07–2020.07. His research focus on the physical mechanism of memristor, memristor-based functions including the memory logics, displays and synapses. His memristor related researches are supported by the Postdoctoral Program for Innovative Talent Support of Chongqing (600 thousand RMB). During past 5years, more than 50 peer reviewed papers were published. Dr. Zhou sincerely thirsts for the communication and cooperation from broad researcher.



Chun Sing Lai (Senior Member, IEEE) received the BEng in electronic and electrical engineering from Brunel University London, UK, and DPhil in engineering science from the University of Oxford, UK in 2013 and 2019, respectively. Dr Lai is currently a Lecturer at the Department of Electronic and Electrical Engineering, Brunel University London, UK. His current interests are in data analytics, power system optimization, energy system modelling, and energy economics for low carbon energy networks and energy storage systems.

Donglian Qi (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the School of Electrical Engineering, Zhejiang University, China, in 2002. She is currently a Full Professor and a Ph.D. Advisor with Zhejiang University. Her recent research interest covers intelligent information processing, chaos system, and nonlinear theory and application.