IEICE TRANSACTIONS

on Electronics

DOI:10.1587/transele.2023CTP0001

Publicized:2024/04/09

A PUBLICATION OF THE ELECTRONICS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

REM-CiM: Attentional RGB-Event Fusion Multi-modal Analog CiM for Area/Energy-efficient Edge Object Detection during both Day and Night

Yuya Ichikawa[†a)], **Ayumu Yamada**[†], **Naoko Misawa**[†], **Chihiro Matsui**[†], ***Nonmembers*** **and Ken Takeuchi**[†], ***Member***

SUMMARY Integrating RGB and event sensors improves object detection accuracy, especially during the night, due to the high-dynamic range of event camera. However, introducing an event sensor leads to an increase in computational resources, which makes the implementation of RGB-event fusion multi-modal AI to CiM difficult. To tackle this issue, this paper proposes RGB-Event fusion Multi-modal analog Computation-in-Memory (CiM), called REM-CiM, for multi-modal edge object detection AI. In REM-CiM, two proposals about multi-modal AI algorithms and circuit implementation are co-designed. First, Memory capacity-Efficient Attentional Feature Pyramid Network (MEA-FPN), the model architecture for RGB-event fusion analog CiM, is proposed for parameter-efficient RGB-event fusion. Convolution-less bi-directional calibration (C-BDC) in MEA-FPN. extract important features of each modality with attention modules, while reducing the number of weight parameters by removing large convolutional operations from conventional BDC. Proposed MEA-FPN w/ C-BDC achieves a 76% reduction of parameters while maintaining mean Average Precision (mAP) degradation to <2.3% during both day and night, compared with Attentional FPN fusion (A-FPN), a conventional BDC-adopted FPN fusion. Second, the low-bit quantization with clipping (LQC) is proposed to reduce area/energy. Proposed REM-CiM with MEA-FPN and LQC achieves almost the same memory cells, 21% less ADC area, 24% less ADC energy and 0.7% higher mAP than conventional FPN fusion CiM without LQC.

key words: *Computation-in-Memory, RGB-event fusion, edge object detection, multimodal AI, event-based vision sensor*

1. Introduction

In many kinds of edge applications such as autonomous driving and robot vision, object detection is an important task. Although object detection has evolved dramatically in recent years with the advancement of deep learning [1], there are still several challenges in edge object detection [2]. Area/energy limitation is one of the challenges [3]. To tackle this issue, Computation-in-Memory (CiM) [4, 5] is a promising Neural Network (NN) accelerator due to high-speed and low-power multiply-accumulate (MAC) calculation with analog approximate computation in the memory array.

Another challenge in edge object detection is the adaptability to various environments such as day and night. For example, when only RGB cameras are used in night conditions, the accuracy degrades severely [6, 7]. Event cameras [8-10] can cope with the night conditions by fusing with RGB cameras owing to their high dynamic range (e.g., 140dB compared to 60dB of RGB camera) [11-15].

Both model architecture and fusion methods that exploit the complementary characteristics of RGB and event data affect mean Average Precision (mAP), a metric of object detection accuracy. As for the model architecture, feature-pyramid network fusion (FPN fusion) [11] shows the mAP improvement by fusing RGB and event data at each layer of the feature extraction module. As for the fusion method, bi-directional calibration (BDC) [14] shows the mAP improvement by extracting important information with an attention mechanism and convolutional calculations. However, BDC-adopted FPN fusion, called Attentional FPN fusion (A-FPN) in this paper, results in a significant increase in parameters of the NN model.

Recently, 256Gb chalcogenide-based cross-point memory using Phase-Change Memory (PCM) has been proposed [16]. The capacity of standalone emerging non-volatile memory (eNVM) is increasing rapidly. On the other hand, the capacity of embedded eNVM is smaller (1-100 MB) than that of standalone eNVM (1-100GB) because of the large

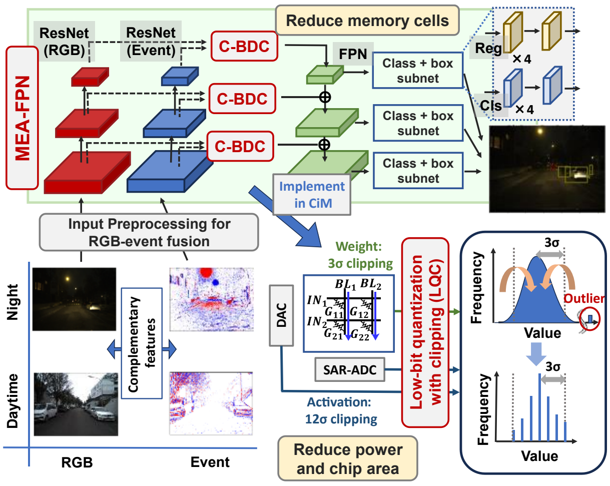

Fig. 1 Overview of proposed REM-CiM. To fuse RGB and event data at edge with area/energy limitation, model architecture (MEA-FPN) and RGB-event fusion module (C-BDC) are proposed. Moreover, appropriate clipping range of weights and activations are investigated for low-bit quantization.

[†] The authors are with Dept. of Electrical Engineering and Information Systems, The University of Tokyo, 113-8656, Japan.
a) E-mail: ichikawa@co-design.t.u-tokyo.ac.jp

overhead of peripheral circuits such as DAC/ADC [17]. The capacities of recently reported CiMs are around 1-10 Mb [4, 18-20]. Although embedded NVM CiM capacity will increase as the technology of NVM integration becomes more mature, the implementation of the RGB-event fusion multi-modal AI on NVM CiM, which requires large memory capacity, is a big challenge.

In this paper, REM-CiM, RGB-Event fusion Multi-modal analog CiM, is proposed to overcome the trade-off between mAP and the number of parameters and to realize multi-modal edge AI (Fig. 1). In REM-CiM, the proposed multi-modal AI algorithms and circuit implementation are co-designed for area/energy-efficient object detection during day and night. The key contributions are:

- An RGB-event fusion model architecture, Memory capacity-efficient Attentional Feature Pyramid Network fusion (MEA-FPN), is proposed to realize the multi-modal AI on edge CiM. Convolution-less bi-directional calibration (C-BDC) in MEA-FPN achieves a 97% reduction in the number of weight parameters compared with BDC by removing large convolution operations, which leads to memory capacity reduction.
- The optimal point of the trade-off between mAP and the number of parameters when considering edge analog CiM implementation is explored by comparing mAPs during day/night and the number of weight parameters among models. The proposed MEA-FPN achieves a 76% reduction of parameters compared with A-FPN while keeping mAP degradation to <2.3%.
- To pursue area/energy/memory-capacity efficiency of analog CiM and realize muti-modal AI on CiM, low-bit quantization with clipping (LQC) is proposed. REM-CiM with MEA-FPN and LQC achieves <150M memory capacity, which indicates the possibility of multi-modal edge CiM.

2. Background and Motivation

2.1 RGB-event fusion method

To effectively utilize the complementary characteristics of RGB and events, it is very important how to fuse them. In [15], late fusion has been proposed. [11] has proposed fusing modalities at multiple stages by utilizing feature pyramid network (FPN) architecture. In these methods, fusion is performed with simple concatenation.

In [12] and [14], the attention module is utilized at the fusion stage to focus on important features and suppress unnecessary ones. In [14], bi-directional calibration (BDC) has been proposed. In BDC, features from one modality are applied to the other. Moreover, important information is extracted in both spatial and channel dimensions with Channel Attention and Spatial Attention [21]. Although BDC achieves mAP improvement, parameter overheads of BDC are so large that it is difficult to implement a BDC-adopted NN model to CiM with energy, area, and memory capacity limitations. In this paper, Convolution-less BDC (C-BDC) is proposed to tackle this challenge.

2.2 Limitation and non-ideality of NVM CiM

Computation-in-Memory (CiM) [4, 5] is a promising NN accelerator. By utilizing Ohm's and Kirchhof's law in weight-embedded memory array, CiM can perform high-speed and low-power analog MAC calculations. By using analog CiM with non-volatile memory (NVM) such as ReRAM, MRAM, PRAM, and Flash, footprint and power consumption are reduced [22, 23]. However, NVM analog CiM has two major issues.

The first issue is a trade-off between accuracy and area/energy due to the bit-resolution of weight memory cells and ADC/DAC. It is reported that ADC/DAC takes a significant portion of the total area/energy consumption [24, 25] and area/energy increases proportionally to ADC/DAC, i.e., activation, bit-resolution [26]. As for weights, the increase in weight bit-resolution leads to the increase in memory area and power consumption of writing operation. To tackle these issues, low-bit quantization is desired.

As for weights, minimizing loss during quantization from floating point to int8 and dequantization from int8 to floating point has been proposed [27]. However, this method does not assume the characteristics of CiM weights (e.g. symmetrical weight distribution with differential pair). As for activations, automatic optimization of the clipping range by considering the clipping range as a parameter has been proposed [28]. However, this quantization method does not consider the quantization of output, which is inevitable when considering ADC in CiM. As shown in these examples, GPU and CiM have fundamentally different weights and quantization methods, so it is necessary to consider clipping and quantization methods suitable for CiM. In this paper, appropriate clipping ranges for weights and activations (i.e. the value of inputs and outputs) of the proposed MEA-FPN are investigated and low-bit quantization with clipping (LQC) is conducted for realizing multi-modal edge CiM.

The second issue is the non-idealities of NVM, such as write variation [18], conductance shift by data-retention [29, 30], and endurance [31]. Due to approximate analog computation, these errors inevitably affect the accuracy. In this paper, the tolerance against these errors is verified.

3. Proposed MEA-FPN Architecture w/ C-BDC

To reduce the number of weight parameters and realize multi-modal AI on edge CiM, an RGB-event fusion model architecture, Memory capacity-efficient Attentional FPN fusion (MEA-FPN), is proposed (Fig. 1). Fig. 2 shows the diagrams of simple concatenation in FPN fusion [11], conventional BDC [14] and the proposed Convolution-less bi-directional calibration (C-BDC). In the proposed MEA-FPN, C-BDC is adopted as an RGB-event fusion module.

Channel Attention (CA) [21] in cross CA [14] makes a channel attention map with average/max-pooling in spatial dimension and Multi-layer Perceptron. Spatial Attention (SA) in cross SA makes a spatial attention map with average/max-pooling in channel dimension and wide (e.g.,

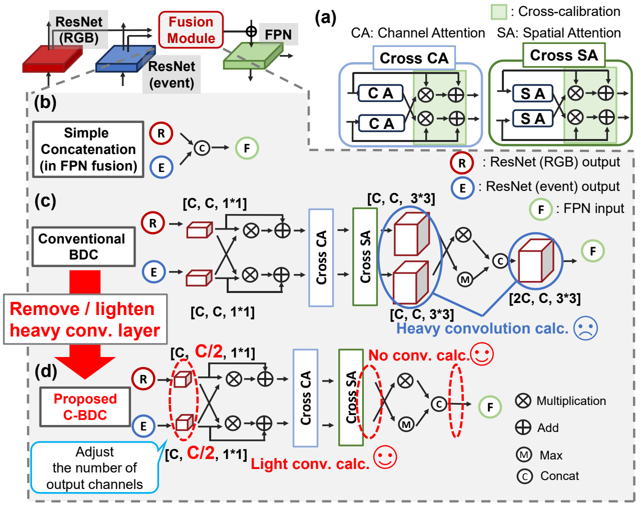

Fig. 2 Diagram of (a) cross Spatial Attention (SA) and cross Channel Attention (CA), (b) simple concatenation in FPN fusion, (c) conventional Bi-directional calibration (BDC) and (d) proposed Convolution-less BDC (C-BDC). In the proposed C-BDC, heavy convolution calc in (c) is removed.

7×7) convolution. By utilizing CA and SA, BDC extracts important features from each modality. Moreover, with the cross-calibration mechanism, BDC transports features from one modality to another (Figs. 2(a) and 2(c)). However, the number of weight parameters in BDC is too large to adopt in CiM (Table I).

To reduce the memory capacity for realizing CiM implementation, C-BDC, a weight parameter-reduced RGB-event fusion module, is proposed (Fig. 2(d)). In C-BDC, the large convolution calculations in BDC are removed except for the first convolution layer. Note that the first convolution is remained just for adjusting the number of channels of C-BDC output to C as shown in Fig. 2(d). On the other hand, cross attention mechanisms are fully utilized for extracting important information and applying features of one modality to the other. By removing large convolutions, the proposed C-BDC achieves a 97% reduction in the number of weight parameters compared with conventional BDC (Table I). This result shows the memory-capacity efficiency of C-BDC in edge multi-modal CiM.

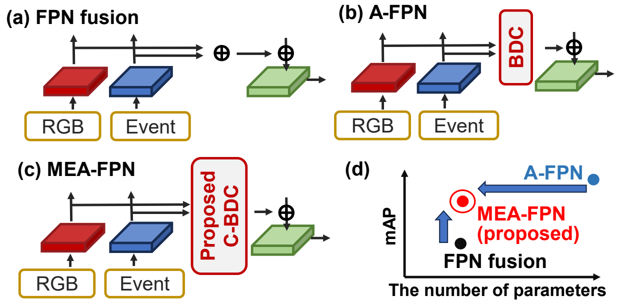

Fig. 3 Diagram of (a) FPN fusion, (b) A-FPN and (c) MEA-FPN. (d) Trade-off between mAP and the number of parameters. MEA-FPN overcomes this trade-off for multi-modal AI on edge CiM.

Table I Comparison of parameters among each module

Module	Params	MACs
ResNet-50	23.5M	25.3G
BDC (Conventional)	210M - 97%	143G
C-BDC (**Proposed**)	5.86M	3.78G

Figs. 3(a)-3(c) show the simple model diagram of FPN fusion, A-FPN, and the proposed MEA-FPN respectively. In A-FPN, the conventional BDC replaces the single RGB-event concatenation in FPN fusion. On the other hand, in the proposed MEA-FPN, C-BDC replaces the concatenation. With middle-fusion architecture, RGB and event features are fused at each stage of the ResNet [32] backbone.

Fused features are fed into the FPN module (Fig. 1). The output of each stage of the FPN module is then fed into the regression and classification subnetwork, and finally the regression boxes and classification results are output.

There is a trade-off between the number of parameters and mAP (Fig. 3(d)). FPN fusion requires fewer parameters, while resulting in relatively low mAP. In contrast, A-FPN achieves higher mAP, while requiring more parameters due to large convolution layers in BDC. A-FPN is suitable when inference is conducted at cloud, with sufficient computational resources. On the other hand, the memory capacity of A-FPN is too large to implement on edge CiM. The proposed MEA-FPN achieves parameter reduction while minimizing the mAP degradation compared with A-FPN, and shows the possibility of implementing multi-modal RGB-fusion AI on edge analog CiM.

4. Evaluation of Proposed MEA-FPN

4.1 Datasets

DSEC dataset is a dataset in driving scenarios including event camera data [33] and contains a lot of night data. However, the object detection label is not publicly released. Therefore, in this work, the dataset provided in [11] is utilized, where over 100,000 objects are labeled using YOLO v5 [34]. In this paper, labels of Car and Pedestrian, which are considered to occur frequently in driving scenarios, are used.

The dataset is split into day and night to measure mAPs during day and night respectively. The number of night data included in the original test dataset is too small (<1500 labels) to measure mAP during the night accurately (Table II). Therefore, a part of the night data in the original training dataset (about 10000 labels) is moved to the test dataset. With this dataset, the impact of the high dynamic range of the event camera on mAP improvement in night conditions

Table II The number of labels in original dataset [11] and this work

		The number of labels	
		Original [11]	This work
Night	Train	32642	22787
	Test	1307	11162
Day	Train	97458	
	Test	22619	

is evaluated.

4.2 Appropriate input preprocessing

Event data are considered as 4D inputs (x,y,p,t), where (x,y) stands for the spatial resolution, p stands for the polarity and t stands for the temporal axis. To fuse with the RGB frame, event representation, where sparse and asynchronous events are converted to dense frames, is necessary. In this work, voxel-grid method [35] is adopted. Voxel-grid retains both temporal and spatial information by dividing each temporal cue into several bins.

In addition, appropriate preprocessing for RGB and event frames is investigated by comparing the 4 points below:

1. Mean & standard deviation (std) of RGB: Whether the mean and std of ImageNet or DSEC is used for the standardization of the RGB frames.
2. RGB normalization: Whether to divide RGB frames by three times the std calculated by all frames (3 sigma) or by the maximum absolute value of each frame (Max).
3. Event standardization: Whether to use the mean and std of all event frames (Global) or those of each event frame (Local) for standardization of event frames.
4. Event clipping: Whether to clip the event frames or not.

Table III shows the comparison results of input preprocessing. The top 2 scores are colored red. With original input processing in [11], loss becomes too large due to the instability of the output of C-BDC, and the model cannot be trained. For RGB standardization (Type 1 vs. Type 2), using the mean and std of DSEC is better than those of ImageNet. Using the mean and std of specific domains (i.e., driving scenario in this work) leads to mAP improvement. The better method of RGB normalization (Type 2 vs Type 3) and event standardization (Type 2 vs Type 4) are 3 sigma and Global respectively. In both methods, all images are divided by the same value, which means that the intensity ratio among images should be preserved. Event clipping leads to score improvement (Type 2 vs Type 5). Clipping reduces the instability of calculation by suppressing outliers. From these results, Type. 2 is utilized for the following experiments as input preprocessing.

4.3 Evaluation setups & metrics

Models are trained to minimize the sum of focal loss, regression loss, and classification loss. ResNet-50 [32] is selected as the backbone. The initial learning rate is set to 0.0001. Adam is selected as an optimizer. In training, the batch number is set to 16 and the epoch number is set to 65. The accuracy of object detection is evaluated by using Average Precision (AP), with setting the threshold of Intersection of Union (IoU) to 50%. In this paper, mAP means the average of the AP of cars and pedestrians. mAPs during day and night are calculated respectively in section 4.4 to investigate the effectiveness of the event camera on mAP in each light condition.

Table III AP comparison between input preprocessing

	RGB		Event		AP: car	AP: Pedes trian
	Standardi zation	Normali zation	Standardi zation	Input clipping		
Original [11]	w/ Mean, std of ImageNet	w/o Norm	-	w/o clip	-	-
Type 1		3 sigma	Global	w/ 3σ clip	0.618	0.271
Type 2	**w/ Mean, std of DSEC**	**3 sigma**	**Global**	**w/ 3σ clip**	**0.651**	**0.299**
Type 3		Max	Global	w/ 3σ clip	0.628	**0.289**
Type 4		3 sigma	Local	w/ 3σ clip	**0.653**	0.288
Type 5		3 sigma	Global	w/o clip	0.627	0.262

4.4 Comparison with conventional models

To investigate the effectiveness of the proposed C-BDC on object detection accuracy under each light condition (day or night) and each label, AP is compared among models (Fig. 5). Moreover, to verify the computational resource and memory-capacity-area efficiency of the proposed MEA-FPN, MACs and the number of weight parameters are also compared among models (Table IV).

To better understand the effectiveness of FPN fusion architecture and multi-modality with RGB-event fusion on mAP improvement, early fusion (Fig. 4(a)) and RGB-only model (Fig. 4(b)) are compared with MEA-FPN. To compare the effectiveness of feature-extraction improvement and fusion-method improvement on the increase in mAP and weight parameters, FPN fusion with ResNet-101, which is deeper than ResNet-50, is also compared with MEA-FPN.

The proposed MEA-FPN achieves 76% parameter reduction compared with A-FPN, while keeping mAP reduction to <2.3% (Table IV). Owing to the much smaller weight parameters of C-BDC compared with BDC, MEA-FPN shows suitability for edge multi-modal CiM. MEA-FPN also achieves more than 3% mAP improvement during both day and night with only 0.3% parameter overhead compared with FPN fusion (Table IV). MEA-FPN also achieves 76% parameter reduction compared with A-FPN, while keeping mAP reduction to <2.3%. The function of C-BDC, i.e. the ability to extract important features and transport one modality feature to the other, plays an important role in mAP improvement during both day and night. Moreover, the proposed C-BDC is suitable for area-limited edge CiM due to the much smaller weight parameters compared with conventional BDC.

The proposed MEA-FPN achieves both better mAP and fewer parameters than FPN fusion with the ResNet-101

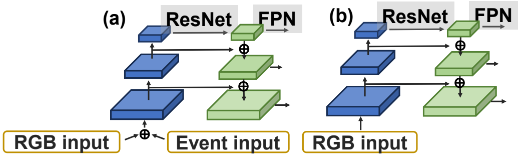

Fig. 4 The model architecture of (a) early fusion and (b) RGB only model.

backbone. Extracting important information from RGB and event features by the proposed C-BDC can achieve better mAP with fewer weight parameters and calculations compared with simply making the feature extraction module deeper. Therefore, C-BDC is necessary for accurate object detection at edge multi-modal CiM, where circuit area and memory capacity are limited.

Early fusion shows better mAP improvement during the night (+3.6%) than daytime (+0.5%) (Table IV). The high dynamic range of the event camera leads to mAP improvement especially during the night. On the other hand, RGB only model achieves higher AP than early fusion in detecting cars during the day (Fig. 5(a)). Even when event data is not effective for detection (e.g., when objects to be detected are hidden by other objects), event features are not suppressed in early fusion, which leads to mAP degradation. Therefore, it is important not only to add the sensors together, but also to extract important features and suppress unimportant ones with the middle-fusion architecture and C-BDC in MEA-FPN for better multi-modal RGB-event fusion during both day and night.

4.5 Effectiveness of cross SA and cross CA

To better understand the impact of the cross-attention modules (Fig. 3 (a)) on multi-modal object detection accuracy, Aps are compared among MEA-FPN, MEA-FPN without cross CA, MEA-FPN without cross SA and FPN fusion (Fig. 6). Both MEA-FPN without cross SA and without cross CA achieve higher AP than FPN fusion, but lower AP than MEA-FPN during both day and night. It is necessary to extract important features in both spatial and channel dimensions with attentional fusion to achieve high AP during both day and night.

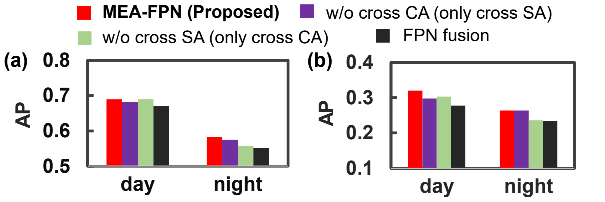

Fig. 6 AP comparison to verify the effectiveness of cross CA and cross SA. (a) AP of car. (b) AP of pedestrian.

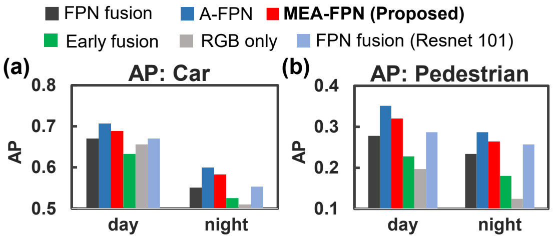

Fig. 5 Comparison about AP of (a) car and (b) pedestrian during day and night among models.

Table IV Comparison of the number of weight parameters, MACs, and mAP during day and night

Feature extraction	Model	Params	MACs	mAP (day)	mAP (night)
ResNet-50	FPN fusion [11]	65.5 M	88.9 G	0.474	0.393
	A-FPN [11], [14]	271M	230G	0.528	0.443
	MEA-FPN (proposed)	-76% 65.7 M	91.2 G +only 0.3%	-2.3% 0.505	+ 3.1% -1.9% 0.424
	Early fusion	36.4 M	62.8 G	0.431	0.353
	RGB only	36.4 M	62.6 G	0.426	0.317
ResNet-101	FPN fusion	104 M	135 G	0.478	0.405

5. Methodology and Evaluation of Proposed LQC

To implement multi-modal AI on edge analog CiM, area, energy and memory capacity are required to be small. However, there is a trade-off between accuracy and area/energy/memory capacity due to the bit-resolution of weights and activations is a major limitation in analog CiM. To achieve the area/energy reduction while maintaining mAP, low-bit quantization with clipping (LQC) is proposed. The novelty of proposed LQC is below: First, the appropriate weight clipping is investigated with consideration of the zero-centered symmetrical characteristics of differential pairs in CiM. Second, to pursue low-bit quantization while maintaining mAP, weight bit-precision sensitivity of each module in MEA-FPN (Fig. 1) is investigated. Third, in contrast to [28], the quantization for both inputs and outputs is conducted to take DAC/ADC into account.

To determine the clipping range in LQC, the appropriate clipping range for weights and activations is investigated respectively (Fig. 1). Note that quantization and clipping of "activation" in this paper means those of both inputs and outputs, as described above. Then, bit-precision sensitivity against weights and activation quantization under the appropriate clipping are compared among models. In addition, to pursue low-bit quantization while maintaining mAP, the weight bit-precision sensitivity of each module in MEA-FPN (Fig. 1) is investigated. From these experiments, the appropriate LQC configuration for the proposed REM-CiM is determined. Moreover, the error-tolerance of proposed and conventional models is investigated by injecting write variation and data retention errors of analog CiM for weights (Fig. 8). Finally, the performance of REM-CiM, with MEA-FPN and LQC, is compared with other CiMs without LQC method.

5.1 Configuration of proposed REM-CiM

Fig. 7 shows the typical weight bit-representation with memory cells, the mapping method of convolution weights, and the representation of weight values in our proposed REM-CiM.

Fig. 7(a) and Fig. 7(b) show bit-parallel weight representation and bit-serial weight representation respectively [36, 37]. In these weight representation methods, weight values are represented by multiple 1-bit memory cells to avoid errors due to the limited signal margins and the device variations of MLC. On the other hand, it is assumed that each memory cell in the proposed CiM can represent the weight values with analog conductance and only one memory cell is used for representing weight value, with reference to [18, 38]. Therefore, the proposed CiM stores weights in its analog conductance without bit-serial or bit-parallel method, shown in Fig. 7(a) or (b).

C_{in}, C_{out}, K represents the size of input channels, the size of output channels, and kernel size respectively. As shown in Fig. 7(c), each C_{in} weights in one kernel are mapped in one column, and the convolution weights at the same place of each kernel are mapped in one array, referring to [39].

Fig. 7(d) shows each weight cell in the CiM array. Each weight of neural network is represented by a differential pair and the value of weight is represented as the difference of analog conductance: $G_{ij}^{+} - G_{ij}^{-}$. Therefore, two memory cells are required to represent one weight value. Note that each memory cell retains analog conductance and positive/negative weight value is represented with a single cell respectively. Fig. 7 (c) shows the cumulative probability of ReRAM cell current [31].

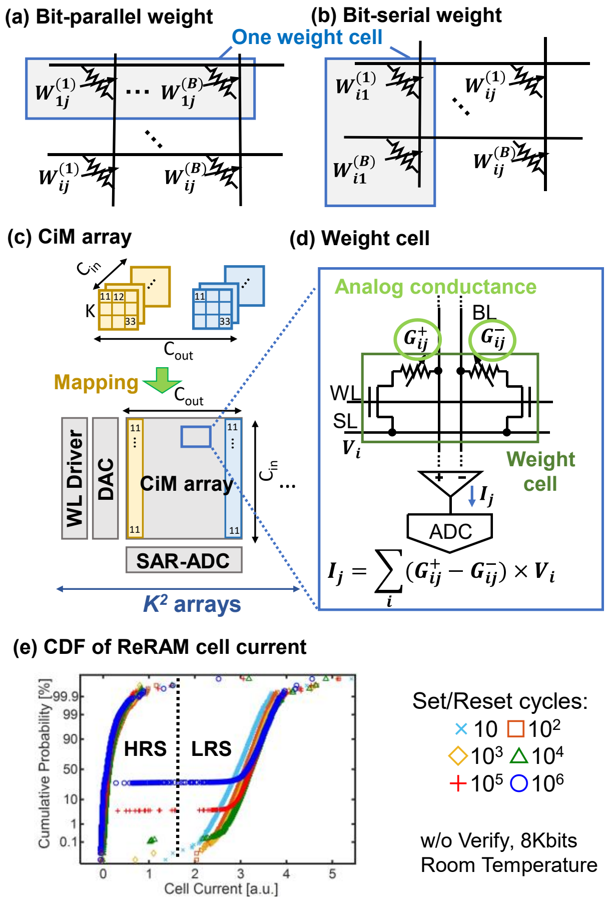

Fig. 7 (a) Bit-parallel weight cell and (b) bit-serial weight cell intypical CiM array. (c) Mapping of convolution layer to CiM. (d) Memory cell assumed in this paper. (e) Cumulative probability of ReRAM cell current [31].

The weight and activation quantization step is determined by peak-to-peak method after clipping (Fig. 8(a)) [38]. As weight is represented by differential pair, the weights are quantized symmetrically around zero.

Write variation is reproduced by adding Gaussian errors with standard deviation σ to weights (Fig. 8(b)). Conductance shift is reproduced by adding a constant value to weights (Fig. 8(c)). Write-verify operation [18] is assumed to performed when writing weights. The baseline of mAP is set to 0.460 in each experiment, which is 1.5% lower than mAP achieved by MEA-FPN with 32-bit precision. The experiment of write variation is conducted with the assumption that the write variation errors injected into weights include the impact of non-linearity. Regarding the tolerance against ADC errors, the report about ADC noise in [38] is referenced. In [38], σ_{ADC} is introduced as the parameter representing ADC noise, and the differential non-linearity (DNL) of each output code follows normal distribution N(1.0 [LSB], σ_{ADC} [LSB]). Considering 4-bit activation in CiM, 0.4 σ_{ADC} of ADC non-linearity is tolerated. This indicates that DNL with an average of 0.5 [LSB] and integrated non-linearity (INL) with an average of 1.0 [LSB] are tolerated respectively. With this consideration, it is assumed that the influence of ADC non-linearity is less than that of weight variation errors.

5.2 Appropriate clipping range of weight & activation

To achieve weight & activation low-bit quantization while maintaining mAP, the appropriate clipping range for the proposed MEA-FPN is investigated (Fig. 9). By setting the clipping range to 3σ, weight bit-precision sensitivity improves from 5-bit to 4-bit (Fig. 9 (a)). On the other hand, the activation clipping range needs to be relatively wide (Fig. 9(b)). With 12σ clipping, activation bit-precision sensitivity improves from 8-bit to 6-bit.In MEA-FPN, the outputs of C-BDC become large (Fig. 10(a)) due to the instability of cross-calibration mechanisms in cross SA and cross CA. In particular, the outliers become much larger than other values, which leads to serious degradation of bit-precision sensitivity. These observations indicate the importance of appropriate activation clipping in MEA-FPN. From these results, 3σ and 12σ are determined as the appropriate clipping range for weights and activation respectively.

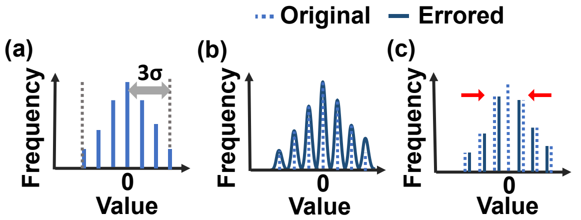

Fig. 8 (a) Histogram of weight values with clipping and quantization. (b) Gaussian error and (c) shift error applied to clipped and quantized weight.

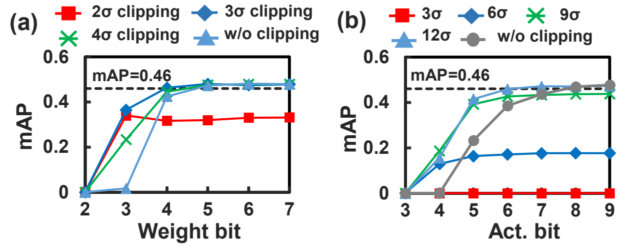

Fig.9 Bit-precision sensitivity of MEA-FPN with various clipping range of (a) weight and (b) activation.

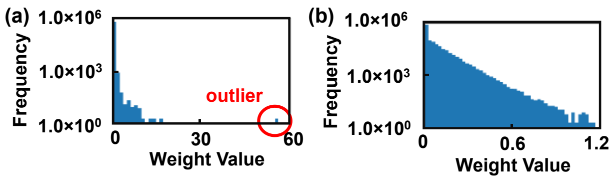

Fig. 10 Input histogram of FPN module in (a) proposed MEA-FPN fusion and (b) conventional FPN fusion.

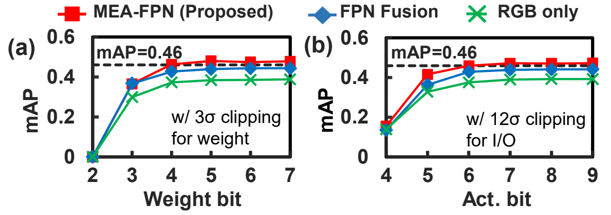

Fig. 11 Comparison of bit-precision sensitivity against (a) weight quantization and (b) activation quantization.

5.3 Appropriate bit-precision of weight & activation for low-bit quantization

To compare the impact of RGB-event fusion and the proposed C-BDC on AP degradation in low-bit quantization, bit-precision sensitivity of weights (Fig. 11(a)) and activation (Fig. 11(b)) is compared among 3 models: MEA-FPN, FPN fusion and RGB only model. Based on the results in section 5.1, 3σ clipping is adopted to weights, and 12σ clipping is adopted to activations. With the appropriate weight & activation clipping, the proposed MEA-FPN can maintain higher mAP with 4-bit and 6-bit quantization for weights and activations respectively.

To reduce CiM memory cells while maintaining mAP, the weight bit-precision sensitivity of each module is also investigated (Fig. 12). The RGB module is a little less tolerant to low-bit quantization compared with others. To avoid wasting the rich RGB information, high bit-resolution is required for RGB-feature extraction. For maintaining mAP when weights of all modules are low-bit quantized, the bit-precision of W_{RGB} is determined as 5-bit in the proposed LQC.

5.4 Comparison of error-tolerance among models

To compare the impact of RGB-event fusion and the proposed C-BDC on the tolerance against write variation and data retention error of analog CiM, the error-tolerance of weights are compared among 3 models (Fig. 13). In this experiment, weights are quantized to 8-bit with 3σ clipping, to ensure that quantization do not affect the mAP degradation and to investigate the mAP degradation driven by NVM errors precisely. The unit of error size "n.s." stands for normalized step, meaning the relative size to weights normalized between -1 and 1. In Fig. 8(a), only the error of write variation is injected to models. In Fig. 8(b), only the error of conductance shift is injected to models. The results show that MEA-FPN tolerates up to 0.03 [n.s.] gaussian error and 0.002 [n.s.] shift error respectively. In other words, to maintain high mAP, gaussian errors should be less than 0.03 [n.s.] and shift error should be less than 0.002 [n.s.].

As reported in [18], the write variation of ReRAM is 0.59 μA and the range of conductance is 30μA when the write-verify operation is performed. From this result, the normalized write variation of each ReRAM cell is supposed to be about 0.02 [n.s.]. [38] shows that the variation of the differential pair when the write-verify operation is performed is supposed to be around 0.03 [n.s.]. With this consideration, it can be said that the proposed MEA-FPN tolerates write variation if the write-verify operation is performed.

Under these errors, MEA-FPN maintains higher mAP than the other models.

5.5 LQC impact on mAP and CiM performance

From the results in this chapter, the appropriate configuration of LQC for MEA-FPN is determined. 3σ and 12σ clipping is adopted for weights and activations respectively. As for weights, W_{RGB} is quantized to 5-bit and others are quantized to 4-bit. As for activations, 6-bit quantization is adopted uniformly.

Table V shows the comparison of each model CiM without LQC and the proposed REM-CiM with MEA-FPN and LQC. The mAPs with write variation are compared considering mapping to analog CiM. The mAPs with write variation & data retention error are also compared considering the case where time has passed since the mapping. Let $ocs, hwif, iob$ and ks represent the output channel size, the product of height and width of input feature, activation bit, and kernel size, respectively, in each matrix operation. It is assumed that every 8 columns are shared in one ADC and weights at different spatial locations of each kernel are mapped to different sub-matrices, referring to [39]. ADC

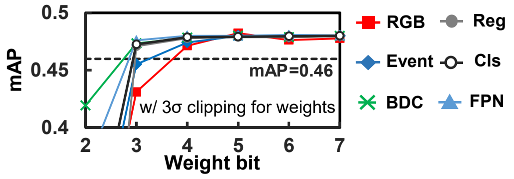

Fig. 12 Weight bit-precision sensitivity of each module in proposed MEA-FPN.

area/energy are assumed to be proportional to the activation bit-precision, referring to [26]. From these assumptions, the relative ADC area/energy is calculated with the following equations:

$$Area \ \propto \sum ocs \cdot iob \cdot ks^2 \qquad (1)$$

$$Energy \ \propto \sum ocs \cdot hwif \cdot iob \cdot ks^2 \qquad (2)$$

The proposed REM-CiM achieves a 25% reduction of ADC area/energy compared with MEA-FPN CiM without LQC. When write variation error is added considering mapping on analog CiM, REM-CiM keeps the mAP reduction to <2.8% compared with MEA-FPN CiM without LQC. A-FPN CiM achieves the best mAP for all error patterns, however, requires the most memory cells (>500Mb) and ADC area/energy. Therefore, it is difficult to implement A-FPN on edge CiM and A-FPN CiM is not suitable for edge usage. On the other hand, REM-CiM achieves the memory capacity around 130Mb, which is compatible with current memory capacity limitation at edge (<100Mb). Considering the rapid evolution of the technology of NVM integration [16], the capacity of embedded eNVM will become larger in the same way as the capacity of standalone NVM. Therefore, it can be said that the implementation of the proposed REM-CiM is feasible even though the capacity of REM-CiM is a little bigger than the current target of 100Mb.

REM-CiM also achieves almost the same memory cells, 19% less ADC area, 24% less ADC energy and 0.7% higher mAP than FPN fusion CiM without LQC, as the arrows in Table V indicate. By co-designing area/energy- efficient algorithm and implementation method of analog CiM, both higher mAP and less area/energy computational resource than conventional method are achieved and implementation of accurate multi-modal AI on edge CiM is realized. The higher mAP of REM-CiM is also maintained when data retention error is also added. This result shows that REM-CiM maintain higher mAP even after time has elapsed.

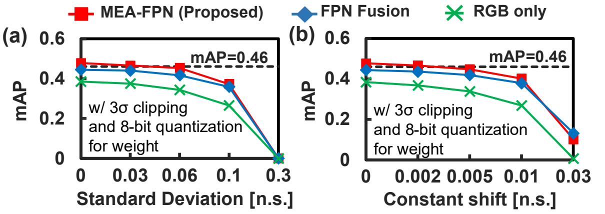

Fig. 13 Comparison of error-tolerance when (a) gaussian or (b) shift errors are injected to each model.

6. Conclusion

In this paper, REM-CiM: RGB-Event fusion Multi-modal CiM is proposed for multi-modal edge object detection during both day and night. In REM-CiM, multi-modal algorithms and circuit implementation are co-designed to realize multi-modal AI on edge analog CiM under the memory capacity limitation. First, memory capacity-reduced RGB-event fusion model architecture, MEA-FPN, is proposed with C-BDC. C-BDC reduces the number of weight parameters by removing large convolution operations, which leads to memory capacity reduction. MEA-FPN achieves a 76% reduction of parameters compared with A-FPN while keeping mAP degradation to <2.3% during both day and night. Second, low-bit quantization with clipping (LQC) is proposed. In LQC, the appropriate clipping range of weight and activation for low-bit quantization is explored. By co-designing algorithms and analog CiM implementation with MEA-FPN and LQC, multi-modal AI on edge CiM is realized. REM-CiM achieves almost the same memory cells, 21% less ADC area, 24% less ADC energy, and 0.7% higher mAP compared with

Table V Comparison between CiMs of each model

		CiM (w/o LQC)				REM-CiM (w/ LQC[1])
Model		RGB only	FPN fusion [11]	A-FPN [11, 14]	MEA-FPN (proposed)	MEA-FPN (proposed)
CiM Configuration	Weight bit-precision	**5-bit** (Fig. 8(a))				W_{RGB} : **5-bit** (Fig. 11) **Other weight: 4-bit** (Figs. 8(a), 10(a))
	Weight clipping	-				**3σ** (Fig. 8(a))
	Activation[2] bit-precision	**8-bit** (Fig. 8(b))				**6-bit** (Figs. 8(b), 10(b))
	Activation clipping	-				**12σ** (Fig. 8(b))
	Tolerable Gauss σ (Write variation)	**<= 0.03 [n.s.]** (Fig. 12(a))				
	Tolerable Const Δ (Data retention error)	**<= 0.002 [n.s.]** (Fig. 12(b))				
mAP	w/o error	0.383	0.437	0.495	0.475	0.454
	w/ Write variation (0.03 [n.s.])	0.375	0.434	0.490	0.469	0.441
	w/ Write variation (0.03 [n.s.]) & Data retention error (0.002 [n.s.])	0.368	0.417	0.479	0.460	0.433
CiM Performance	Required number of memory cells	72.8M	131M	542M	131M	131M
	ADC Area (Normalized)	0.57	0.95	1.65	1	0.75
	ADC Energy (Normalized)	0.57	0.99	1.14	1	0.75

+0.7%

-21%

-24%

[1] LQC configuration is determined by setting mAP baseline to 0.460

[2] In activation quantization and clipping, both inputs and outputs are quantized/clipped.

FPN fusion CiM without LQC.

Acknowledgments

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Z. Zou et al., “Object Detection in 20 Years: A Survey,” Proceedings of the IEEE, vol. 111, no. 3, pp. 257-276, 2023.

[2] Z. Chang et al., “A Survey of Recent Advances in Edge-Computing-Powered Artificial Intelligence of Things,” IEEE IoT Journal, vol. 8, no. 18, pp. 13849-13875, 2021.

[3] L. Liu et al., “Equinox: A Road-Side Edge Computing Experimental Platform for CAVs,” MetroCAD, pp. 41-42, 2020.

[4] M. Chang et al., “A 73.53TOPS/W 14.74TOPS Heterogeneous RRAM In-Memory and SRAM Near-Memory SoC for Hybrid Frame and Event-Based Target Tracking,” ISSCC, pp. 426-428, 2023.

[5] N. Verma et al., “In-Memory Computing: Advances and Prospects,” IEEE SSC-M, vol. 11, no. 3, pp. 43-55, 2019.

[6] J. Lin and F. Zhang, “R3LIVE: A Robust, Real-time, RGB-colored, LiDAR-Inertial-Visual tightly-coupled state Estimation and mapping package,” ICRA, pp. 10672-10678, 2022.

[7] Z. Wu et al., “Robust RGB-D Fusion for Saliency Detection,” 3DV, pp. 403-413, 2022.

[8] G. Gallego et al., “Event-Based Vision: A Survey,” IEEE TPAMI, vol. 44, no. 1, pp. 154-180, 2022.

[9] P. Lichtsteiner *et al.*, "A 128 X 128 120db 30mw asynchronous vision sensor that responds to relative intensity change," in ISSCC, pp. 2060-2069, 2006.

[10] T. Finateu et al., “5.10 A 1280×720 Back-Illuminated Stacked Temporal Contrast Event-Based Vision Sensor with 4.86µm Pixels, 1.066GEPS Readout, Programmable Event-Rate Controller and Compressive Data-Formatting Pipeline,” ISSCC, pp. 112-114, 2020.

[11] A. Tomy et al., “Fusing Event-based and RGB camera for Robust Object Detection in Adverse Conditions,” ICRA, pp. 933-939, 2022.

[12] L. Sun et al., “Event-based fusion for motion deblurring with cross-modal attention,” ECCV, pp. 412-428, 2022.

[13] P.Shi, *et al,* “EVEN: An Event-Based Framework for Monocular Depth Estimation at Adverse Night Conditions”, arXiv preprint arXiv:2302.03680, 2023.

[14] Z. Zhou, et al., “RGB-Event Fusion for Moving Object Detection in Autonomous Driving,” arXiv preprint Arxiv:2209.08323v2.

[15] S. Tulyakov et al., “Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion,” CVPR, pp. 17734-17743, 2022.

[16] F. Pellizzer, A. Piirovano, R. Bez, and R. Mayer, “Status and Perspectives of Chalcogenide-based Cross-Point Memories (Invited),” IEEE International Electron Devices Meeting (IEDM), 2023.

[17] N. Lepri et al., “In-memory computing for machine learning and deep learning,” IEEE Journal of the Electron Devices Society, 2023.

[18] R. Mochida et al., “A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture,” VLSI Tech., pp. 175-176, 2018.

[19] J. Han et al., "ERA-LSTM: An Efficient ReRAM-Based Architecture for Long Short-Term Memory," in IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 6, pp. 1328-1342, 1 2020.

[20] C. -X. Xue et al., "24.1 A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," 2019 IEEE ISSCC, 2019, pp. 388-390.

[21] S. Woo et al., “Cbam: Convolutional block attention module,” ECCV, pp. 3-19, 2018.

[22] L. Song et al., "PipeLayer: A Pipelined ReRAM-Based Accelerator for Deep Learning," in HPCA, 2017, pp. 541-552.

[23] Q. Liu et al., “33.2 A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing,” ISSCC, pp. 500-502, 2020.

[24] P. Chen et al., “RIMAC: An Array-level ADC/DAC-Free ReRAM-Based In-Memory DNN Processor with Analog Cache and Computation,” ASP-DAC, pp. 1-6, 2023.

[25] H. Jiang, et al., “A 40nm Analog-Input ADC-Free Compute-in-Memory RRAM Macro with Pulse-Width Modulation between Sub-arrays,” VLSI Technology and Circuits, pp. 266-267, 2022.

[26] S. Yu et al., “Compute-in-Memory with Emerging Nonvolatile-Memories: Challenges and Prospects,” CICC, pp. 1-4, 2020.

[27] H. Wu, P. Judd, X. Zhang, M. Isaev, and P. Micikevicius, “Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation,”, *arXiv preprint* https://arxiv.org/abs/2004.09602.

[28] J. Choi, S. Venkataramani, V. V. Srinivasan, K. Gopalakrishnan, Z. Wang and P. Chuang, “Accurate and efficient 2-bit quantized neural networks,” *Proceedings of Machine Learning and Systems*, 1, 348-359, 2019.

[29] S. Fukuyama et al., “Comprehensive Analysis of Data-Retention and Endurance Trade-Off of 40nm TaOx-based ReRAM,” IRPS, pp. 1-6, 2019.

[30] Y.-H. Lin et al., “Performance Impacts of Analog ReRAM Non-ideality on Neuromorphic Computing,” IEEE Transactions on Electron Devices, vol. 66, no. 3, pp. 1289-1295, 2019.

[31] K. Taoka, N. Misawa, S. Koshino, C. Matsui and K. Takeuchi, “Simulated Annealing Algorithm & ReRAM Device Co-optimization for Computation-in-Memory,” 2021 IEEE International Memory Workshop (IMW), Dresden, Germany, pp. 1-4, 2021.

[32] K. He et al., “Deep Residual Learning for Image Recognition,” CVPR, pp. 770-778, 2016.

[33] M. Gehrig et al., “DSEC: A Stereo Event Camera Dataset for Driving Scenarios,” IEEE RALs, vol. 6, no. 3, pp. 4947-4954, 2021.

[34] Ultralytics. Yolov5. [Online]. Available: https://github.com/ultralytics/yolov5.

[35] A. Z. Zhu et al., “Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion,” CVPR, pp. 989-997, 2019.

[36] Q. Liu et al., "33.2 A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," 2020 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2020, pp. 500-502

[37] A. Parmar, K. Prasad, N. Rao and J. Mekie, "An Automated Approach to Compare Bit Serial and Bit Parallel In-Memory Computing for DNNs," 2022 IEEE International Symposium on Circuits and Systems (ISCAS), Austin, TX, USA, 2022, pp. 2948-2952

[38] A. Yamada, N. Misawa, C. Matsui and K. Takeuchi, "LIORAT: NN Layer I/O Range Training for Area/Energy-Efficient Low-Bit A/D Conversion System Design in Error-Tolerant Computation-in-Memory," 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD), San Francisco, CA, USA, pp. 1-9, 2023.

[39] X. Peng et al., “Optimizing Weight Mapping and Data Flow for Convolutional Neural Networks on Processing-in-Memory Architectures,” IEEE TCAS-I, vol. 67, no. 4, pp. 1333-1343, 2020

Yuya Ichikawa Received the B.S degree in Information and Communication Engineering from the University of Tokyo in 2022. He is now a master course student in Takeuchi Laboratory in the department of Electrical Engineering and Information Systems, the University of Tokyo. His current research interests include RGB-event fusion multimodal AI and Computation-in-Memory system.

Ayumu Yamada Received the B.S. degree in Electrical Engineering from the University of Tokyo in 2022. He is now a master course student in Takeuchi Laboratory in the department of Electrical Engineering and Information Systems, the University of Tokyo. His current research interests include Computation-in-Memory (CiM) system, emerging non-volatile memories, neuromorphic computing, and Bayesian machine learning.

Naoko Misawa Received the M.S. degree from Imperial College London in 2012. She is currently an academic staff in Takeuchi Laboratory in the department of Electrical Engineering and Information Systems, Graduate School of The University of Tokyo. Her research interests include emerging non-volatile memories, neuromorphic computing, and Vision Transformer.

Chihiro Matsui is currently a Project Associate Professor in the Department of Electronics Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo. Her research interest includes system, circuit, and device co-design with emerging non-volatile memories for enterprise applications. She earned her B.S. and M.S. degrees in Physics from Ochanomizu University, Tokyo, Japan, in 2003 and 2005, respectively, and her Ph.D. degree in Information Security Sciences from Chuo University, Tokyo, Japan, in 2018. She was a Project Assistant Professor of Research and Development Initiative at Chuo University from 2018 to 2020 and a Project Assistant Professor in the Department of Electronics Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo from 2020 to 2023.

Ken Takeuchi is currently a Professor at Department of Electrical Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo. He is now working on data-centric computing such as computation in memory, approximate computing, datacenter scale computing, AI chip design and brain-inspired memory. He received the B.S. and M.S. degrees in Applied Physics and the Ph.D. degree in Electric Engineering from The University of Tokyo in 1991, 1993 and 2006, respectively. In 2003, he also received the M.B.A. degree from Stanford University. Since he joined Toshiba in 1993, he had been leading Toshiba's NAND flash memory circuit design for fourteen years. He was an Associate Professor at Department of Electrical Engineering and Information Systems, Graduate School of Engineering of The University of Tokyo from 2007 till 2012. He was a Professor at Department of Electrical, Electronic and Communication Engineering, Faculty of Science and Engineering of Chuo University from 2012 till 2020. In 2020, he rejoined The University of Tokyo. He designed six world's highest density NAND flash memory products such as 0.7um 16Mbit, 0.4um 64Mbit, 0.25um 256Mbit, 0.16um 1Gbit, 0.13um 2Gbit and 56nm 8Gbit NAND flash memories. He holds 228 patents worldwide including 124 U.S. patents. Especially, with his invention, “multipage cell architecture”, presented at Symposium on VLSI Circuits in 1997, he successfully commercialized world's first multi-level cell NAND flash memory in 2001. He has authored numerous technical papers, one of which won the Takuo Sugano Award for Outstanding Paper at ISSCC 2007. He is serving as the program chair of Asian Solid-State Circuits Conference (A-SSCC) in 2023. He served as the symposium chair/co-chair of Symposium on VLSI Circuits in 2021/2020. He served as the program chair/co-chair of Symposium on VLSI Circuits in 2019/2018. He has also served on the program committee member of International Solid-State Circuits Conference (ISSCC), Custom Integrated Circuits Conference (CICC), Asian Solid-State Circuits Conference (A-SSCC), International Memory Workshop (IMW), International Conference on Solid State Devices and Materials (SSDM) and Non-Volatile Memory Technology Symposium (NVMTS). He served as a tutorial speaker at ISSCC 2008, forum speaker at ISSCC 2015, SSD forum organizer at ISSCC 2009, 3D-LSI forum organizer at ISSCC 2010, Ultra-low voltage LSI forum organizer at ISSCC 2011 and Robust VLSI System forum organizer at ISSCC 2012.