This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3260116

A 40nm 2TOPS/W Depth-Completion Neural Network Accelerator SoC with Efficient Depth Engine for Realtime LiDAR Systems

Miao Sun[1], Yingjie Cao[2], Jian Qian[1], Jie Li[1], Sifan Zhou[3], Ziyu Zhao[3], Yifan Wu[4], Tao Xia[1]
Yajie Qin[1], Lei Qiu[4], Shunli Ma[1], Patrick Yin Chiang[1], Shenglong Zhuo[1]
[1]State Key Laboratory of ASIC & System, Fudan University, Shanghai, China, Email: 18112020006@fudan.edu.cn
[2]TiMESiNTELLi Technologies, Shanghai, China
[3]School of Automation, Southeast University, Nanjing, China
[4]College of Electrical Information and Engineering, Tongji University, China

***Abstract*—Light Detection and Ranging (LiDAR) is becoming a critical requirement for future computer vision applications, such as AR/VR (iPhone-LiDAR) and ADAS (Automotive-LiDAR). A depth point-cloud input has different characteristics than a conventional RGB image input, such that the CNN depth-inference implementation is unique when compared with a standard super-resolution CNN(SR-CNN). In this brief, we present a heterogeneous AI-accelerator SoC, which is specific to depth image completion computation. Three key innovations are introduced to improve SoC's performance. First, to accommodate the unique input data structure of a depth input, a fully-filled dataflow management engine is proposed to pre-process the RGB+Depth input, significantly improving processing element utilization (PEU). Second, to improve the efficiency of the instruction configurations of the CNN accelerator, a hardware-tiling co-processor is proposed that performs the tiling strategy of the CNN accelerator, assigning each sub-job to the PE array directly, therefore reducing the time for task assignments. Third, due to the large number of vector operations required for the post-process in the neural network, a RISC-V core is incorporated to execute vector computations better. The SoC is implemented in 40nm CMOS process, achieving 2TOPs/W energy efficiency with 34fps throughput under VGA-resolution output for real-time LiDAR systems.**

***Index Terms*—depth completion, depth engine, RISC-V Extended Vector, DSA, On-Chip co-processor scheduler.**

I. Introduction

THREE dimensional point-cloud data has proven its value in various applications such as autonomous vehicles (AV) and Simultaneous Localization and Mapping (SLAM). Recent LiDAR systems based on direct time-of-flight (dToF) and single photon avalanche diodes (SPADs) technology show a detection range of hundreds of meters and ranging accuracy down to one centimeter. With this new type of sensor, a new data structure called RGB-D is created by the fusion of RGB data generated by the traditional image sensor and point cloud generated by LiDAR. Due to the cost of capturing depth information being higher than RGB images, the minimization design of SPAD array is a popular trend [1]. In [2], a 256x8 SPAD imager is fabricated to detect 30m range with 28fps. In [3], a 128x128 SPAD imager is proposed by implementing

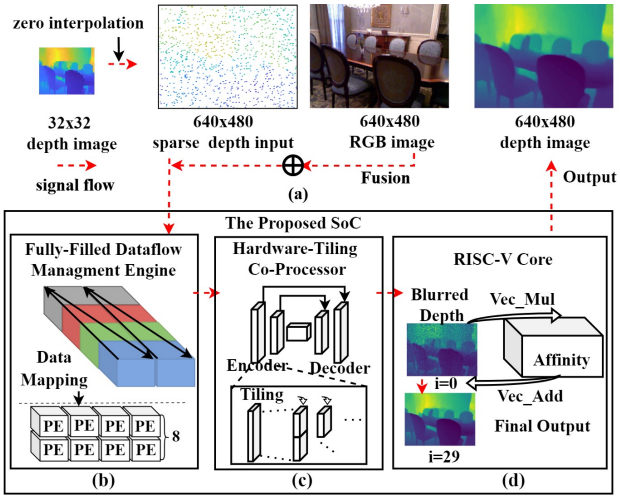

Fig. 1. Detailed computation flow of the proposed depth completion. (a) The input and output of depth completion. (b) The proposed fully-filled dataflow management engine for pre-processing. (c) The proposed hardware-tiling co-processor for the neural network backbone. (d) The designed RISC-V Core for post-process.

128 dynamic time-to-digital converter (TDC) to collect data. In [4], a 168x63 SPAD sensor with passive quenching and recharge front-end circuitry, which is capable of refining strong background light. However, the spatial resolution of state-of-the-art dToF SPAD sensors is limited to hundreds of thousands of pixels since it is difficult to reduce the size and fill factor of SPAD pixels. As a result, it makes LiDAR hard to detect objects such as cars and pedestrians at hundreds of meters. One standard solution is to use an AI network to upscale and complete such sparse images, such as a SR-CNN accelerator.

Another problem is that the sparse characteristic of the point cloud brings obstacles to alignment between RGB imagers and LiDAR sensors. This misalignment of RGB and point-cloud data results in obscure edges or even errors in object detection. Therefore, depth completion is vital for SPAD LiDAR systems to guarantee the consistency of pixel resolution between RGB images and point-cloud. Some research has been conducted

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3260116

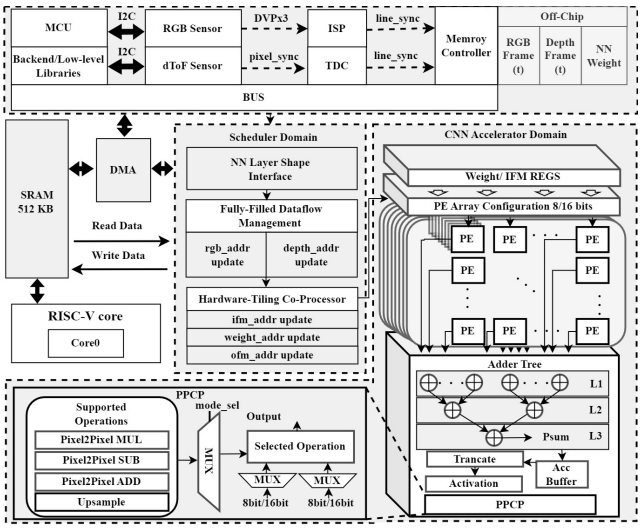

Fig. 2. Overall architecture of the accelerator SoC. The weight parameters of Neural network(NN) and raw data are loaded on the off-chip DRAM.

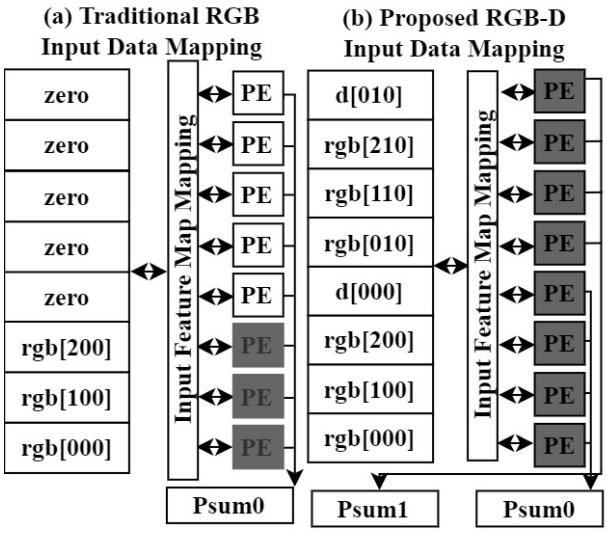

Fig. 3. The data mapping description of fully-filled dataflow management engine (FF-DME) and different number of the output partial sum(Psum). (a) Data mapping on DRAM. (b) Data mapping on PE array.

in this field [5]–[9], realizing depth up-scaling by machine learning(ML) techniques. In [5], by employing the sparse point cloud and the grey-scale data as the inputs of the 3D convolution neural network (CNN), the temporal noise is filtered out and the spatial density is increased. To achieve a better end-to-end super-resolution image quality, [6] takes the optical phase plate pattern and its spatial pattern distribution function as the neural network's prior knowledge, which combines the optical design and the reconstruction neural network to better utilize the illumination pattern information. [7] proposes a two-step interpolation scheme for high-speed 3D sensing and a high-quality imaging system with a frame rate over 1k-fps is demonstrated. A multi-feature neural network using the first depth map and the second depth map is designed to get the up-sampling depth [8] and the feasibility of object segmentation on a small batch is proven [9]. Most of the current research mainly focusing on algorithm design, however, it is necessary for LiDAR systems to guarantee real-time data processing capability, ensuring safe operation during harsh environments. Therefore, the hardware latency, computational load, and efficiency energy ratio in the real scene should be fully considered. Another drawback of the current neural networks demonstrated in the latest works is that they only contain convolution layers, which ignore other non-convolution sufficient modules in computer vision (CV), such as the spatial propagation network(SPN) structure of [10] adopted in this work. In this paper, a depth completion SoC shown in Fig.1 is proposed that includes extended-vector embedded RISC-V for vector computations, a depth engine for the RGB-D input convolution computations in a depth completion neural network, and a hardware tiling co-processor for scheduling the sub-jobs of each convolution layer in CNN. The remainder of this paper is organized as follows. In Section II, the overall SoC architecture with details of designed modules is demonstrated. The measurement results of the SoC are shown in Section III and its performance is concluded in Section IV.

II. SoC Implementation

Fig.2 shows the hardware architecture of the proposed accelerator SoC. The image signal processor (ISP) transfers the input data to the off-chip DRAM after receiving the RGB and depth data from the image sensor and dToF depth sensor, respectively. The four-channel RGB-D data is pre-fetched and aligned by a specific order in the SRAM and then dispatched into the processing element (PE) array by the direct memory access (DMA) module. The CNN accelerator integrates 512 MACs in 32 groups while the 512KB ping-pong SRAM stores the immediate features and weights. Each PE group is configurable for either sixteen 8x8bit MACs or four 16x16 MACs. In Pixel-to-Pixel Co-Processor (PPCP) unit, some element-wise operations, such as add, subtract, multiplication, and up-sampling, are implemented. Finally, the post-processing of the spatial propagation is performed by the RISC-V core.

A. Depth Engine

In the accelerator design, PE utilization is a vital metric, which guarantees that each operation is fully loaded. For different designs, the upper limit of the utilization depends on its parameter setting and the shape of the convolution layer. Therefore, some optimization methods should be adopted to make sure the accelerator is compatible with the algorithm. As shown in Fig.3, the length of the atomic operations at one time in a PE array is limited to eight. To fill up all PEs, five zeros are padded after the original RGB value on the memory footprint, therefore, the PE utilization is only 37.5% for a three-channel RGB-only input. To save the on-chip memory cost and improve the PE utilization, for RGB-D input, a specific data mapping is designed. Firstly, adjacent two pixels are realigned in the depth dimension, which is consistent with the atomic number in this design. Secondly, the partial sum accumulator from one channel is divided into two

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3260116

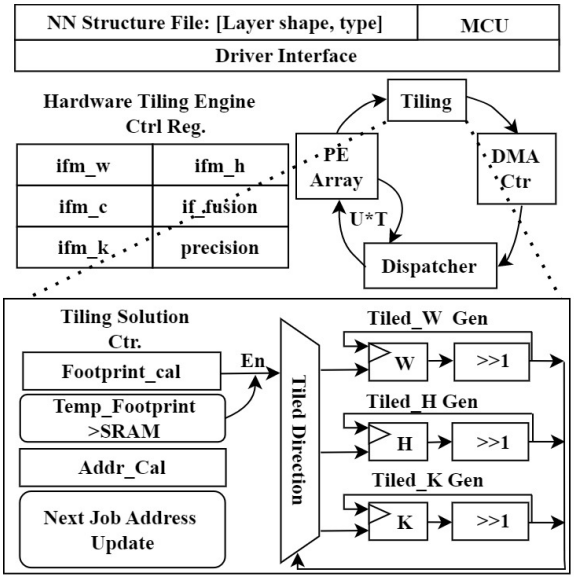

Fig. 4. The state machine of the scheduler. The brief configuration registers including the width, height, channel and kernel of input feature map(ifm_w, ifm_h, ifm_c, ifm_k) are listed. In the tiling stage, the tiled width, height and kernel(Tiled_W, Tiled_H, Tiled_K) are updated to calculate the current footprint on SRAM for an acceptable tiling solution.

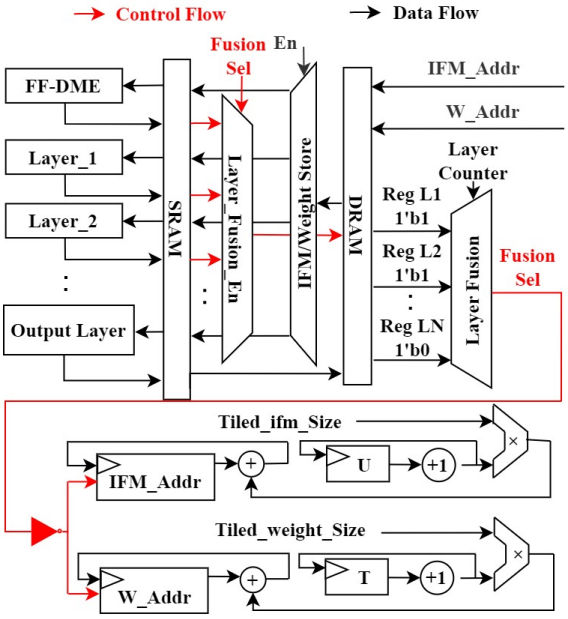

Fig. 5. The block diagram of the dispatcher. The address of input feature map and weight(IFM_Addr, W_Addr) are updated each time depending on the size of the tiled input feature map and weight(Tiled_ifm_Size, Tiled_weight_Size).

partial sums when the input data type is RGB-D, such that the number of output data to the SRAM is doubled. By rearranging the data store on L1-memory and PE partial sum method, the accelerator utilization for RGB-D input is improved to 100% in this work.

B. Scheduler Design

For existing scheduling strategies in research works, many tiling methods are discussed within a single layer, while the

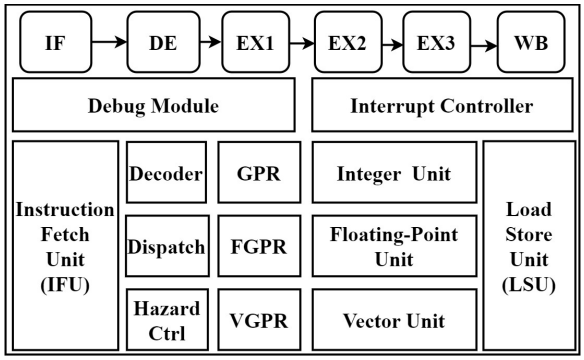

Fig. 6. The block diagram of the RISC-V core. For the Integer Unit, Floating-Point Unit and Vector Unit, the 32 32-bit General Purpose Registers(GPR), 32 32-bit Floating Point General Purpose Registers(FGPA) and 32 256-bit Vector General Purpose Registers(VGPR) are provided for high-level applications.

tiling job configuration time cost from MCU to deep learning accelerator (DLA) module is ignored. Generally, a single convolution job is tiled into tens of sub-jobs for edge-computing platforms. Therefore, the configuration time from firmware should be considered and optimized for edge-computing devices. In this work, a hardware-tiling co-processor is incorporated to reduce the time cost of job tiling and instruction commanding. This hardware-tiling co-processor contains four stages to process either a single layer or several fused layers, as depicted in Fig.4. The tiling stage is responsible for conducting a tiling strategy based on the shape of the current layer, where a dichotomy is adopted in this SoC. After the tiling size combination is determined, the corresponding feature map is transferred between L1 and L2 caches by a DMA module with the dispatcher (Fig.5), sending the current job configurations to the CNN accelerator. In layer-fusion mode, the data throughput within the layers is reduced, lowering the bandwidth required for the memory system.

C. The Architecture of RISC-V and Decomposition of SPN

The computation decomposition of SPN in RSIC-V. The architecture of RISC-V is shown in Fig.6. The designed RISC-V core is composed of single-issue, in-order execution units, and a six-level pipeline RISC-V core, integrating a floating-point unit, integer unit, and vector unit for general computer-vision processing. The vector processor can be configured to execute on eight double, sixteen single, and thirty-two half-precision integer computations, which are compiled into the RISC-V RVV integer instructions.

The SPN process is decomposed as Fig.7. An affinity matrix based on a spatial propagation process is decomposed into vector operations. Every pixel in the blurred depth is refined by its eight nearest-neighbor pixels by the corresponding coefficient from the affinity matrix, which is obtained from the neural network backbone where each channel represents the coefficient in a specific direction. By iterating this spatial propagation refinement process thirty times, the blurred depth is rendered to a high-resolution and high-precision depth output.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems--II: Express Briefs. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2023.3260116

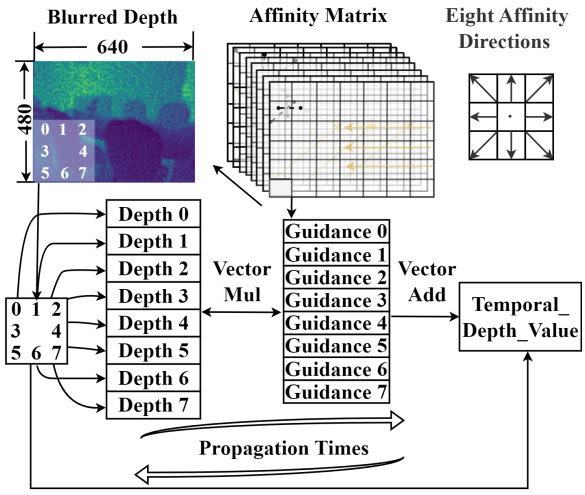

Fig. 7. Decomposed spatial propagation in the proposed RISC-V core.

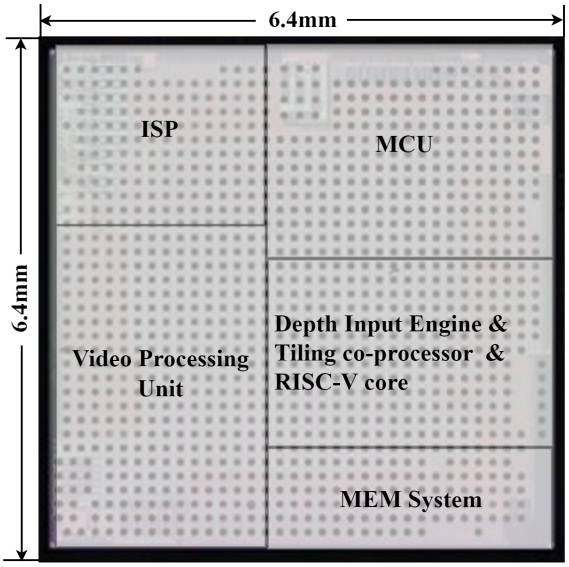

Fig. 8. Chip Micrograph

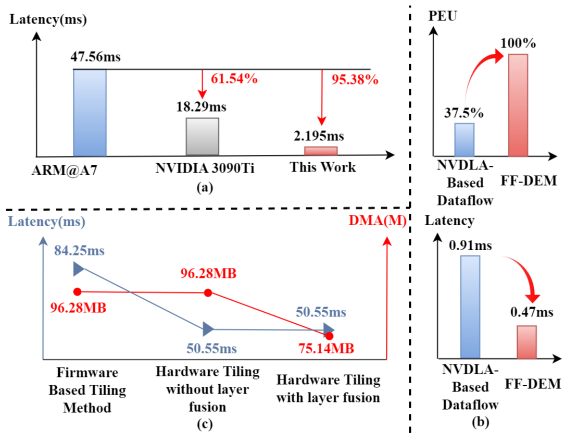

Fig. 9. (a) Measurements and comparisons of post-process in this design. (b) Performance analysis of Fully-Filled Dataflow Management Engine (FF-DME) for VGA-resolution input. (c) Performance analysis of Hardware Tiling Co-Processor.

TABLE I
PERFORMANCE SUMMARY OF THE DEPTH COMPLETION NN

Block Name	Latency (ms)	GOPs (M)	PE Utilization	DRAM Access (MB)
FF-DEM	0.47	126.56	1	5.56
Encoder	18.58	4387.5	0.963	34.74
Decoder	31.5	12561.7	0.97	34.84
Spatial Propagation	2.195	281.25	-	0.151
Entire Neural Network	**Latency (ms)**	**GOPs (G)**	**Efficiency (TOPs/W)**	**Output Resolution**
	52.745	16.95	2	640x480

III. EXPERIMENTAL RESULTS

Fig.8 shows the chip micrograph of the proposed SoC. The measurements are benchmarked versus other common architectures and listed in Fig.9(a). Compared with an NVIDIA 3090Ti, the proposed post-process iteration achieves up to a 2.6x speed-up, while obtaining a 21.6x speed-up when compared with an ARM@A7 core. As shown in Fig.9(b), for a VGA-resolution depth input, the fully-filled dataflow management engine obtains a 100% PEU, in contrast with a 37.5% PEU observed for a traditional NVDLA-based dataflow. Furthermore, by increasing the number of output accumulators, the latency of the input layer decreases to 0.47ms from 0.91ms, when compared with NVDLA. The implementation of the hardware-tiling co-processor improves the configuration time by a 1.6x speed-up. For the consistent layers whose weight sizes are less than half of the on-chip SRAM, the layer sequence is fused to save the external memory access (EMA) overhead. In this work, 8 layers are fused among 27 layers, such that 21% of the EMA area is reduced when compared with the no-fusion tiling method(Fig.9(c)). The measurement results for the proposed fully-filled dataflow management engine, encoder module, decoder module, and benchmarks of the entire neural network are summarized in Table I. The depth completion neural network achieves 28.62ms latency, 34.958MB EMA size, and 97% PEU on average, resulting in a depth-completed VGA (640x480) resolution from a sparse 32x32 point-cloud. The proposed full precision neural network is constructed by Pytorch framework and trained on two NVIDIA 3090Ti graphic cards for 40 epochs on the NYUD dataset [11]. The CPU configuration of the training host is Intel core i7. After the model quantization, the final completed depth point cloud exhibits a depth-rms error-rate of 0.118m, which exhibits 0.009m less accuracy when compared with the float32 model evaluated on the NYUD ground-truth dataset. The tested RISC-V core benchmark is summarized and compared with other similar works in Table II. Compared with [12], to achieve a high frequency, a more dedicated pipeline is designed. Compared with some advanced technology works in [13] and [14], the performances of CoreMark and Dhrystone are considerable.

Table III summarizes the performance of the SoC and the comparison with the state-of-the-art. This proposed AI-accelerator SoC is optimized for a low-resolution depth input while integrating heterogeneous multi-core accelerators for neural networks with a large number of operators. The ASIC

TABLE II
PERFORMANCE SUMMARY OF THE RISC-V CORE

Parameter	Unit	This work	[12]	[13]	[14]
Technology Node	nm	40	40	22	28
Clock Frequency	MHz	500	198.02	961	1000
Pipeline Stage	-	6	4	5	9
CoreMark	per MHz	2.37	4.13	2.37	3.77
Dhrystone	per MHz	1.8	1.71	2.13	-
Core Area	mm^2	0.38	0.036	16	4.86
Power	mW	136.9	1.745	-	-

TABLE III
COMPARISONS WITH THE STATE-OF-THE-ART

Parameters	This work	[15]	[16]	[13]	[17]
Process / Platform	40nm	65nm	Kintex-7 XC7K410T	Intel 22FFL	12nm
Application	Depth Completion	SR	SR	DNN	SR
Depth Input Supported	Yes	No	No	No	No
Layer Fusion Method	Hardware Tiling	Selective Cache	No	No	Depth First
Vector Support	Yes	No	No	Yes	No
Energy Efficiency (TOPs/W)	2	1.1	0.5002 (4x)	0.1036	4.3-15.8
Power(W)	316@400MHZ	NA	NA	278@200MHz	NA
Precision	8/16bit	8/16bit	13bit	8bit	8bit
Throughput	640x480@34fps	1080p@25fps	2880x1280	NA	1080p
Clock Frequency (MHz)	400	250-930	200	130	930
SRAM Size (KB)	512	572	921	1024	1780
Die Area (mm^2)	6.4x6.4	4x4	NA	16	4.5

is fabricated in a 40nm process and occupies a 40.6mm2 area. The energy efficiency of 2TOPS/W for the CNN accelerator with the hardware-tiling co-processor and PPCP is achieved while operating at 400MHz. Compared with [15], [16], a higher energy efficiency ratio and smaller SRAM footprint are achieved, meanwhile, this work supports vector computations with an integrated RISC-V core. Compared with [13], this work is a more specific accelerator design for depth completion application and the better heterogeneous design trade-off between DLA and RISC-V is considered.

IV. Conclusion

In this paper, a real-time depth-completion neural network accelerator SoC designed in 40nm CMOS technology is reported. Firstly, the Fully-Filled Dataflow Management Engine is proposed to improve PE array utilization when the input data is in the RGB-D format. Secondly, an on-chip scheduler is designed to reduce the time consumption of the convolution layer tiling and configuration. Furthermore, a RISC-V core with a vector processor is integrated into this SoC to support the post-process operations in the depth completion algorithm. With the proposed innovations and optimized hardware optimization, a power efficiency of 2TOPS/W is achieved. Also, the neural network accelerator SoC shows a real-time processing capability of 34fps while the output depth map is VGA resolution.

Acknowledgment

The authors acknowledge the efforts of the layout team, SoC team and algorithm team of TiMESiNTELLi Technologies.

References

[1] K. Morimoto and E. Charbon, "High fill-factor miniaturized SPAD arrays with a guard-ring-sharing technique," *Optics express*, vol. 28, no. 9, pp. 13 068–13 080, 2020.

[2] P. Keränen and J. Kostamovaara, "256 × 8 SPAD Array With 256 Column TDCs for a Line Profiling Laser Radar," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 11, pp. 4122–4133, 2019.

[3] C. Zhang, S. Lindner, I. M. Antolovic, M. Wolf, and E. Charbon, "A CMOS SPAD imager with collision detection and 128 dynamically reallocating TDCs for single-photon counting and 3D time-of-flight imaging," *Sensors*, vol. 18, no. 11, p. 4016, 2018.

[4] O. Kumagai, J. Ohmachi, M. Matsumura, S. Yagi, K. Tayu, K. Amagawa, T. Matsukawa, O. Ozawa, D. Hirono, Y. Shinozuka *et al.*, "7.3 A 189× 600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LiDAR systems," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64. IEEE, 2021, pp. 110–112.

[5] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-photon 3D imaging with deep sensor fusion." *ACM Trans. Graph.*, vol. 37, no. 4, pp. 113–1, 2018.

[6] Q. Sun, J. Zhang, X. Dun, B. Ghanem, Y. Peng, and W. Heidrich, "End-to-end learned, optically coded super-resolution SPAD camera," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 2, pp. 1–14, 2020.

[7] I. Gyongy, S. W. Hutchings, A. Halimi, M. Tyler, S. Chan, F. Zhu, S. McLaughlin, R. K. Henderson, and J. Leach, "High-speed 3D sensing via hybrid-mode imaging and guided upsampling," *Optica*, vol. 7, no. 10, pp. 1253–1260, 2020.

[8] A. Ruget, S. McLaughlin, R. K. Henderson, I. Gyongy, A. Halimi, and J. Leach, "Robust super-resolution depth imaging via a multi-feature fusion deep network," *Optics Express*, vol. 29, no. 8, pp. 11 917–11 937, 2021.

[9] G. Mora-Martín, A. Turpin, A. Ruget, A. Halimi, R. Henderson, J. Leach, and I. Gyongy, "High-speed object detection with a single-photon time-of-flight image sensor," *Optics express*, vol. 29, no. 21, pp. 33 184–33 196, 2021.

[10] X. Cheng, P. Wang, and R. Yang, "Learning Depth with Convolutional Spatial Propagation Network," 2018.

[11] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[12] S. Bora and R. Paily, "A high-performance core micro-architecture based on risc-v isa for low power applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 6, pp. 2132–2136, 2021.

[13] A. Gonzalez, J. Zhao, B. Korpan, H. Genc, C. Schmidt, J. Wright, A. Biswas, A. Amid, F. Sheikh, A. Sorokin *et al.*, "A 16mm 2 106.1 GOPS/W Heterogeneous RISC-V Multi-Core Multi-Accelerator SoC in Low-Power 22nm FinFET," in *ESSCIRC 2021-IEEE 47th European Solid State Circuits Conference (ESSCIRC)*. IEEE, 2021, pp. 259–262.

[14] P.-F. Chiu, C. Celio, K. Asanović, D. Patterson, and B. Nikolić, "An out-of-order risc-v processor with resilient low-voltage operation in 28nm cmos," in *2018 IEEE Symposium on VLSI Circuits*, 2018, pp. 61–62.

[15] J. Lee, D. Shin, J. Lee, J. Lee, S. Kang, and H.-J. Yoo, "A full HD 60 fps CNN super resolution processor with selective caching based layer fusion for mobile devices," in *2019 Symposium on VLSI Circuits*. IEEE, 2019, pp. C302–C303.

[16] J.-W. Chang, K.-W. Kang, and S.-J. Kang, "An energy-efficient FPGA-based deconvolutional neural networks accelerator for single image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 281–295, 2018.

[17] K. Goetschalckx and M. Verhelst, "Depfin: A 12nm, 3.8tops depth-first cnn processor for high res. image processing," in *2021 Symposium on VLSI Circuits*, 2021, pp. 1–2.