A 1920×1080 129fps 4.3pJ/pixel Stereo-Matching Processor for Pico Aerial Vehicles

Pingcheng Dong[1, 2, 3], Zhuoyu Chen[1], Ke Li[1], Lei Chen[1], Kwang-Ting Cheng[2, 3], Fengwei An[1]
[1]Southern University of Science and Technology, China; [2]The Hong Kong University of Science and Technology, Hong Kong;
[3]AI Chip Center for Emerging Smart System (ACCESS), Hong Kong; Email: anfw@sustech.edu.cn

ESSCIRC 2023- IEEE 49th European Solid State Circuits Conference (ESSCIRC) | 979-8-3503-0420-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ESSCIRC59616.2023.10268790

***Abstract*—This paper presents an energy-efficient stereo matching processor that fully leverages disparity and RGB images to achieve high-speed and low-power depth prediction in 28nm CMOS process. A data compression-based guided image filter is proposed to compensate for the accuracy degradation caused by insufficient aggregation paths with slightly extra hardware resources. Two SRAM architectures tailored for the aggregation and multi-linebuffer in stereo matching achieve a compact chip area of only 1.26mm², which enables 30-129fps dense depth throughput on full HD resolution (1920×1080) with only 57-147mW power dissipation. Besides, its normalized energy dissipation outperforms the state-of-the-art in 28nm CMOS by 9.6× even with a larger on-chip SRAM footprint.**

Keywords*—Stereo vision, semi-global matching, guided image filtering, pico aerial vehicles.

I. Introduction

Stereo vision has consistently served as a fundamental component for a myriad of visual applications, including pose estimation, robot navigation, and autonomous vehicles. Presently, the mainstream approaches can be categorized into two branches: the first exploits powerful deep learning algorithms on GPU devices, which could yield remarkable performance but accompanied by exceptionally high computational requirements [1], even amounting to over 700GMACs at HD resolution; the second involves designing dedicated hardware architectures in low-power and high-speed fashions, to accelerate traditional stereo matching algorithms exemplified by semi-global matching (SGM) on FPGA or ASIC platform [2]. Nonetheless, in the context of pico aerial vehicles (PAV), the limitations pertaining to area and endurance preclude the acceleration on GPUs. Thus, designing stereo vision processors is still the preferred strategy in practical applications.

Although several stereo-matching processors have been proposed for autonomous vehicles, they prevalently rely on improving aggregation algorithms in semi-global matching (SGM) with extra hardware resources [2] or bringing in guided image filtering (GIF) [3] for higher accuracy [4]. In contrast, PAVs highly entail real-time and low-power processing due to limited space. As shown in Fig. 1, a typical stereo-matching processor oriented toward PAVs encounters three major challenges: 1) A multi-bank solution for high-bandwidth memory access in aggregation leads to poor convergence in timing performance and a large chip area. 2) Frequent multi-linebuffer data caching results in excessive SRAMs, occupying a significant portion of chip areas and power consumption. 3) Applying GIF for greater accuracy incurs enormous memory utilization and severe area/power overheads.

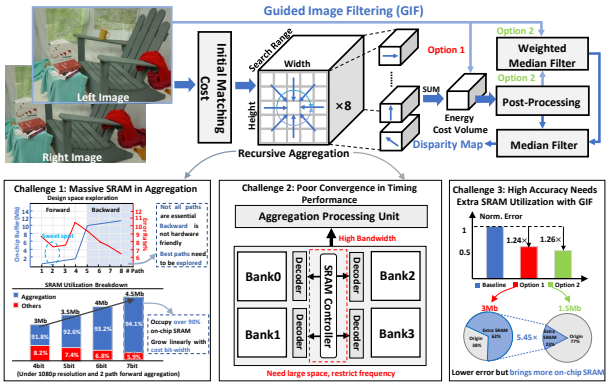


Fig. 1. Challenges in a stereo matching processor for PAVs.

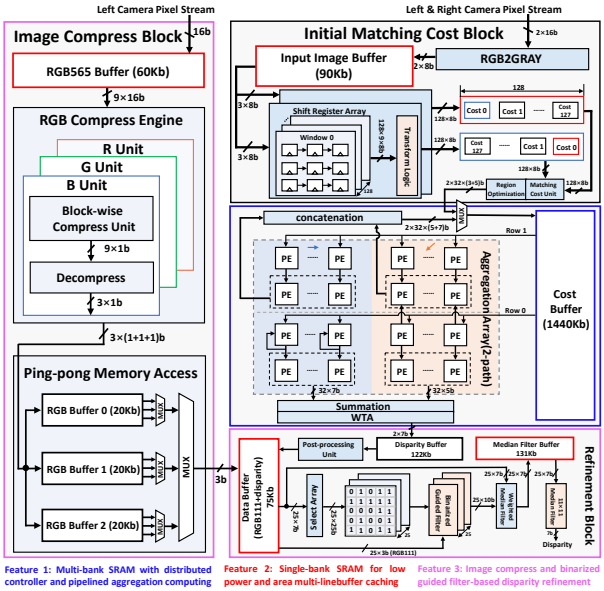


Fig. 2. Overall architecture of the stereo matching processor.

The stereo-matching processor developed in this work features three main characteristics. Firstly, we introduce distributed multi-bank SRAM controllers and pipelined aggregation computing schemes based on the homogeneity of read and write behaviors in the aggregation process. Furthermore, to alleviate the area overhead generated by the multi-linebuffer, we employ a sliced bitwise write structure to unify multiple linebuffers into a single-bank SRAM. Lastly, we propose a binarized guided filter architecture, achieving 16× RGB data compression and effectively avoiding numerous division operations.

II. Architecture

Figure 2 shows the overall architecture in which the initial matching cost block calculates the 128×2 initial costs by census transform and hamming distance between the left and right grayscale images. As demonstrated in [5], we can compress the costs into 32×2 and aggregate them in two paths through design space exploration. Then, the aggregation array optimizes the data in the costs buffer in a pipeline fashion. In the other branch, the left image in RGB565 format, which serves as the guided image, is binarized to RGB111 in the

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on November 17,2023 at 14:07:25 UTC from IEEE Xplore. Restrictions apply.

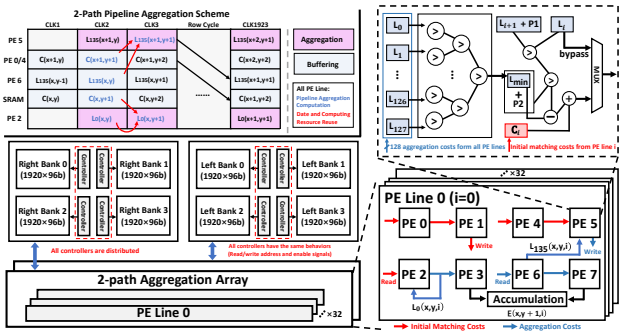


Fig. 3. Multi-bank SRAM with distributed controllers and a 2-path pipelined aggregation scheme.

image compress block based on the range and median in each non-overlapped local window. The refinement block improves the disparity map in the post-processing unit simultaneously, and the select array generates 25 binary maps through bitwise comparison in each 5×5 7-bit disparity window. After RGB and disparity binarization, several guided image filters are performed on 25 binary maps with the guided image in RGB111 to generate 25 weights for a 5×5 weighted median filter. Finally, an 11×11 median filter removes the noises.

A. Distributed Controllers and Pipelined Aggregation

The aggregation in SGM generally demands extra support for high bandwidth on-chip SRAM access. For instance, the 30fps depth prediction requires at least 46.45Gb/s for HD resolution, significantly exacerbating area and power overheads. To alleviate this problem, a customized SRAM with multi-bank SRAM access and disconnecting read transistor and RBL is proposed in [6]. Although innovatively customizing multi-bank SRAM can reduce power consumption, partitioning banks on blocks brings challenges in timing convergence, and the customized SRAM methodology suffers from limited generality.

To tackle this challenge, we first analyze the process of the high-bandwidth SRAM in aggregation from the perspective of disparity dimensions and discover that the SRAM read/write behavior of each disparity exhibits homogeneity, i.e., identical read/write signals and addresses. Thus, we propose partitioning the required SRAM into multiple banks in terms of data bit width, where each bank has 1920×96-bit. However, this still leads to poor convergence in the timing performance, primarily due to multiple banks sharing a single SRAM controller, resulting in excessive fanout. Therefore, we replicate the same controller for each bank in a distributed manner, providing the same read/write signals and addresses. In Fig. 3, the distributed controllers for each bank sacrifice a small amount of logical resources but attain a compact layout area with greater timing convergence, and this approach boosts frequency by 1.54× to 183.8Gb/s.

Since each bank contains a part of the initial costs, the PEs are distributed in the vicinity of each bank and aggregate the related initial or aggregated costs. Besides, multi-path aggregation is a recursive optimization strategy. It results in a loop-carried dependency, i.e., the current cost needs to be updated based on its previous one. Moreover, two levels of data dependency exist in the proposed work due to two directions. One is temporal, where the aggregation in the 0° path requires data from the previous cycle, and the other level is spatial such as 135° aggregation necessitates data from the previous row. To resolve these dependencies, we first unroll the recursive aggregation loop, allowing PE units near each bank to concurrently update the 0° costs at (x, y+1) and the 135° costs at (x+1, y+1) within the bank. As can be seen in Fig. 3, the updated costs are broadcasted to all PE units at the end of the cycle. Utilizing this pipelined approach, the computing resources can effectively cooperate with the multi-bank SRAM while extensively reusing the intermedia results.

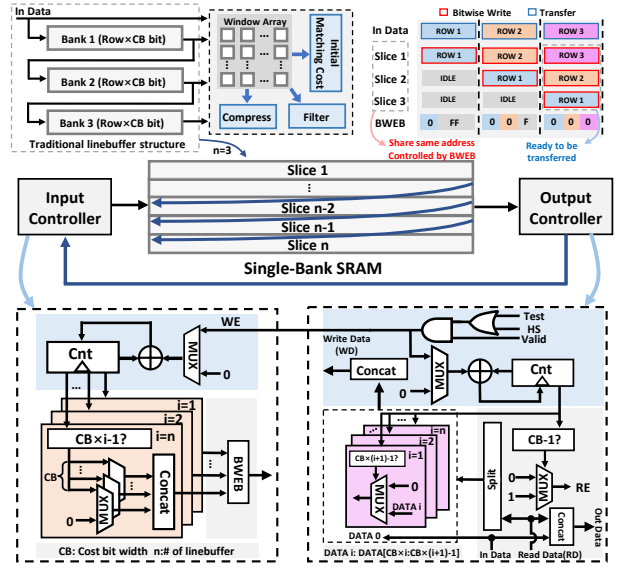


Fig. 4. Single-bank SRAM for multi-linebuffer caching.

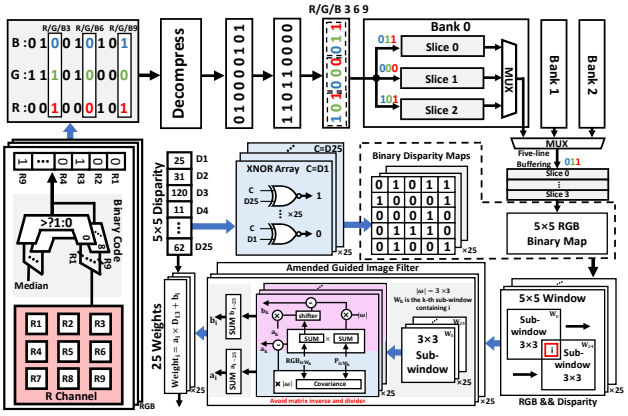


Fig. 5. RGB-Compress and Binarized Guided Image Filtering.

B. Single-Bank SRAM for Multi-Linebuffer Caching

The multi-linebuffer structure is a widely used technique in traditional image processing, especially in the domain of stereo vision [7]. This work aims to explore a single-bank, rather than a multi-bank, SRAM while satisfying data bandwidth demands and simultaneously multi-lined caching to reduce the chip area further. Although the bandwidth of the multi-linebuffer SRAM is considerably smaller than that of aggregation, the frequent utilization of window caching exacerbates the overhead of chip area and power dissipation. To this end, we propose a single-bank SRAM controller tailored for multi-line buffers.

As shown in Fig. 4, the output of each slice is connected to the input of the next adjacent slice, imitating the

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on November 17,2026 at 14:07:35 UTC from IEEE Xplore. Restrictions apply.

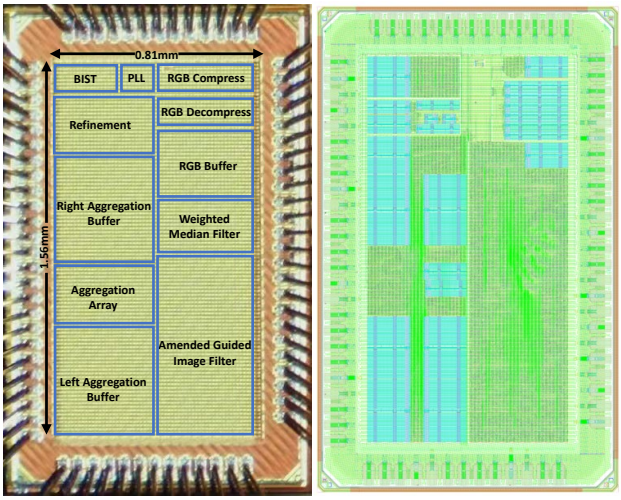


Fig. 6. Die photograph and chip layout.



Fig. 7. Demonstration system.

interconnection of line buffers. Moreover, the input controller is responsible for activating individual slices by controlling the segment of the bitwise-write-enable signal based on the row counter. Meanwhile, the output controller splits the read data of SRAM into multiple components, subsequently allocating them to the corresponding port of the slice. Upon reaching total capacity in the last SRAM slice, all outputs of the slice together with *In_Data* are concatenated to form a column of window data.

C. RGB-Compressed and Binarized Guided Image Filter

As discussed in Figure 1, two primary options exist for applying the GIF to improve accuracy. The first option encompasses filtering each cost layer generated by the aggregation. In contrast, the alternative option applies the GIF to a single disparity layer incorporated into the refinement stage. Given the considerable bandwidth during the aggregation stage, option 1 is not hardware-friendly as it results in a much larger SRAM utilization. In contrast, option 2 can attain a comparable enhancement ability while saving computation and memory, as proved in [8].

Nonetheless, GIF typically requires guidance from a reference image. Option 2 generally employs an RGB input image for guidance, introducing extra memory to buffer the initial input RGB until the refined disparity map is completed. To minimize this memory overhead, we divide the RGB565 image into non-overlapping 3×3 blocks (Fig. 5), perform a bitwise comparison for data binarization, and use the median values as a threshold to reconstruct an RGB111 map with a 16× SRAM compression ratio. But this compression scheme leads to data transfer and block size inconsistencies since the transmission of compressed RGB111 data is 3×3 blockwise and non-overlapping. At the same time, the GIF in the refinement stage requires a sequential and overlapped 5×5 window RGB111 pixel stream. To handle these data inconsistencies, we proposed a ping-pong memory access-based architecture in Fig. 5. The non-overlapping blockwise compression results in two pipelined stalls after each RGB data binarization. We introduce a decompress module to combine the RGB column data and provide sequential output, which could resolve the data transfer conflicts.

On the other hand, we alternately write three R/G/B111 column data into three SRAM banks for buffering, but the GIF demands access to five data in a column. Thus, we design a Ping-Pong memory accessing structure in two levels. The GIF selects two banks containing five rows in the first level, then identifies relevant five slices in two banks, resolving the block size inconsistency.

Besides, the GIF incorporates a substantial amount of division operations to obtain weight, which occupies almost 45% of the chip area and intensifies the power overhead. However, the range of the denominator in GIF tends to be constant due to binarization. We convert all divisions into multiplications by expanding the same scaling factor. Significantly, the overall computing resources are reduced by over 70%, and there is no loss in accuracy even without the matrix inverse and regularization terms.

III. Measurement Results

Fig. 6 shows the die photograph and chip layout, the stereo matching processor in this paper is fabricated in TSMC 28nm CMOS process with only a 1.26mm^2 chip area. In the test setup shown in Fig. 7, the left and right images are captured by ov5640 binocular cameras and transmitted via two Digital Video Port (DVP) interfaces. Subsequently, the host sends calibration parameters through USB, and the image rectification is performed on the Spartan-7 FPGA to obtain rectified left and right images on the same epipolar plane. The FPGA then sends pixel streams to the chip for stereo matching,

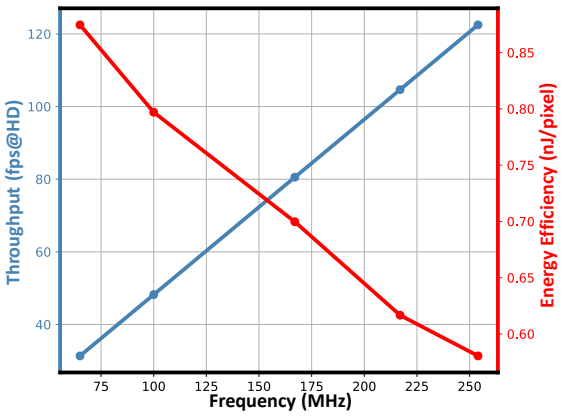


Fig. 8. Measured throughput and energy efficiency versus different frequencies.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on November 17,2026 at 14:07:35 UTC from IEEE Xplore. Restrictions apply.

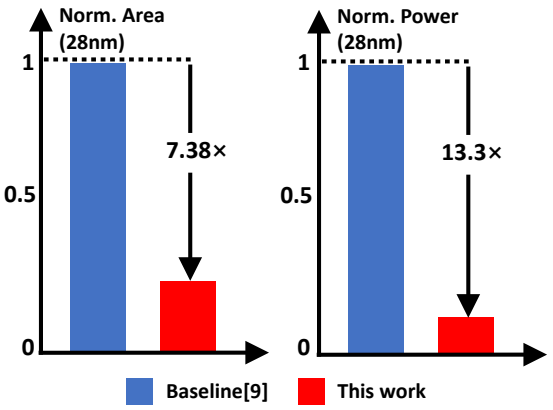


Fig. 9. Comparison of normalized chip area and power dissipation with baseline work [9] in TSMC 28nm process.

	ISSCC 2017 [6]	VLSI 2018 [9]	CICC 2022 [7]	This work
Method	SGM	NG-SGM	Block Patchmatch	rSGM+GIF
Technology/ FPGA	40nm	28nm	Xilinx ZCU102	28nm
Area	10.8mm²	9.3mm²	-	1.26mm²
On-chip memory	1064Kb	1568Kb	10.26Mb	1857Kb
Search range	128+2bit fractional	30976	128	128
Frequency	170MHz	180MHz	350MHz	65-254MHz
Throughput & voltage	30fps@0.75V*	30fps@1.3V*	165.7fps*	30fps@(0.9V, 65MHz)* 129fps@(0.9V, 267MHz)*
Power	836mW@30fps*	760mW@30fps*	3.4W*	57mW@30fps* 147mW@129fps*
Normalized energy#	26.2nJ@30fps*	69nJ@30fps*	76.2nJ*	7.16pJ@30fps* 4.3pJ@129fps*
Error rates	7%@KITTI	4.7%@Middlebury	6.31%@KITTI	7.48%@KITTI

*: 1920×1080 resolution, #: Power/(search range × # of pixels)

Fig. 10. Comparison to the state-of-the-art stereo processors.

and the final resulting depth map is displayed on the screen via the HDMI interface.

The measurement throughput and energy efficiency results with varying frequencies are shown in Fig. 8. The maximum operating frequency attains 254MHz, where the throughput can reach 129fps, and the energy efficiency is only 0.55nJ/pixel. Since the processor in this work features a fully pipelined architecture without any off-chip memory, the throughput and energy efficiency exhibit positive and negative linear relationships with operating frequency under the standard supply voltage of 0.9V. Moreover, the proposed stereo matching processor improves the chip area and power dissipation by 7.38× and 13.3×, respectively, compared to the baseline work [9] under the same TSMC 28nm CMOS process in Fig. 9.

The detailed comparison results with other state-of-the-art designs in [6], [7], and [9] are summarized in Fig. 10. Proposed processor achieves a 7.48% outlier for depth in KITTI automotive benchmark. Although the patch match method in [7] can reach the best accuracy, it brings in an extensive SRAM utilization (10.26Mb) which is not practicable for resource-constrained devices. In contrast to the ASIC designs presented in [6] and [9], despite the largest SRAM usage in the proposed work, its chip area remains at a mere 1.26mm^2 with the help of compact architectures and algorithm-oriented memory designs. It consumes only 57-147mW/7.16-4.3pJ to process the depth map at 30-129 fps for HD resolution. It is worth noting that the power and normalized energy of the proposed processor achieves the new state-of-the-art for 30fps stereo matching processors in both ASIC and FPGA.

IV. Conclusion

This paper reported a 129fps and 4.3pJ/pixel stereo matching processor in TSMC 28nm CMOS process. The distributed multi-bank SRAM controllers with pipelined computing scheme boost the bandwidth to 183.8Gb/s and alleviate the area overhead. The proposed single-bank SRAM for multi-linebuffer data caching further saves numerous on-chip SRAM and helps achieve a compact chip area with 1.26mm^2. The binarized guided image filtering refinement based on RGB compression attains a 16× RGB image compression ratio and 45% chip area reduction while enhancing the accuracy to a 7.48% error rate in the KITTI benchmark. The proposed stereo-matching processor provides promising architecture design methods for low-power and high-speed depth estimation on resource-constrained applications.

References

[1] A. Kendall et al., "End-to-End Learning of Geometry and Context for Deep Stereo Regression," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 66-75, doi: 10.1109/ICCV.2017.17.

[2] Y. Lee and H. Kim, "A High-Throughput Depth Estimation Processor for Accurate Semiglobal Stereo Matching Using Pipelined Inter-Pixel Aggregation," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 411-422, Jan. 2022, doi: 10.1109/TCSVT.2021.3061200.

[3] K. He, J. Sun and X. Tang, "Guided Image Filtering," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 6, pp. 1397-1409, June 2013, doi: 10.1109/TPAMI.2012.213.

[4] C. Ttofis, C. Kyrkou and T. Theocharides, "A Low-Cost Real-Time Embedded Stereo Vision System for Accurate Disparity Estimation Based on Guided Image Filtering," in IEEE Transactions on Computers, vol. 65, no. 9, pp. 2678-2693, 1 Sept. 2016, doi: 10.1109/TC.2015.2506567.

[5] P. Dong, Z. Chen, Z. Li, Y. Fu, L. Chen and F. An, "A 4.29nJ/pixel Stereo Depth Coprocessor With Pixel Level Pipeline and Region Optimized Semi-Global Matching for IoT Application," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 1, pp. 334-346, Jan. 2022, doi: 10.1109/TCSI.2021.3100071.

[6] Z. Li et al., "A 1920×1080 30fps 2.3TOPS/W stereo-depth processor for robust autonomous navigation," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2017, pp. 62-63, doi: 10.1109/ISSCC.2017.7870261.

[7] H. Wang, W. Zhou, X. Zhang and X. Lou, "A 39pJ/label 1920x1080 165.7 FPS Block PatchMatch Based Stereo Matching Processor on FPGA," 2022 IEEE Custom Integrated Circuits Conference (CICC), Newport Beach, CA, USA, 2022, pp. 1-2, doi: 10.1109/CICC53496.2022.9772830.

[8] Z. Ma, K. He, Y. Wei, J. Sun and E. Wu, "Constant Time Weighted Median Filtering for Stereo Matching and Beyond," 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 2013, pp. 49-56, doi: 10.1109/ICCV.2013.13.

[9] Z. Li, J. Wang, D. Sylvester, D. Blaauw and H. -S. Kim, "A1920 × 1080 25FPS, 2.4TOPS/W Unified Optical Flow and Depth 6D Vision Processor for Energy-Efficient, Low Power Autonomous Navigation," 2018 IEEE Symposium on VLSI Circuits, Honolulu, HI, USA, 2018, pp. 135-136, doi: 10.1109/VLSIC.2018.8502347.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on November 17,2026 at 14:07:35 UTC from IEEE Xplore. Restrictions apply.