CogniVision: A mW Power envelope SoC for Always-on Smart Vision in 40nm

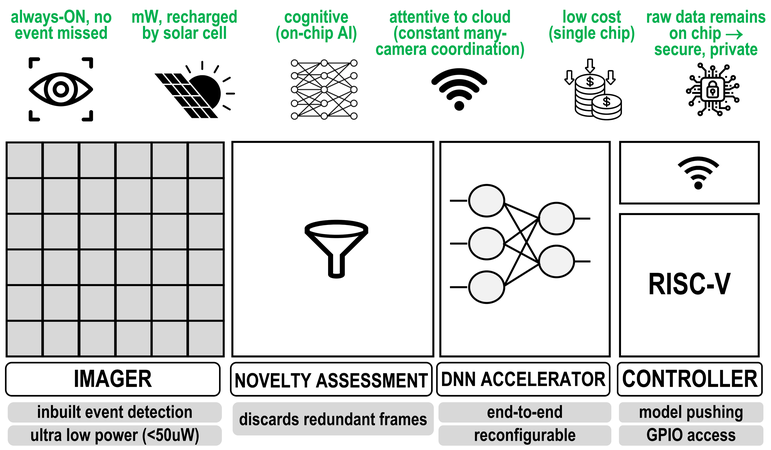
Animesh Gupta*, Japesh Vohra* and Massimo Alioto, National University of Singapore
(* Equal Contributing Authors)

ABSTRACT

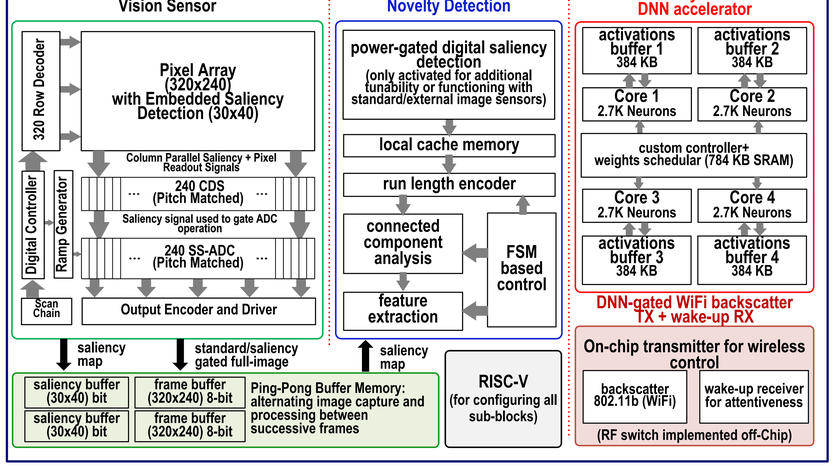
A full vision system on chip with hierarchical execution from sensor to AI and communications is presented. Always-on system power is aggressively reduced to the mW level through activity reduction, gating any subsequent vision pipeline stage starting from the lowest semantic level of the scene. The system comprises an imager with dual-architecture in/near-sensor saliency detection, on the-fly novelty detection, DNN accelerator with on-chip scheduler for weight memory reduction, WiFi transmitter, wake-up receiver for cloud pushed DNN model update and software programmable orchestration via RISC-V. Average 2.1-mW power at 30 fps is achieved with a SqueezeNet V1.0 model running entirely on chip.

MOTIVATION

REQUIREMENTS OF HIGHLY-DISTRIBUTED CAMERAS

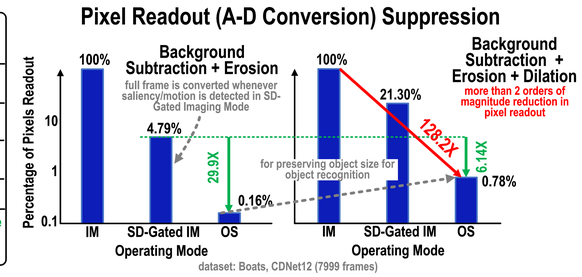


VISION SYSTEM ARCHITECTURE

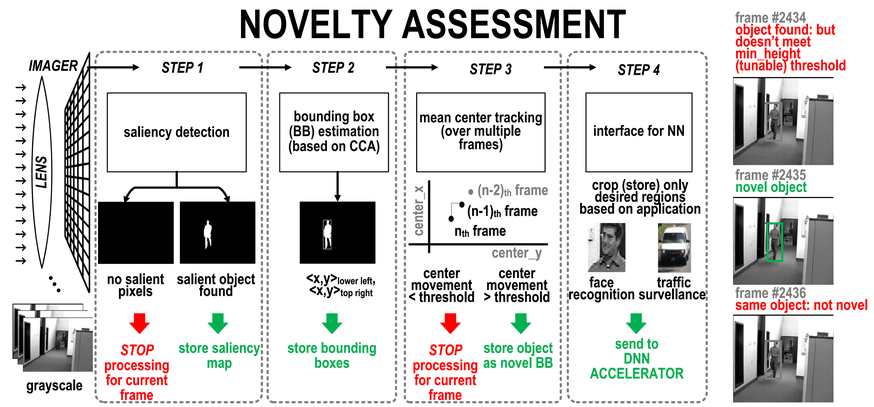


EVENT DETECTOR IMAGER

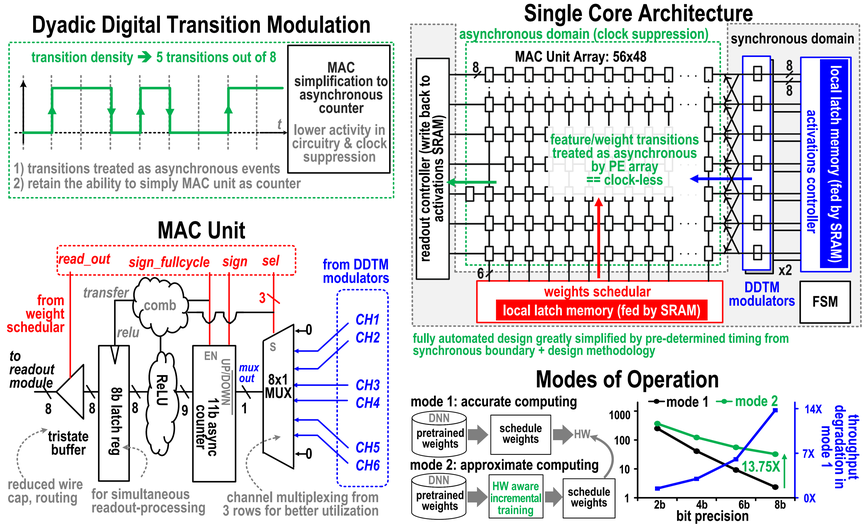
	conventional	this work
Readout Suppression	NO	YES
Memory Update Suppression	NO	YES
Neural Network Suppression	YES	YES
Data Locality	NO	YES
Morphological Filters	Erosion	Programmable Erosion + Dilation



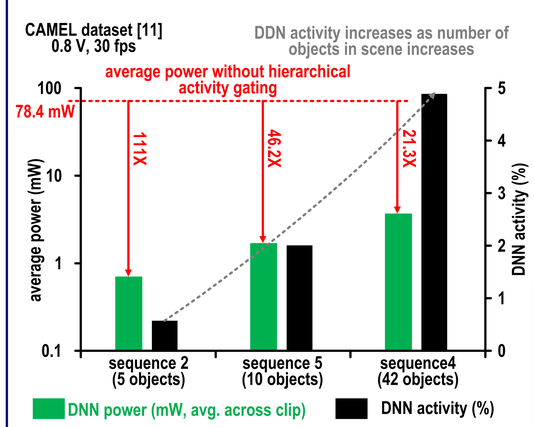
NOVELTY ASSESSMENT

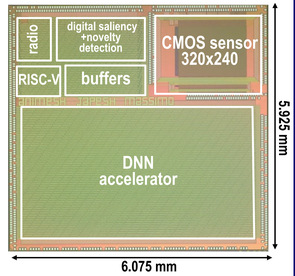


DNN ACCELERATOR



IMPLEMENTATION DETAILS





Testchip Summary	
technology	40 nm
die area	36 mm²
memory	~3 MB
supply voltage	0.7-1.1 V
clock frequency	DNN: 500 MHz other: 5 MHz

COMPARISON TABLE

	this work	ISSCC'23 [1]	VLSI'23 [2]	VLSI'23 [3]	ISSCC'22 [4]	VLSI'22 [5]	VLSI'21 [6]	JSSC'21 [7]
technology	40	40	28	12	40	22	22	40
area (mm²)	36	7.75	9.72	1.76	184	8.76	3.42	26.88
application	General AoV	General AoV	General AoV	General AoV	Edge AI	Micro Robot Navigation	AoV FD+FR	Image Recognition
computation type	**time-domain MAC**	digital MAC	ACIM	DCIM	ACIM	digital MAC	digital MAC	digital MAC
CMOS sensor	**on-chip**	external	**on-chip**	external	external	external	external	external
resolution	320x240	320x240	128x128	128x128	-	640x480	320x160	640x480
imager-to-comms hierarchal execution	**YES**	NO	NO	NO	NO	NO	NO	NO
MAC precision (A=activations, W=weights)	2-8b (W) 8b (A)	1/4/8b (W, A)	1-8b (W, A)	4/8 (W, A)	2-8b (W, A)	8/16 (W, A)	1/16 (W) 16 (A)	8b (W, A)
memory (kB)	3086	1024	100	512	5325	1428 + 2048 NVM	1176	1152
DNN supply (V)	0.6-1.1	0.9	1	0.6	-	0.5-1.0	0.65	0.58-0.7
frequency (MHz)	350	31	200	100-200	-	0.056-190	180-220	0.153
peak TOPS	**3 (8b)**	0.00126 [1]	-	0.0512 (8b)	**17 (8b) 60 (2b)**	0.51 (8b)	0.006 (1b) 0.005 (16b)	-
TOPS/W	**29.45 (8b) [2]**	1.514 [1]	25.86 (8b) 709.3 (1b)	46.8 (8b) 124.1 (4b)	3.3 (8b) 10.9 (2b)	12.1 (8b) 3.5 (16b)	0.32 (16b) 1.07 (1b)	1.5 (8b)
peak fps	30 [2]	5	-	15	90	554	20	5
power (fps) (mW)	**2.1mW (30) [2]**	0.83 (5)[3]	7.3 (-)	2.86 (15) 0.87 (2)	3730 (90)[3]	158 (554)[3] 10.2 (20)[3]	17.23 (20) 2.11 (5) 0.64 (1)	0.17 (0.16)[3]
energy/frame.pixel (nJ)	**1.16 [2] (full system)**	2.16[3]	-	11.64	-	1.66[3]	12.5	3.46[3]
DNN type	SqueezeNet v1.0	MobileNet-V1 0.25	ResNet 20	MobileNet-v1	AlexNet	ResNet-8	Custom	Custom

One addition + one multiplication = two operations. 1) 740k cycles at 128x128 MobileNet V1 0.25 0.9V 31MHz, 2) at 0.8V 3) imager power not included

REFERENCES:

[1] D. Garrett, ISSCC 2023, [2] W. Xie, VLSI 2023, [3] E. Chang, VLSI 2023, [4] L. Fick, ISSCC 2022, [5] Q. Zhang, VLSI 2022, [6] P. Jokic, VLSI 2021, [7] H. An, JSSC 2021, [8] A. Sayal et al., JSSC 2021, [9] A. Gupta, CICC 2022, [10] Y. Wang, CVPR 2014, [11] E. Gebhardt, AVSS 2018

TO KNOW MORE: