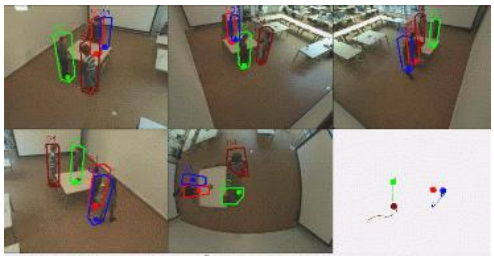
Annex D – Additional diagrams and figures

Table I. Additional examples of state-of-the-art deep neural networks with related accuracy and hardware cost in GPU implementations.

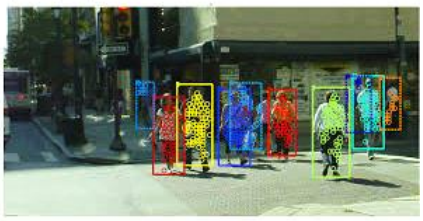
deep neural network	task	accuracy	memory, computational effort
ResNet [HZR2016]	10 million images from 1,000 categories	3.57% top-5 error rate	12GB memory requirement, 0.5-seconds/image in best-in-class GPUs
ResNet + Faster R-CNN [HZR2016]	detect objects from 20 different categories	83.8%	12-GB GPU memory and 0.5 seconds/object detection on 256x256 images (similar to [DLH2016], [LAE2015])
DeepLab-V2 [CPK2016]	segmenting images containing various objects from 20 categories	79.7%	6-GB memory cost and 2 seconds/image with best-in-class GPUs
VGG16 [ZLL2016]	pedestrian detection	9.6% miss rate	6-GB memory and 0.5 seconds/pedestrian
multi-domain network [NH2016]	tracks objects across video frames	3% error rate	6-GB memory and 1 frame/s in best-in-class GPUs
RCNN-based face detector [CHW2016], [WOJ2015]	detects and align faces	98.35%	6-GB memory and 30 images/s in best-in-class GPUs
GoogleNet [SKP2015], [WZL2016]	face verification	99.63%	6-GB memory, 20 images/s rate in best-in-class GPUs
GoogleNet [ZLL2016], [ZGW2016]	facial emotion classification	97.3%	6-GB memory, 20 images/s rate in best-in-class GPUs

Internet of
(visual) Things

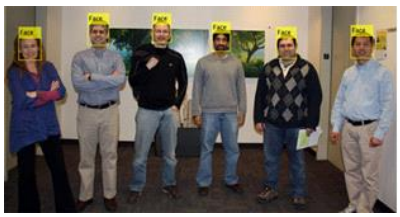


object detection/classification, visual indexing (heavy loads, vehicles…)

intelligent/predictive transportation (MRT…)



crowd volume/motion monitoring for predictive infrastructure/access mgmt



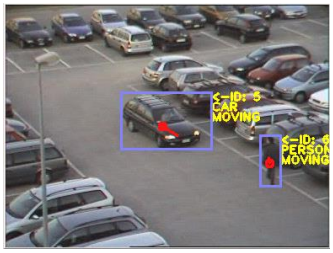
targeted human search, person/face recognition, occupancy monitoring

vehicle/pedestrian danger prediction (e.g., collision)

warehouse management

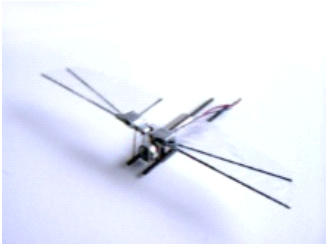
ubiquitous surveillance (abandoned object detection…)

crowd sentiment monitoring (normal, frantic, panicking…)



augmented-reality information-enriched surveillance

human activity monitoring (congestion, congregation, loitering)



vision in autonomous ultra-lightweight UAVs

object removal detection

border control and secure neighborhoods

monitoring for high-productivity manufacturing

Fig. D1. Societal impact of CogniVision: examples of applications that are enabled by (or benefit from) cognitive cameras.

a

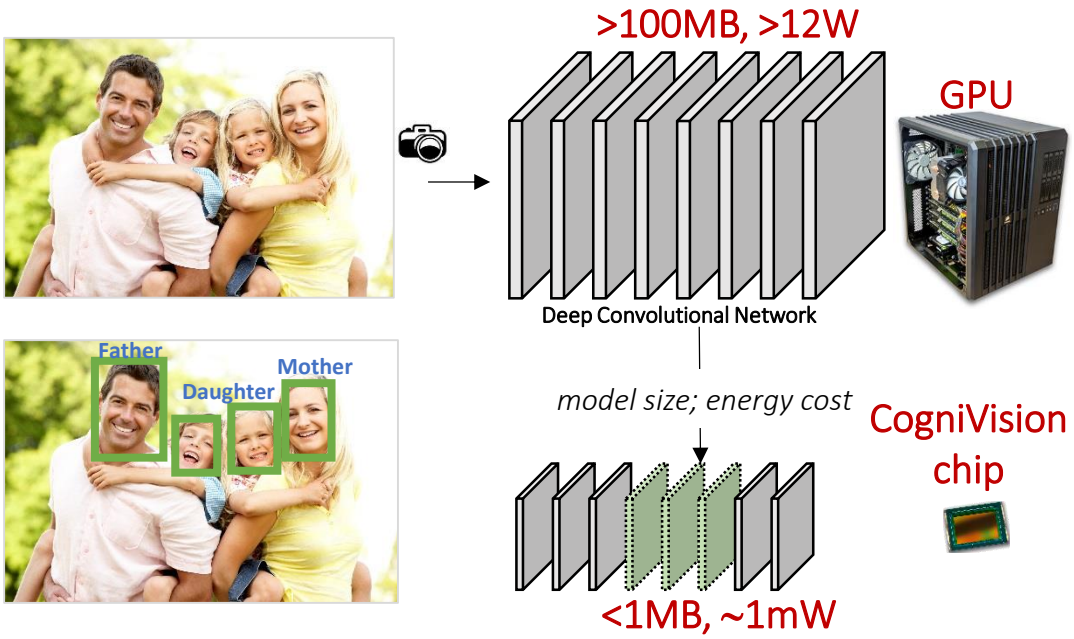


Fig. D2. Memory size and power requirements of GPU-scale and CogniVision chip-scale vision and example (face recognition).

Fig. D3. The industrial interest in embedded vision is growing rapidly, as testified by the large number of enterprises that joined the Embedded Vision Alliance [EVA].

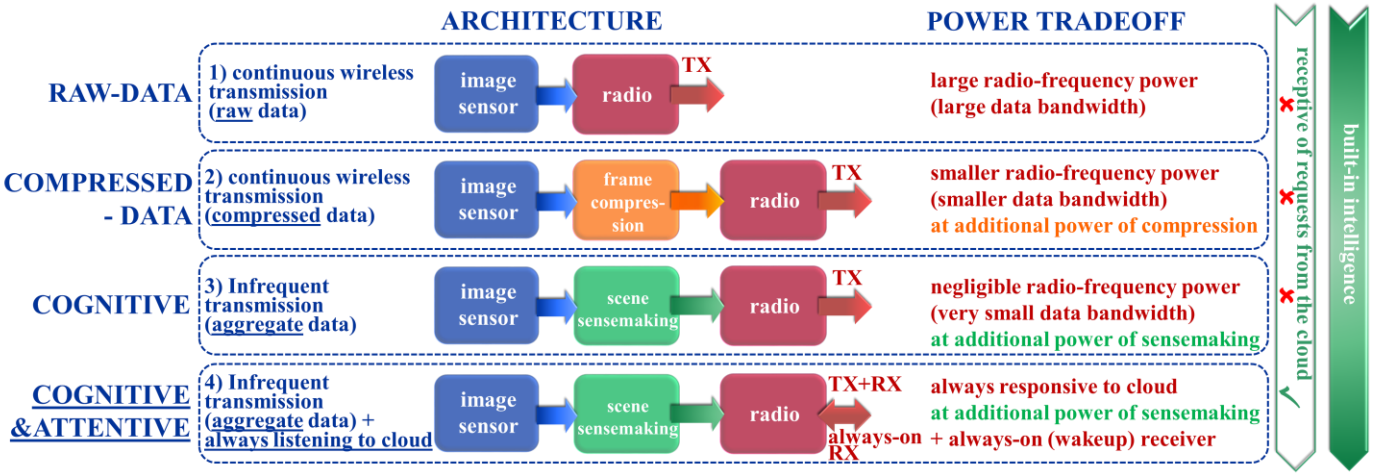


Fig. D4. General architectures for untethered cameras (CogniVision adopts the "cognitive&attentive" architecture to drastically reduce the radio-frequency transmitted power, and enables continuous responsiveness to the cloud requests through ultra-low power always-on receiver).

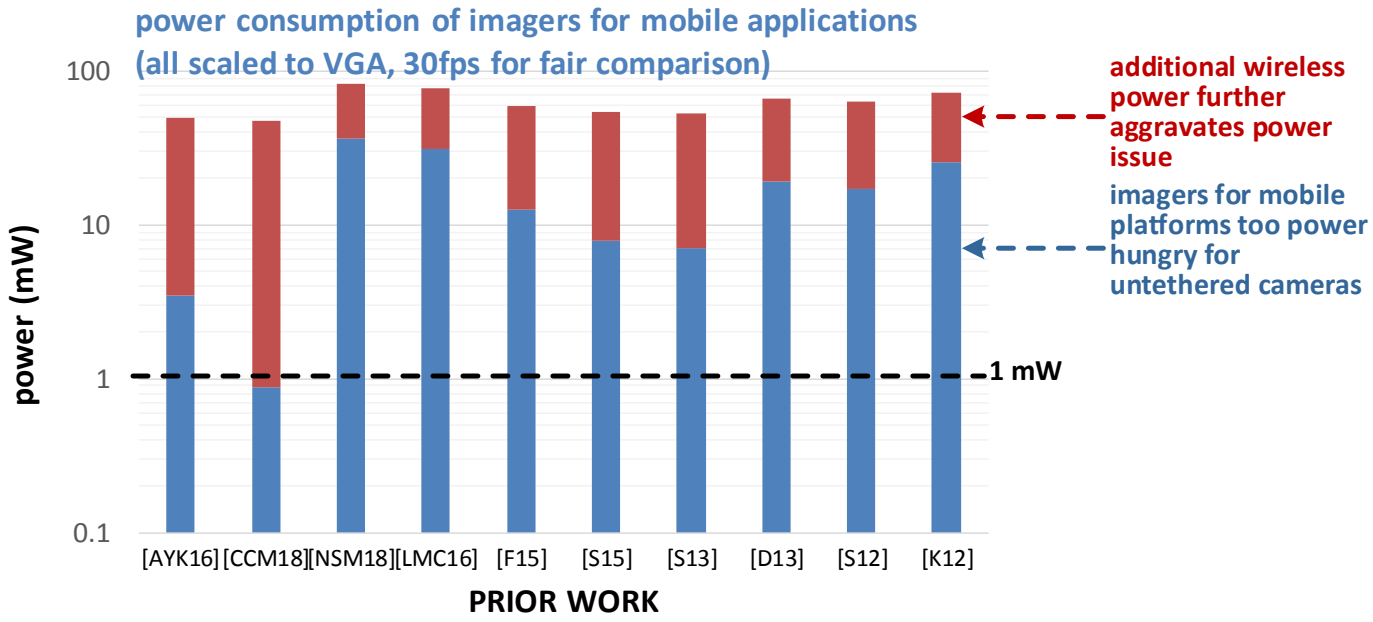


Fig. D5a. Power consumption of state-of-the-art imagers for mobile applications and additional wireless power (assuming an optimistic 5 nJ/bit - representative of best-in-class radios [ITT16]). In these plots, for fair comparison the power of imagers is scaled to VGA format at 30 frame/second by optimistically retaining the same energy/pixel at such requirements.

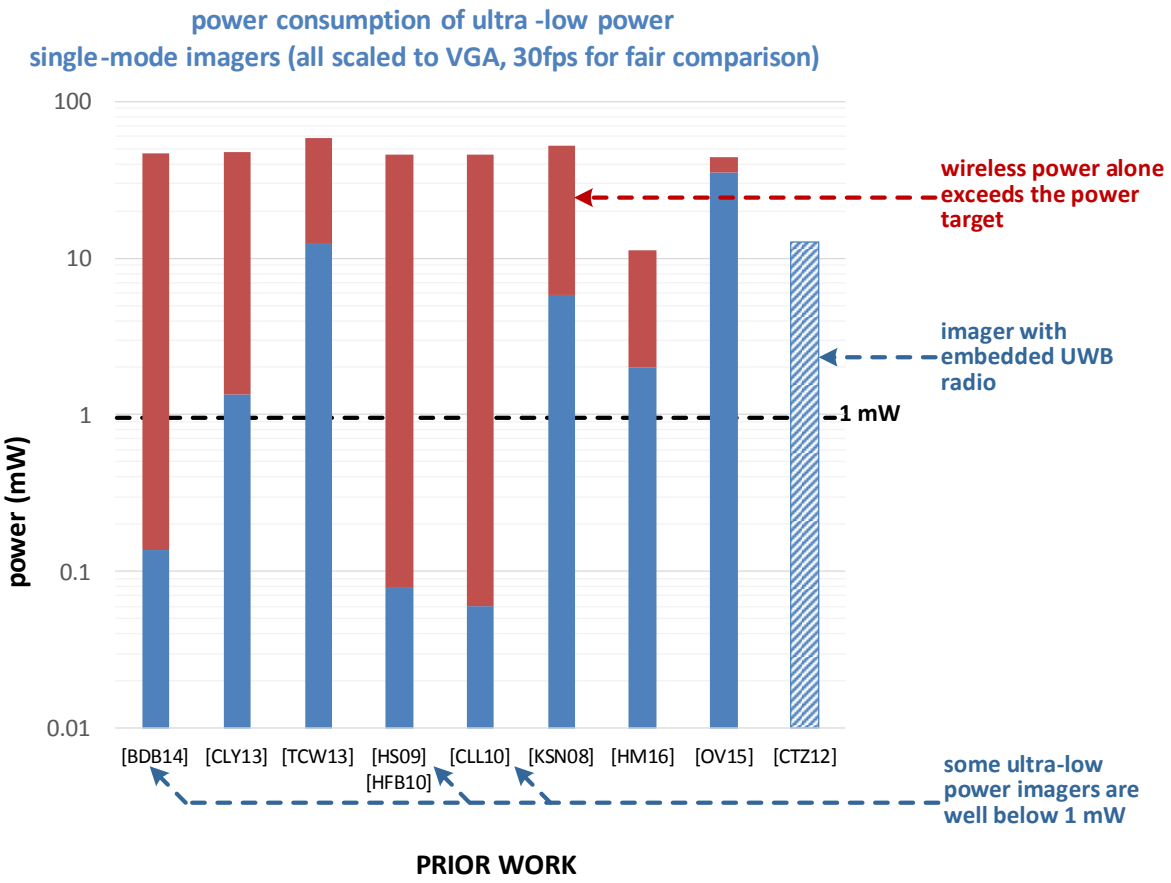


Fig. D5b. Power consumption of state-of-the-art ultra-low power imagers for always-on cameras and additional wireless power. As a result, the architecture #1 in Fig. 3 is unsuitable for sub-mW power budget.

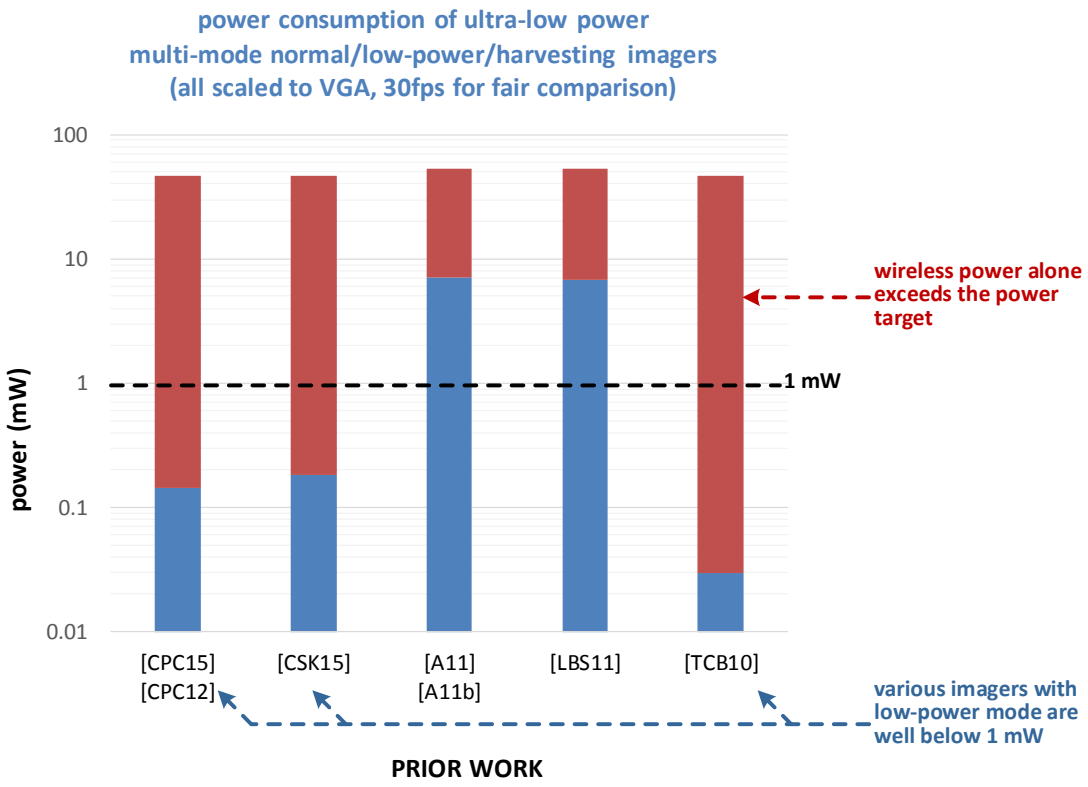


Fig. D5c. Power consumption of state-of-the-art imagers with ultra-low power multi-mode (e.g., high-accuracy mode activated only if illumination changes) and additional wireless power. Again, the architecture #1 in Fig. 3 is unsuitable for sub-mW power budget.

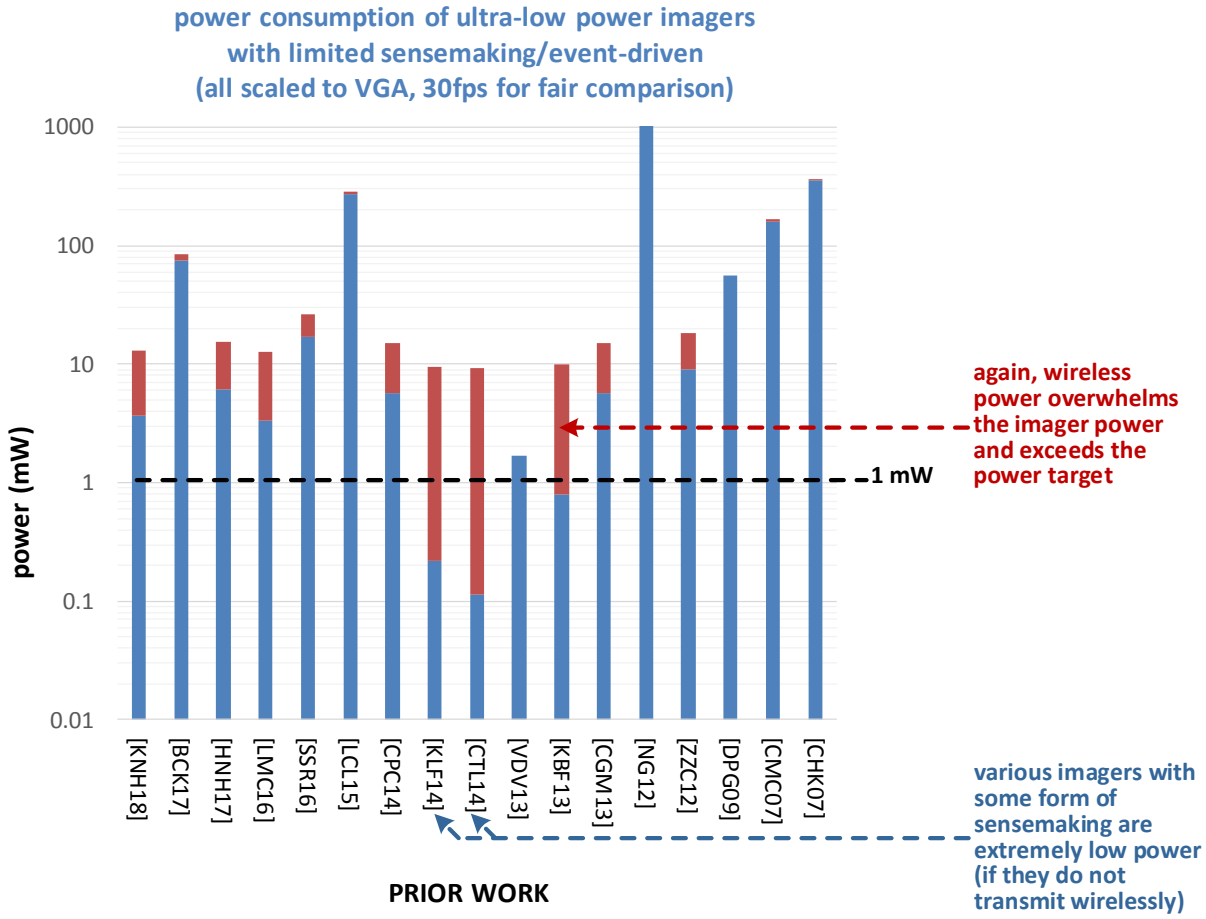


Fig. D5d. Power consumption of state-of-the-art imagers with some limited form of sensemaking (e.g., undetected motion inhibits image sensing and reduces power) and additional wireless power. Again, the architecture #1 in Fig. 3 is unsuitable for sub-mW power budget.

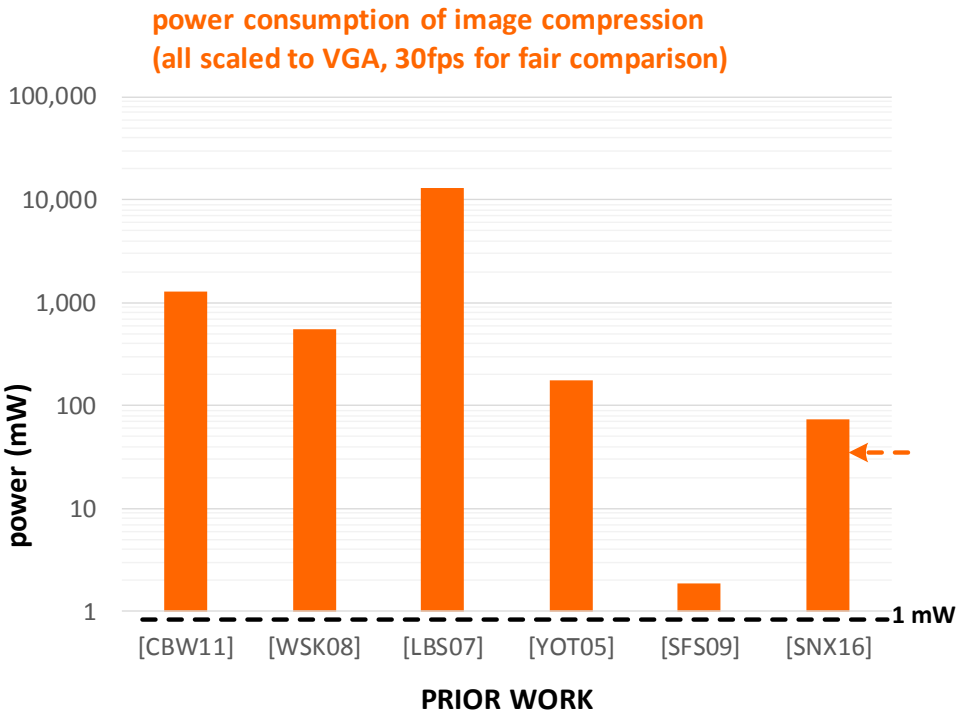


Fig. D6. The power consumption of state-of-the-art image compression accelerators (e.g., MPEG) alone exceeds the power target of untethered cameras. As a result, the architecture #2 in Fig. 3 is unsuitable for sub-mW power budget.

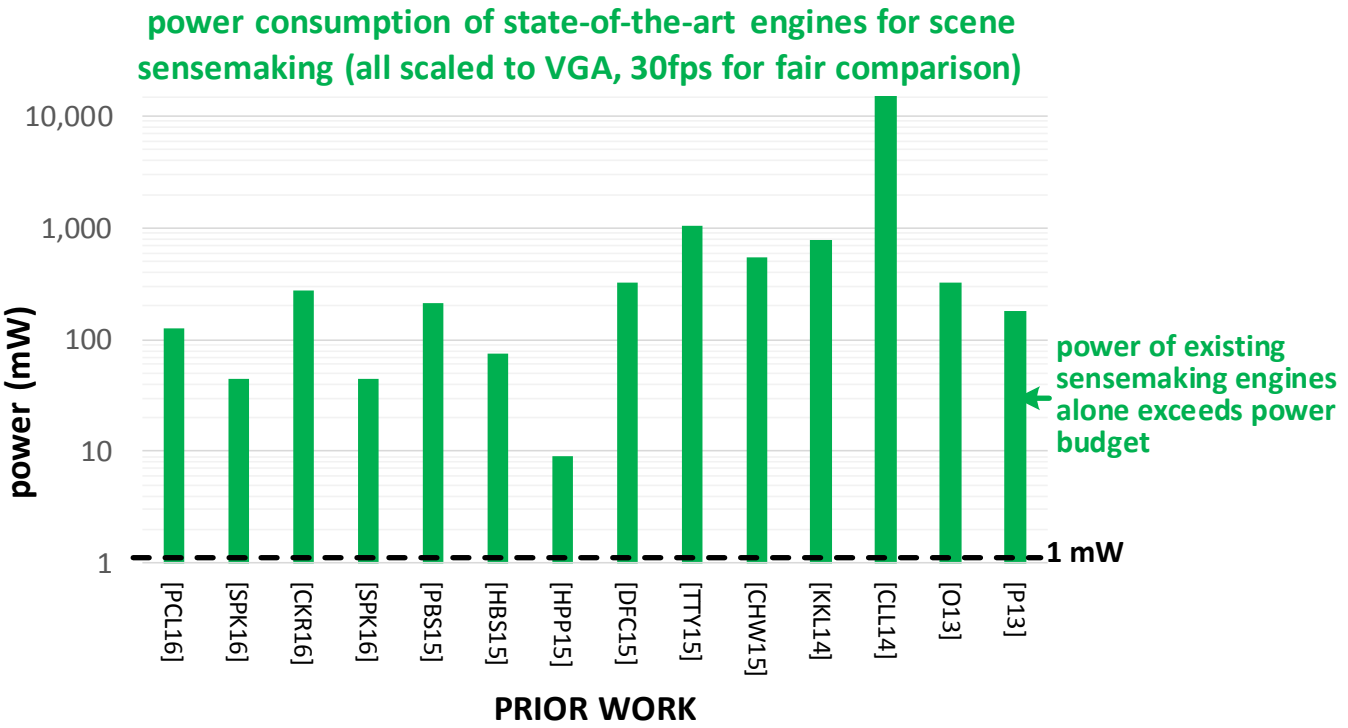


Fig. D7. The power consumption of state-of-the-art engines for sensemaking (e.g., deep learning, object recognition) alone exceeds the power target of untethered cameras. As a result, the architecture #3 in Fig. 3 based on existing stand-alone components is unsuitable for sub-mW power budget (i.e., system co-design is necessary to further reduce power).

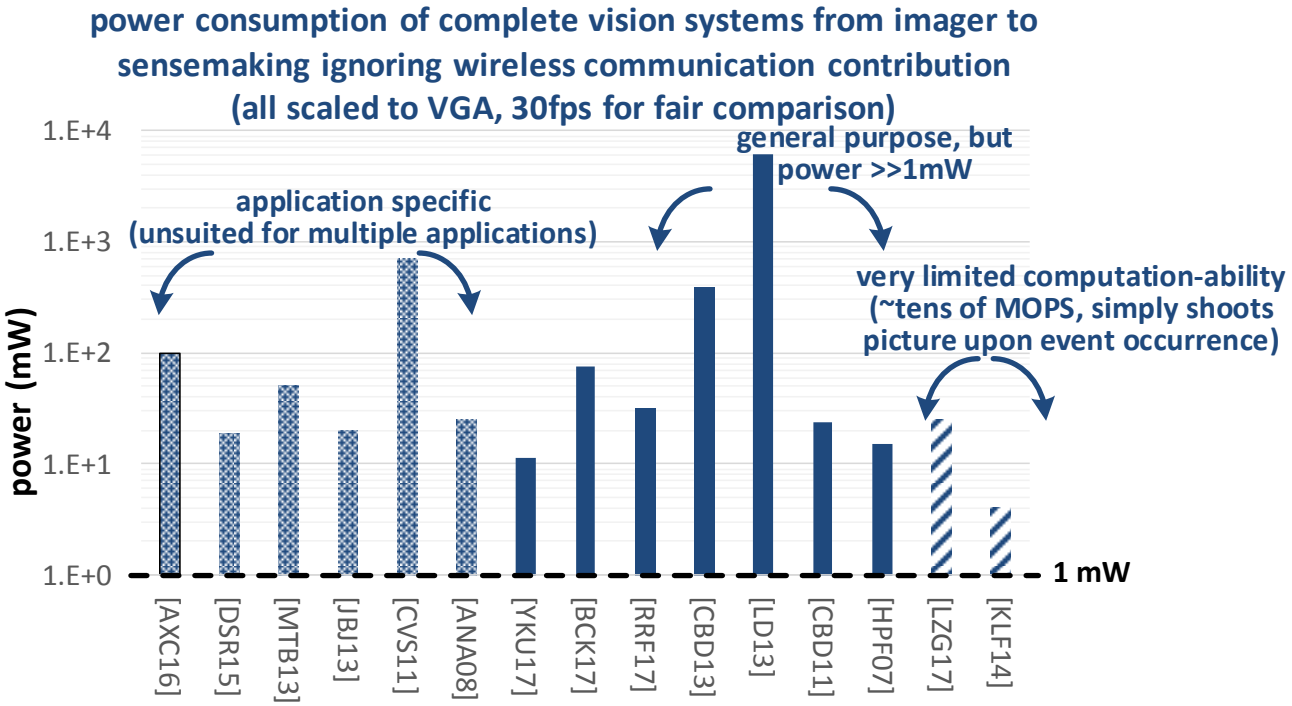


Fig. D8. Integrated research prototypes: power consumption in complete and vision systems is invariably beyond 10mW when fairly scaled at same VGA resolution and 30fps framerate. The demonstrations that are in the mW have very limited computation-ability (tens of MOPS, compared to the targeted 20,000MOPS), which only allows for shooting a picture upon the occurrence of simple events. CogniVision aims to fill this gap, allowing mW power while assuring suitability for a wide range of applications, as permitted by the reprogrammable deep learning accelerator and the adequate throughput to complete meaningful vision tasks at the targeted 30fps frame rate.

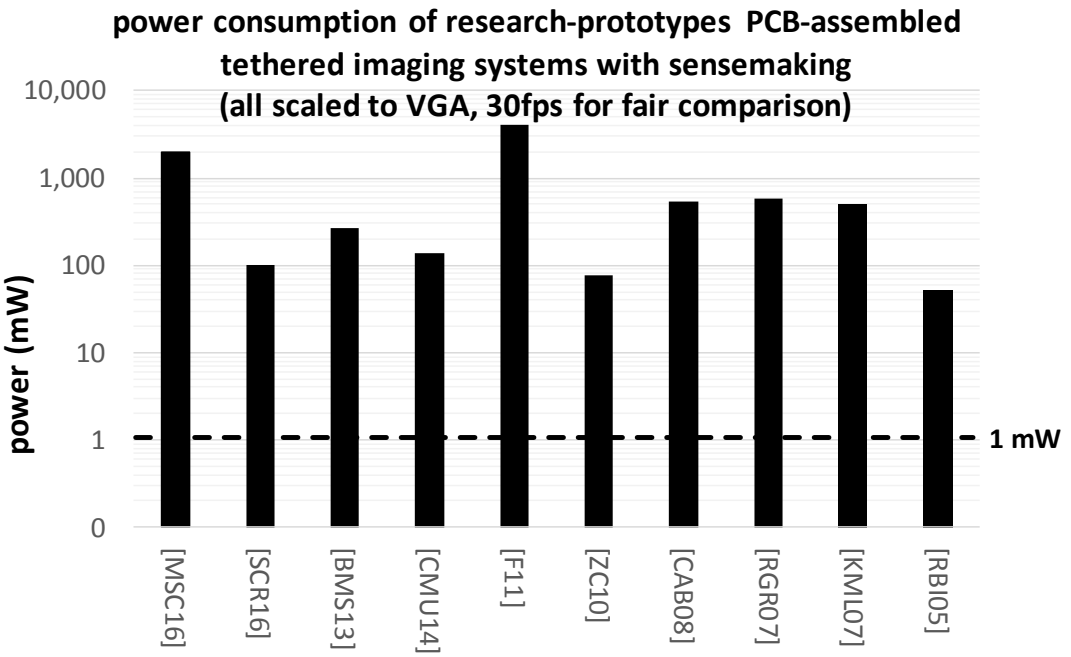


Fig. D9. PCB-assembled research prototypes: cameras in real conditions consume a power that is much larger than 1 mW, and hence unsuited for energy-autonomous cameras.

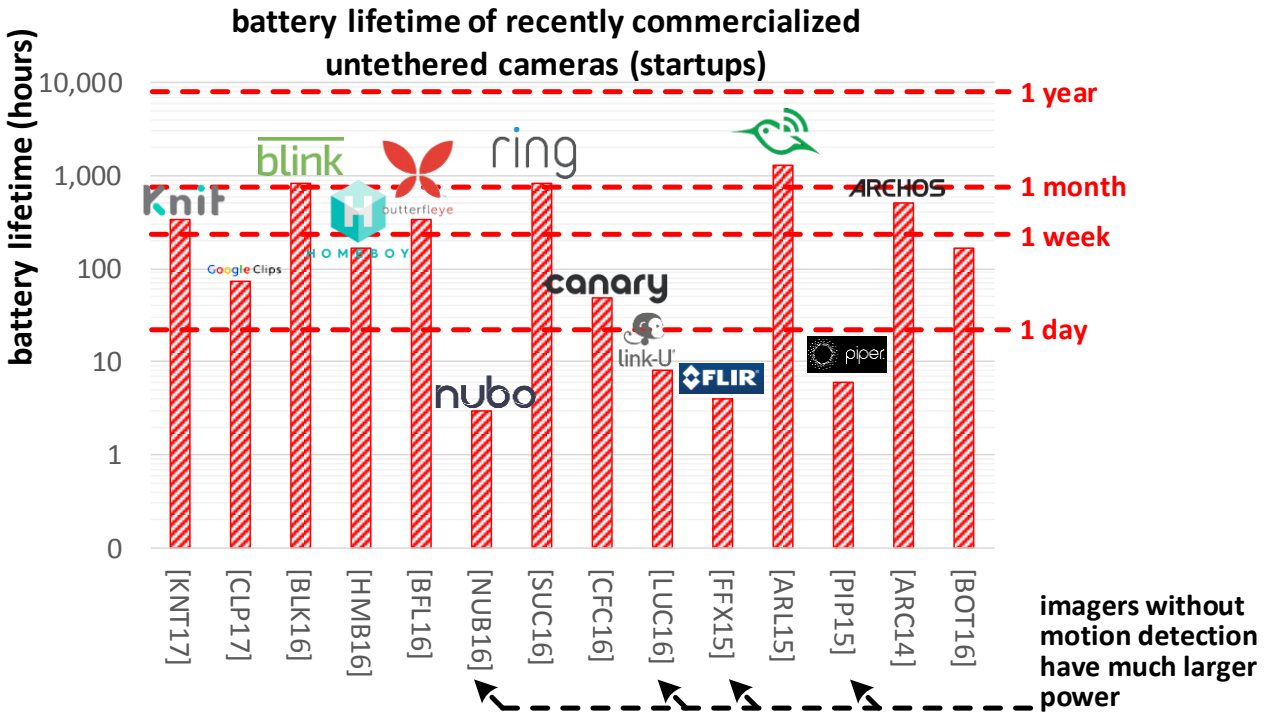


Fig. D10. Battery lifetime of untethered cameras (PCB-assembled). Most of them were released during the review of the white paper of this proposal, showing a very broad interest in untethered cameras. Their lifetime is reported as per their datasheet or based on Amazon users' reviews where available. Their lifetime is definitely inadequate for distributed sensing, and justifies the "CogniVision" research program, which aims to enable nearly-perpetual lifetime via energy harvesting in a small form factor (<<100mm^3).

Table II. *Leading researchers and their research work in areas affine to the "CogniVision" program (**prominent researchers highlighted in bold**).*

	institution / company	leading researchers	publication samples	research scope and limitations
ultra-low power imagers	University of Michigan, Ann Arbor (USA)	**E. Yoon**	[CPC15], [CPC14], [CPC12], [CHK07]	adaptive imagers with some embedded low-level intelligence, but ignores the fundamental problem of very large wireless power
	HKUST (Hong-Kong)	**A. Bermak**	[TCW13], [LBS11], [CBW11], [TCB10]	sub-mW power achieved only in imagers with extremely poor resolution (e.g., 128 × 96 pixels), inadequate in real applications
	University of Michigan, Ann Arbor (USA)	**D. Blaauw, D. Sylvester**	[KLF14], [KBF13], [HFB10], [HS09]	ultra-low power achieved only in imagers with very low resolution (e.g., 128 × 128 pixels) and frame rate (e.g., 0.5 fps), and very limited processing for event-triggered picture shooting (tens of MOPS), both inadequate in targeted applications
	NTU (Singapore)	**S. Chen, K.-S. Low, H. Zhuang**	[ZZC12]	ultra-low power achieved only in imagers with very low resolution (e.g., 64 × 64 pixels), sensemaking heavily constrained by the event-driven sensing framework (deep learning and state-of-the-art video processing algorithms cannot be applied)
	Samsung Advanced Institute of Technology (Korea)	**D.-S. Park**	[CSK15]	multi-mode imagers, but ignores the fundamental problem of very large wireless power
	University of Idaho (USA)	**S. U. Ay**	[A11], [A11b]	ultra-low power achieved only in imagers with extremely poor resolution (e.g., 50 × 50 pixels), ignores the fundamental problem of very large wireless power
	Purdue University (USA)	**E. Culurciello**	[CTZ12]	Includes ultra-low power radio, but power is still 10X larger than needed; ultra-low power achieved only in imagers with extremely poor resolution (e.g., 64 × 64 pixels), inadequate in targeted applications
	Johns Hopkins University (USA)	**R. Etienne-Cummings**	[CMC07]	ultra-low power achieved only in imagers with extremely poor resolution (e.g., 90 × 90 pixels), sensemaking heavily constrained by the event-driven sensing framework (deep learning and state-of-the-art video processing algorithms cannot be applied)
	FBK (Italy)	M. Gottardi	[GMJ09]	ultra-low power achieved only in imagers with extremely poor resolution (e.g., 128 × 64 pixels), sensemaking heavily constrained by the event-driven sensing

				framework (deep learning and state-of-the-art video processing algorithms cannot be applied)
	ARC-sr (Austria)	T. Delbruck	[LPD08]	ultra-low power achieved only in imagers with extremely poor resolution (e.g., 128 × 128 pixels), sensemaking heavily constrained by the event-driven sensing framework (deep learning and state-of-the-art video processing algorithms cannot be applied)
	UC Louvain (Belgium)	D. Bol, N. Couniot,	[BDB14]	focused on imager, ignores the fundamental problem of very large wireless power
	NTHU (Taiwan)	C.-C. Hsieh	[CLY13]	focused on imager, ignores the fundamental problem of very large wireless power
	Yonsei University (Korea)	J. Lee, G. Han	[CLL10]	focused on imager, ignores the fundamental problem of very large wireless power
	Nara Institute of Science and Technology (Japan)	M. Nunoshita, J. Ohta	[KSN08]	focused on imager, ignores the fundamental problem of very large wireless power, ultra-low power achieved only in imagers with extremely poor resolution (e.g., 128x96 pixels)
	Himax Technologies, Inc. (Taiwan)	N/A	[HM16]	ultra-low camera power only in environments with virtually no motion in the scene, unsuitable for public spaces (excessive power)
	OmniVision Technologies, Inc. (USA)	N/A	[OV15]	ultra-low camera power only in environments with low/steady lighting and no motion in the scene, unsuitable for public spaces (excessive power)
	Gdansk University of Technology (Poland)	R. Piotrowski	[JBJ13]	Low-power low-resolution imagers with on-chip low-level analog feature extraction, no mid/high-level sensemaking, no reprogrammability
	Columbia University (USA)		[G15], [NSF15]	very large (>10 cm), no intelligence, ignores the fundamental problem of very large wireless power
accelerators for scene sensemaking	KAIST (Korea)	**H.-J. Yoo, Lee-Sup Kim**	[SLL17], [PCL16], [SPK16], [PBS15], [HBS15], [O13], [P13], [WSK08], [KLK08], [LKK11],	focused on sensemaking only (no imager/cameras), high-accuracy at power consumption 100X larger than allowed in targeted applications

			[LOK10], [HPP15], [PBS15], [KKL14], [HBS15], [PCL16], [LKK08], [OLK09], [OPK11], [OKP12]	
	MIT	**V. Sze**	[CKR16]	focused on sensemaking only (no imager/cameras), power consumption 10X larger than allowed in targeted applications
	Toshiba (Japan)	H. Hayashi, T. Miyamori	[SPK16]	focused on sensemaking only (no imager/cameras), power consumption 100X larger than allowed in targeted applications
	NTU (Taiwan)	L.-G. Chen	[CHW15]	focused on sensemaking only (no imager/cameras), power consumption 100X larger than allowed in targeted applications
	ETHZ (Switzerland)	L. Benini	[RRL16], [LLR16], [PCR17]	efficient architectures for low-power triggering, processing/sensemaking not as efficient as best-in-class accelerators for deep learning; hierarchical processing is explored
	KULeuven (Belgium)	M. Verhelst	[MV16], [MV17]	general-purpose energy-efficient accelerators for deep learning with scalable precision (but no automatic quality control)
	STMicroelectronics (France)	N/A	[DCB17]	general-purpose energy-efficient accelerators for deep learning with scalable precision (but no automatic quality control)
	Stanford University	M. Horowitz, W. Dally	[HLM16]	general-purpose energy-efficient accelerators for deep learning for high performance, and energy efficiency not on par with best in class
	Hokkaido University	M. Motomura	[UAH18]	TSV-less 3D stacked deep learning acceleration for high-speed, highly-parallel systems
	INRIA	**O. Temam**	[DFC15], [CLL14], [CDS14], [LCL15]	focused on sensemaking only (no imager/cameras), power consumption >100X larger than allowed in targeted applications

imagers for mobile platforms (related, but different application domain)	TSMC (Taiwan)	C. Chao, F.-L. Hsueh	[LMC16]	focused on imager for mobile applications (large power, ignores the problem of very large wireless power)
	NHK Science & Technology Research Labs (Japan)	T. Hayashida, H. Shimamoto	[F15]	focused on imager for mobile applications (large power, ignores the problem of very large wireless power)
	SONY (Japan)	Y. Inada, H. Wakabayashi, T. Hirayama, N. Fukushima	[S15], [S13], [KNH18], [HNH17], [NSM18]	focused on imager for mobile applications (large power, ignores the problem of very large wireless power), and recently on 3D stacking
	Toshiba (Japan)	R. Okamoto, S. Kousai	[D13]	focused on imager for mobile applications (large power, ignores the problem of very large wireless power)
	Shizuoka University (Japan)	S. Kawahito	[S12]	resolution-scalable, but focused on imager for mobile applications (large power, ignores the problem of very large wireless power)
	Samsung Electronics (Korea)	C.-Y. Choi, G.-S. Han	[K12]	focused on imager for mobile applications (very large power)

Complete vision system from image sense to sensemaking	GeorgiaTech	J. Romberg, **A. Raychowdhury,** S. Mukhopadhyay	[XCR16], [AXC16], [DSR15]	photovoltaic cell-powered always-on camera with gesture recognition capability, 4-5 cm wide + 7 cm-wide solar cell, 100s mW power
	Université Blaise Pascal (France)	C. Bourrasset	[BMS13]	wired camera (no wireless communication, not energy autonomous), 6-7 cm wide
	Carnegie Mellon University	N/A	[CMU14]	wired camera (no wireless communication, not energy autonomous), 5.5 cm wide
	ETHZ (Switzerland)	**L. Benini**	[KML07], [MTB13], [RRF17]	focused on low-power trigger and hence on the low end of vision sensors; processing/sensemaking not as efficient as best-in-class accelerators for deep learning; hierarchical processing is explored
	KAIST (Korea)	H. J. Yoo	[BCK17], [MTB13], [BCK17]	low-end and application specific systems with limited computation-ability and no reprogrammability (e.g., fixed face recognition)
	Sony (Japan)	N/A	[YKU17]	3D stacked image sensor and processor for high-performance/high-speed imagine (unsuited for distributed vision)

	University of Manchester (UK)	P. Dudek	[CBD13], [LD13], [CBD11]	low-end and application specific systems with limited computation-ability and no reprogrammability (e.g., loiterer detection), or high-speed high-power smart imagers (unsuited for distributed vision)
	Fraunhofer Institute (Germany)	N/A	[F11]	untethered camera with compression, 8 cm long, very large power (4 W)
startups and large enterprises in 2015-2016	Blink	N/A	[BLK16]	no intelligence (only sends 10-s clips when motion is detected), low-power (3 mW) only in unrealistically fixed scene, very large power (25 mW) in public spaces and other realistic conditions, 7-cm wide
	HomeBoy	N/A	[HMB16]	no intelligence (only sends 30-s clips when motion is detected), 2-month operation in unrealistically fixed scene (much shorter in realistic conditions), 7-cm wide
	Butterfleye	N/A	[BFL16]	motion detector, limited intelligence to discard false events, sends (or records) up to 30ss clips when triggered, 2-week operation, 9-cm wide
	Google CLIPS	N/A	[CLP17]	limited intelligence to trigger video shooting upon event occurrence, but no interaction with cloud, no control on the type of events
	Knit Health	N/A	[KNT17]	limited intelligence to trigger video shooting upon motion detection, limited to recording (no interaction with cloud, no control on the type of events)
	Arlo (Netgear)	N/A	[ARL15]	motion detector, sends (or records) up to 30s clips when triggered (no intelligence), 3-6 month operation in unrealistically fixed scene (much shorter in realistic conditions), 7-cm wide

Table III. *Recent and on-going worldwide research programs on areas related to CogniVision.*

research program title	funding agency	year of completion	scope	limitations, differences
Reconfigurable Imaging (ReImagine) [REC16]	DARPA (USA)	N/A (~2022)	enabling software-reconfigurable imagers through highly-reconfigurable pixel architectures with distinct imaging modes in different regions of interest	focused on imager only, no sensemaking and vision, high-performance imagers (large power, not suited for untethered cameras)
Hercules (High-Performance Real-time Architectures for Low-Power Embedded Systems)	H2020 (EU)	2020	integrated framework for cutting-edge heterogeneous multi-core platforms for real-time computation	one of the two targeted applications is a visual recognition system for the avionic domain, focused on very high speed computation, not on complete and ultra-low power vision systems
MicroLearn: Micropower Deep Learning	Swiss National Foundation	N/A (~2020)	ultra-low power accelerators for deep learning	focused on deep learning accelerators, no integration of complete vision systems
Smart Cyber-Physical Systems	H2020 (EU)	2020	H2020-ICT-2014-1 smart cyber-physical systems (including vision)	focused on low-end sensing platforms, no focus on general-purpose ubiquitous deep learning accelerators
Visual Cortex on Silicon [NSF13]	National Science Foundation (USA)	2018	understanding the fundamental comprehension mechanisms used in the visual cortex	no camera demonstration, no chip demonstration, only sensemaking based on simulations and off-the-shelf components
Systems of Neuromorphic Adaptive Plastic Scalable Electronics (SyNAPSE) [SYN09]	DARPA (USA)	2017	new chip design mimicking brain's power-saving efficiency, with 100x less power for complex processing than state-of-the-art chips (spurred TrueNorth neuromorphic chip)	focused on large-scale compute-intensive sensemaking (cloud/datacentre level), no camera
COgnitive & Perceptive CAMeraS [COP13]	European Union	2016	ultra-low power computer architectures for cameras, based on many-core/GPU platform, focused on application-network-software-architecture interface	no imagers/cameras, no specialized hardware, very large power Watt range [COP13b], no chip demonstration (only FPGA prototyping)
Vision-in-Package [CSE15]	Swiss National Science Foundation (Switzerland)	2015	ultra-low power imaging system with imager and ARM Cortex M4, perform face detection, facial landmark tracking, person identification	wired camera (not energy autonomous, not ubiquitous), 1.85-cm wide, assembled on printed circuit board, 3–4 fps
IcyCAM [CSE15b]	CSEM (Switzerland)	2015	single-chip miniaturized camera	wired camera (not energy autonomous, not ubiquitous), imager and with general-purpose processor integrated on same silicon chip, but much larger power (80 mW at ¼ of VGA)

Supervised Autonomous Fires Technology (SAF-T) [SAF13]	Office of Naval Research (USA)	2015	visual processing algorithms and hardware/software platform for remote weapon stations (targeting, tracking and fire control)	focused on algorithms, no camera demonstration, no chip demonstration, only sensemaking based on simulations and off-the-shelf components
NeoVision2 [NEO09]	DARPA (USA)	2012	focused on neuroscience-inspired visual algorithms for detection, recognition, and tracking of many different classes of objects in live video imagery	focused on algorithms inspired by the design principles employed by mammalian vision systems, no camera demonstration, no chip/hardware demonstration

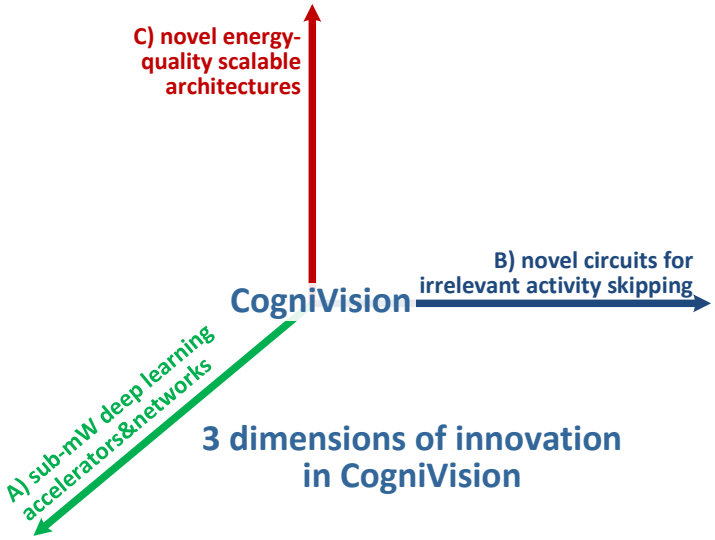
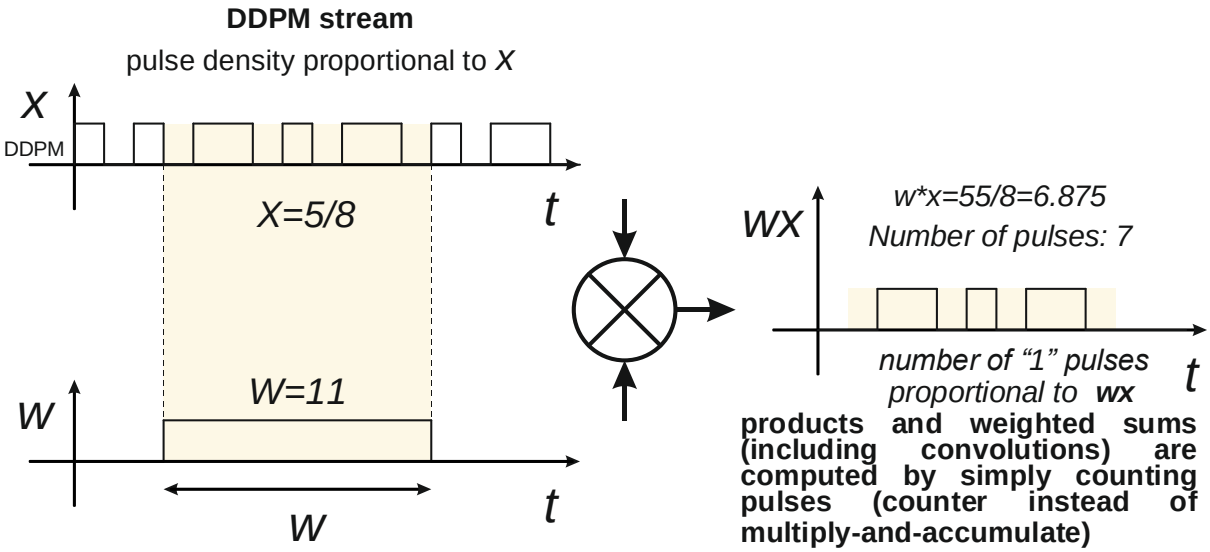
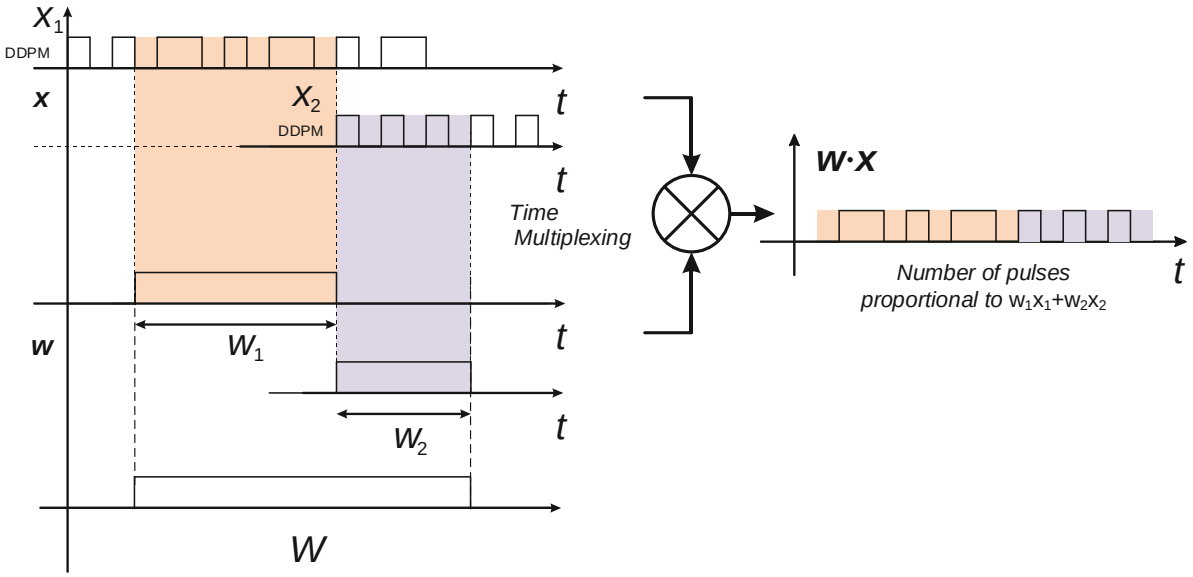


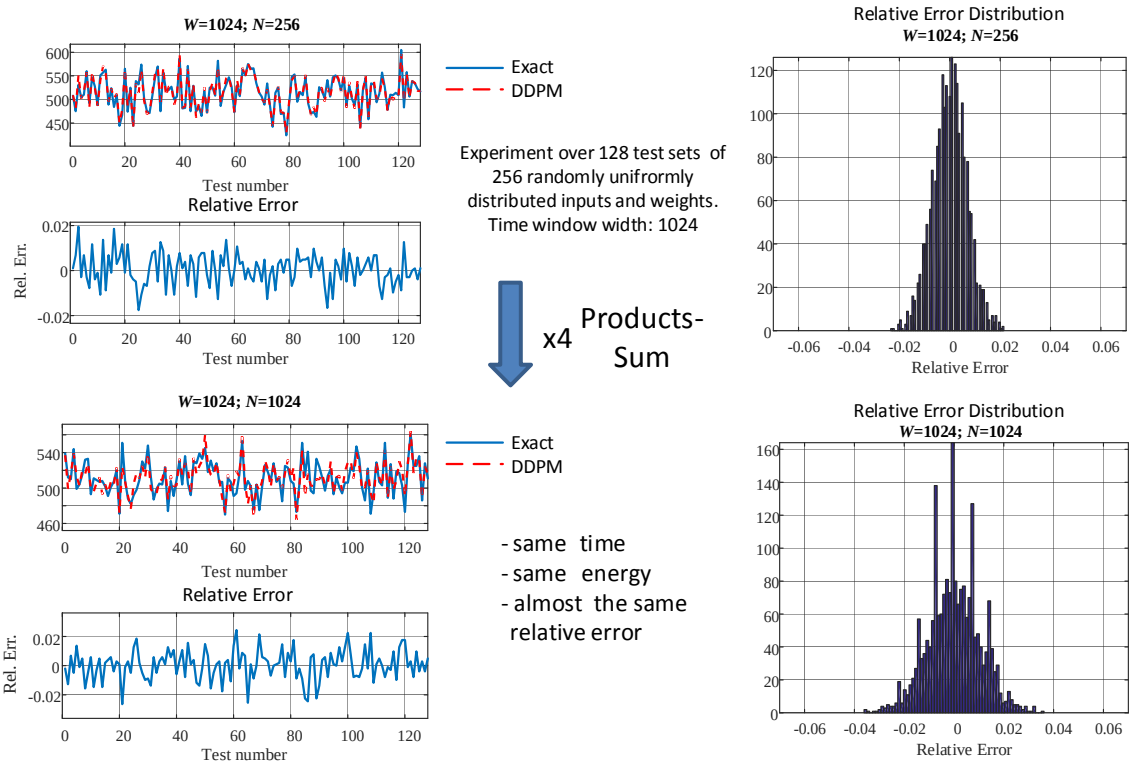
Fig. D11. The three dimensions of innovation in CogniVision.



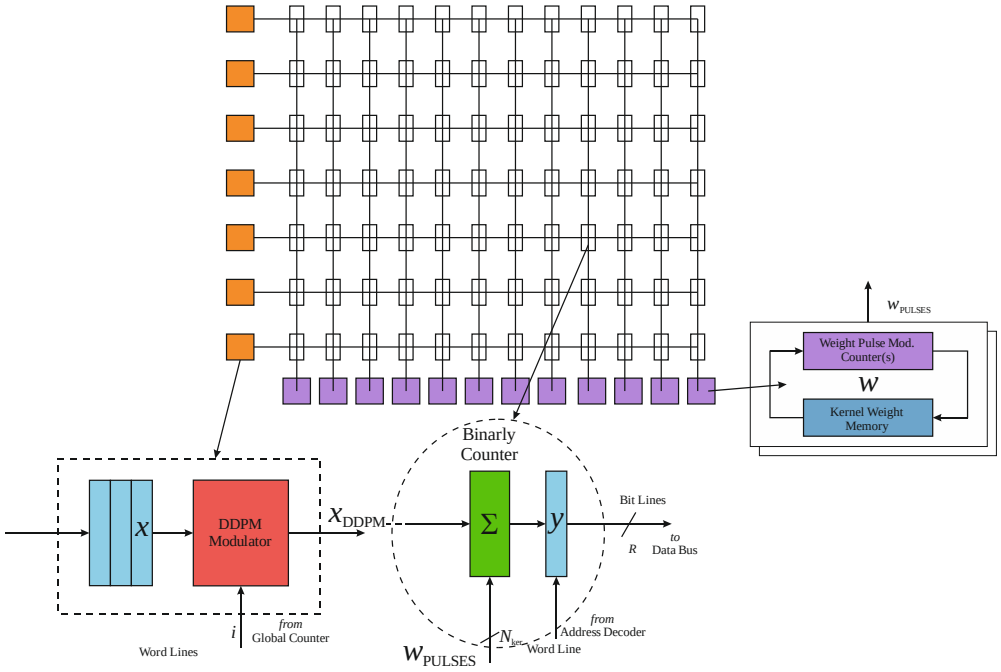
(a)



(b)

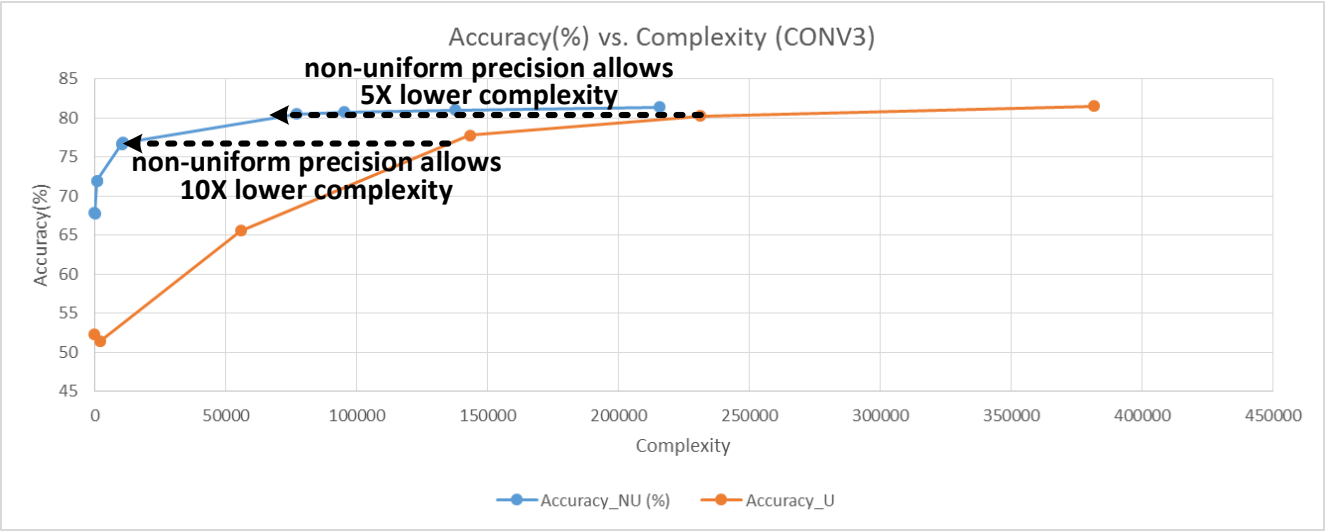


(c)

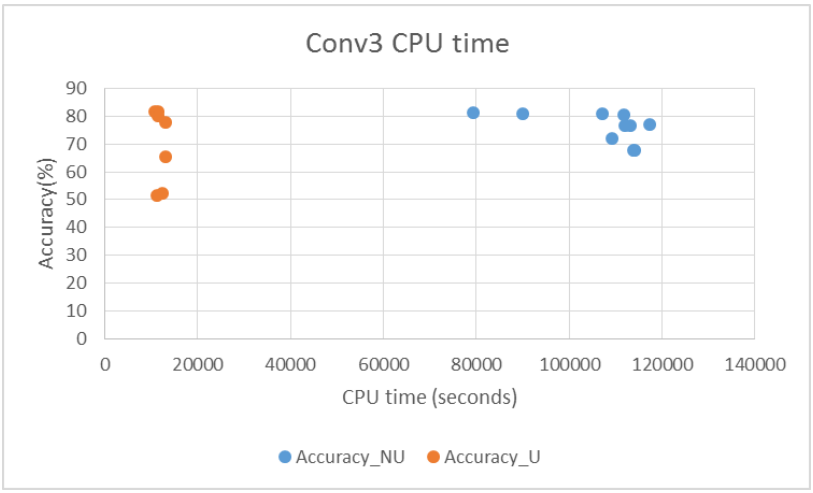


(d)

Fig. D12. a) General Dyadic Digital Pulse Modulation (DDPM) operation [C17]. Interestingly, the DDPM modulation can be effectively used to perform products, weighted sums (and hence convolutions for deep learning) with very low hardware cost, which consists of simple pulse counters (see on the right side of the figure). b) CogniVision leverages this fundamental and new observation to simplify each neuron into a counter, replacing the conventional energy-hungry method to compute convolution through multiply and accumulate.
c) Example of numerical simulation showing that the computational complexity and the relative accuracy are independent of the number of weighted products (i.e., complexity of the network), thanks to the DDPM approach. In this example, the results of 128 sets of N=256-terms weighted sums and N=1024-terms weighted sums are computed according to the proposed DDPM technique, and are compared with the results of the conventional computation, showing an error which is almost always less than 2% in both cases, independently of the number of weights. The targeted error can be easily reduced (1 additional bit of accuracy for doubled W) by increasing W, at the expectable cost of increased computation time.
d) Resulting DDPM architecture of deep learning accelerators (see preliminary results in relevant section with 50TOPS/W expected energy efficiency in 28nm CMOS technology).



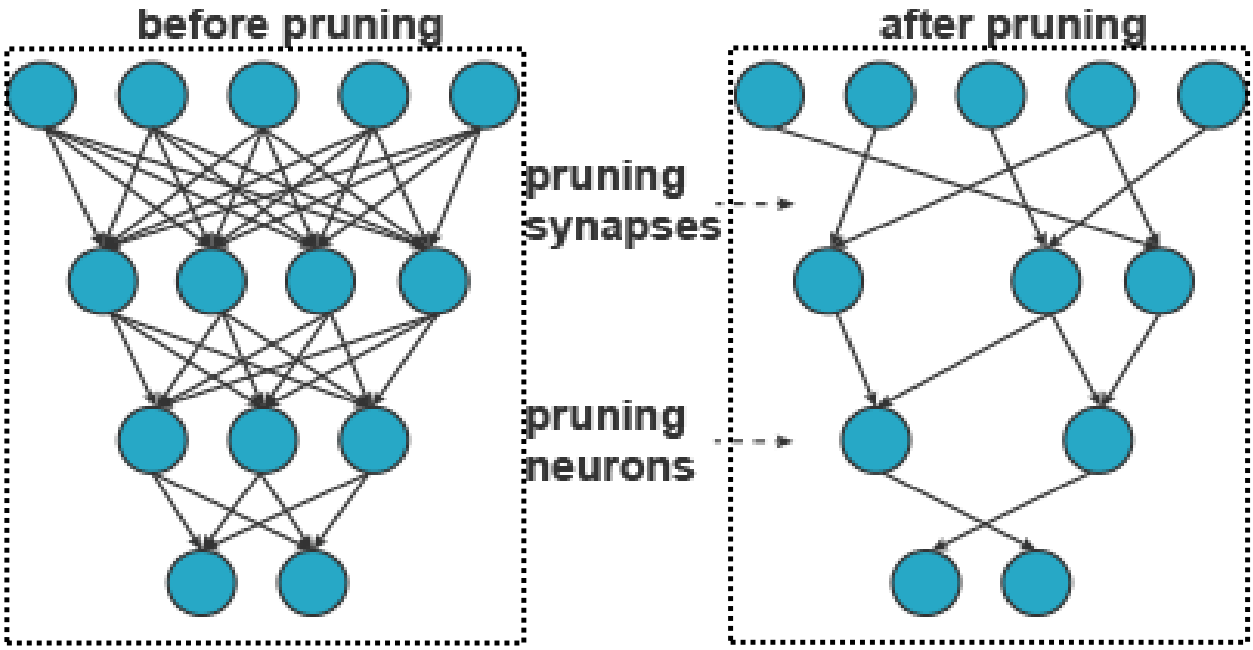
(a)



(b)

Fig. D13. a) Example with CIFAR10-trained neural network based on conventional uniform (U) and proposed non-uniform (NU) precision across neurons in convolutional layer #3 (results in other layers are equivalent or better). For a given accuracy over CIFAR-10 benchmark, non-uniform precision allows 5-10X reduction in complexity (i.e., overall number of computed bits, and hence gate count) compared to conventional uniform.

b) The penalty of non-uniform precision training is a 10X increase in the offline training time. This increase in offline training time is amortize across all devices performing inference. The increased offline training time can be dealt with by using commercially available cloud services (e.g., Amazon), which permit to temporarily scale up the server speed for training at larger cost. In other words, non-uniform precision allows a tradeoff between cost at training time (usually very small, in view of the large number of devices sharing the same network) and the complexity and power at inference time.



(a)

PROPOSED 2-PHASE NETWORK POWER-AWARE COMPRESSION APPROACH

Phase I: hard thresholding over connections and sub-network fine-tuning.

Apply hard thresholding over gradients magnitude calculated at each neuron to select the most informative ones (with large gradient magnitude). The hard thresholding preserves the top k neurons with the largest magnitude and disables the others by zeroing their parameters. Then, fine-tune the alive neurons to compensate the performance loss caused by the reduction in the number of filters. The loss function is calculated in a way specific to the application, and also combines both the accuracy and the power to achieve a desired balance between energy and quality (power-aware).

Phase II: neuron re-activation

The disabled neurons are re-activated and all the parameters are learned by training the entire network. The goal of this phase is to restore the truncated neurons and re-train the network to escape from some incorrectly compressed network models.

The above two phases are performed iteratively until there is no change over the neuron selection.

The final operation is the one in phase I to produce a compressed network. The proposal of such a gradient-based compression approach is based on the general intuition that the gradient magnitude passing through each neuron could reflect the "informativeness" of each neuron during the optimization process [Z16].

(b)

Fig. D14. a) Model pruning to remove redundant parameters and reduce the size of a deep learning model. In this example, both the connections between different layers of the model and redundant parameters (the neurons) are pruned based on the iterative hard thresholding method. As a result, more than 50% of the parameters (shown as the connections) are pruned.
b) Details on the proposed two-phase power-aware compression approach.

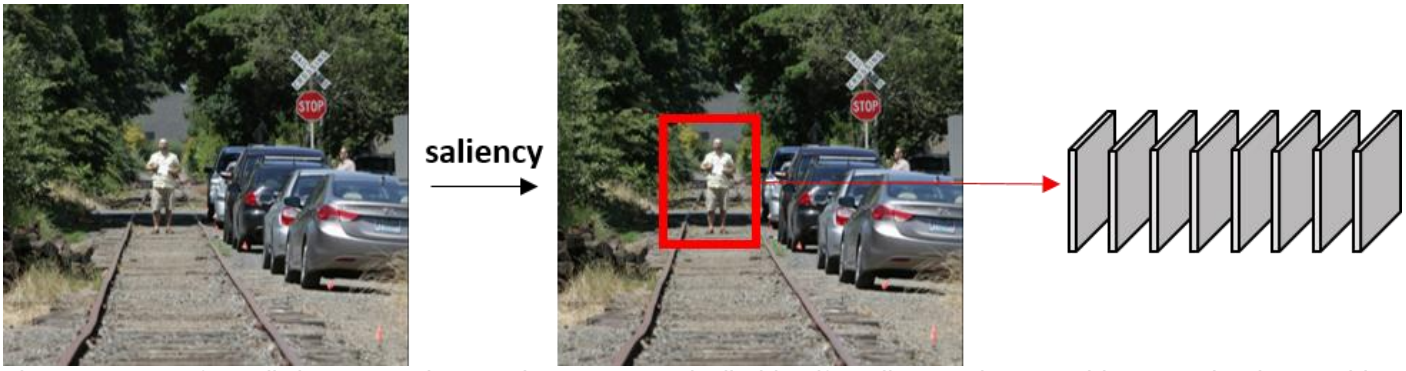


Fig. D15. Use of small deep neural networks to automatically identify salient regions. In this example, the machine learning circuit automatically focuses its attention on the person in the image (highlighted in red), discarding other irrelevant regions to avoid unnecessary computation (again, a form of irrelevant computation skipping).

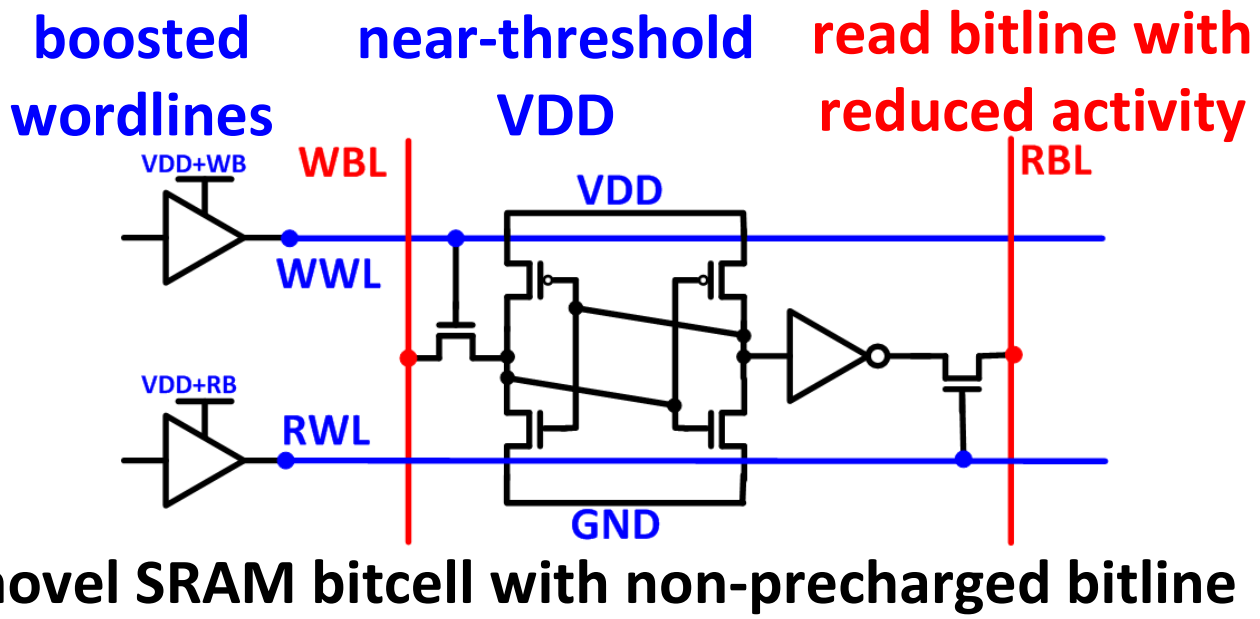


Fig. D16. Novel on-chip SRAM memory bitcell with unconventional non-precharged bitline for 70-80% reduced bitline activity (and 40% reduced power) to store features, pixels and weights. As opposed to existing 6T and 8T bitcells, the proposed bitcell is able to drive the read bitline to ground and to the supply voltage, thus avoiding the need for precharge and the resulting high bitline activity encountered in conventional pre-charged SRAMs.

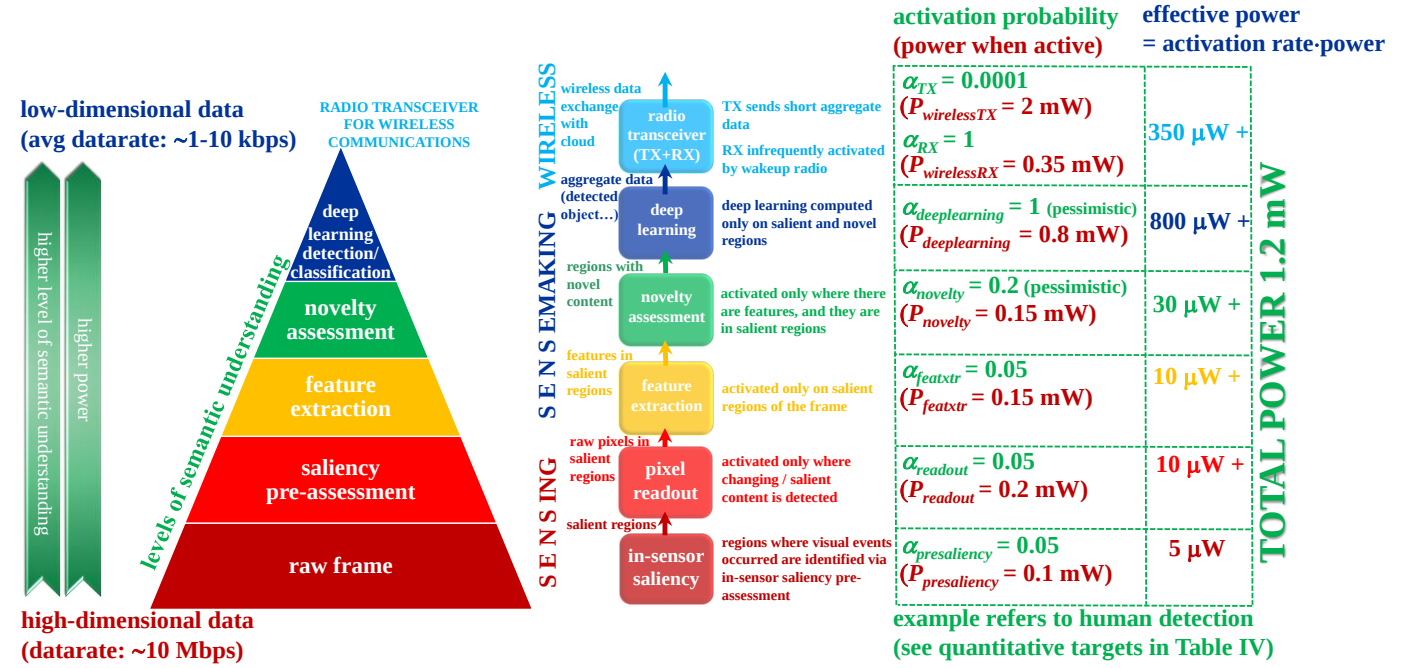


Fig. D17. In CogniVision, irrelevant activity is stopped at the lowest possible level of semantic understanding. The sooner it is stopped, the lower its power cost as higher levels of semantic understanding are associated with larger power. Every task has low activation rate (i.e., it is executed on a small fraction of the frame), reducing effective power by the same factor. The numerical example on the right refers to human detection in an indoor environment (maximum up to 20 humans in the field of view, 500-1,000lux light level), and uses preliminary deep learning logic-level simulations and detailed power calculations/estimates reported in Table IV.

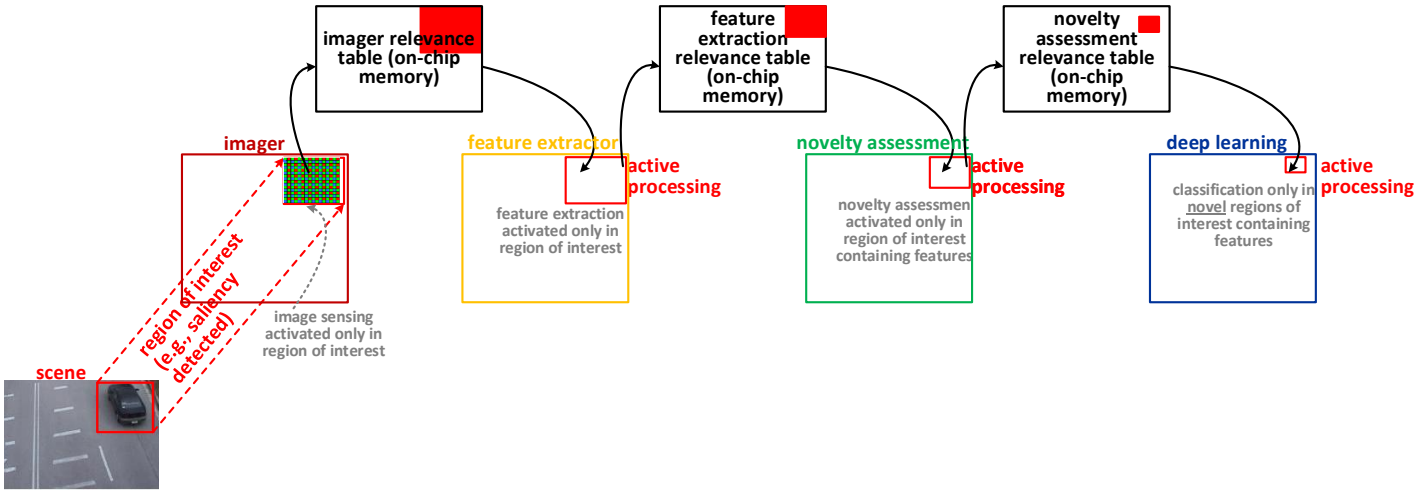
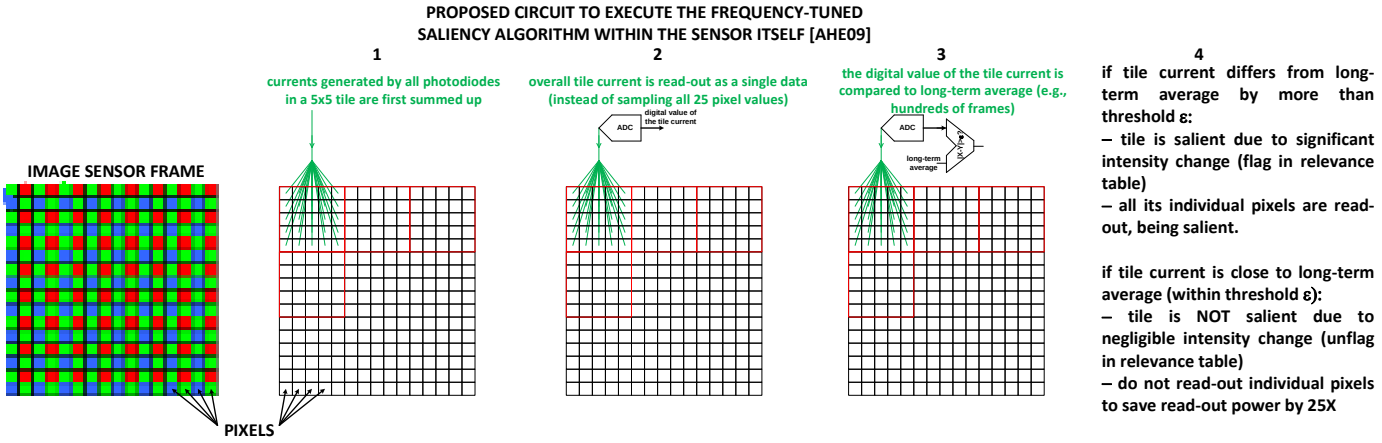
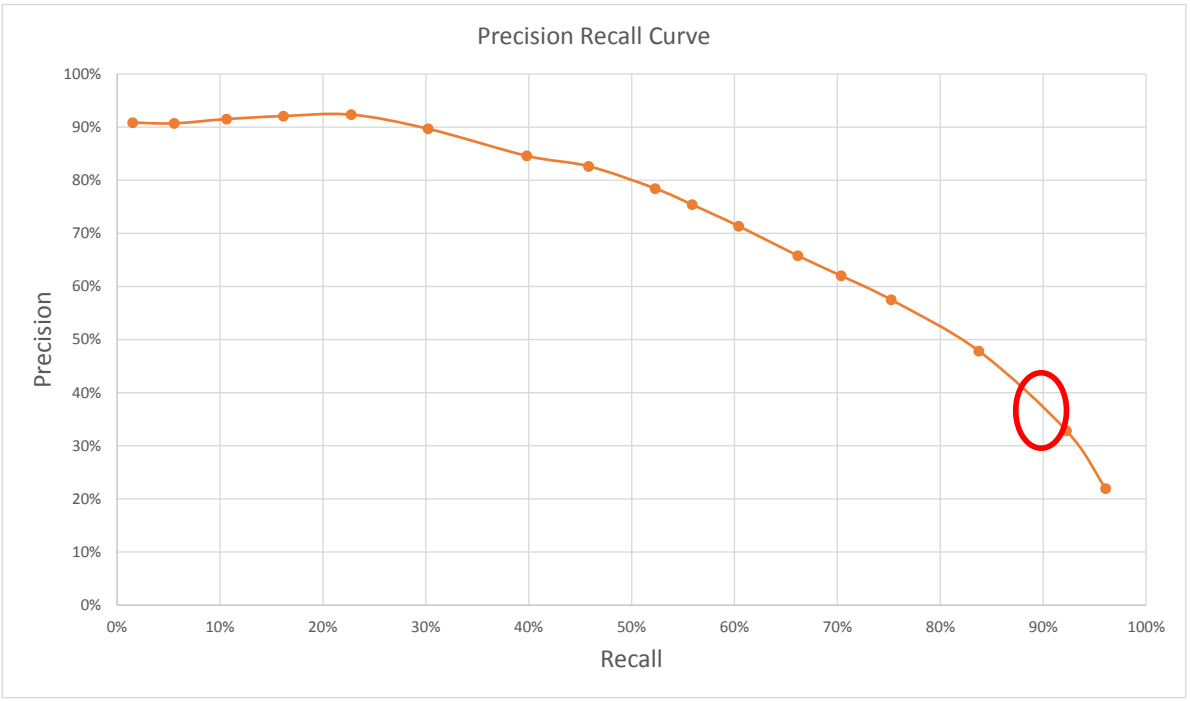


Fig. D18. In CogniVision, each sub-system in Fig. D14 (e.g., imager, feature extractor…) generates a small relevance table (e.g., few kb at most), where the frame portions/tiles where relevant activity is taking place. The output of the relevance table is taken up by the next sub-system (e.g., feature extractor after imager) to skip computation that pertains to irrelevant regions (i.e., where the bits in the relevance table are tagged as irrelevant, which are left blank in this figure).
This mechanism involves all sub-systems to avoid the waste of power observed in conventional vision systems on a chip that re-compute the entire frame every time a single event occurs (e.g., appreciable motion in a pixel).



(a)



(b)

Fig. D19. a) Circuit principle of the proposed in-sensor saliency detector: if the overall 5x5 pixel tile current changes significantly, it means that the intensity in the tile has changed appreciably, hence the tile is salient. In this case, the imager relevance table in Fig. D18 is updated, flagging the corresponding tile as relevant (i.e., salient). All individual pixels in the tile are read-out normally.

If the overall 5x5 pixel tile current is similar to its long-term average, no appreciable change is detected and the tile is non-salient. In this case, pixels do not need to be read out, thus reducing number of read-outs and imager power by 25X.

b) Numerical analysis of in-sensor saliency detector through benchmark in []. The precision vs recall plot for various values of the threshold ε in Fig. D19a shows that lower thresholds improve Recall (higher), at the cost of worse precision (lower). To avoid skipping potentially salient regions, Recall is more important than Precision and hence needs to be favored.

The point highlight in red (ε=0.02) is an example of reasonable tradeoff, where Recall is quite high (92%), and Precision is fairly low (33%), but still reasonable in terms of impact on power. Indeed, the resulting increase in false positives (i.e., activity of feature extractor) has minor impact on the overall power saving, since only 2-3% of tiles turn out to be salient anyway (i.e., activity and power are drastically reduced in spite of the presence of false positive

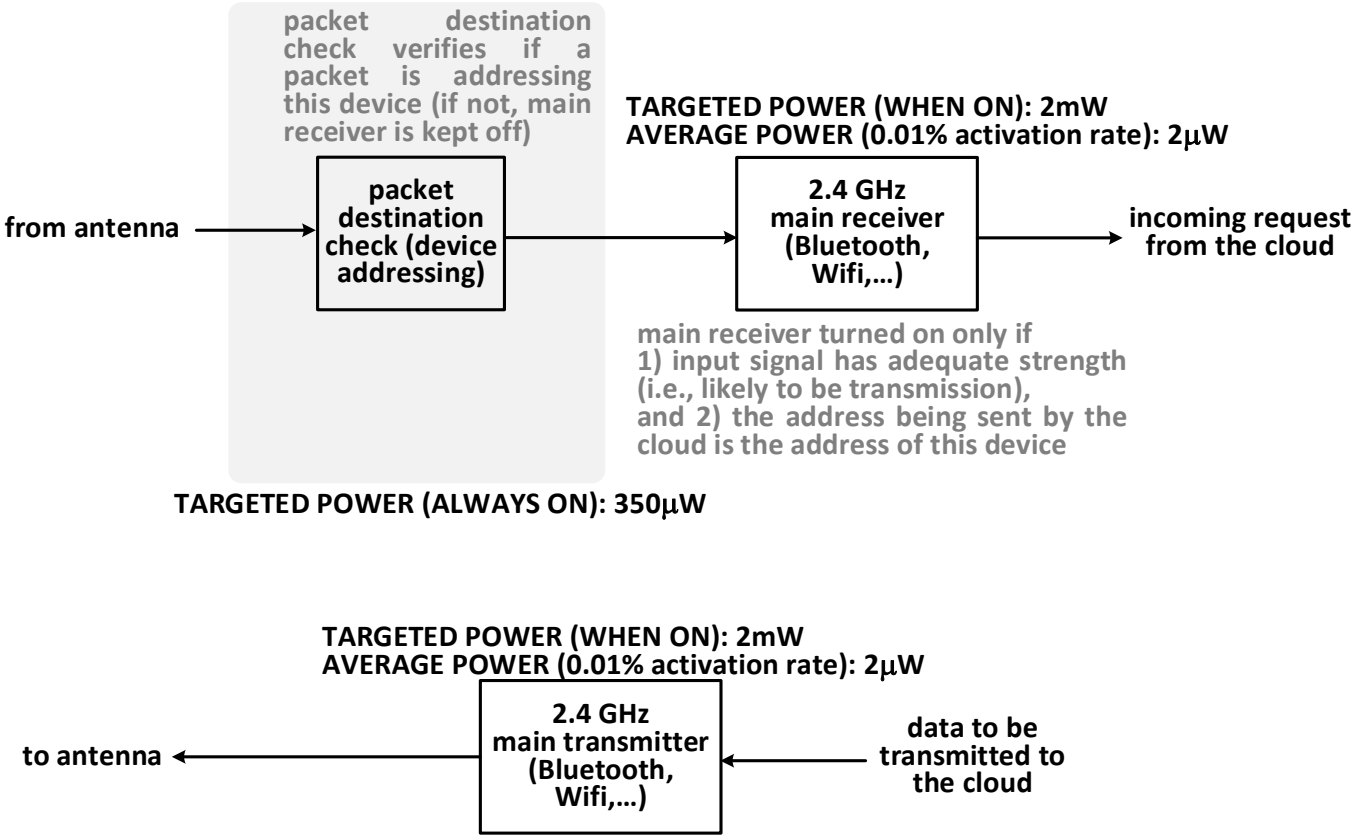


Fig. D20. Architecture of always-on receiver at the ISM band of 2.4GHz.

The receiver power in the always-on part is estimated to be 300-350μW from preliminary simulations in 180nm CMOS. The receiver and the transmitter are expected to consume 2mW when ON, but their infrequent activation reduces their average power by two orders of magnitude (i.e., few uWs), under realistic activation rates in the order of 0.01% (i.e., communication between cloud and camera occurs every 10,000 frames, or equivalently every 33 seconds - or longer - at 30frames/s).

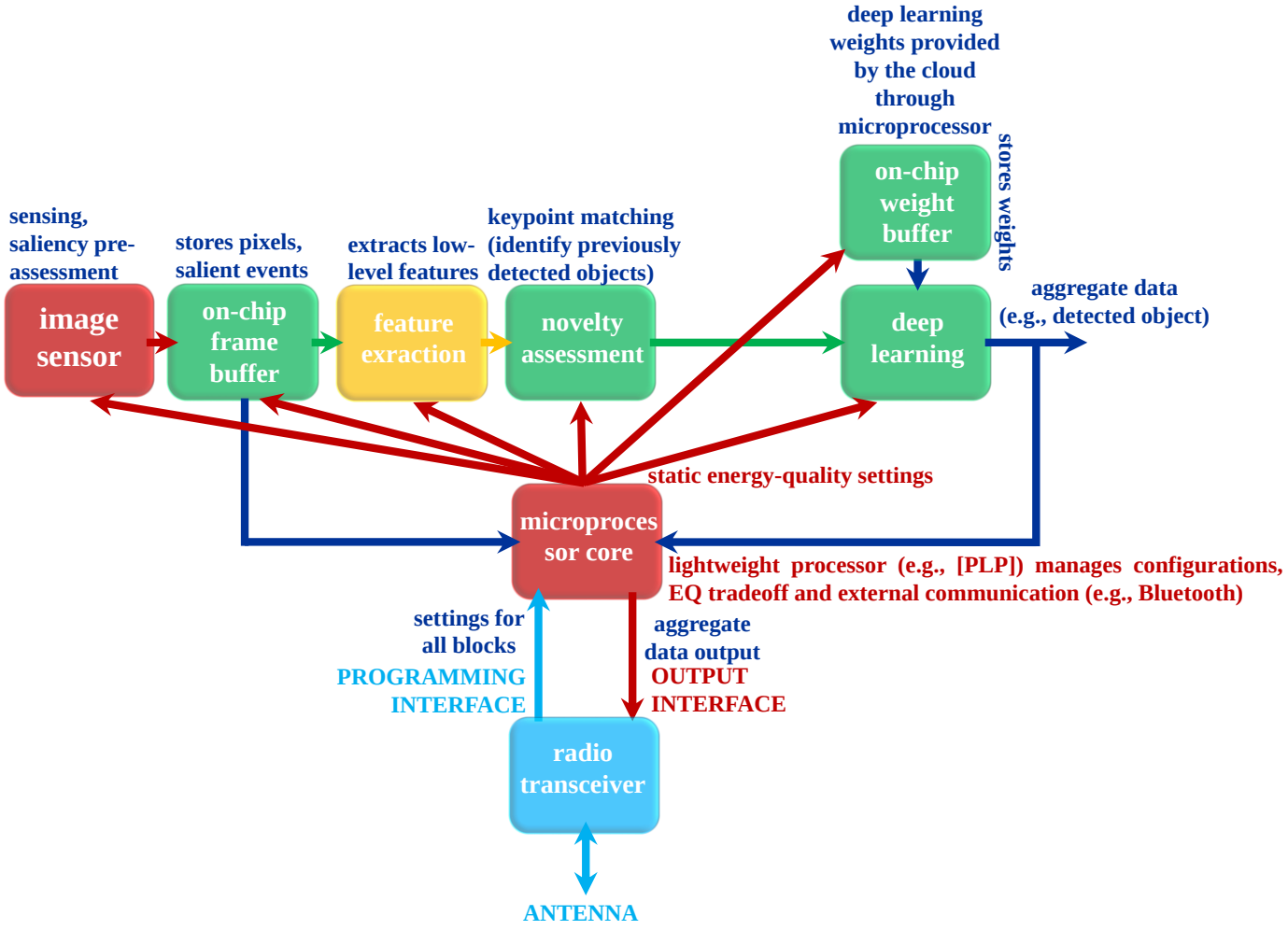


Fig. D21. In-principle architecture of CogniVision. The System on Chip communicates with the external world through a radio transceiver, which is connected to the low-performance microprocessor managing the chip settings via a) a programming interface that provides the settings (including the weights for deep learning) as per the cloud's requests, b) an output interface for wireless transmission (e.g., ZigBee).

Fig. D22. Gantt chart: project launch, integration, exploration & demonstration, energy-centric techniques

(Mx.y = milestone y in sub-project x; Dx.y = deliverable y in sub-project x)

		Year 1				Year 2				Year 3				Year 4				Year 5			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Project launch - phase 0	**Hiring, procurement and collaborative SW environment setup**		M0.1 (D0.1)																		
	0.1 Recruitment of majority of the manpower																				
	0.2 Requisition of major equipment essential for the Programme																				
	0.3 Setup of collaborative SW environment																				
Sub-project 1	**System modeling, exploration, integration, demonstration**																				
	1.1 System simulation/modeling framework								D1.1												
	1.2 System on board (SoB) integration/characterization (round #1)												M1.2								
	1.3 System on board (SoB) integration/characterization (round #2)																M1.3				
	1.4 System on chip (SoC) partitioning, chip level simulation environment																M1.4				
	1.5 System on chip (SoC) optimization and integration																		D1.5		
	1.6 SoC characterization and in-field validation																				M1.6
Sub-project 2	**Energy-centric circuit techniques and interaction at imager-sensemaking and wireless-sensemaking boundary**																				
	2.1 Imager and transceiver architectural exploration				D2.1																
	2.2 Imager and transceiver tapeout and testing (round #1)								D2.2				M2.2								
	2.3 Imager and transceiver tapeout and testing (round #2)												D2.3				M2.3				
	2.4 Imager and transceiver final revision for SoC, silicon demonstration																		M2.4		

	2.5 Final characterization and validation					M2.5
Sub-project 3	**Energy-centric machine learning-circuit co-design**					
	3.1 Deep learning model compression	M3.1a	M3.1b	D3.1		
	3.2 Energy-aware deep learning network design and training			D3.2		
	3.3 Saliency model			M3.3		
	3.4 In-field model fine-tuning, validation and integration					M3.4
Sub-project 4	**Irrelevant activity skipping/EQ-scalable sensemaking circuits/architectures**					
	4.1 Activity skipping architectures/circuits			M4.1		
	4.2 EQ-scalable architectures/circuits			M4.2		
	4.3 Feature extraction, novelty assessment, deep learning, SRAM tapeout and testing (round #1)		D4.3	M4.3		
	4.4 Feature extraction, novelty assessment, deep learning, SRAM tapeout and testing (round #2)			D4.4	M4.4	
	4.5 Final characterization and validation					M4.4
Project control – task 5	**Project control and reviews**					
	5.1 Internal review meetings with Advisory Board	M5.1	M5.2	M5.3	M5.4	M5.5
	5.2 Mid-term review			M5.6		
	5.3 Final review					M5.7

Table IV. Detailed targets for the final demonstration and measure of the success of the project in three visual tasks (ImageNet classification, human detection and object detection). Detailed operating conditions, dataset, neural network targets and chip performance targets are provided for each of them.

task	1) ImageNet image classification (0.5MobileNet network [MBN17])	2) human detection* (detect and localize the presence of persons within a frame)	3) object detection* (detect and localize objects of a specific category in a frame)
testing dataset	public Imagenet database [ILSVRC]	live scenes captured in EA lobby @ NUS (additional scenes from public space in Singapore, subject to approval)	live scenes captured in EA lobby @ NUS (additional scenes from public space in Singapore, subject to approval)
operating condition**	500-1,000lux light level, wall-projected ImageNet samples	500-1,000lux light level, up to 20 humans in the field of view	500-1,000lux light level, up to 10 objects in the field of view
adopted network	standard MobileNet [MBN17]	structure similar to AlexNet (5CONV+3FC), retrained for human detection via innovative compression techniques	structure similar to AlexNet (5CONV+3FC), retrained for object detection via innovative compression techniques
accuracy target	60% (reference: 57.2% in AlexNet)	detect 85% of 300 persons in a single frame	80% over 10 categories (person, car, chair, dog, bicycle, bird, bus, table,motorbike,monitor)
throughput target	30fps, projected images with 256x256 resolution (ImageNet benchmark)	30fps VGA resolution	30fps VGA resolution
model size (weight memory)***	1.3E6 (1.3 MB after innovative network compression and weight binarization)	6E6 (0.75 MB after innovative network compression and weight binarization)	6E6 (0.75 MB after innovative network compression and weight binarization)
# operations/frame****	76E6	114E6	114E6
targeted throughput @ 30fps (ops/frame * framerate)	2,280MOPS	3,420MOPS	3,420MOPS
targeted CogniVision power*****	1mW (dominant contribution: 0.56mW deep learning accelerator)	1.2mW (dominant contribution: 0.8mW deep learning accelerator)	1.2mW (dominant contribution: 0.8mW deep learning accelerator)

* Detection is here performed on a frame basis (no tracking). Occlusion is not dealt with in these demonstrations, as no elegant solution has been found in the preliminary exploration we have performed in this area (due to the complexity of the task). If strictly needed, occlusion can be addressed in the cloud by occasionally having the cognitive camera send all the keypoints for frames where there is activity, and have the cloud deal with occlusion. Another possible approach is to generate a deep network that is able to perform this task within the capabilities of the CogniVision system on chip (i.e., MB-range weight memory, 20,000MOPS computational throughput).
** Range of conditions that have been used in deep learning simulations to estimate the achievable accuracy in preliminary exploration (Caffe framework [BKL]), same as target conditions at CogniVision deployment
*** Weight memory evaluated after training and compressing the AlexNet network for the accuracy target in the table (see proposal for the details of the techniques introduced to reduce the model size)
**** Number of operations (additions, multiplications, comparisons) per frame evaluated from the actual structure of the compressed network in the table, then scaled to VGA by realistically assuming a complexity increase (i.e., neurons, number of computations) by 12X compared to AlexNet at its 256x256 resolution (12X was evaluated by retraining the network with the same structure for VGA resolution).
***** Power of deep learning accelerator is obtained as TOPS/(TOPS/W) where TOPS=1,000,000 MOPS is indicated in the table, and the energy efficiency TOPS/W = 50 from logic simulations of the DDPM accelerator in 28nm CMOS. Dominant wireless power is dictated by the receiver, and is 350μW from the preliminary results discussed in the text.

Estimates in this table are generated under the following **assumptions**:
- the popular FOM of the imager is 10pJ/pixel (in line with reasonably good imagers with similar pixel size of 5μm and technology, although not best-in-class as this FOM is not critical to the overall power as shown in the example in Fig. D17)
- the energy/pixel of the feature extractor is estimated to be 22pJ/pixel in in 28nm CMOS (i.e., only 2X lower than recent silicon demonstration from our team [APA17], which is pessimistic compared to the preliminary simulation results obtained with the new feature extractor architecture that will be explored in the project). Such pessimistic assumption will not impact the overall power estimate significantly, as the dominant contributions come from the deep learning accelerator and the radio transceiver
- the energy/frame in novelty assessment is equal to the energy in the feature extractor (estimated to be comparable from high-level simulations)
- the deep learning accelerator has an energy efficiency of 50TOPS/W, as found from post-synthesis logic simulations of a preliminary Verilog description of a small-scale DDPM accelerator (16x8 neurons) in 28nm CMOS
- memory energy per access is 30fJ/bit, in line with circuit simulations of an SRAM in 28nm CMOS
- transmitted wireless power is assumed to be 2mW (reduced to 2μW by the realistic activation rate of 0.01%, which corresponds to one transmission every 10,000 frames, or equivalently 33s)
- pixel activation probability in pre-saliency assessment is 5% (pessimistic, as it can be as low as 3.5% depending on the specific video, using the benchmark in [CDT12])
- novelty assessment identifies 20% features as novel on average (pessimistic, this has been observed to be down to 5% through the benchmark in [CDT12])
- no energy saving from irrelevant activity skipping is pessimistically being considered in the deep learning accelerator, as this seems to be dependent on the network from a preliminary analysis. Deeper analysis will be carried out during the execution of the project.
The above assumptions immediately lead to the numerical results in Fig. D17, by simply multiplying each power contribution by the corresponding activation rate (see above assumptions).

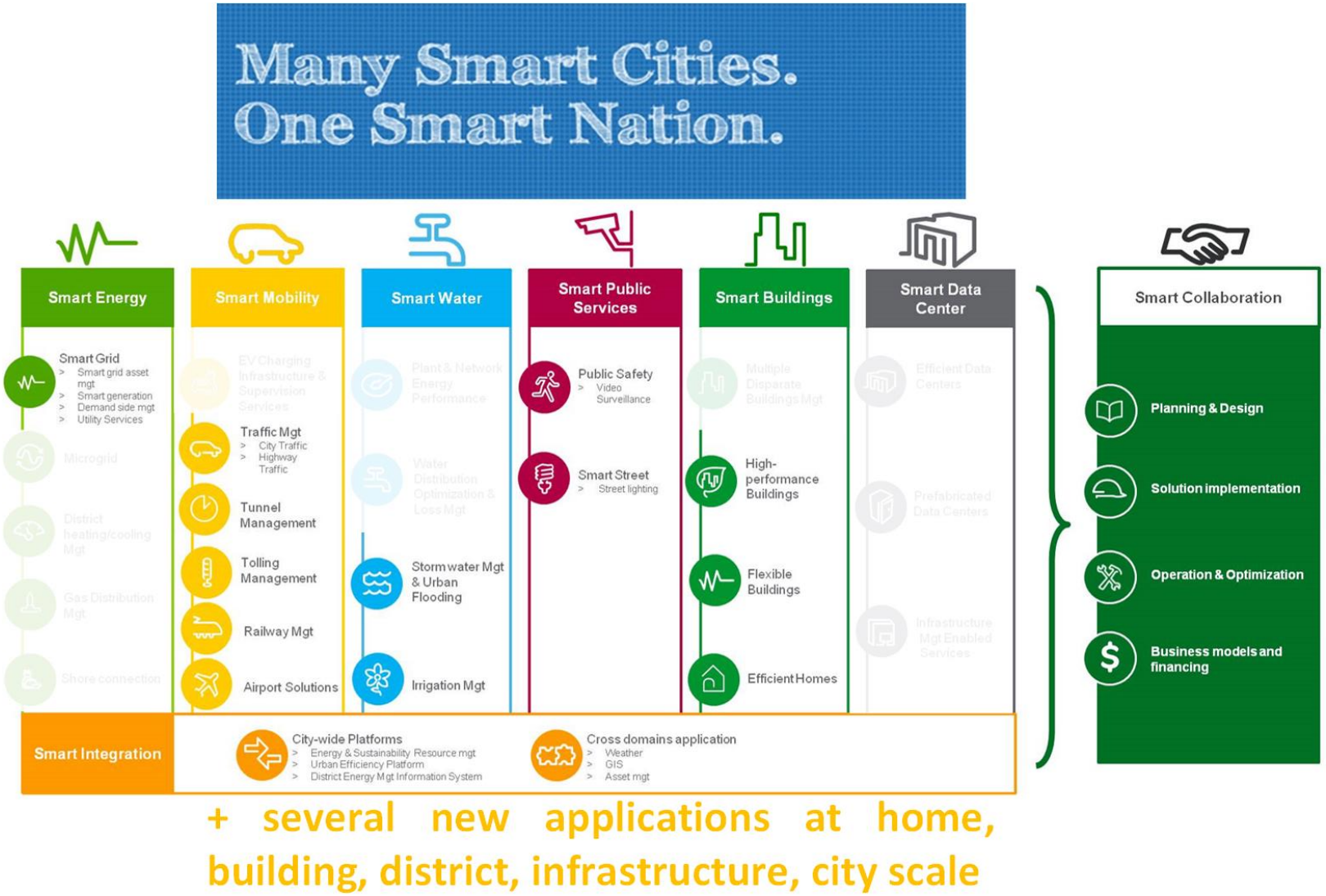


Fig. D23. The alignment of the "CogniVision"CRP program and the Smart Nation vision (cognitive cameras can be used to address several challenges and accelerate the fulfilment of the Smart Nation end goals).

Table V. Industrial collaborations of team members and adoption of their research work in areas that are relevant to the CogniVision project

team member	companies (Singapore)	research topic	notes
Prof. Massimo ALIOTO	Intel	ultra-low power digital signal processing	research collaboration
	Mediatek	energy-quality scalable circuits	Intellectual Property sharing, full fabrication support
	TSMC (Taiwan)	ultra-low power circuits for IoT	Intellectual Property sharing, full fabrication support
	Huawei, NeuroMem Technologies and several others	ultra-low power frontends for vision	possible licensing of previously developed vision technologies (under discussion)
Prof. FENG Jiashi	Huawei, Qihoo 360, Adobe, Snap on	deep learning and computer vision (vehicle detection, scene parsing, human pose estimation, …)	research collaboration
	Panasonic R&D	face verification/detection	adopted in Panasonic Face Pro system (most accurate face recognition in NIST IJB-A benchmark) and will be used in the surveillance system managed by the Singapore Ministry of Home Affairs
Prof. YEO Kiat Seng	GlobalFoundries, Samsung	RF device characterization and modeling	inductor design for RF, transformers, varactors, VCOs, RF transistors
	MediaTek, Panasonic, LTA, A*STAR, Broadcom, Infineon	RF transceiver architectures and power amplifiers	· has demonstrated the world's smallest on-chip low-pass filter (US Patent) with the broadest stop-band up to 52 times the cut-off frequency, i.e., 110GHz · 36G/24G front-end transceiver architectures with carrier suppression and ultra-low unwanted emissions, power amplifier and linearization techniques using active and passive devices
Prof. Luca Benini	Greenwaves Technologies	parallel-ultra-low power digital processor for computer vision, deep-learning accelerator	commercially licensed
	Google, Micron, STMicroelectronics, Mentor Graphics, Cadence	PULP open source platform for near-sensor analytics	publicly acknowledged adoption
Prof. CHEN Shoushun	Samsung	High Dynamic Range CMOS Image Sensor System with Adaptive Integration Time and Multiple Readout Channels" (US Patent)	commercialization in progress: signed NDA and disclosed patent details
	HILLHOUSE TECHNOLOGY PTE LTD, (Singapore-based startup company)	A High Speed Motion Detection Image Sensor" (US patent US 9,628,738 B2 granted in July 2017)	twelve-year exclusive licensing
Prof. Dennis SYLVESTER	founded two startups: 1) Ambiq Micro in 2010 based in Austin, TX 2) CubeWorks in 2013 based in Ann Arbor, MI	1) ultra-low power components for wearables and IoT 2) Michigan Micro Mote (M3) platform (one M3 design includes imaging	1) raised $90M to date, lead is Kleiner Perkins (VC that led Google funding) 2) Intel Capital is lead funder

		based on infrequent triggering)	