FULL PROPOSAL SUBMISSION TO CRP 20TH CALL

All information is treated in confidence. The information is furnished to the National Research Foundation with the understanding that it shall be used or disclosed for evaluation, reference and reporting purposes.

Proposal ID: CRP20-2017-0006

Proposal Title: CogniVision – Energy-autonomous always-on cognitive and attentive cameras for distributed real-time vision with milliwatt power consumption

Budget Requested *(Excluding Indirect Costs)*: **S$ $6,362,550.00**	**Period of Support: 5 years**	**Host Institution: NUS**

	Project Team Members *(Please add/delete rows where necessary)*				
Role	**Name** (Please include ORCID for Lead PI and all Team PIs)	**Designation**[1]	**Department/ Institution**	**% effort within this project**[2]	**% of time committed on the project**[3]
Lead PI	Massimo ALIOTO *(ORCID 0000-0002-4127-8258)*	Associate Professor	Electrical and Computer Engineering / NUS	40%	30%
Team PI (1)	Kiat Seng YEO *(ORCID 0000-0002-4524-707X)*	Professor	Engineering Product Development / SUTD	20%	20%
Team PI (2)	FENG Jiashi *(ORCID 0000-0001-6843-0064)*	Assistant Professor	Electrical and Computer Engineering / NUS	25%	25%
Collaborator (1)	Dennis SYLVESTER *(ORCID 0000-0003-2598-0458)*	Professor	EECS/University of Michigan, Ann Arbor (US)	5%	5%
Collaborator (2)	CHEN Shoushun *(ORCID 0000-0002-5451-0028)*	Assistant Professor	School of EEE / NTU	5%	5%
Collaborator (3)	Luca BENINI *(ORCID 0000-0001-8068-3806)*	Full Professor	IIS/ETH Zürich	5%	5%
			Total:	100%	

[1] For A*STAR's researchers, please indicate RSE grade.
[2] Represent % effort spent by the researcher in the project relative to his/her other team members. **The total must add up to 100%.**
[3] Represent % effort spent by the researcher in the project relative to his/her other job scope and other research grants. Lead PI and Team PIs are expected to commit a proportionate amount of their time in ensuring the success of the project **(at least 25% of the total time for lead PI and at least 20% for Team-PIs)**.

Rebuttal letter (1 page)

REVIEWER: The proposal develops capabilities at the interfaces between energy and performance, algorithms and hardware, and has potential for wide impact on the design of future smart systems. The Full Proposal should address the following:

- *Elaborate on the deployment architecture and the desired energy efficiency of the novel circuits for on-chip deep learning;*

We thank the Reviewer for the suggestion. In this full proposal, the proposed architecture for on-chip deep learning has been detailed, and its energy efficiency target has been clearly specified (50TOPS/W or better, i.e. 10X better than prior art with accuracy targets in line with real applications). Quantitative goals have been defined for individual blocks, as well as overall power/accuracy/memory/throughput goals have been defined in three well-defined use cases.

REVIEWER: Augment the team with more established expertise on deep learning and system architecture.

As suggested by the Reviewer, Prof. Luca Benini (ETHZ) has been added as collaborator. Prof. Benini is well known to be one of the maximum experts in architectures for deep learning in the world.

REVIEWER: Provide details of how the 4.5GHz data link would be designed and set up because it is a non-standard data link. The cameras would need to be talking to a customised kerbside radio access point (are any limitations to how far it can be away to talk to many such cameras in its vicinity?) which will probably not be ultra-low power;

We thank the Reviewer for catching this typo, the targeted carrier frequency has been corrected to 2.4GHz, which is widely used for wireless communications and networks, being an unlicensed band utilized by many existing standards. As correctly pointed out by the Reviewer, the radio transceiver developed in this project is fully compatible with most common IEEE standards in the 2.4GHz band (i.e., 802.11x including WiFi, Bluetooth, etc.)

The kerbside radio is assumed to be a conventional router that serves as a gateway. To limit the power of the transmitter in the cameras to mWs, a distance of few tens of meters (e.g., 20) is assumed. As correctly pointed out by the Reviewer, conventional wireless camera that transmits entire frames would not be low ultra-power. However, cognitive cameras perform on-chip computation and hence transmit only aggregate data (i.e., short packets) upon the occurrence of events (i.e., infrequently), thus determining a very low activation rate for the transmitter (0.001 or lower). In turn, this translates into an average transmitter power of μWs or less, which indeed justifies the cognitive camera approach.

REVIEWER: Clarify whether the System-on-Chip (SoC) work could be scaled to work at HD resolution at 30 fps processing rate.

HD resolution would entail a throughput increase by 3X within the same power budget. In principle, this objective might be feasible along the execution of the project. This will become clear once the proposed techniques and the underlying tradeoffs are well understood and proven on silicon. However, committing to such goal seems risky at this juncture, and pursuing VGA resolution at the same 30fps rate is a goal that our team is comfortable with.

NOTE ON THE BUDGET: in this full proposal submission, the direct cost of the project has been increased from S$5,66M to S$6.36M, compared to the original white paper. This is explained by the previous adoption of outdated salary tables, and the previously incorrect exclusion of the customary annual salary increases that adjust the EOM salary to inflation every year.

Research Proposal (20 pages)

Research Objectives

MOTIVATION AND THE GRAND-CHALLENGE

The **grand goal** of "CogniVision" is to enable the unprecedented capability of performing ubiquitous real-time vision through novel silicon chips that are untethered, always-on and nearly-perpetual, ultra-miniaturized (<100 mm^3), inexpensive (~1$). From a broad viewpoint, CogniVision introduces a new class of cameras that are "cognitive" and "attentive". CogniVision cameras are **cognitive** as they are able to constantly make sense of the scene through extremely energy-efficient circuits for best-in-class machine learning algorithms, i.e. deep learning based on convolutional networks [GBC16]. In the last few years, deep learning and convolutional networks have been extensively demonstrated to achieve outstanding accuracy, and to exhibit an uncommon degree of flexibility as they can be restructured (e.g., adjusting number of layers and weight values) to perform a very wide range of vision tasks. Indeed, **deep learning** has become the **de facto standard framework for image and video processing**, with remarkable success in content understanding [KSH2012], face detection [CHW2016], [WOJ2015], [SKP2015], [WZL2016], object detection [HZR2016] and tracking [NH2016], image classification [HZR2016] and segmentation [CPK2016], pedestrian detection [ZLL2016], loiterer detection [LNA16], abandoned luggage detection [SI18] (see examples in Table I in Annex D). Deep learning is an ideal framework for silicon accelerators due to easy upgradeability, and generality of its framework.

A given neural network is able to perform either a specific or a range of tasks (e.g., multi-task networks) [C93], [GBC16], [R17], but it cannot cover the entire range of all possible applications of distributed vision. To achieve broad coverage, the straightforward solution of storing a wide variety of networks on the same cognitive camera chip is not feasible, given the large amount of memory generally required for each network (e.g., 6-12 GBs in Table I), and the limited memory available on chip (various MBs, currently). Also, this approach would prohibit important capabilities such as

1) respond to time-varying requirements of the "cloud" server gathering the output of many cameras (e.g., request to perform a new task or occasionally send entire frames, as triggered by events captured by neighboring cameras, based on global understanding of the cloud)

2) upgrade the neural network, using its innate ability to be refined via retraining with new data

3) save power when degraded quality in processing (e.g., approximations) is tolerable for less visually demanding tasks (e.g., optical character recognition simpler than object detection).

A suitable approach to achieve these capabilities is to allow the cloud to push neural network configurations onto individual cameras, which in turn need to be responsive and receptive of the related commands from the cloud. Accordingly, cognitive cameras also need to be **attentive**, i.e. listen to commands wirelessly sent by the cloud, hence requiring an always-on radio receiver.

In general, nearly-perpetual always-on operation is pursued by harvesting power from the environment, which limits the power consumption of CogniVision cameras to ~1 milliwatt to maintain the system volume well below 100mm^3 (e.g., provided by a 0.1-mm thick, 5-cent, 1-2 cm wide organic photovoltaic foil attached to a wall [INF], with a stacked 0.4-mm equally sized battery [BTV] and on-foil printed antenna [GSI], all commercially available). Reducing the **power consumption of cognitive cameras down to the 1mW range** is the **fundamental objective** of this project. This entails a power reduction by at least 20-30X compared to the most power-efficient existing cameras that constantly monitor the scene with resolution and frame rate that are adequate for distributed monitoring and surveillance [PSC] (e.g., VGA resolution, 30 frames/s).

Cognitive cameras with power down to 1mW will be enabled by drastically limiting the amount of data transmitted wirelessly to the server cloud that makes sense of the scene, thus substantially reducing the traditionally large power due to the transmission of entire video frames (e.g., 40-50 mW with MPEG-compressed VGA frame, Bluetooth Low Energy transmission [G15b]). This is accomplished by embedding substantial sensemaking capability (e.g., object detection) into the

camera silicon chip, leveraging the recent impetuous advances in deep learning and convolutional neural networks [HZR2016], [DLH2016], [LAE2015], [ZLL2016] (widely adopted by Google, Facebook, Microsoft). As paradigm shift, CogniVision **moves sensemaking from the cloud to cognitive cameras**, keeping the power in the mW range in spite of the traditionally high computational complexity of deep learning. This will be achieved via innovation on energy-efficient circuits/architectures for sensemaking (see "Approach" section), including a novel digital energy-quality scalable architecture for general-purpose on-chip acceleration of convolutional networks with energy efficiency of 50TOPS/W or better, i.e. 10-20X more energy-efficient than the state of the art. Its ability to execute any convolutional network makes it applicable to the very wide (and ever-expanding) range of applications of convolutional networks, as long as the network fits the on-chip available memory and processing array size, as discussed in the "Subprojects" section.

Being "attentive", CogniVision cameras have also the capability to be responsive to the cloud, and occasionally be **reprogrammed by the cloud** in the following ways: 1) **transmit a short series of frames** to be processed directly by the cloud (e.g., if the visual task exceeds the cognitive capabilities of the camera); 2) **update the neural network** to a different one (i.e., uploading layer structure and weights), when the cloud requests a substantial change in the visual task executed by the camera (e.g., the cloud needs to identify very specific objects in a given area being covered by some of the cameras); 3) **statically adjust on-chip energy-quality knobs** that can save energy in vision tasks where lower processing accuracy or arithmetic precision are tolerable (e.g., less demanding visual tasks such as optical character recognition, as compared to more challenging tasks such as object detection – see details in the "Innovation and unifying framework" section).

As side benefit, cognitive cameras **solve the traditional issue of data deluge** in distributed vision systems. Indeed, frames from cameras are traditionally transmitted wirelessly to the cloud, involving large data volumes (~20 cameras exhaust the capacity of a wireless LAN, Internet video traffic is increasing alarmingly fast [CIS15]). This is avoided in cognitive cameras, as the transmitted data volume is reduced by several orders of magnitude (from preliminary simulations, they transmit at a data rate of ~1-10kbps on average, as opposed to several MBs in traditional cameras).

Regarding the **timeliness of the CogniVision project**, embedding vision in energy-autonomous nodes has been pursued for a decade [AMC06] with very limited success, due to the excessive power consumption required by on-chip processing. We are now witnessing the **convergence of three technology trends**, which are reshaping the areas of machine learning for computer vision and ultra-low power chips. On one hand, deep convolutional neural networks have made tremendous advances in terms of vision capability, although at substantial power and memory cost that is beyond the capabilities of energy-autonomous systems [ZLL2016], [ZGW2016], [SKP2015], [LAE2015], [KSH2012]. Their power power is now reaching the tens of mW range after two very intense years of research in deep learning accelerators [DL18]. Simultaneously, fundamental advances have been recently made in the area of energy-quality scalable integrated circuits and systems (including deep learning accelerators and vision processors), where substantial reduction in the intensity of computation and energy is achieved when moderate reduction in the quality of processing/sensing (e.g., arithmetic precision) is tolerable by the vision task at hand [A17b], [A16], [D15b], [FKB14] (see upcoming IEEE JETCAS journal special issue led by the PI [A18], and his recent book [A17]). Also, fundamental advances have been recently made in image sensor design, introducing the ability to embed simple in-sensor processing with low energy cost, limiting the expensive centralized processing requiring full frame readout [BLK16], [HMB16], [BFL16], [CPC15], [CSK15]. As convergence of the above trends, CogniVision leverages the well-known **exceptional robustness of deep learning/vision against inaccuracies to exploit energy-quality scaling and simple in-sensor processing**, which justify the timeliness of the project.

Recent market trends confirm the timeliness of CogniVision, and the expectable importance that smart untethered cameras will have in the years to come. For example, in December 2017 Amazon has acquired the wireless camera company Blink [F17]; in October 2017 Google has

released the CLIPS wireless camera [SL17]. Although the capabilities of such cameras are currently limited (e.g., actual lifetime from 3-5 hours with continuous shooting [GC17], [BLKa] to 2-5 weeks [BLKb], they simply record clips when event occur), this clearly shows a technological and market interest in ubiquitous vision. In 2017 Qualcomm announced the intention to pursue a research project on low-resolution (320x240) cameras for smart toys/appliances [QCM17] with low recognition capabilities (e.g., single object detection, ambient light sensing). None of the available cameras can interact with the cloud in real time (i.e., they are not "attentive"). As another example, in March 2018 Sony and other companies formed the NICE alliance to support the creation of a prospective generation of cameras with on-board analytics [NIC18], [NICE18b].

Ubiquitous cognitive cameras can provide **novel technological capabilities and societal benefits**, enabling for the first time situational awareness with fine spatial granularity across wide areas (from building to city scale). Fig. D1 in Annex D show examples of targeted applications, such as ubiquitous/augmented surveillance, vehicle/pedestrian detection, intelligent transportation, crowd monitoring, industrial plant monitoring, warehouse management, detection of dangerous objects, disaster management, among the others. In short, **CogniVision empowers the Internet of Things** (IoT) (i.e., ubiquitous sensor augmentation of the Internet [A17], [IOT14]) **with the sense of vision**, for the first time. As IoT is the next "big wave" of technology (45% annual growth, global value of 11T$ by 2025 [MCK15]), CogniVision will leverage its capabilities and potential growth to **create economic value in Singapore, accelerating the Smart Nation vision** [SN].

TECHNICAL CHALLENGES AND REQUIRED INNOVATION

The goal of CogniVision is pursued by embedding real-time scene sensemaking ("cognitive") and always-on radio receiver ("attentive") in a mW-power budget, addressing the following challenges.

A. Enabling sub-mW deep learning accelerators and power-aware neural networks

Deep learning hardware acceleration is well known to be compute-intensive. For example, the popular AlexNet network requires 122MB of memory, 1.14E9 multiplications/additions per frame, and the DRAM memory alone consumes a power of 12.8W at 20 frames/s, which is well beyond the power budget of mobile and IoT devices [HMD16]. In the last two years, the chip design community has aggressively pushed towards the conception of deep learning accelerators with power down to tens of mW under popular benchmarks such as AlexNet [DL18], , [UAH18], [[BWL17], [BCK17], [MV17], [WLL17], [SLL17], [DCB17], [YOT17], [AUO17], [PCR17], [MV16], [CKR16], [SPK16], [CKR16]. To aggressively reduce the power below 1mW while maintaining a throughput that accommodates for most typical vision tasks (target is ~10X the throughput required for AlexNet at 30fps and VGA resolution, i.e., 20 GOPS – 20 billions of operations/s), an inter-disciplinary approach cutting across neural network algorithms, digital architectures and circuits is needed. Innovative digital architectures beyond current implementations of convolutions through adders and multipliers are needed, to undercut the power cost of such power-dominant operators. Innovative neural network compression/training methods to reduce the model size down to MBs are needed to fit weights on chip, avoiding the large power of off-chip memories [HMD16], [HZC17]. Power/architecture-aware neural networks and training methods are needed to incorporate power into the network training loop, instead of conventionally focusing on mere accuracy. New opportunities to save power are available by focusing computation on "informative" frame regions.

B. Introducing innovative ultra-low power techniques to suppress irrelevant activity

The scene typically exhibits substantial redundancy in the temporal and spatial dimension. Our analysis of a large video dataset [CDN] has shown that only 3-5% of the frame changes between subsequent frames, and only 5% of such small fraction is truly novel (i.e., a new object is coming in) and hence deserves to be processed. Hence, a fundamental challenge is to truly **exploit such temporal and spatial sparsity of relevant and fresh information in each frame**, suppressing irrelevant activity in most regions of the frame where no new event/object is taking place. Instead,

today's low-power imagers and commercial cameras suppress computation only in the infrequent case where no pixel has changed in the frame [CPC14], [CSK15], [BLK16], [HMB16], [BFL16].

To address this challenge, innovative techniques are needed to empower all vision sub-systems with the **new capability to perform low-level, inexpensive and fine-grain** (e.g., small pixel tiles) **assessment of the relevance of the frame content**, before other more energy-hungry activity is performed at higher levels of semantic understanding. For example, motion or saliency should be assessed before feature extraction or classification, since the former can be performed locally and at much lower power. Similarly, innovation is needed in radiofrequency circuits for ultra-low power always-on wireless communications, to assure that the **radio receiver is always listening to the cloud**, while consuming only a **few hundred μWs** to fit the mW power budget.

C. Enabling innovative energy-quality scalable architectures for ultra-low power vision

As a further dimension that can be leveraged to reduce power, vision and machine learning algorithms are well known to be **robust against computation quality degradation** [A17], [A17c], [CMR10], (e.g., arithmetic precision, early termination of iterative algorithms). This translates into the (currently unexploited) opportunity to degrade quality and hence reduce energy in all levels of semantic understanding, when the visual task being executed allows it [MV17]. The concept of energy-quality scaling is general and is currently being explored [A18]. For example, several deep learning accelerators and neural networks with scalable arithmetic precision were proposed in the last two years [DL18]. Precision is statically optimized for a given machine learning task to minimize power, while meeting the classification accuracy requirement for the visual task at hand. Having clearly exhausted its potential [DL18], as uniform precision scaling (i.e., same precision for all neurons in the same network layer) needs to be superseded by more general approximations. Also, to allow aggressive power reductions, energy-quality scaling needs to be extended to all stages in the sensing-sensemaking chain, from the image sensor to the feature extractor, up to the machine learning engine. The challenge is in devising novel algorithm, architectural and circuit methods to insert energy-quality knobs that substantially reduce power, while slightly degrading quality and introducing minimal circuit overhead. As an example of the potential benefit of energy-quality scaling, at the end of 2017 the **PI's research group demonstrated the first energy-quality chip for feature extraction with 20X lower power** compared to the state of the art [APA17].

EXPECTED OUTCOMES AND SIGNIFICANCE

CogniVision introduces a **paradigm shift** enabling for the first time distributed and ubiquitous vision. In terms of **technological impact**, it empowers cameras with the following capabilities:

- UNTETHERED: CogniVision cameras are untethered in view of their ultra-low power operation, with mW power budget within the means of commercial energy harvesters
- UBIQUITOUS: CogniVision cameras can be deployed ubiquitously and unobtrusively thanks to their small size, as enabled by their single-chip integration and mW-power operation, which keeps the energy harvester and storage small (e.g., 0.1-mm thick, 1-2 cm wide photovoltaic foil [INF], with a stacked 0.4-mm equally sized battery [BTV])
- ALWAYS-ON: their ultra-low power consumption and suitability for harvesting enables energy autonomy and continuous operation for uninterrupted visual monitoring
- LOW COST: CogniVision cameras leverage the low cost of mass-produced standard CMOS chips and commercial harvesters, with a cost at volume in the dollar range
- ON-CHIP ANALYTICS: thanks to its processing energy efficiency (target: 50 TOPS/W), CogniVision brings deep learning from the cloud into cameras, enabling local data analytics
- DATA DELOUGE AVOIDANCE: CogniVision cameras perform significant computation on chip and transmit only aggregate information, reducing the wireless bandwidth by several orders of

magnitude (~10,000X). This eliminates the traditional issue of heavy network utilization of wireless cameras, and enables integration in existing wireless networks (no upgrade needed).

- MITIGATION OF PRIVACY ISSUES: privacy issues are mitigated since aggregate information is mostly transmitted to the cloud, transcending individuals.

In terms of **societal benefits**, the distributed/ubiquitous vision capability of CogniVision enables for the first time the ability to achieve continuous and pervasive situational awareness, and at very different scales (e.g., from building, to district and city scale). This capability has potentially very strong implications in terms of improved security, safety, infrastructure planning and dynamic optimization, shared services and urban resource allocation, location mapping of objects and subjects, crowd behavioral monitoring, context-enriched social media and augmented reality, disaster management, real-time visual search, among the many others. Our society can indeed greatly benefit from such non-intrusive technologies that can recognize and locate objects, situations and contexts of interest, and signal if greater attention or human intervention is needed.

From an **economic impact** point of view, CogniVision bridges image sensors and integrated accelerators for vision, and hence can make an economic impact in both markets, and open new market opportunities in the broad area of the IoT. The market size of traditional imagers is rapidly growing at CAGR of 10.3% reaching 17.5 B$ in 2020 [MAM15]. Much larger growth opportunities are expected in network cameras (CAGR of 43% until 2021 [TEC17], [GIA15], [GAR16]) and in the much wider IoT market (45% CAGR until 2025 [MCK15], [IFS16]). Vision in embedded cameras has become strategic, and has triggered the formation of the Embedded Vision Alliance [EVA] with 60+ companies (Fig. D3). CogniVision will leverage the above unparalleled capabilities, growth opportunities, and the convergence with the growing enterprise fabric in the IoT space to **create substantial economic value in Singapore**, and **accelerate the Smart Nation vision** [SN].

CogniVision will leverage **synergy with local industry**, with both semiconductor manufacturers and distributed vision system integrators, **and Singaporean ministries/agencies** (see letters of support). This will assure industrial relevance of the research outcomes, strategic positioning in the existing technological ecosystem, easier de-risking towards mass-production, strong alignment with real use cases, and true deployment in Singapore for in-field testing.

Approach

STATE OF THE ART AND RESEARCH LANDSCAPE

Networks of **massively distributed untethered cameras** with small size and very long lifetime (e.g., decades) were conceptualized a decade ago [AMC06], [KGS05], [A08], assuming that the camera technology would be eventually feasible. Today, such capability is **not yet available, due to the excessive power** of existing silicon chips for vision, which largely exceeds the 1-mW target. Fig. D4 (Annex D) summarizes the available architectures of untethered cameras.

The “raw-data” architecture #1 in Fig. D4 comprises an image sensor and a radio transmitting all raw video frames to the cloud. From Fig. D5a in the Annex D, conventional imagers for mobile platforms alone consume mWs or tens of mW [LMC16], [F15], [S15], [D13], largely exceeding the power target of untethered cameras. Hence, specialized ultra-low power image sensors are a necessity. Such imagers (see Fig. D5b in Annex D) typically achieve low power consumption at a severe resolution penalty (e.g., 64 x 64 pixels) [BDB14], [CLY13], [TCW13]. When fairly scaled to the same VGA resolution and 30 frame/s, various imagers [BDB14] can meet the above 1-mW power budget. Lower power is achieved by specialized imagers with multi-mode operation (Fig. D5c in Annex D) and limited sensemaking (see Fig. D5d) [CGM13], [KBF13], [CPC14], [CPC15], [CPC12], [CSK15], [CPC14], [KLF14], [CTL14]. As an example, the specialized sensor in [CGM13] performs in-pixel adaptive background subtraction through in-pixel low-pass filtering, and performs detection of rapidly changing pixels, providing a 2-bit 64x64 pixel output image at 13 frames/s and 33-μW power (although with poor resolution). Among the imagers that are capable of motion detection, the multi-mode sensor in [KBF13] performs motion detection with temporal averaging in

specific locations to detect slow object motion, while performing conventional motion detection in others, consuming 1.1µW at QVGA/30fps thanks to the suppression of the frame read-out when no motion is detected (power in normal mode is 29µW). Among imagers capable of feature extraction, the specialized sensor in [CPC14] is triggered by motion sensing and extracts Histogram of Gradient features from the captured image for the detection of objects of interest, with power consumption of 51µW at 256x256 resolution, 15 frames/s. As example of imagers capable of analog-to-information conversion (AIC), [CTL14] compresses non-overlapping 4×4 pixel blocks, and extracts mean and gradient via a capacitor array, consuming 110µW at QVGA resolution, 30 frames/s (mean, gradient and pixels are sent only if the gradient is large enough that it carries significant information). Event-driven imagers can capture faster events than all above time-driven sensors, but their power is at least in the order of mW when scaled at VGA resolution, due to the relatively large bias current [GMJ09], [LPD08], [RRL16]. The architecture #1 invariably exceeds the mW-power target by ~50X or more, considering the wireless power of a best-in-class radio with 5 nJ/bit [ITT16], due to the large amount of data produced in a frame (see Fig. D5a-d). Hence, the **architecture #1 in Fig. D4 with raw frame video streaming is unsuitable for mW cameras**.

The "compressed-data" architecture #2 in Fig. D4 substantially reduces the radio power by compressing frames (e.g., using H.264 or HEVC encoding), which is an intensive task entailing hundreds of mW in commercial and most research prototypes [CBW11], [WSK08], [SNX16], down to mWs in an extremely efficient research prototype [SFS09]. Under common compression ratios of 50 and best-in-class radios with energy of 5nJ/bit, the VGA bandwidth of ~2Mbps leads to a typical wireless power of 8-10mW, which added to the compression power exceeds the mW budget (see Fig. D6 in Annex D). Again, this makes the architecture #2 in Fig. D4 unsuitable as well.

The "cognitive" architecture #3 in Fig. D4 **with on-chip sensemaking is potentially viable for untethered cameras**, as it transmits only aggregate information, making the radio power negligible. However, existing specialized accelerators for scene analysis consume from ten to a few hundred mWs [SPK16], [HBS15], [DFC15], [CLL14], [CKR16], [CDS14], [LCL15], [HPP15], [PBS15], [KKL14], [PCL16] (see Fig. D7). Since 2016, several research prototypes of deep learning acceleration were demonstrated, with power from tens to hundreds of mWs on realistic workloads [UAH18], [MV17], [SLL17], [DCB17], [HLM16], [MV16], [CKR16] (i.e., ImageNet classification [IMG] rescaled at VGA, 30frames/s). Thus, the main challenge addressed in this project is to **enable sensemaking with power below 1 mW** (see details in next section). To drastically reduce the power consumption due to off-chip memories (~1,000X larger than on-chip memories [HZC17]), aggressive neural network compression techniques were recently introduced to reduce the memory requirement down to the MB, as available on chip [HZC17], [BWL17], [HMD16], [IHM15].

The additional ability to be "**attentive**" is achieved by the architecture #4 in Fig. D4, through the inclusion of an always-on on-chip radio receiver whose power needs to be significantly lower than the targeted mW power. Conventional wireless receivers for a range of tens of meters (e.g., Bluetooth) consume from several mWs to a few tens of mWs [W18], [LDB17], [BSM17], [ISS17], [KFC17], [LNZ17], [CLB15], [PPW15], [LKH14], [DLS10]. To reduce the receiver power, wake-up radios with power of several tens of µWs have been proposed to allow continuous monitoring of the wireless channel and detect transmission, before turning on the main receiver to complete the reception [SCK16], [BY15], [YJC12], [HBH12], [PGR09]. Unfortunately, most wake-up radios require the addition of an off-chip high-Q resonators (e.g., bulk acoustic wave, crystal), whose cost and off-chip connection are incompatible with the requirements of sensor nodes [SCK16]. The very few wake-up radios that do not require high-Q resonators [SCK16] are not suitable either, as their intrinsically limited capability to reject interferences would cause frequent false positive transmission detections, in public environments where several tens or hundreds of radios can overlap in the same area (e.g., smartphones, wearables, wifi at 2.4GHz). Also, proprietary solutions (e.g., frequency diversity [HBH12]) to ignore the transmission of other wireless devices and focus

only on the transmission from the cloud basestation are not feasible, as the CogniVision cameras need to fit existing communication standards used in commercial basestations, for obvious compatibility reasons. Hence, novel transceivers with average power consumption of hundreds of μWs (targeted: 350μW) are needed to fit the mW power budget, as discussed in the next section.

Regarding the state of the art of **research-stage untethered vision systems** with imaging and on-board intelligence (Figs. D8-D9 in Annex D), most of them consume a power from tens of mWs to a few hundred mWs [BCK17], [YKU17], [LZG17], [RRF17], [AXC16], [RRL16], [LLR16], [DSR15], [SYH14], [CBD13], [MTB13], [IBJ13], [CLB13], [CBW13], [LD13], [CVS11], [CBD11], [ANA08], [HPF07]. Only very few are in the mW range or lower, when fairly scaled at the same VGA resolution and operating in a public space with reasonably frequent events [RRL16], [LLR16], [KLF14], [CLB13], [CBW13]. However, most of them (with the only exception of [KLF14]) are actually application specific and hence cannot be used as a vision platform across different applications. Also, their vision computational capabilities are very limited and can deliver a throughput in the order of tens of MOPS or lower (MOPS=millions of operations per second), whereas non-trivial vision tasks at VGA resolution and 30frames/s require thousands of MFLOPS or more [PCL16], [SPK16], [CKR16], [SPK16], [PBS15], [HBS15] (see, e.g., examples in Table IV). CogniVision aims to fill the power and on-chip computation-ability gap in existing cameras, simultaneously targeting mW power and 20,000MOPS to cover a wide range of tasks.

As further sign of rapidly growing interest in the area of distributed vision, several **companies and startups** have recently released their first prototypes of untethered cameras [KNT17], [BLK16], [HMB16], [BFL16], [NUB16], [SUC16], [CFC16], [LUC16], [FFX15], [ARL15], [PIP15], [ARC14]. As shown in Fig. D9, their lifetime is still very short (from hours to weeks) and inadequate for distributed vision, the size is in the 5-10 cm scale, and the cost is in the hundreds of dollars range. In large-sized companies, various **industrial research&development efforts and startup acquisitions** have recently been carried out. As mentioned above, in December 2017 Amazon has acquired the wireless security camera company Blink [F17] (lifetime of a few weeks); in October 2017 Google released the CLIPS wireless camera [SL17] (lifetime of hours or days), in 2017 Qualcomm has announced the intention to pursue the Computer Vision Module research project to enable low-end untethered cameras for smart toys and appliances (equivalent power of 10mW when fairly scaled to VGA), and is currently hiring researchers in the field [QCM17]; in March 2018, Sony/Nikon/Scenera/Foxconn/Wistron formed the Network of Intelligent Camera Ecosystem to create a new generation of smart cameras [NIC18]. Other companies that are currently collaborating with the team members are also starting exploring the area (not publicly disclosed), due to the potentially large market of distributed vision. Finally, well-known efforts on machine learning accelerators (e.g., Google's TPU, IBM's TrueNorth) target datacenter-scale applications and power levels that are several orders of magnitude larger, hence they are not relevant to the area investigated in CogniVision. A summary of current **industrial interest and collaborations with our team** in the area of distributed vision is detailed in Table IV in Annex D.

Table II in Annex D presents the analysis of the **research landscape** in ultra-low power silicon chips for vision, leading researchers and limitations of previous work. From Table II, there is no available research outcome enabling mW cognitive and attentive cameras with significant computation-ability (e.g., tens of thousands of MOPS). The effort has indeed been fragmented into the optimization of individual components, and has not involved the integration of machine learning into a fully integrated ultra-low power imaging system on chip. **CogniVision aims to fill this research gap**. Table III summarizes related **research programs funded by DARPA, NSF, EU** and others. From this table, ubiquitous vision has recently become a very hot topic, but research is being focused mostly on individual algorithms (Virtual Cortex on Silicon, SAF-T, NeoVision2, SyNAPSE), imagers (REImagine), computer architectures (COPCAM). Research programs on cameras (Vision-in-Package, IcyCAM) target only wired systems, due to less ambitious power

targets than CogniVision. Again, this project is distinctively focused on **on-chip vision system co-design (from imager to processing) with aggressive mW power budget**.

COGNIVISION: INNOVATION FRAMEWORK (including preliminary results)

The **ambitious mW power target is pursued** by introducing innovation in **three dimensions** (Fig. D11 in Annex D), corresponding to the challenges in the "Technical challenges" section.

A. NOVEL SUB-MW DEEP LEARNING ACCELERATORS & NEURAL NETWORKS: a novel class of energy-efficient deep learning accelerators and innovative deep neural networks will be investigated, from circuit to algorithm level. The **proposed class of deep learning accelerators** enables unprecedented energy efficiency in the dominant energy of convolutions and products, leveraging on the drastic memory energy reduction allowed by novel compressed neural networks that can fit the memory available on chip (instead of being conventionally off chip). The proposed approach is based on the Dyadic Digital Pulse Modulation (DDPM) [C17], which provides a non-binary representation of an integer number x consisting of a digital bitstream with a 1's density proportional to x over any time interval. In DDPM, the number of pulses in a time interval w is proportional to the product x·w as in Fig. D12, with a resolution that increases with width w. Hence, products and weighted sums (including convolutions) can be simply computed by counting pulses.

More quantitatively, consider N DDPM-encoded input features in a convolutional network x_i [M12], [GBC16], which are multiplexed over time windows with different width w_i, and with a fixed total duration $W = \sum_{i=1}^{N} w_i$, as in Figs. D12a-b. The resulting total number of pulses is proportional to the weighted sum $y = \sum_{i=1}^{N} x_i w_i$, and can be computed by a binary counter. Interestingly, the total computation time is independent of the number of weights N, and the resolution in each product is determined by each w_i, while the overall accuracy of the end result can be proven to be constant and set by W. Hence, the computation time and energy are constant and independent of the number of weights N, and depend only on the targeted output resolution, which is set by the total duration W (see example in Fig. D12c). This property allows for combining a large number of products in nearly-constant time, providing at least an order of magnitude complexity reduction compared to conventional multiply-and-accumulate (quadratic in the number of multiplications and thus kernel size, which is typically between 3 and 11 [D15b]). This also allows to achieve pre-defined accuracy in the final result, and have a predictable accuracy-computation time tradeoff. Considering that DDPM modulators are very simple [C17] and the weighted sum is computed by simple binary counters, the proposed approach is well suited for very efficient implementations of large-scale deep learning accelerators based on the novel architecture in Fig. D12d. In this architecture, the input data is converted into 1-bit DDPM streams, and forwarded to neurons via a multiplexer network. Neurons are simple binary counters activated by pulses encoding the weights. Our preliminary post-synthesis simulations show that an **energy efficiency** of 50TOPS/W can be achieved in 28nm CMOS, which is **at least 10X better than state-of-the-art** accelerators whose accuracy has been proved to be adequate for real applications [DL18].

At the neural network **algorithm level**, innovation will be introduced both at the network compression and at training time. Compressed power-aware networks will be generated by **introducing for the first time the energy cost within the training objective of the deep learning** model. To this aim, reinforcement learning (RL) will be introduced to achieve power-aware model training, using circuit power models for the deep learning building blocks, and hence **closing the training loop with circuit-level information**, as opposed to conventional designs where circuit and network designs do not interact with each other. At training time, the novel approach of non-uniform precision will be introduced to leverage the fact that different weights and filters have different importance in terms of final deep learning model output. This fact has been extensively exploited in pruning [HMD16], whereas precision has been kept uniform across weights. In CogniVision, **for the first time we introduce the notion of non-uniform precision** by allocating higher arithmetic precision (i.e., energy) to most important weights, while scaling down the

precision in other weights. Our preliminary results on CIFAR-10 dataset [CFR] (Fig. D13) promise up to 10X circuit and energy reduction at same accuracy, compared to conventional uniform precision approaches. Interestingly, the non-uniform precision adjustment approach matches well the intrinsic capability of the DDPM architecture to assign a different precision to different weights during DDPM weight encoding, and makes the overhead of non-uniform precision irrelevant. The synergy between DDPM and non-uniform precision offers a fundamental advantage, as adopting multiple precisions in conventional accelerators is typically expensive [DL18].

The inherent redundancy of deep learning networks will be removed by developing **novel "deep compression" techniques** consisting of pruning, weight quantization (including binarization [C16]) and information theory-based coding. Among the novel ideas that will be investigated, pruning based on **hard thresholding** of its parameters (e.g., with small activations) will be explored, as shown in the example in Fig. D14a and described in the proposed approach detailed in Fig. D14b. Iterative hard-thresholding (based on gradients) approach to identify the task-specific redundant neurons and compress the deep network model by removing those neurons. The approach would search for the redundant neurons within the network model based on magnitude information about the back-propagated gradient. From our preliminary results, such deep compression framework can reduce a state-of-the-art deep neural network model by 1,000-2,000x, thus reducing the memory requirement from the GB range down to sub-MB, with negligible performance drop. This is a 2-4X improvement over the results demonstrated with recent and popular compression techniques, which can achieve 500X compression in AlexNet IHM15], [HMD16]. Further energy reductions will be pursued at the algorithm level by embedding novel techniques that make deep learning data-adaptive, allocating energy on “important” or “informative” regions of the frame. Accordingly, attention mechanisms for automated detection of critical parts of frames will be investigated. Leveraging on our current exploratory work, small deep neural networks with memory can be used to select regions of interest (e.g., Recurrent Neural Network with Long Short-Term Memory), as in Fig. D15. The **model essentially learns which parts in the images are relevant for the task at hand**, and attributes higher importance to them. According to our preliminary results, deep models with attention show that a bird out of 200 species can be recognize at the accuracy of 70% by introducing an LSTM-based attention network, which can focus its “attention” to a small region of only 40x40 pixel. We have also observed that the insertion of gating functions can further increase the image recognition accuracy by 5%. Combining compression and attention models, model size reductions exceeding 1,000x were observed.

As other fundamental sub-system necessary for deep learning acceleration, a **novel class of static RAM (SRAM) on-chip memories with non-precharged bitline** will be introduced to reduce the bitline switching activity. Indeed, the latter is well known to be responsible for the largest power contribution [FKB14], [R13], due to the constant bitline precharge at the supply voltage, which determines a bitline transition regardless of the value stored in the accessed bitcell [R13], [WH11]. Instead, the novel SRAM bitcell in Fig. D16 does not require any bitline precharge since it is able to drive the bitline to both ground and the supply voltage. Accordingly, if the same value is being read in adjacent memory accesses (e.g., due to the well-known spatial correlation between adjacent pixels [S10]), the bitline will not change value and hence will give negligible contribution to the power. Preliminary circuit simulations in 28nm showed 70-80% bitline activity reduction compared to a conventional precharged SRAM. For a typical SRAM where the bitline accounts for more than 50% of the overall power [FKB14], [R13], the adoption of the proposed SRAM for the frame buffer is expected to lead to 40% power reduction. Interestingly, this method permits to reduce activity by 75% even without bit correlation across memory accesses, as pairs of random and uncorrelated values with 0.5 switching probability clearly lead to a bitline activity of 0.25 (i.e., bits in adjacent accesses assume the same value with probability of 0.75). Hence, the same technique allows about the same power saving even for the weight memory.

B. NOVEL CIRCUITS FOR IRRELEVANT ACTIVITY SKIPPING: conventional vision systems leverage the temporal sparsity of the frame information content to suppress processing (e.g., frame is not processed if it is the same as the previous one, e.g. [LZG17], [QCM17], [AXC16], [CSK15], [CPC14]). However, spatial redundancy is largely unexploited, as existing approaches re-compute the entire frame even when appreciable motion is detected in a single pixel [RRF17], [CLP17], [RRL16], [BLK16], [KLF14], [CBD13], wasting a vast amount of processing. In CogniVision, **both temporal and spatial information sparsity are simultaneously exploited to skip irrelevant activity on selected parts of the frame** that are changing, novel and salient. This will be achieved by introducing the scheme in Fig. D17 and novel circuit techniques in all sub-systems to inhibit their irrelevant activity (from imager, to feature extraction, classification and wireless communication).

Regarding the architecture in Fig. D17, **un-necessary energy-hungry tasks are stopped via inexpensive assessment of their relevance** at the least abstract (i.e., lowest-energy) level of understanding. For example, computation in a given region is stopped if there is no salient content (e.g., tile), or if there is no feature extracted in that region, or if extracted features are not novel as they correspond to an object that pre-existed in the previous frame (rotated/translated/resized) (Fig. D17). As shown in Fig. D17, the classifier utilization and power are reduced by activating it only for frame regions that contain features, as well as salient and novel information content. Each level of abstraction generates its conventional output and an additional **"relevance table"** (i.e., on-chip small memory) identifying the tiles where relevant content is being detected (see below and Fig. D18), to let the next (i.e., upper in Fig. D17) sub-system skip the irrelevant frame portions.

In regard to the irrelevant activity detection in each sub-system, the image sensor will be enriched with a **novel in-sensor saliency detector circuit**, which distinguishes tiles of pixels that change in intensity over time, while identifying and ignoring the background. The proposed in-sensor saliency detector executes the frequency-tuned saliency algorithm [AHE09] with very simple circuit techniques described in Fig. D19, which consists in the comparison of pixels with their long-term average. Such comparison highlights the important changes compared to the background or to objects that have remained in the frame for a long time and are hence progressively blending with the background. Interestingly, this approach generalizes conventional motion detection as the latter is simply obtained by performing no time averaging (i.e., the proposed in-sensor approach includes conventional motion detection as particular case). As in Fig. D19a, the proposed in-sensor saliency detector circuit has a fundamental difference compared to the algorithm in [AHE09], as it can monitor (squared) tiles of pixels instead of individual pixels, and hence permits to monitor intensity changes with fewer read-outs and hence lower read-out power. As an example, if a 5x5 tile is chosen as in Fig. D19a, the overall current generated by the corresponding photodetectors within the pixels is read out, instead of reading all 25 pixel currents. This reduces the number of read-outs by 25X, and the power by the same factor, while maintaining an accuracy of 92% (Fig. D19b). This **drastically reduces activity, compared to conventional vision systems where the imager invariably reads out all pixels** whenever some event is occurring within the frame, and rigidly process all of them at a higher level of semantic understanding to identify events.

In CogniVision, the feature extractor is based on the ORB algorithm [R11], and detects keypoints (i.e., low-level "point of interest", e.g. corner, blob [S10]). As in Fig. D17, the feature extractor is enriched with the new capability to skip keypoint extraction in portions of the relevance table that are tagged as irrelevant. The keypoints in irrelevant areas are not re-computed, as the ones coming from the previous frame are reused. In CogniVision, we will **leverage our results published in late 2017 [APA17] with the first ORB chip demonstration**, whose power is **well below 1mW for the first time, and 20X lower** than the next best in class. Architectural evolution in CogniVision to further reduce power by >3X is discussed in the next section.

Similarly, the **new capability of assessing novelty of keypoints** in salient portions is introduced in CogniVision (see Fig. D17) through a novel mechanism that is based on the fundamental

observation that novel keypoints are those that cannot be matched to the keypoints in the previous frame. In other words, novelty assessment is simplified into the well-known keypoint matching problem across adjacent frames (although usually matching is performed between a frame and an image database [S10]). A **novel low-complexity approach to perform real-time inter-frame keypoint matching** will be explored in CogniVision, leveraging the fact that ORB generates keypoint in strict order, where ranking is dictated by corner measure [R11], [APA17]. The proposed approach is based on the consideration that the ranking of keypoints across adjacent frames is strongly correlated, i.e. an important keypoint likely remains important in the next frame, and hence in the top part of the ranked keypoint list. Accordingly, matching can be performed by confining the comparison of keypoints with similar ranking in adjacent frame (40 out of 400 in our experiments), instead of exhaustively compare all possible pairs of the 400 available keypoints. From preliminary ORB simulations in the OpenCV environment [OCP], complexity of keypoint matching can be brought down by an order of magnitude, and hence close to the complexity (i.e., power) of the feature extractor. Similarly, the relevance table generated by novelty assessment confines the deep learning computation in the new frame to the activations in the output feature map of each layer that are affected by the novel content, whereas other activations will be retained (i.e., not re-computed, but stored on chip) from the previous frame. For example, preliminary simulations with AlexNet network required 3MB for all activations, which can be stored on chip. Although its power benefits are not accounted for in the estimates in this proposal (due to the difficulty to have a solid architectural-level estimate), we expect this will add at least 2X energy efficiency improvement.

Irrelevant activity skipping will be consistently performed at the wireless communication level as well (top of Fig. D17), so that the power-hungry main receiver to make CogniVision “attentive” to cloud’s requests is turned on only when the cloud is truly transmitting. As discussed in detail in the next section, this will be achieved through **innovative radio-frequency techniques at the circuit level** (operation at the 2.4GHz ISM band is targeted, for compatibility reasons with standards such as BlueTooth, WiFi, etc.) At circuit level, ultra-low voltage operation will be pursued through circuits that leverage transistor operation at the lower boundary of the near-threshold region (i.e., 0.5V supply instead of conventional 1.2-3V [LNZ17], [PPW15], [CLB15], [YJC12]). This drastically reduces the transistor gate-source voltage and hence the minimum supply voltage and power (essentially by the voltage reduction factor, i.e. 2.5-6X), at the cost of an order of magnitude wider transistors. As side benefits, the transconductance/current ratio is improved over conventional designs at larger voltages, and latch-up immunity is substantially improved due to the intrinsic inhibition of the parasitic bipolar transistor at 0.5V [YDB10]. As opposed to conventional standalone radios, the overall area of the CogniVision system on a chip is clearly dominated by the image sensor and the deep learning array, thus making the larger area of the radio acceptable. As another challenge posed by near-threshold operation, on-chip parasitic and noise models delivered by silicon foundries are no longer reliable, and proprietary modeling approaches are needed. On this, we will leverage the extensive modeling research work that our team members have carried out in the last decade [CYC15], [OYC14], [YDB10].

The transceiver is a single-chip solution with on-board antennas (printed on top of the flexible solar cell hosting the chip), operating at the 2.4GHz frequency range based on On-Off Keying (OOK) modulation. The OOK transmitter includes a 2.4GHz Voltage-Controlled Oscillator (VCO) and an OOK switch with an antenna driver stage. The receiver consists of OOK power decoder/detector, comparator and a driver to interface with the baseband chipset. The communication distance of up to a few tens of meters (targeted: 20m) with two separate compact antennas for Transmit (TX) and Receive (RX) will mitigate the requirement of complex TX/RX switch at the receiver front-end. To save power, a wakeup and a sleep mode can be selected on the receiver. Due to the crowded 2.4 GHz frequency band, a secure link will be established between the transceiver and the wireless basestation (off-the-shelf). Under non-functional state, the transmitter and receiver are in sleep/idle mode consuming a negligibly small dc power with the

driver stages in shutdown state. To wake up and initiate the secure communication, the transmitter can initially send a known bit sequence which will be detected at the receiver front-end and compared internally. Once the bit sequence is matched, the detector stage shall power up the driver stage and establish the communication path. Combined with the above near-threshold power reduction, this is expected to bring at least an order of magnitude lower power from previous exploration (4mW), thus reducing the receiver power to hundreds of uWs (our target is 350uWs).

C. INNOVATIVE ENERGY-QUALITY (EQ) SCALABLE ARCHITECTURES: deep learning and vision algorithms are well known to be **resilient against noise and inaccuracies**, as exemplified by lower precision [D15b], [HMD16], [GAG15], and approximations [VRR14], [IHM15]. This offers the opportunity to deliberately degrade quality of sensing and sensemaking, and hence reduce the energy consumption, if the visual task at hand allows. The energy-quality scaling concept has been pioneered by the lead PI [A18], [A17], [A17b], [A17c], [FKB14], [A16], and provides the cloud with an **additional (optional, but very effective) knob that can reduce the power consumption for tasks that are not particularly critical, or not particularly visually demanding**. Such knobs are statically set by the cloud for a specific task and neural network, but can be occasionally changed by leveraging the fact that CogniVision cameras are "attentive", and can hence be occasionally reconfigured. The **energy-quality knob optimization is performed offline while training the neural network**, via the same methods that are used to adjust the arithmetic precision in deep learning accelerators [MV17], [HMD16], [MV16]. If the user is more interested in minimizing the training effort, all knobs can be simply set at maximum quality and ignored. Accordingly, the values of the energy-quality parameters optimized while training the neural network are integral part of the CogniVision system configuration for a specific task, along with the weights of the neural network.

The **innovation brought in CogniVision on this dimension** lies in the explicit tune-ability of energy-quality knobs in all sub-systems, from the image sensors to deep learning. This capability is not available in current vision systems, and is an additional opportunity to reduce power for a specific task. According to the experimental chip results recently published by our team [APA17] on feature extraction, 3X power reduction is achieved from energy-quality scaling alone. Similar or better power reductions by 4-5X are achieved in deep learning accelerators with adjustable precision [DL18]. Accordingly, energy-quality scaling is expected to provide substantial power savings. However, accurately quantifying such power savings through simulations is computationally extremely intensive, and its accurate exploration can be feasibly performed by using the **CogniVision system on chip as a valuable tool to gain a better understanding of the energy-quality tradeoff in real-world applications**. The following knobs will be considered in CogniVision in each sub-system in Fig. D14:

- IMAGE SENSOR: three knobs will be considered, the tile size in Fig. D19, the threshold ε for saliency detection in the same figure, and the analog-to-digital converter (ADC) resolution. Larger tiles and higher thresholds ignore more local events and save power, at the cost of lower saliency assessment accuracy. Similarly, lower ADC resolution saves power in the read-out (typically 2X for each one-bit resolution reduction [FFA14]) Among the other dimensions that will be explored in this sub-project, the resolution of the Analog-to-Digital Converter (ADC) for the readout will be adjusted via resolution-scalable architectures (see, e.g., [FFA14]).
- FEATURE EXTRACTOR: the same energy-quality knobs that have been explored in [APA17] will be embedded, as they have been proved to be very effective.
- NOVELTY ASSESSMENT: one knob will be considered, i.e. the number of bits of the keypoint descriptor that are used to compare and match keypoints (see previous subsection).
- DEEP LEARNING: the adjustment of (non-uniform) precision is the main knob, as in the CIFAR-10 example in Fig. D13. From this plot (and several others, omitted), non-uniform precision adjustment permits to trade off energy and quality on a very wide range, thanks to the much more graceful quality degradation in Fig. D13a (10X at 5% quality degradation.

COGNIVISION: SUBPROJECTS

The project is structured in four sub-projects, which all converge into the final demonstration in sub-project #1 of the CogniVision system on chip (see **in-principle architecture** in Fig. D21). Sub-projects are organized in an inter-disciplinary manner, and are centered around the interaction between sub-systems and levels of abstraction.

1. System modeling, exploration, integration, demonstration of cognitive/attentive cameras (led by M. Alioto, joined by all)

This sub-project addresses the system-level challenges and unifies the efforts of the other sub-projects into a **cohesive modelling, design and verification framework**. Regarding the system modelling, a high-level simulation framework will be developed and shared among all PIs to evaluate the functionality, the performance and the energy efficiency of individual components, as well as their impact at the system level. Energy per operation will also be modelled using proprietary models, to preliminarily estimate the benefit of each innovative technique before performing time-consuming circuit and architectural design. The same environment will be used to share a common database of benchmarks for quantitative assessment, and to perform experiments in a controlled environment shared by all researchers in the team. Tentatively, the environment will be in OpenCV-Python [OCP] as a compromise between Python's code readability (as needed in collaborative efforts) and availability of OpenCV libraries (which has also been used by the PIs to generate some preliminary results). Such environment will also be used to generate test vectors for chip testing.

This sub-project also covers the **system design, integration and demonstration** aspects in CogniVision, once the above preliminary exploration is performed, and circuit/architectural techniques are investigated and developed for silicon implementation in other sub-projects. System integration will be first performed as a System on Board (SoB), assembling the stand-alone chips that are generated in the various sub-projects for two silicon rounds. The final demonstration is instead performed in the form of a single System on Chip (SoC). Accordingly, chip design partitioning and floorplan will be preliminarily performed, and a mixed-signal simulation/verification environment will be developed to verify the design from behavioral down to gate-level and some selected circuit simulations, when designs become available over time for the blocks in the CogniVision SoC. Also, this sub-project focuses on the silicon infrastructure for chip configuration and testing, based on the CogniVision chip architecture in Fig. D21. Once verified and taped out, the CogniVision chip will be fabricated by a commercial silicon foundry (e.g., GlobalFoundries) and tested in a real-world environment to assure that the ultimate quantitative targets in Table IV are achieved. The targeted use cases in this table are well within the capabilities of CogniVision, both in terms of memory (2MBs) and throughput (<20,000MOPS). The on-chip microprocessor (tentatively PULPino by ETHZ, also team collaborator [PLP]) in Fig. D21 does not affect the performance, as it is only configures the accelerators and weights into the on-chip memory.

2. Energy-centric circuit techniques and interaction at imager-sensemaking and wireless-sensemaking boundary (led by K. S. Yeo, joined by PI M. Alioto and collaborator S. Chen)

In sub-project #2, the interaction of sensemaking with the image sensor on one side, and the wireless interface on the other side is investigated, according to Fig. D11. From the perspective of the irrelevant activity skipping, imager architectures with in-sensor saliency and relevance table generation will be explored, while systematically taking its interaction with feature extraction into account (Fig. D17). The image sensor will include novelty (the above in-sensor saliency detection circuitry), whereas the pixel and array architecture will be taken from prior designs from Prof. Yeo's group [CAB08], [WHY12] to de-risk the demonstration, considering that the energy efficiency of the imager is not critical for the system. Also, the wireless communication circuits will be developed while incorporating their interaction with sensemaking, in particular with the deep network configuration, which is uploaded by the cloud into the on-chip memory for reconfiguration purposes.

In this sub-project, the image sensor and wireless transceiver are first explored from an architectural point of view. This is followed by two rounds of chip demonstration and testing to first validate the fundamental ideas and translate it into circuits, and then refine the design in preparation for the final System on Chip (SoC) demonstration. In the latter phase, the effort is focused mostly on the fine-tuning and integration with the other blocks in Fig. D21. A characterization of the final prototype will be performed, and correlated with silicon measurements in the two previous versions, evaluating the effect of process/voltage/temperature corners.

3. Energy-centric machine learning-circuit co-design (led by J. Feng, joined by M. Alioto and the collaborator Prof. Luca Benini)

This sub-project focuses on the algorithm-circuit interaction, through the investigation of a novel class of deep neural networks that will be designed and trained by including power consumption as explicit metric/cost function, as opposed to conventional machine learning methods focusing on pure accuracy [HVD2015]. Also, a novel class of ultra-efficient deep learning accelerators based on the DDPM modulation (Fig. D12) will be investigated.

In this sub-project, we investigate systematic **energy-aware model design and training** schemes, introducing the energy cost within the training objective of the deep learning model. Being circuit/architecture parameters within the network optimization loop, this creates an interdependence and ultimately a synergy that is of particular interest for this sub-project. At the same time, low-activity SRAM memories will be explored and demonstrated. Machine learning circuit techniques will be explored that **smartly allocate energy between training and sensemaking**, adding run-time criteria for early termination of the computation, without incurring further unnecessary energy cost while accuracy is plateauing. The developed energy-centric machine learning algorithm-circuit co-design will be validated in terms of accuracy and energy in applications for processing images at the resolution from 1,000x1,000 to 80x80 to assess the scalability of the proposed techniques. The resulting models will be validated and integrated in the final silicon prototype first in a controlled environment, and then in a real-world setting. Benchmarks provided by our project partners (see letters of support from agencies) will be used to this purpose, covering human and object recognition, in addition to the popular AlexNet benchmark (Table IV).

4. Irrelevant activity skipping/EQ-scalable sensemaking circuits/architectures (led by Alioto, joined by all, including the collaborator D. Sylvester)

This sub-project focuses on the circuit and architectural implications on the sensemaking of the three research directions in Fig. D11. Regarding the irrelevant activity skipping, the processing elements in Fig. D17-D21 will be organized both logically (architecture) and physically (floorplan) in a regular fashion that maps the imager tiles (see sub-project #2) onto the sub-systems that perform the corresponding computation. To this aim, **novel chip design methodologies pursuing vertical integration from physical level to architecture** will be developed in this sub-project, with the goal of assuring data locality (to limit the large energy cost of signal distribution) and maximizing the reuse of memory accesses (to limit the large energy cost of multiple accesses to the same memory address). In regard to the energy-quality scalability, this novel capability will be introduced in all components of the SoC. The fundamental vision algorithm parameters will be evaluated as primary candidates for being used as energy-quality knobs, and their impact on energy and quality will be preliminarily assessed through high-level simulations (e.g., OpenCV [OCP]. Also, this sub-project involves the translation of the expected research results into measurable chip demonstrators of saliency pre-assessment, feature extraction, novelty assessment, and deep learning in Fig. D17. These circuits are designed and tested in two rounds, respectively for initial validation and further refinement. The very final version of their design will be integrated in the final System on Chip (SoC) demonstration, and its characterization will be again cross-correlated with the silicon measurements in the two previous versions, evaluating the effect of process/voltage/temperature corners and in both a controlled and real-world environment.

Program Plan

PROJECT MANAGEMENT STRUCTURE AND GOVERNANCE

Massimo Alioto will be the Lead PI and will coordinate the contributions from the PIs, and the interaction with the industrial and agency partners. The PIs will have monthly meetings to track the progress of the overall program. An **advisory board** will be formed to provide strategic directions (e.g., alignment with technology and Singaporean ecosystem), independent views and valuable criticism to the project. The board meets once a year (or more, upon need), and consists of the following members: Dr. MIN Kian Boon (Deputy Director, Singapore Ministry of Home Affairs), Dr. Tan Khen Sang (Senior Advisor Executive, Mediatek Singapore), Ma Mun Thoh (Senior Associate Director, NUS Industry Liaison Office), Dr. John Gustafson (CTO of Ceranovo, previously Director of Intel Labs, Santa Clara), Shengmei Sheng (Panasonic), Tang Min (Huawei R&D, Singapore)

The majority of students and staff at NUS will be in **closely-tied lab space and co-supervised** by PI and co-PIs, as facilitated by the spatial contiguity of the labs of PIs Alioto and Feng. The outcome of all research activities will converge on a **shared simulation/exploration/design server environment**, where updates on models, benchmarks, in-house software tools and chip design will be instantly available to all PIs. This will accelerate the progress beyond the coarse time granularity of meetings, and ensure cohesiveness. To facilitate teamwork, an internal **software collaborative environment** will be created to create a repository, a knowledge base for the entire team, and the medium to quickly share findings among PIs and share results over the web (e.g., publications, news, industrial engagement).

As summarized in the Gantt chart in Fig. D22, the **project plan** is organized around six main activities: the project launch (phase 0), four technical sub-projects (1-4), and a project control structure (5). Sub-project #1 is focused on the system-level view, from modelling to final system silicon demonstration. Sub-projects #2-4 are focused on the interactions within the CogniVision system: #2 covers the imager and the transceiver, along with their interaction with sensemaking, #3 covers embedded machine learning algorithms and their interaction with circuits, #4 covers the circuits/architectures for sensemaking and their interaction with the system through activity skipping and EQ scalability. Their interdependence and risk mitigation are summarized below.

0. Hiring, procurement, collaborative SW environment setup (led by M. Alioto)

Task 0.1. Recruitment of most of manpower is initiated before the start, and completed by Y1Q2.

Task 0.2. Procurement of essential equipment will be completed by Y1Q2.

Task 0.3. Collaborative software environment (e.g., MS SharePoint) is setup by Y1Q2.

1. System modeling, exploration, integration, demonstration (led by M. Alioto)

Task 1.1. System modelling environment is developed to support the selection of most promising techniques from sub-projects 2-4, and their preliminary architecture/system-level assessment

Task 1.2. 1st silicon stand-alone prototype of imager/transceiver and various sensemaking blocks are assembled on Printed Circuit Board (PCB) and tested for preliminary assessment of techniques

Task 1.3. 2nd silicon stand-alone prototypes are assembled on PCB for component assessment

Task 1.4. As preliminary work on SoC design, system is partitioned into modules, and mixed-signal simulation environment and verification flow are defined

Task 1.5. CogniVision SoC is designed/verified, integrating the imager/transceiver/sensemaking from T2.3-2.4 (D2.3 and refinement in T2.4), 4.4 (D4.4), and algorithms from Tasks 3.1, 3.2, 3.3

The main source of risk is posed by possible escaped design bugs that make the chip inoperable. This will be mitigated through testing ports to test/bypass any individual block in the SoC.

2. Energy-centric circuit techniques and interaction at imager-sensemaking and wireless-sensemaking boundary (led by K. S. Yeo)

Task 2.1. Imager/transceiver architectures explored for in-sensor processing, low-power radio

Task 2.2 Circuit-level aspects in T2.1 are investigated, 1st imager/transceiver prototype is designed

Task 2.3 Circuit-level issues arising in the prototype in Task 2.2 are addressed, leading to design/testing of a 2nd prototype for further refinement

Task 2.4 Final revision, verification and silicon demonstration will be conducted in the CogniVision SoC, solving timing and signal integrity aspects arising from integration

Some risk is in the delayed manufacturing due to delays in the foundry shuttle run, which will be mitigated by relying on multiple foundries, and choosing the one with more frequent shuttle runs.

3. Energy-centric machine learning-circuit co-design (led by J. Feng)

Task 3.1 Deep learning model compression is investigated, exploring new techniques to automatically locate redundancy, remove redundant connections and quantize the parameters

Task 3.2. Energy-aware deep learning networks are investigated in terms of design and training, introducing novel units, skipping connections, and including energy in the training loop

Task 3.3 Saliency front models to automatically detect informative frame regions are investigated, introducing gating functions to selectively pass salient regions to deep learning

Task 3.4 The deep learning models, design and training techniques in Tasks 3.1-3.3 are amalgamated with circuit-level aspects in Tasks 4.1-4.4 for energy-optimal integration on chip

Excessive accuracy drop is a possible risk, which is mitigated by explicitly managing the energy-accuracy tradeoff, and balancing model compression, redundancy clearing and quantization.

4. Irrelevant activity skipping/EQ-scalable sensemaking circuits/architect. (led by M. Alioto)

Task 4.1. Circuit- and architectural-level techniques to enable activity skipping in all blocks of sensemaking are investigated, modeled, verified and coordinated to minimize the overall energy

Task 4.2. Circuit- and architectural-level techniques to enable energy-quality scalability in all blocks of sensemaking are investigated, modeled, verified and coordinated to minimize energy

Task 4.3 Circuit- and architectural-level aspects in all blocks for sensemaking are investigated, and 1st silicon prototype is designed, manufactured (by silicon foundry) and tested to validate them

Task 4.4 Circuit-level issues arising in the prototype in Task 4.3 are fixed, and further across-block energy optimization is performed, leading to design/testing of a 2nd prototype for further refinement

Task 4.5 Final revision of sensemaking blocks, verification and silicon demonstration is conducted in the CogniVision SoC, solving issues arising from integration (e.g., timing, supply integrity)

The main source of risk is posed by possible escaped design bugs that make the chip inoperable. This will be mitigated through testing ports to test/bypass any individual block in the SoC.

5. Project control and reviews (led by Alioto)

Task 5.1. Annual meetings are held with the Advisory board to assess the progress of the project

Task 5.2. Mid-term review and meeting take place to assess if all models and fundamental components in their first silicon iteration have been successfully designed and tested

Task 5.3. Final review and meeting take place to assess if the research, models, methodologies and all components have come together as SoC with sub-mW power and targeted accuracy.

BUDGET DESCRIPTION AND JUSTIFICATION

The main expenditures are allocated to manpower (61%) and OOE (22%). OOE mostly covers the cost of silicon manufacturing, due to multiple tapeouts to de-risk design before the final system integration. The **total project value ($7.5M) above 5M$** is justified by the chip design-intensive nature of the project, whose credible demonstration requires integrated circuit design skills, substantial R&D effort, experimental validation. The manpower budget is (Total = S$3,931,920):

- *Sub-Project 1*
  - 1RF (yr 1-5) contributing to system-level aspects and integration → tasks {1.1-1.6}
  - 1RF (yr 1-5) contributing to system simulation and design → tasks {1.1, 1.5, 1.6}
  - 1 lab officer (1 day/week) for equipment/computers setup, management and monitoring
- *Sub-Project 2*
  - 1RF (yr 1-5) works on research on imager/transceiver circuit/architecture → tasks {2.1-2.5}

- 1RA (yr 1-5) contributing to imager → tasks {2.1, 2.2, 2.3, 2.4, 2.5}
- 1RA (yr 1-5) contributing to transceiver → tasks {2.1, 2.2, 2.3, 2.4, 2.5}

- *Sub-Project 3*
- 1RF (yr 1-3) research on deep learning models and saliency → tasks {3.1-3.3}
- 1RF (yr 3-5) research on deep learning training, benchmarking → tasks {3.3, 3.4}
- 1RA (yr 1-3) development of deep learning models and saliency → tasks {3.1-3.3}
- 1RA (yr 3-5) development of deep learning training, benchmarking → tasks {3.3, 3.4}

- *Sub-Project 4*
- 1RF (yr 1-5) contributing on architectural and system-level activity skipping/EQ → tasks {4.1-4.5}
- 1RA (yr 1-5) circuit-level optimization, verification of activity skipping/EQ → tasks {4.1-4.5}
- 1RA (yr 1-5) gate-level optimization, testing of activity skipping/EQ → tasks {4.1-4.5}

Research Scholarships

- 1RS (yr 1-4) on energy-autonomous integrated system modelling, design and optimization for real-time video processing → tasks {1.2, 1.3, 1.4, 1.5, 1.6}
- 1RS (yr 1-4) on energy-aware integrated circuit design for machine learning and real-time on-chip analytics → tasks {1.2, 1.3, 1.4, 1.5, 1.6}

Budget for Equipment (Total = S$374,120.66):

- GPU workstations/servers: deep learning network training, system simulations (Total = S$ 70 K)
- Measurement equipment for testchip characterization (Total = S$ 209,121) comprising National Instruments integrated equipment for timing characterization, testing, power characterization
- Racks and network switch for servers (Total = S$ 5 K)
- Servers for chip design, necessary for circuit simulation/design, 5 server blades are needed for 5 simultaneous designers (Total = S$ 75 K)
- Workstations: 5 workstations with monitors for 5 research staff (2 RF, 3 RA) (Total = S$ 15 K)

Other Operating Expenses (Total = S$ 1,402,500):

- Books/ebooks and journals: Books/ebooks and journals for research purposes (Total = S$ 2.5 K)
- Computer peripherals/accessories: computer accessories (external HD for backup, NAS, other peripherals for productivity, storage, etc.) are needed for ordinary needs (Total = S$ 7 K)
- Consumables: materials&consumables, postage, photocopying for ordinary tasks, printer cartridges, photocopies, document exchange (Total = S$ 12.5 K)
- License for CAD tools for chip design: CAD software tool licenses (e.g., Cadence, Synopsys, Mentor Graphics) for circuit design exploration, simulation, design and verification, as well as the integrated demonstrators. Licenses will be shared across the PIs (Total = S$ 150 K)
- Local conferences/workshops/seminars: registration fees for scientific events (Total = S$ 5 K)
- Maintenance fees: cost of equipment recalibration or fix (Total = 10 K)
- Printed Circuit Board fabrication, chip packaging, miscellaneous electronics (Total = S$ 40.5 K)
- Publication fees (Total = S$ 10 K)
- Silicon manufacturing for chip prototyping: testchip fabrication in CMOS technology (targeted: 28 nm). Two rounds of prototyping are needed for imager/transceiver and sensemaking (S$ 200K/tapeout in 28mm2). Merge into final chip takes 1.5X the area of each (Total = S$ 1,100 K)
- Visiting professors (collaborators) (Total = S$ 60 K for 3 months, salary of $20 K/month)
- Software license, cloud services for collaborative environment (e.g., SharePoint, Total = S$5 K)

Overseas Travel (OT, Total = S$174 K): PIs will travel to conferences and visit collaborator.

The budget is strengthened by the additional **contribution from industrial partners**: Mediatek (S$50 K, see letter of commitment) and Panasonic (S$600 K, see letter of commitment).

Role of team members

The role of the PIs and their expertise are summarized below, along with the areas of the project that they will interact on. The **Industrial interactions of team members** are in Table V.

Prof. Massimo Alioto, lead PI, is a leader in the area of energy-efficient integrated circuit design, holding numerous worldwide records in the field (see group website). His Green IC research group tapes out 10+ chips a year (14 last year) to prove new concepts and ideas in the area of low-power chip design. As relevant to this project, he has pioneered energy-quality scalable integrated circuits (see Sept 2018 IEEE JETCAS special issue, led by him), and has worked with the two academic groups (UCBerkeley, University of Michigan) that first demonstrated millimeter-sized integrated systems with nearly-perpetual operation. He is also active in the IoT area, with 250 publications overall, 50 talks in the last 5 years, and the first book on chip design for IoT. Prof. Alioto is Deputy Editor in Chief of two IEEE journals (TVLSI, JETCAS), ISSCC TPC member, and IEEE Fellow for "contributions on energy-efficient circuits". Leveraging on his expertise, Prof. Alioto will lead sub-project #1 and #4, and join sub-projects #2 to create a well-coordinated interaction of sensing/wireless/sensemaking, and #3 for deep learning architecture-algorithm interaction.

Prof. Yeo Kiat Seng, co-PI, is a widely known authority in low-power RF/mm-wave IC design, and on image sensors more recently. He is author of 600 publications, 7 books and holds 38 patents. He is currently the Associate Provost for Graduate Studies at the Singapore University of Technology and Design, and member of Board of Advisors of the Singapore Semiconductor Industry Association. He was previously Head of Division of Circuits and Systems and Founding Director of VIRTUS of the School of Electrical and Electronic Engineering at NTU Singapore. Prof. Yeo holds or has held key positions in many international conferences as Advisor, General Chair, Co-General Chair and Technical Program Chair. He was awarded the Public Administration Medal (Bronze) on National Day 2009 by the President of the Republic of Singapore. Prof. Yeo is an IEEE Fellow. He will lead sub-project #2 on image sensors and wireless communications.

Dr. Feng Jiashi, co-PI, has rich research experience with computer vision, machine learning (including deep learning). His Learning and Vision research group (20+ people) has published over 60 papers on machine learning and computer vision in the past 5 years. Dr. Feng received the winner award for emotion recognition in the wild challenge 2016, best paper prize from TASK-CV with ICCV'2015 and best technical demo prize from ACM MM'2012. He served as the technical program chair for ACM ICMR'2017 and area chair for ACM MM'2017. Dr. Feng will lead sub-project #3 on energy-centric machine learning-circuit co-design. Dr. Feng will also collaborate with Prof. Alioto and Prof. Benini on the deep learningalgorithm-architecture interaction.

The **collaborator Prof. Dennis Sylvester** is a prominent researcher in the field of energy-efficient circuits and has demonstrated the imaging system with lowest power to date. He has a stable collaboration with Prof. Alioto since 2011, as documented by several joint publications on ultra-low energy processing and sensing systems, and research staff exchange. Prof. Sylvester will contribute to sub-project #1 and #4 on the across-layer integration of multiple algorithms on silicon.

The collaborator Prof. Chen Shoushun, was the Program Director of Smart Sensors, under VIRTUS, IC Design Centre of Excellence, NTU (Singapore). He leads a Smart Sensors group, aiming to investigate smart sensory systems, combining new circuit architectures and energy-efficient signal processing algorithms. He is currently on one-year leave from NTU to lead a start-up company developing innovative image sensors, which has been spun off from his research effort at NTU. His team has designed 30+ CMOS image sensors, one of which was launched in space in the VELOX-I nanosatellite in 2014. Prof. Chen will contribute to sub-project #1 and #2 (and #4 to a minor extent), in particular on aspects related to imager sensors.

The **collaborator** Prof. Luca Benini is Professor at ETH Zurich (Switzerland), and a worldwide leader in energy-efficient computer and specialized deep learning architectures. He has served as Chief Architect for the Platform2012/STHORM project in STmicroelectronics in 2009-2013. He has published more than 700 papers and 4 books. He is a Fellow of the IEEE and the ACM, and a member of the Academia Europaea. Prof. Benini will contribute to sub-project #1 and #3, and in particular on architectural aspects related to deep learning.

The team will collaborate with **industrial partners and agencies supporting various aspects of the project**, from in-kind contribution of 0.7M$ in terms of silicon manufacturing support, to real-world datasets, domain expertise and hardware/cloud services for large-scale computation (see letters of support). Their support assures relevance to industrial interest, and alignment with the fast-changing landscape of distributed sensing. Industrial partners cover the key areas that the proposal aims to make an impact on. Mediatek is a leading company in low-energy integrated systems for mobile platforms and IoT. Panasonic is a well-known leader in distributed vision and imaging, among the other fields. The Singaporean Ministry of Home Affairs is also a key project partner, with strong testbedding and deployment capabilities and domain expertise. All industrial collaborators are physically located in Singapore.

Outcomes & Deliverables (see Gantt chart in Fig. D22 in Annex D)

Year 1

M0.1 (t_0+6 months). Complete recruitment, requisition of major equipment, software environment

M2.1 (t_0+1 year). Definition of imager and transceiver architecture, and modelling

M3.1a (t_0+1 year). Deep learning compression (50x smaller, <10% accuracy drop w.r.t Table IV)

M5.1 (t_0+1 year). Internal review meeting with project Advisory Board

Year 2

D1.1 (t0 + 2 years). Completion of system simulation/model framework and related software

D2.2 (t0 + 2 years). Imager and transceiver (round #1) chip tape out (VGA, 30 fps)

M3.1b (t_0+2 years). Deep learning compression (200x smaller, accuracy drop <5% w.r.t. targets in Table IV)

D4.3 (t_0+2 years). Sensemaking chip tapeout (round #1) with <400μW feature extraction, <400μW novelty assessment, 2mW deep learning at full AlexNet activity, SRAM with 70% activity reduction

M5.2 (t_0+2 years). Internal review meeting with project Advisory Board

Year 3

M1.2 (t_0+3 years). Completion of testing of PCB-assembled components (round #1)

M2.2 (t_0+3 years). Completion of imager characterization and demo (round #1)

D2.3 (t_0+3 years). Fine-tuned imager and transceiver (round #2) chip tape out

D3.1 (t_0+3 years). Demo on deep learning compression with >1,000x smaller size, and accuracy drop <2% w.r.t. targets in Table IV

D3.2 (t_0+3 years). Demo on deep learning with >10x power reduction w.r.t. state of the art

D3.3 (t_0+3 years). Demo on saliency detection with 10x less computational cost and accuracy drop less than 2% for targets in Table IV

M4.1 (t_0+3 years). Completion of exploration of activity skipping and architectures/circuits definition

M4.2 (t_0+3 years). Completion of exploration of activity skipping and EQ-scalable circuits

M4.3 (t_0+3 years). Demo of sensemaking chip tapeout (round #1) with 200μW feature extraction, 200μW novelty assessment, <1mW deep learning at full activity, SRAM with 70% activity reduction

D4.3 (t_0+3 years). Sensemaking chip tapeout (round #2) with <150μW feature extraction, <150μW novelty assessment, <1mW deep learning at full AlexNet activity, SRAM with 70% activity reduction

M5.3 (t_0+3 years). Internal review meeting with project Advisory Board

M5.6 (t_0+3 years). Mid-term review (see quantitative targets below)

Year 4

M1.4 (t_0+4 years). Completion of SoC partitioning, floorplan, simulation/verification environment

M2.3 (t_0+4 years). Completion of imager characterization and demo (round #1)

M4.4 (t_0+4 years). Demo of sensemaking chip tapeout (round #2) with <150μW feature extraction, <150μW novelty assessment, <1mW deep learning at full AlexNet activity, SRAM

M5.4 (t_0+4 years). Internal review meeting with project Advisory Board

Year 4

D1.5 (t_0+4.5 years). CogniVision final SoC chip tapeout with power targets as in Table IV

M1.6 (t_0+5 years). Final characterization and demo of CogniVision SoC chip (power as in Table IV)

M2.4 (t_0+4.5 years). Completion and tapeout of final imager/transceiver for system integration

M2.5 (t_0+5 years). Completion of characterization of sensing part of CogniVision and demo

M3.4 (t_0+5 years). Completion of in-field testing of deep learning (see Table IV) and demo on models with reduced size (>1,000X), <2% accuracy drop w.r.t. targets in Table IV

M4.4 (t_0+5 years). Completion of characterization of sensemaking part of CogniVision and demo

M5.5 (t_0+5 years). Internal review meeting with project Advisory Board

M5.7 (t_0+5 years). Final review (see quantitative targets below and Table IV)

Milestones at the mid-term review (end of year 3): chip demo of feature extractor with 100 μW power, novelty assessment engine with 100 μW power, deep learning engine with <1mW power (see Table IV), imager with 100 μW power at VGA resolution, 30 fps and activity rate of average NeoVision2 benchmark. Deep learning model with 1,000X reduced size with <2% accuracy degradation in face and object detection, compared to targets in Table IV.

Milestones at the completion of the program: system on chip demonstration of a complete cognitive camera (from sensor to sensemaking) with average power consumption in the order of 1 mW in the three use cases in Table IV (see also power targets). An international workshop will be held at the end of the program and co-located with a leading IEEE conference.

IMPACT OF THE RESEARCH TO SINGAPORE

The success of CogniVision will provide a **unique technological competitive advantage**, in view of the demonstration of the first camera chip with nearly-perpetual operation, fully untethered, energy-harvested, millimeter-sized, capable of on-chip real-time sensemaking, low cost ($ range). The on-chip sensemaking also fundamentally solves the challenges of data delouge and privacy, which are currently faced with distributed (tethered) cameras. Accordingly, CogniVision **accelerates the Smart Nation vision**, and contributes to make Singapore a global hub for IoT sensing technologies, and in particular **high added-value technologies** such as visual sensing.

To reach the intended impact, **local enterprises** working on or using distributed sensors (e.g., belonging to the recently formed IoT Consortium of the Singapore Semiconductor Industry Association (SSIA)), will be engaged during the project via demonstration in our labs (end of year 3). On a global scale, the Embedded Vision Alliance [EVA] will be engaged at the end of year 4 to reach out to leading companies in image sensing applications. These companies can indeed be technological or venture partners in the successive translation of CogniVision into a commercial technology. The support of agencies is key to the success of the project (see letter of support from Singapore MHA), as Singapore is a natural testbed for CogniVision, and will benefit from the introduction of ubiquitous vision capability in the Smart Nation vision (see alignment in Fig. D23). Their expertise will facilitate alignment with compelling applications and use cases.

At the end of the project, a **workshop** will be organized to share findings and to demonstrate the outcomes of CogniVision. To make our technologies widely available, we will consider the **opportunity of spinning off a company based in Singapore for commercialization of CogniVision**. The CogniVision project will leverage the **synergy with local industry** in the IoT space, starting from the project industrial partners, which cover the key areas related to CogniVision, i.e. system integration (Panasonic) and chips for IoT (Mediatek). As key factor that promises significant impact of CogniVision is the **relevance to a very wide range of diverse applications and verticals**, ranging from consumer to security, smart cities, industry, and others.