This is the Pre-Published Version.
This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use (https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms), but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1038/s41928-020-00501-9.

1 # Near-sensor and in-sensor computing

2 Feichi Zhou[1] and Yang Chai[1, *]

3 [1]Department of Applied Physics, The Hong Kong Polytechnic University, Hung Hom,
4 Kowloon, Hong Kong, P. R. China

5 *Corresponding author: E-mail: ychai@polyu.edu.hk (Y.C.)

6

7 ## Abstract

8 The number of nodes in sensory network is rapidly growing, and generates a huge
9 amount of redundant data, which forces frequent data exchange in sensory terminals
10 and computing units, occupies large computation, memory and communication
11 resources, and consumes substantial energy during data transfer. To efficiently process
12 massive data and decrease power consumption, it is quite necessary to develop new
13 computing paradigms that are close to or inside sensory networks, and can reduce the
14 redundant data movement between sensing and processing units for enhancing speed,
15 power efficiency and integration density. Here we propose near-sensor and in-sensor
16 computing paradigms for moving part of computation tasks to sensory terminals. We
17 classify their functions into low-level and high-level processing, and discuss the
18 implementation of near/in-sensor computing for different physical sensing systems. We
19 also analyse the existing challenges and provide possible solutions for hardware
20 implementation of integrated sensing and processing units using advanced
21 manufacturing technologies.

22

23 The ever-increasing and ubiquitous sensor nodes distributed in the Internet of Things
24 generate large volume of data, which are continuously growing at a rapid pace. The
25 number of sensory nodes is predicted to reach 75 billion by 2025 and surge to 125
26 billion by 2030.[1] In these sensory nodes, a large portion of generated raw data are
27 unstructured and redundant. In typical designs, the sensory systems are physically
28 separated from computing units because of different requirements and manufacturing
29 technologies of sensing and computation units, where sensing functions are realized in
30 noisy analogue domain, and computing is usually executed in digital format with von
31 Neumann architecture. As a result, voluminous quantities of raw data are locally
32 acquired from sensor terminals, and transferred between sensing and computation units
33 or cloud computing, which poses significant challenges for energy consumption,
34 response time, data storage, communication bandwidth, and security.

35 The data proliferation from ubiquitously distributed sensors gives rise to massive
36 increases in information processing demands, especially for the sensor-rich platforms
37 (*e.g.*, intelligent vehicles, autonomous and micro robots, mobile medical, wearable
38 electronics) and the applications with strict delay requirements (*e.g.*, real-time video
39 analysis, cooperative autonomous driving), and makes the computing architecture more
40 data-centric instead of computation-centric. This data-centric computing paradigm has
41 a number of new characteristics that are dramatically different from conventional
42 computing, and thus demands new computational paradigms, which in turn require new
43 hardware platforms to match such characteristics for achieving high performance and
44 energy efficiency. To process information more efficiently using the same or less power,
45 we need to develop new computing paradigms to shift some of the computational tasks
46 close to the sensory devices, which reduces redundant data movement, and generates,
47 collects and consumes the data locally. [2]

48 Near-sensor and in-sensor computing paradigms emerge as effective alternatives for
49 efficient sensory data processing by residing computing units at sensor endpoint or
50 equipping sensors with computing capabilities to reduce or eliminate data transfer and
51 conversion at the sensor/processor interface. **Fig. 1** illustrates the comparison of
52 conventional sensory, near-sensor and in-sensor computing architectures. In
53 conventional architecture, the analogue sensory data are firstly converted to digital
54 signal through analogue-to-digital conversion (ADC), temporally stored in a memory
55 unit, and then fetched from memory to processing units, in which the data transmission
56 and conversion result in inefficient power use and high latency. In the **near-sensor**
57 **computing** architecture, the processing units or the accelerators locate beside sensors
58 to execute specific computational tasks at the sensor endpoints, which can resolve the
59 bottlenecks between sensors and processors by optimizing sensor/processor interface,
60 minimizing data transfer and conversion, and reducing redundant data. In the **in-sensor**
61 **computing** architecture, individual self-adaptive sensor or multiple connected sensors
62 can directly process sensory information, which further eliminates the sensor/processor
63 interface and combines sensing and computing functions.

64 In this perspective, we will illustrate the functions of near/in-sensor computing
65 paradigms, introduce their design strategies, architectures, and representative examples,
66 and identify key challenges and future research directions. We will also provide
67 possible ways for the hardware deployment of the near/in-sensor computing
68 architectures.

69 **Near/in-sensor computing for low-level sensory processing**

70 Integrated sensory/computing systems are featured with hierarchical and
71 feedforward nature, ranging from low-level sensory information to high-level abstract
72 representation. Low-level information processing involves selectively encoding

73 spatiotemporal features from unstructured sensory signals and provides essential
74 information to complicated circuits for high-level processing. At this stage, the outputs
75 of low-level processing are still the representations of the sensory signals. The low-
76 level processing can preliminarily and selectively extract useful data from large volume
77 of raw data by suppressing unwanted noise or distortion, or enhance the feature for
78 further processing, which are important processing steps in data-intensive applications.
79 The low-level sensory processing, such as noise suppression (filtering), background
80 extraction, feature enhancement, motion extraction, *etc*., can effectively reduce the
81 computational load and improve the efficiencies for high-level processing tasks, and
82 can serve as the interface between sensing and other high-level processing units,
83 enabling parallel and real-time processing for delay-sensitive applications. Low-level
84 sensory processing units typically include sensor arrays, readout circuits, ADCs and
85 processing units. The sensor arrays are usually connected with clock multiplexed
86 circuits, and one processor handles all the sensory data in series. **Fig. 2** schematically
87 illustrates low-level processing for visual, auditory, and olfactory signals before and
88 after processing.

89 Among various sensors, image sensors can be fabricated with CMOS-compatible
90 process over a large scale. With the increase of image pixel and frame rate, the image
91 processing has become a typical data-intensive computing. Low-level image processing
92 involves edge and contrast enhancement, noise reduction (Fig. 2a). The image
93 processing with pulse-domain based algorithms only requires simple logic operation
94 and circuit implementation.[3,4] A vision chip by planar system-on-chip (SoC) integration
95 consists of photodiode arrays, pulse frequency modulation circuits, and simple 1-bit
96 ADC. Compared with conventional spatial filtering algorithms, *e.g.*,
97 Gradient/Laplacian methods or Gaussian filtering (for edge enhancement) and

98 histogram processing methods (for contrast enhancement) that require complex circuits
99 with adders and multipliers,[5,6] the pulse domain algorithm eliminates the use of adders
100 or multipliers, thus reducing circuitry complexity and improving the fill factor.

101 Compared with digital processing, analogue computing can directly process analogue
102 signals without ADC. Conventional sensors usually compute a linear function of signal
103 intensity, while analogue processing circuits connected with the sensors can perform
104 nonlinear mapping functions, spatiotemporal filtering, and adaption. For example,
105 adaptive image sensors can employ logarithmic output respect to light illumination,
106 making the image contrast independent on background change. Additionally, they can
107 serve as filters to output a low gain for static and low-frequency stimuli, and a high gain
108 for transient and high-frequency stimuli. [7-9]

109 Emerging sensors can be used for low-level image processing with in-sensor
110 configuration. Different from near-sensor approach that alters the output using external
111 circuits, the sensors *in-situ* respond to external stimuli and output different
112 characteristics. We designed and demonstrated an optoelectronic memory for
113 neuromorphic vision sensor with both light-intensity-dependent and time-dependent
114 plasticity, which allows to directly perform low-level processing for analogue sensing
115 data. We utilize self-adaptive characteristics of the sensor, which adaptively reduces
116 the amplitude in dark pixel, and retains the features in bright pixel. In this approach, the
117 sensor array presents *in-situ* image pre-processing, including image contrast
118 enhancement and background smoothing, which is also proved to improve image
119 recognition efficiency. [10] Wang *et al.* developed a vision sensor with gate-tuneable
120 positive (ON) and negative photo-responses (OFF) to emulate the characteristics of
121 bipolar cells in human retina. The reconfigurable sensor array is constructed to extract
122 the edges in the image through the combination of excitatory and inhibitory interactions

123 between neighbouring pixels.[11] The in-sensor configuration can greatly simplify the
124 circuitry of sensory processing and eliminate the sensor/processor interface.

125 The low-level auditory processing includes frequency decomposition, noise
126 suppression and signal enhancement, which are vital to extract clean signals for
127 subsequent high-level processing (Fig. 2b). Similar to near-sensor analogue processing
128 in visual signals, researchers adopted analogue circuits near auditory sensor to emulate
129 the functions of cochlea.[12-14] Multiple auditory sensors are sampled and connected with
130 analogue spatiotemporal and adaptive bandpass filtering circuits for frequency
131 decomposition and noise suppression. [12,13,15] Spike-coding-based processing with
132 address event representation can benefit real-time and event-based auditory
133 processing.[16] The sensors and processing circuits are usually integrated through printed
134 circuit board (PCB) or planar SoC. [16,17]

135 For olfactory sensing, an essential step is to cancel the DC baseline in a
136 heterogeneous chemosensor array (Fig. 2c), which usually has a large variation in
137 baseline among different types of sensors. Olfactory chips have been fabricated with
138 planar SoC integration, in which the olfactory sensors are connected with adaptive
139 circuits for baseline cancellation. [18] The adaptive elements enable the sensors to be self-
140 adapted within a working range of the circuit for different odours.

141 For other low-level sensory processing, thermal artificial nociceptor was
142 demonstrated by connecting a thermoelectric module to a diffusive memristor with
143 threshold switching, which can respond to damaging stimuli by sending “warning”
144 signals. [19] The low-level processing, *e.g.*, base cancellation, filtering and noise
145 suppression, can be also extended to other sensory processing, such as
146 electroencephalography, electrocardiography, tactile sensing, *etc*.

147 In-sensor computing allows feature enhancement through self-adaptive
148 characteristics of individual sensors, while near-sensor computing processes sensory
149 signals by transferring data to adjacent computation units with short distance. In term
150 of circuitry complexity, the in-sensor computing design is simpler, but it is restricted
151 by limited functions and specific application scenarios.

152 **Near/in-sensor computing for high-level sensory processing**

153 Low-level sensory processing is responsible for optimizing the features in raw and
154 unstructured data that are difficult to be identified; while it also requires high-level
155 processing for abstract representation of sensory data, *e.g.*, recognition, classification,
156 and localization (**Fig. 3a**). High-level sensory processing involves the cognitive
157 processes that enable to identify "what" or "where" of the input signals.

158 The accelerators based on deep neural network (DNN) and convolutional neural
159 network (CNN) have been extensively used for image/speech recognition or
160 classification. The accelerator efficiency is limited by dynamic random-access memory
161 (DRAM) accesses for inputs and outputs. Several near-sensor approaches have been
162 proposed for efficient processing with optimized sensor/accelerator interface. A near-
163 sensor CNN accelerator for image recognition (ShiDianNao) can dramatically reduce
164 the energy costs and shift processing close to the sensor. [20] In this approach, all the
165 shared weights are directly stored in small on-chip SRAM, exhibiting 60× more energy
166 efficient and approximately 30× faster than previously reported neural network
167 accelerator.[21] However, the accelerator based on digital processing units restricts the
168 performance because of the ADC interface.

169 The convolutional operation can be directly implemented at the sensor endpoint
170 without ADCs for matrix-vector multiplication (MVM) and multiply-and-accumulation

171 (MAC) operations, which can accelerate the computing and reduce the workload of
172 ADC, because convolutional operations consume substantial computation resources
173 and dominates the running time. [22-24] In RedEye design[24], MVM and convolutional
174 operations are implemented in analogue domain through charge-sharing tuneable
175 capacitor. The energy consumption per frame is 44.3% and 45.6% lower than GPU and
176 CPU, respectively. However, the sensors are required to be sampled first without real-
177 time readouts. In addition, the sensors and processors are integrated on PCB. The MVM
178 operations can be implemented with analogue memory synaptic arrays near the sensors,
179 corresponding to the synaptic plasticity in the neural network [25-27] (**Fig. 3b**). The current
180 through a memory is the multiplication of voltage and conductance following Ohm's
181 law; and the resulting currents are summed along a row or column through Kirchhoff's
182 law.

183 Real-time detection and learning for colour-mixed image recognition can be realized
184 by connecting h-BN/WSe_2 photodetector and h-BN/WSe_2 synaptic transistor in series.[28]
185 The artificial synapse presents distinguishable synaptic weights and plasticity under
186 lights with different wavelengths. An optical neural network is further constructed for
187 colour-mixed image recognition. However, the configuration still lacks large-scale
188 integration and completed processing tasks. Spiking neural network (SNN) provides
189 another promising solution to enhance the efficiency by processing time-encoded
190 neural signals in parallel. **Fig. 3c** shows an illustration of near-sensor SNN based on
191 memory synaptic array through spiking-time-dependent plasticity (STDP) learning rule.
192 A near-sensor architecture with image sensors, CMOS neuron circuit and memory array
193 was integrated on PCB for image recognition. [29]

194 To further accelerate the hardware implementation of deep learning algorithms,
195 reconfigurable sensor arrays can be constructed for efficient in-sensor MAC operation

196 (**Fig. 3d**). For a sensor array with $m \times n$ sensor elements, the stimuli to sensory elements

197 can be represented as $\boldsymbol{S}$ vector, $\boldsymbol{S} = (S_1, S_2, \ldots, S_m)$, and $\boldsymbol{R_{mn}}$ is the responsivity matrix of

198 sensory array. The output vector $\boldsymbol{I}$ can be expressed as:

199
$$\mathbf{I} = \boldsymbol{R} \times \boldsymbol{S} = \begin{bmatrix} \mathrm{I}_1 \\ \mathrm{I}_2 \\ \vdots \\ \mathrm{I}_n \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{bmatrix} \begin{bmatrix} \mathrm{S}_1 \\ \mathrm{S}_2 \\ \vdots \\ \mathrm{S}_m \end{bmatrix} \quad (1)$$

200 Depending on the type of external stimuli (light, pressure, chemicals, electromagnetic

201 field, *etc*), the responsivity $\boldsymbol{R}$ can be variously physical parameters. The multiplication

202 of stimulation and responsivity occurs at an individual sensor through sensing process,

203 and the resulting currents are summed along interconnected sensory elements through

204 Kirchhoff's law. The summation of all the currents produced in each sensor element

205 can be expressed as:

206
$$I_n = \sum_{m=1}^{m} I_{mn} = \sum_{m=1}^{m} R_{mn} S_m \quad (2)$$

207 The responsivity of individual sensor is designed to be modulated and updated through

208 external modulation, which emulates the change of synaptic weight ratios in neural

209 network during the learning process.

210 The in-sensor computing architecture greatly simplifies hardware design, and can

211 effectively perform high-level information processing, *e.g.*, classification, recognition,

212 and autoencoding. Researchers used conventional semiconductors or emerging two-

213 dimensional (2D) materials to construct a reconfigurable photodiode array for in-sensor

214 image processing. [30,31] In the example of a sensory neural network based on 2D

215 semiconductors, one pixel consists of three interconnected photodiodes as subpixels.

216 Both photo-sensing and MAC operation can be realized in the sensor array. The polarity

217 and magnitude of photoresponsivity of WSe_2 photodiodes can be modulated by

218 applying positive or negative gate voltage. Thus, the weight ratios can be tuned during

219 the training process. With the rational design of in-sensor computing architecture, an
220 ultrafast image recognition within nanosecond level can be realized by eliminating the
221 interface between sensor and computing units. [30]

222 Olfactory sensors and neuromorphic circuits can be integrated on a planar SoC chip.
223 [18,32] The implementation of spiking neural network (SNN) STDP learning enables
224 simple odour classification tasks. The synapse array based SNN implementation
225 provides an effective method to simplify the learning circuits and improve the energy
226 efficiency, which can be further adopted in the near-sensor olfactory processing. A
227 large number of olfactory sensors can be integrated with processing units in a three-
228 dimensional (3D) configuration with shorter interconnect length, compared with planar
229 SoC integration.[33]

230 Tactile sensing can be processed with near-sensor computing architectures for
231 perceptual learning and recognitions.[34-36] Pressure sensors are integrated with artificial
232 neuron and synaptic devices to demonstrate an artificial spiking afferent nerve with
233 learning capabilities.[34] The pressure-dependent spiking-rate-dependent plasticity
234 (SRDP) and weight updating enable the feature learning, which can be further
235 employed in SNN for touched pattern recognition and movement detection. This
236 configuration provides potential hardware design for near-sensor computing with
237 simplified and efficient circuitry.

238 Sound localization and speech recognition are two primary functions for high-level
239 auditory processing. The neural network based on resistive switching memories can be
240 used for efficient auditory localization and recognition. The SNN implementations
241 exhibit learning capabilities of spatiotemporal patterns. The localization is determined
242 by the interaural time difference between left and right ears. [37] The neural network
243 based on resistive switching memories can be potentially further integrated with

244 auditory sensors for efficient sound localization. In addition, the near-sensor
245 architectures based on neuromorphic computing devices for both olfactory and auditory
246 processing remain small-scale demonstration with a few devices and lack large-scale
247 integration for the complete neural network.

248 **Integration technology of near/in-sensor computing**

249 The manufacturing of sensory and computing units relies on different materials
250 systems, device structures, circuit design and processing technologies. To implement
251 the near/in-sensor computing systems, heterogeneous integration of sensory and
252 computing units is a practical way for the hardware architectures with high performance
253 and high integration intensity. The integration on PCBs presents long distance between
254 sensing and processing units and low integration density. To meet the demand of low-
255 power and high-speed near/in-sensor computing architecture, it requires new
256 integration technologies for sensing and processing units. Depending on the physical
257 length between sensor and processing elements, there are a few types of integration
258 technologies. In-sensor computing architecture performs information processing inside
259 the sensory networks, showing zero physical distance between sensing and computing
260 elements. For near-sensor computing, the integration approaches include 3D monolithic
261 integration, planar SoC, 3D heterogeneous integration, 2.5D heterogeneous integration,
262 *etc*. In these technologies, the physical length between sensor and computation
263 functions ranges from hundreds nanometre to millimetre scale.

264 3D monolithic integration adopts microfabrication process to construct on-chip
265 interconnects between different device layers on a single substrate, which offers high
266 integration density and more communication bandwidth. The monolithic system
267 consists of functional layers (sensor, memory, computing and communication, *etc*) in a
268 3D stacked configuration, where each layer is connected through local inter-layer vias

269 with the length of tens to hundreds of nanometres. Short interconnects help to greatly
270 reduce parasitic resistance-capacitance time delay and power consumption, and largely
271 increase integration density. Shulaker *et al.* demonstrated a 3D monolithic chip for gas
272 sensing, data storage and processing, in which they achieve the connectivity more than
273 1000 times greater than the conventional 3D stacking chip by through-Si-vias (TSVs).
274 [33] The sensors on the top layer can collect data and pass to the underlying memory layer.
275 A classification accelerator with carbon nanotube field-effect transistor conducts a pre-
276 identification among ambient vapours. This architecture enables sensory data to be
277 directly stored into memory in parallel and perform processing at high speed.
278 Conventional semiconductor process usually relies on epitaxy at high temperature,
279 which greatly restricts the selection of materials systems and processing technologies
280 for the 3D monolithic integration. Thus, it is important to develop new materials and
281 devices with low-temperature process, reasonably high performance, and compatibility
282 with existing process technologies for 3D monolithic integration.

283 Planar SoC with both sensing and processing units on a single substrate has been
284 extensively applied for near-sensor computing. Various types of sensor chips (visual,
285 olfactory, and auditory) have been demonstrated with planar SoC technology to
286 perform feature extraction and learning functions for high-level processing. [18,38] Most
287 of those chips have a size of millimetre and interconnect length of tens of microns. In-
288 sensor computing is a promising alternative approach in a single planar chip, which can
289 simplify the circuits and allow high integration density.

290 In 3D heterogeneous integration scheme (**Fig. 4c**), sensing and processing units can
291 be built with different manufacturing processes on different wafers, and are integrated
292 with advanced packaging technologies (*e.g.*, TSVs, die-to-die interconnects), which
293 allows to combine incompatible manufacturing on a single chip with relatively larger

pitch size (a few micrometres to tens of micrometres) compared with monolithic 3D integration.[39,40] Researchers have integrated CMOS image sensors with accelerator for feature extraction and complete CNN for image classification, respectively. [41] The chip stacks sensor, memory and computing layers in a vertical structure connected with TSV. The 3D heterogeneous integration technology can be also extended to flexible electronics.[42,43] The relatively long interconnect length limits the performance of 3D stacked chip.

2.5D packaging, also called Chiplet (**Fig. 4d**), is one of special heterogeneous integration, which enables the integration of different chips side-by-side on silicon, glass or organic interposer with TSV or redistribution layer technologies. All functional circuit blocks (Chiplets) are integrated and mounted on an interposer with <1 mm separation for high-speed communication, as a midpoint between planar SoC and 3D heterogeneous integration. Chiplets are connected through interposer with shorter interconnects (<1 mm) and fine pitch (< 4 μm) interconnects.[44] It allows to integrate disparate technologies (*e.g.,* sensor, memory, logic, and communication) with short development period, low cost, low design complexity and chip failure risks. Researchers have demonstrated heterogeneous integration for neural sensing with front-end low-level processing. [45] Interposer carries various circuit chips, which can increase the system functionality and present a potential integration technology for near-sensor computing. More importantly, Chiplet technology can be relied on existing matured chips, exhibiting great advantages of reduced fabrication/design cost, reasonably high performance, and short time-to-market.

Outlook

Near/in-sensor computing is an interdisciplinary research, covering materials, devices, circuits, architectures, algorithms, and integration technologies. Compared

319 with near/in-memory computing paradigm, the computing in or near sensors is more
320 complex, because it needs to handle large amount and various types of signals in
321 different scenarios. Successful deployment of near/in-sensor computing will need co-
322 development and co-optimization of sensors, devices, integration technologies, and
323 algorithms.

324 ***Multi-modal sensors***

325 The performance of conventional sensor is normally evaluated through sensitivity,
326 response time, dynamic range, error tolerance, *etc*. While the sensory devices in near/in-
327 sensor computing paradigm are required to have self-adaptation and self-identification
328 characteristics for efficiently processing the information in combination with
329 algorithms. These sensor devices cannot only communicate information across the
330 sensor network, but also cooperate together to perform more complex tasks, like signal
331 processing, data aggregation and compression. The development of these intelligent
332 sensor devices requires the innovation of device physics and sensing mechanisms.[10]
333 Current investigations mainly focus on single type of sensory processing. Human
334 perception system can simultaneously sense and process different types of information
335 in a very small perceptive field and complex environments. Therefore, it is highly
336 desirable to develop intelligent devices and systems for the fusion of different sensory
337 processing in a real-time manner, including visual, auditory, olfactory, tactile, *etc.* [46]
338 These integrated systems can benefit future applications, such as robotics, intelligent
339 vehicles, wearable electronics, *etc*. It requires a dynamic hardware reconfiguration of a
340 sensor node in a single chip to accommodate a particular sensing method or a universal
341 hardware platform that can adaptively fit to different sensor conditions, and algorithms
342 for the same platform.

343

344 ***Computing devices***

345 For computing devices, the hardware needs to intimately work together with
346 algorithms. Conventional Si CMOS electronics cannot exhibit high efficiency for
347 neural network algorithms because of their intrinsically digital characteristics.
348 Emerging neuromorphic computing devices, such as two-terminal resistive switching
349 memories with analogue multiple resistance states, tuneable plasticity high symmetry
350 and linearity, high speed, low operation energy, small footprint and high stackability,
351 are regarding as promising candidates for hardware implementation of artificial neural
352 network and executing in-memory computing for cognitive tasks (*e.g*., object
353 recognition, association, adaptation, and learning, etc.). [47-49] The design of computing
354 sensors can be utilized for in-memory computing by further integrating sensing
355 functions in these devices.

356 ***Processing and materials***

357 Disparate manufacturing processes raise grand challenges for the integration of
358 sensing and computing units for near/in-sensor computing architectures. Both 3D
359 monolithic and heterogeneous integration are involved with multiple functional
360 layers/chips and different materials. To avoid adverse effects of high-temperature
361 process on the functionalities of existing devices, we need to employ low-temperature
362 process for high reliability of the integrated systems. In 3D stacked chips, high built-in
363 stresses raise reliability issues, thus requiring the development of highly reliable low-
364 temperature bonding and interconnect process to minimize the coefficient of thermal
365 expansion mismatch between stacked chips. To reduce the parasitic time delay, it is
366 also necessary to decrease the thickness of active devices and passive components. One-
367 dimensional carbon nanotubes and 2D layered semiconductors with ultrathin body have
368 been successfully transferred onto arbitrary substrates at low temperature. However, it

369 still remains a challenge for large-scale and high-quality materials growth, and
370 processing compatibility with existing manufacturing technologies.

371 ***Integration***

372 The location of the processing unit close to individual sensor in a planar
373 configuration will unavoidably occupy the area that reserved for sensors, reducing the
374 footprint for sensing external environment and affecting signal-to-noise ratio. For
375 example, the fill factor of CMOS image sensors is limited by the area occupation of
376 readout and processing circuits. An ideal solution is to integrate the sensing and
377 processing or readout functions in a 3D monolithic configuration, where sensors can be
378 placed on top layer to ensure full exposure to the ambient environment for high
379 sensitivity, and the processing units are arranged underneath the sensor layer with the
380 shortest distance to sensors for high communication bandwidth, low latency and high
381 fill factor.

382 ***Algorithms***

383 The practical implementation requires the development of more efficient algorithms
384 that can be embedded in near/in-sensor computing systems. The algorithms for sensor
385 terminals must be extremely simple and efficient given the highly constrained
386 conditions. For examples, the signals collected from sensory terminals are usually
387 temporal events, which can be converted to spike trains for direct SNN implementation
388 and event-driven processing. It also requires algorithms for high-level processing to
389 classify spatiotemporal pattern with CNN or SNN.

390 **Conclusions**

391 The near/in-sensor computing paradigms represent future trends of hardware
392 implementation for intelligently sensory processing. To enable low-level and high-level

processing functions near or in the sensor, it requires advanced hardware architecture and efficient algorithms. The direct processing at the sensor endpoint is beneficial for high area-, time- and energy-efficiencies, exhibiting great potentials for real-time and data-intensive applications. In-sensor processing is especially significant to realize the real-time processing by eliminating massive data transfer and conversion. Near-sensor processing is enabled by advanced integration technologies and new computing algorithms close to sensor. In-sensor processing requires to develop emerging devices with new functions and mechanisms, and new computing algorithms. Although in-sensor computing shows huge potentials, most of the existing devices still remain on the investigation stage. A complete processing and large-scale integration with peripheral control units have rarely been demonstrated, which are of great significance for future in-sensor processing architectures.

Acknowledgements

This work was supported by Research Grant Council of Hong Kong (15205619), and the Hong Kong Polytechnic University (1-ZVGH and ZG6C).

Competing interests

The authors declare no competing interests.

Author contributions

Y.C. conceived the project. F. Z. performed literature research and prepared the figures. F. Z. and Y.C. carried out comparative analysis and wrote the manuscript.

416 **References**

417 1 Truong, T. P., Le, H. T. & Nguyen, T. T. A reconfigurable hardware platform
418 for low-power wide-area wireless sensor networks. *J. Phys. Conf. Ser.* **1432**,
419 012068 (2020).
420 2 Chai, Y. In-sensor computing for machine vision. *Nature* **579**, 32-33 (2020).
421 3 Taherian, F. & Asemani, D. Design and implementation of digital image
422 processing techniques in pulse-domain. *2010 IEEE Asia Pacific Conference on*
423 *Circuits and Systems*, 895-898 (2010).
424 4 Kagawa, K. et al. Pulse-domain digital image processing for vision chips
425 employing low-voltage operation in deep-submicrometer technologies. *IEEE J.*
426 *Select. Topics Quantum Electron.* **10**, 816-828 (2004).
427 5 Wilson, G. & Premson, Y. FPGA Implementation of Hardware Efficient
428 Algorithm for Image Contrast Enhancement Using Xilinx System Generator.
429 *Procedia Technol.* **24**, 1141-1148 (2016).
430 6 Mukherjee, D. & Mukhopadhyay, S. Fast hardware architecture for fixed-point
431 2D Gaussian filter. *Int. J. Electron. Commun.* **105**, 98-105 (2019).
432 7 Delbrück, T. & Mead, C. Analog VLSI phototransduction by continuous-time,
433 adaptive, logarithmic photoreceptor circuits. (1995).
434 8 Ruedi, P.-F. et al. A 128×128 pixel 120-db dynamic-range vision-sensor chip
435 for image contrast and orientation extraction. *IEEE J. Solid-State Circuits* **38**,
436 2325-2333 (2003).
437 9 Lichtsteiner, P. & Delbruck, T. A 64x64 AER logarithmic temporal derivative
438 silicon retina. *Research in Microelectronics and Electronics* **2**, 202-205 (2005).
439 10 Zhou, F. et al. Optoelectronic resistive random access memory for
440 neuromorphic vision sensors. *Nat. Nanotechnol.* **14**, 776-782 (2019).
441 11 Wang, C.-Y. et al. Gate-tunable van der Waals heterostructure for
442 reconfigurable neural network vision sensor. *Sci. Adv.* **6**, eaba6173 (2020).
443 12 Hasler, P., Smith, P. D., Graham, D., Ellis, R. & Anderson, D. V. Analog
444 floating-gate, on-chip auditory sensing system interfaces. *IEEE Sens. J.* **5**, 1027-
445 1034 (2005).
446 13 Ellis, R., Yoo, H., Graham, D. W., Hasler, P. & Anderson, D. V. A continuous-
447 time speech enhancement front-end for microphone inputs. *IEEE International*
448 *Symposium on Circuits and Systems* **2**, II-II (2002).
449 14 Wen, B. & Boahen, K. A 360-channel speech preprocessor that emulates the
450 cochlear amplifier. *IEEE International Solid State Circuits Conference*, 2268-
451 2277 (2006).
452 15 Lyon, R. F. & Mead, C. An analog electronic cochlea. *IEEE Trans. Acoust.,*
453 *Speech, Signal Processing* **36**, 1119-1134 (1988).
454 16 Jiménez-Fernández, A. et al. A binaural neuromorphic auditory sensor for
455 FPGA: A spike signal processing approach. *IEEE transactions on neural*
456 *networks and learning systems* **28**, 804-818 (2016).
457 17 Kumar, N., Himmelbauer, W., Cauwenberghs, G. & Andreou, A. G. An analog
458 VLSI chip with asynchronous interface for auditory feature extraction. *IEEE*
459 *Transactions on Circuits and Systems II: Analog and Digital Signal Processing*
460 **45**, 600-606 (1998).
461 18 Koickal, T. J. et al. Analog VLSI circuit implementation of an adaptive
462 neuromorphic olfaction chip. *IEEE Trans. Circuits Syst. {I}* **54**, 60-73 (2007).
463 19 Yoon, J. H. et al. An artificial nociceptor based on a diffusive memristor. *Nat.*
464 *Commun.* **9**, 1-9 (2018).

465 20 Du, Z. et al. in *Proceedings of the 42nd Annual International Symposium on*
466 *Computer Architecture.* 92-104.
467 21 Chen, T. et al. Diannao: A small-footprint high-throughput accelerator for
468 ubiquitous machine-learning. *ACM SIGARCH Computer Architecture News* **42**,
469 269-284 (2014).
470 22 Chen, Z. et al. Processing Near Sensor Architecture in Mixed-Signal Domain
471 With CMOS Image Sensor of Convolutional-Kernel-Readout Method. *IEEE*
472 *Trans. Circuits Syst. {I}* (2019).
473 23 Liu, Z. et al. A 1.8 mW Perception Chip with Near-Sensor Processing Scheme
474 for Low-Power AIoT Applications. *IEEE Computer Society Annual Symposium*
475 *on VLSI (ISVLSI)*, 447-452 (2019).
476 24 LiKamWa, R., Hou, Y., Gao, J., Polansky, M. & Zhong, L. RedEye: analog
477 ConvNet image sensor architecture for continuous mobile vision. *ACM*
478 *SIGARCH Computer Architecture News* **44**, 255-266 (2016).
479 25 Li, C. et al. Analogue signal and image processing with large memristor
480 crossbars. *Nat. Electron.* **1**, 52 (2018).
481 26 Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on
482 memristive systems. *Nat. Electron.* **1**, 22-29 (2018).
483 27 Yao, P. et al. Face classification using electronic synapses. *Nat. Commun.* **8**, 1-
484 8 (2017).
485 28 Seo, S. et al. Artificial optic-neural synapse for colored and color-mixed pattern
486 recognition. *Nat. Commun.* **9**, 1-8 (2018).
487 29 Chu, M. et al. Neuromorphic hardware system for visual pattern recognition
488 with memristor array and CMOS neuron. *IEEE Trans. Ind. Electron* **62**, 2410-
489 2419 (2014).
490 30 Mennel, L. et al. Ultrafast machine vision with 2D material neural network
491 image sensors. *Nature* **579**, 62-66 (2020).
492 31 Kyuma, K. et al. Artificial retinas—fast, versatile image processors. *Nature* **372**,
493 197-198 (1994).
494 32 Hsieh, H.-Y. & Tang, K.-T. VLSI implementation of a bio-inspired olfactory
495 spiking neural network. *IEEE Trans. Neural Netw. Learn. Syst.* **23**, 1065-1073
496 (2012).
497 33 Shulaker, M. M. et al. Three-dimensional integration of nanotechnologies for
498 computing and data storage on a single chip. *Nature* **547**, 74-78 (2017).
499 34 Tan, H. et al. Tactile sensory coding and learning with bio-inspired
500 optoelectronic spiking afferent nerves. *Nat. Commun.* **11**, 1-9 (2020).
501 35 Kim, Y. et al. A bioinspired flexible organic artificial afferent nerve. *Science*
502 **360**, 998-1003 (2018).
503 36 Wan, C. et al. An artificial sensory neuron with tactile perceptual learning. *Adv.*
504 *Mater.* **30**, 1801291 (2018).
505 37 Wang, W. et al. Learning of spatiotemporal patterns in a spiking neural network
506 with resistive switching synapses. *Sci. Adv.* **4**, eaat4752 (2018).
507 38 Cottini, N., Gottardi, M., Massari, N., Passerone, R. & Smilansky, Z. A 33μW
508 64×64 Pixel Vision Sensor Embedding Robust Dynamic Background
509 Subtraction for Event Detection and Scene Interpretation. *IEEE J. Solid-State*
510 *Circuits* **48**, 850-863 (2013).
511 39 Kundu, S. & Chattopadhyay, S. *Network-on-chip: the next generation of*
512 *system-on-chip integration*. (CRC press, 2018).
513 40 Samal, S. K., Nayak, D., Ichihashi, M., Banna, S. & Lim, S. K. Monolithic 3D
514 IC vs. TSV-based 3D IC in 14nm FinFET technology. *2016 IEEE SOI-3D-*

515 *Subthreshold Microelectronics Technology Unified Conference (S3S)*, 1-2
516 (2016).
517 41 Amir, M. F., Ko, J. H., Na, T., Kim, D. & Mukhopadhyay, S. 3-D stacked image
518 sensor with deep neural network computation. *IEEE Sens. J.* **18**, 4187-4199
519 (2018).
520 42 Wang, Y. et al. Monolithic integration of all-in-one supercapacitor for 3D
521 electronics. *Adv. Energy. Mater.* **9**, 1900037 (2019).
522 43 Shi, J. et al. Smart Textile-Integrated Microelectronic Systems for Wearable
523 Applications. *Adv. Mater.* **32**, 1901958 (2020).
524 44 Zhang, X. et al. Heterogeneous 2.5 D integration on through silicon interposer.
525 *Appl. Phys. Rev.* **2**, 021308 (2015).
526 45 Hu, Y.-C. et al. An advanced 2.5-D heterogeneous integration packaging for
527 high-density neural sensing microsystem. *IEEE Trans. Electron Devices* **64**,
528 1666-1673 (2017).
529 46 Wang, M. et al. Gesture recognition using a bioinspired learning architecture
530 that integrates visual data with somatosensory data from stretchable sensors.
531 *Nat. Electron.*, 1-8 (2020).
532 47 Yao, P. et al. Fully hardware-implemented memristor convolutional neural
533 network. *Nature* **577**, 641-646 (2020).
534 48 Lin, P. et al. Three-dimensional memristor circuits as complex neural networks.
535 *Nat. Electron.* **3**, 225-232 (2020).
536 49 Ielmini, D. & Wong, H.-S. P. In-memory computing with resistive switching
537 devices. *Nat. Electron.* **1**, 333-343 (2018).

538

539

540

541

542

543

544

545

546

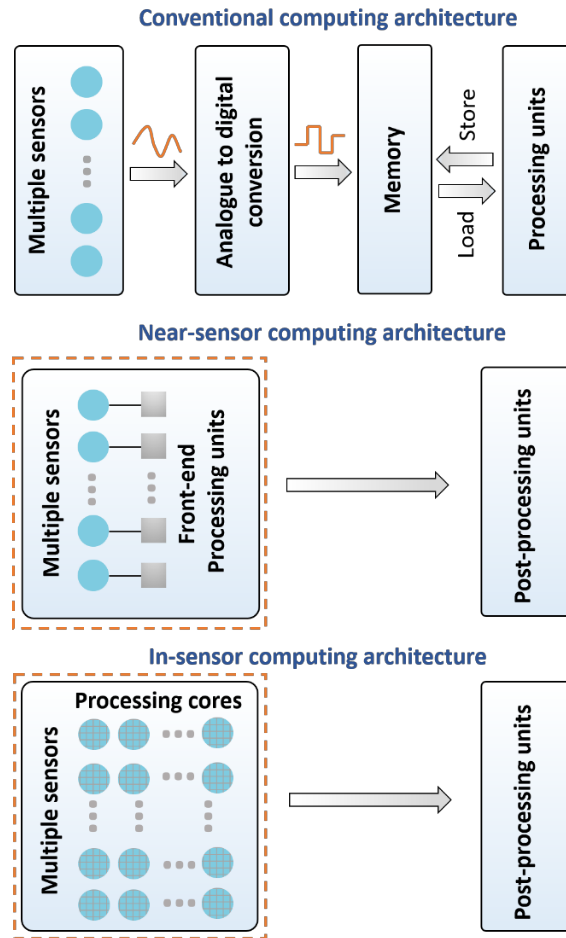
547

548

549

550

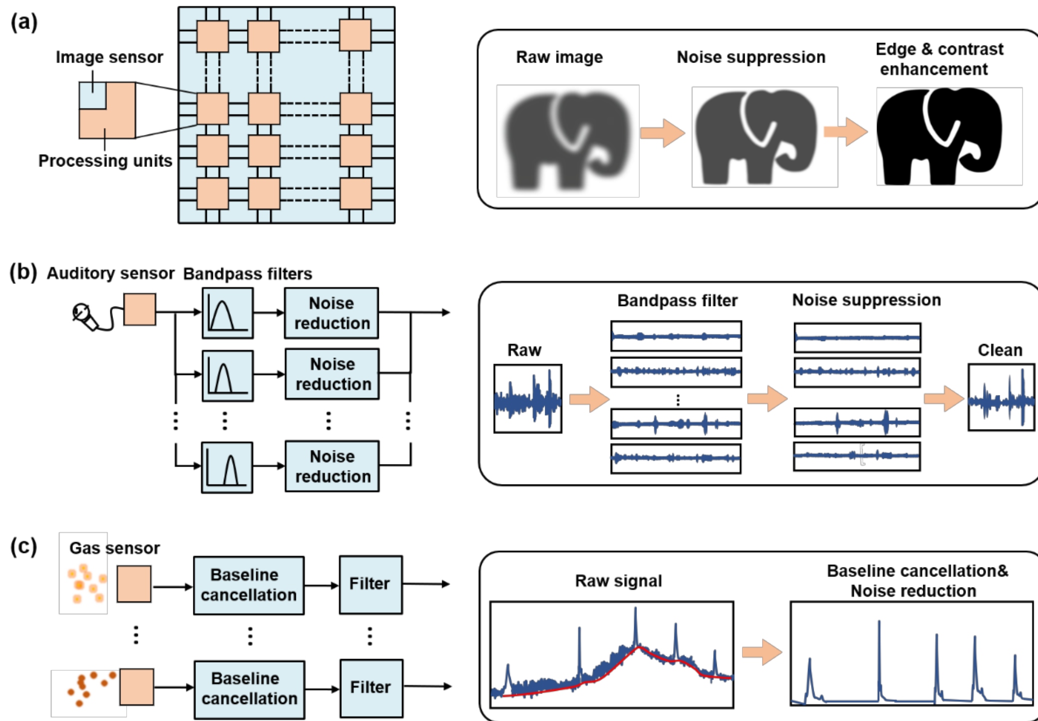
551 **Figures and captions**



552

553 **Fig. 1 | Conventional sensory architecture, near-sensor computing and in-sensor computing**
554 **architectures for processing sensory data.** In conventional computing architecture, analogue
555 outputs from the sensors are first converted to digital signals that are stored in memory. Processing
556 units load data from memory units, and then transmit output signals back to memory for storage. In
557 the near-sensor computing architecture, individual sensors are connected to front-end processing
558 units through advanced integrated circuit packaging technologies for real-time readout and
559 processing. The front-end near-sensor processing units implement a portion of processing tasks,
560 which are then further transmitted to post-processing units for more complicated processing. In the
561 in-sensor computing architecture, the processing functions are embedded the sensors for front-end
562 processing. The sensors can collaborate together to perform information processing, data
563 aggregation and compression, eliminating data transmission between sensors and processors.

564



565

566

567 **Fig. 2 | Illustrations of low-level sensory processing architectures and functions.** (a) Low-level
568 image processing: noise suppression, edge extraction and contrast enhancement. For the near-sensor
569 computing, processing units are directly connected to pixels in an image sensor. For the in-sensor
570 computing, image sensing and processing are fused in sensor itself. (b) Low-level auditory
571 processing. The raw auditory signal is filtered through bandpass filters with noise suppression in
572 each channel to obtain clean signals for further processing. (c) Low-level olfactory processing. The
573 variation of baseline in the raw sensory data can affect the differentiation of gas types in high-level
574 processing. During the low-level processing, baseline is removed from the body signals.

575

576

577

578

579

580

581

582

583

584

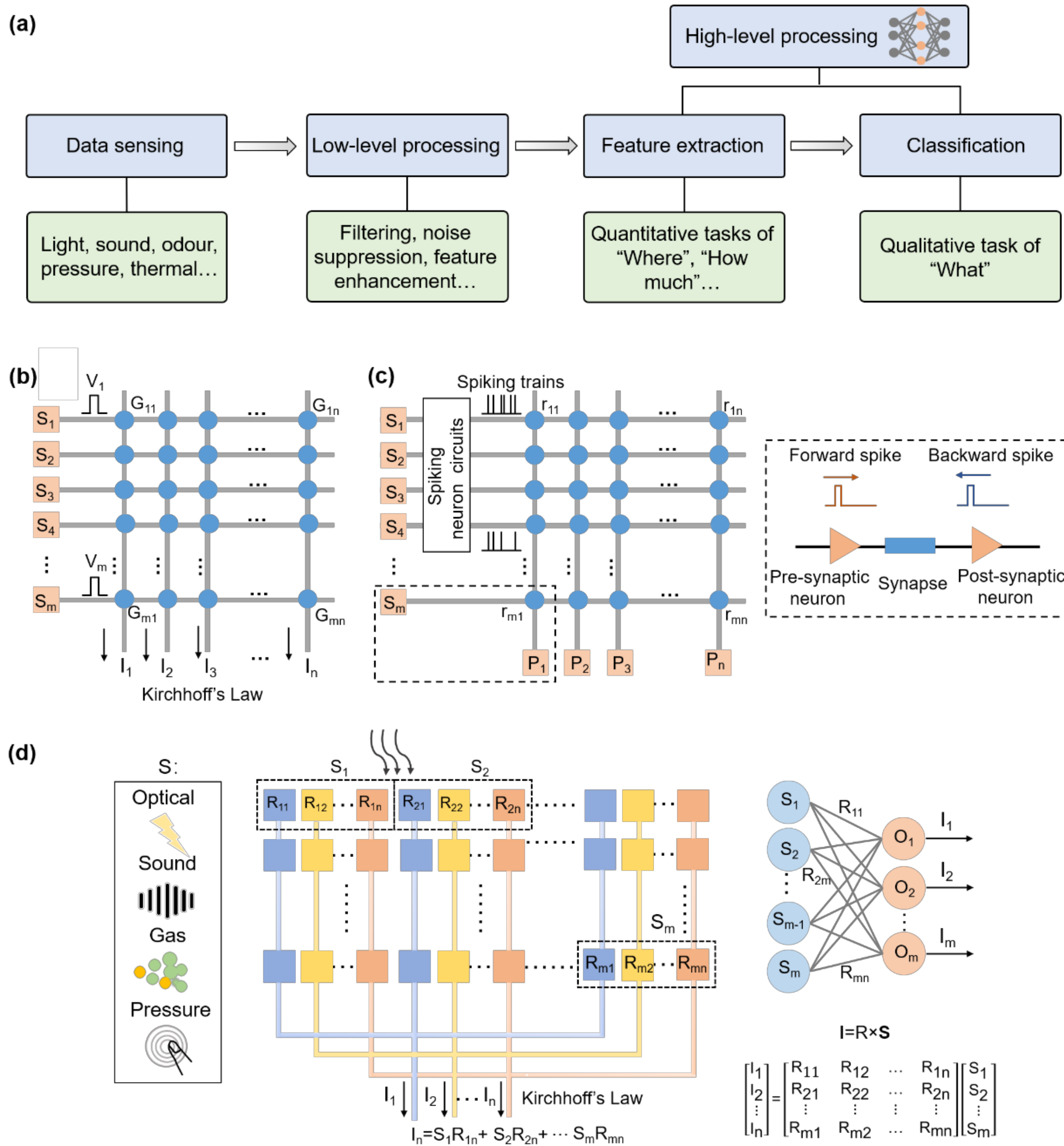
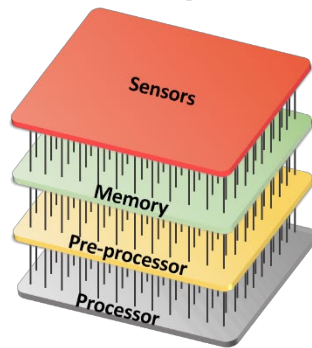
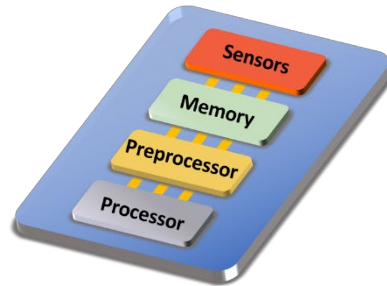


Fig. 3| Near-sensor and in-sensor high-level sensory processing. (a) Data flow of sensory processing from low-level processing to high-level processing. Low-level processing generally involves filtering, noise suppression and feature enhancement, which are local operations. High-level processing is associated with feature extraction and recognition processes for abstract representation, such as quantitative and qualitative determinations of "where" and "what". (b) Near-sensor architecture of SNN implementation through STDP learning with memory synaptic array, where *r* stands for the synaptic weight. The sensory information is coded into spike trains through rate-coding, which is further inputted into synaptic arrays. A presynaptic neuron is connected to a postsynaptic neuron via synapses. (c) Schematic illustrations of in-sensor computing architecture with reconfigurable sensors for MAC operations in the neural network. ***S*** stands for the stimuli to sensor elements, ***R*** is the sensor responsivity, and ***I*** is the summation of output currents. The relationship of ***S***, ***R*** and ***I*** can be expressed with matrix-vector multiplication. $\boldsymbol{I} = (I_1, I_2,\ldots,I_n) = \boldsymbol{R_{mn}}\cdot\boldsymbol{S} = \boldsymbol{R_{mn}}\cdot(S_1, S_2,\ldots,S_m)$, where $\boldsymbol{R_{mn}}$ is the responsivity matrix, and ***I*** and ***S*** are output and input vectors. (d) Near-sensor architecture of MAC operation with memory synaptic array in CNN. A vector of voltage outputs $\boldsymbol{V} = (V_1, V_2,\ldots,V_n)$ from a sensor is directly inputted to the rows of a memory array. $\boldsymbol{G_{mn}}$ is the conductance matrix and output vector $\boldsymbol{I} = (I_1, I_2,\ldots,I_n) = \boldsymbol{G_{mn}}\cdot\boldsymbol{V} = \boldsymbol{G_{mn}}\cdot(V_1, V_2,\ldots,V_m)$.

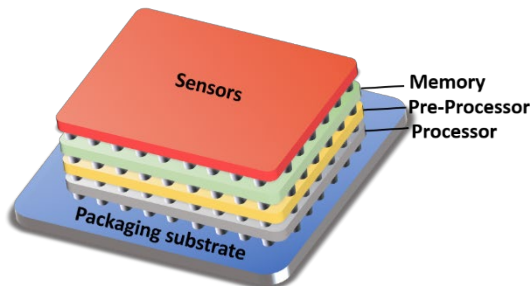
3D monolithic integration



Planar SoC integration



3D heterogeneous integration



2.5D chiplet integration

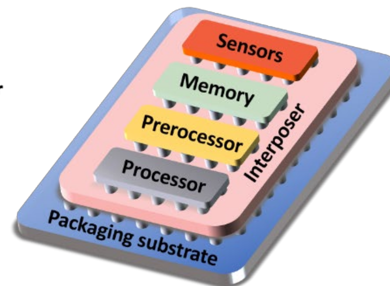


Fig. 4 | Integration technologies for near-sensor and in-sensor computing. (a) 3D monolithic integration system generally connects different functional layers of sensor, memory, and processors in a 3D stacked configuration via inter-layer vias. (b) The functional units are integrated on a planar SoC chip with planar wire connection. (c) In 3D heterogeneous integration, different functional units are fabricated on different wafers, which are further integrated with advanced packaging technologies (*e.g.*, TSVs, die-to-die, die-to-wafer, and wafer-to-wafer interconnects). (d) 2.5D Chiplets with specific functions are connected through interposer, which is a midpoint between 2D and 3D packaging integration.