香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

ICTA 2023, Hefei, China

HQ OWL

ASIC for AI Embodied in Tomorrow's Robots – Why & Why Not?

Patrick Yue 俞捷

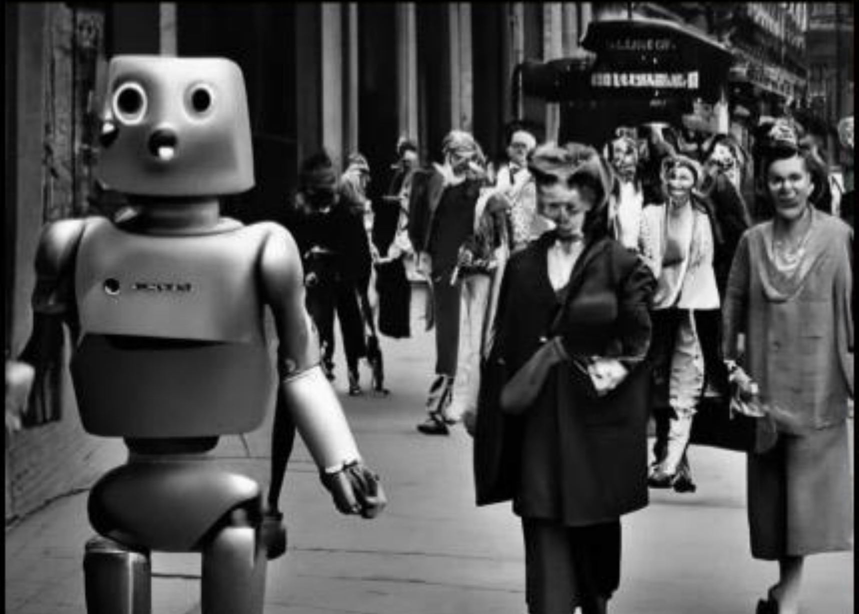
Integrated Circuit Design Center (ICDC) & Optical Wireless Lab (OWL)
Department of Electronic and Computer Engineering (ECE)
The Hong Kong University of Science and Technology (HKUST)

ASICs – AI Specific Interated Circuits

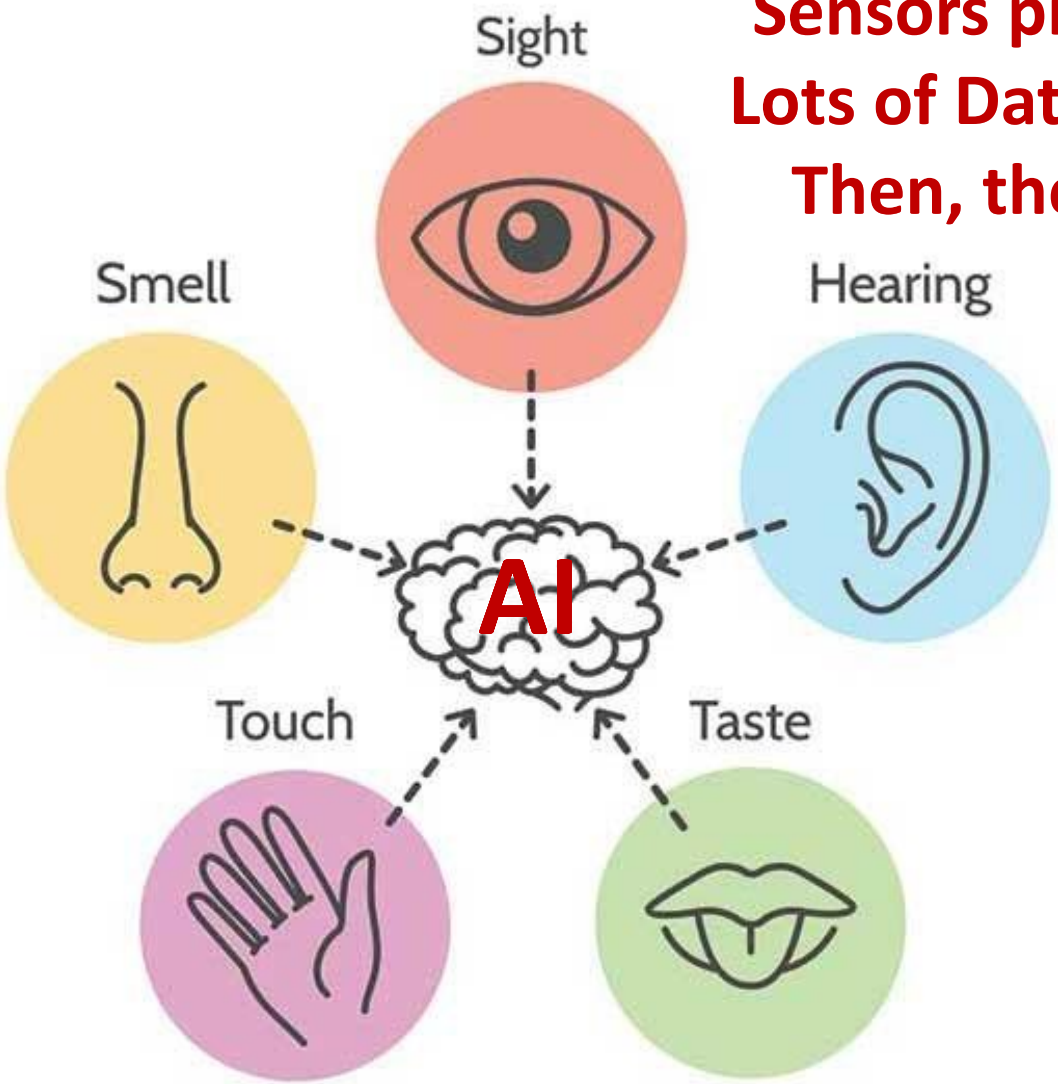
ASIC DESIGN


Run fast in the wrong way
is worse than “laying flat”

AI Embodied in Tomorrow’s Robots Why & Why Not?

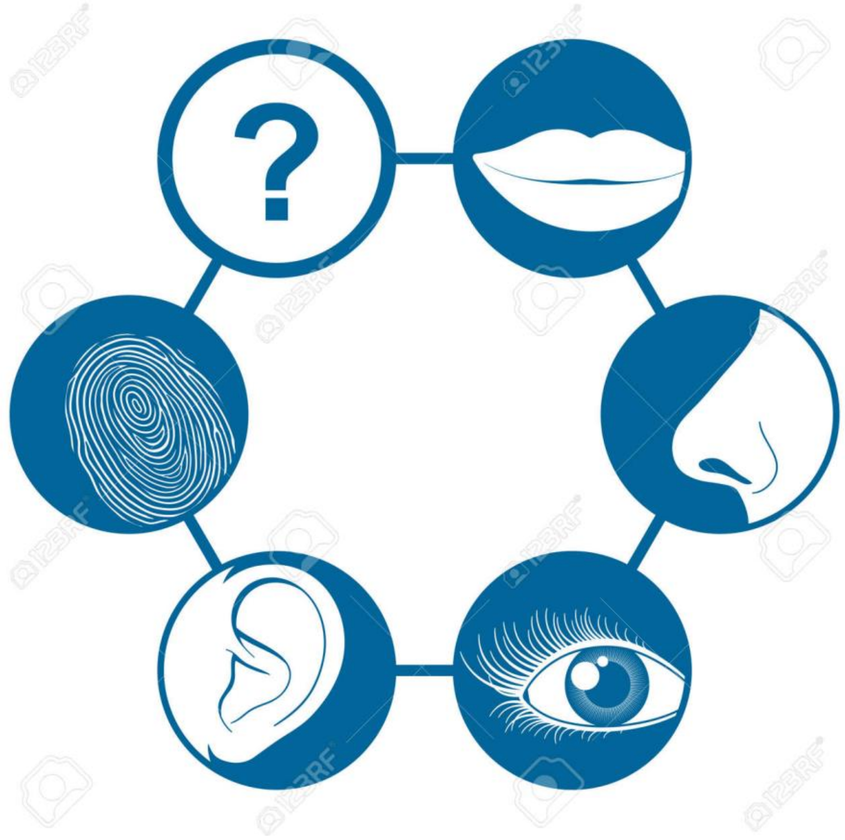


Tomorrow's Smart Robots Empowered with AI & Senses

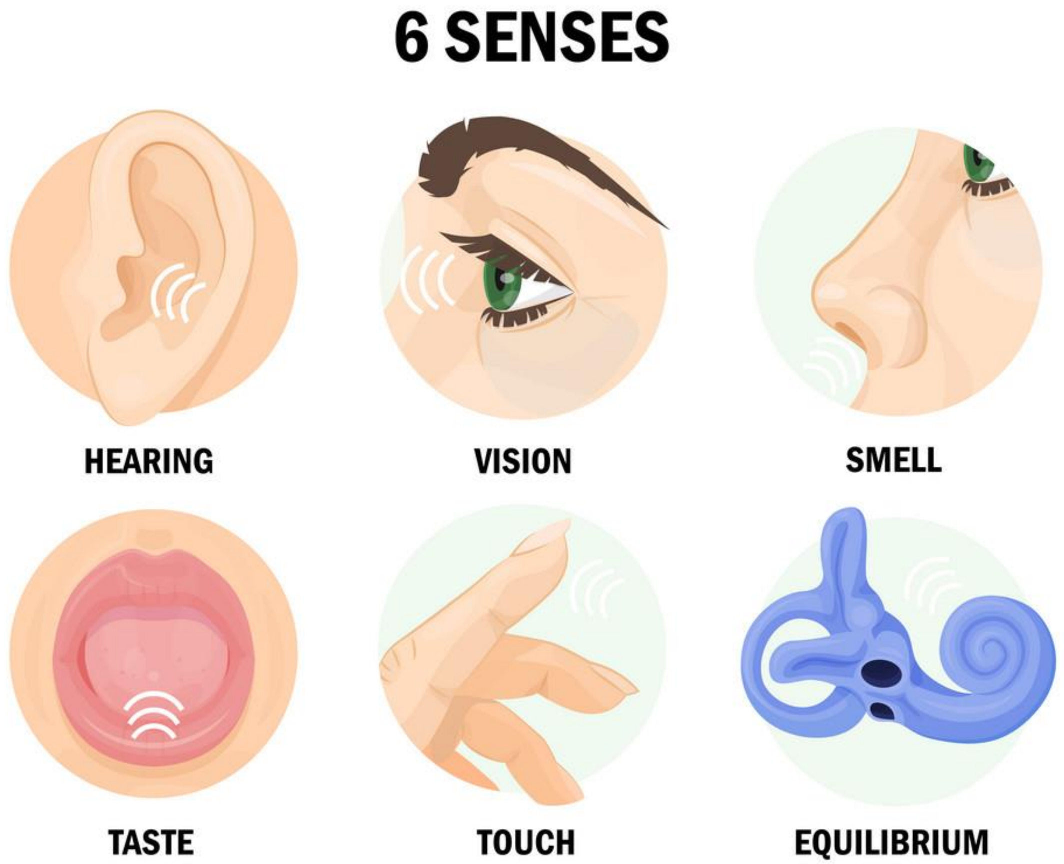


Sensors provide data
Lots of Data! Big Data!
Then, there is LLM

What's the 6th sense?



Some Suggested that 6th Sense is Equilibrium, but…

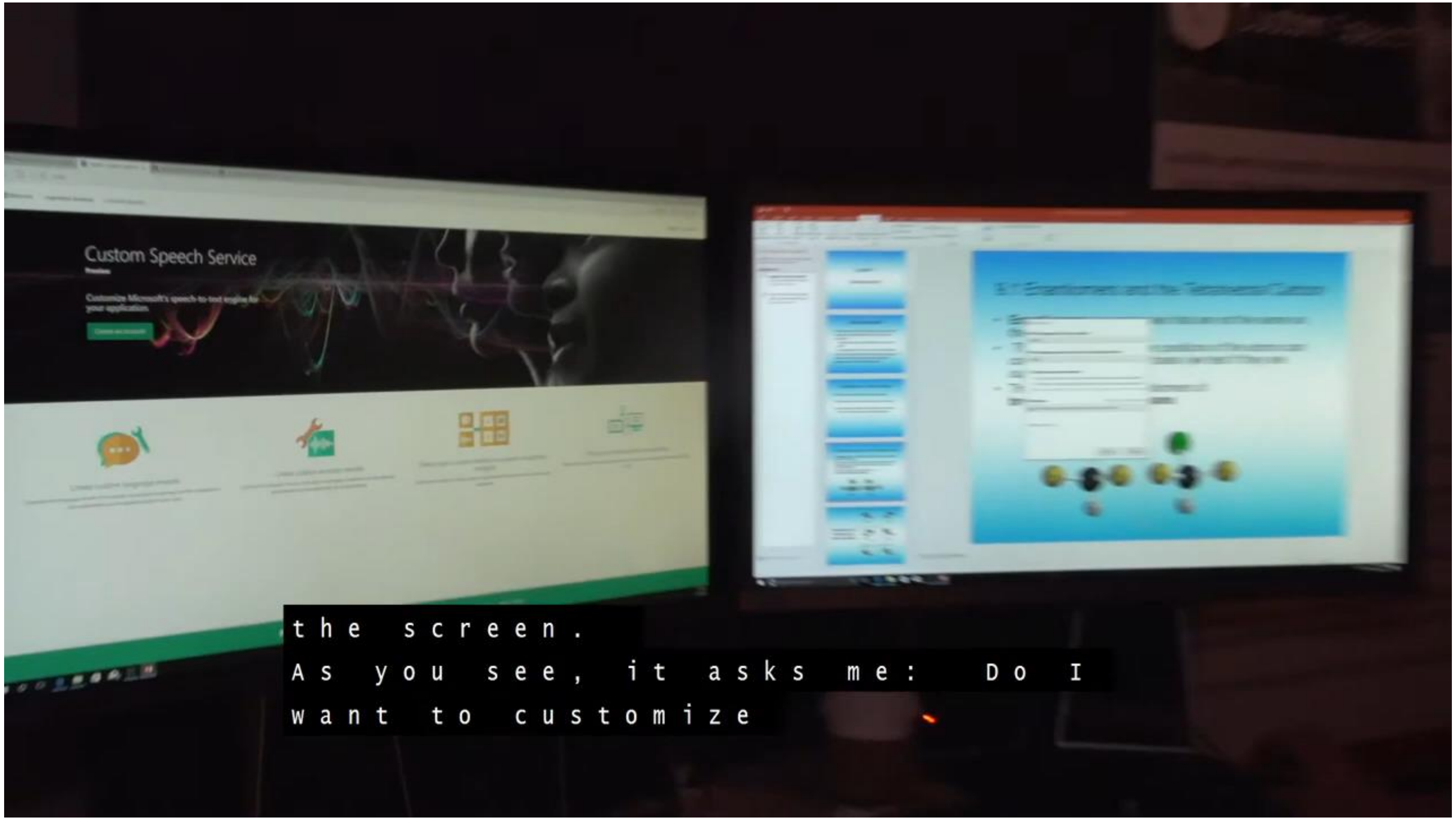


The 1st Sense – Vision *(A. Dai, Stanford, Princeton & TU Munich, 2017)*

https://www.bilibili.com/video/BV17E411Y7mC/?spm_id_from=333.337

Dai, A. *et al.*, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," 2017 *CVPR*.

The 2nd Sense – Speech *(M. Seltzer, Microsoft, 2017)*



https://www.youtube.com/watch?v=80CPeMLJMEw

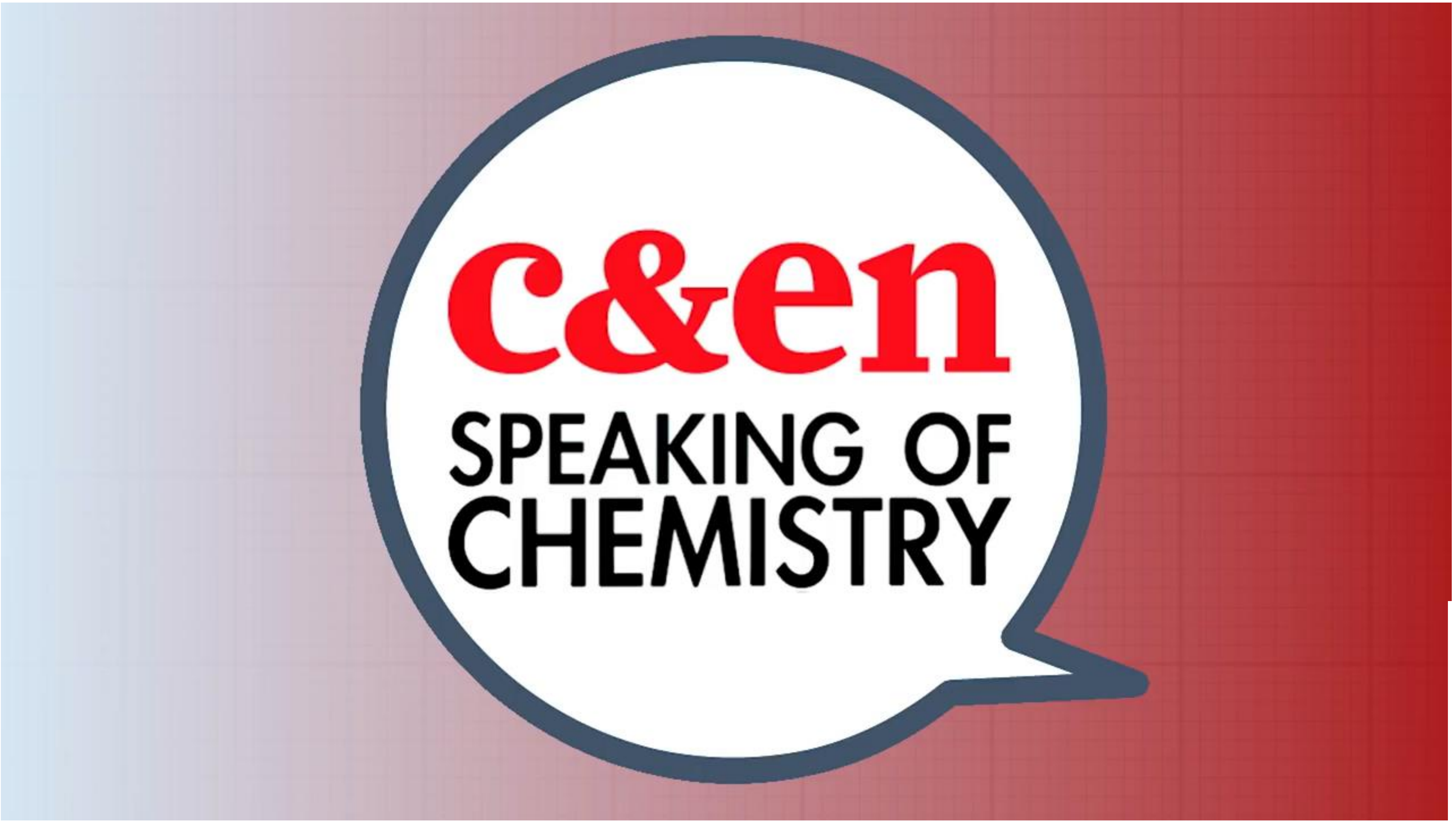
The 4th Sense – Taste *(H. Miyashita, Meiji University, 2022)*

https://www.youtube.com/watch?v=P-V3EqQEuyQ

The 3rd Sense – Smell *(J. McGann, Rutgers University, 2015)*

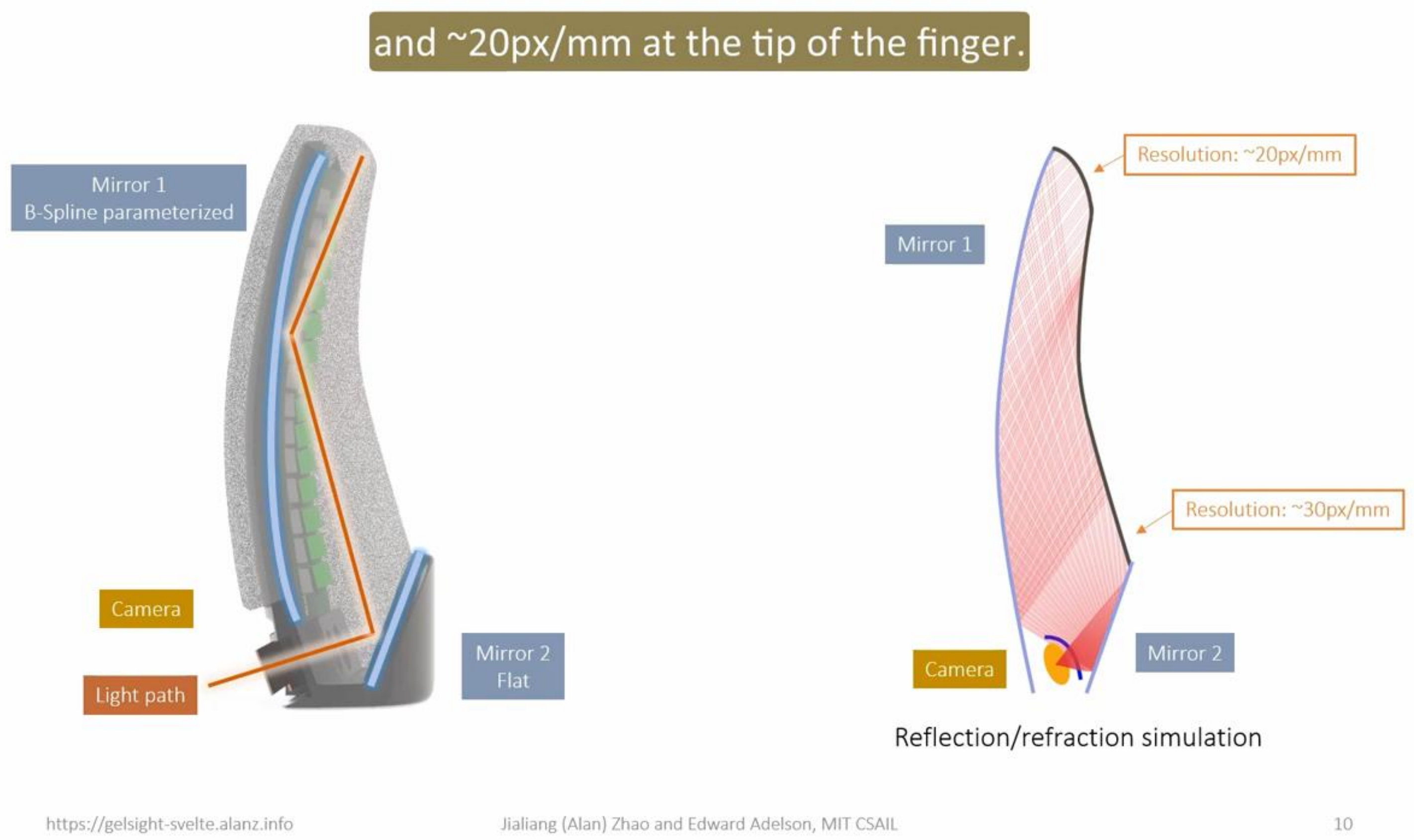
https://www.youtube.com/watch?v=zaHR2MAxywg&t=99s%E2%80%8B

The 5th Sense – Touching *(Z. Bao, Stanford University, 2016)*



https://www.bilibili.com/video/BV1M4411j7wm/?spm_id_from=333.337

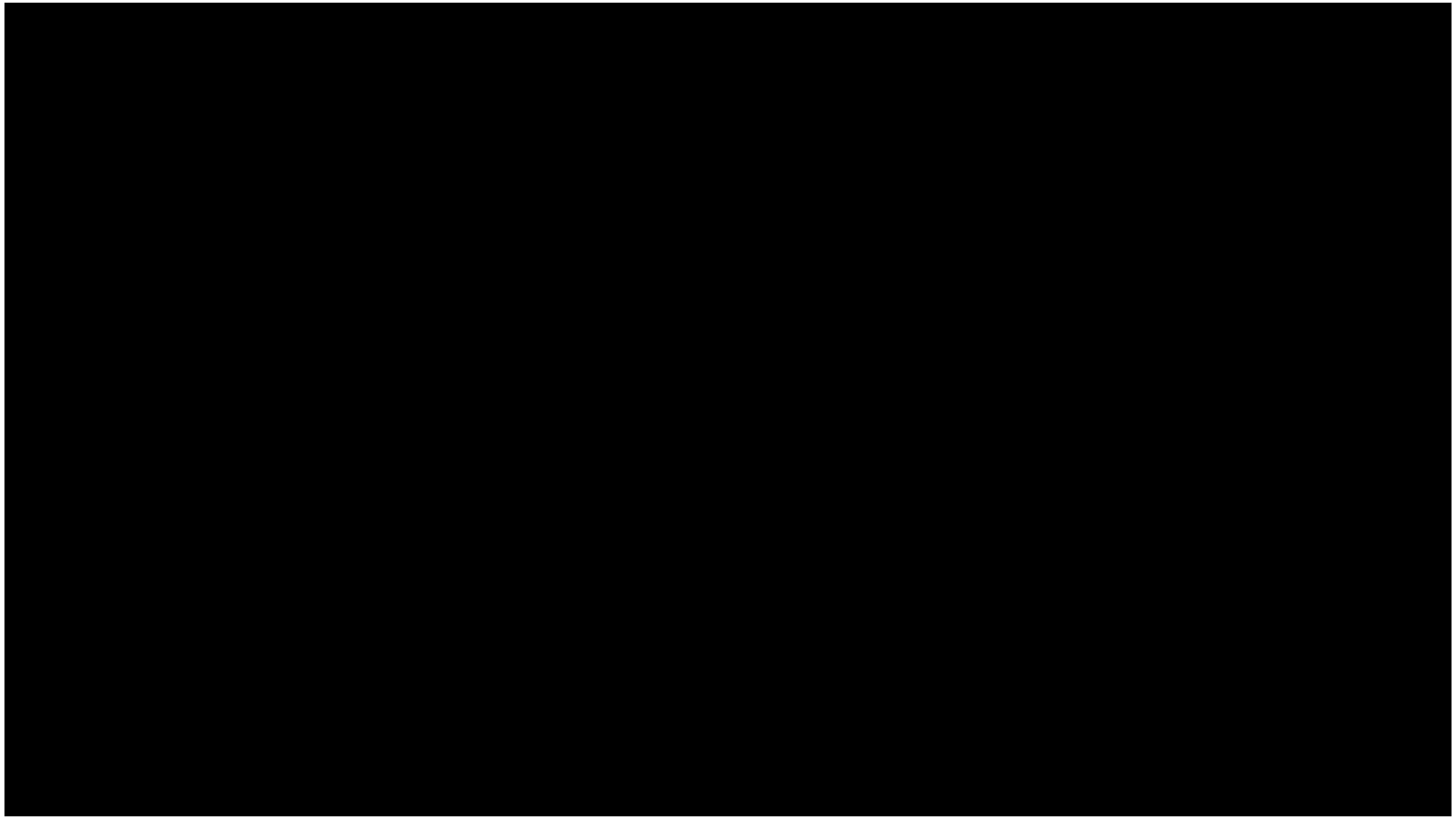
The 5th Sense – Touching *(J. Zhao & E.H. Adelson, MIT, 2023)*



https://www.youtube.com/watch?v=yI6WDzfYD8Q&t=175s

Zhao, J., & Adelson, E. H. "GelSight Svelte Hand: A Three-finger, Two-DoF, Tactile-rich, Low-cost Robot Hand for Dexterous Manipulation," arXiv preprint arXiv:2309.10886, 2023
Zhao, J., & Adelson, E. H. "GelSight Svelte: A Human Finger-shaped Single-camera Tactile Robot Finger with Large Sensing Coverage and Proprioceptive Sensing," arXiv preprint arXiv:2309.10885, 2023

The 6th Sense – Indoor Positioning *(A. Arun, UCSD, 2022)*



https://www.youtube.com/watch?v=JjalvBHqC94&t=111s%E2%80%8B

Arun *et al.*, "P2SLAM: Bearing Based WiFi SLAM for Indoor Robots," IEEE Robotics and Automation Letters, 7(2), 3326-3333, 2022.

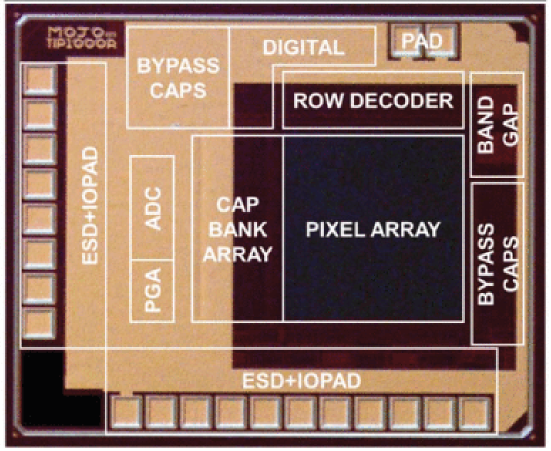
So, how does influence the chip design?

[3] Vision SoC by Mojo Vision, 40nm CMOS ULP, 21000μm^2

Imager	
Technology	CIS BSI FEOL 90nm/BEOL 65nm
Supply Voltage	3.3V (pix.), 1V (ana.), 0.8V (dig.)
Die Size	1.0mm × 1.3mm
Array Size	256 × 256 monochrome
Pixel Size	1.75μm × 1.75μm
Readout Modes	Progressive 4, 6 and 8 bits, Subsampling 2×,4× and 8×
Chip I/O	SPI, Custom high-speed serial
Conv. Gain	130μV/e$^-$
Dynamic Range	53dB[a] (gain=1×), 42dB[a] (gain=8×)
Random Noise	4.6e$^-_{rms}$ at gain = 8×
Wake-up Time	<150μs from deep-sleep mode
ADC FOM	20fJ/conv-step (with ref.) at 8 bits
Energy Efficiency	21pJ/frame/pixel[b] at 8 bits
	13pJ/frame/pixel[b] at 4 bits
Power	61μW[b], 82μW[c] at 8 bits, 44FPS
	75μW[b], 95μW[c] at 4 bits, 88FPS

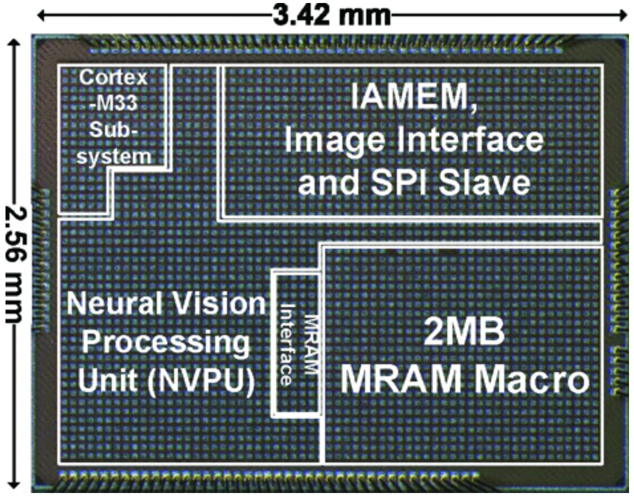
a EMVA Standard 1288 definition
b Excludes power to send data off-chip
c Includes power to send data off-chip (5.5pF load)

Imager Processor	
Architecture	Linear contrast filter + Serial image convolution filters
Technology	40nm CMOS ULP
Supply Voltage	0.8V
Max data rate	27Mbps
Area	21000μm^2
Energy Efficiency	34pJ/frame/pixel
Power	111μW (edge detection), 0.6μW (contrast filter) at 50FPS



- Singh, Rituraj, et al. "34.2 a 21pJ/frame/pixel imager and 34pJ/frame/pixel Image Processor For A Low-vision Augmented-reality Smart Contact Lens." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021.

[7] Vision SoC by University of Michigan, 22nm CMOS, 2.56mm x 3.42mm

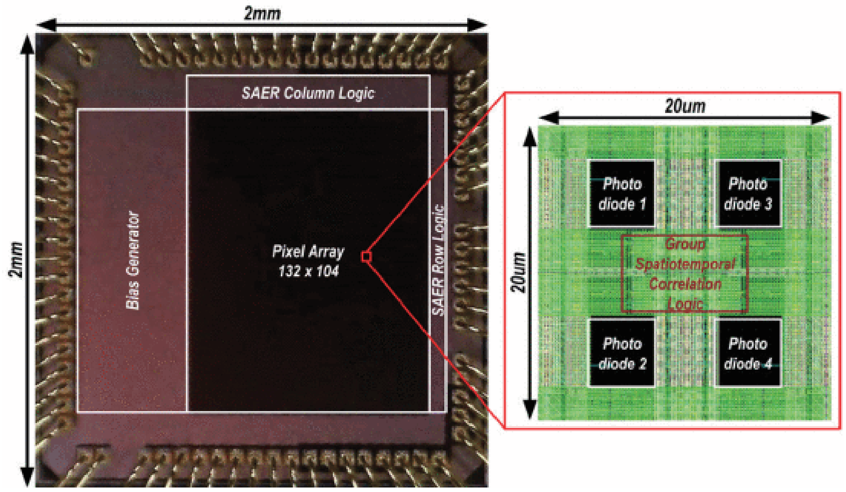


	VLSI '17 [4]	JSSC '19 [3]	ISSCC '19 [9]	ISSCC '21 [10]	VLSI '21 [11]	This Work
Technology	28nm CMOS	65nm CMOS	22nm FDSOI	22nm FDSOI	40nm CMOS	22nm CMOS
Application	Multi-View Depth	Visual-Inertial Odometry	Image Processing	DNN for IoT	Edge NN Inference and Weight Tuning	**CNN and Non-CNN Vision for *Micro-Robot* Navigation**
Processing Architecture	Dedicated Accelerator	Dedicated Accelerator	Streaming-based CGRA	CMP and NN Accelerator	Systolic Array	Hybrid Systolic 2D-Mapping PE Array
Programmability	Not programmable	Not programmable	Only for Image Processing	General Purpose and NN	Only for NN	**General Purpose, Image Processing and NN**
Die Area	5.96mm^2	20mm^2	4.9mm^2	12mm^2	29.2mm^2	8.76mm^2
SRAM	582.5 kB	854 kB	690 kB	1728 kB	0.5 MB	1428 kB
On-Chip NVM	N/A	N/A	N/A	4MB MRAM	2MB RRAM	**2MB MRAM**
Voltage	0.9V	1.0V	0.8V	0.5 ~ 0.8V	1.1V	0.5 ~ 1.0V
Frequency	300MHz	62.5MHz	5 ~ 220MHz	32kHz ~ 450MHz	200MHz	56kHz ~ 190MHz
Power	380mW	24mW	10.7 ~ 101.4mW	1.7μW ~ 49.4mW	126mW	468μW ~ 158mW
Peak Image Proc. Perf.	Not Reported	**59.1GOPS**[1]	**145GOPS**[2]	Not Reported	N/A	**207GOPS**[2] @ 1.0V, 180MHz
INT8 Peak NN Perf.	N/A	N/A		**32.2GOPS**[3] **(NN Accelerator)**	**920GOPS**[3]	**511GOPS**[3] @ 1.0V, 190MHz (146GOPS[2] INT16)
Feature Detection Efficiency	N/A	**0.33nJ/pix**[1] Shi-Tomasi corner and **1.40nJ/pix**[1] sparse LK flow, 71fps Wide-VGA (752x480)	Not Reported	Not Reported	N/A	**0.22nJ/pix**, 3.4mW, 50fps VGA, Harris corner @ 0.52V, 20MHz
Feature Tracking Efficiency	N/A		**1.16nJ/pix**, 30fps VGA, dense LK flow	Not Reported	N/A	**0.22nJ/pix**, 1.6mW, 23fps VGA, sparse LK flow (≤ 100 features) @ 0.52V, 20MHz
Dense Depth Estimation Efficiency	**0.042nJ/pix**, 32fps 2K, stereo local matching	N/A	Not Reported	Not Reported	N/A	**0.055nJ/pix**, 10.9mW, 10fps VGA, stereo local matching (depth level = 64) @ 0.62V, 50MHz
INT8 NN Inference Efficiency	N/A	N/A	Not Reported	**1.3TOPS/W**[3,4] **(NN Accelerator)**	**2.2TOPS/W**[3,4]	**12.1TOPS/W**[3,5] (3.5TOPS/W[2,5] INT16)

[1] OP definition not specified. Energies calculated from Fig. 11 and Table III in [3]. [2] 16b/32b mult/add = 1OP. [3] 8b MAC = 2OPs. Our 16b MAC (16b mult, 32b add) is normalized to 7OPs as per [12]. [4] Sparsity not specified. [5] 80% weight sparsity and 50% input activation sparsity.

- **Zhang, Qirui, et al. "A 22nm 3.5 TOPS/W Flexible Micro-Robotic Vision SoC with 2MB eMRAM for Fully-on-Chip Intelligence." 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2022.**

[9] Vision SoC by iniVation AG & iniLabs GmbH & INI UZHĐ, 65nm 1P9M, 2mm x 2mm



		This Work	[5]	[3]	[2]	[1]
Technology		**65nm 1P9M**	90nm 1P9M BSI	0.18µm 1P6M	0.35µm 2P4M	0.35µm 2P4M
Resolution		**132x104**	640x480	240x180	128x128	128x128
Chip Size (mm^2)		**2x2**	8x5.8	5x5	4.9x4.9	6.3x6
Pixel Size (µm2)		**10x10**	9x9	18.5x18.5	31x30	40x40
Fill Factor (%)		**20**	-	22	10.5	8.1
Power Supply (V)		**1.2**	2.8 & 1.2	3.3 & 1.8	3.3	3.3
Power (mW)	High Activity	**4.9@180Meps** a	50@300Meps	14	-	24
	Low Activity	**0.25@100keps** a	27@100keps	5	4	-
Normalized Power b	Dynamic (pJ/event)	**26**	77	-	-	-
	Static (nW/pixel)	**18**	88	-	-	-
Max Event Rate (Meps)		**180**	300	12	20	2
Readout Efficiency (event/clock)		**best: 4, worst: 0.25** c	best: 6.7, worst: 0.077 d	-	-	-

a The power includes bias generator and IO power and was measured using identical bias configuration.

b The normalized power is calculated as:

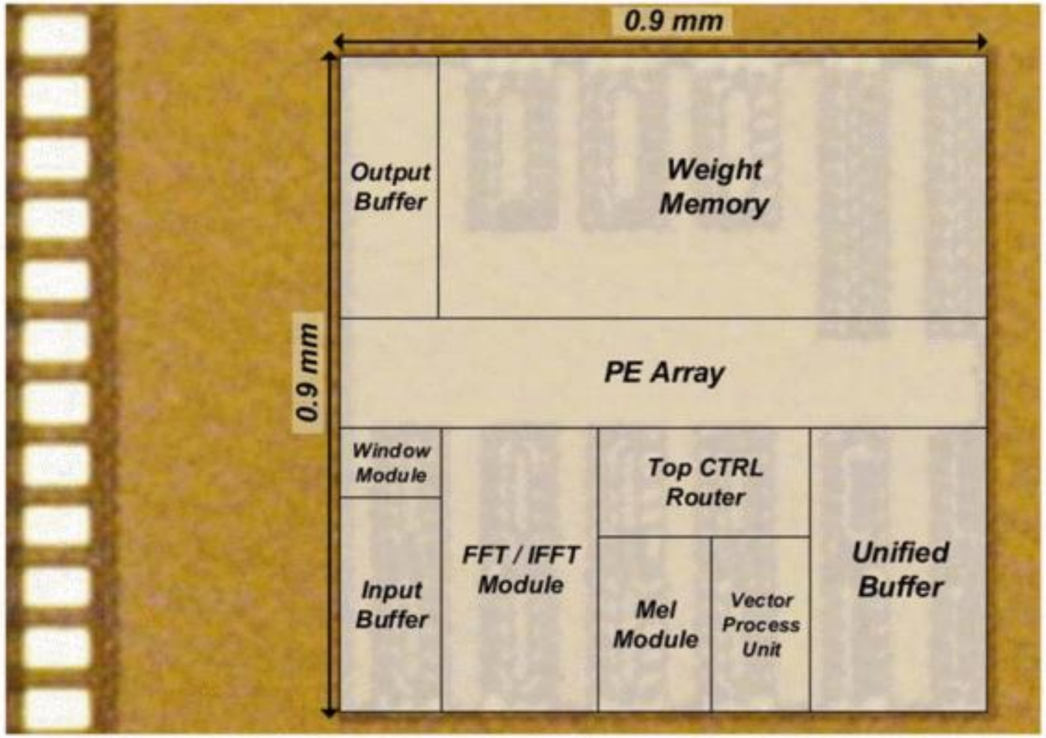
Dynamic Energy = $(P_H - P_L) / (R_H - R_L)$, Static Power = $(P_L - R_L \cdot \text{Dynamic Energy}) / N_P$, where P_H is power at high activity, P_L is power at low activity, R_H is event rate at high activity, R_L is event rate at low activity, N_P is total number of pixels.

c The best case is when all events are in the same row of groups, the worst case is when all events are in different rows of groups, where a minimum of 4 clocks per row is needed.

d The best case is when all events are in the same column, the worst case is when all events are in different columns, where a minimum of 13 clocks per column is needed.

- **Li, Chenghan, et al. "A 132 by 104 10µm-Pixel 250µW 1kefps Dynamic Vision Sensor With Pixel-parallel Noise And Spatial Redundancy Suppression." 2019 Symposium on VLSI Circuits. IEEE, 2019.**

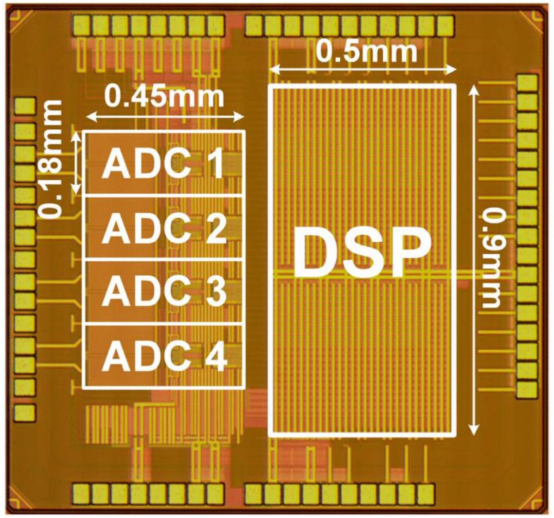
[10] Speech SoC by Seoul National University, 28nm LP, 0.9mm x 0.9mm



Technology	28nm LP
Core Area	0.9 × 0.9 mm²
On-Chip SRAM	35kB
Supply Voltage	0.8V – 1.1V
Frequency	2.5MHz
Power	740μW (@ 0.8V, 2.5MHz)
Efficiency	11.75μJ/frame

- **Park, Sungjin, et al. "22.8 A0. 81 mm 2 740μW Real-Time Speech Enhancement Processor Using Multiplier-Less PE Arrays for Hearing Aids in 28nm CMOS." 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023.**

[11] Speech SoC by University of Michigan & Intel, 40nm LP CMOS, 0.94mm²

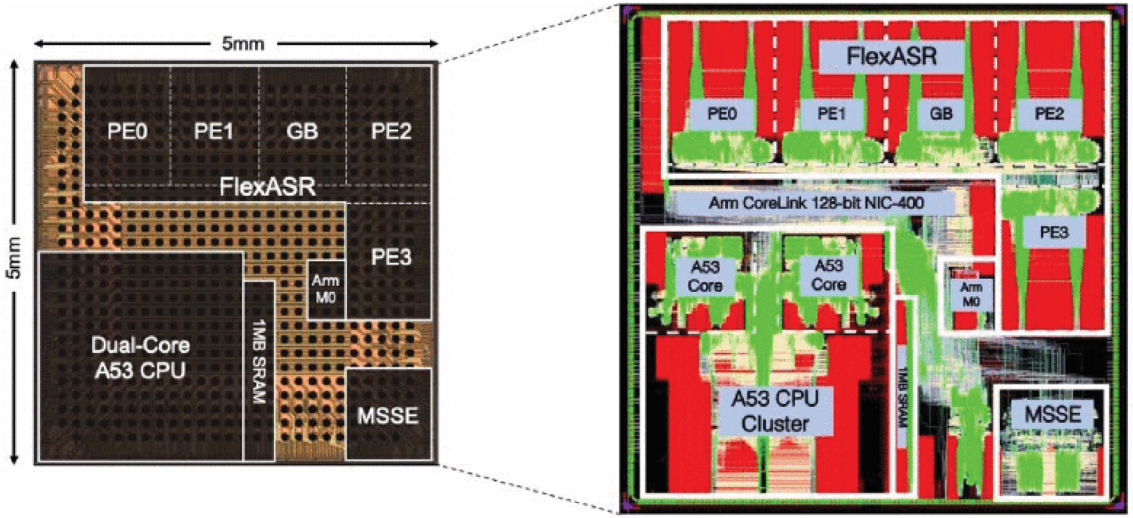


	This Work	[4] Kang	[1] Lee	[8] Liu	[3] Sainath	Google Home
Implementation	**Analog mic. +Single chip**	Analog mic. +Single chip	Analog mic. +Single chip	Digital mic. +Multichip	Digital mic. +Software	Digital mic. +Multichip
Technology	**40nm LP CMOS**	40nm LP CMOS	40nm GP CMOS	90nm CMOS	-	-
Area (mm²)	**0.94**	0.89	1.1	0.47	N/A	N/A
VDD (Analog / Digital)	**1.0V / 0.7V**	1.0V / 0.7V	1.0V / 0.55V	- / 0.33V	-	-
# Signal Sources	**4**	4	8	2	2	2
Functionality	**ADCs, adaptive beamforming, feature extraction**	ADCs, adaptive beamforming, feature extraction	ADCs, fixed beamforming, feature extraction	Adaptive beamforming (fixed steering), feature extraction	Adaptive beamforming, feature extraction, classification	ADCs, beamforming, feature extraction, classification
DR [8kHz BW]	**80 / 65dBA***	83dBA	85dBA	-	-	108dBA
BW	**8kHz**	8kHz	8kHz	8kHz	8kHz	8kHz
Beamforming Type	**Adaptive GABF**	Adaptive RGSC	Fixed delay-and-sum	Adaptive Griffiths-Jim	Adaptive filter-and-sum with trained coefficients	-
DOA Correction	**Yes**	No	No	No	N/A	N/A
Multi-mode Operation	**Yes**	No	No	No	No	N/A
Feature Type	**Log-Mel filter bank energy**	Log-Mel filter bank energy	Log-Mel filter bank energy	FFT-based Log filter bank	Convolutional long short-term memory DNN filter bank	-
# Features	**40**	40	60	8	128	-
AFE Power Consumption	**48 / 23μW***	367μW	0.81mW	-	-	-
DSP Power Consumption	**109 / 49μW****	280μW	3.1mW	0.1mW***	-	4.4mW****

*CTNSSAR / NSSAR, **GABF full / DTDAS only, *** Excludes ADCs, **** Calculated from datasheets, only includes MEMS microphones, ADCs.

- **Kang, Taewook, et al. "A Multimode 157μW 4-Channel 80dBA-SNDR Speech-Recognition Frontend With Self-DOA Correction Adaptive Beamformer." 2022 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 65. IEEE, 2022.**

[12] Speech SoC by Cornell University & Harvard University & Tufts University, 16nm FFC, 5mm x 5mm

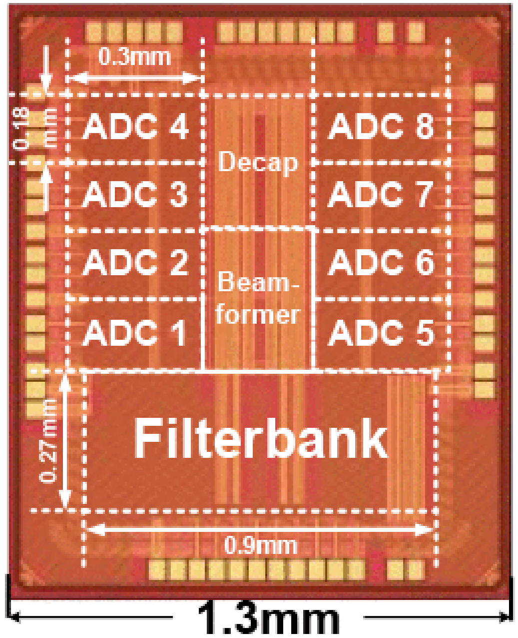


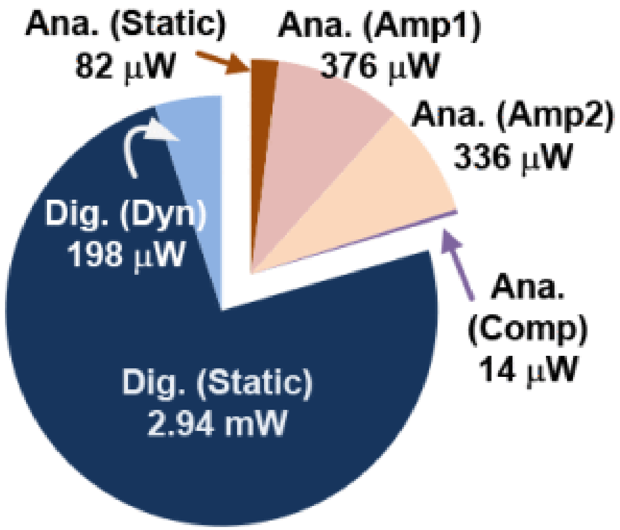
Technology	TSMC 16nm FFC
Die Area	$25mm^2$
Total SRAM	9.8MB
Gate Count	11M
Clock Domains	6
Power Domains	5
Supply Voltage	0.55 – 1V
Packaging	Flip-chip BGA-672

	Dual-A53	MSSE	FlexASR
Workloads	FFT, Application Logic	Markov Random Field	Attention-based seq2seq models
Data Type	FP64	FxP32	FP8
Area	$6.21mm^2$	$1.31mm^2$	$8.84mm^2$
SRAM	2.41MB	0.103MB	5.03MB
Voltage	0.55 - 1V	0.55 - 1V	0.55 - 1V
Frequency	354 – 775MHz	287 – 651MHz	130 – 573MHz
Power @ Fmax/0.8V	50.4mW	42.2mW	214mW
Power Eff.	-	4.33 – 17.6 Gsamples/s/W	2.6 – 7.8 TFLOPS/W

- **Tambe, Thierry, et al. "9.8 A 25mm 2 SoC for IoT Devices with 18ms Noise-robust Speech-to-text Latency via Bayesian Speech Denoising and Attention-based Sequence-to-sequence DNN Speech Recognition In 16nm Finfet." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021.**

[15] Speech SoC by University of Michigan & Intel, 40nm GP, 1.5mm x 1.3mm





General specifications	
Process	40nm GP
Die size	1.5x1.3mm^2
Package	48-pin QFN
Analog VDD	1.0V
Digital VDD	0.55V
Main Clock Freq.	2.048MHz
SD ADC specifications (1ch)	
Area	300x180µm^2
Sampling Freq.	2.048MHz
Bandwidth	8kHz
Power	91µW
SNDR	84dB
Input impedance	60kΩ

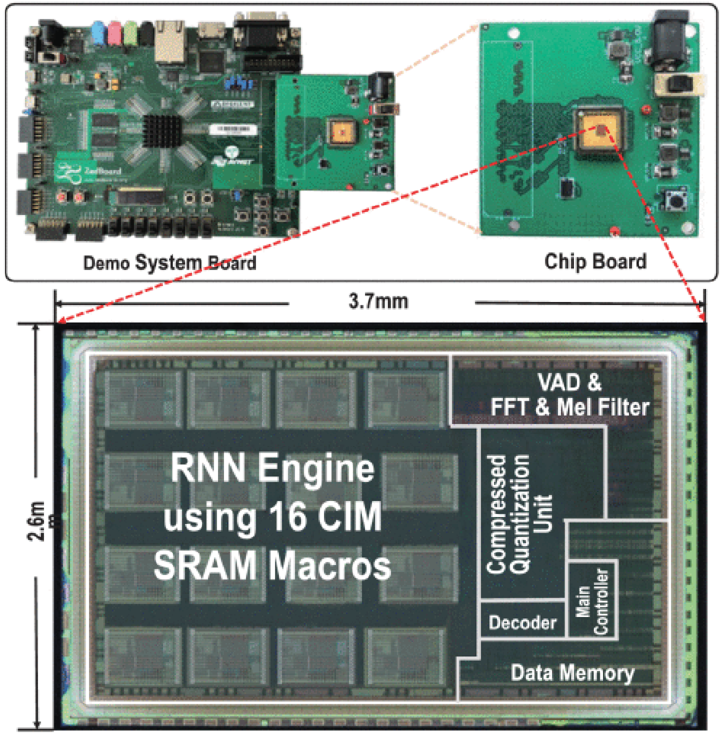
Beamforming	
Method	Freq. selective True time
No. Channels	8
Time resolution	3.42µs
Steering resolution (Linear 1" array)	2.62 °
Feature extractor	
Extraction type	Mixed signal
Feature type	Mel filter-bank
No. Features	60
Frequency range	70–8000Hz
Window time	25ms
Overlapping time	10ms
Output data width	24bit

- **Lee, Seungjong, et al. "An 8-element Frequency-selective Acoustic Beamformer And Bitstream Feature Extractor With 60 Mel-frequency Energy Features Enabling 95% Speech Recognition Accuracy." 2020 IEEE Symposium on VLSI Circuits. IEEE, 2020.**

[16] Speech SoC by Tsinghua University & National Tsing Hua University & TsingMicro Tech, 65nm CMOS, 3.7mm x 2.6mm

Chip Summary	
Process	65nm CMOS
Supply Voltage	0.9 - 1.1V
Frequency	5 - 75MHz
Core Size	$3.1\times2mm^2$
Die Size	$3.7\times2.6mm^2$
Logic Gates (NAND2)	1M
SRAM Capacity	10KB
SRAM-CIM Macro Capacity	16×4Kb
Latency per Inference	127.3us - 1.91ms
Energy per Inference	3.36uJ - 49.2uJ
Energy per Neuron	5.1 - 148.2 pJ/Neuron
Arithmetic Energy Efficiency	6.45 - 11.7 TOPS/W

(a) Chip Summary



(b) Chip Photograph

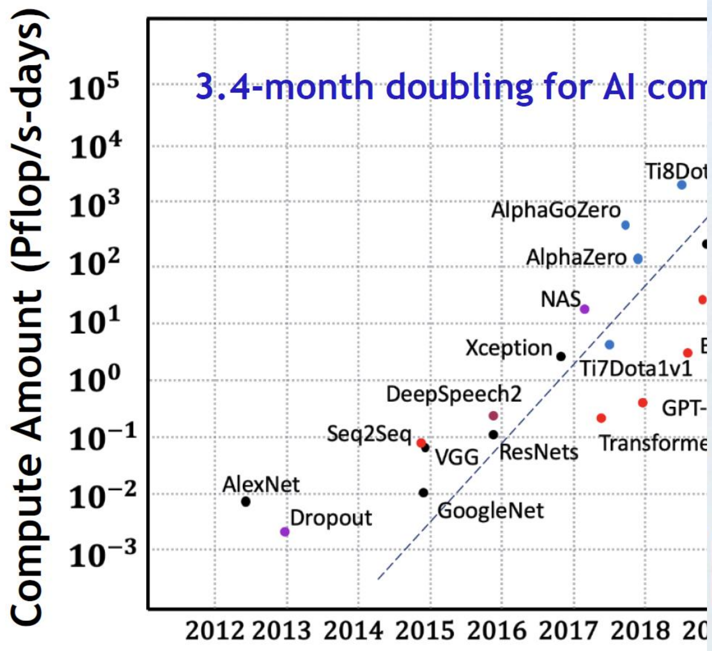
- Guo, Ruiqi, et al. "A 5.1 pJ/neuron 127.3 us/Inference RNN-based Speech Recognition Processor using 16 Computing-in-Memory SRAM Macros in 65nm CMOS." 2019 Symposium on VLSI Circuits. IEEE, 2019.

Let’s step back and think a little…

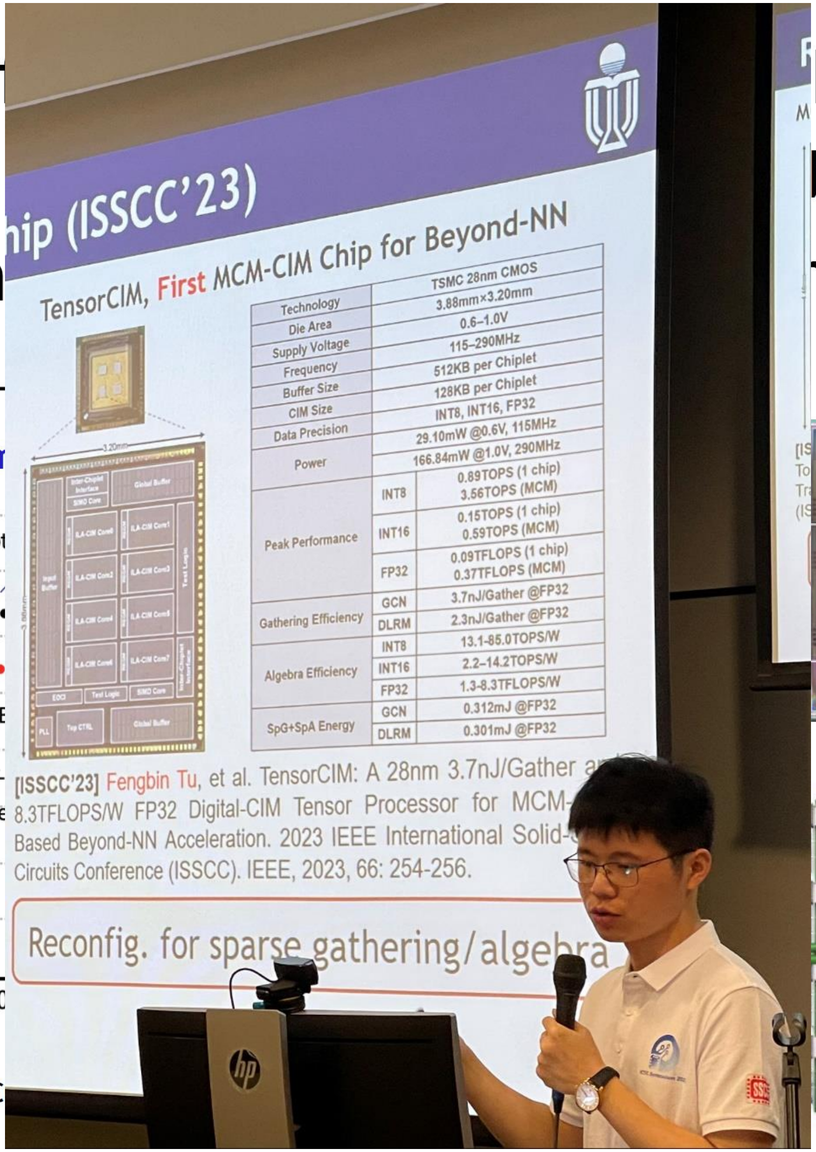
(Remember: don’t run in the wrong way)

AI Processors: The Strong Body Under the Master Mind

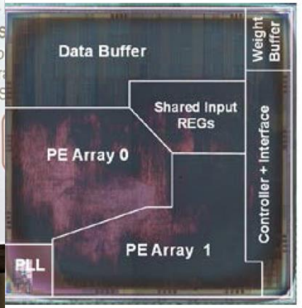
- The development o… high compute amount.
  - 3.4-month doubling … bling of Moore's Law.
- AI chips construct h… uctures of the AI era.



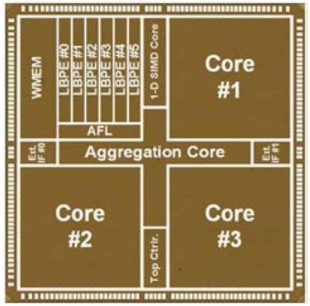
https://openai.com/blog/ai-and-c



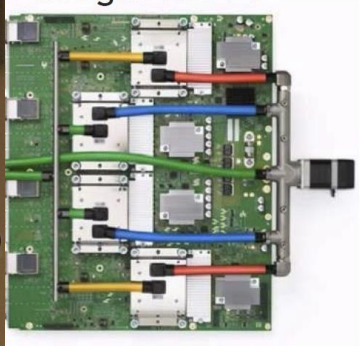
THU Thinker



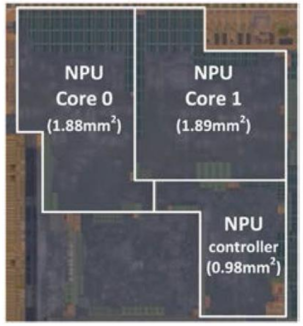
KAIST UNPU



Google TPUv4

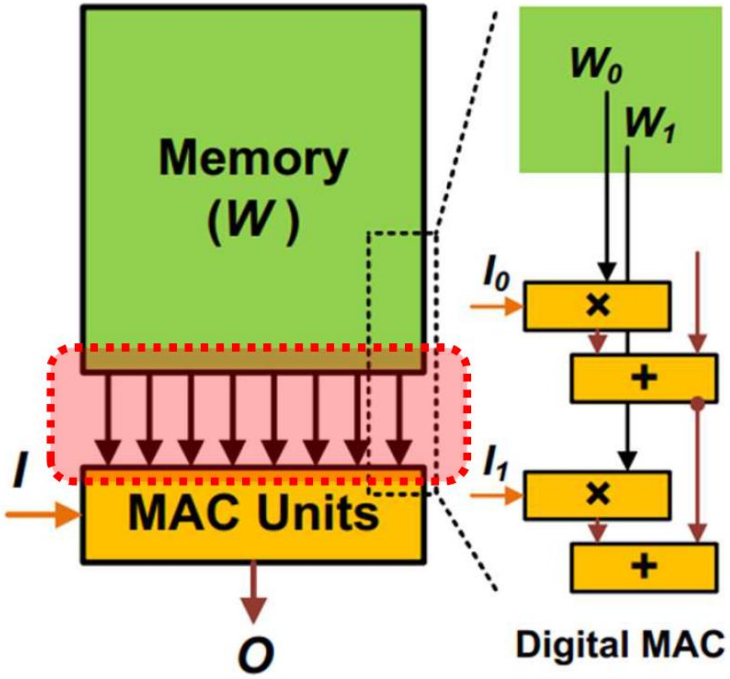


Samsung NPU

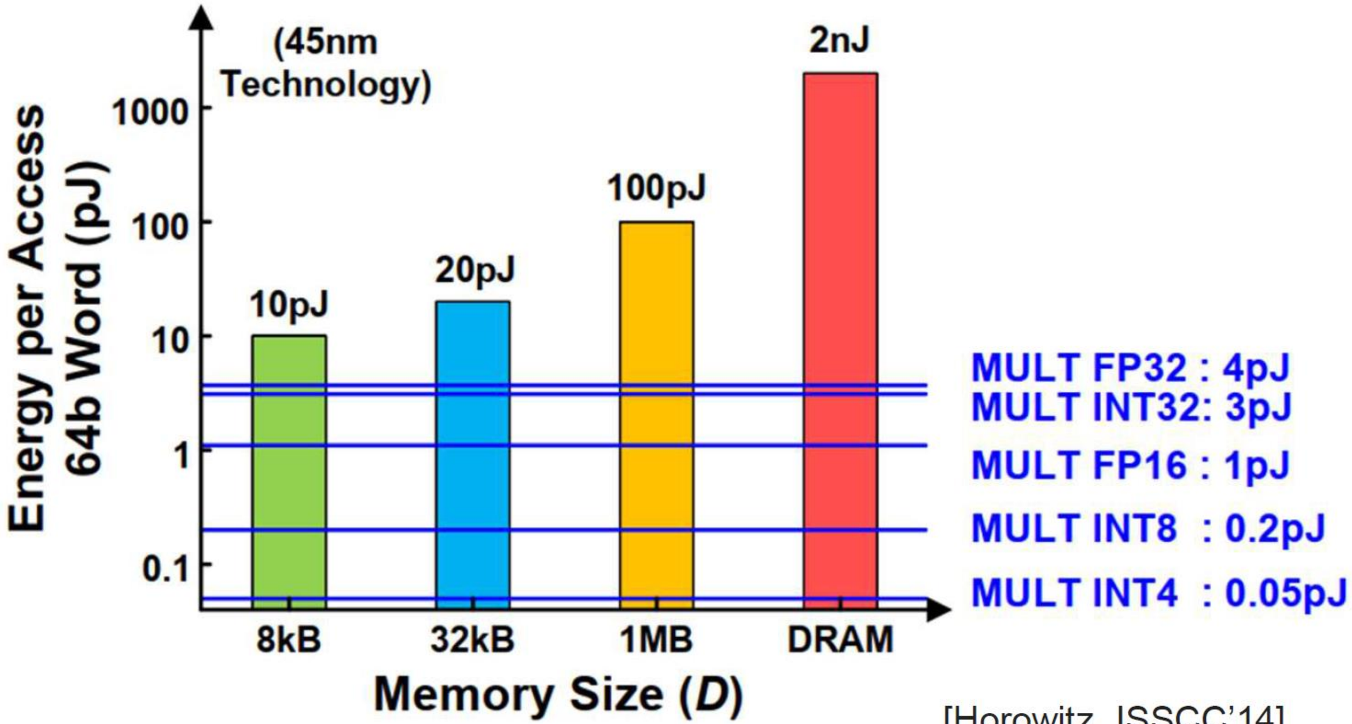


The I/O Bottleneck in Von Neumann AI Chip

- Due to the increasing size and computation of AI models, conventional digital AI chips usually suffer from massive data movements between separate compute and memory.



Memory Access Energy > Compute Energy

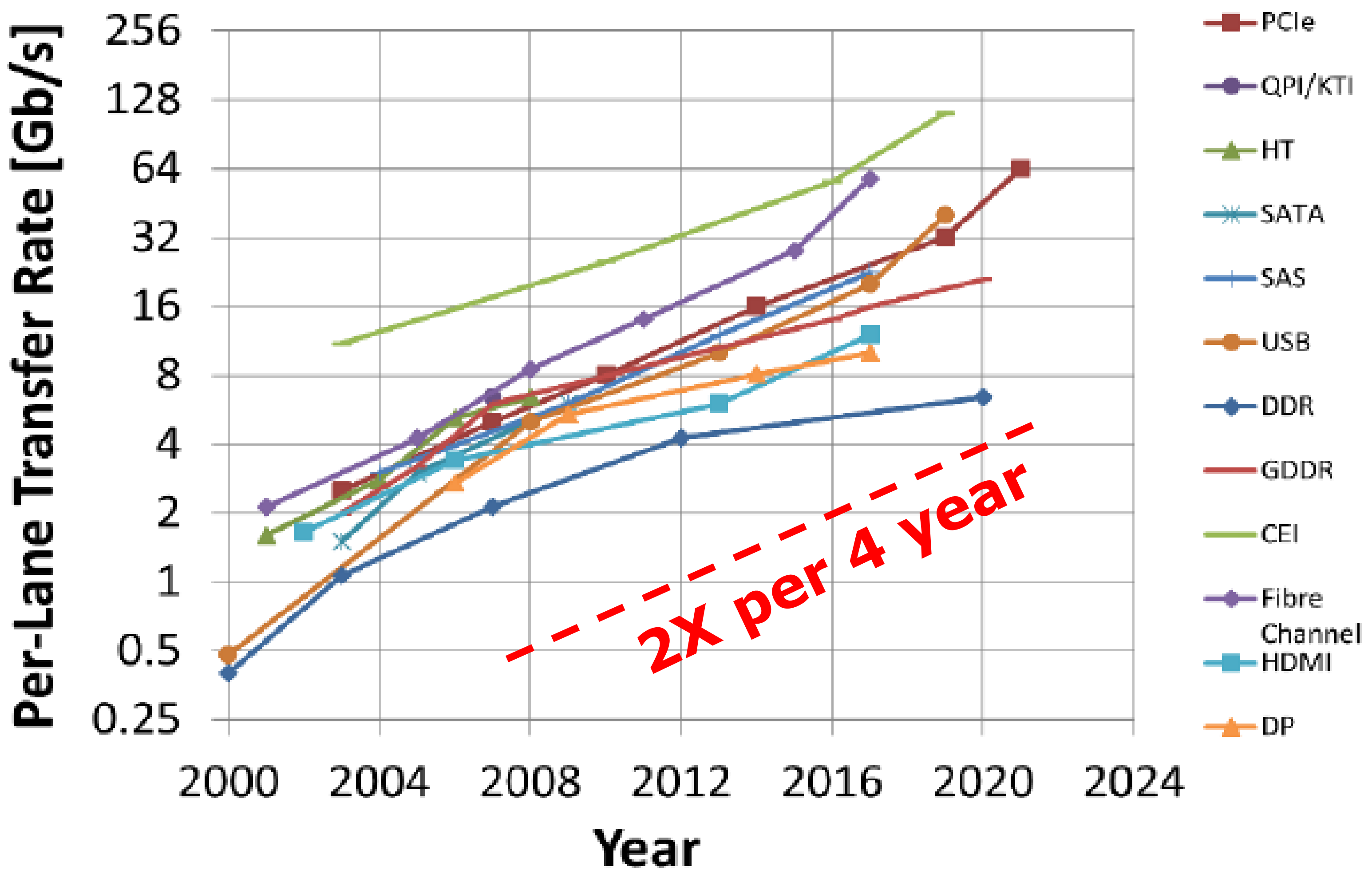


[Horowitz, ISSCC'14]
[Jia et al., ISSCC'21]

But the interconnects (high-speed I/Os) are not energy-efficient and dense enough

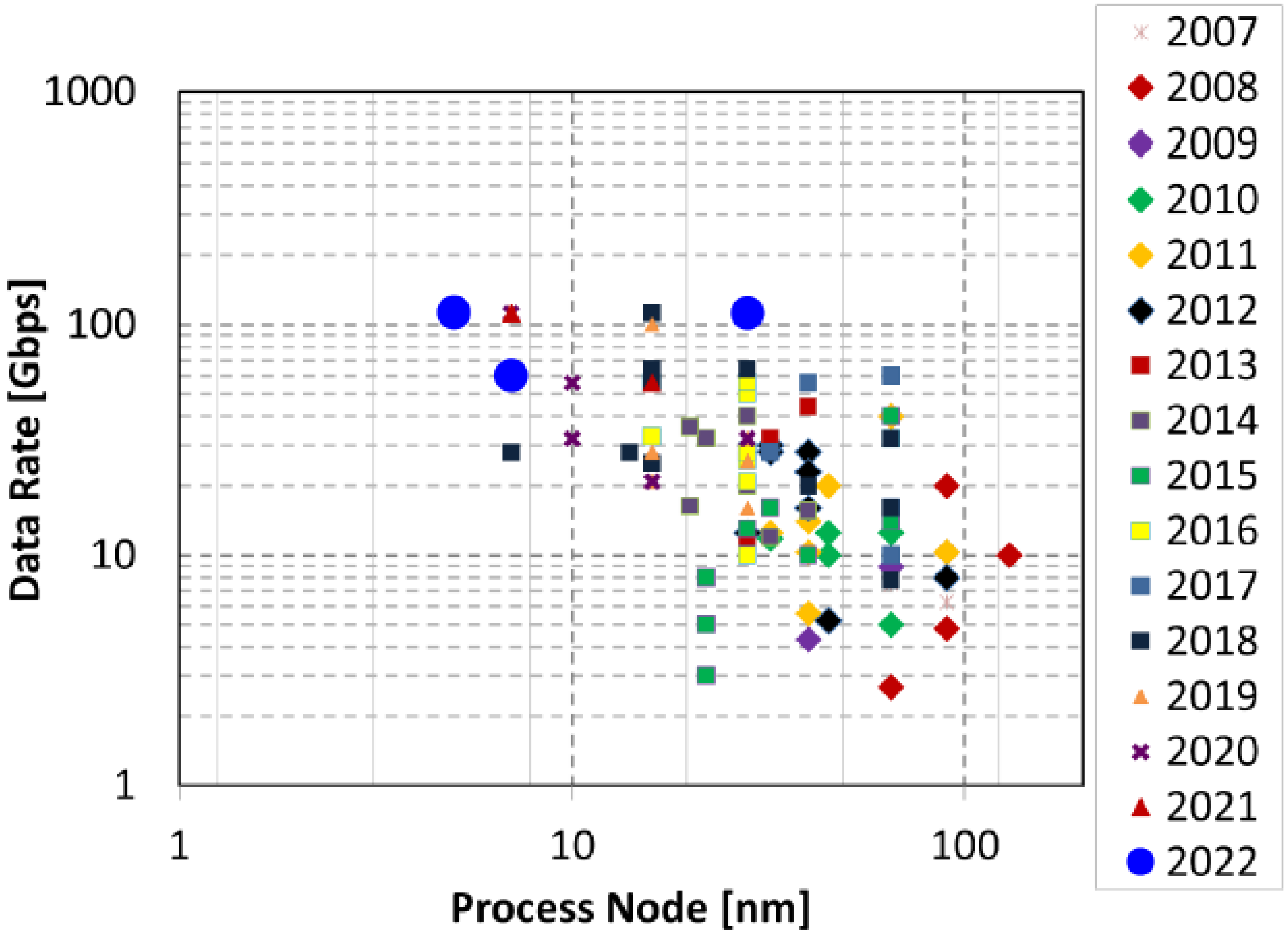
Hence, CIM design has taken the center stage

High-Speed I/O Trend: Per-Lane Data Rate Growth



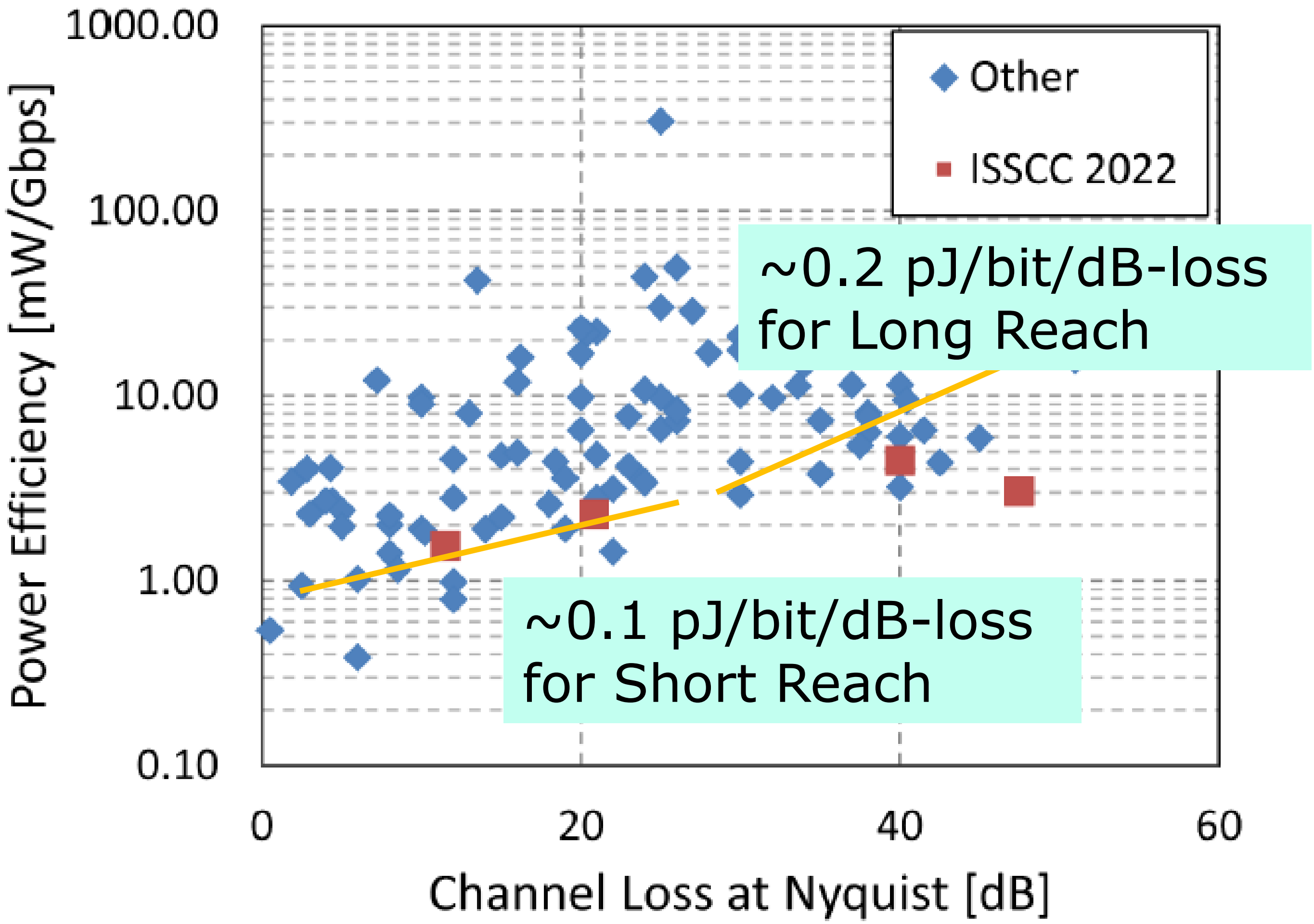
- Data-rate per pin has approximately X2 every four years across various I/O standards ranging from DDR, to graphics, to high-speed Ethernet.

Data Rate vs. Process Node



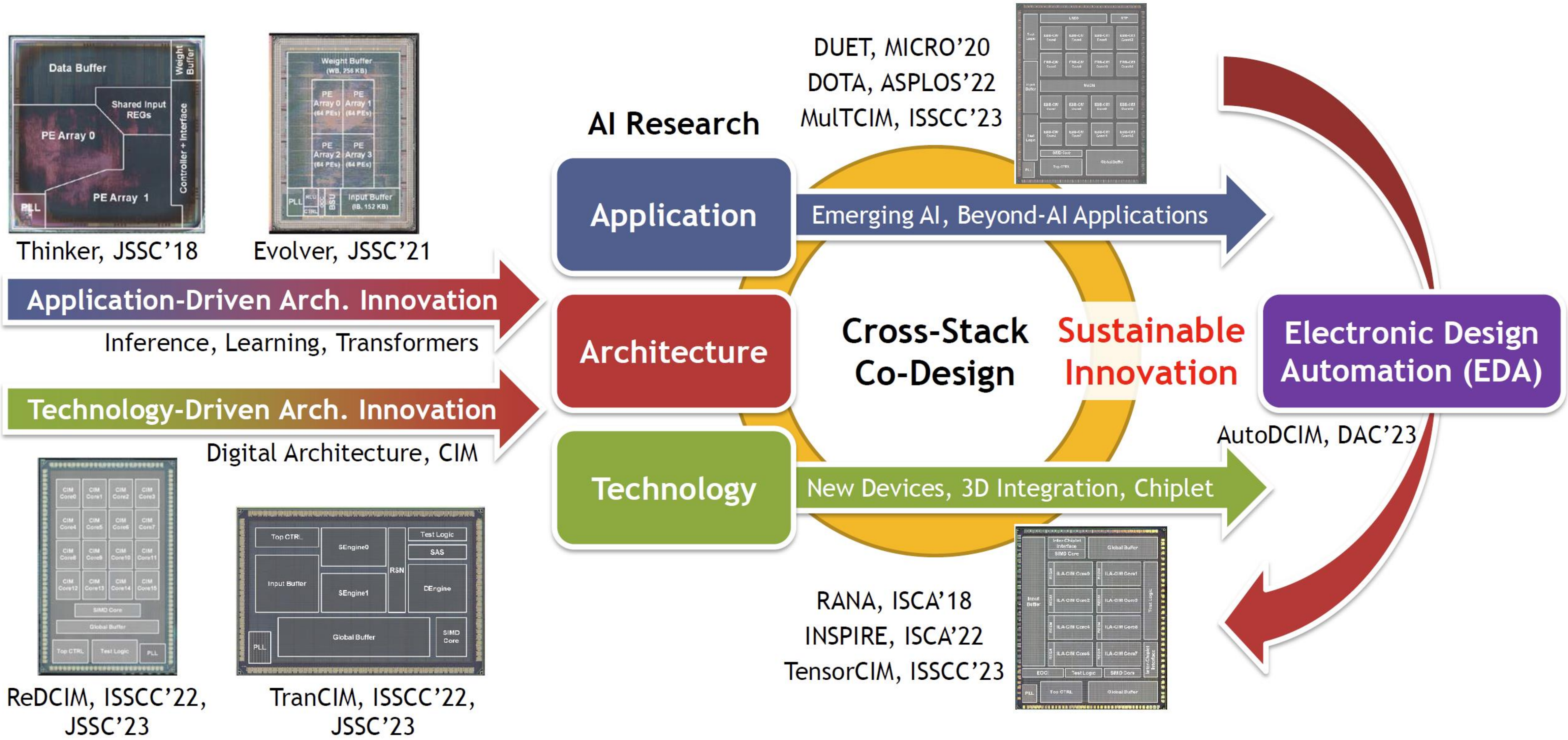
- PAM4 transceivers have kept pace at 56Gb/s and 112Gb/s while taking advantage of CMOS scaling below 10 nm for more aggressive channel loss compensation.

Trends: Bit Efficiency vs. Channel Loss



- Channel loss compensation by different equalization techniques is essential
- Power efficiency better than ~1-2 pJ/bit only for Short Reach with < 20 dB loss
- Power efficiency ~5-10 pJ/bit for Long Reach with up to 50 dB loss

Outlook for AI Chip Research Directions



Smart Construction Robots

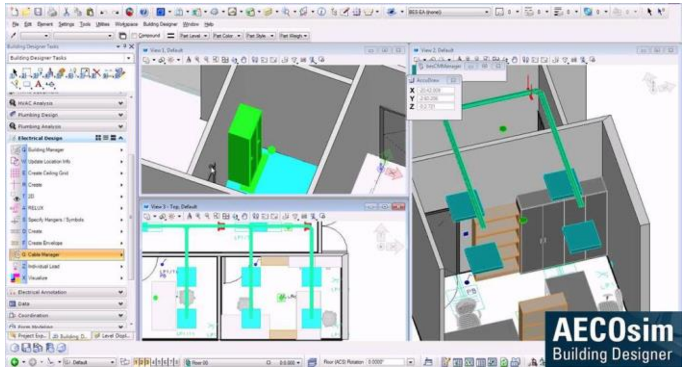
- Robot + AI: a new flashpoint for the development of the construction industry *

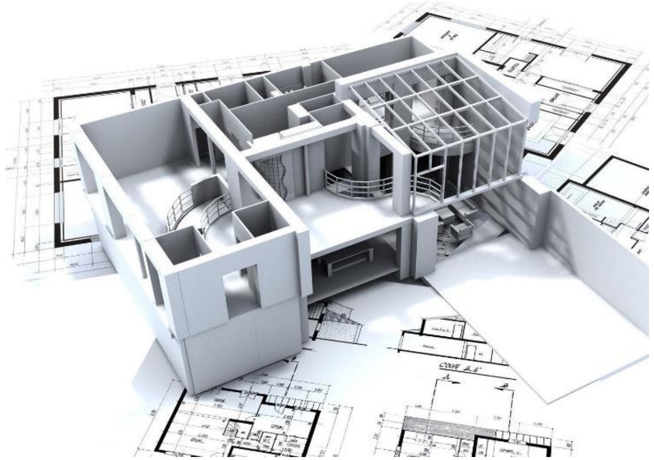





*https://www.bzlrobot.com/list.html

- BIM: the basis for efficient, fast, and autonomous intelligent operations
- The future market demand and scale are huge: reach an estimated $9 billion by 2025 with a CAGR of 9% to 11% from 2019 to 2025 *

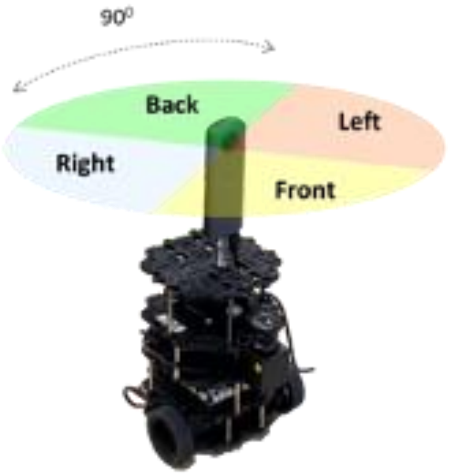


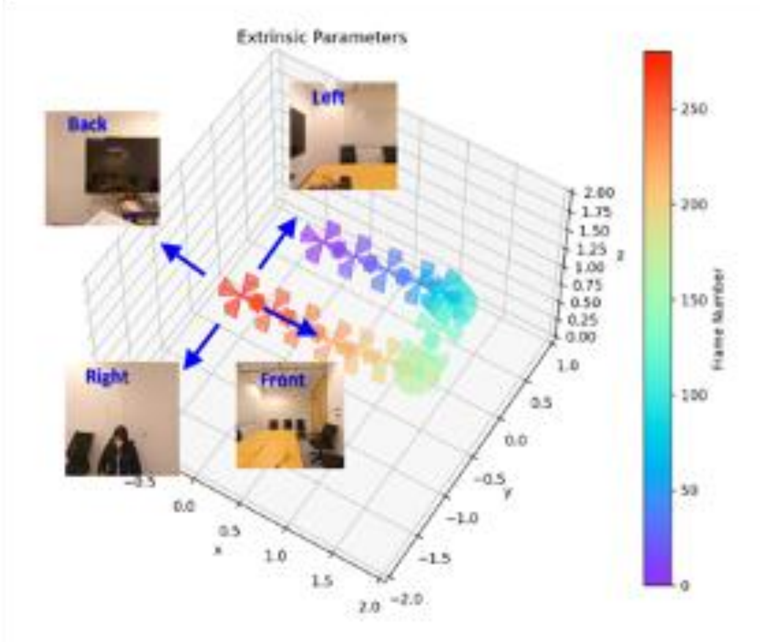


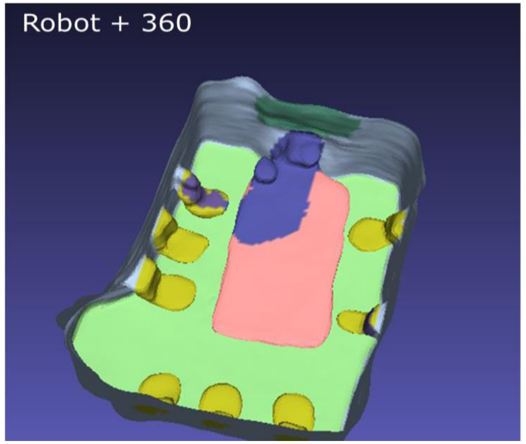
*https://--ResearchAndMarkets.comwww.businesswire.com/news/home/20210412005399/en/Global-Building-Information-Modeling-Market-2021-2025-Trends-Forecasts-and-Competitive-Analysis-

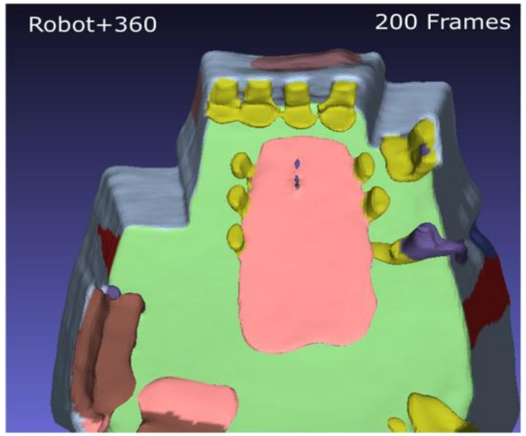
Enabling Technologies in Our Smart Construction Robots

1. Positioning: VLP   2. Path planning: RL   3. 3D scene: CDRNet & 360 Camera



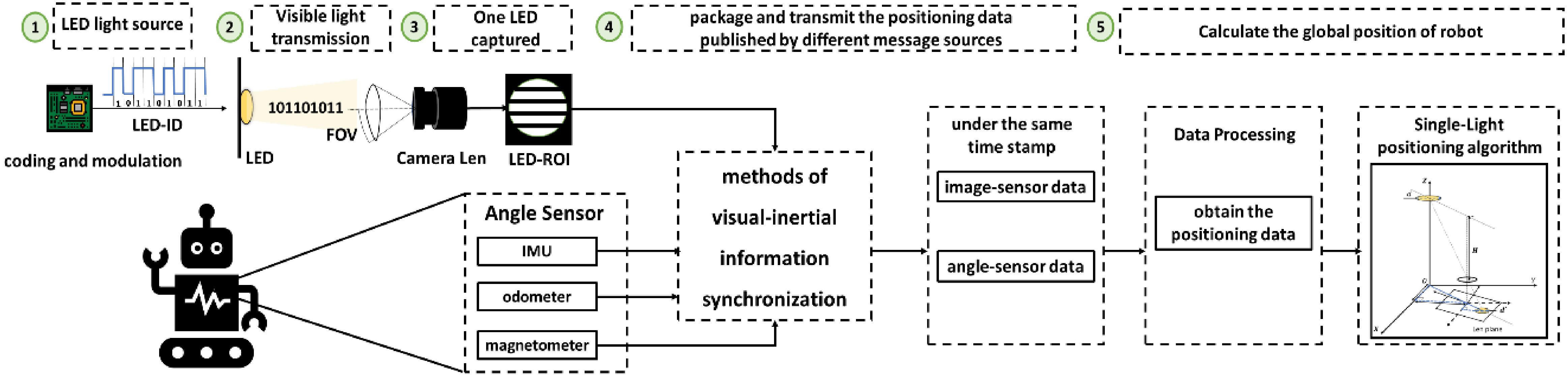


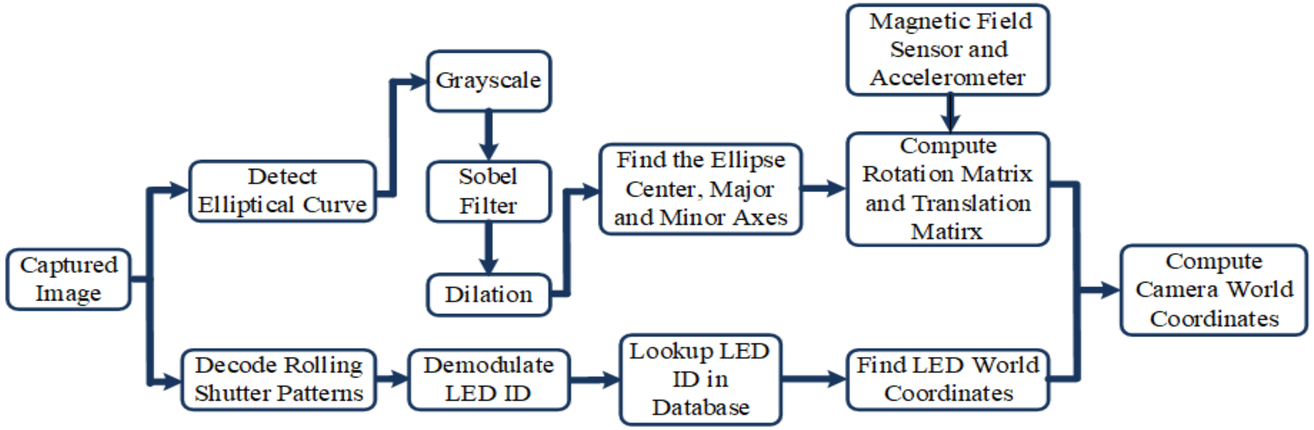


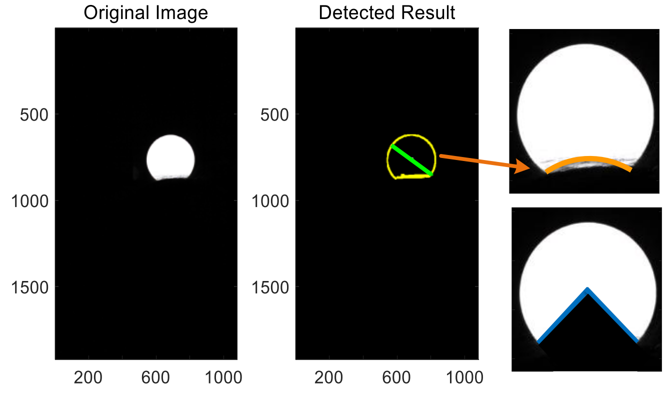


- Visualization of perspective views and the corresponding pose using VLP
- 3D reconstructed model using CDRNet with input images captured using a 360 camera mounted on the robot
- Reinforcement learning for multi-robot path planning and collaborative task

Visible Light Positioning (VLP) for Smart Construction Robots

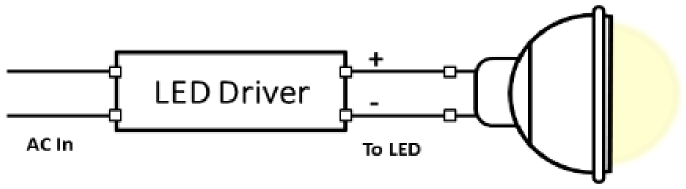


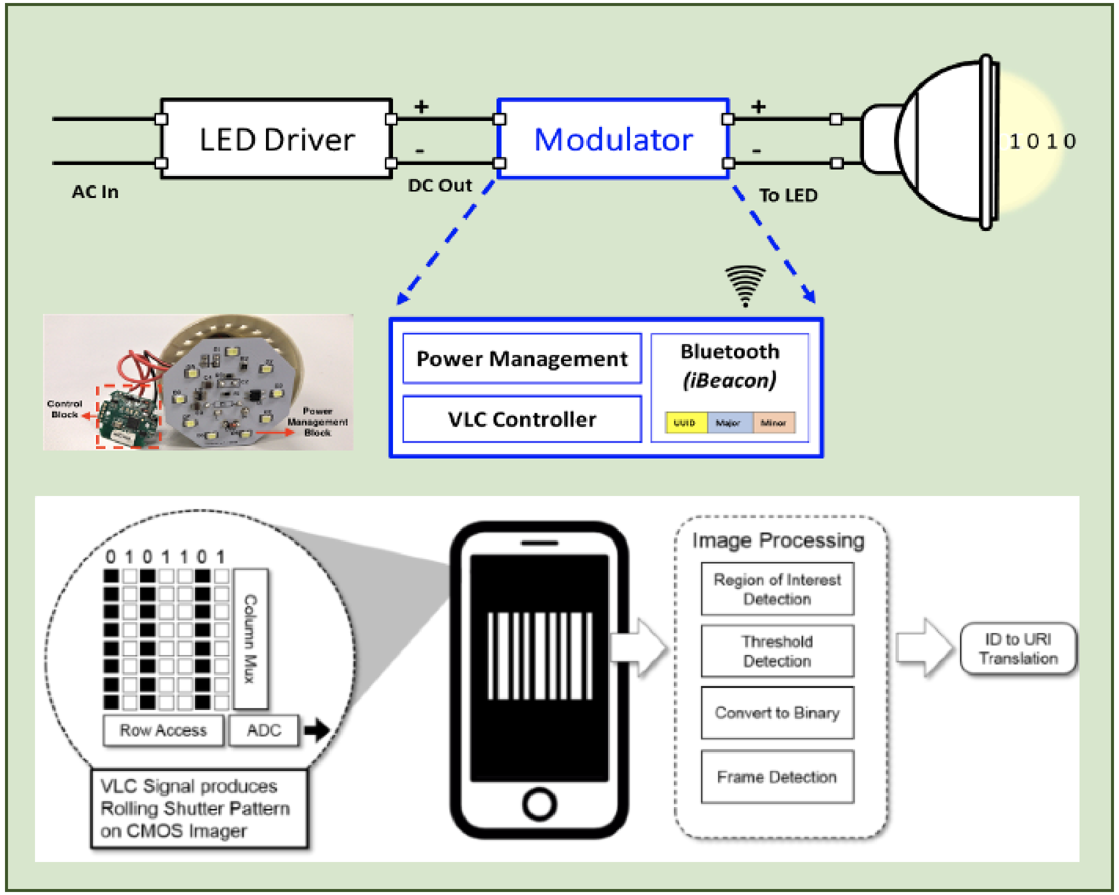




W. Guan *et al.*, “Robust Robotic Localization Using Visible Light Positioning and Inertial Fusion,” *IEEE Sensor Journal*, 2022.

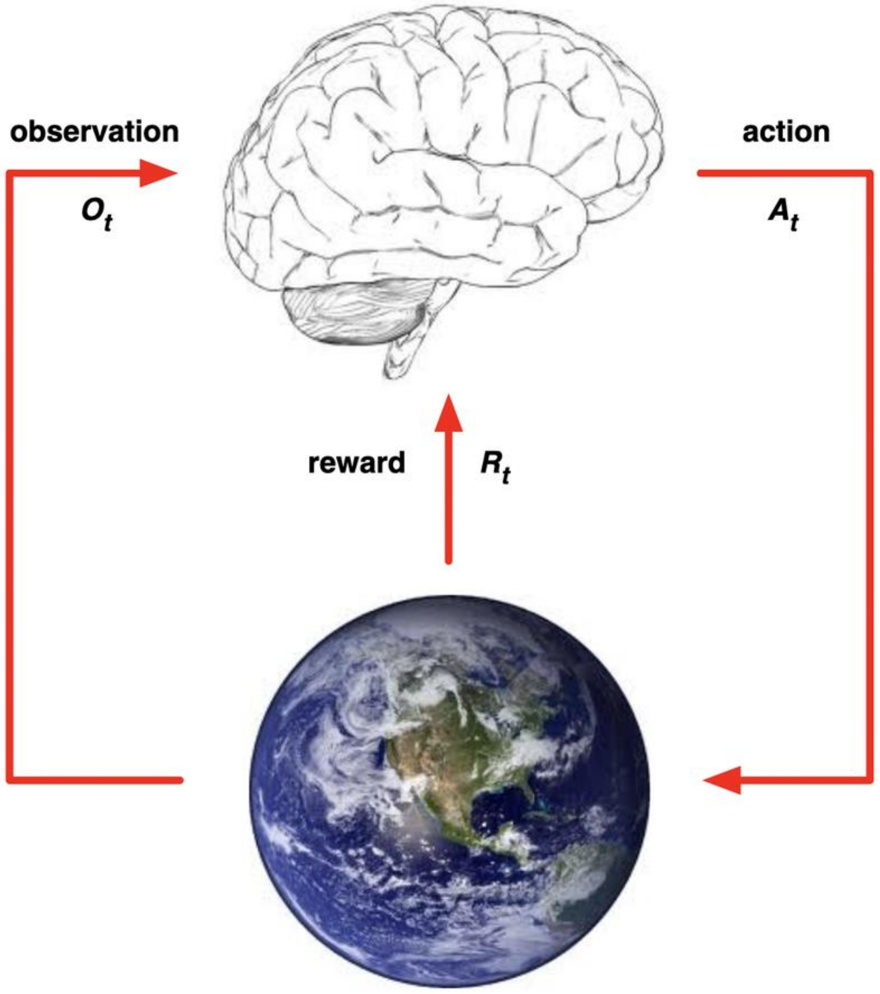
Enabling VLP in Ordinary LED Lighting





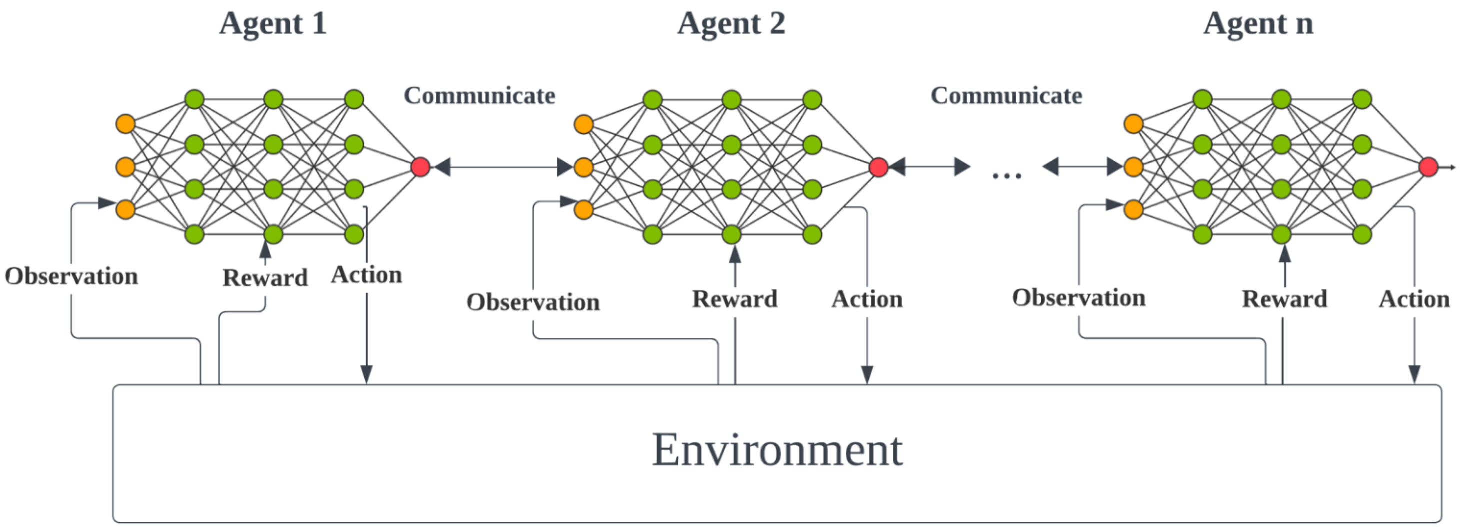
Reinforcement Learning (RL)

- An approach where an agent learns to make optimal decisions by interacting with an environment through trial and error
- Enables adaptive and versatile behavior across diverse tasks and domains
- Advantages
  - Ability to learn from experience and adapt to changing environments
  - Can handle complex decision-making problems
  - Offers potential for breakthroughs in various domains such as autonomous systems



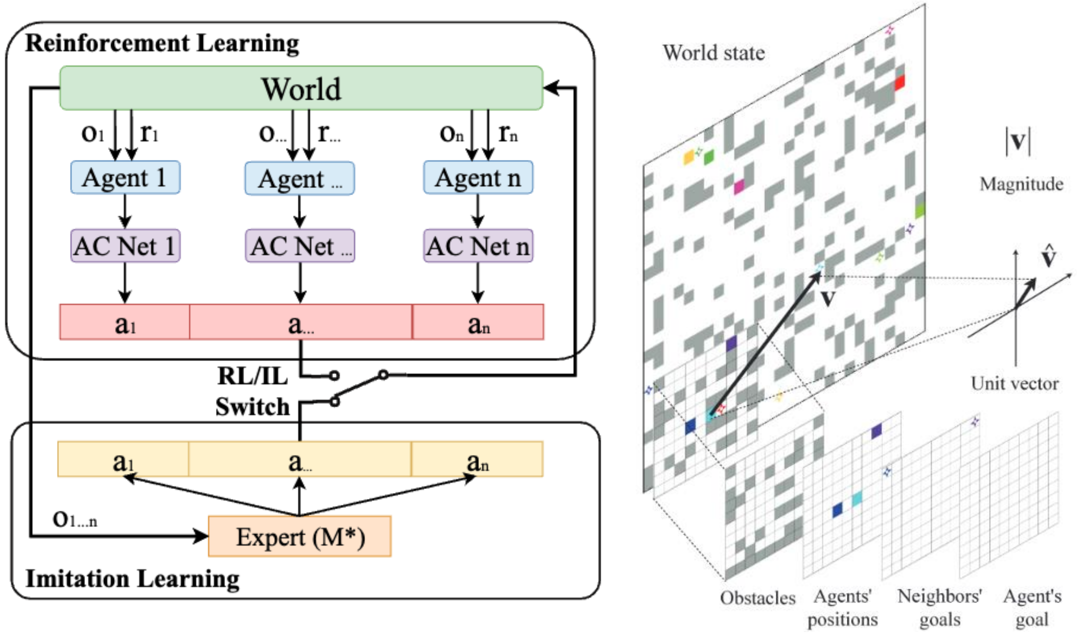
https://www.researchgate.net/figure/Figure-from-deep-minds-introduction-to-reinforcement-learning-course-Silver-2015-To_fig2_368665532

Multi-Agent Reinforcement Learning (MARL)

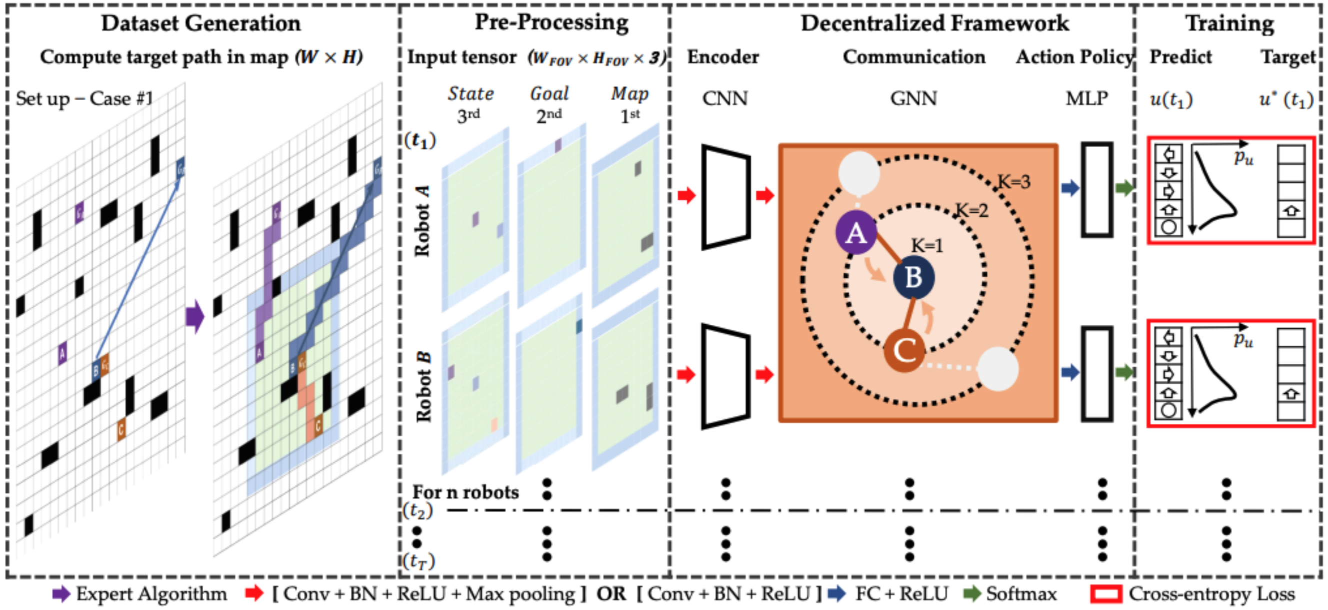


- ❑ **Trains agents to make decisions in a coordinated manner**
- ❑ **Enables agents to handle complex scenarios, strategic interactions in various domains**
- ❑ **Important applications**
  - ➢ **Multi-robot coordination: Coordinating actions of autonomous robots for tasks like exploration, surveillance, or swarm behavior**
  - ➢ **Traffic management: Optimizing traffic flow and reducing congestion in for self-driving vehicles**
  - ➢ **Resource allocation: Optimizing resource distribution in scenarios like power grids**

Multi-Agent Path Finding with RL



***PRIMAL*, Sartoretti et al., RA-L'19**



***GNN for Decentralised Multi-agent Path Planning*, Li et al., IROS'20**

❑ Limitations

- Using broadcast communication which caused waste of bandwidth and energy
- Rely on expert algorithms that do not scale well on complex scenarios

Results on Optimized MARL

- Scenario: Multi-agent path finding
  - Warehouse robotics
  - Multi-robot exploration in hazardous areas
- Experiment on optimizing Field-of-View (FOV)
  - FOV impacts agents' perception, navigation, opportunity awareness and communication in MARL
  - FOV affects coordination and communication through overlapping views
- Tested the performance in metrics such as success rate, number of communications, etc.
  - FOV size does not always correlate with improved performance; increasing it may weaken performance
  - Smaller FOV sizes can be more effective as performance does not decrease proportionally with FOV size

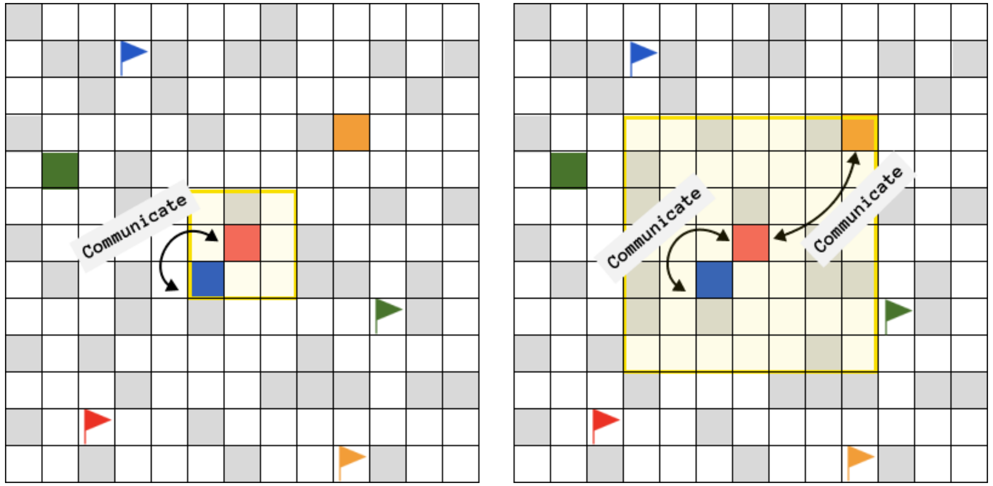
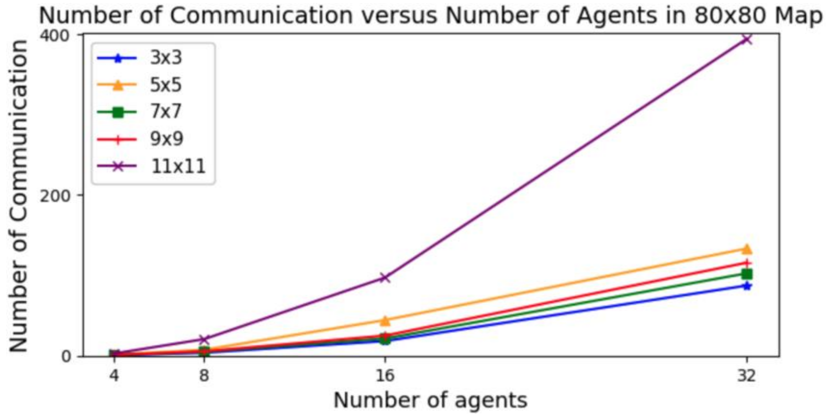
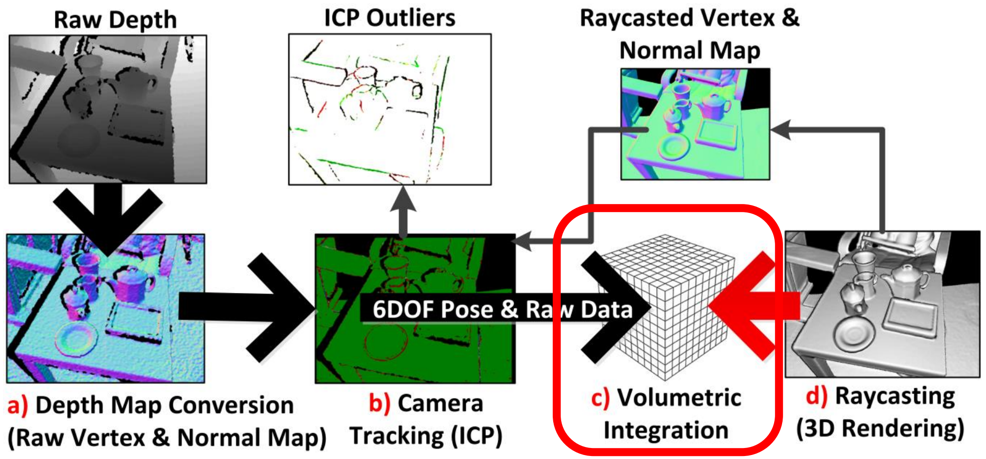


Illustration of 3x3 FOV (left) and 7x7 FOV (right)

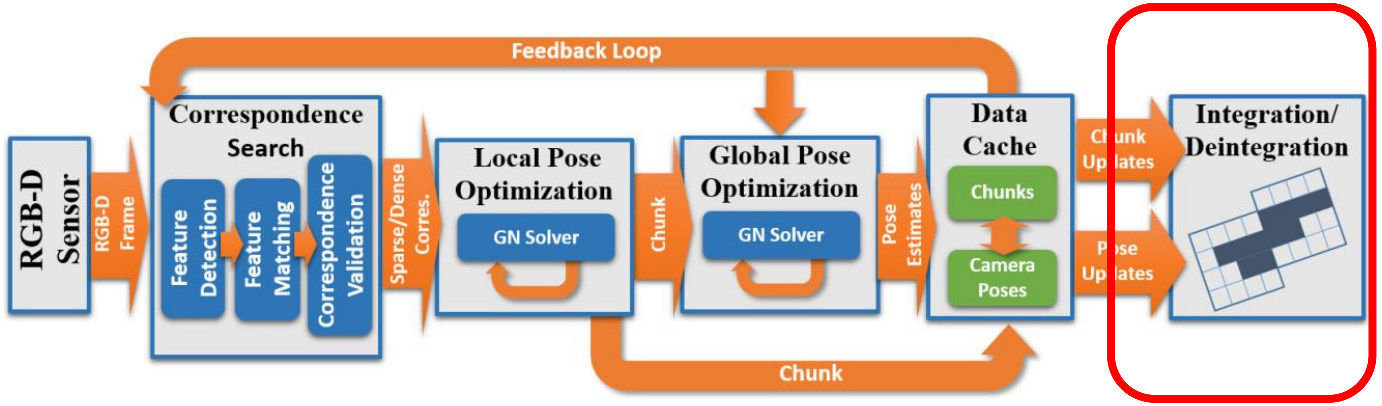


H. C. Cheng, L. *Shi, C. P. Yue*, "Optimizing Field-of-View for Multi-Agent Path Finding via Reinforcement Learning: A Performance and Communication Overhead Study," *62nd IEEE Conference on Decision and Control (CDC)*, in press.

Construct 3D Models by Fusing Depth Map

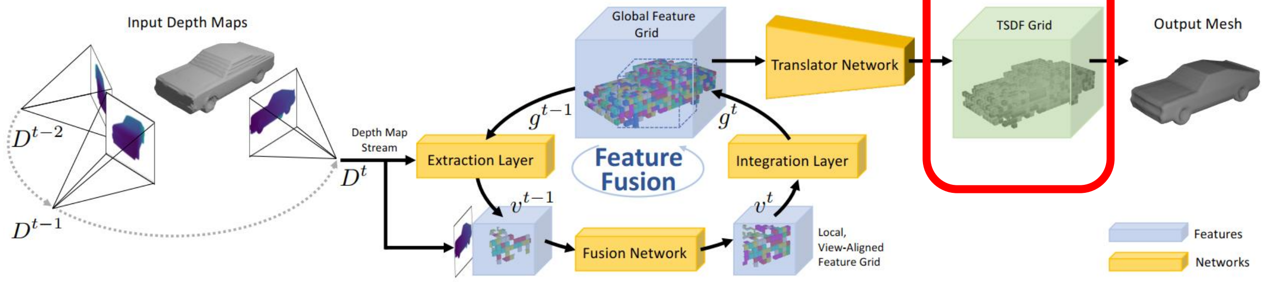


KinectFusion, **Newcombe et al., ISMAR'11**



BundleFusion, **Dai et al., ACM ToG'17**

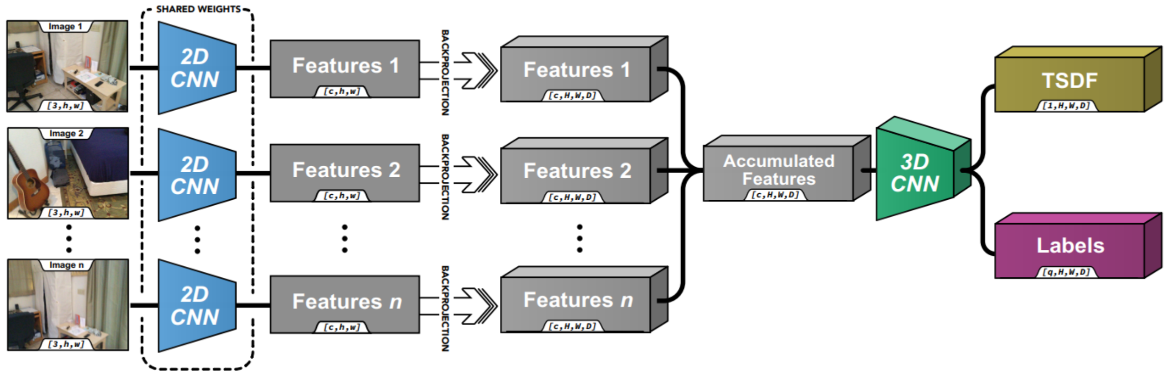
Truncated Signed Distance Function (TSDF)

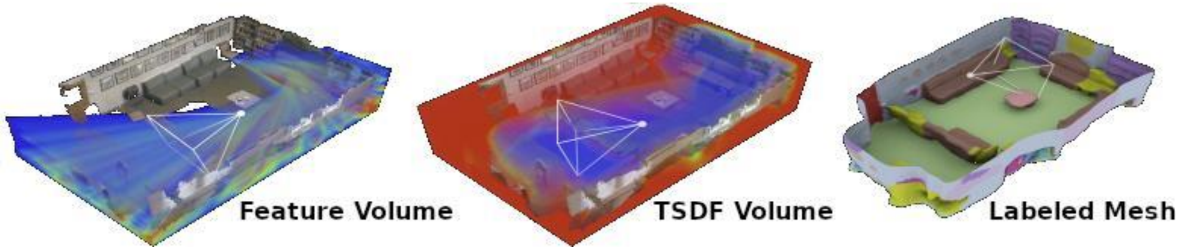


NeuralFusion, **Weder et al., CVPR'21**

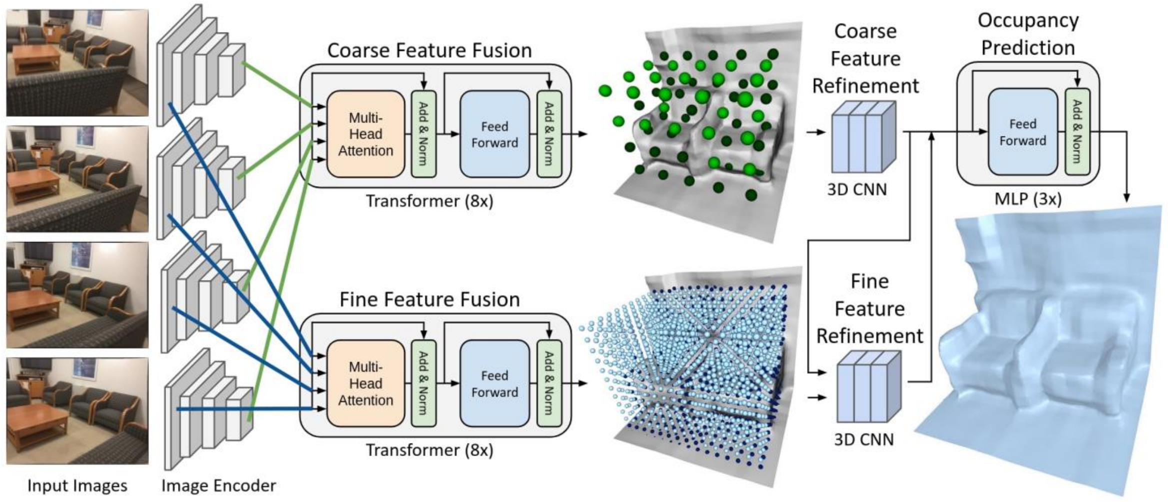
- **Limitations**
  - **Range sensor is infeasible for light-weight hardware**
    - **Costly**
    - **Power hungry**
  - **Depth map suffers from noise and low albedo issues**

Vision-Based Learning Methods





***Atlas*, Murez et al., ECCV'20**



***TransformerFusion*, Božic et al., NeurIPS'21**

- **Limitations**
  - **Implicitly learn TSDF without any prior knowledge**
  - **Global volume average, not real-time**
    - Not a fit for the real-world SLAM usage

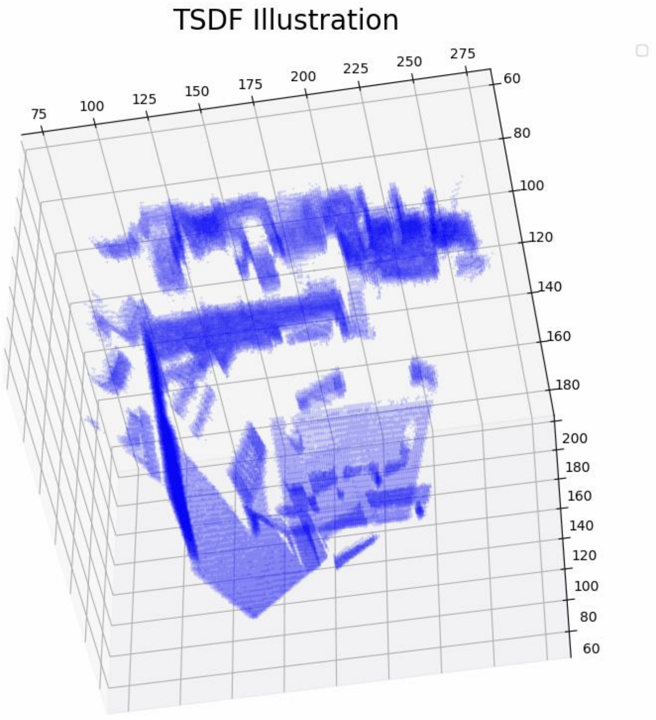
Prior Works and Limitations on 3D Reconstruction

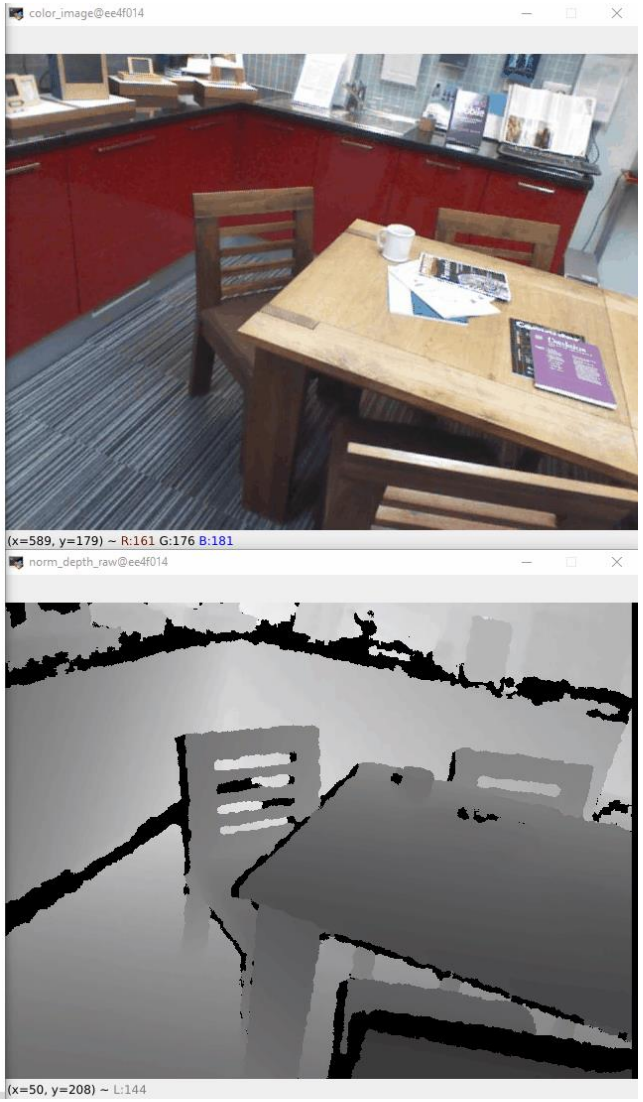
- Recent 3D reconstruction works are mostly based on cameras equipped with depth sensors (RGBD camera or LiDAR)
- Construct truncated signed distance function (TSDF) representation for surface reconstruction
  - Microsoft Kinect (KinectFusion)
- Limitations
  - Cost ☹
  - Latency ☹

See video

5x speed

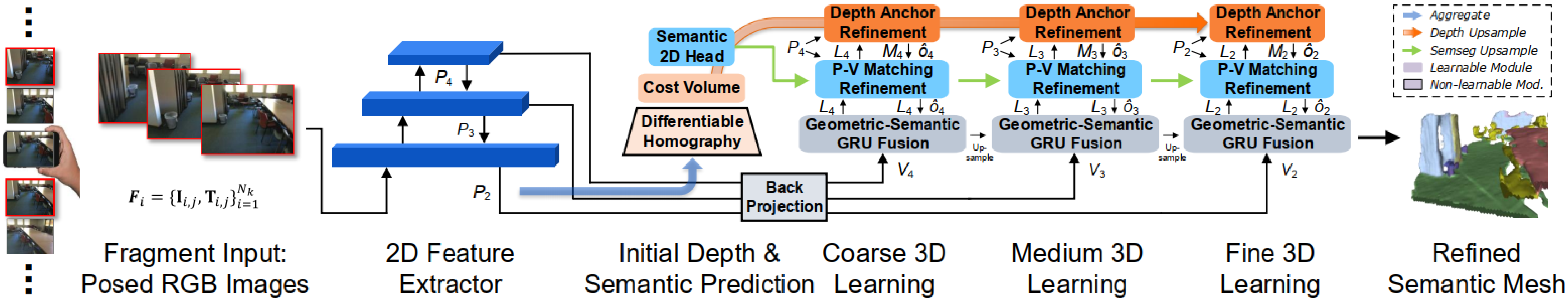
RGB Image





Depth Map

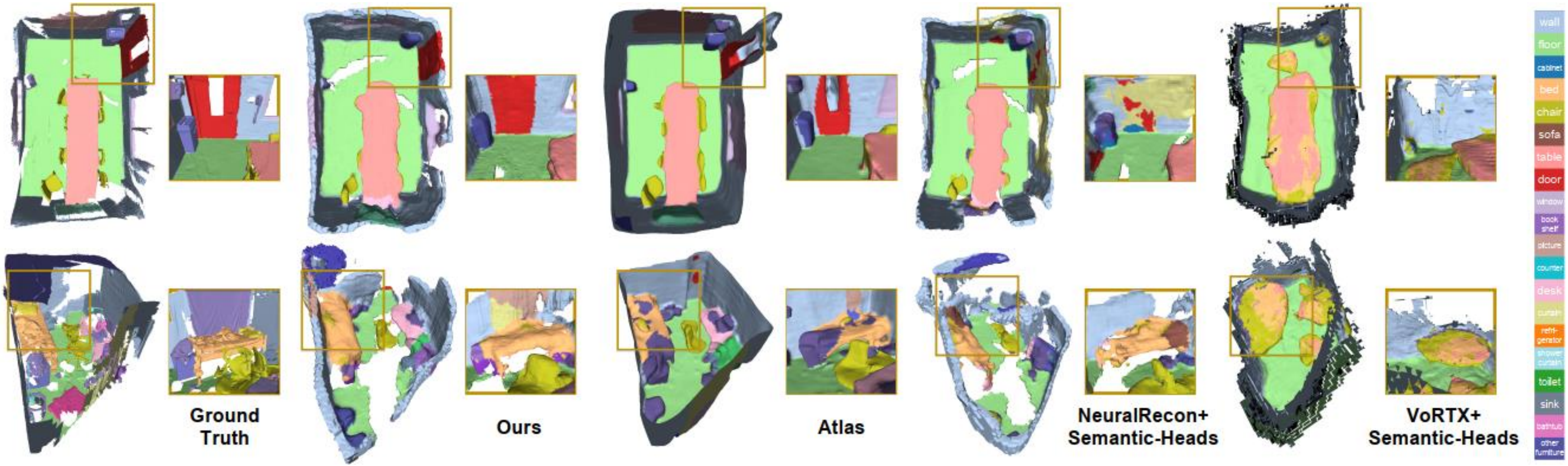
Proposed 3D Perception Pipeline: CDRNet



- Sparse 3D convolution to improve efficiency
- Temporal: Using gated recurrent units (GRU) to infer a local truncated signed distance field (TSDF) volume and merge into global features
- Spatial: 2D explicit inference as a prior knowledge to refine the 3D feature

Z. Hong *et al.*, "Cross-Dimensional Refined Learning for Real-Time 3D Visual Perception from Monocular Video," *IEEE/CVF ICCV workshops*, 2023.

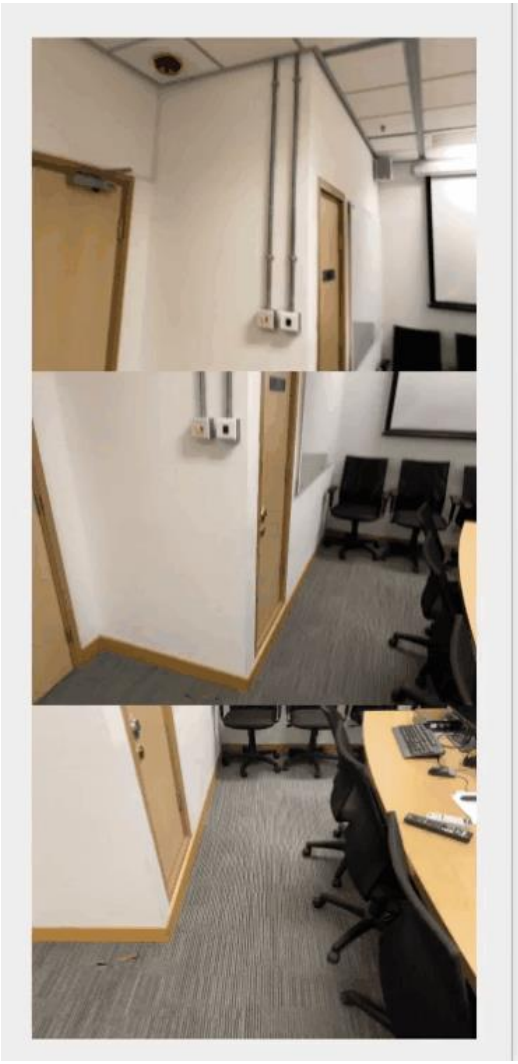
Results on A Large-Scale Public Dataset

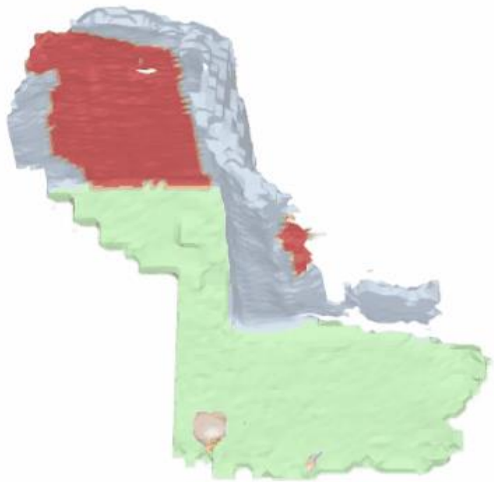


- Evaluated on the ScanNet dataset
  - 2.5M RGB-D images and 1513 3D scans in total
- Achieves the best 3D perception performance, sometimes surpasses the ground truth

A. Dai *et al.*, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," *IEEE/CVF CVPR*, 2017.

Real-Time 3D Perception Using CDRNet





See video

FPS: 158, Fragmented Mesh Rate: 2.38/sec

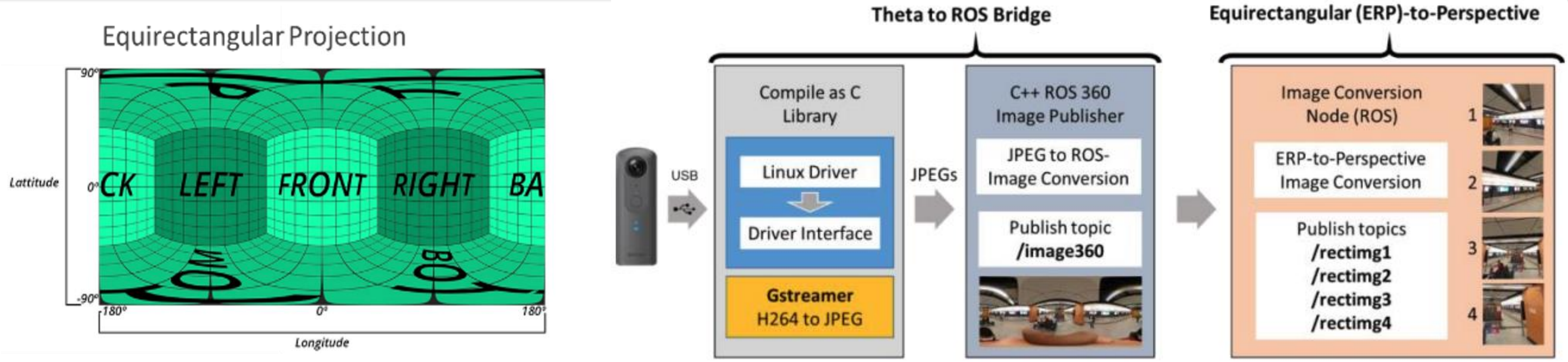
3D Reconstruction using 360 Camera

- Advantages of using 360 cameras
  - Captures full spherical view, enabling a comprehensive coverage of the environment
  - More time-efficient in capturing the environment than perspective cameras
- Using 360 cameras is a common practice in industries for monitoring purposes (e.g., BIM)
- Challenges
  - Calibrating 360 camera
  - Integration with existing deep learning pipelines and workflows

The 360 Camera used in our experiments

Equirectangular Projection (ERP)

ERP Image Conversion

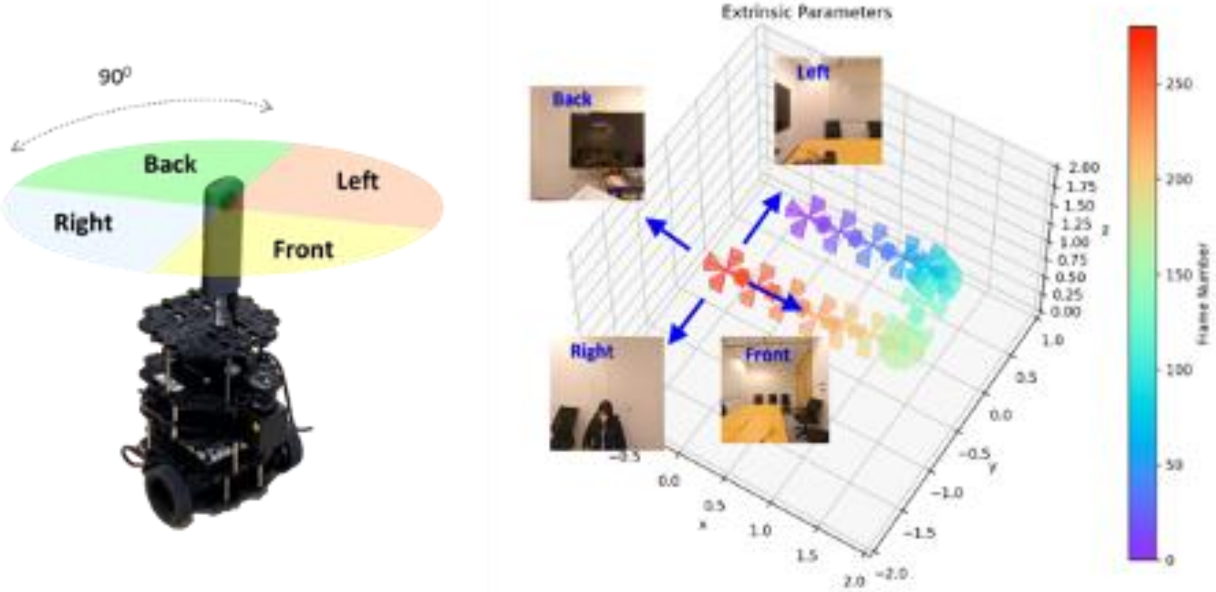


- Convert Equirectangular Projection (ERP) images to 4 perspective images
- Each view has a field of view of 90°
- Treat the 4 views as part of cube-maps, resembling four virtual cameras pointing in different directions
- The final perspective images are compatible with established deep learning pipelines

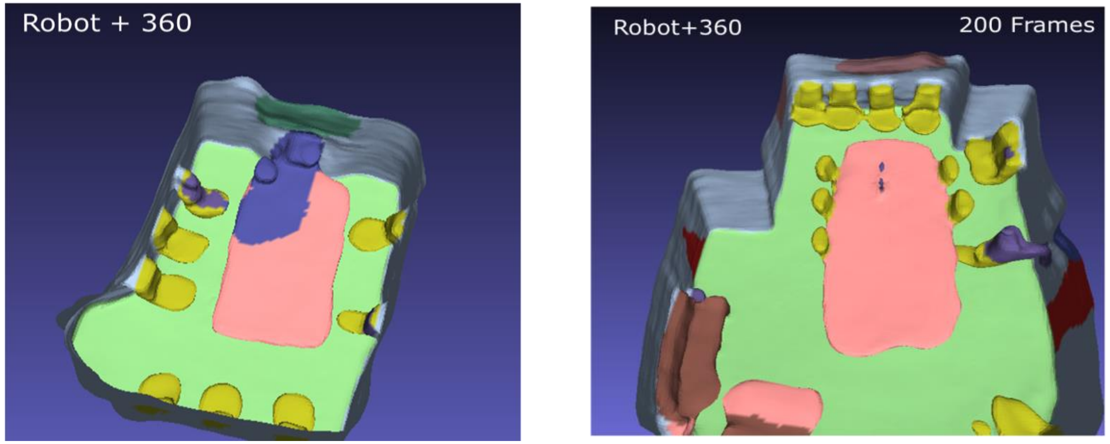
H. C. Cheng *et al.*, "Leveraging 360° Camera in 3D Reconstruction: A Vision-based Approach," *2023 2nd International Conference on Video and Signal Processing*, in submission.

Pose Estimation & 3D Reconstruction Results

- IMU and LiDAR are used to calculate the pose (rotation & translation) of the camera
- The 4 perspective images' poses are transformed from ERP's pose with rigid body rotation
- Atlas: Convolutional Neural-Network-based TSDF Estimation Model
  - Input: perspective images converted from ERP and their corresponding poses



Visualization of perspective views and the corresponding pose

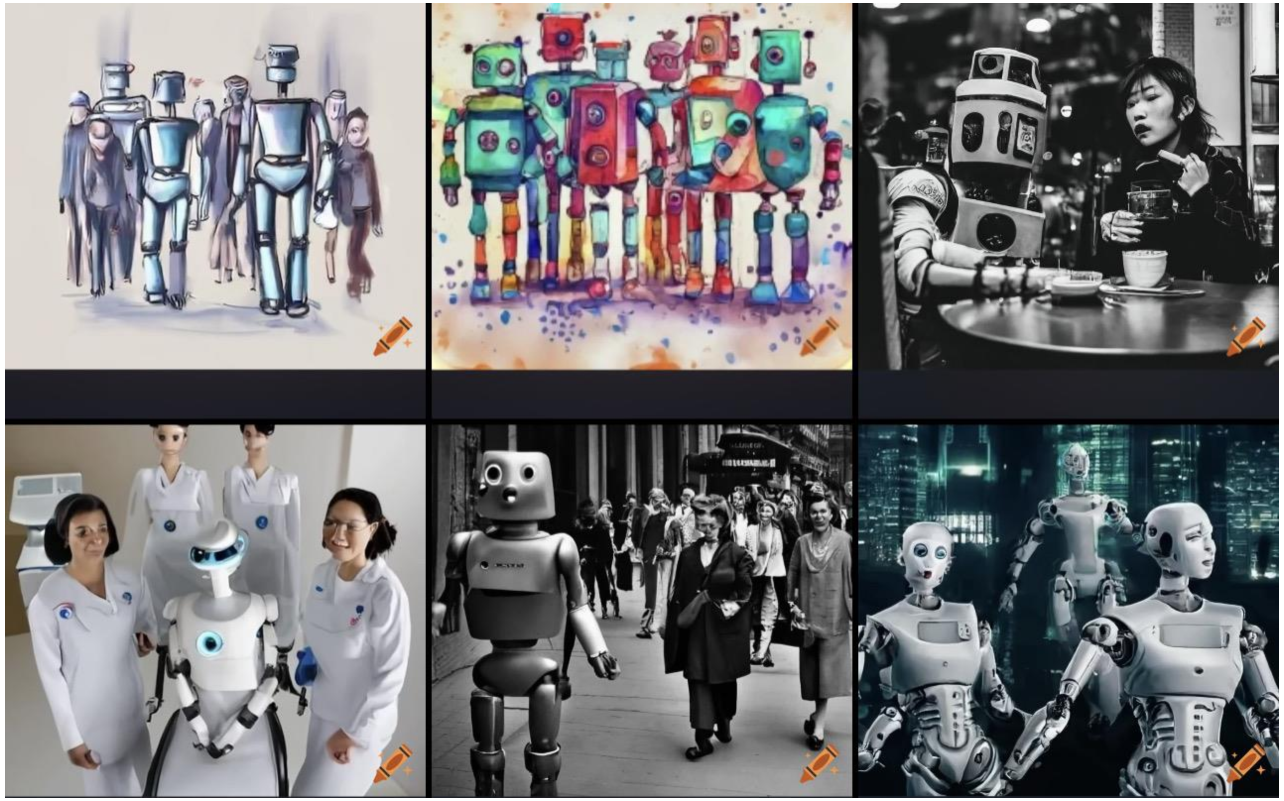


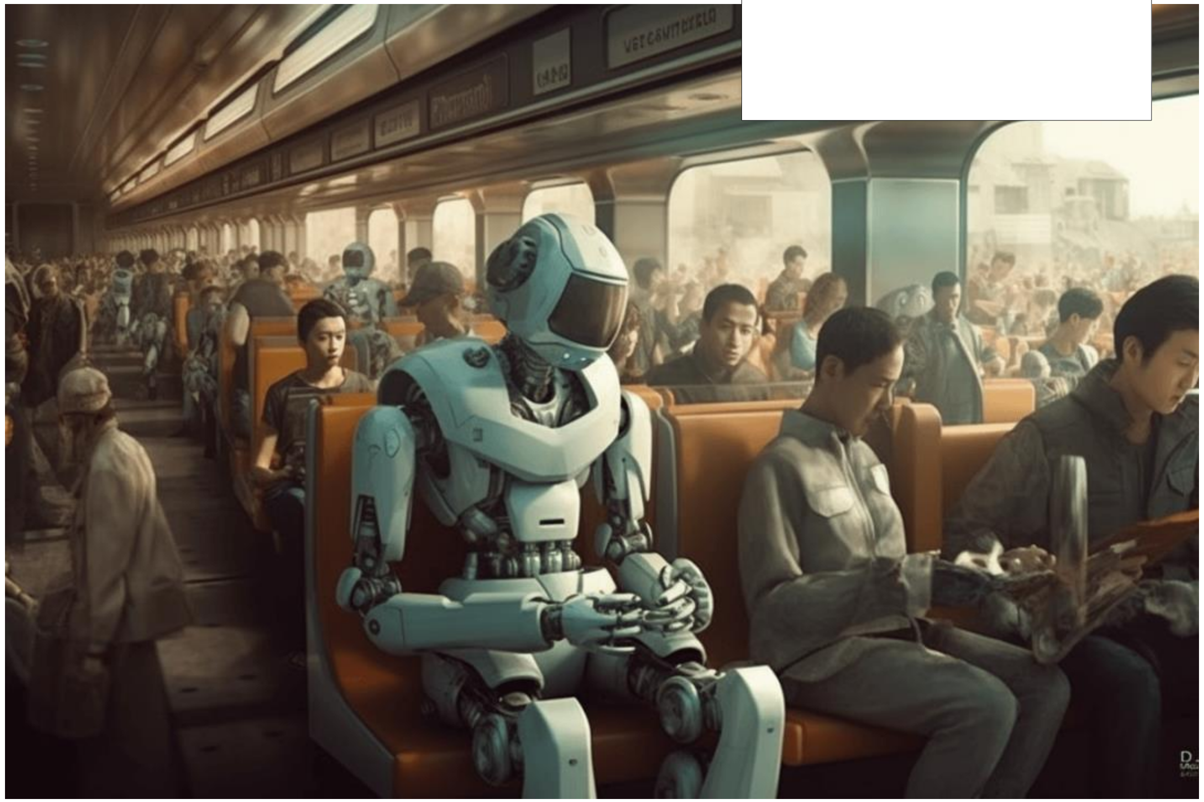
The reconstructed scenes in HKUST

Market Driven=> Application-Specific => Power Awareness

- The demand for AI algorithm computing power has surged, and the energy consumption of hardware systems may lead to "Demand exceeds Supply"
  - The amount of algorithm calculations doubles every 3-4 months, and the hardware computing power doubles every 18-24 months
  - The domestic advanced technology is stuck, and it is urgent to break through the limitations of advanced technology blockade on intelligent computing energy efficiency
- Smart construction robots with embodied AI and multi-robot human collaboration have been investigated
  - indoor positioning, multi-agent planning systems, 3D perception, and 360° imaging

谢谢!! 🙏🙏





香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

HQ OWL

Representative Papers for the Senses

Senses	Papers
The 1st Sense – Vision	Cao, Qiankai, and Jie Gu. "A Sparse Convolution Neural Network Accelerator for 3D/4D Point-Cloud Image Recognition on Low Power Mobile Device with Hopping-Index Rule Book for Efficient Coordinate Management." 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2022. Ma, X., Zhao, R., & Zhou, J. "Convolutional Neural Network (CNN) Accelerator Chip Design," *In 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID) (pp. 211-215). IEEE,* 2019
The 2nd Sense – Speech (Speak and Hear)	Ambrogio, Stefano, et al. "An Analog-AI Chip for Energy-efficient Speech Recognition and Transcription." *Nature 620.7975 (2023): 768-775.*
The 3rd Sense - Smell	Bahremand, Alireza, et al. "The Smell Engine: A System For Artificial Odor Synthesis In Virtual Environments." 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE, 2022. Benazzouz, Yazid, and Rachid Boudour. "Integration Of Smell Into The Digital World." 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). IEEE, 2019.
The 4th Sense – Taste	Kamata, Yusaku, and Tomokazu Ishikawa. "A Study On The Influence Of Animation On The Sense Of Taste." 2023 Nicograph International (NicoInt). IEEE, 2023. Zülfikar, İdil Esen, Hamdi Dibeklioğlu, and Hazim Kemal Ekenel. "A Preliminary Study On Visual Estimation Of Taste Appreciation." 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2016.
The 5th Sense - Touching	Balaji, A. N., & Peh, L. S. "AI-On-Skin: Towards Enabling Fast and Scalable On-body AI Inference for Wearable On-Skin Interfaces," *Proceedings of the ACM on Human-Computer Interaction, 7(EICS), 1-34,* 2023
The 6th Sense - Positioning	Y. Kim *et al.* "A 0.55 V 1.1 mW Artificial Intelligence Processor With On-Chip PVT Compensation for Autonomous Mobile Robots." *IEEE Transactions on Circuits and Systems I: Regular Papers 65.2 (2017): 567-580.*

Representative Papers for the Senses from ISSCC & VLSI

Senses	ISSCC Papers	VLSI Papers
The 1st Sense – Vision	[1] Garrett, David, et al. "A 1mW Always-on Computer Vision Deep Learning Neural Decision Processor." 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023.speech [2] Murakami, Hirotaka, et al. "A 4.9 Mpixel Programmable-Resolution Multi-Purpose CMOS Image Sensor for Computer Vision." 2022 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 65. IEEE, 2022. [3] Singh, Rituraj, et al. "34.2 a 21pJ/frame/pixel imager and 34pJ/frame/pixel Image Processor For A Low-vision Augmented-reality Smart Contact Lens." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021. [4] Venkatasubramanian, Rama, et al. "2.6 A 16nm 3.5 B+ Transistor> 14TOPS 2-to-10W Multicore SoC Platform for Automotive and Embedded Applications with Integrated Safety MCU, 512b Vector VLIW DSP, Embedded Vision and Imaging Acceleration." 2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020. [5] Xu, Chen, et al. "5.1 A Stacked Global-shutter CMOS Imager with SC-type hybrid-GS Pixel and Self-knee Point Calibration Single Frame HDR and On-chip Binarization Algorithm For Smart Vision Applications." 2019 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2019.	[6] Rüedi, P-F., R. Quaglia, and H-R. Graf. "A 90 mw at 1 fps and 1.33 mw at 30 fps 120 db Intra-scene Dynamic Range 640× 480 Stacked Image Sensor For Autonomous Vision Systems." 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2023. [7] Zhang, Qirui, et al. "A 22nm 3.5 TOPS/W Flexible Micro-Robotic Vision SoC with 2MB eMRAM for Fully-on-Chip Intelligence." 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2022. [8] Sumi, Hirofumi, et al. "Advanced Multi-nir Spectral Image Sensor With Optimized Vision Sensing System And Its Impact On Innovative Applications." 2021 Symposium on VLSI Circuits. IEEE, 2021. [9] Li, Chenghan, et al. "A 132 by 104 10µm-Pixel 250µW 1kefps Dynamic Vision Sensor With Pixel-parallel Noise And Spatial Redundancy Suppression." 2019 Symposium on VLSI Circuits. IEEE, 2019.

Representative Papers for the Senses from ISSCC & VLSI

Senses	ISSCC Papers	VLSI Papers
The 2nd Sense – Speech (Speak and Hear)	[10] Park, Sungjin, et al. "22.8 A0. 81 mm 2 740µW Real-Time Speech Enhancement Processor Using Multiplier-Less PE Arrays for Hearing Aids in 28nm CMOS." 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023. [11] Kang, Taewook, et al. "A Multimode 157µW 4-Channel 80dBA-SNDR Speech-Recognition Frontend With Self-DOA Correction Adaptive Beamformer." 2022 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 65. IEEE, 2022. [12] Tambe, Thierry, et al. "9.8 A 25mm 2 SoC for IoT Devices with 18ms Noise-robust Speech-to-text Latency via Bayesian Speech Denoising and Attention-based Sequence-to-sequence DNN Speech Recognition In 16nm Finfet." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021.	[13] Dosho, Shiro, et al. "A Compact 0.9 uW Direct-Conversion Frequency Analyzer for Speech Recognition with Wide-Range Q-Controlable Bandpass Rectifier." 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2023. [14] Nako, E., et al. "Experimental Demonstration of Novel Scheme of HZO/Si FeFET Reservoir Computing With Parallel Data Processing For Speech Recognition." 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2022. [15] Lee, Seungjong, et al. "An 8-element Frequency-selective Acoustic Beamformer And Bitstream Feature Extractor With 60 Mel-frequency Energy Features Enabling 95% Speech Recognition Accuracy." 2020 IEEE Symposium on VLSI Circuits. IEEE, 2020. [16] Guo, Ruiqi, et al. "A 5.1 pJ/neuron 127.3 us/Inference RNN-based Speech Recognition Processor using 16 Computing-in-Memory SRAM Macros in 65nm CMOS." 2019 Symposium on VLSI Circuits. IEEE, 2019.

Reference for Senses

- The 1st Sense – Vision

- Dai, A. *et al., “*ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," 2017 *CVPR.*
- Ma, X., Zhao, R., & Zhou, J. "Convolutional Neural Network (CNN) Accelerator Chip Design," *In 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID) (pp. 211-215). IEEE,* 2019
- Murez, Zak, *et al.* "Atlas: End-to-End 3d Scene Reconstruction from Posed Images." *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. Springer International Publishing, 2020.
- Sun, Jiaming, *et al*. "NeuralRecon: Real-time Coherent 3D Reconstruction from Monocular Video." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
- Stier, Noah, *et al.* "Vortx: Volumetric 3d Reconstruction with Transformers for Voxelwise View Selection and Fusion." *2021 International Conference on 3D Vision (3DV). IEEE*, 2021.
- Newcombe, Richard A., *et al.* "Kinectfusion: Real-time Dense Surface Mapping and Tracking." *2011 10th IEEE international symposium on mixed and augmented reality. IEEE*, 2011.
- Dai, Angela, *et al.* "Bundlefusion: Real-time Globally Consistent 3D Reconstruction using On-the-Fly Surface Reintegration." *ACM Transactions on Graphics (ToG)* 36.4 (2017): 1.
- Weder, Silvan, *et al.* "Neuralfusion: Online Depth Fusion in Latent Space." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021.
- Bozic, Aljaz, *et al.* "Transformerfusion: Monocular RGB Scene Reconstruction using Transformers." *Advances in Neural Information Processing Systems 34* (2021): 1403-1414.

Reference for Senses

- The 2nd Sense – Speech (Speak and Hear)
  - “Tech Showcase: Customizing Speech Recognition for Higher Accuracy Transcriptions,” www.youtube.com. https://www.youtube.com/watch?v=80CPeMLJMEw
  - Ambrogio, Stefano, et al. "An Analog-AI Chip for Energy-efficient Speech Recognition and Transcription." *Nature 620.7975 (2023): 768-775.*
- The 3rd Sense - Smell
  - “The Science Behind Smell (How Your Nose Works!),” www.youtube.com. https://www.youtube.com/watch?v=zaHR2MAxywg&t=99s%E2%80%8B
  - Bahremand, Alireza, *et al.* "The Smell Engine: A System For Artificial Odor Synthesis In Virtual Environments." *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE*, 2022.
  - Benazzouz, Yazid, and Rachid Boudour. "Integration Of Smell Into The Digital World." *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). IEEE,* 2019.
- The 4th Sense – Taste
  - “Japan Researchers Develop Electric Chopsticks to Enhance Salty Taste,” www.youtube.com. https://www.youtube.com/watch?v=P-V3EqQEuyQ
  - Kamata, Yusaku, and Tomokazu Ishikawa. "A Study On The Influence Of Animation On The Sense Of Taste." *2023 Nicograph International (NicoInt). IEEE,* 2023.
  - Zülfikar, İdil Esen, Hamdi Dibeklioğlu, and Hazim Kemal Ekenel. "A Preliminary Study On Visual Estimation Of Taste Appreciation." *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW).* IEEE, 2016.

Reference for Senses

- The 5th Sense - Touching
  - “How to Make Electronic Skin with Stanford’s Zhenan Bao—Speaking of Chemistry Roa_哔哩哔哩_bilibili,” www.bilibili.com. https://www.bilibili.com/video/BV1M4411j7wm/?spm_id_from=333.337
  - Zhao, J., & Adelson, E. H. "GelSight Svelte Hand: A Three-finger, Two-DoF, Tactile-rich, Low-cost Robot Hand for Dexterous Manipulation," *arXiv preprint arXiv:2309.10886,* 2023
  - Zhao, J., & Adelson, E. H. "GelSight Svelte: A Human Finger-shaped Single-camera Tactile Robot Finger with Large Sensing Coverage and Proprioceptive Sensing," *arXiv preprint arXiv:2309.10885,* 2023
  - Balaji, A. N., & Peh, L. S. "AI-On-Skin: Towards Enabling Fast and Scalable On-body AI Inference for Wearable On-Skin Interfaces," *Proceedings of the ACM on Human-Computer Interaction, 7(EICS), 1-34,* 2023
- The 6th Sense - Positioning
  - Arun *et al.,* "P2SLAM: Bearing Based WiFi SLAM for Indoor Robots," *IEEE Robotics and Automation Letters,* 7(2), 3326-3333, 2022.
  - Y. Kim *et al.* "A 0.55 V 1.1 mW Artificial Intelligence Processor With On-Chip PVT Compensation for Autonomous Mobile Robots." *IEEE Transactions on Circuits and Systems I: Regular Papers 65.2 (2017): 567-580.*

Our Publications

- W. Guan et al., “Robust Robotic Localization Using Visible Light Positioning and Inertial Fusion,” IEEE Sensor Journal, 2022.
- Y. Wang et al., “High Precision Indoor Robot Localization Using VLC Enabled Smart Lighting,” Optical Fiber Communication Conference, M1B. 2021.
- Z. Hong et al., “Cross-Dimensional Refined Learning for Real-Time 3D Visual Perception from Monocular Video,” IEEE/CVF ICCV workshops, 2023.
- H. C. Cheng et al., “Leveraging 360° Camera in 3D Reconstruction: A Vision-based Approach,” 2023 2nd International Conference on Video and Signal Processing, in press.
- H. C. Cheng et al., “Optimizing Field-of-View for Multi-Agent Path Finding via Reinforcement Learning: A Performance and Communication Overhead Study,” 62nd IEEE Conference on Decision and Control (CDC), in press.

References on Reinforcement Learning

- B. Sarkar et al., "PantheonRL: A MARL Library for Dynamic Training Interactions, " In Proceedings of the 36th AAAI Conference on Artificial Intelligence (Demo Track), 2022.
- G. Sartoretti et al., "PRIMAL: Pathfinding via Reinforcement and Imitation Multi-Agent Learning," IEEE Robotics and Automation Letters (RA-L), vol. 4, no. 3, pp. 2378–2385, 2019.
- Q. Li et al., "Graph neural networks for decentralized multi-robot path planning," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 11 785–11 792, 2020.
- V. Mnih et al., "Playing Atari with Deep Reinforcement Learning," NeurIPS Deep Learning Workshop, 2013.
- V. Mnih et al., "Asynchronous Methods for Deep Reinforcement Learning, " International Conference on Machine Learning (ICML), pp. 1928-1937, 2016.
- A. Das et al., “Tarmac: Targeted multi-agent communication,” International Conference on Machine Learning (ICML), pp. 1538–1546, 2019.

Reference from ISSCC

- **Vision**
  - Garrett, David, et al. "A 1mW Always-on Computer Vision Deep Learning Neural Decision Processor." 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023.speech
  - Murakami, Hirotaka, et al. "A 4.9 Mpixel Programmable-Resolution Multi-Purpose CMOS Image Sensor for Computer Vision." 2022 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 65. IEEE, 2022.
  - Singh, Rituraj, et al. "34.2 a 21pJ/frame/pixel imager and 34pJ/frame/pixel Image Processor For A Low-vision Augmented-reality Smart Contact Lens." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021.
  - Venkatasubramanian, Rama, et al. "2.6 A 16nm 3.5 B+ Transistor> 14TOPS 2-to-10W Multicore SoC Platform for Automotive and Embedded Applications with Integrated Safety MCU, 512b Vector VLIW DSP, Embedded Vision and Imaging Acceleration." 2020 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2020.
  - Xu, Chen, et al. "5.1 A Stacked Global-shutter CMOS Imager with SC-type hybrid-GS Pixel and Self-knee Point Calibration Single Frame HDR and On-chip Binarization Algorithm For Smart Vision Applications." 2019 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2019.
- **Speech**
  - Park, Sungjin, et al. "22.8 A0. 81 mm 2 740µW Real-Time Speech Enhancement Processor Using Multiplier-Less PE Arrays for Hearing Aids in 28nm CMOS." 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023.
  - Kang, Taewook, et al. "A Multimode 157µW 4-Channel 80dBA-SNDR Speech-Recognition Frontend With Self-DOA Correction Adaptive Beamformer." 2022 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 65. IEEE, 2022.
  - Tambe, Thierry, et al. "9.8 A 25mm 2 SoC for IoT Devices with 18ms Noise-robust Speech-to-text Latency via Bayesian Speech Denoising and Attention-based Sequence-to-sequence DNN Speech Recognition In 16nm Finfet." 2021 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 64. IEEE, 2021.

Reference from VLSI

- **Vision**
  - Rüedi, P-F., R. Quaglia, and H-R. Graf. "A 90 mw at 1 fps and 1.33 mw at 30 fps 120 db Intra-scene Dynamic Range 640× 480 Stacked Image Sensor For Autonomous Vision Systems." 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2023.
  - Zhang, Qirui, et al. "A 22nm 3.5 TOPS/W Flexible Micro-Robotic Vision SoC with 2MB eMRAM for Fully-on-Chip Intelligence." 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2022.
  - Sumi, Hirofumi, et al. "Advanced Multi-nir Spectral Image Sensor With Optimized Vision Sensing System And Its Impact On Innovative Applications." 2021 Symposium on VLSI Circuits. IEEE, 2021.
  - Li, Chenghan, et al. "A 132 by 104 10μm-Pixel 250μW 1kefps Dynamic Vision Sensor With Pixel-parallel Noise And Spatial Redundancy Suppression." 2019 Symposium on VLSI Circuits. IEEE, 2019.
- **Speech**
  - Dosho, Shiro, et al. "A Compact 0.9 uW Direct-Conversion Frequency Analyzer for Speech Recognition with Wide-Range Q-Controlable Bandpass Rectifier." 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2023.
  - Nako, E., et al. "Experimental Demonstration of Novel Scheme of HZO/Si FeFET Reservoir Computing With Parallel Data Processing For Speech Recognition." 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2022.
  - Lee, Seungjong, et al. "An 8-element Frequency-selective Acoustic Beamformer And Bitstream Feature Extractor With 60 Mel-frequency Energy Features Enabling 95% Speech Recognition Accuracy." 2020 IEEE Symposium on VLSI Circuits. IEEE, 2020.
  - Guo, Ruiqi, et al. "A 5.1 pJ/neuron 127.3 us/Inference RNN-based Speech Recognition Processor using 16 Computing-in-Memory SRAM Macros in 65nm CMOS." 2019 Symposium on VLSI Circuits. IEEE, 2019.