ASK 2025 논문집

Annual Symposium of KIPS 2025

신진학자 워크숍

Enabling Ubiquitous AI via Adaptive and Efficient On-Device ML

공태식 조교수(UNIST)

Enabling Ubiquitous AI via Adaptive and Efficient On-Device ML

Taesik Gong

U*AI Ubiquitous Artificial Intelligence Laboratory UNIST

Who am I?

Taesik Gong

Assistant Professor
@ CSE & AIGS, UNIST
2024.08 ~

https://taesikgong.com/

Experience

- Visiting Scholar, **University of Cambridge,** Cambridge, UK, 2024
- Research Scientist, **Nokia Bell Labs,** Cambridge, UK, 2023-2024
- Research Intern, **Google Research**, NYC, USA, 2022
- Research Intern, **Microsoft Research**, Beijing, China, 2019
- Research Intern, **Nokia Bell Labs**, Cambridge, UK, 2018

Education

- **KAIST**: Ph.D., School of Computing, 2023
- **KAIST**: M.S., School of Computing, 2017
- **Yonsei University**: B.S., Computer Science, 2016

Research areas

- **Human-Centered AI**
- **Adaptive & Personalized AI**
- **On-Device AI Systems**

Selected Publications

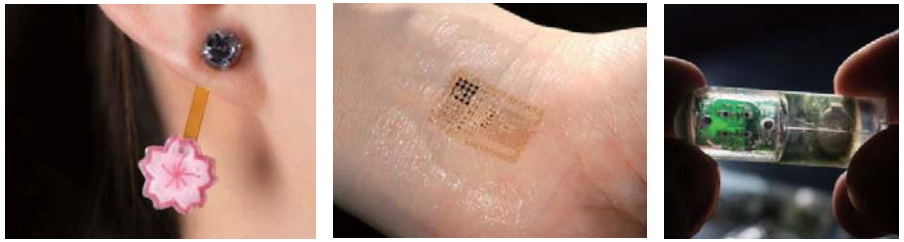
- **AI/ML**: ICLR '25, NeurIPS '24, EMNLP '24, CVPR '24, NeurIPS '23, NeurIPS '22
- **Ubiquitous Computing**: SenSys '25, UbiComp '24, UbiComp '23, CHI '22, SenSys '19, UbiComp '19

Our Lab's Mission: Making AI Ubiquitous

Foundational AI

deepseek

Ubiquitous AI



Security
Low latency
Personalization

On-Device AI: The Backbone of Ubiquitous AI

- Running AI models on edge devices without cloud servers

Tesla Autopilot 4.0
(2023.03)

Google Gemini Nano
(2023.12)

Samsung Live Translate
(2024.01)

Apple Intelligence
(2024.07)

Our Research Directions

Direction 1

Human-Centered AI Applications

- Knocker (UbiComp '19)
- MilliCat (MobileHCI '20)
- MyDJ (CHI '22)
- MIRROR (UbiComp '24)

Direction 2

Adaptive & Personalized AI

- NOTE (NeurIPS '22)
- SoTTA (NeurIPS '23)
- AETTA (CVPR '24)
- Lens (ICLR '25)

Direction 3

Resource-Efficient On-Device AI

- DAPPER (UbiComp '23)
- DEX (NeurIPS '24)
- Self-Replay (SenSys '25)
- Synergy (TMC '25)

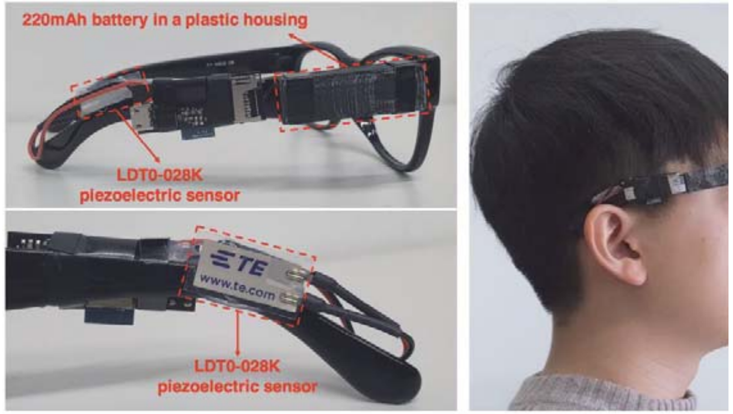
Direction 1: Human-Centered AI Applications

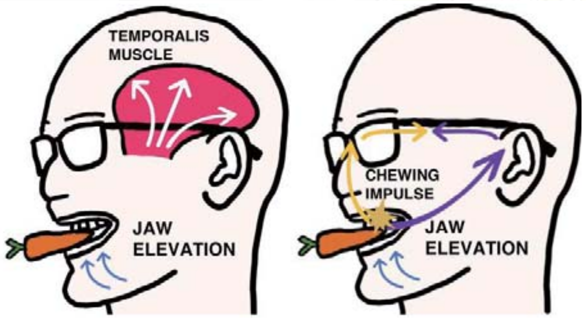
"How can we enrich users' daily lives with on-device AI?"

Object Interaction (UbiComp '19)

*Featured by KBS, MBC, YTN

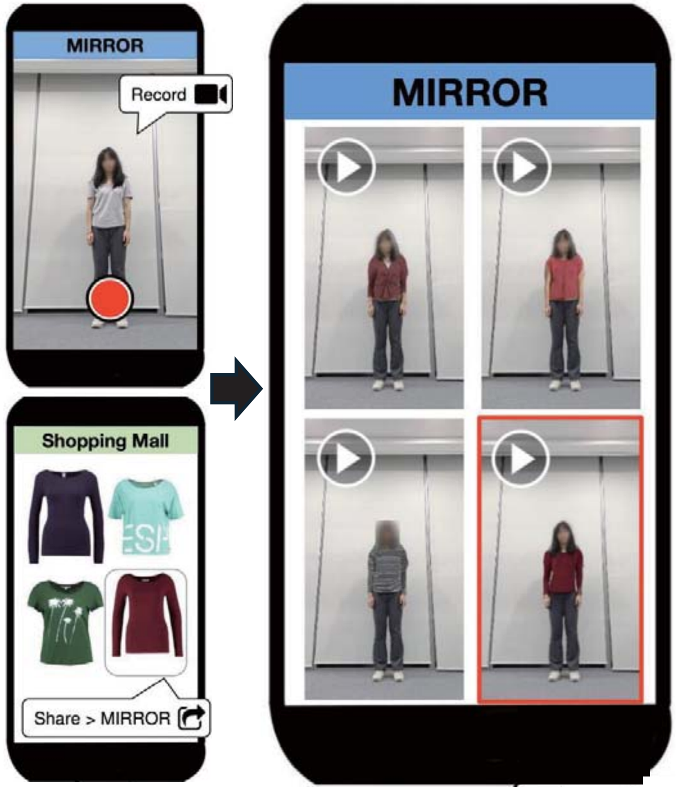
Eating Tracking (CHI '22)





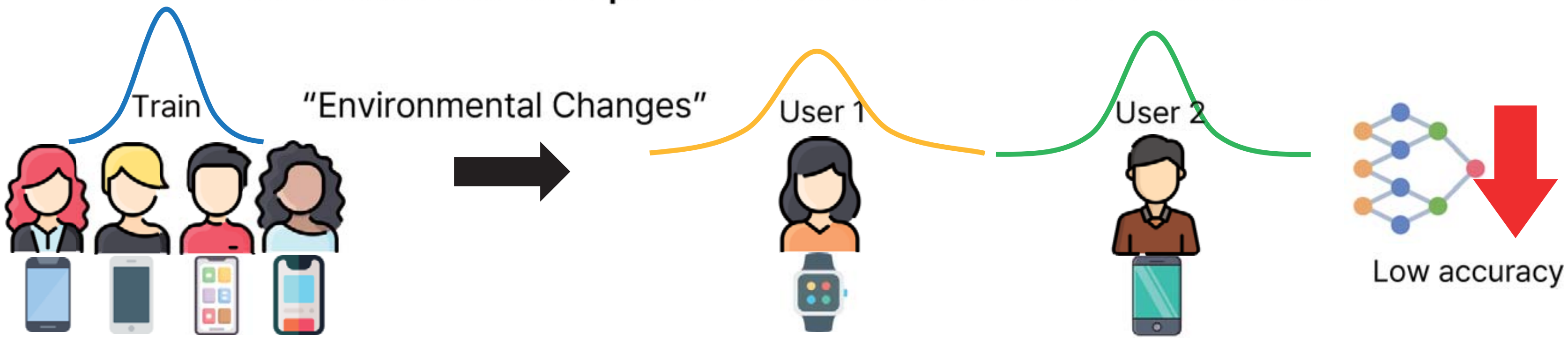
*Best Paper Honorable Mention

Virtual Try-On (UbiComp '24)



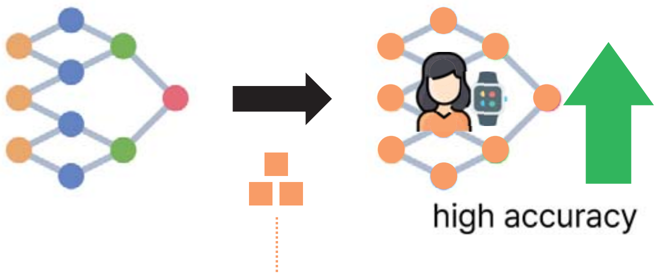
Direction 2: Adaptive & Personalized AI

"How can we adapt AI to different environments?"



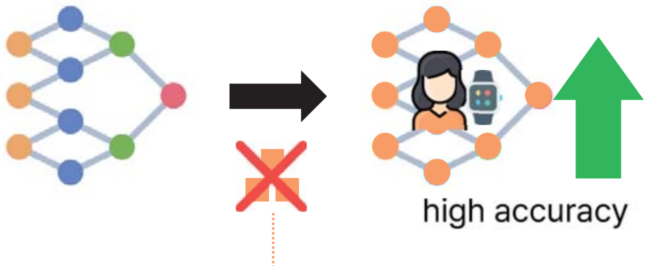
Requirements: (1) accuracy ↑ (2) computation ↓ (3) user burden ↓

Few-Shot Adaptation
(SenSys '19, TMC '22, UbiComp '23)



Adaptation with one or two samples

Test-Time Adaptation
(NeurIPS '22, NeurIPS '23, CVPR '24)

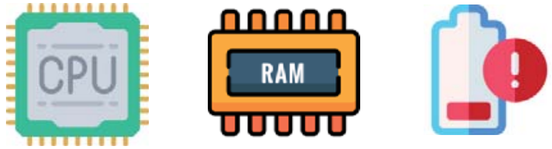


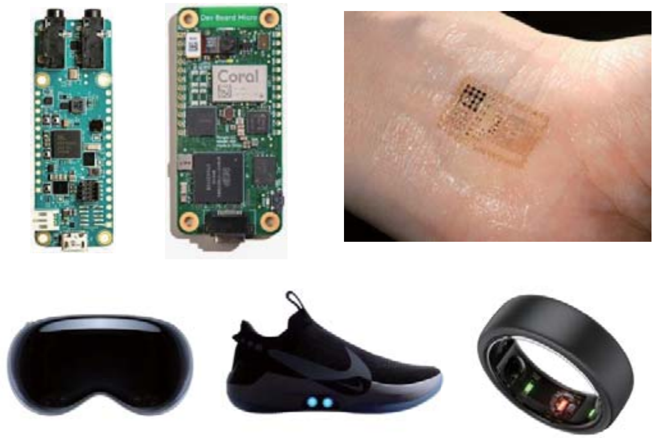
Adaptation without data collection

Direction 3: Resource-Efficient On-Device AI

"How can we support AI in a resource-efficient manner?"

Extremely Limited Resources

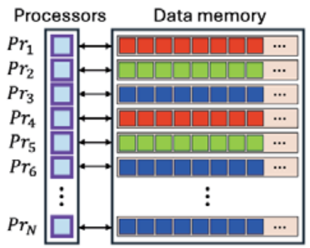




Performance Estimator (UbiComp '23)

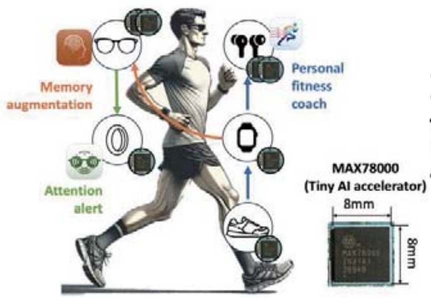
396× ↓ latency
40% ↑ accuracy

Tiny AI Accelerator (NeurIPS '24)



21× ↑ utilization
3% ↑ accuracy

Wearable Collaboration (TMC '25)



23× ↑ TPUT
74% ↓ latency
16% ↓ power

Our Research Directions

Direction 1

Human-Centered AI Applications

- Knocker (UbiComp '19)
- MilliCat (MobileHCI '20)
- MyDJ (CHI '22)
- MIRROR (UbiComp '24)

Direction 2

Adaptive & Personalized AI

- NOTE (NeurIPS '22)
- SoTTA (NeurIPS '23)
- AETTA (CVPR '24)
- Lens (ICLR '25)

Direction 3

Resource-Efficient On-Device AI

- DAPPER (UbiComp '23)
- DEX (NeurIPS '24)
- Self-Replay (SenSys '25)
- Synergy (TMC '25)

This talk' s focus

SoTTA: Robust Test-Time Adaptation on Noisy Data Streams

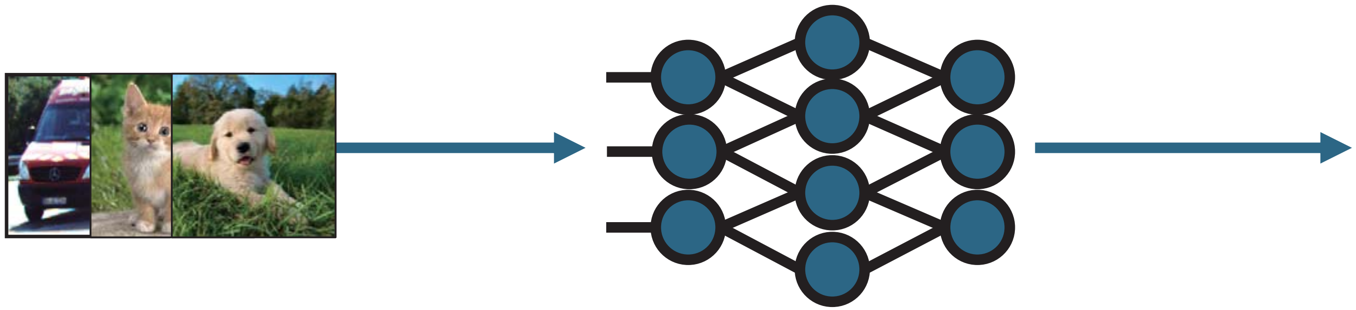
Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee

NeurIPS 2023

Illustration of Test-Time Adaptation (TTA)

Incoming Test Samples

Test & Adapt



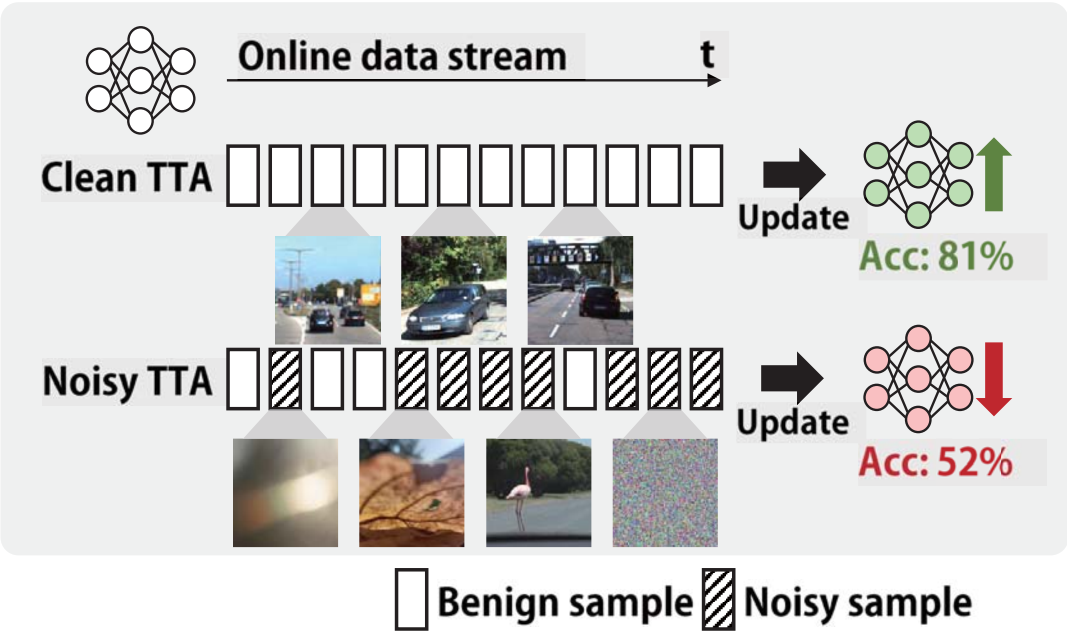
Dog
Cat

Accuracy

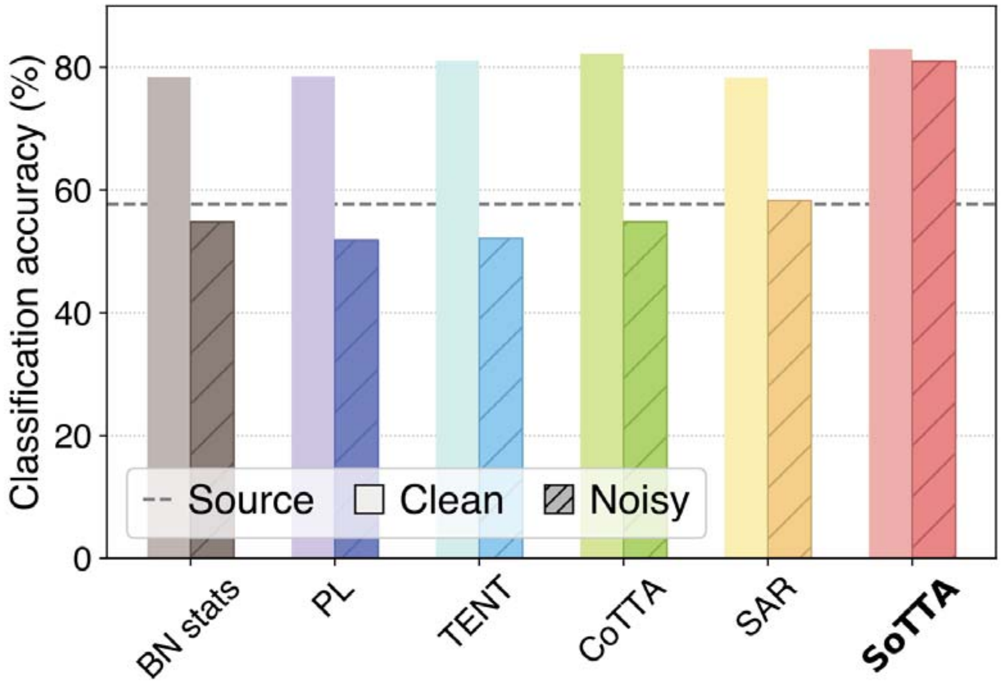
TTA gradually adapts to unseen environments as it’s being used

Test Samples Can be Unexpectedly Diverse in the Wild

Example: Autonomous driving scenario



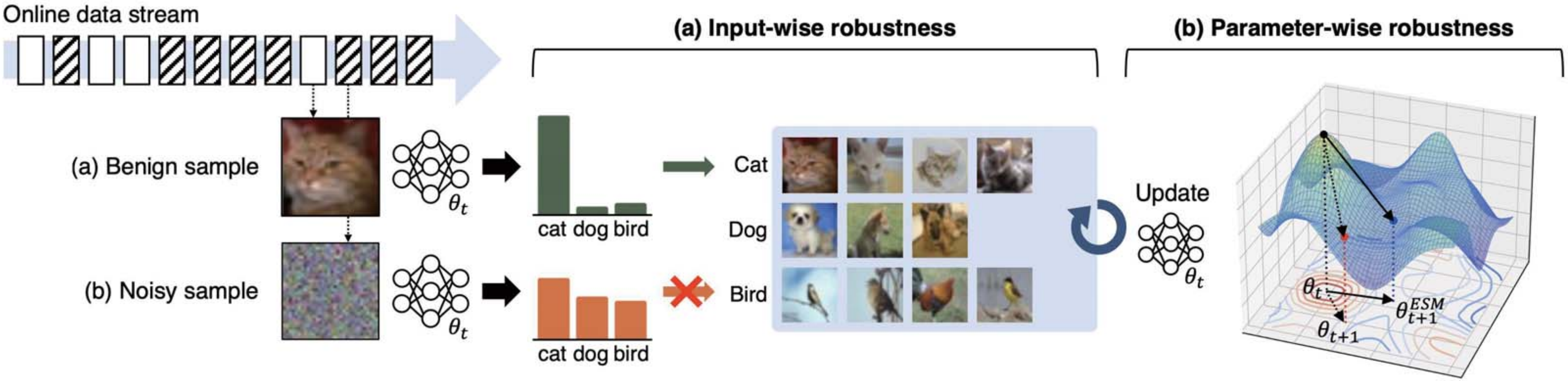
Prior methods fail with noisy test data



→Models are contaminated with noisy samples

SoTTA: Screening-out Test-Time Adaptation

Goal: reduce the impact of noisy samples in TTA

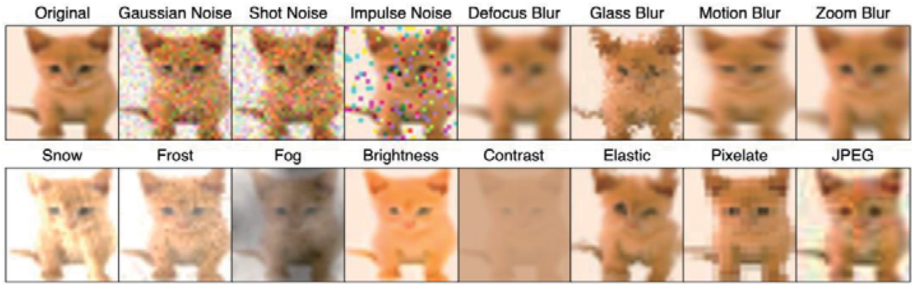


High-confidence Uniform-class Sampling (HUS)
: avoids selecting noisy samples when updating the model

Entropy-Sharpness Minimization (ESM)
: makes parameters resilient to weight perturbation caused by noisy samples

Evaluation with five scenarios (CIFAR10-C)

TTA benchmark: CIFAR10 + 15 Corruptions



Five noisy sample scenarios

(a) Benign. (b) Near. (c) Far. (d) Attack. (e) Noise.

Accuracy ↑ (%)

Method	Benign	Near	Far	Attack	Noise	Avg.
Source	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0
BN Stats [27]	78.4 ± 0.3	76.6 ± 0.4	75.2 ± 0.3	55.9 ± 1.4	54.8 ± 0.8	68.2 ± 0.5
PL [17]	78.5 ± 0.3	73.4 ± 0.2	69.8 ± 1.5	66.3 ± 1.3	51.8 ± 0.9	68.0 ± 0.6
TENT [38]	81.0 ± 0.4	74.3 ± 0.9	71.2 ± 1.0	68.9 ± 0.9	52.1 ± 0.4	69.5 ± 0.4
LAME [1]	55.9 ± 0.5	56.4 ± 0.6	55.5 ± 0.4	55.9 ± 0.5	54.9 ± 0.6	55.7 ± 0.5
CoTTA [39]	82.2 ± 0.2	78.4 ± 0.4	74.5 ± 1.2	69.5 ± 1.5	54.8 ± 1.3	71.9 ± 0.4
EATA [28]	**82.4 ± 0.2**	63.9 ± 0.4	56.3 ± 0.5	70.9 ± 0.6	36.0 ± 0.8	61.9 ± 0.2
SAR [29]	78.3 ± 0.7	72.4 ± 8.8	73.3 ± 3.9	56.2 ± 1.8	58.3 ± 0.3	67.7 ± 2.4
RoTTA [44]	75.5 ± 0.7	77.7 ± 0.6	77.1 ± 1.1	78.4 ± 0.7	73.6 ± 0.5	76.5 ± 0.7
SoTTA	82.2 ± 0.3	**81.4 ± 0.5**	**81.6 ± 0.6**	**84.5 ± 0.2**	**80.0 ± 1.4**	**81.9 ± 0.5**

- Most existing TTA methods show performance degradation under noisy test streams
- SoTTA is robust to noisy streams and outperforms the best baseline by 5.4%p

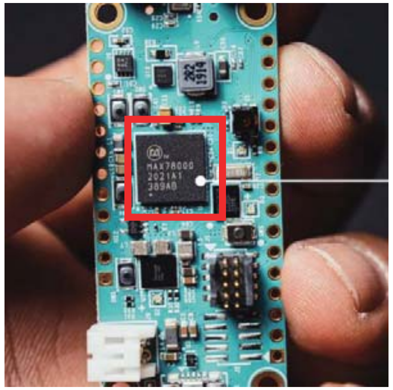
DEX: Data Channel Extension for Efficient CNN Inference on Tiny AI Accelerators

Taesik Gong, Fahim Kawsar, Chulhong Min

NeurIPS 2024

Tiny AI Accelerators: New On-Device AI Platforms

Tiny AI Accelerator
(MAX78000, 8mm × 8mm)



Omnibuds by Bell Labs
https://omnibuds.tech/

Tiny AI Accelerators   Microcontroller units (MCUs)

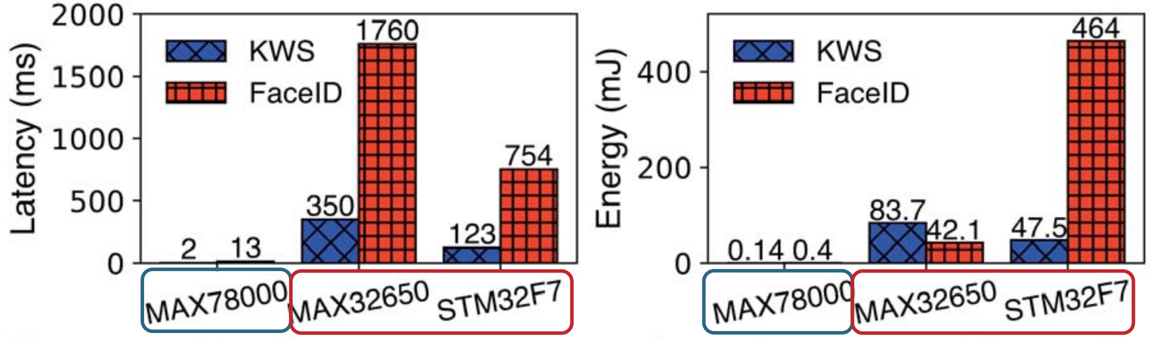
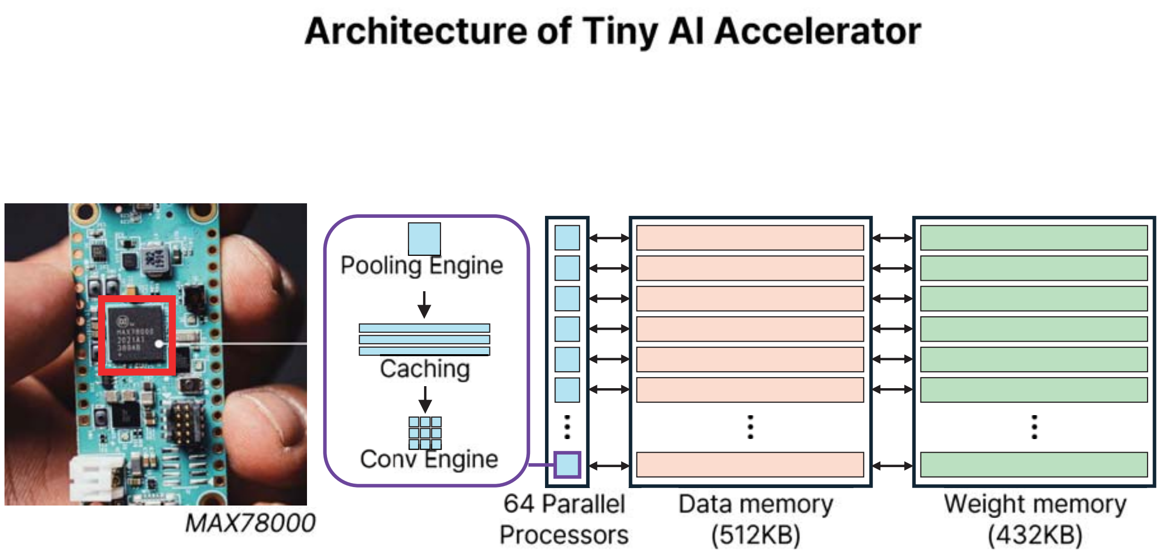


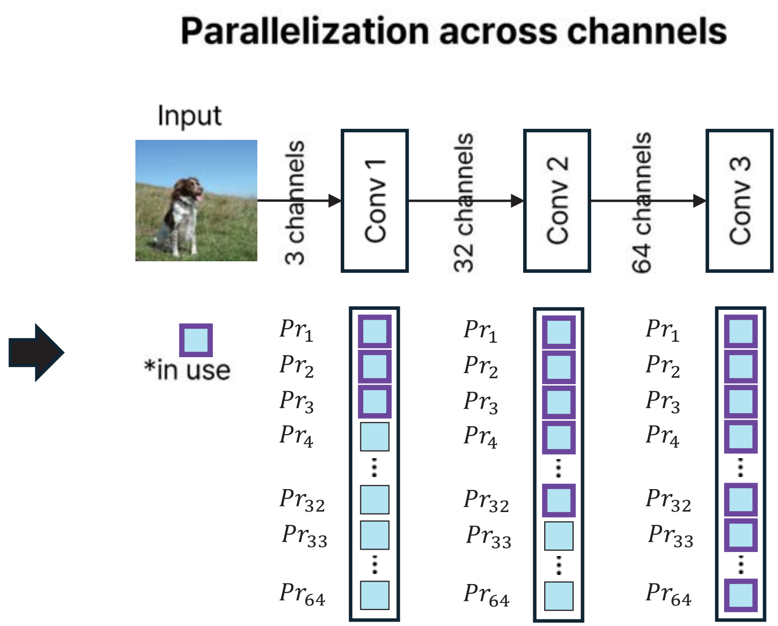
Figure 3: Performance comparison between AI accelerator (MAX78000) and MCUs (MAX32650 and STM32F7).

- 62~175× faster inference
- 105~1160× less energy consumption

→ Opportunity of (1) reduced latency, (2) lower power cost, and (3) improved privacy for on-device AI

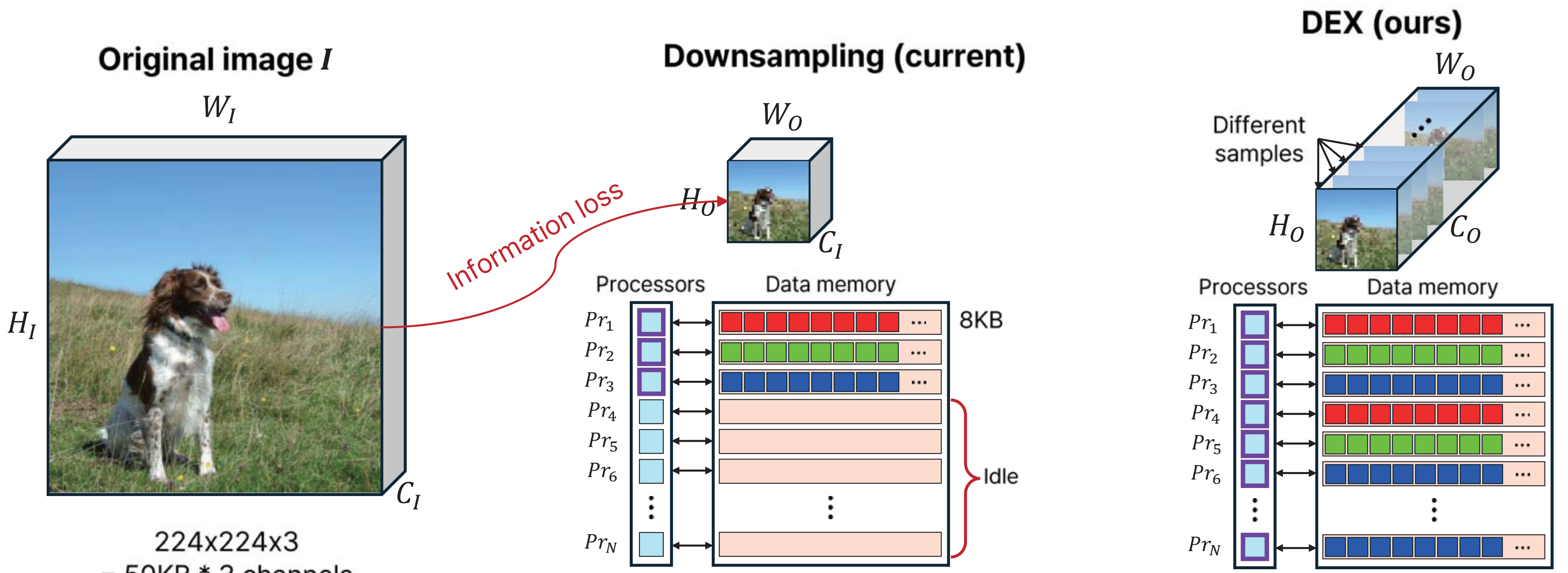
Why Are Tiny AI Accelerators Fast? Parallelization





Parallel data access and processing are the keys to fast inference

Tiny AI Accelerator Lacks Data Memory



224x224x3
= 50KB * 3 channels

*exceeds data memory limit

Underutilized
processors and memory

Improves accuracy with **additional spatial information**
with the same inference latency

DEX: Result

UNIST | U*AI Ubiquitous Artificial Intelligence Laboratory

Accuracy

Dataset	Method	SimpleNet	WideNet	EfficientNetV2	MobileNetV2	AVG (%)
ImageNette	Downsampling	57.8 ± 1.2	61.8 ± 0.2	51.3 ± 0.5	62.0 ± 0.7	58.2
	CoordConv	58.0 ± 1.1	61.7 ± 0.2	51.9 ± 0.1	61.6 ± 0.3	58.3
	CoordConv (r)	55.4 ± 1.5	61.4 ± 0.2	51.7 ± 1.0	61.2 ± 1.1	57.4
	DEX (ours)	**61.4 ± 0.6**	**65.6 ± 0.6**	**56.8 ± 0.5**	**64.4 ± 0.6**	**62.0**
Caltech101	Downsampling	54.6 ± 2.1	55.8 ± 1.2	38.6 ± 0.9	51.4 ± 1.6	50.1
	CoordConv	53.8 ± 1.6	56.5 ± 0.1	38.7 ± 0.2	49.8 ± 0.5	49.7
	CoordConv (r)	52.7 ± 0.5	56.0 ± 1.7	38.2 ± 1.0	49.7 ± 1.2	49.1
	DEX (ours)	**56.9 ± 1.3**	**61.1 ± 1.4**	**45.9 ± 1.9**	**53.3 ± 1.7**	**54.3**
Caltech256	Downsampling	19.8 ± 0.6	20.8 ± 0.5	14.7 ± 0.4	22.4 ± 1.0	19.4
	CoordConv	19.8 ± 0.5	21.3 ± 0.8	14.8 ± 0.8	22.7 ± 0.8	19.6
	CoordConv (r)	20.0 ± 1.6	20.9 ± 0.6	14.5 ± 0.3	22.7 ± 0.4	19.5
	DEX (ours)	**22.8 ± 0.5**	**22.9 ± 0.9**	**18.3 ± 0.9**	**26.3 ± 0.5**	**22.6**
Food101	Downsampling	16.0 ± 0.4	17.7 ± 0.7	12.1 ± 0.2	22.4 ± 0.6	17.1
	CoordConv	16.1 ± 0.8	17.7 ± 0.3	12.0 ± 0.1	21.7 ± 0.3	16.9
	CoordConv (r)	16.3 ± 0.4	17.3 ± 0.6	12.0 ± 0.6	20.9 ± 0.3	16.6
	DEX (ours)	**18.4 ± 0.4**	**20.9 ± 0.4**	**16.4 ± 0.1**	**23.3 ± 1.1**	**19.8**

Latency

Model	Method	InputChan	Size (KB)	InfoRatio (×)	ProcUtil (%)	Latency (μs)
SimpleNet	Downsampling	3	162.6	1.0	4.7	2592 ± 1
	CoordConv	5	162.9	1.0	7.8	2592 ± 2
	CoordConv (r)	6	163.0	1.0	9.4	2592 ± 2
	DEX (ours)	64	171.2	21.3	100.0	2591 ± 1
WideNet	Downsampling	3	306.4	1.0	4.7	3820 ± 1
	CoordConv	5	306.9	1.0	7.8	3820 ± 0
	CoordConv (r)	6	307.1	1.0	9.4	3819 ± 1
	DEX (ours)	64	319.3	21.3	100.0	3818 ± 1
EfficientNetV2	Downsampling	3	742.4	1.0	4.7	11688 ± 2
	CoordConv	5	743.0	1.0	7.8	11685 ± 3
	CoordConv (r)	6	743.2	1.0	9.4	11689 ± 1
	DEX (ours)	64	759.6	21.3	100.0	11690 ± 2
MobileNetV2	Downsampling	3	1317.8	1.0	4.7	3553 ± 4
	CoordConv	5	1318.2	1.0	7.8	3554 ± 1
	CoordConv (r)	6	1318.4	1.0	9.4	3554 ± 2
	DEX (ours)	64	1330.7	21.3	100.0	3552 ± 3

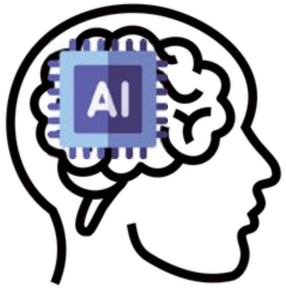
DEX improves accuracy by **3.5%p**
while keeping the **inference latency the same** on the tiny AI accelerator

Ongoing & Future Work

Personal Multi-Modal LLM Agents

Personalization + On-device LLM

Human memory augmentation



LLM + context analysis

AI-powered ingestible pill

On-device AI + medical problem

We are always open to collaborations—please feel free to reach out!

Prof. Taesik Gong

CSE & AIGS, UNIST

taesik.gong@unist.ac.kr

SoTTA: Impact of individual components

Input-wise robustness: High-Confidence Uniform-Class Sampling (HUS)
Parameter-wise robustness: Entropy-Sharpness Minimization (ESM)

Method	Benign	Near	Far	Attack	Noise	Avg.
Source	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0	57.7 ± 1.0
HC	34.9 ± 4.8	13.6 ± 0.3	17.6 ± 3.8	16.9 ± 1.6	16.8 ± 0.2	20.0 ± 2.0
UC	66.4 ± 3.0	62.1 ± 0.8	56.5 ± 2.0	70.0 ± 3.9	59.5 ± 3.0	62.9 ± 0.7
HC + UC (HUS)	69.8 ± 1.1	61.7 ± 1.3	58.4 ± 0.5	40.9 ± 5.5	58.9 ± 2.6	57.9 ± 0.8
ESM	**82.6 ± 0.2**	77.9 ± 0.4	72.8 ± 0.7	83.4 ± 0.2	60.5 ± 1.8	75.4 ± 0.5
HC + ESM	82.3 ± 0.2	80.9 ± 0.6	74.9 ± 2.4	83.5 ± 0.2	68.7 ± 7.0	78.0 ± 2.0
UC + ESM	82.2 ± 0.2	78.0 ± 0.4	75.9 ± 0.5	84.3 ± 0.1	77.7 ± 0.7	79.6 ± 0.2
HUS + ESM (SoTTA)	82.2 ± 0.3	**81.4 ± 0.5**	**81.6 ± 0.6**	**84.5 ± 0.2**	**80.0 ± 1.4**	**81.9 ± 0.5**

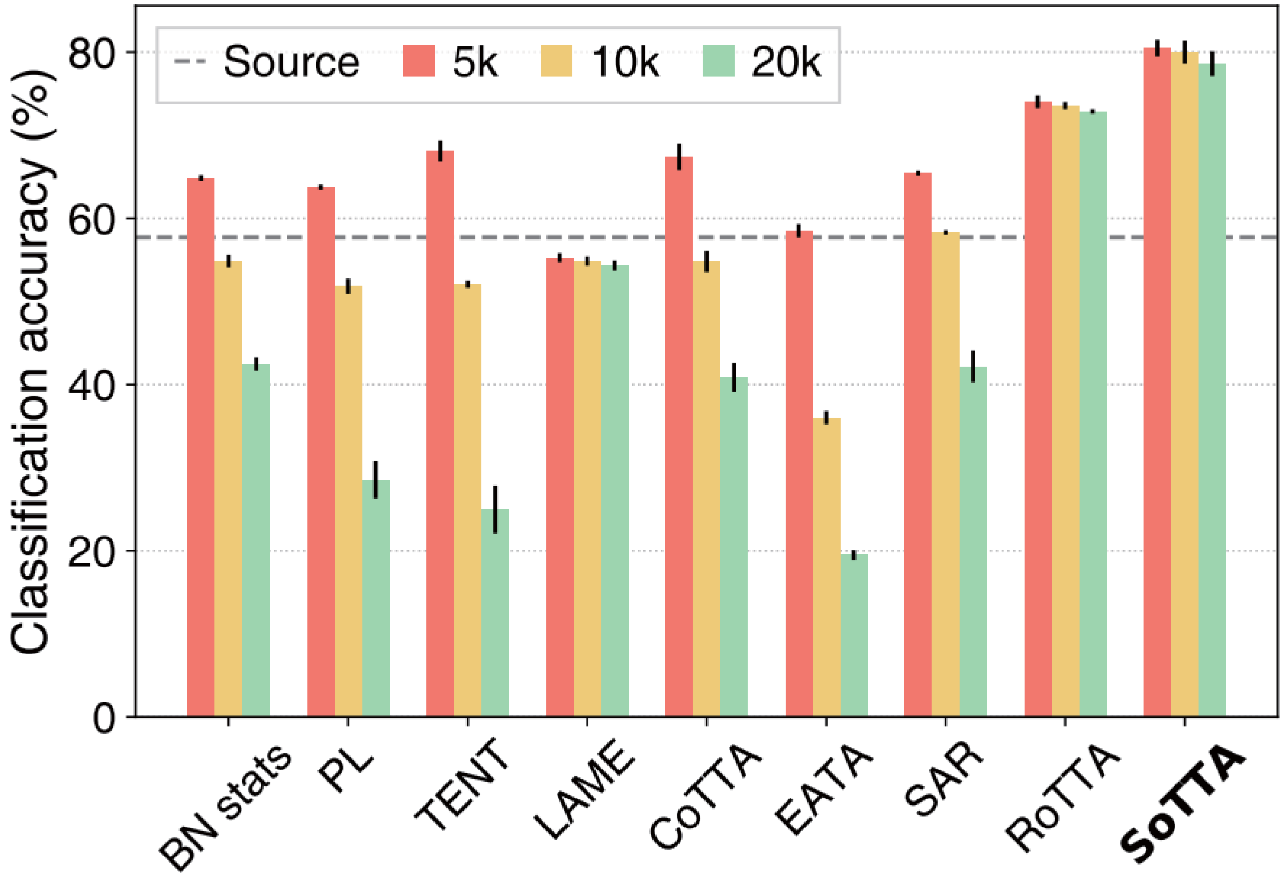
- **The accuracy is improved as we sequentially added each approach of SoTTA**
- **Ensuring both input-wise and parameter-wise robustness via HUS and ESM is a synergetic strategy**

SoTTA: CIFAR100-C & ImageNet

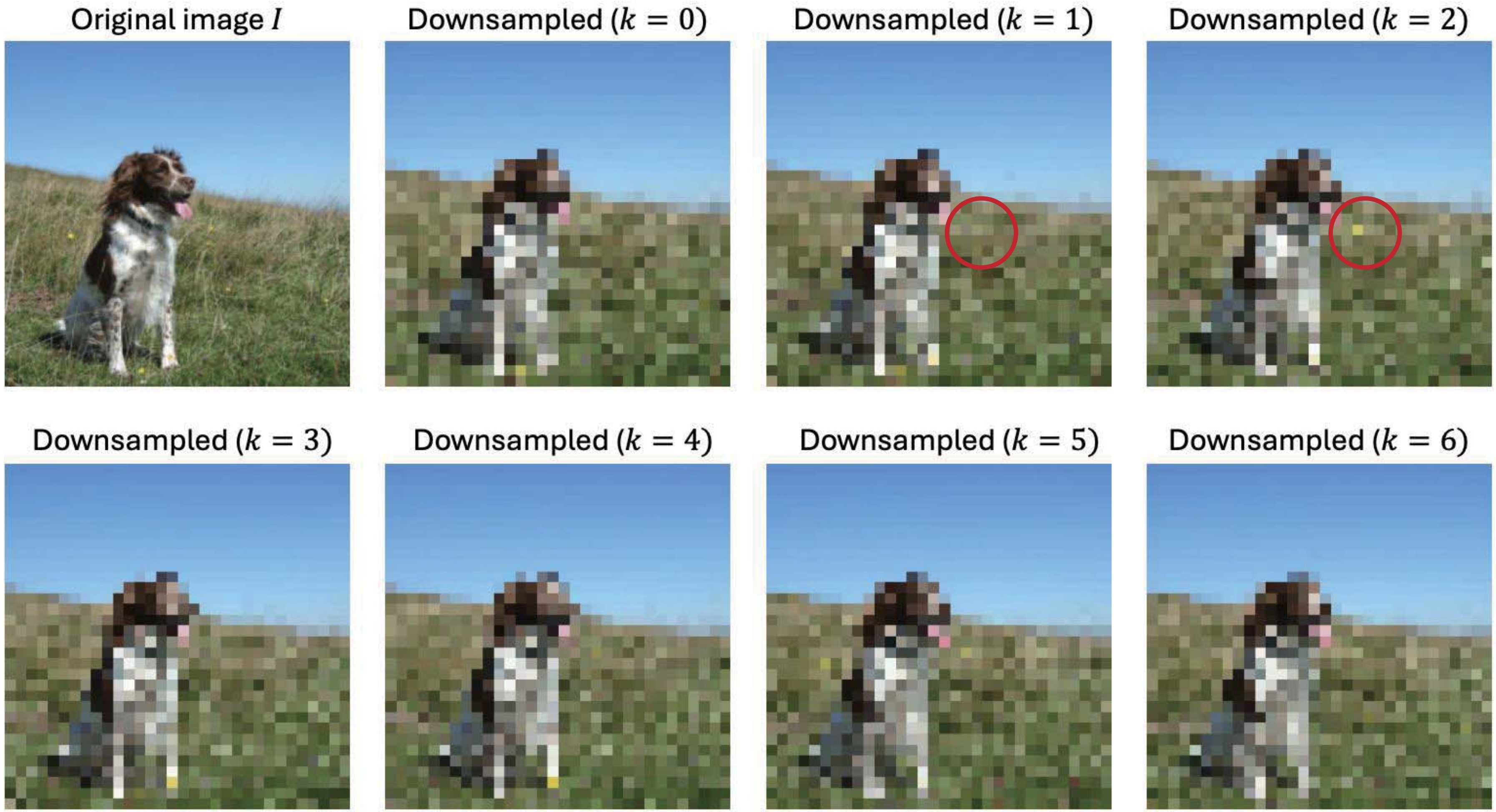
Method	Benign	Near	Far	Attack	Noise	Avg.
Source	33.2 ± 0.4	33.2 ± 0.4	33.2 ± 0.4	33.2 ± 0.4	33.2 ± 0.4	33.2 ± 0.4
BN Stats [27]	53.7 ± 0.2	50.8 ± 0.1	46.8 ± 0.1	29.2 ± 0.4	28.3 ± 0.3	41.8 ± 0.1
PL [17]	56.6 ± 0.2	48.0 ± 0.3	42.8 ± 0.7	39.0 ± 0.4	23.8 ± 0.6	42.1 ± 0.3
TENT [38]	59.5 ± 0.0	46.4 ± 1.4	40.0 ± 1.3	31.9 ± 0.7	20.0 ± 0.9	39.5 ± 0.7
LAME [1]	31.0 ± 0.5	31.5 ± 0.5	30.8 ± 0.7	31.0 ± 0.6	31.1 ± 0.7	31.1 ± 0.6
CoTTA [39]	55.8 ± 0.4	50.0 ± 0.3	42.4 ± 0.4	37.2 ± 0.2	27.3 ± 0.3	42.6 ± 0.2
EATA [28]	23.5 ± 1.9	6.1 ± 0.3	4.8 ± 0.5	3.7 ± 0.6	2.4 ± 0.2	8.1 ± 0.3
SAR [29]	57.3 ± 0.3	55.4 ± 0.1	51.2 ± 0.1	34.4 ± 0.3	38.1 ± 1.2	47.3 ± 0.3
RoTTA [44]	48.7 ± 0.6	49.4 ± 0.5	49.8 ± 0.9	51.5 ± 0.4	48.3 ± 0.5	49.6 ± 0.6
SoTTA	**60.5 ± 0.0**	**57.1 ± 0.2**	**59.0 ± 0.4**	**61.9 ± 0.0**	**58.6 ± 1.0**	**59.4 ± 0.3**

Method	Benign	Near	Far	Attack	Noise	Avg.
Source	14.6 ± 0.0	14.6 ± 0.0	14.6 ± 0.0	14.6 ± 0.0	14.6 ± 0.0	14.6 ± 0.0
BN Stats [27]	27.1 ± 0.0	18.9 ± 0.1	14.8 ± 0.0	17.4 ± 0.8	12.8 ± 0.0	18.2 ± 0.1
PL [17]	30.5 ± 0.1	6.9 ± 0.0	5.1 ± 0.2	18.1 ± 1.3	3.4 ± 0.6	12.8 ± 0.2
TENT [38]	27.1 ± 0.0	18.9 ± 0.1	14.8 ± 0.0	17.4 ± 0.8	12.8 ± 0.0	18.2 ± 0.1
LAME [1]	14.4 ± 0.0	14.4 ± 0.1	14.4 ± 0.0	14.0 ± 0.6	14.3 ± 0.0	14.3 ± 0.1
CoTTA [39]	32.2 ± 0.1	23.3 ± 0.2	17.6 ± 0.2	28.3 ± 1.3	16.0 ± 0.9	23.4 ± 0.2
EATA [28]	38.0 ± 0.1	25.6 ± 0.4	23.1 ± 0.1	26.1 ± 0.1	20.7 ± 0.2	26.7 ± 0.0
SAR [29]	36.1 ± 0.1	27.6 ± 0.3	23.5 ± 0.4	26.8 ± 1.0	22.0 ± 0.4	27.2 ± 0.2
RoTTA [44]	29.7 ± 0.0	25.6 ± 0.4	29.2 ± 0.2	32.0 ± 1.2	31.2 ± 0.2	29.5 ± 0.3
SoTTA	**39.8 ± 0.0**	**27.9 ± 0.3**	**36.1 ± 0.1**	**41.1 ± 0.1**	**39.0 ± 0.1**	**36.8 ± 0.0**

SoTTA: Impact of the number of noisy samples



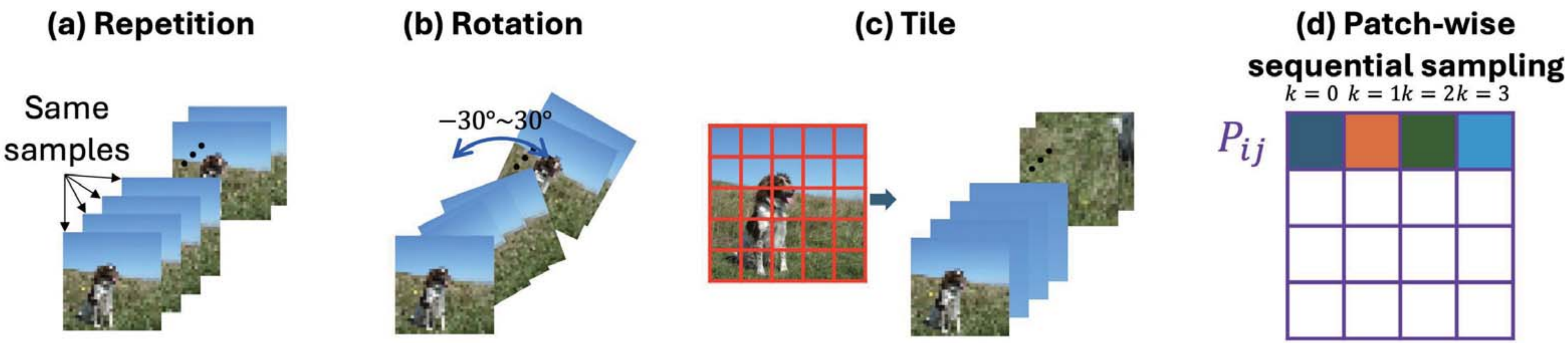
DEX: example images



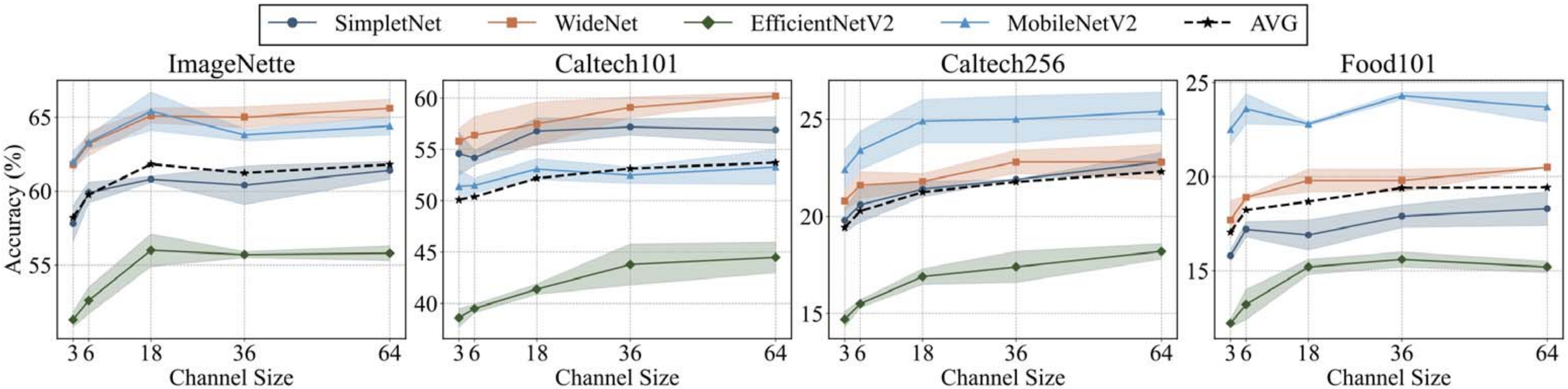
DEX: Comparison of data extension strategies

Table 4: Comparison of data extension strategies.

Method	InputChan	InfoRatio (×)	Accuracy
Downsampling	3	1.0	57.8 ± 1.2
Repetition	64	1.0	56.3 ± 0.8
Rotation	64	1.0	55.7 ± 0.6
Tile per channel	64	21.3	39.3 ± 0.9
Patch-wise seq.	64	21.3	60.4 ± 1.5
DEX	64	21.3	**61.4 ± 0.6**



DEX: Accuracy of DEX varying the channel size



ASK 2025 논문집

Annual Symposium of KIPS 2025

신진학자 워크숍

Cyclic-Consistent Modality Translation between MRI and CT using Diffusion Models

최기환 조교수(서울과학기술대학교)

Cyclic-Consistent Modality Translation between MRI and CT using Diffusion Models

Kihwan Choi
Dept. of Applied Artificial Intelligence
Seoul National University of Science & Technology

May 30, 2025

서울과학기술대학교
SEOULTECH SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

Education & Work

B.S. in Electrical Engineering (1998. 3 ~ 2004. 2)

M.S. in EECS (2004. 3 ~ 2006. 2)

- Wireless Networks (Advisor: Sunghyun Choi)

M.S./Ph.D. in Electrical Engineering (2006. 9 ~ 2014. 4)

- Large-Scale Optimization, Medical Image Reconstruction (Advisors: Lei Xing and Stephen Boyd)

M.S. in Statistics (2011.3 ~ 2013. 1)

- Statistical Learning, Compressed Sensing

SW Solution Lab. (2014.4 ~ 2017. 2)

- Vision for Autonomous Driving / Neural Processing Unit

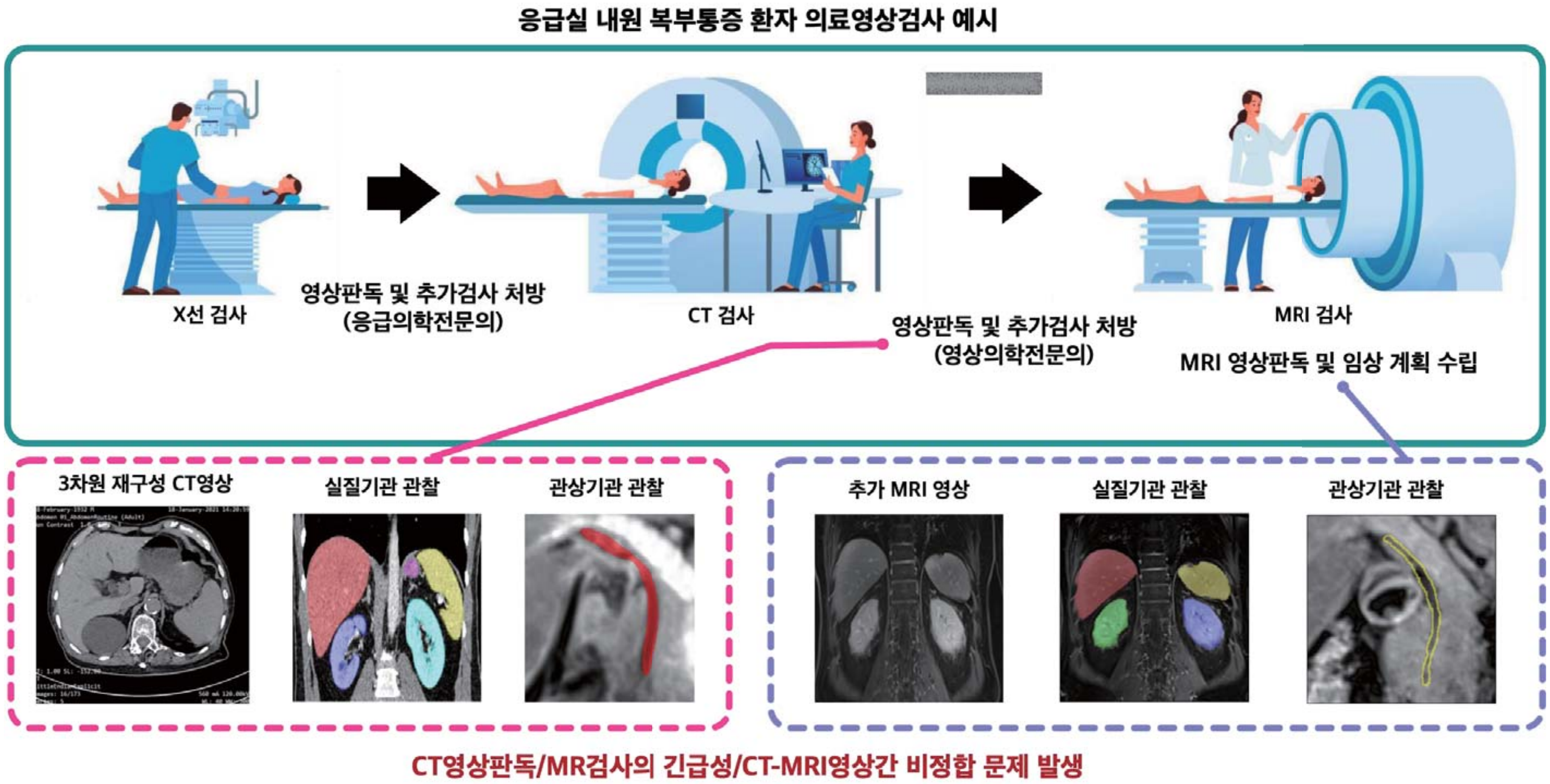
Center for Bionics (2017. 3 ~ 2023.8)

- AI for Medical Image Processing and Diagnosis

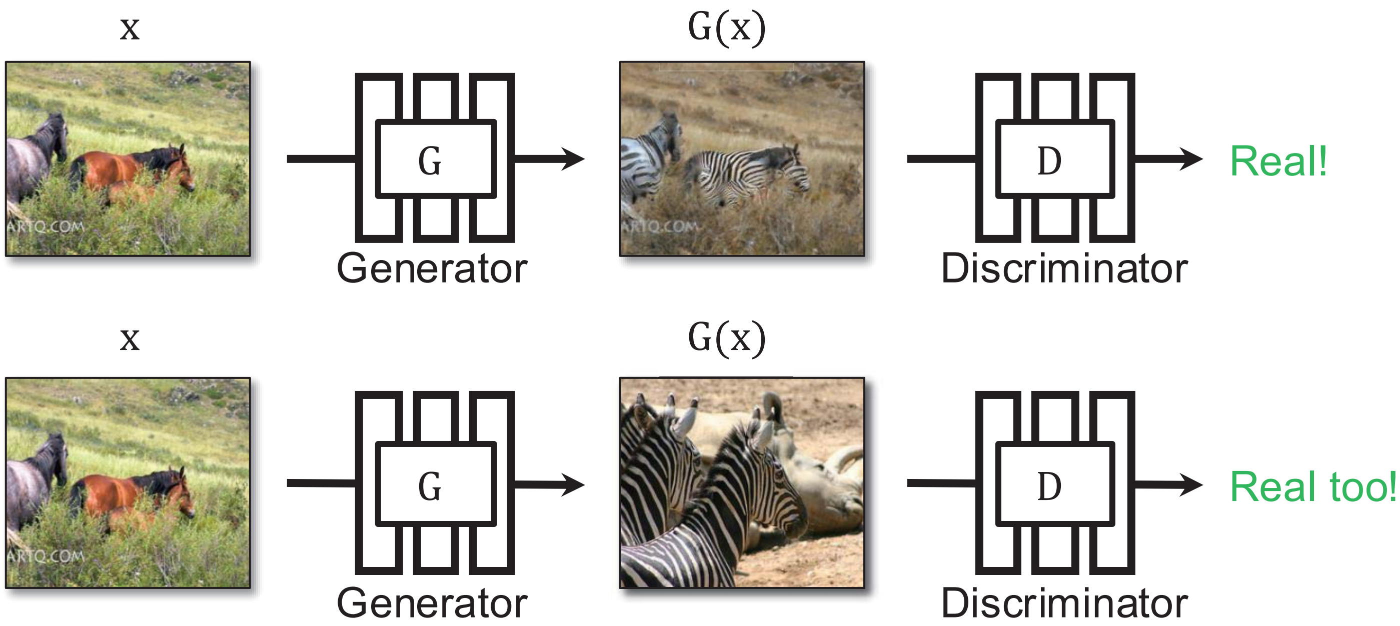
Department of Applied Artificial Intelligence (2023. 9 ~ Present)

- Biomedical AI System Laboratory (BAISLab)

Demand on Modality Translation in ER



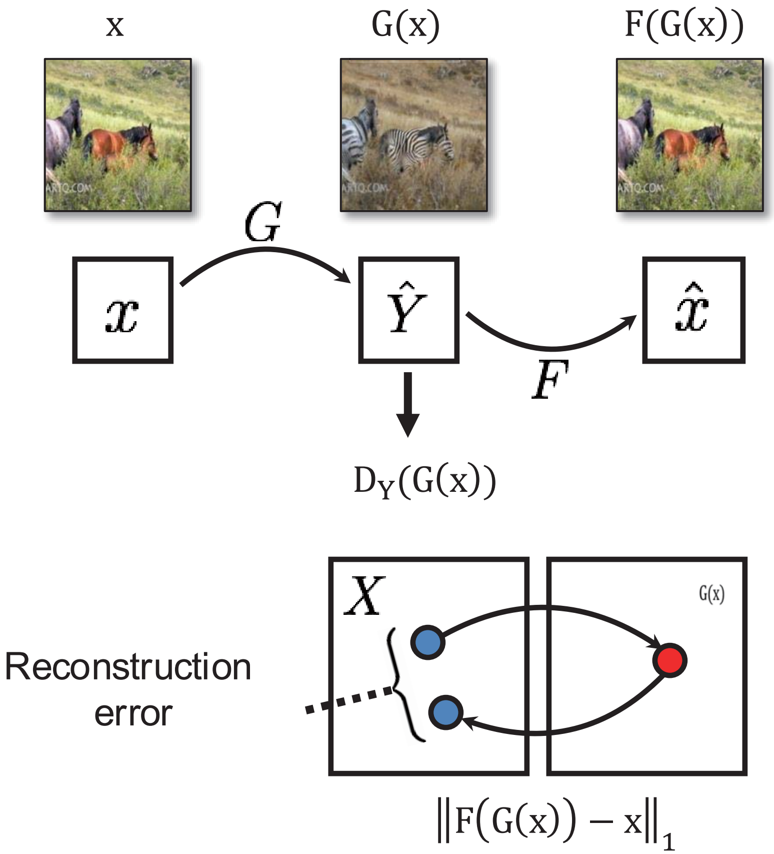
Background: GAN with Unpaired Data

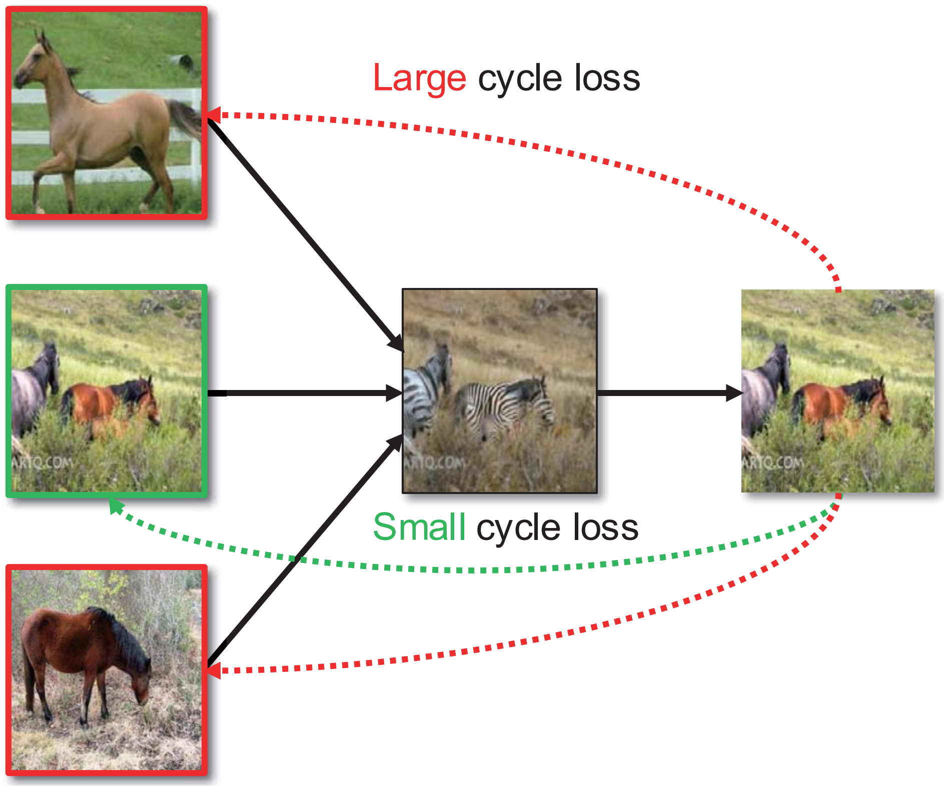


GANs do **not** force output to correspond to input!

Background: CycleGAN

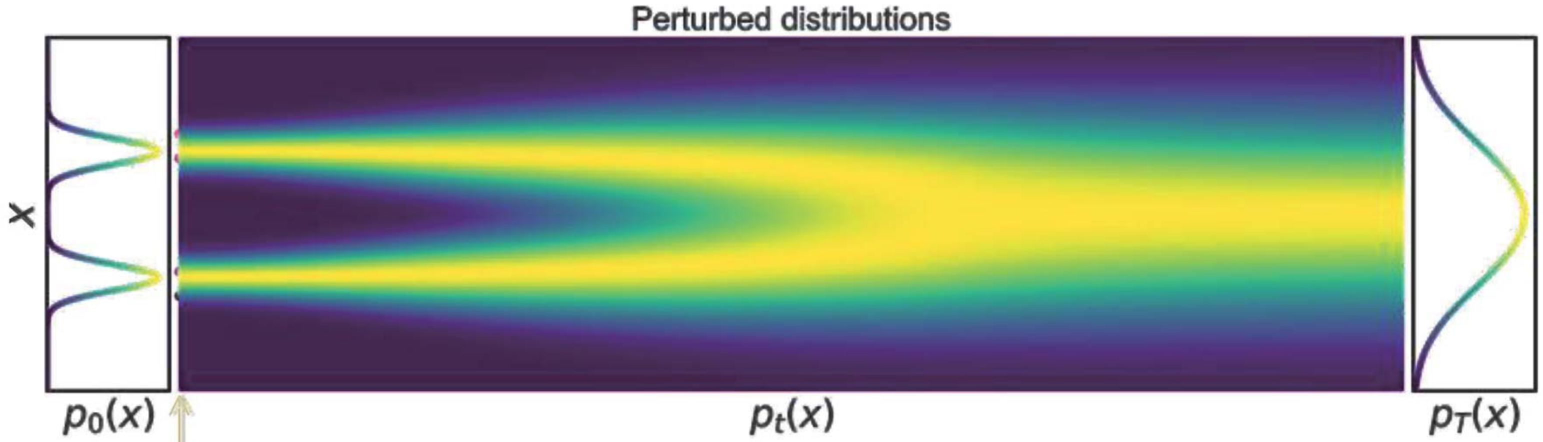
Cycle Consistency Loss

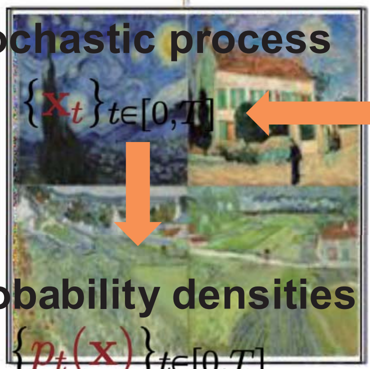




[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Perturbing data with stochastic processes





Stochastic process

$\{\mathbf{x}_t\}_{t\in[0,T]}$

Probability densities

$\{p_t(\mathbf{x})\}_{t\in[0,T]}$

Stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{x}_t = \boxed{\boldsymbol{f}(\mathbf{x}_t, t)}\,\mathrm{d}t + g(t)\,\boxed{\mathrm{d}\mathbf{w}_t}$$

Deterministic drift

Infinitesimal noise

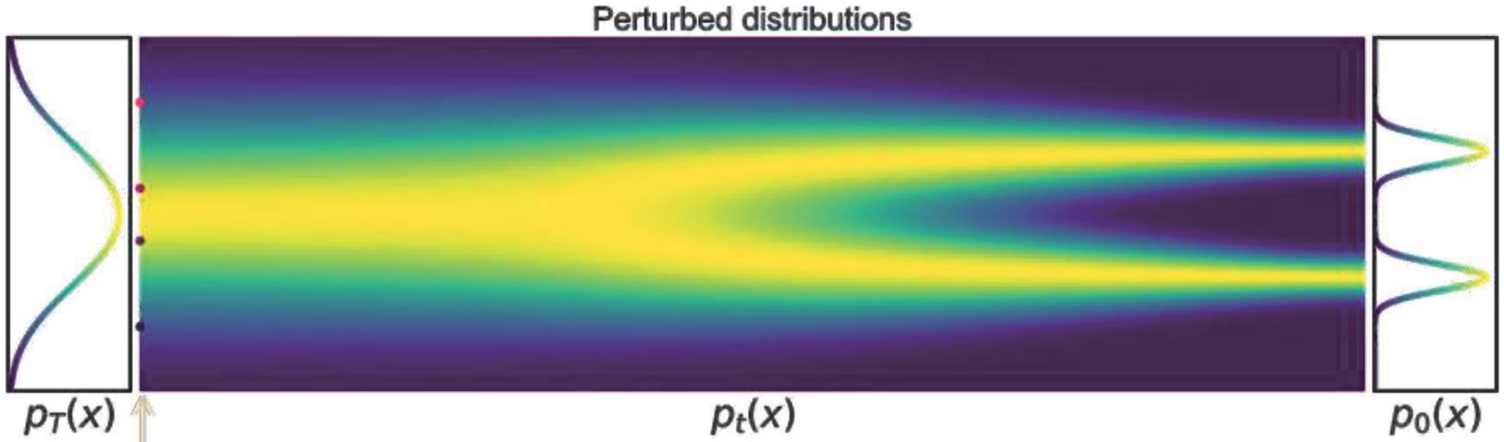
WLOG: Toy SDE

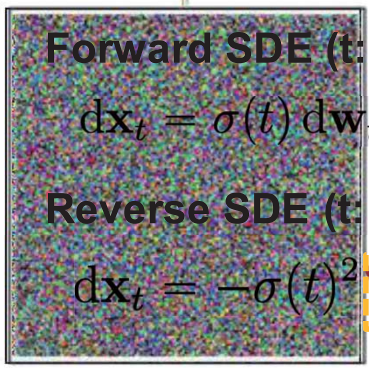
$$\mathrm{d}\mathbf{x}_t = \sigma(t)\,\mathrm{d}\mathbf{w}_t$$

$$p_T(\mathbf{x}) \approx \pi(\mathbf{x})$$

Background: Score-Based Diffusion Models

Generation via reverse stochastic processes





$\pi(\mathbf{x})$
$\approx$
$p_T(\mathbf{x})$

Forward SDE (t: 0→T)

$$\mathrm{d}\mathbf{x}_t = \sigma(t)\,\mathrm{d}\mathbf{w}_t$$

Reverse SDE (t: T→0)

Infinitesimal noise in the reverse time direction

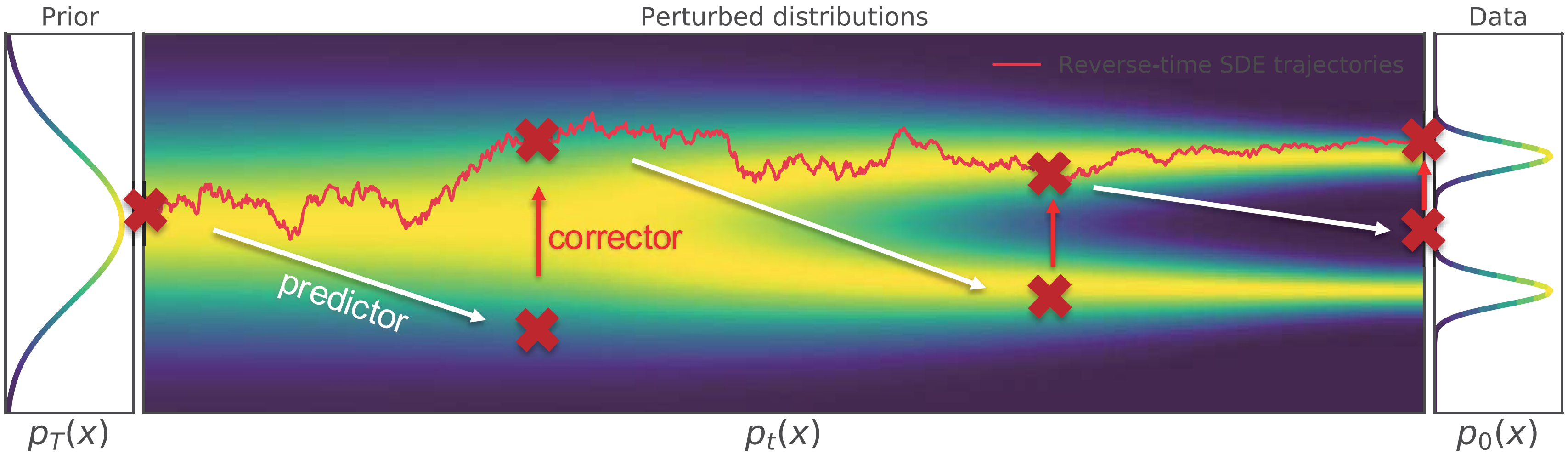
$$\mathrm{d}\mathbf{x}_t = -\sigma(t)^2 \boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)}\,\mathrm{d}t + \sigma(t) \boxed{\mathrm{d}\bar{\mathbf{w}}_t}$$

Score function!

Background: Score-Based Diffusion Models

Predictor-Corrector sampling methods

- Predictor-Corrector sampling.
  - **Predictor:** Numerical SDE solver
  - **Corrector:** Score-based MCMC



Background: Score-Based Diffusion Models

Score-based generative modeling via SDEs

- Time-dependent score-based model

- Training: $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$

$$\mathbb{E}_{t \in \mathcal{U}(0,T)}[\lambda(t)\mathbb{E}_{p_t(\mathbf{x})}[\|\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) - \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)\|_2^2]]$$

- Reverse-time SDE

$$\mathrm{d}\mathbf{x} = -\sigma^2(t)\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)\mathrm{d}t + \sigma(t)\mathrm{d}\bar{\mathbf{w}}$$

- Euler-Maruyama (analgous to Euler for ODEs)

$$\mathbf{x} \leftarrow \mathbf{x} - \sigma(t)^2 \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)\Delta t + \sigma(t)\ \mathbf{z} \quad (\mathbf{z} \sim \mathcal{N}(\mathbf{0}, |\Delta t|\ \mathbf{I}))$$

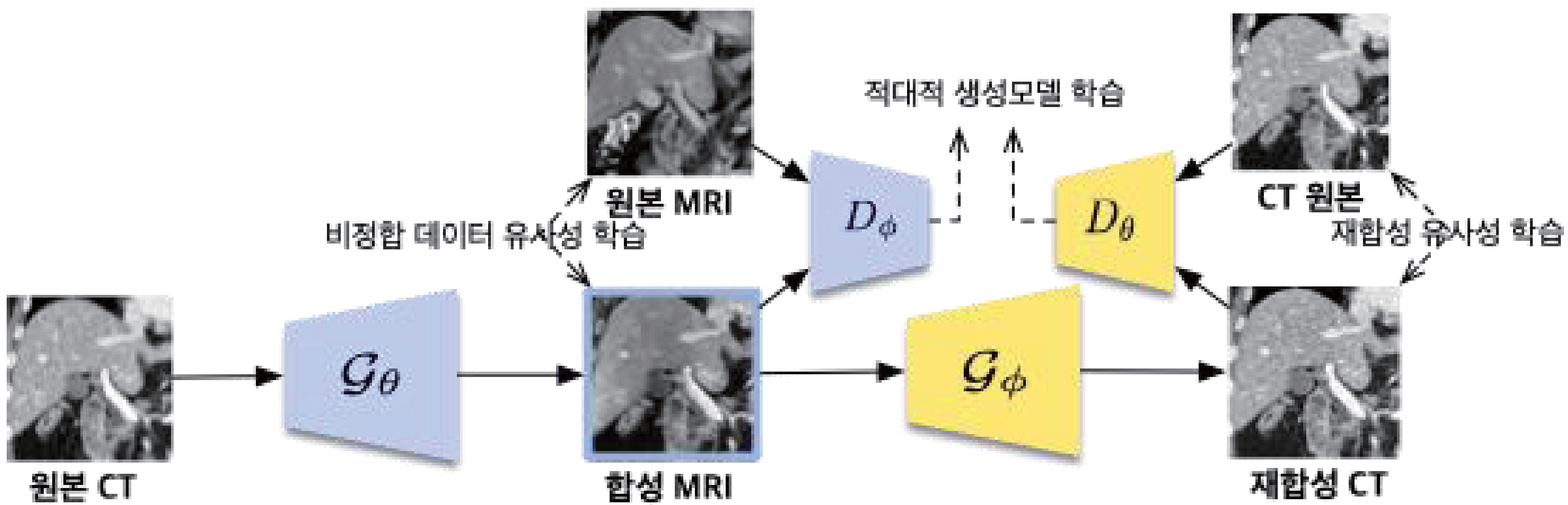
$$t \leftarrow t + \Delta t$$

Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations.” ICLR 2021.

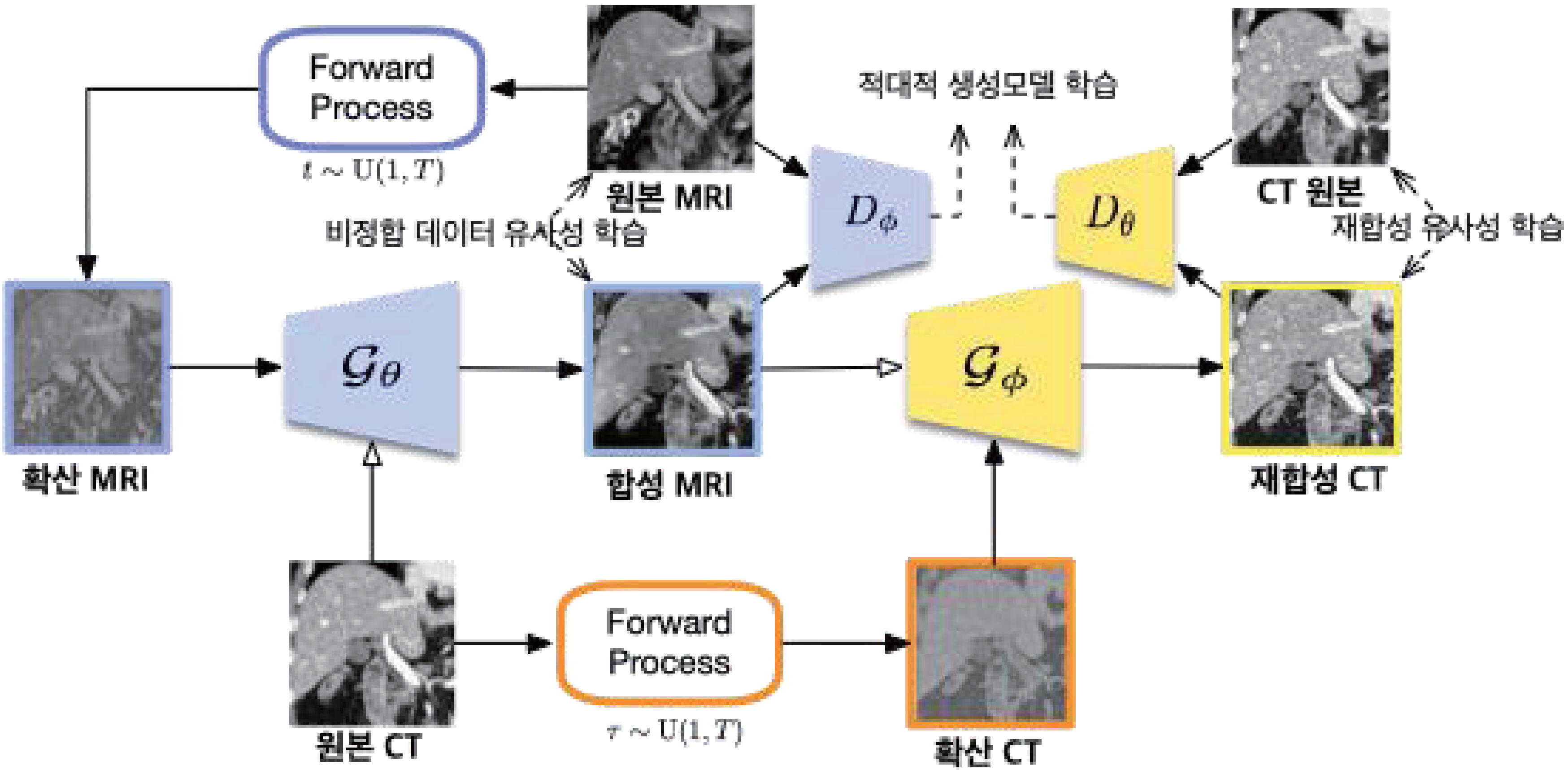
Problem Specification

- **CT2MR Image Translation**
  - Has not been actively explored compared to MR2CT
  - CT is first option at ER, while MR is secondary
  - **Objective**: MR-less prediction of invisible anatomy (e.g. tubular organs)
  - **Constraints**: no change in visible anatomy (e.g. solid organs)
  - Collaboration with **Korea University Medical Center**

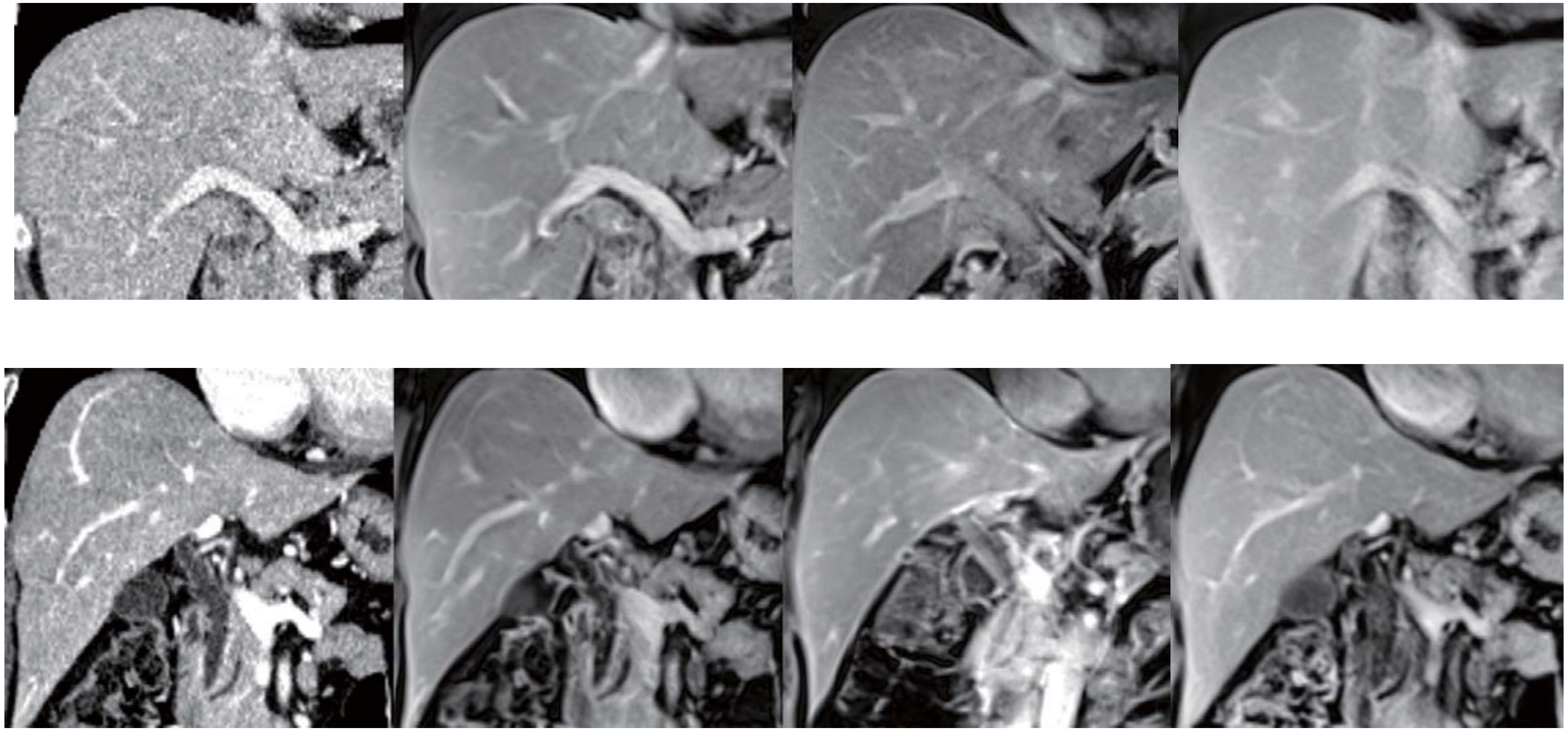
CycleGAN for CT-MRI Translation



Cyclic Diffusion Model for CT-MRI Translation



Cyclic Diffusion Model for CT-MRI Translation

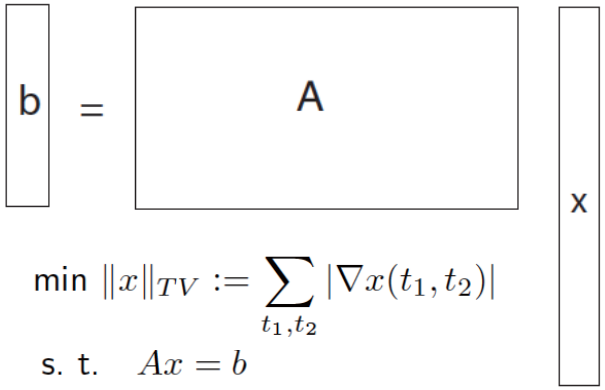


CT | Synthetic MR (ours) | Real MR (not matched) | SOTA (SynDiff)

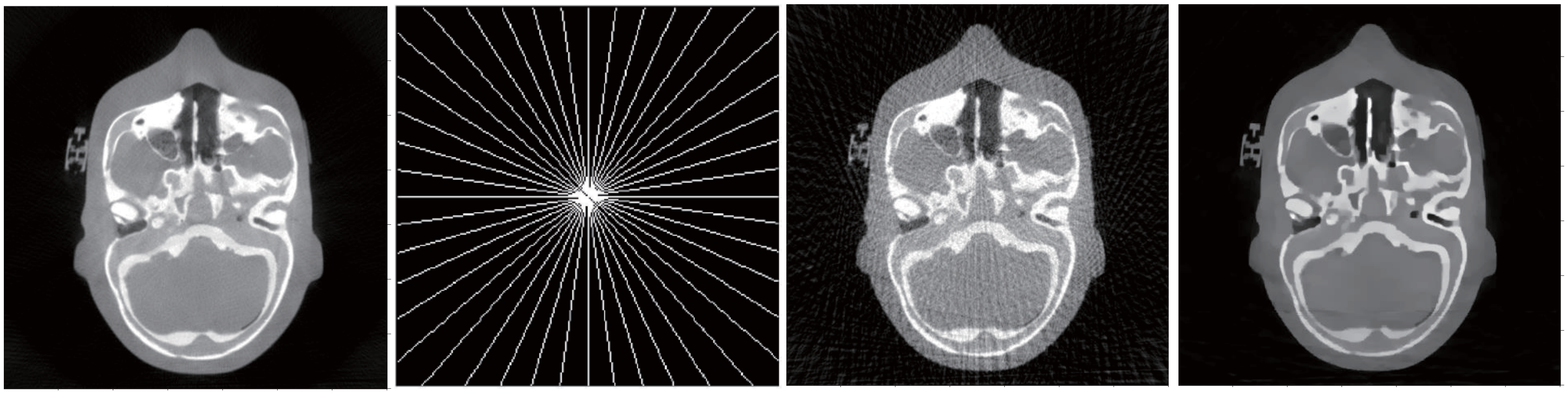
Previous Research: Sparse-View CT

“Compressed sensing based CBCT reconstruction with a first-order method”
Medical Physics 2010

- Undersampled measurements
  - Fundamental theorem of algebra and Nyquist theorem: original signal cannot be recovered w/o aliasing
- Under some conditions we can perfectly recover signal
  - Signal can be expressed with sparse representations



$$\min \|x\|_{TV} := \sum_{t_1,t_2} |\nabla x(t_1,t_2)|$$
$$\text{s. t.} \quad Ax = b$$

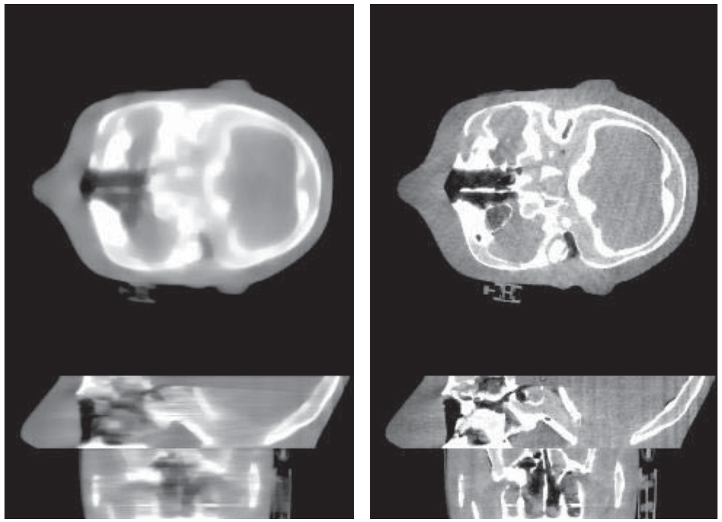


Original | **Undersampled Measurements** | **Conventional** | **CS Reconstruction**

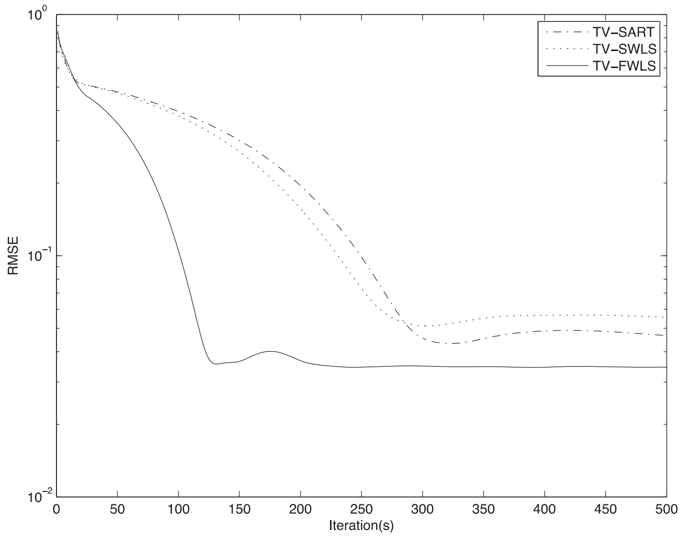
Previous Research: Sparse-View CT

"A Fourier-based compressed sensing technique for accelerated CT image reconstruction" ***Physics in Medicine & Biology*** 2014

- Fast and Accurate Compressed Sensing for CT Imaging
  - Optimization solver becomes slow and inefficient when Hessian matrix is ill-conditioned: $\lambda_{\max}(A^T A) \gg \lambda_{\min}(A^T A)$
  - Approach: Fourier-domain preconditioning inspired by conventional FBP:
    $\tilde{A} := H^{1/2}\mathcal{F}WA$ and $\lambda_{\max}(\tilde{A}^T \tilde{A}) \approx \lambda_{\min}(\tilde{A}^T \tilde{A})$



Naïve CS **Ours**



Convergence Rate

Previous Research: Self-Supervised Denoising

"Self-supervised inter-and intra-slice correlation learning for low-dose CT image restoration" ***Expert Systems with Applications*** 2022

- Self-Supervised Image Denoising
  - Applied self-supervised learning to denoise CT images **without references**
  - Trained to recover partially blinded inputs: $\mathcal{L}_{\text{intra}}(G;X) = \sum_{J\in\mathcal{J}} \mathbb{E}_{X_{J^c}} \mathbb{E}_{X_J|X_{J^c}} \|g(\mathbf{x}_{J^c}) - \mathbf{x}_J\|_{\ell_2}$
  - Similarity between denoised images and thicker slices: $\mathcal{L}_{\text{inter}}(G;X) = \sum_{J\in\mathcal{J}} \mathbb{E}_X \|[G(f_J(\mathbf{x}))]_J - \tilde{\mathbf{x}}_J\|_{\ell_1}$
  - Two-stage training strategy: offline pretraining and online finetuning

Noisy Input | **Offline Pretrained** | **Online Finetuned**

Previous : Self-Supervised Denoising

"Self-supervised denoising of projection data for low-dose cone-beam CT"
Medical Physics 2023

- Self-Supervised Projection Denoising
  - Ground-truth **not acquirable** in CBCT with flat panel detector (due to scattering)
  - Applied self-supervised learning to denoise CBCT projections **without references**
  - Trained to recover partially blinded inputs
  - Considered both pixel-wise and view-wise statistical correlations

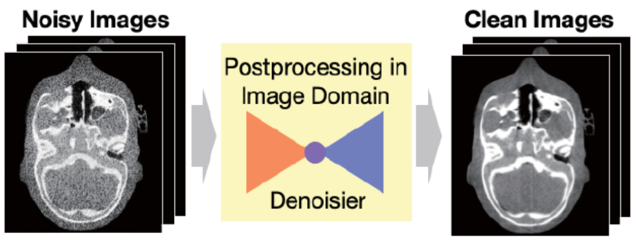
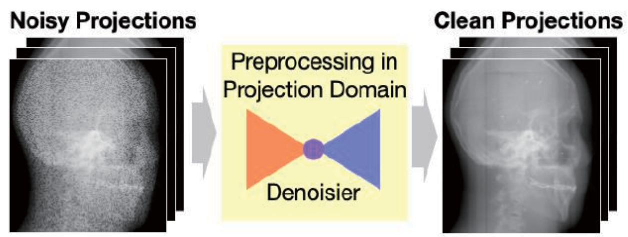
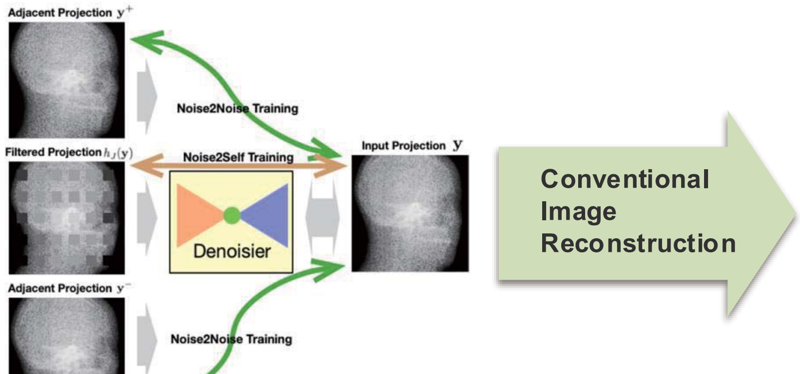


Image Postprocessing

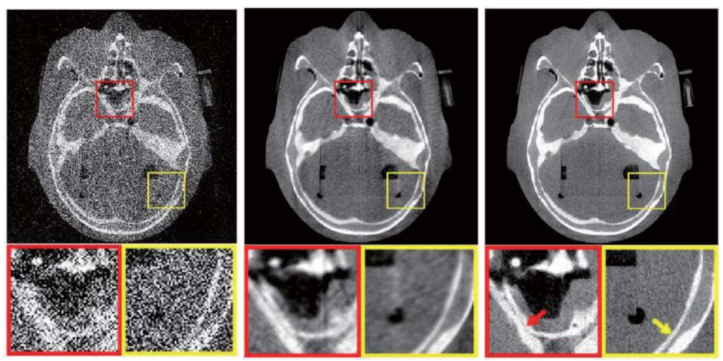
vs



Data Preprocessing



(Noise+Adjacent)2Self Training


No preprocessing | Ours | Ground Truth

신진학자 워크숍

Pathfinding Future BCI Systems Through Full-Stack Design Space Exploration

이현준 교수(한양대학교)

Pathfinding Future BCI Systems Through Full-Stack Design Space Exploration

Hunjun Lee

E-Mail: hunjunlee@hanyang.ac.kr
Web: hunjunlee.github.io

Assistant Professor
@Hanyang University

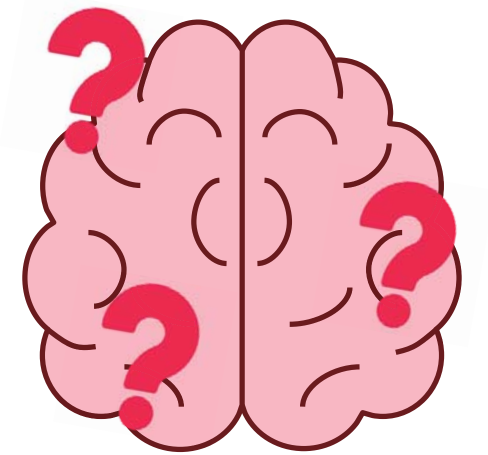
Research interests

- **In Silico Brain Modeling Processor**
  - A flexible digital circuit design **[MICRO’19]**
  - Event-driven brain simulation **[ASPLOS’21]**
  - Speculative brain simulation **[HPCA’22]**

- **AI Algorithms & Hardware Performance Evaluations**
  - SNN vs. ANN **[Neurocomputing’21]**

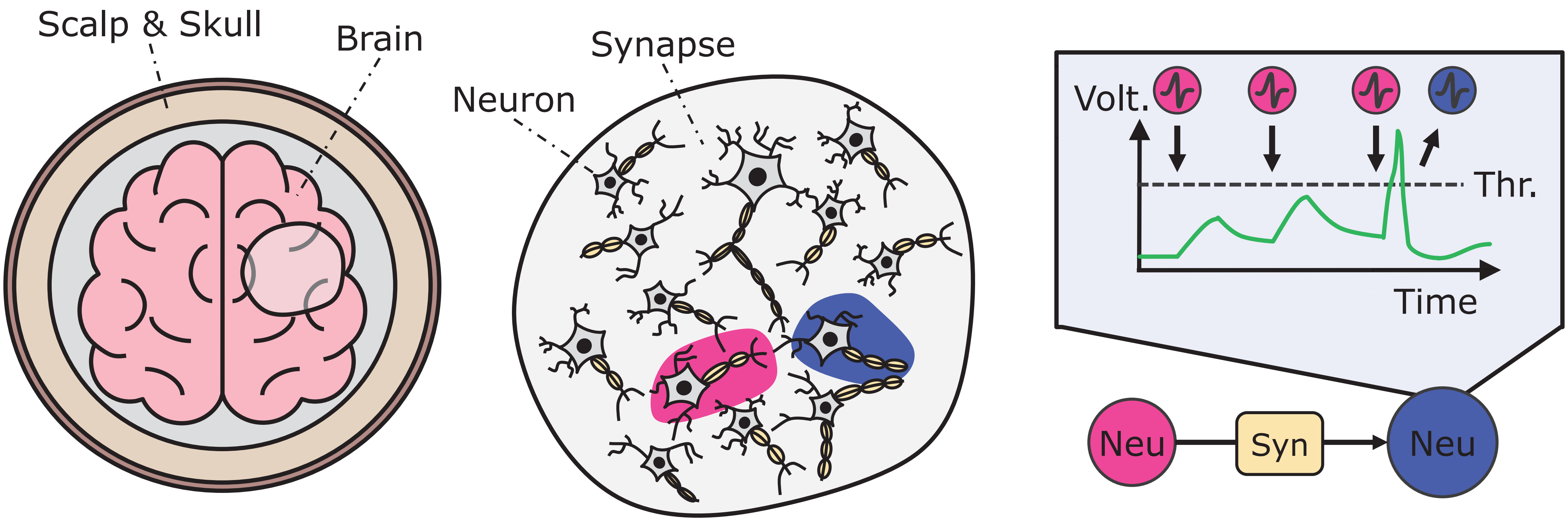
- **Analog-Based Process-in-Memory Architecture**
  - 3D NAND Flash-based PIM **[MICRO’22]**

- **In Vivo BCI Signal Processor**
  - Spike-driven BCI processor **[MICRO’24]**
  - Learning-enabled BCI processor **[ISCA’25 (To Appear)]**



Overall structure of the brain

- **The brain consists of a biological neural network**



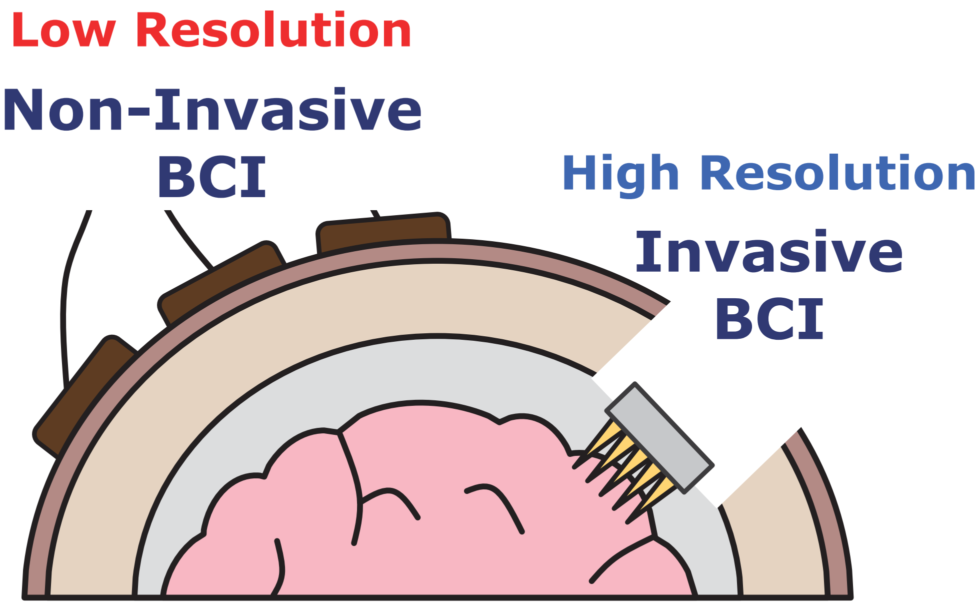
Human Brain Structure

Biological Neural Network

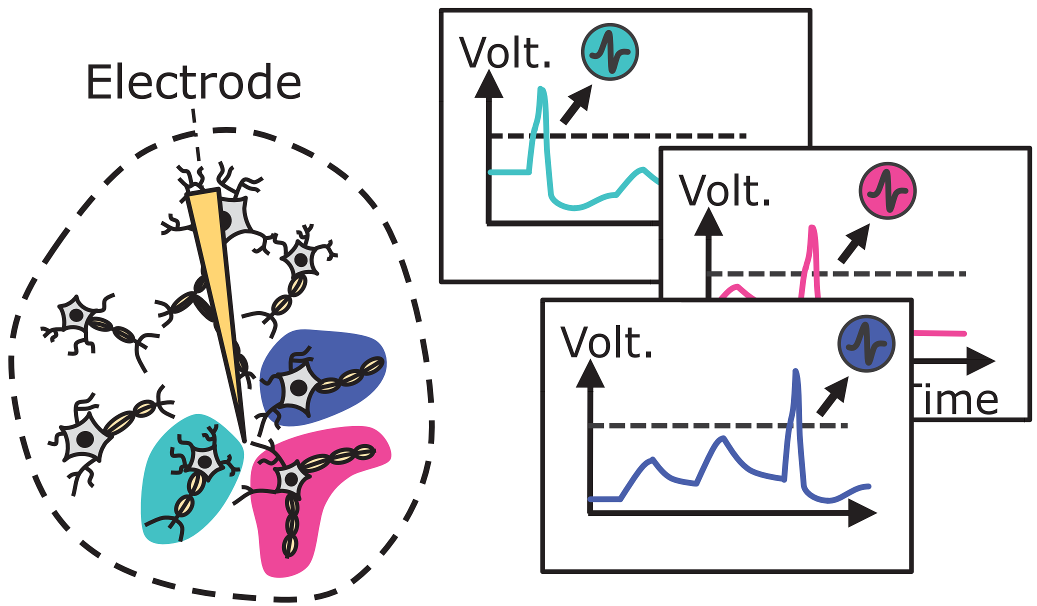
Internal Mechanism

Brain-computer interfacing

- **Brain-computer interfaces (BCIs) are electrophysiological devices that directly record and stimulate the neurons**



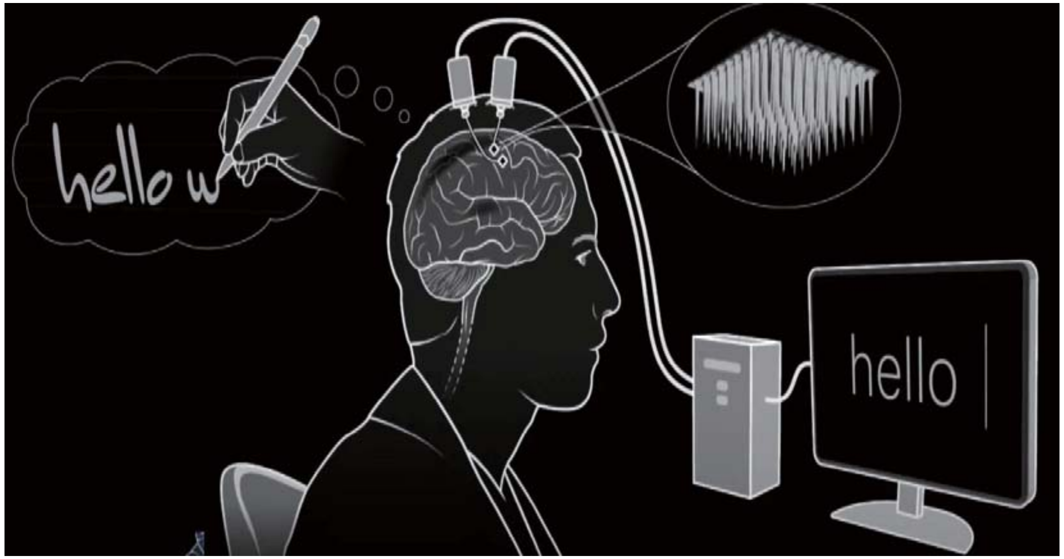
BCI Implementation Variants



Brain signals recorded using invasive BCIs

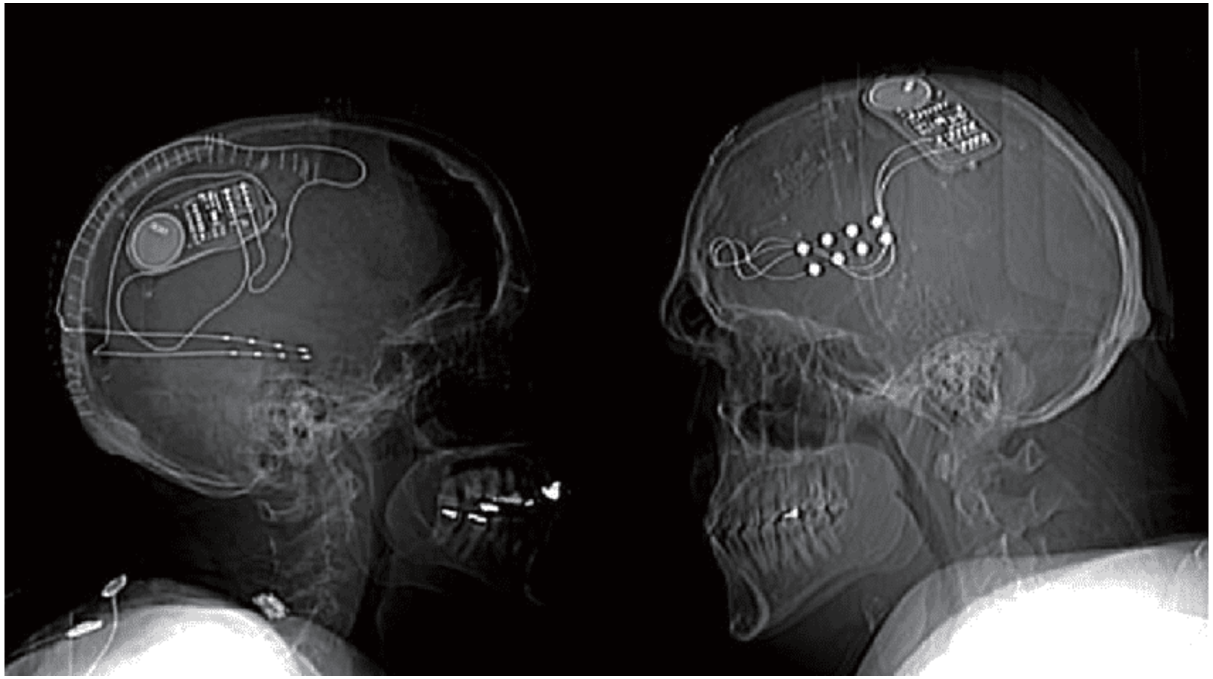
Use case #1: Neural prosthesis

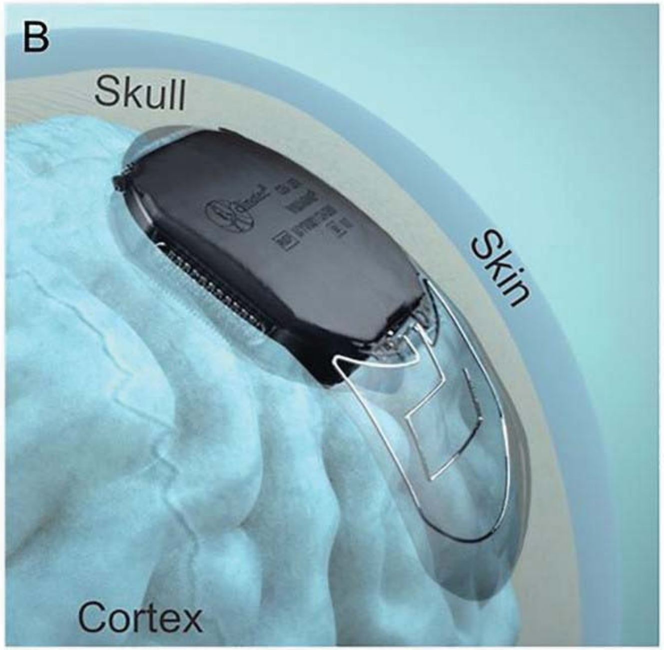
- **BCI signals reveal intended body movements by decoding signals at the motor cortex**
  – Enables various applications including texting, game playing, robot arm movements



Use case #2: Seizure prevention

- **BCI devices help cure neurological disorders by stimulating the brain at the onset**
  - There are multiple FDA-approved medical devices (e.g., Neuropace, GBrain)





Increasing brain computer interface market size

synchron

BRAINGATE
TURNING THOUGHT INTO ACTION

INTEGRA
LIMIT UNCERTAINTY

COMPUMEDICS® NeuroScan™

NEURALINK

CADWELL®
Helping you help others

CorTec

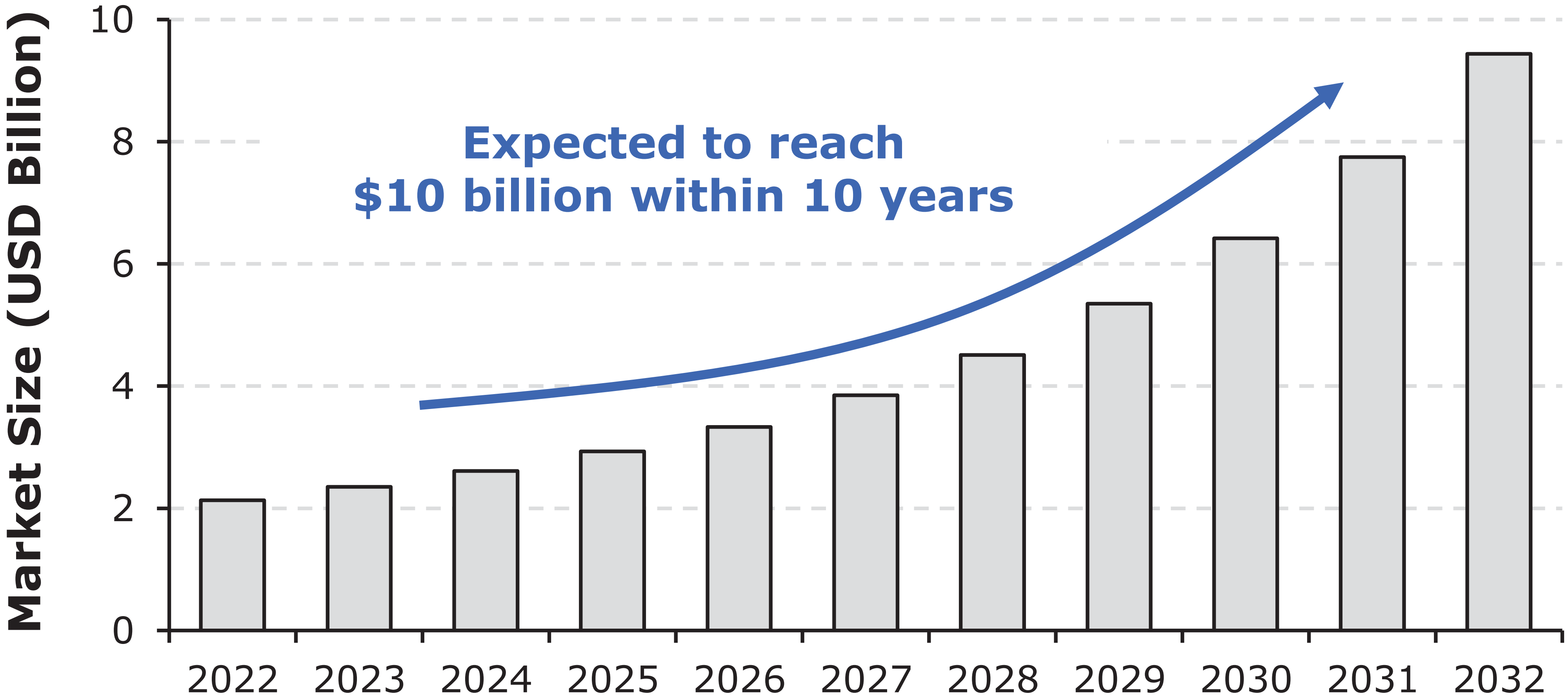
natus.

Medtronic
Alleviating Pain · Restoring Health · Extending Life

BRAIN PRODUCTS
Solutions for neurophysiological research

NeuroSky®
Body and Mind. Quantified.

Increasing brain computer interface market size



[source: Precedence Research]

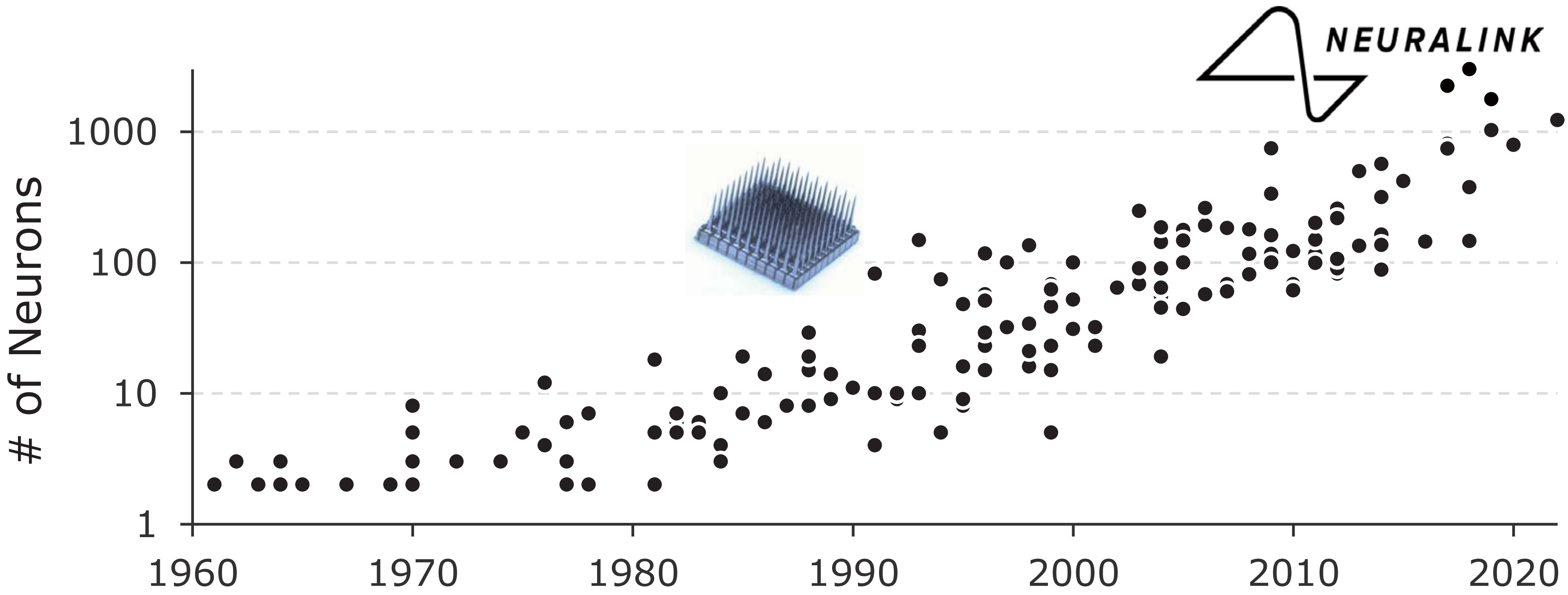
The recent trends resolved various challenges in realizing practical BCIs

Reduced surgical risks w/ automated surgery

- **Neuralink developed a robot to implant the BCI to the human brain (FDA-approved)**
- **Can be implanted with a small burr hole in the brain**

Scaling trends of the electrodes

- **The number of recorded neurons is increasing at a rapid rate**
  - The electrode scales 2x every 6 years (2x every 2 years recently)

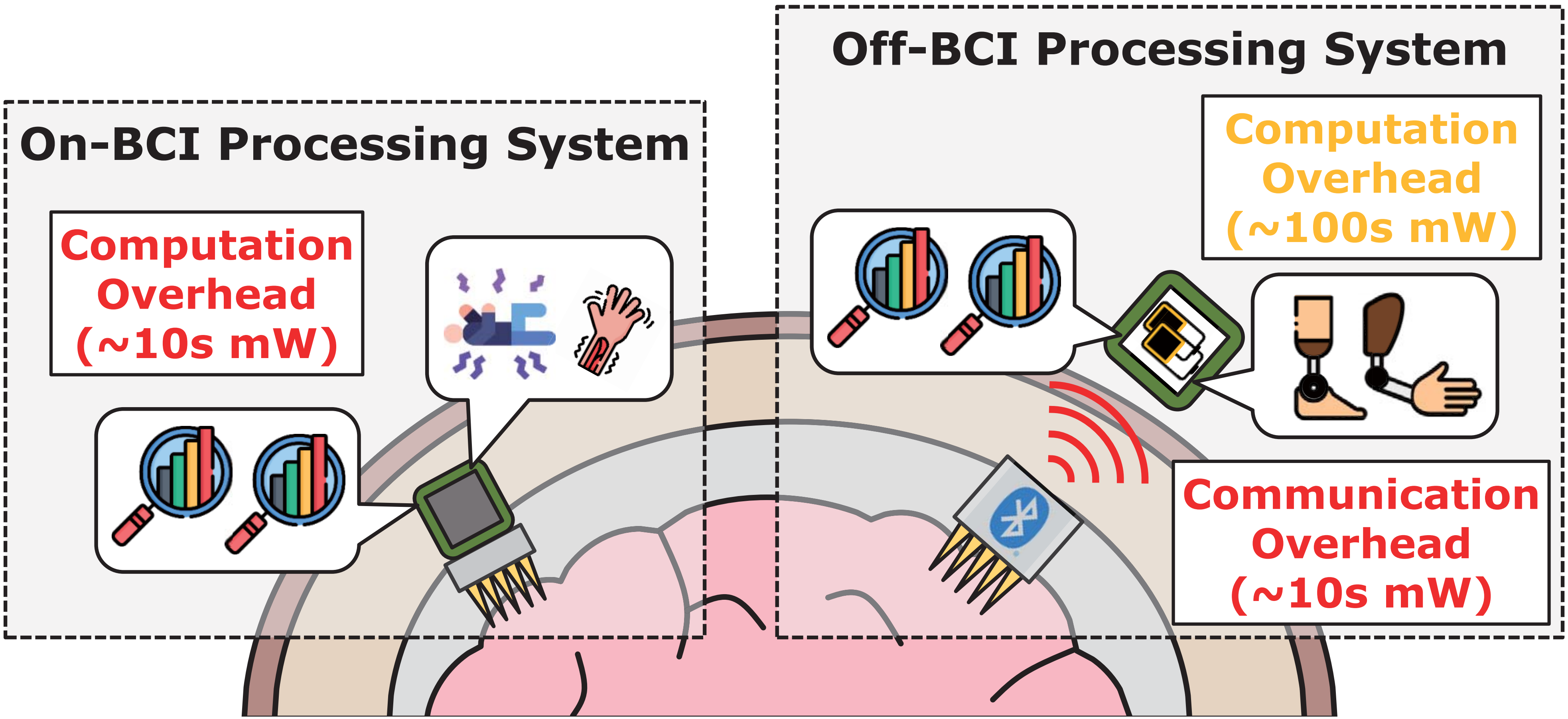


Opensource datasets

- **Communities are releasing neurophysiological datasets as an open problem**
- **Neuralink are releasing part of the monkey datasets to the research communities**

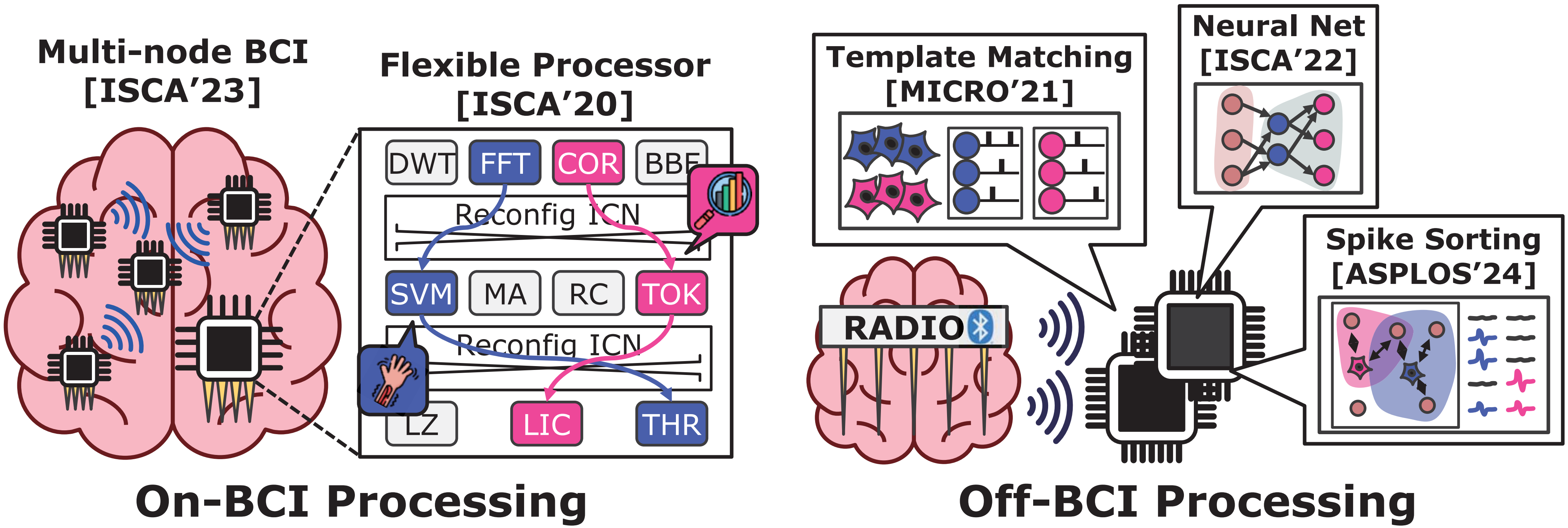
Right time to design a processing system for this new type of I/O

Architectural Perspective: Processing System for Brain-Computer Interfacing



Architectural Perspective: Processing System for Brain-Computer Interfacing

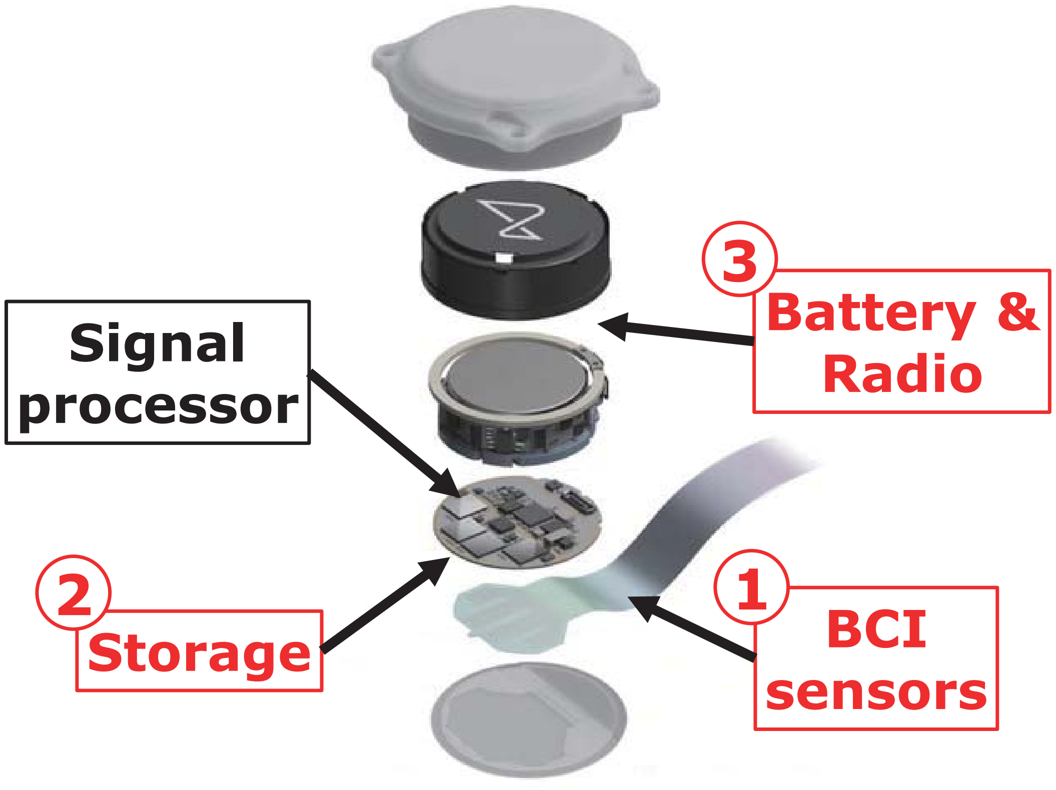
- There are a few architectural studies that focus on **designing a dedicated processor** for brain-computer interfacing



We need a full-stack design space exploration to find the best system!

Full-stack design space exploration

- We should fully explore various design points including the (1) BCI signals, (2) storage, and (3) battery & radio



Research plans

- **Sensor: "Spike-driven architecture" for BCI processing**
  [NeuroLobe – MICRO'24]
  - **Rearchitect a neuromorphic-style processor** for the purpose of supporting various BCI algorithms
- **Storage: "Learning-enabled" NVM-assisted BCI system**
  [MemBrain – ISCA'25 (Accepted)]
  - **Propose an NVM-driven acceleration system to** handle BCI processing with learning support
- **Battery & Radio: "Communication and power-aware" BCI scheduling system**
  [Ongoing]
  - **Design a low-cost scheduler and** to handle battery and thermal imbalance among distributed BCI nodes

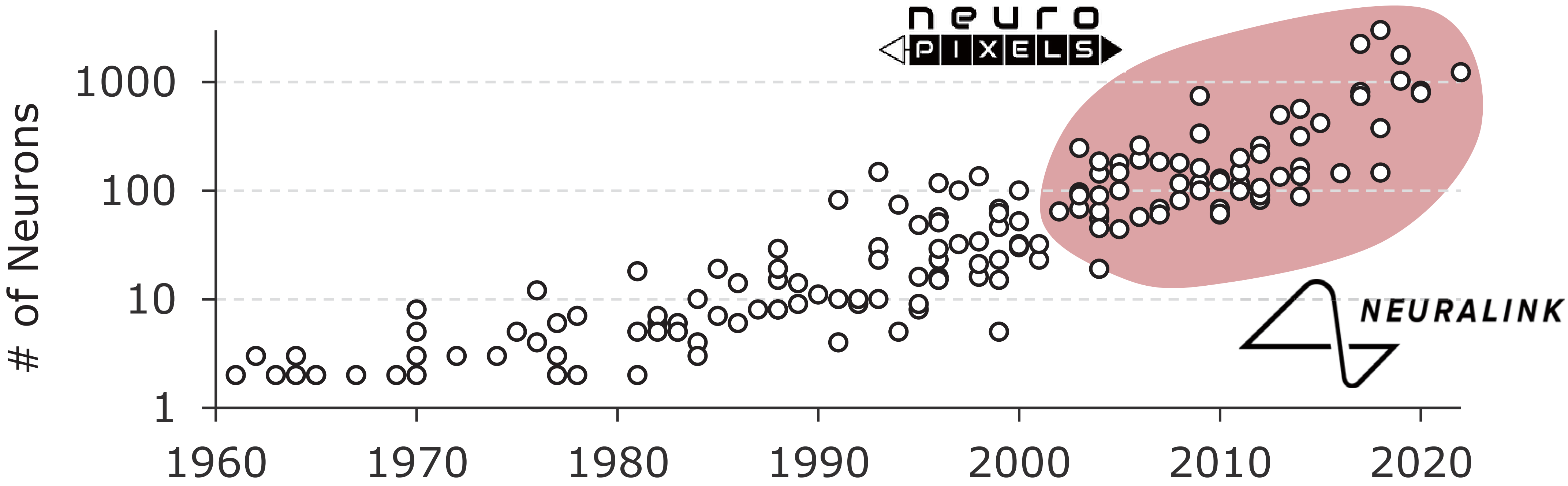
Research plans

- **Sensor: "Spike-driven architecture" for BCI processing**
  [NeuroLobe – MICRO'24]
  - **Rearchitect a neuromorphic-style processor** for the purpose of supporting various BCI algorithms
- **Storage: "Learning-enabled" NVM-assisted BCI system**
  [MemBrain – ISCA'25 (Accepted)]
  - **Propose an NVM-driven acceleration system to** handle BCI processing with learning support
- **Battery & Radio: "Communication and power-aware" BCI scheduling system**
  [Ongoing]
  - **Design a low-cost scheduler and** to handle battery and thermal imbalance among distributed BCI nodes

Challenge: Stevenson's scaling law

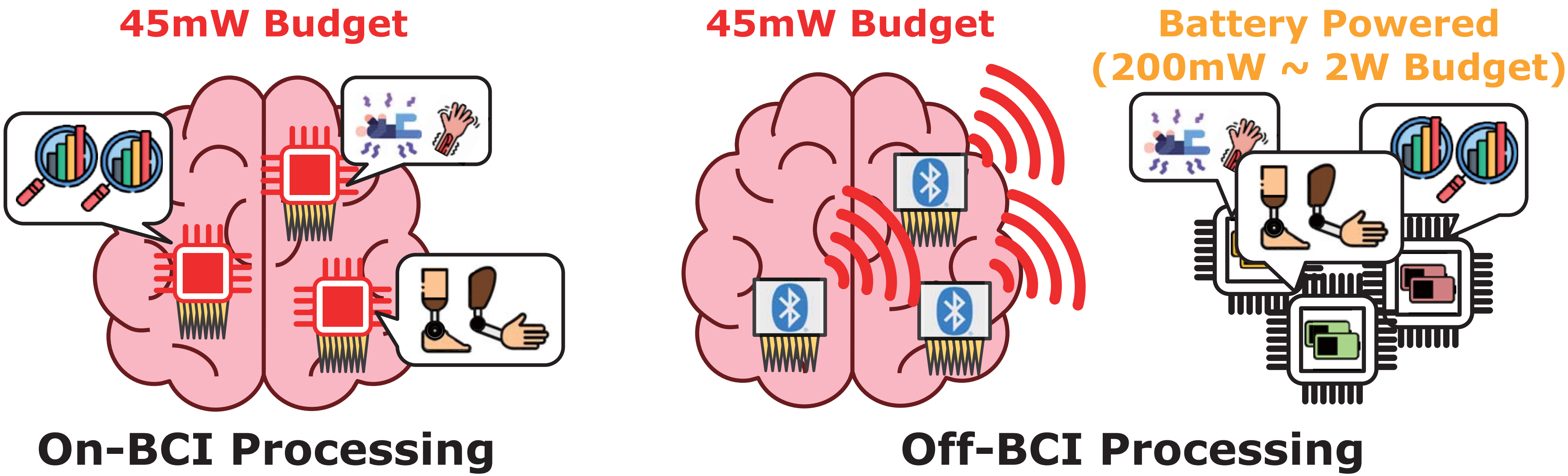
- **Invasive BCIs scale up to record a larger number of neurons**

1. **High communication overhead** to transfer the BCI signals (> ~10s mW)
2. **High computation overhead** to process the signals



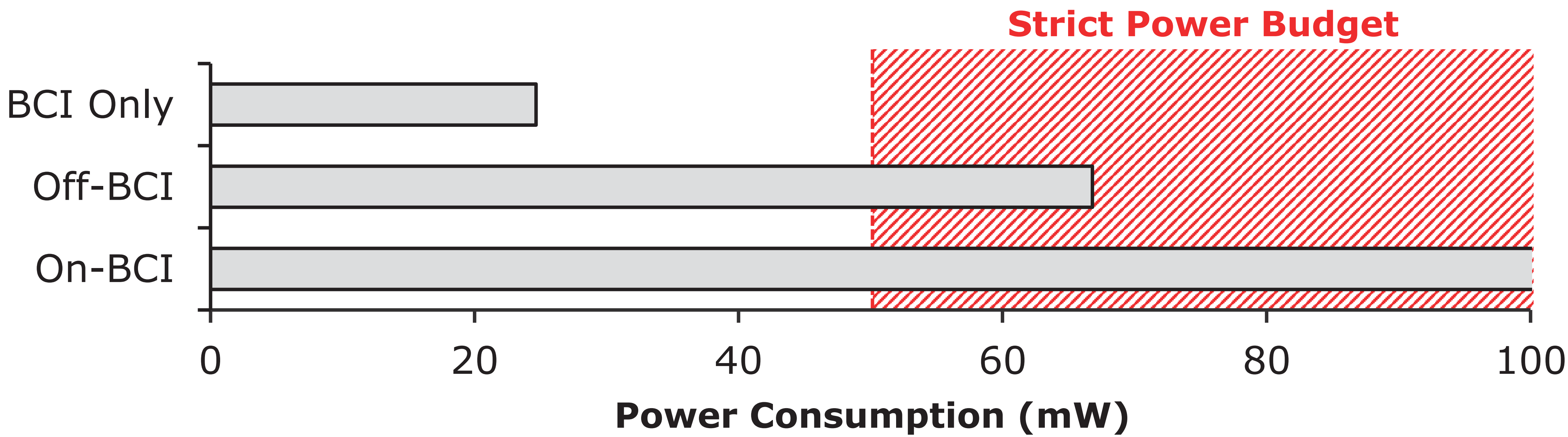
Challenge: Stevenson's scaling law

- The BCI processor **violates the thermal budget** as the number of electrodes scale (>200 Mbps)



Challenge: Stevenson’s scaling law

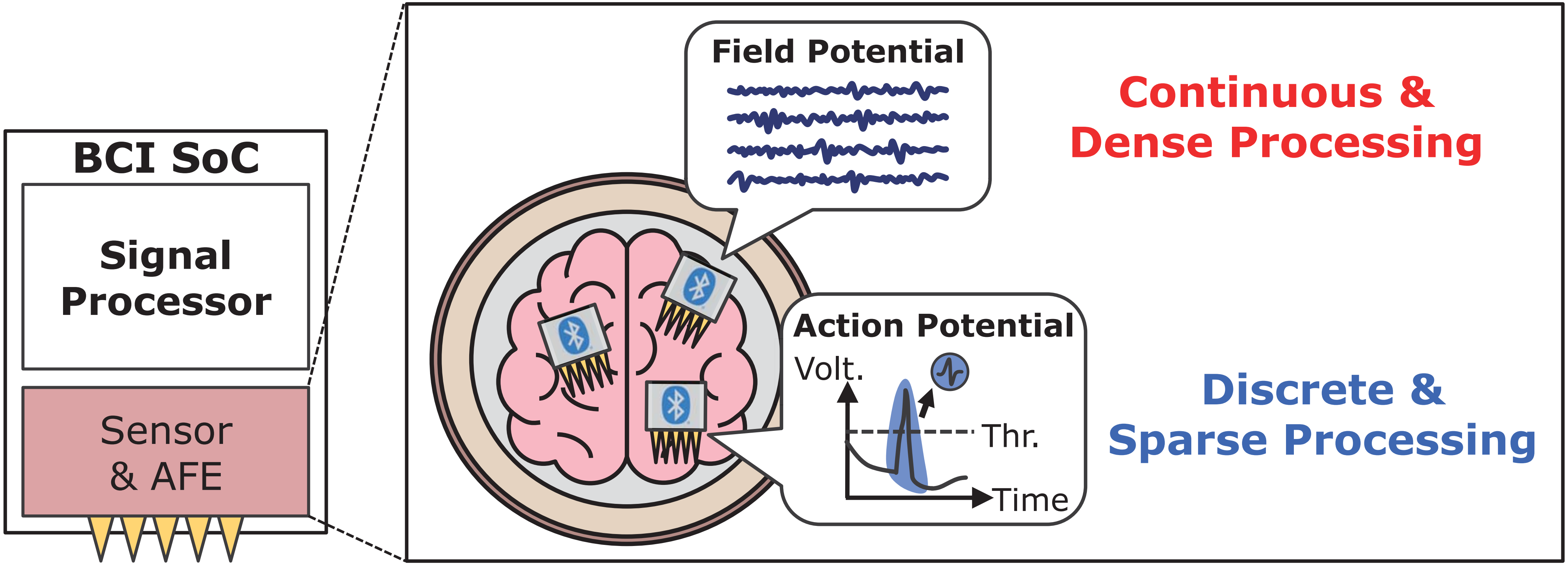
- The BCI processor **violates the thermal budget** as the number of electrodes scale (>200 Mbps)



The system should support scaled-up BCI within the power budget

Solution: Spike-driven processing system

- **Utilize only the spiking nature of the BCI signals?**

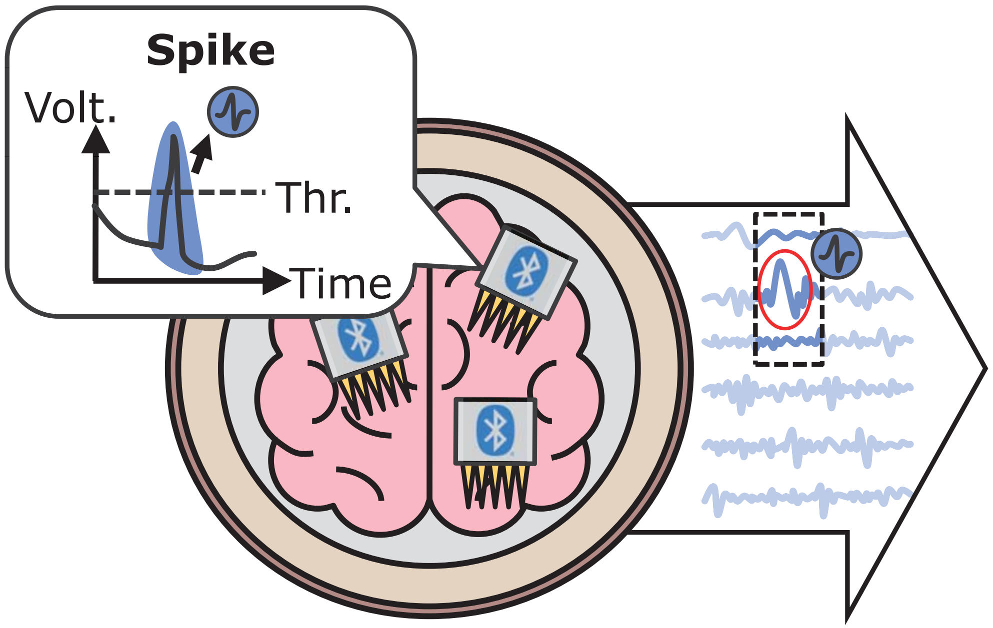


Reduce the computation & communication overhead using spikes

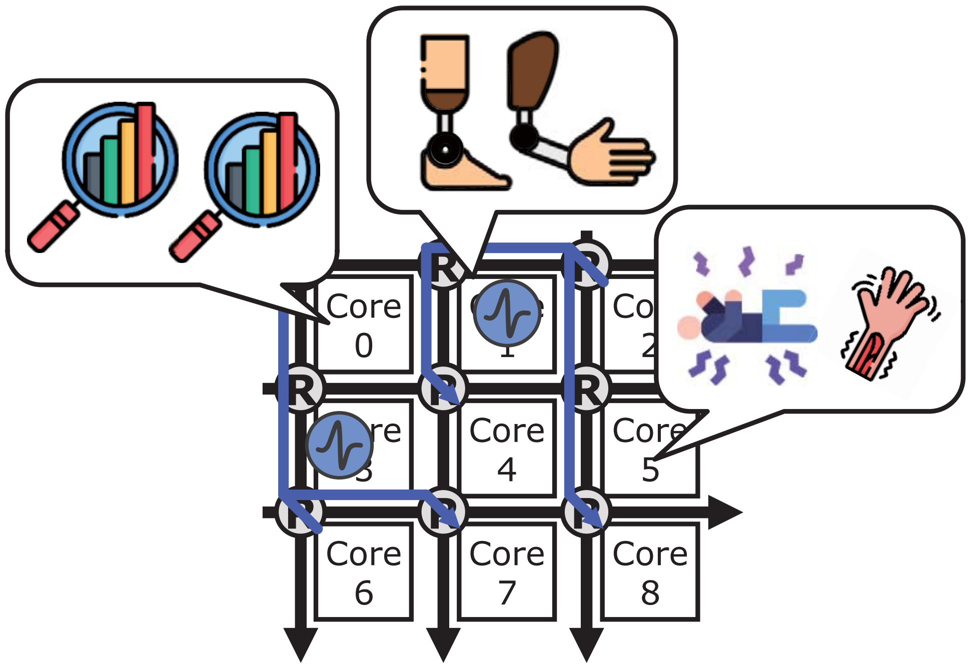
Solution: Spike-driven processing system

- **We utilize the spiking nature of the BCI signals**

1. **Low communication overhead** by transferring only spike signals
2. **Efficient event-driven computation** using a neuromorphic processor

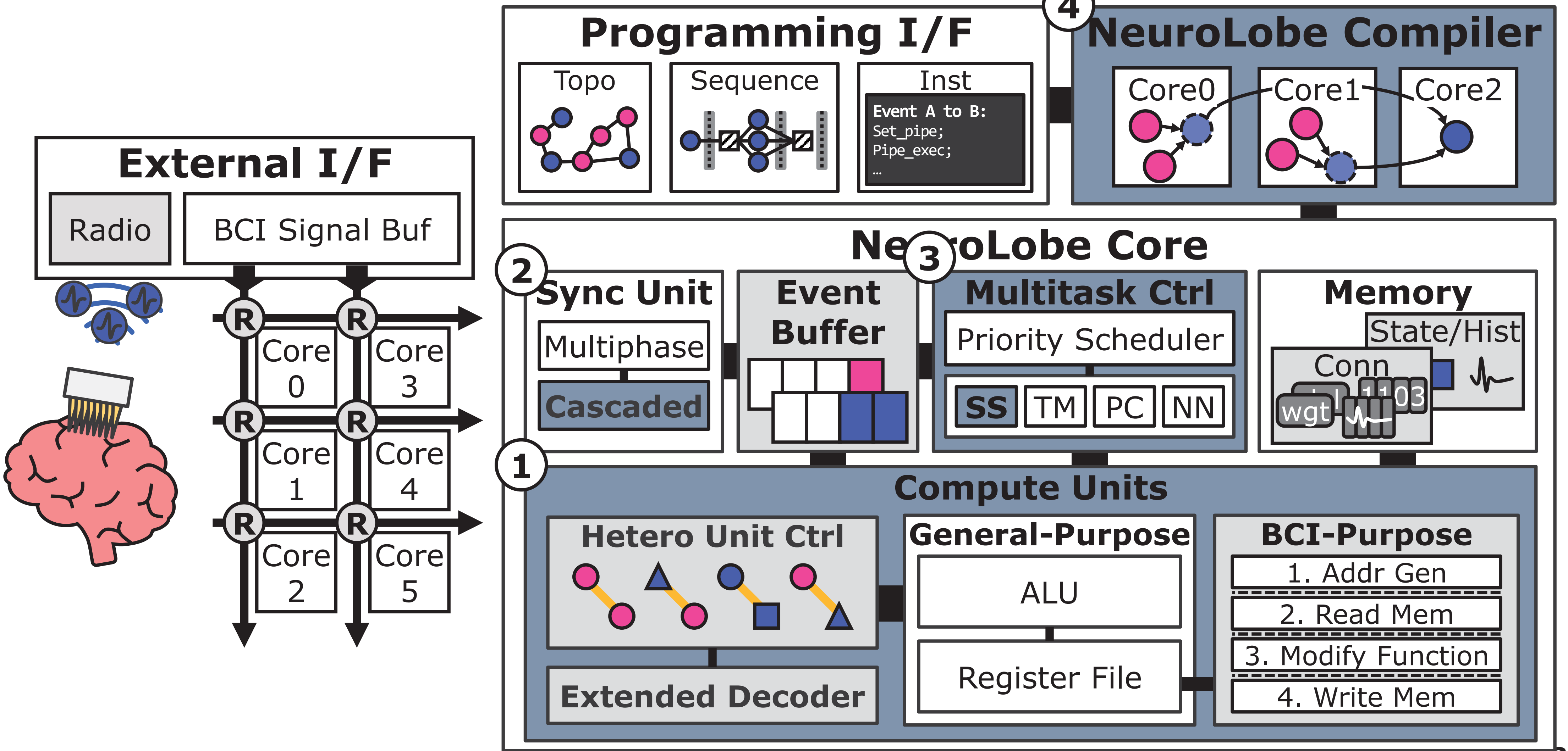


BCI Implant Site



Sparse Signal Accelerator

Overview: NeuroLobe system



Research plans

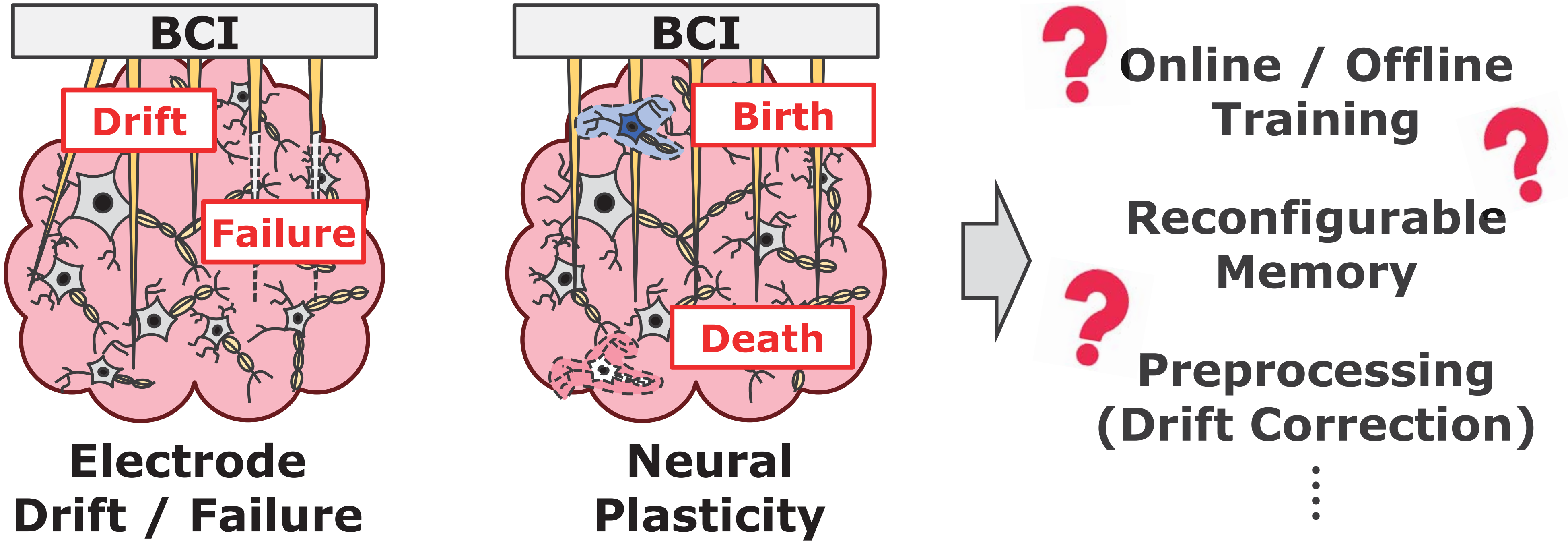
- **Sensor: "Spike-driven architecture" for BCI processing**
  [NeuroLobe – MICRO'24]
  – **Rearchitect a neuromorphic-style processor** for the purpose of supporting various BCI algorithms

- **Storage: "Learning-enabled" NVM-assisted BCI system**
  [MemBrain – ISCA'25 (Accepted)]
  – **Propose an NVM-driven acceleration system to** handle BCI processing with learning support

- **Battery & Radio: "Communication and power-aware" BCI scheduling system**
  [Ongoing]
  – **Design a low-cost scheduler and** to handle battery and thermal imbalance among distributed BCI nodes

Challenge: Adaptive Processing Support

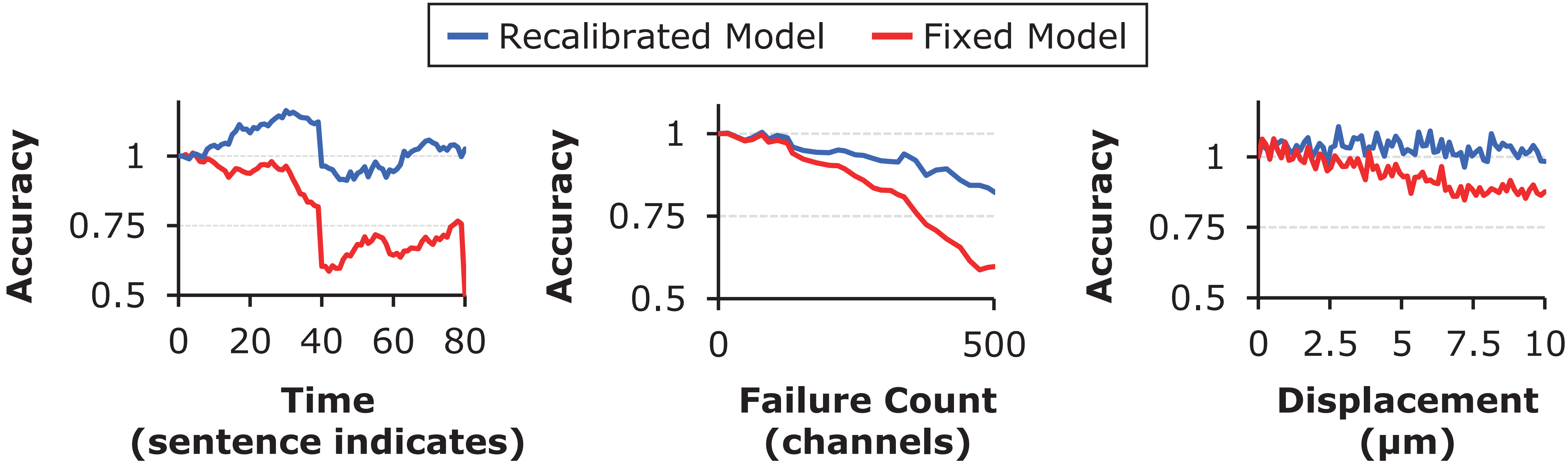
- **The BCI signals change continuously and abruptly** in practical use cases



The processor should adapt to the continuously changing BCI signals

Challenge: Adaptive Processing Support

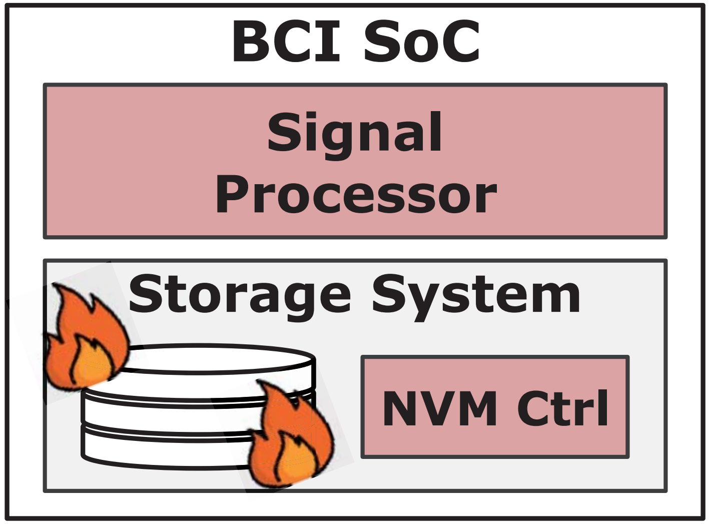
- The system should **continuously update the model parameters** to sustain sufficient accuracy over time

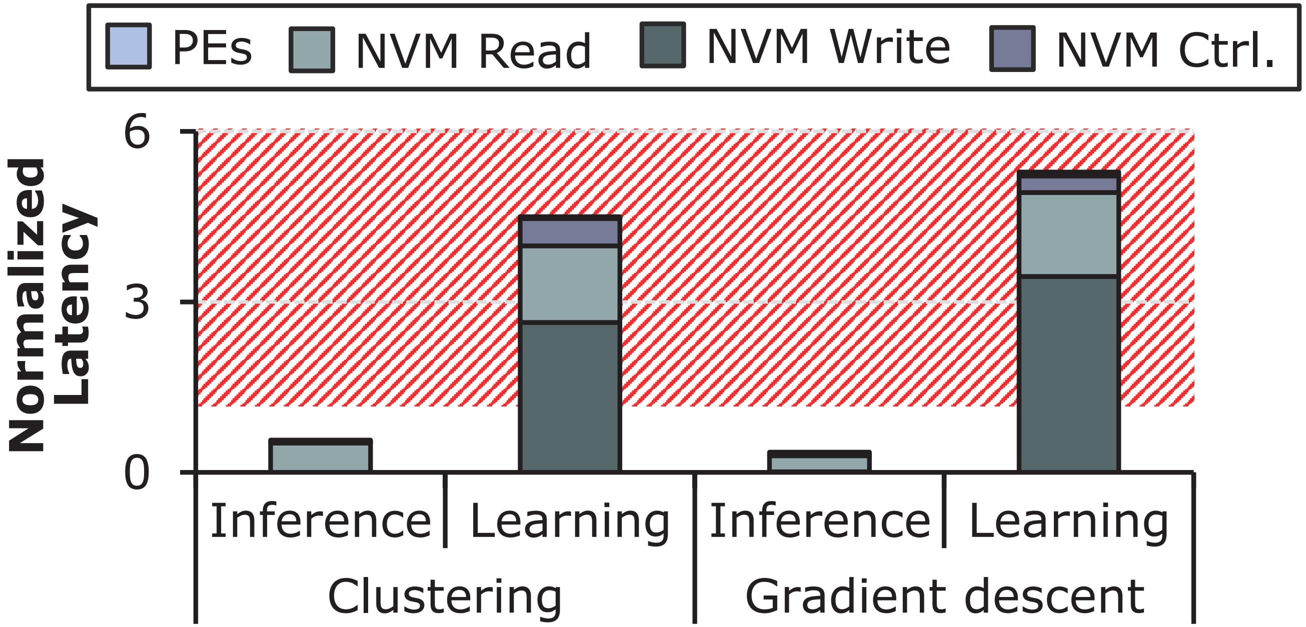


The system demands real-time recalibration to mitigate the accuracy drop

Challenge: Adaptive processing support

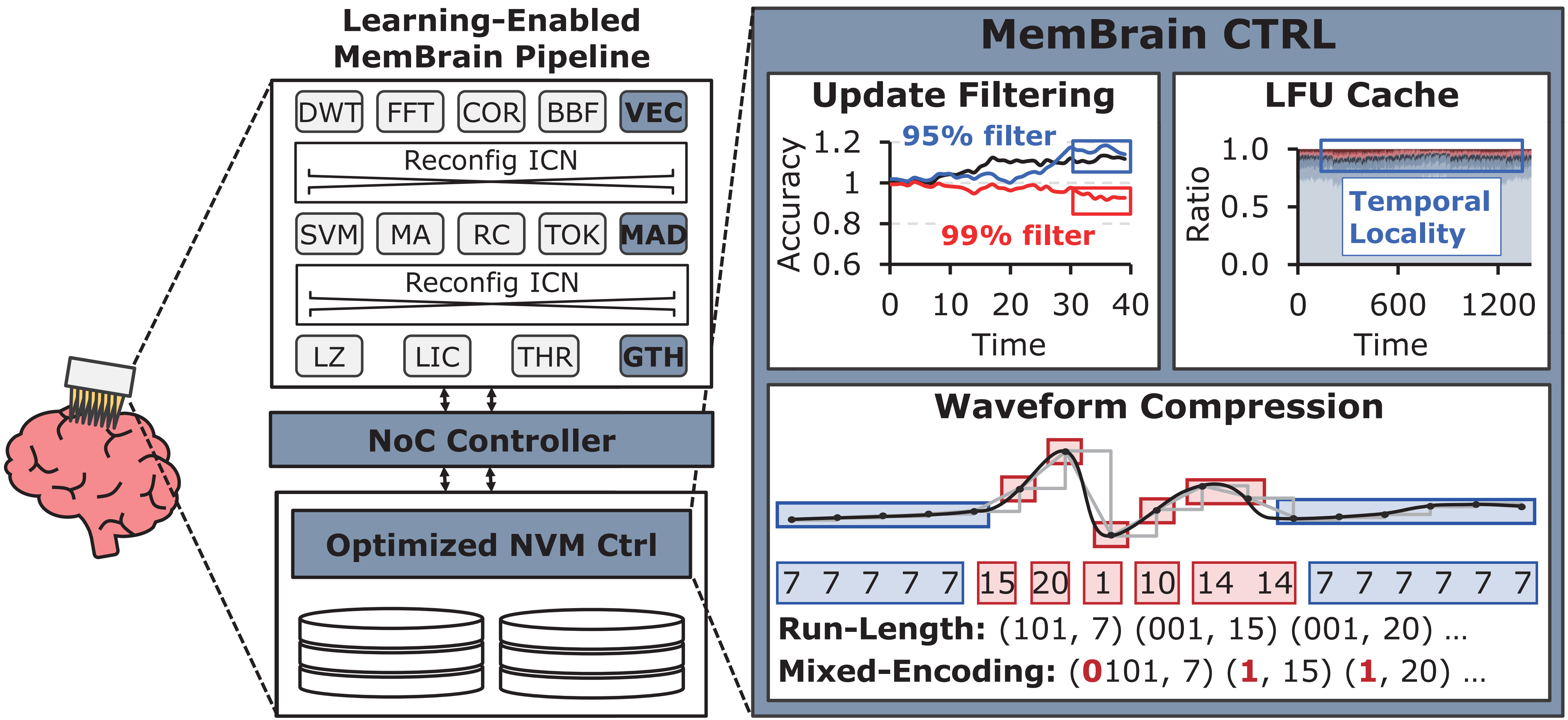
- Continual learning incurs **excessive write operations** to the NVM devices





The NVM Write becomes the major performance overhead

Overview: MemBrain system



What's Next?

- **Sensor: "Spike-driven architecture" for BCI processing**
  [NeuroLobe – MICRO'24]
  – **Rearchitect a neuromorphic-style processor** for the purpose of supporting various BCI algorithms

- **Storage: "Learning-enabled" NVM-assisted BCI system**
  [MemBrain – ISCA'25 (Accepted)]
  – **Propose an NVM-driven acceleration system to** handle BCI processing with learning support

- **Battery & Radio: "Communication and power-aware" BCI scheduling system**
  [Ongoing]
  – **Design a low-cost scheduler** and to handle battery and thermal imbalance among distributed BCI nodes

Thank You!

Any Questions?

Hunjun Lee

Computer Architecture (CArch) Lab.
Department of Computer Science
Hanyang University
ITBT 405-2

E-mail: hunjunlee@hanyang.ac.kr