Cross-Device Collaborative Test-Time Adaptation

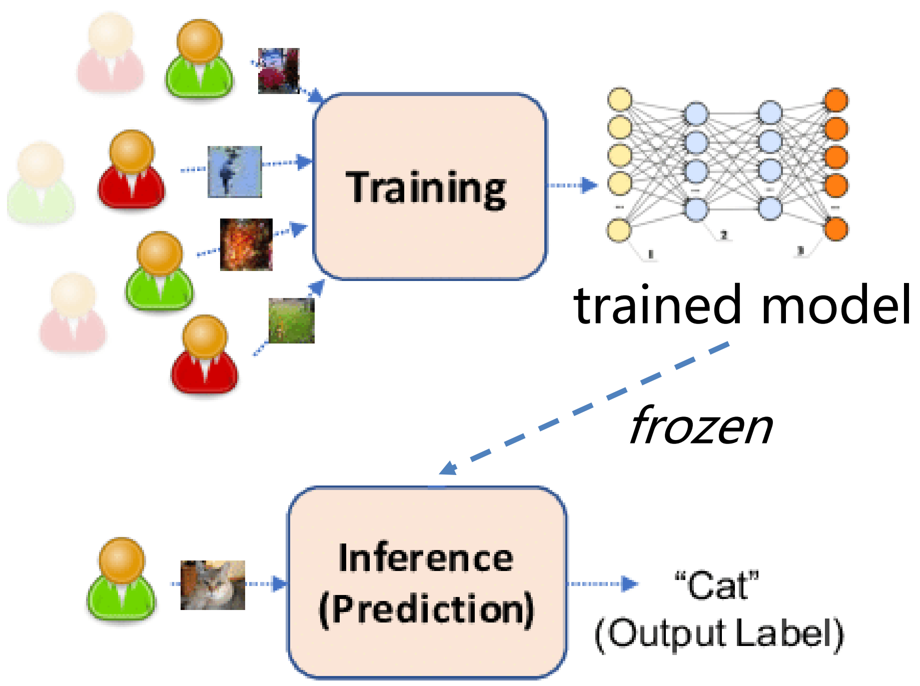
Guohao Chen*, Shuaicheng Niu*, Deyu Chen, Shuhai Zhang, Changsheng Li†, Yuanqing Li, Mingkui Tan†

South China University of Technology, Nanyang Technological University

Pazhou Laboratory, Beijing Institute of Technology

12 December 2024

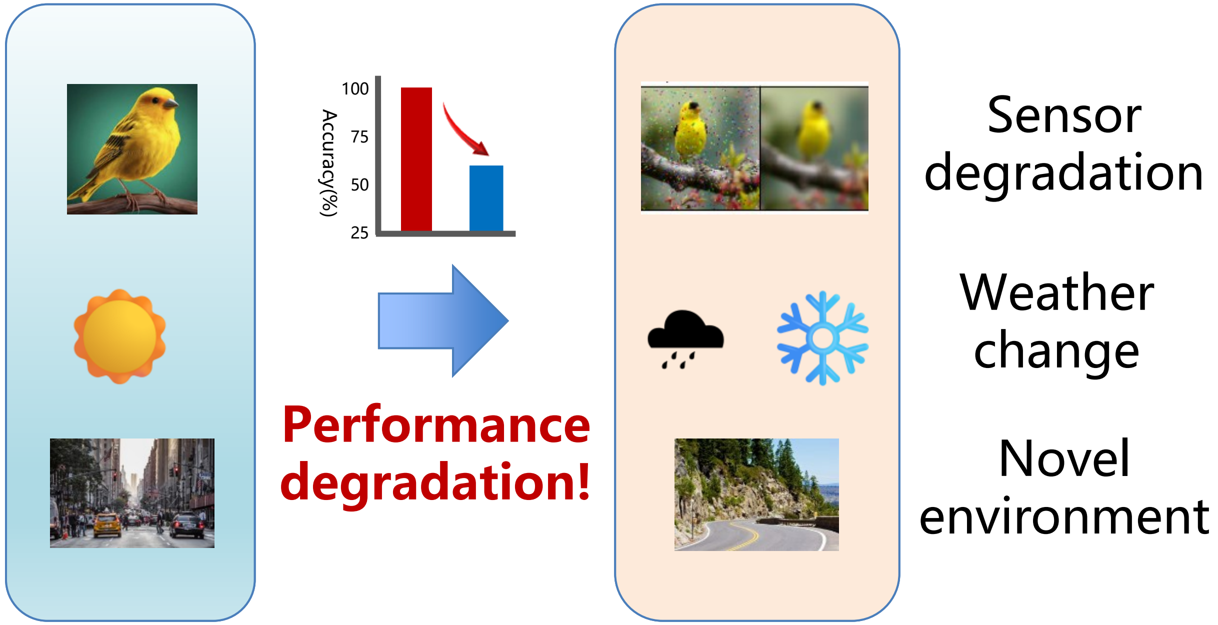
Background: Conventional Deep Learning



Inference with frozen knowledge[1]

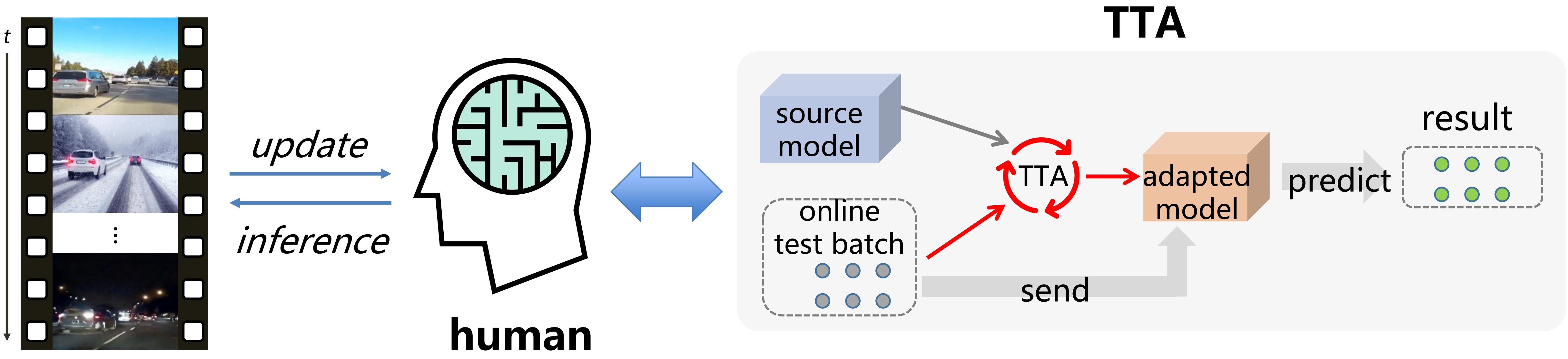
Data Shifts

in-distribution (ID) out-of-distribution (OOD)



[1] Deep Learning on Private Data.

Human Intelligence and Test-Time Adaptation (TTA)

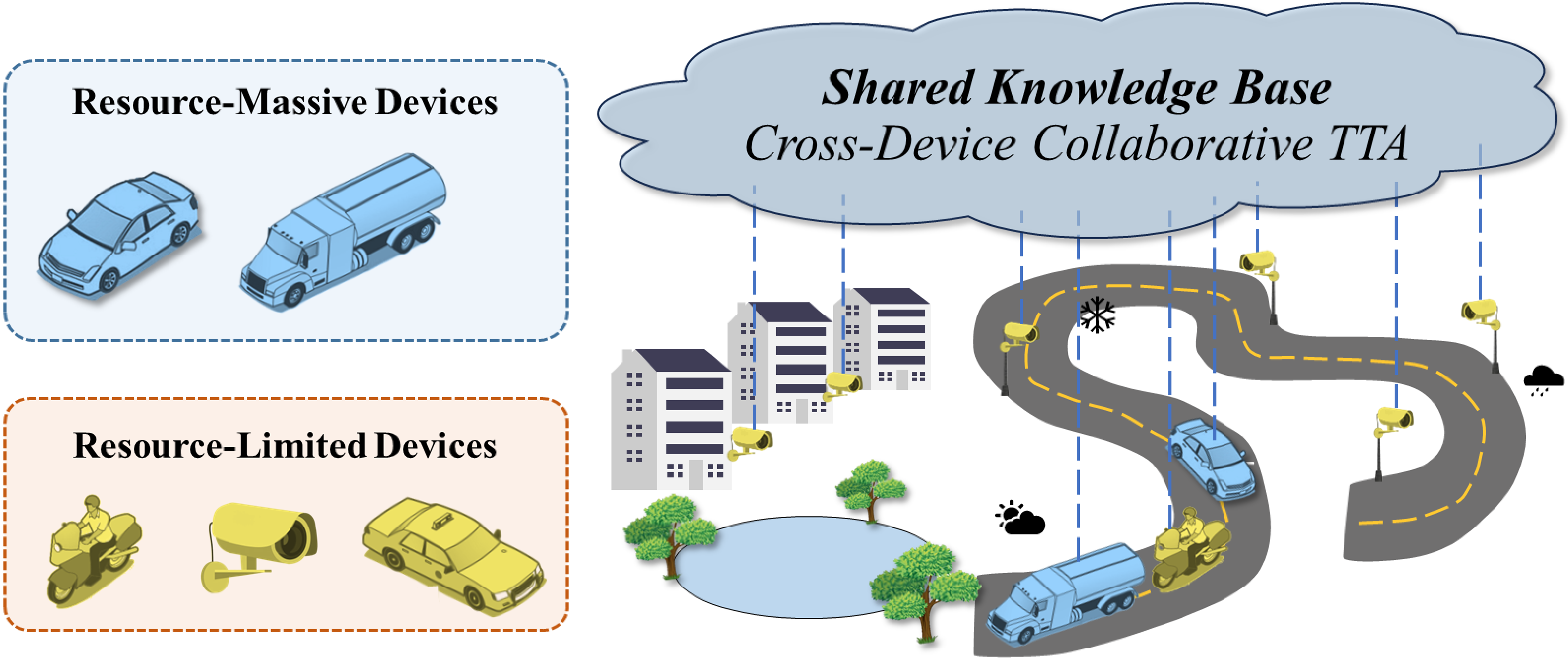


Inference with continuous learning

- TTA **(online) learns** from testing data
- TTA updates model via **self-/un-supervised** objectives before prediction

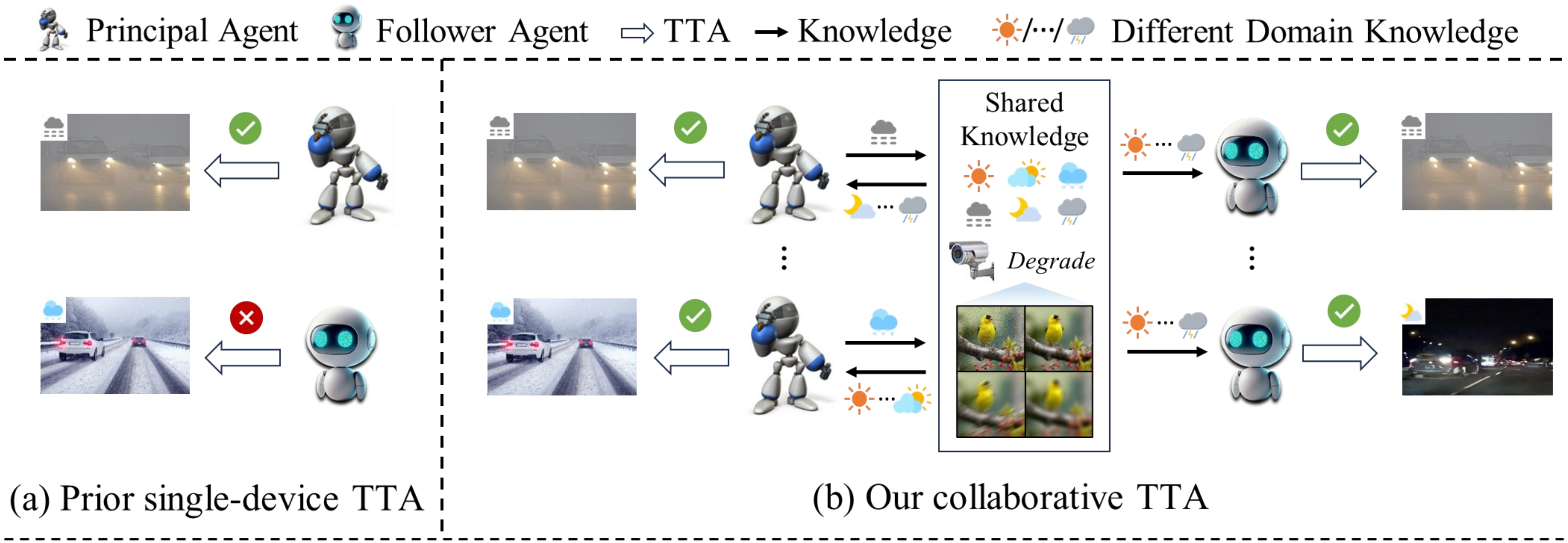
The TTA Figure is borrowed from *Uncovering Adversarial Risks of Test-Time Adaptation.*

Practical Use Case



Multiple devices across multiple constantly changing scenarios

Limitations of Existing TTA



Main limitations when applying TTA to multi-device system:

1. Under continuously changing distributions, they tend to forget what it has learned in long-term scenarios.
2. Adaptation is conducted on each device independently, useful knowledge from other devices is ignored.
3. Rely on backpropagation for model updates, which is infeasible on resource-limited devices.

Technical Challenges to Resolve

Goal: ***enable cross-device and cross-scenario collaborative adaptation so that "one device adapts, all devices benefit."***

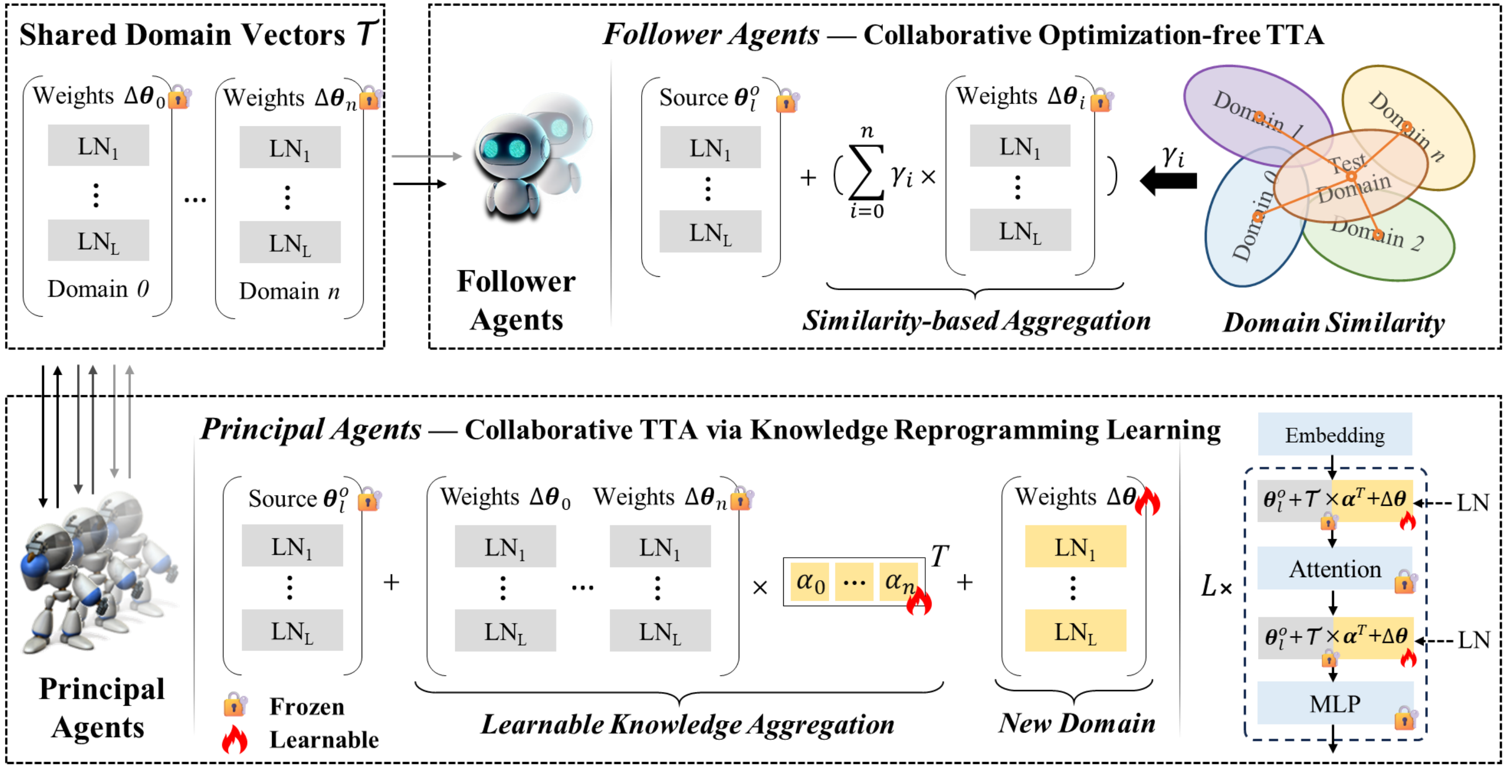
Main Insights:

1. Resource-abundant devices: learn, accumulate, share, and utilize knowledge based on back-propagation.
2. Resource-limited devices: adaptively aggregate the shared knowledge in a forward-only manner.

Key Challenges:

1. How to adapt continuously without forgetting learned knowledge from previously encountered domains?
2. How to facilitate knowledge sharing among devices in a data-free manner for privacy preservation?
3. How to exploit various shared domain knowledge in a backpropagation-based and forward-only manner?

Methods Overview



Training-Free Adaptation

$\boldsymbol{\gamma}$ for knowledge reprogramming

Determine $\boldsymbol{\gamma}$ via domain similarities

Decoupled Optimization

$\boldsymbol{\alpha}$ for knowledge reprogramming

$\Delta\boldsymbol{\theta}$ for new knowledge learning

Methods

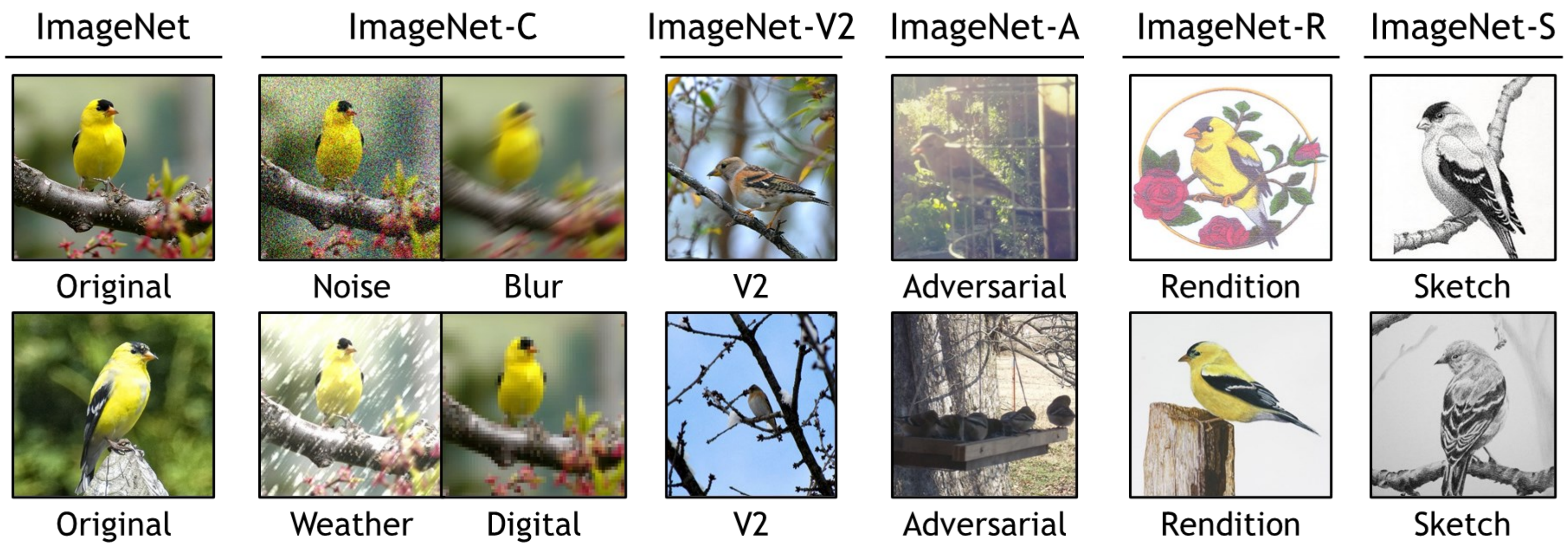
1. **Domain shift detector:** detects domain changes with $D(\phi_d, \hat{\phi}_t) > z$ to save/share learned vectors in $\mathcal{T}$ without forgetting
2. **Knowledge reprogramming TTA:** exploits shared knowledge and learn new knowledge by $\boldsymbol{\theta}_l = \boldsymbol{\theta}_l^o + \sum_{i=0}^{N} \alpha_i \Delta\boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}$
3. **Training-free TTA:** aggregates shared knowledge $\boldsymbol{\theta}_l = \boldsymbol{\theta}_l^o + \sum_{i=0}^{N} \gamma_i \Delta\boldsymbol{\theta}_i$ based on domain similarities with forward only

Experiments

- Main Arguments
  - We learn without forgetting.
  - Collaboration is helpful on BP devices.
  - Collaboration is helpful on FP devices.
  - Our method is computation/memory efficient.

- Main Experiments
  - Lifelong Adaptation to verify accumulation.
  - TTA *vs.* Collaborative TTA on BP devices.
  - TTA *vs.* Collaborative TTA on FP devices.
  - Extra computation/memory of our method.

- Datasets



Experiments

- Main Arguments
  - We learn without forgetting.
  - Collaboration is helpful on BP devices.
  - Collaboration is helpful on FP devices.
  - Our method is computation/memory efficient.

- Main Experiments
  - Lifelong Adaptation to verify accumulation.
  - TTA *vs.* Collaborative TTA on BP devices.
  - TTA *vs.* Collaborative TTA on FP devices.
  - Extra computation/memory of our method.

✓ Lifelong Adaptation to verify knowledge accumulation.

Round	1	2	3	4	5	6	7	8	9	10	Average
NoAdapt	29.9	29.9	29.9	29.9	29.9	29.9	29.9	29.9	29.9	29.9	29.9
CoTTA [9]	44.9	40.6	35.8	32.7	30.4	28.9	27.7	27.1	27.2	26.5	32.2
EATA [3]	60.4	60.0	59.6	59.4	59.3	59.1	59.0	58.8	58.7	58.6	59.3
SAR [4]	59.1	60.6	60.9	61.2	61.3	61.4	58.3	60.4	60.8	61.1	60.5
+ CoLA (Ours)	59.1	62.4	63.6	64.3	64.7	64.9	65.2	65.1	65.3	65.4	$64.0_{(+3.5)}$
ETA [3]	61.4	58.7	54.5	50.2	46.2	44.1	38.8	38.0	36.7	35.1	46.4
+ CoLA (Ours)	62.0	63.9	64.8	65.1	65.3	65.3	65.3	65.3	65.4	65.4	$64.8_{(+18.4)}$
DeYO [2]	59.8	48.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	10.9
+ CoLA (Ours)	61.7	62.5	63.6	64.5	65.0	65.1	65.3	65.5	65.5	65.5	$64.4_{(+53.5)}$

No Degradation!

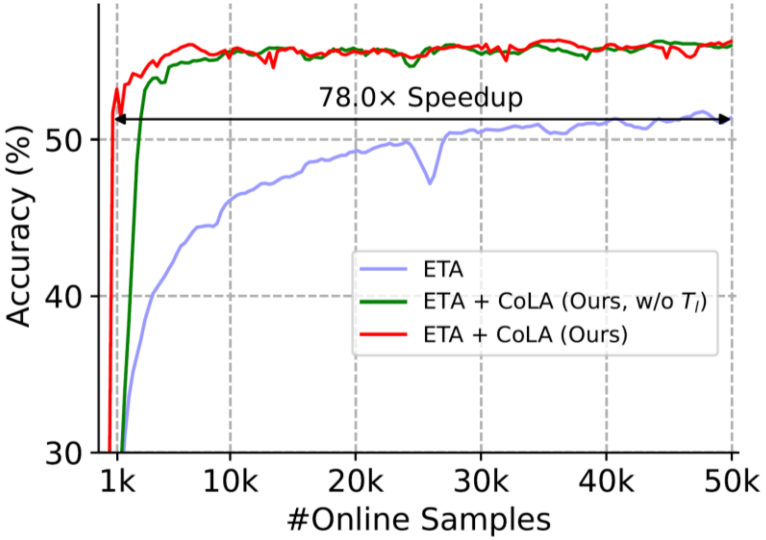
Experiments

- Main Arguments
  - We learn without forgetting.
  - Collaboration is helpful on BP devices.
  - Collaboration is helpful on FP devices.
  - Our method is computation/memory efficient.

- Main Experiments
  - Lifelong Adaptation to verify accumulation.
  - TTA *vs.* Collaborative TTA on BP devices.
  - TTA *vs.* Collaborative TTA on FP devices.
  - Extra computation/memory of our method.

✓ TTA *vs.* Collaborative TTA on BP devices.

	Device 1 (Adapt →)				Device 2 (Adapt →)				Device 3 (Adapt →)				
Method	Noise	Blur	Weat.	Digit.	Blur	Noise	Digit.	Weat.	Weat.	Digit.	Blur	Noise	Avg.
NoAdapt	8.2	28.4	36.1	41.7	28.4	8.2	41.7	36.1	36.1	41.7	28.4	8.2	28.6
CoTTA [29]	28.9	41.3	50.2	47.6	36.3	37.1	50.4	52.7	50.4	55.0	42.5	38.7	44.2
EATA [17]	53.5	57.0	68.1	67.2	58.1	52.2	67.0	68.5	69.4	67.7	57.9	51.5	61.5
SAR [18]	50.4	54.4	66.3	64.5	55.1	48.3	64.0	66.4	66.5	64.3	55.4	47.7	58.6
+ CoLA (Ours)	50.4	58.0	69.4	68.7	55.0	55.0	67.1	70.5	66.3	65.3	58.8	55.5	61.7
ETA [17]	55.2	56.9	67.5	66.0	59.8	51.7	65.0	67.4	70.3	67.8	58.0	49.4	61.2
+ CoLA (Ours)	55.2	60.0	70.9	69.3	59.5	56.3	68.8	70.9	70.2	68.1	59.7	55.3	**63.7**
DeYO [13]	56.3	49.9	68.1	67.8	55.6	46.7	67.2	69.0	71.1	68.8	51.1	4.3	56.3
+ CoLA (Ours)	56.2	55.1	71.2	70.2	54.8	54.5	70.0	71.5	71.0	69.0	53.7	54.3	62.6



Experiments

- Main Arguments
  - We learn without forgetting.
  - Collaboration is helpful on BP devices.
  - Collaboration is helpful on FP devices.
  - Our method is computation/memory efficient.

- Main Experiments
  - Lifelong Adaptation to verify accumulation.
  - TTA *vs.* Collaborative TTA on BP devices.
  - TTA *vs.* Collaborative TTA on FP devices.
  - Extra computation/memory of our method.

✓ TTA vs. Collaborative TTA on FP devices.

	Noise			Blur				Weather				Digital				
Method	Gaus.	Shot	Imp.	Def.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Avg.
NoAdapt	9.5	6.7	8.2	29.0	23.4	33.9	27.1	15.9	26.5	47.2	54.7	44.1	30.5	44.5	47.8	29.9
T3A [12]	9.5	7.0	8.7	23.3	23.3	31.2	25.9	11.9	24.2	44.0	52.2	41.0	30.1	43.0	47.0	28.2
T3A* [12]	9.5	6.5	8.1	29.8	24.1	34.3	28.2	16.0	26.9	49.0	55.5	44.5	33.1	44.5	48.2	30.5
LAME [2]	9.3	6.5	8.0	28.6	23.0	33.3	26.6	15.2	26.0	45.9	54.1	43.6	29.3	44.0	47.4	29.4
CoLA (SAR)	55.2	56.0	56.8	57.3	49.1	59.9	58.5	65.8	65.8	72.2	77.1	66.2	65.9	72.2	69.4	63.2
CoLA (ETA)	55.7	57.3	56.9	58.5	46.2	59.4	63.4	69.1	66.5	73.1	77.6	66.3	69.2	73.1	69.9	**64.1**
CoLA (DeYO)	56.6	57.7	57.5	58.2	47.7	55.5	39.0	69.6	67.2	73.5	78.0	67.0	70.4	73.5	70.3	62.8

Experiments

- Main Arguments
  - We learn without forgetting.
  - Collaboration is helpful on BP devices.
  - Collaboration is helpful on FP devices.
  - Our method is computation/memory efficient.

- Main Experiments
  - Lifelong Adaptation to verify accumulation.
  - TTA *vs.* Collaborative TTA on BP devices.
  - TTA *vs.* Collaborative TTA on FP devices.
  - Extra computation/memory of our method.

✓ Extra computation/memory of our method.

Method	BP	Acc.	Time (s)	Mem. (MB)
NoAdapt	✗	9.5	50	816.6
T3A [12]	✗	9.5	158	909.9
CoLA (Eqn. 3)	✗	55.7	51	821.9
EATA [17]	✔	49.5	113	7439.3
SAR [18]	✔	44.0	202	7429.9
ETA [17]	✔	51.9	109	7429.6
+ CoLA (Eqn. 2)	✔	54.3	112	7435.3

Experiments

- Other Discussions
  - Effectiveness of CoLA on single-domain TTA
  - Robustness of CoLA against potential harmful knowledge.
  - Scalability of CoLA with more collaborative devices.
  - Efficiency of CoLA with increasing domain vectors.
  - Prompt tuning with CoLA using multiple hard prompts.
  - Advantages of CoLA over FedAvg in collaborative TTA.
  - Robustness of CoLA under small batch sizes.
  - Effectiveness of CoLA in mitigating error accumulation.
  - …

Thank you!