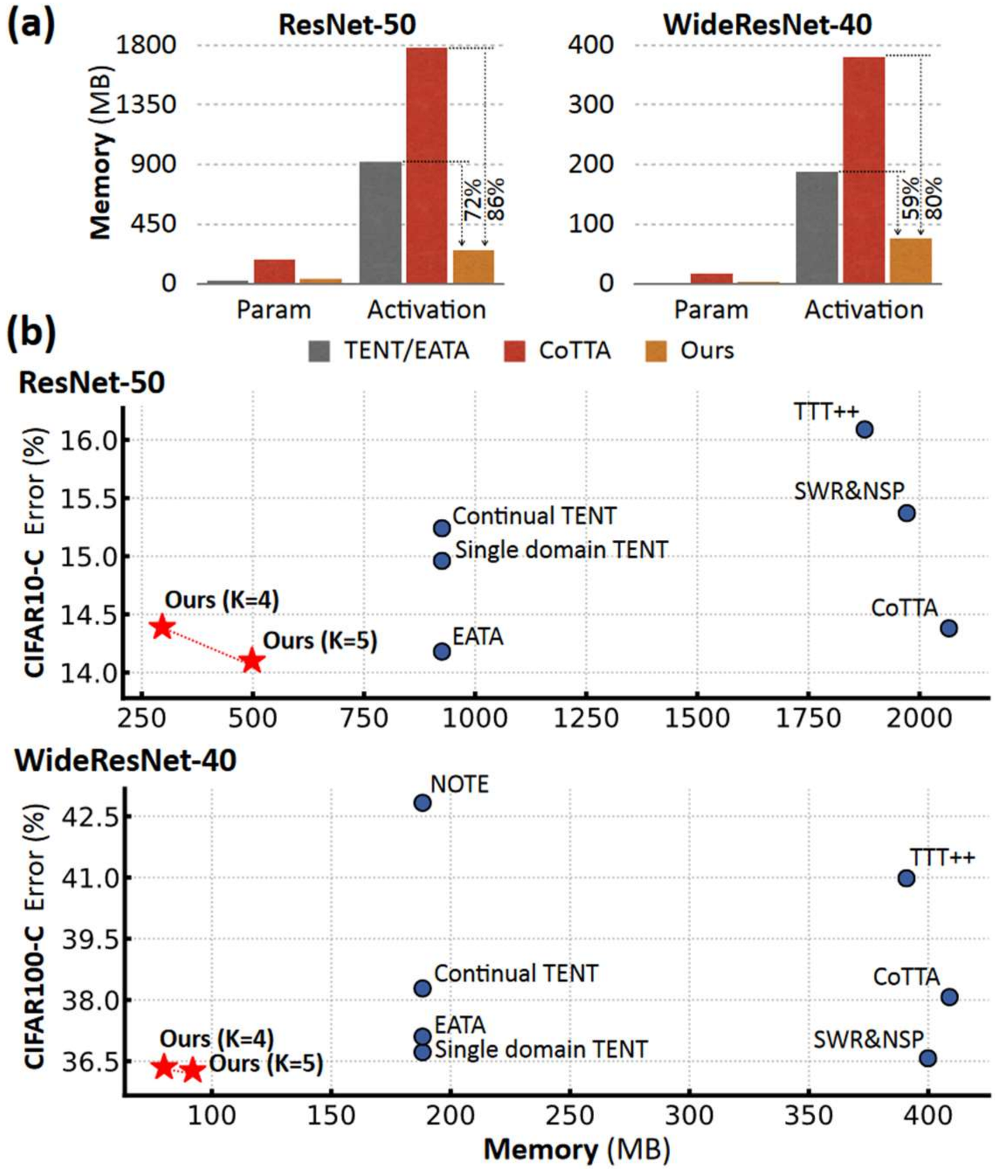
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

EcoTTA: Memory-Efficient Continual Test-time Adaptation via Self-distilled Regularization

Junha Song[1,2*], Jungsoo Lee[1], In So Kweon[2], Sungha Choi[1†]
[1]Qualcomm AI Research[‡], [2]KAIST

CVPR 2023

Motivation



Reducing memory cost
in Test Time Adaptation

Methods

- **Architecture Comparison**

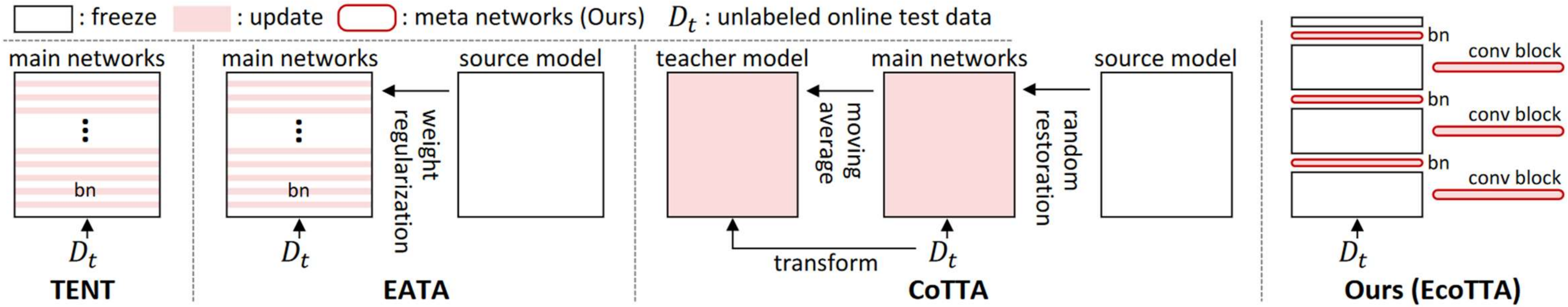
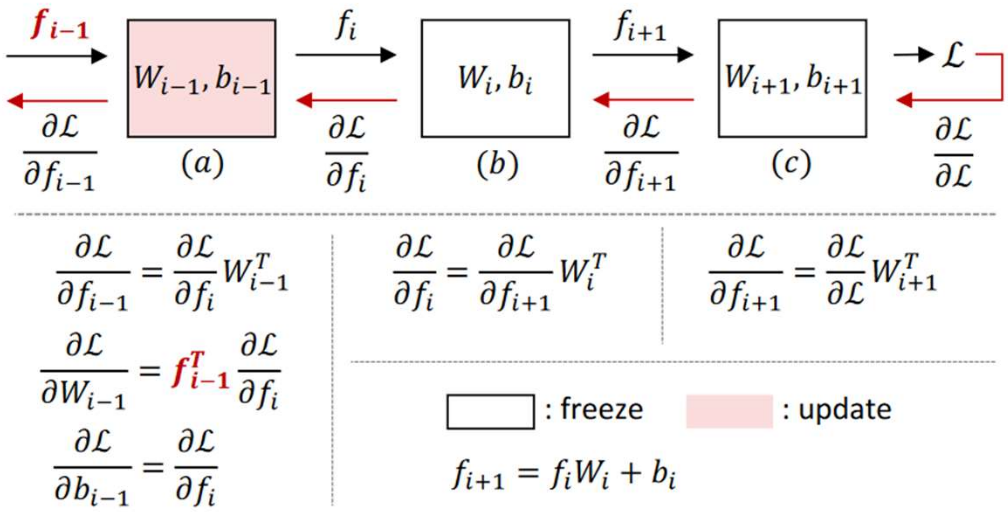


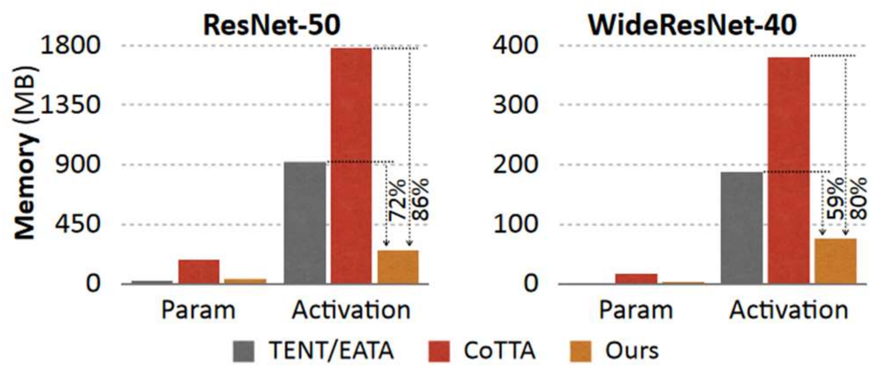
Figure 2. **Architecture for test-time adaptation.** We illustrate TTA methods: TENT [65], EATA [50], CoTTA [66], and Ours (EcoTTA). TENT and EATA update *multiple* batch norm layers, in which large activations have to be stored for gradient calculation. In CoTTA, an entire network is trained with additional strategies for continual adaptation that requires a significant amount of both memory and time. In contrast, our approach requires a minimum size of activations by updating only *a few* layers. Also, stable long-term adaptation is performed by our proposed regularization, named self-distilled regularization.

Methods

Given a linear layer $f_{i+1} = f_i \mathcal{W} + b$

The gradient can be calculated as $$\frac{\partial \mathcal{L}}{\partial f_i} = \frac{\partial \mathcal{L}}{\partial f_{i+1}} \mathcal{W}^T, \quad \frac{\partial \mathcal{L}}{\partial \mathcal{W}} = f_i^T \frac{\partial \mathcal{L}}{\partial f_{i+1}}.$$





Methods

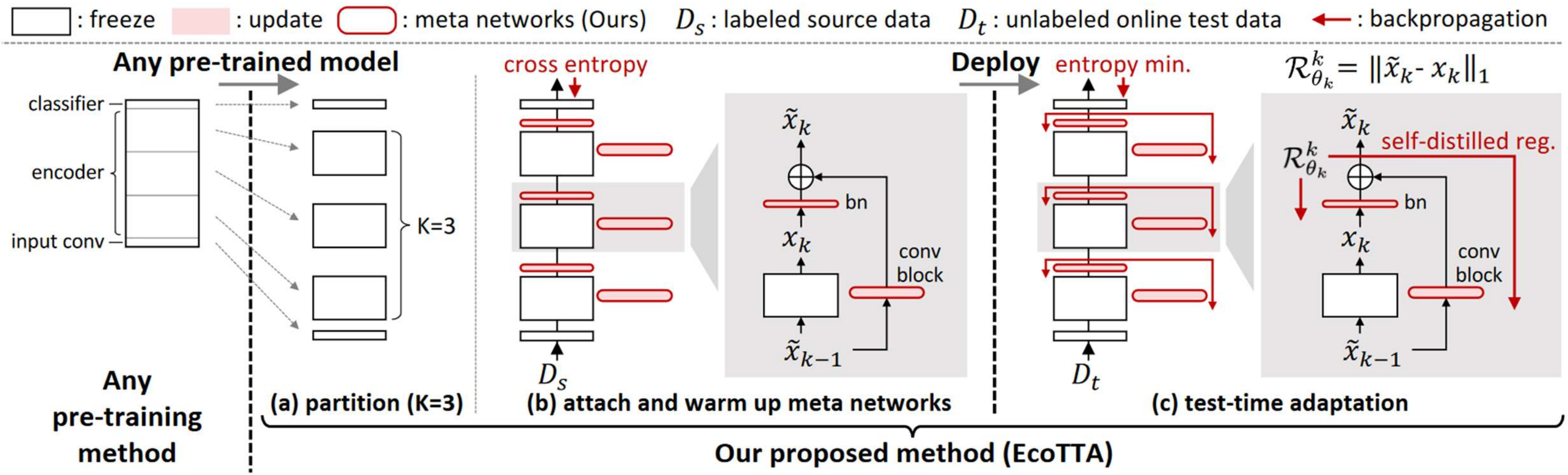


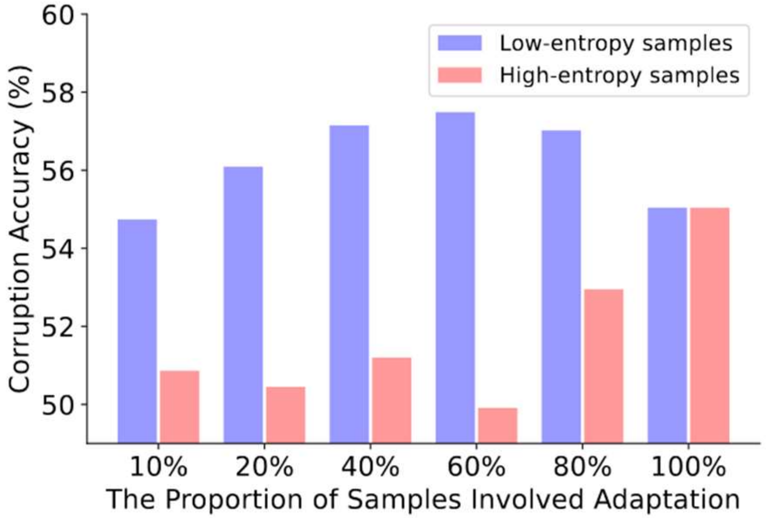
Figure 3. **Overview of our approach. (a)** The encoder of the pre-trained model is divided into K parts (*i.e.,* model partition factor K). **(b)** Before deployment, the meta networks are attached to each part of the original networks and pre-trained with source dataset $\mathcal{D}_s$. **(c)** After the model is deployed, *only* the meta networks are updated with unsupervised loss (*i.e.*, entropy minimization) on target data $\mathcal{D}_t$, while the original networks are frozen. To avoid error accumulation and catastrophic forgetting by the long-term adaptation, we regularize the output $\tilde{x}_k$ of each group of the meta networks leveraging the output x_k of the *frozen* original network, which preserves the source knowledge.

Methods

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

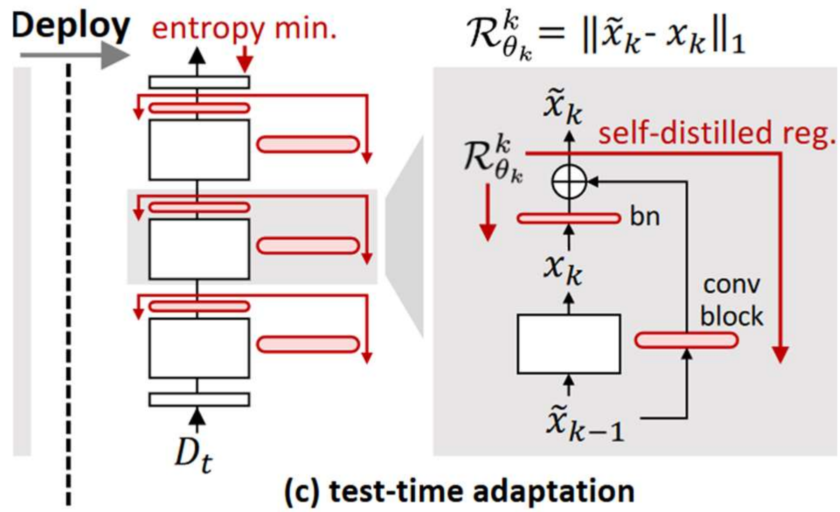
Selected Entropy Minimization

$$\mathcal{L}^{ent} = \mathbb{I}_{\{H(\hat{y})<H_0\}} \cdot H(\hat{y}),$$



Self-distilled Regularization

$$\mathcal{R}^k_{\theta_k} = \|\tilde{x}_k - x_k\|_1 \, .$$



$$\mathcal{L}^{total}_{\theta} = \mathcal{L}^{ent}_{\theta} + \lambda \sum_{k}^{K} \mathcal{R}^k_{\theta_k},$$

Experiments

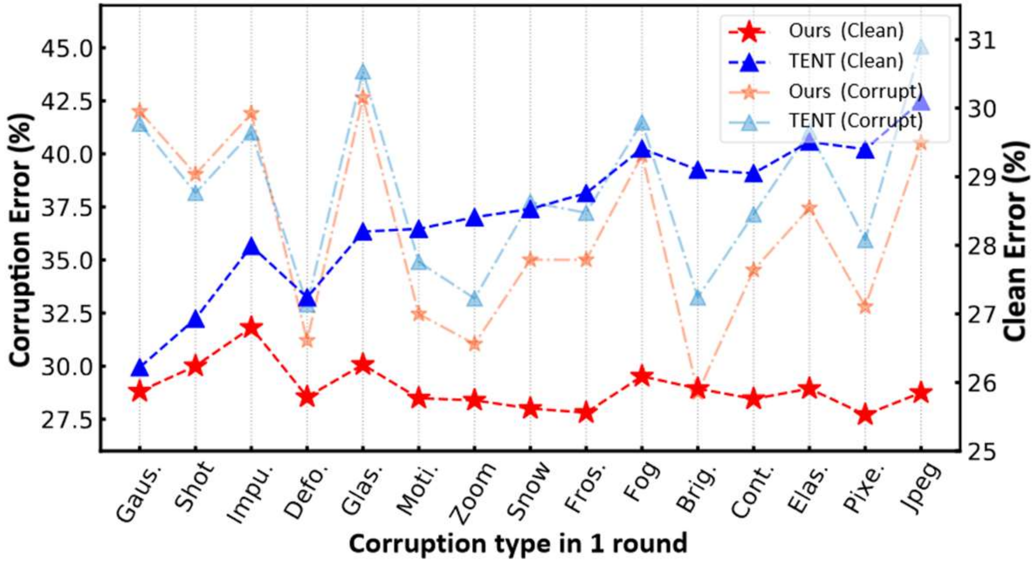
Method	WideResNet-40 (AugMix) Avg. err ↓	WideResNet-40 (AugMix) Mem. (MB)	WideResNet-28 Avg. err ↓	WideResNet-28 Mem. (MB)	ResNet-50 Avg. err ↓	ResNet-50 Mem. (MB)
Source	36.7	11	43.5	58	48.8	91
BN Stats Adapt [49]	15.4	11	20.9	58	16.6	91
Single do. TENT [65]	12.7	188	19.2	646	15.0	925
Continual TENT	13.3	188	20.0	646	15.2	925
TTT++ [42]	14.6	391	20.3	1405	16.1	1877
SWR&NSP [9]	12.1	400	17.2	1551	15.4	1971
NOTE [17]	13.4	188	20.2	646	-	-
EATA [50]	13.0	188	18.6	646	14.2	925
CoTTA [66]	14.0	409	17.0	1697	14.4	2066
Ours (K=4)	12.2	80 (80, 58%↓)	16.9	404 (76, 38%↓)	14.4	296 (86, 68%↓)
Ours (K=5)	**12.1**	92 (77, 51%↓)	**16.8**	471 (72, 27%↓)	**14.1**	498 (76, 46%↓)

(a) **CIFAR10-C with severity level 5**

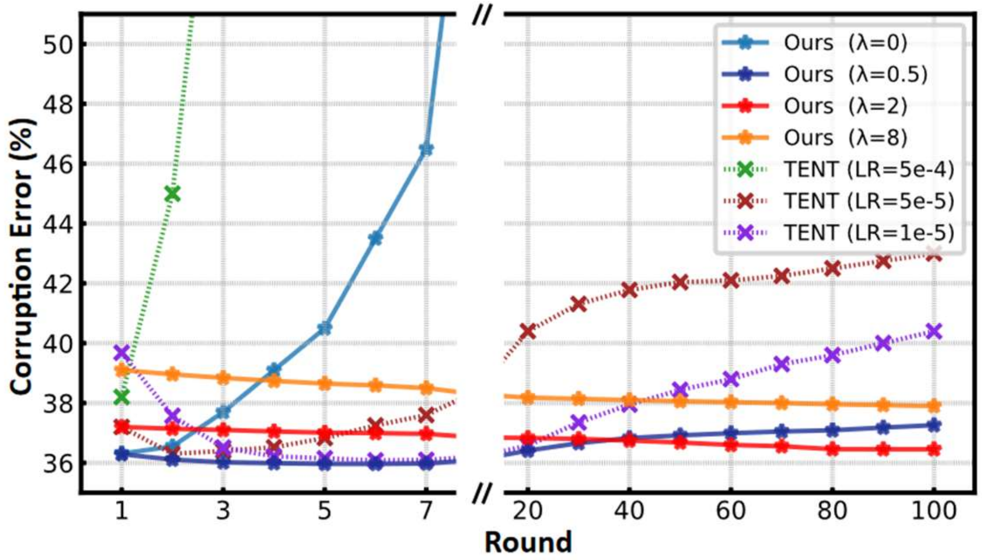
Method	WideResNet-40 (AugMix) Avg. err ↓	WideResNet-40 (AugMix) Mem. (MB)	ResNet-50 Avg. err ↓	ResNet-50 Mem. (MB)
Source	69.7	11	73.8	91
BN Stats Adapt [49]	41.1	11	44.5	91
Single do. TENT [65]	36.7	188	40.1	926
Continual TENT	38.3	188	45.9	926
TTT++ [42]	41.0	391	44.2	1876
SWR&NSP [9]	36.6	400	44.1	1970
NOTE [17]	42.8	188	-	-
EATA [50]	37.1	188	39.9	926
CoTTA [66]	38.1	409	40.2	2064
Ours (K=4)	36.4	80 (80, 58%↓)	39.5	296 (86, 68%↓)
Ours (K=5)	**36.3**	92 (77, 51%↓)	**39.3**	498 (76, 46%↓)

(b) **CIFAR100-C with severity level 5**

Experiments

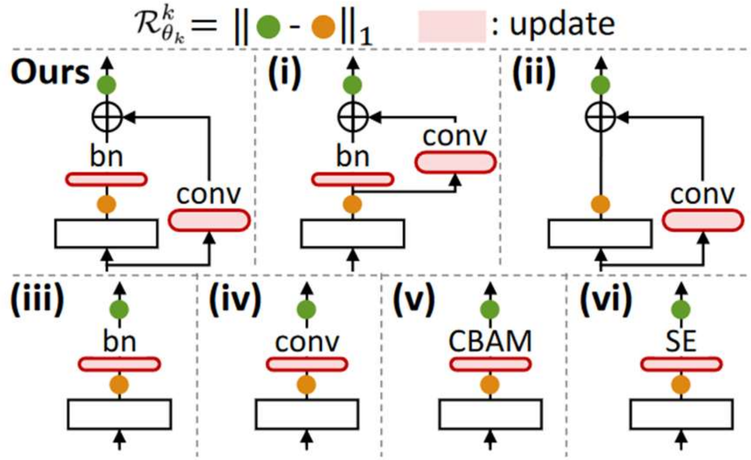


(a) **Catastrophic forgetting effect**



(b) **Error accumulation effect**

Experiments



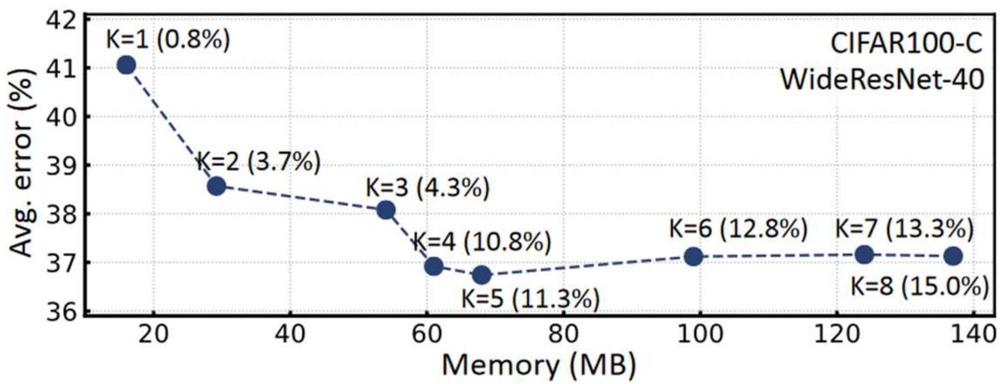
(a) **Visualization of networks variants**

Avr. err	CIFAR10-C	CIFAR10-C	CIFAR100-C
Arch	**WRN-28**	**WRN-40**	**WRN-40**
(i)	<u>18.1</u>	<u>12.6</u>	<u>37.2</u>
(ii) Ours w\o BN	18.7	13.7	38.2
(iii) Ours w\o Conv	20.7	14.9	40.1
(iv) Conv	60.6	73.3	77.2
(v) CBAM [67]	21.4	15.1	40.9
(vi) SE [30]	22.3	16.2	40.5
Ours	**16.8**	**12.1**	**36.3**

(b) **Meta network design (K=5)**

Model	#Block	Avg. err
WRN-28 (12) CIFAR10-C	3,3,3,3	<u>17.3</u>
	4,4,2,2	17.9
	2,2,4,4	**16.9**
WRN-40 (18) CIFAR10-C	4,4,5,5	<u>12.8</u>
	6,6,3,3	13.7
	3,3,6,6	**12.2**
WRN-40 (18) CIFAR100-C	4,4,5,5	<u>36.9</u>
	6,6,3,3	38.5
	3,3,6,6	**36.4**

(c) **# of blocks of each partition (K=4)**



Thanks