ICLR

New Insights for the Stability-Plasticity Dilemma in Online Continual Learning

Dahuin Jung, Dongjin Lee, Sunwon Hong, Hyemi Jang, Ho Bae*, Sungroh Yoon*

SEOUL
NATIONAL
UNIVERSITY

Motivation

- Stability-Plasticity Dilemma in Online Continual Learning
  - Stability: Retention of previous knowledge ⇔ Plasticity: Ability to learn new knowledge
    - *The plasticity of online continual learning* is **more vulnerable** than offline continual learning

    → because the training signal that can be obtained from streaming data is **limited**

Motivation

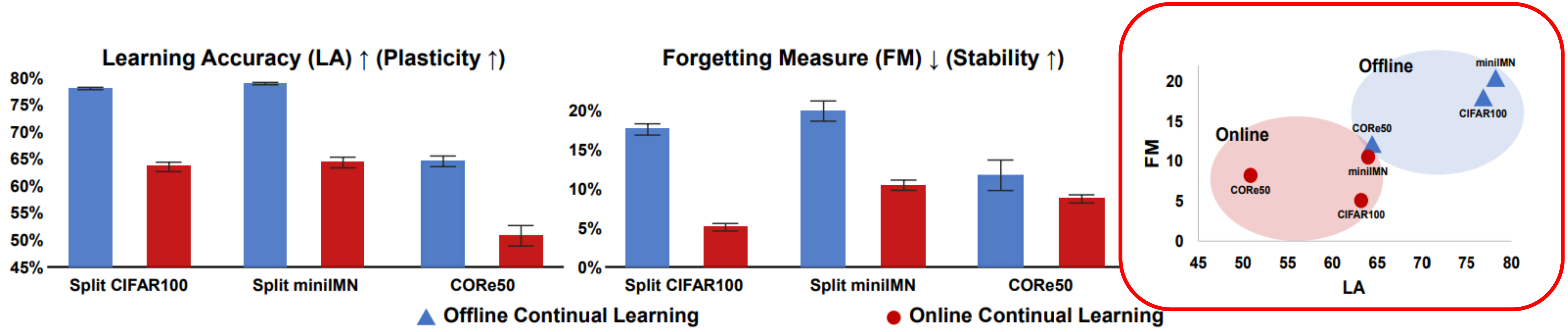
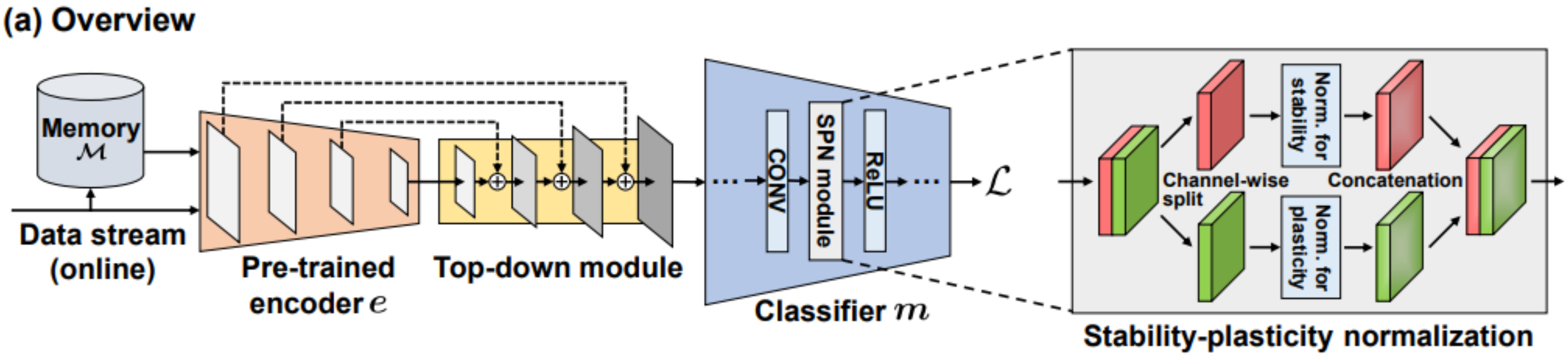


Figure 1: Comparison results of ER-Ring on three CL benchmarks in offline and online CL in bar (left and middle) and scatter (right) plots. For offline CL, plasticity is relatively high, whereas stability is low. In contrast, for online CL, stability is relatively high, whereas plasticity is low. It shows the difference in trend between offline and online CL in terms of the stability-plasticity dilemma. Further

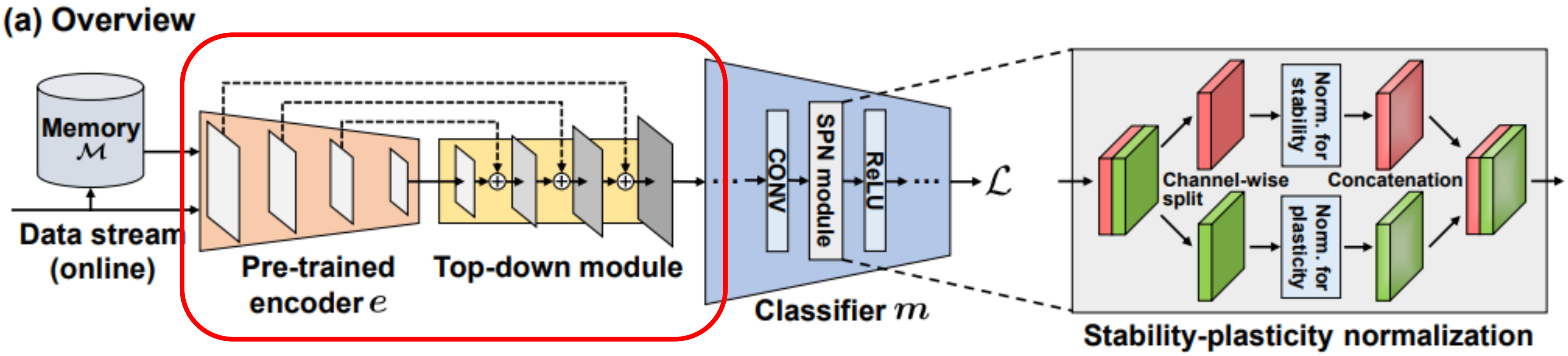
Method

- We propose a multi-scale feature adaptation network (**MuFAN**) for online CL, which consists of three components to obtain both high stability and plasticity.
  1. **As input:** multi-scale feature maps exploited from a pre-trained model
  2. **As loss:** novel structure-wise distillation loss across tasks
  3. **As an architecture:** novel stability-plasticity normalization module



$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{ER}} + \lambda_{\text{D-CTN}}\mathcal{L}_{\text{D-CTN}} + \lambda_{\text{D-CSD}}\mathcal{L}_{\text{D-CSD}}$$

Method I. Multi-scale Feature Maps As Input

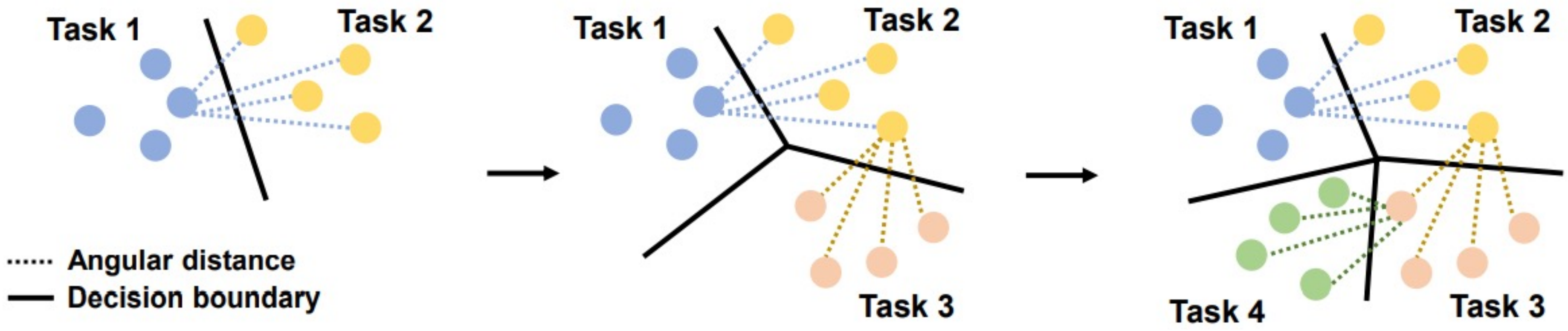


- Rather than leveraging a raw RGB image, we accelerate classifier training by leveraging an aggregated feature map from the meaningful spaces of the pre-trained encoder.
  - Why an aggregated multi-scale feature map from a pre-trained model?
    - The low-level features extracted by shallow layers encode more pattern-wise and general information, whereas the high-level features extracted by deeper layers contain more contextual information that focuses on prominent features.

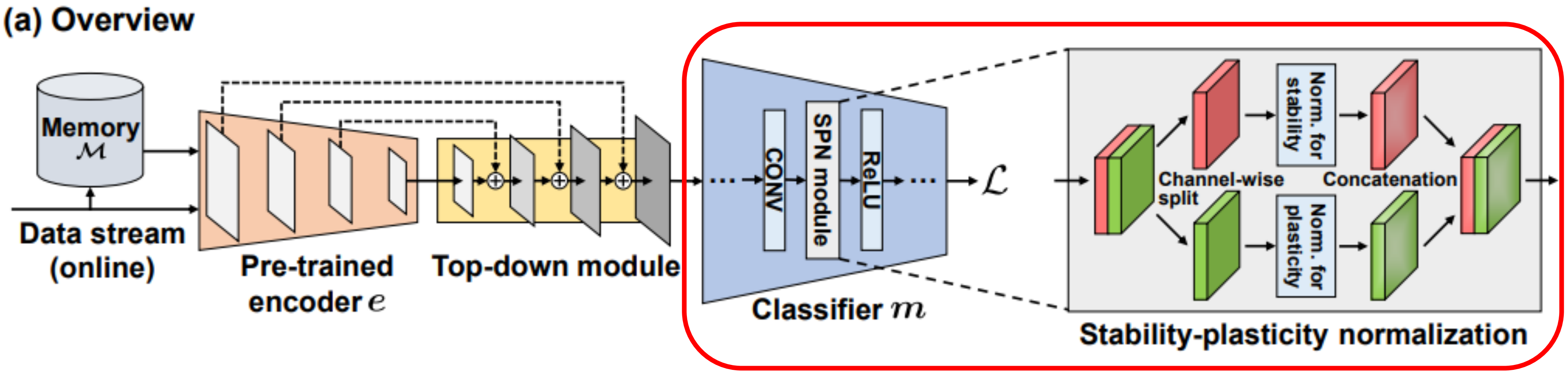
Method II. Structure-wise Distillation Loss ($\mathcal{L}_{\mathrm{D-CSD}}$)

- Most distillation losses in CL have been point-wise; for each data point, they regularized the change by distilling individual outputs.
  - We present **a structure-wise distillation loss $\mathcal{L}_{\mathrm{D-CSD}}$** for CL.
    - We compute a structure-wise potential ψ for a tuple of data samples across tasks from the replay buffer and distill relations through the potential.

(b) Cross-task Structure-wise Distillation



Method III. Stability-Plasticity Norm Module



- We propose a new stability-plasticity normalization (SPN) module that sets one normalization operation efficient for stability and another normalization operation efficient for plasticity *in a parallel manner*.
  - A SPN module splits the feature map $\boldsymbol{a}$ of the classifier m into halves along the channel dimension and applies a different normalization operation to **each half feature map**.

Experimental Setup

- Benchmarks:
  - Split SVHN: 5 tasks
  - Split CIFAR100: 20 tasks
  - Split miniImageNet: 20 tasks
  - CORe50: 10 tasks
- CL scenarios:
  - Online task-incremental scenario
  - Online task-free scenario
- Architecture:
  - **As the pre-trained encoder**, we use an ImageNet-pretrained EfficientNet-lite0 or a COCO-pretrained SSDlite, which uses MobileNetV3 as a backbone, considering model complexity.

Experimental Results

MuFAN achieves the state-of-the-art performance

- Online task-incremental scenario

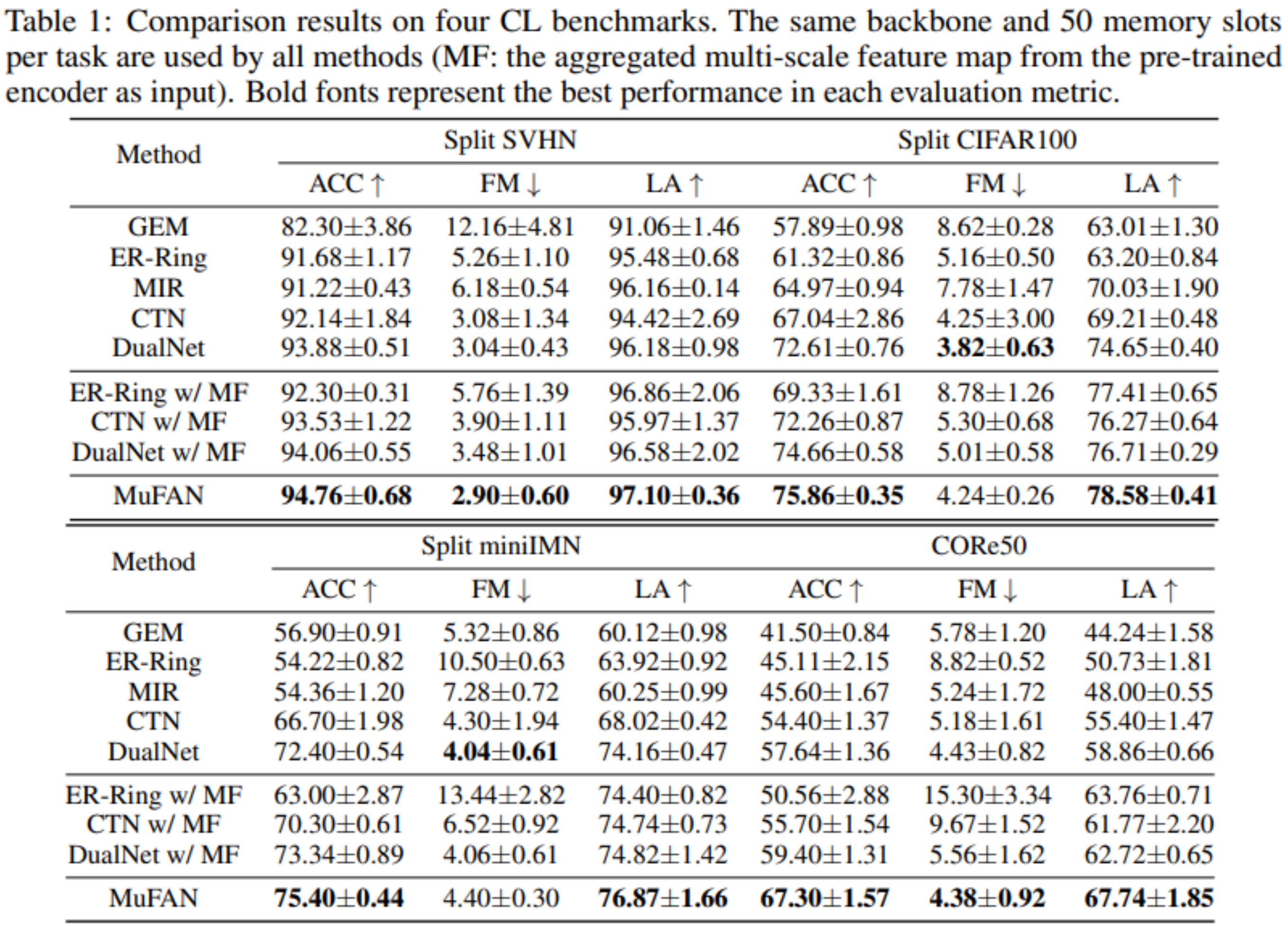


Table 1: Comparison results on four CL benchmarks. The same backbone and 50 memory slots per task are used by all methods (MF: the aggregated multi-scale feature map from the pre-trained encoder as input). Bold fonts represent the best performance in each evaluation metric.

Method	Split SVHN			Split CIFAR100		
	ACC ↑	FM ↓	LA ↑	ACC ↑	FM ↓	LA ↑
GEM	82.30±3.86	12.16±4.81	91.06±1.46	57.89±0.98	8.62±0.28	63.01±1.30
ER-Ring	91.68±1.17	5.26±1.10	95.48±0.68	61.32±0.86	5.16±0.50	63.20±0.84
MIR	91.22±0.43	6.18±0.54	96.16±0.14	64.97±0.94	7.78±1.47	70.03±1.90
CTN	92.14±1.84	3.08±1.34	94.42±2.69	67.04±2.86	4.25±3.00	69.21±0.48
DualNet	93.88±0.51	3.04±0.43	96.18±0.98	72.61±0.76	**3.82±0.63**	74.65±0.40
ER-Ring w/ MF	92.30±0.31	5.76±1.39	96.86±2.06	69.33±1.61	8.78±1.26	77.41±0.65
CTN w/ MF	93.53±1.22	3.90±1.11	95.97±1.37	72.26±0.87	5.30±0.68	76.27±0.64
DualNet w/ MF	94.06±0.55	3.48±1.01	96.58±2.02	74.66±0.58	5.01±0.58	76.71±0.29
MuFAN	**94.76±0.68**	**2.90±0.60**	**97.10±0.36**	**75.86±0.35**	4.24±0.26	**78.58±0.41**

Method	Split miniIMN			CORe50		
	ACC ↑	FM ↓	LA ↑	ACC ↑	FM ↓	LA ↑
GEM	56.90±0.91	5.32±0.86	60.12±0.98	41.50±0.84	5.78±1.20	44.24±1.58
ER-Ring	54.22±0.82	10.50±0.63	63.92±0.92	45.11±2.15	8.82±0.52	50.73±1.81
MIR	54.36±1.20	7.28±0.72	60.25±0.99	45.60±1.67	5.24±1.72	48.00±0.55
CTN	66.70±1.98	4.30±1.94	68.02±0.42	54.40±1.37	5.18±1.61	55.40±1.47
DualNet	72.40±0.54	**4.04±0.61**	74.16±0.47	57.64±1.36	4.43±0.82	58.86±0.66
ER-Ring w/ MF	63.00±2.87	13.44±2.82	74.40±0.82	50.56±2.88	15.30±3.34	63.76±0.71
CTN w/ MF	70.30±0.61	6.52±0.92	74.74±0.73	55.70±1.54	9.67±1.52	61.77±2.20
DualNet w/ MF	73.34±0.89	4.06±0.61	74.82±1.42	59.40±1.31	5.56±1.62	62.72±0.65
MuFAN	**75.40±0.44**	4.40±0.30	**76.87±1.66**	**67.30±1.57**	**4.38±0.92**	**67.74±1.85**

- Online task-free scenario

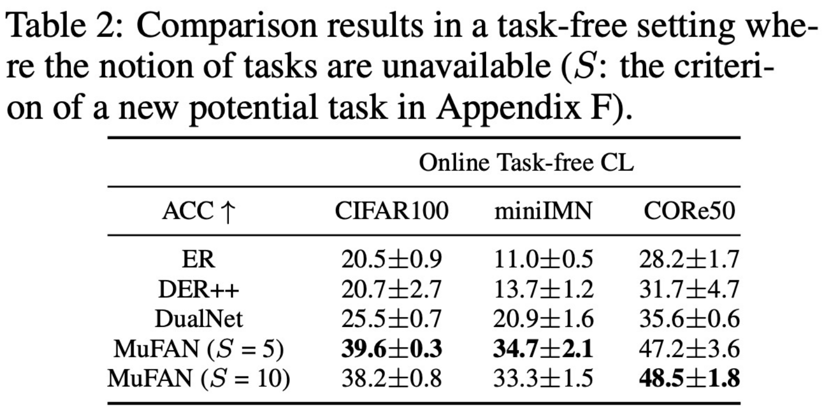


Table 2: Comparison results in a task-free setting where the notion of tasks are unavailable (S: the criterion of a new potential task in Appendix F).

	Online Task-free CL		
ACC ↑	CIFAR100	miniIMN	CORe50
ER	20.5±0.9	11.0±0.5	28.2±1.7
DER++	20.7±2.7	13.7±1.2	31.7±4.7
DualNet	25.5±0.7	20.9±1.6	35.6±0.6
MuFAN ($S = 5$)	**39.6±0.3**	**34.7±2.1**	47.2±3.6
MuFAN ($S = 10$)	38.2±0.8	33.3±1.5	**48.5±1.8**

Discussion

- Ablation study on each proposed component

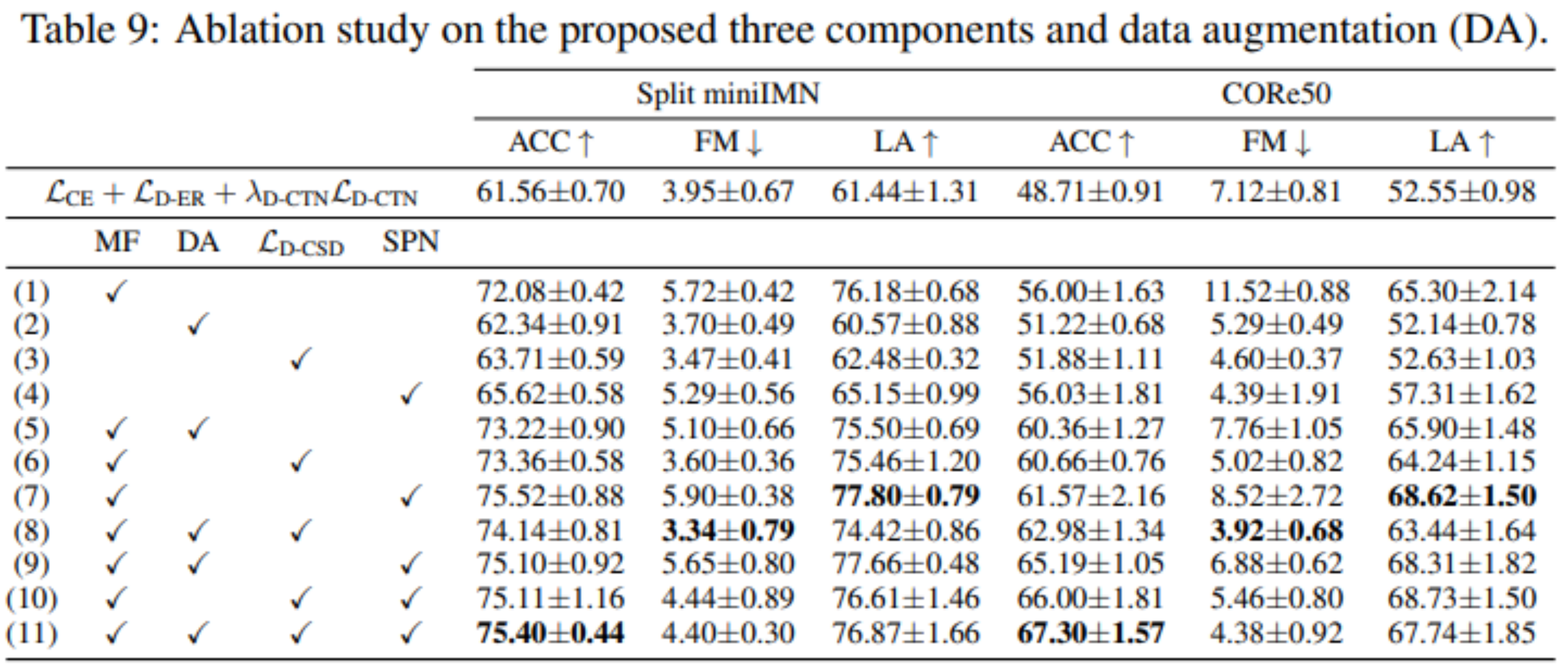


Table 9: Ablation study on the proposed three components and data augmentation (DA).

					Split miniIMN			CORe50		
					ACC ↑	FM ↓	LA ↑	ACC ↑	FM ↓	LA ↑
$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{D-ER}} + \lambda_{\text{D-CTN}}\mathcal{L}_{\text{D-CTN}}$					61.56±0.70	3.95±0.67	61.44±1.31	48.71±0.91	7.12±0.81	52.55±0.98
	MF	DA	$\mathcal{L}_{\text{D-CSD}}$	SPN						
(1)	✓				72.08±0.42	5.72±0.42	76.18±0.68	56.00±1.63	11.52±0.88	65.30±2.14
(2)		✓			62.34±0.91	3.70±0.49	60.57±0.88	51.22±0.68	5.29±0.49	52.14±0.78
(3)			✓		63.71±0.59	3.47±0.41	62.48±0.32	51.88±1.11	4.60±0.37	52.63±1.03
(4)				✓	65.62±0.58	5.29±0.56	65.15±0.99	56.03±1.81	4.39±1.91	57.31±1.62
(5)	✓	✓			73.22±0.90	5.10±0.66	75.50±0.69	60.36±1.27	7.76±1.05	65.90±1.48
(6)	✓		✓		73.36±0.58	3.60±0.36	75.46±1.20	60.66±0.76	5.02±0.82	64.24±1.15
(7)	✓			✓	75.52±0.88	5.90±0.38	**77.80±0.79**	61.57±2.16	8.52±2.72	**68.62±1.50**
(8)	✓	✓	✓		74.14±0.81	**3.34±0.79**	74.42±0.86	62.98±1.34	**3.92±0.68**	63.44±1.64
(9)	✓	✓		✓	75.10±0.92	5.65±0.80	77.66±0.48	65.19±1.05	6.88±0.62	68.31±1.82
(10)	✓		✓	✓	75.11±1.16	4.44±0.89	76.61±1.46	66.00±1.81	5.46±0.80	68.73±1.50
(11)	✓	✓	✓	✓	**75.40±0.44**	4.40±0.30	76.87±1.66	**67.30±1.57**	4.38±0.92	67.74±1.85

ICLR

Thank you!

For any question, feel free to ask via e-mail

VERI TAS LUX MEA

SEOUL
NATIONAL
UNIVERSITY