Poster Session
WED-AM-344

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Few-Shot Class-Incremental Learning via Class-Aware Bilateral Distillation

Linglan Zhao[1*], Jing Lu[2*], Yunlu Xu[2], Zhanzhan Cheng[2†], Dashan Guo[1], Yi Niu[2], Xiangzhong Fang[1]

[1]Department of Electronic Engineering, Shanghai Jiao Tong University [2]Hikvision Research Institute

https://github.com/LinglanZhao/BiDistFSCIL

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

HIKVISION

Overview

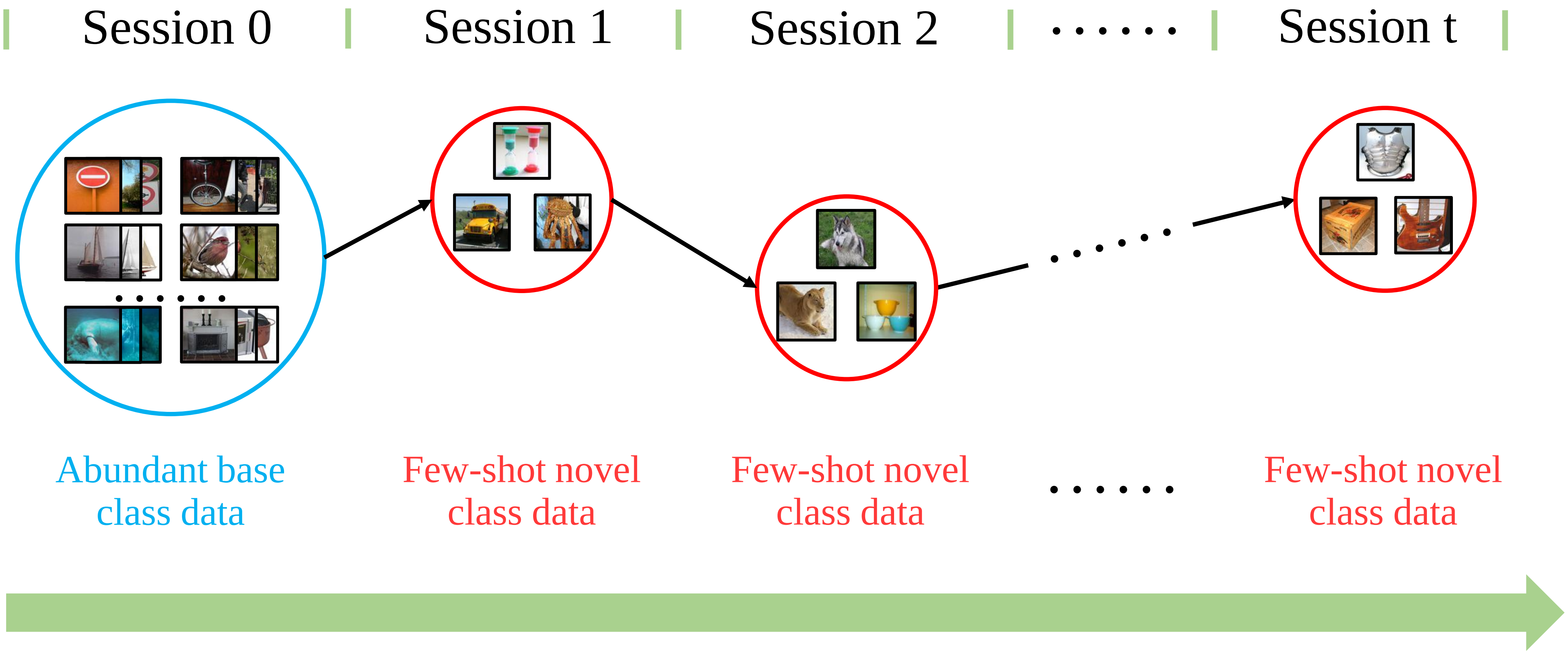
In this paper, we:

- propose a **class-aware bilateral distillation** method to adapt the vanilla knowledge distillation technique for FSCIL.
- propose a **two-branch framework** to be conveniently applied to arbitrary pre-trained models without sophisticated meta-training.
- achieve state-of-the-art performance on three popular FSCIL datasets.

(FSCIL is short for Few-Shot Class-Incremental Learning)

Task Definition

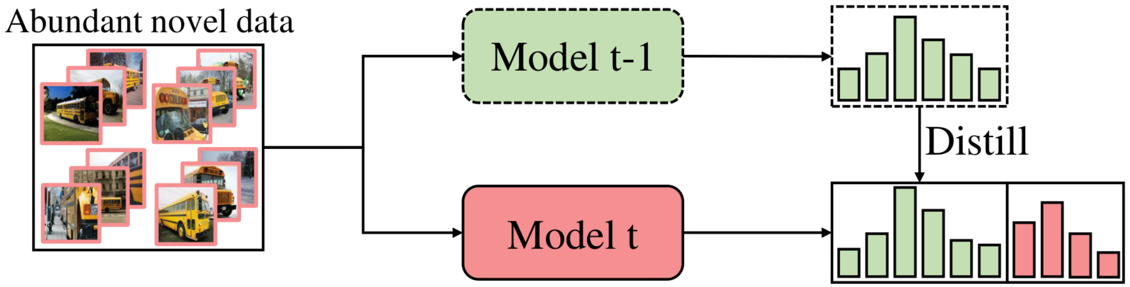
- Few-Shot Class-Incremental Learning (FSCIL):



- Incrementally recognize all the encountered classes

Motivation

- Vanilla knowledge distillation in Class-Incremental Learning (CIL) is **not suitable** for Few-Shot Class-Incremental Learning (FSCIL) task.



Vanilla distillation:

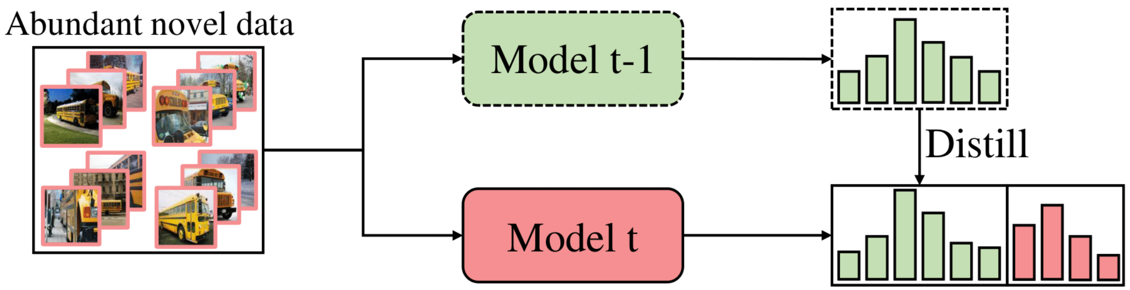
Drawing knowledge from only previous model (t-1)

Reasons

- Finetuning on few-shot novel class data in FSCIL results in the **unique overfitting issue**.
- Vanilla distillation for CIL causes **aggravated catastrophic forgetting** in FSCIL.

Motivation

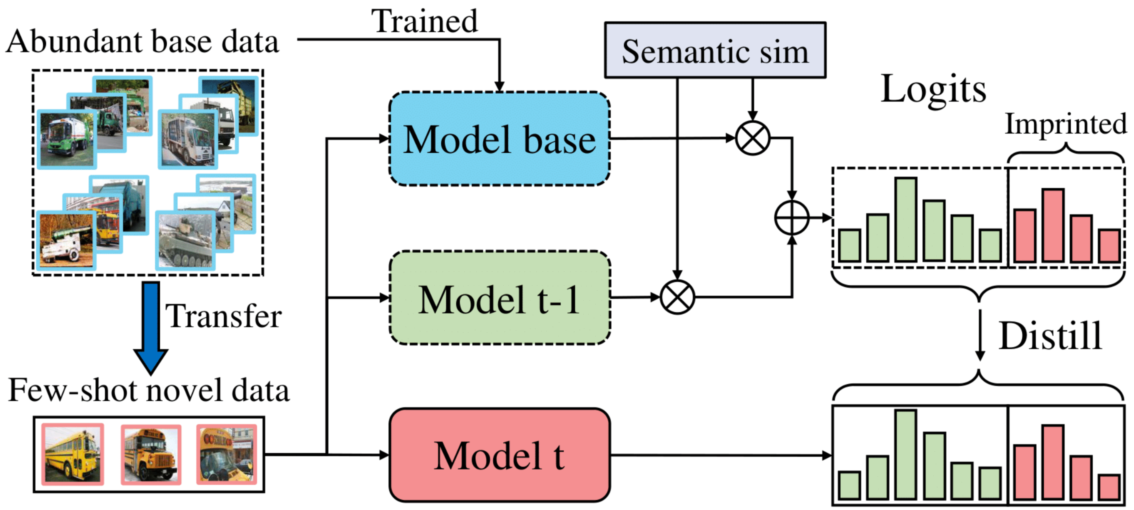
- Class-Incremental Learning (CIL):



<u>Vanilla distillation:</u>

Drawing knowledge from only previous model (t-1)

- Few-Shot Class-Incremental Learning (FSCIL):



<u>Our distillation:</u>

Drawing knowledge dynamically from dual teachers

Method

➢ Class-Aware Bilateral Distillation (CABD)

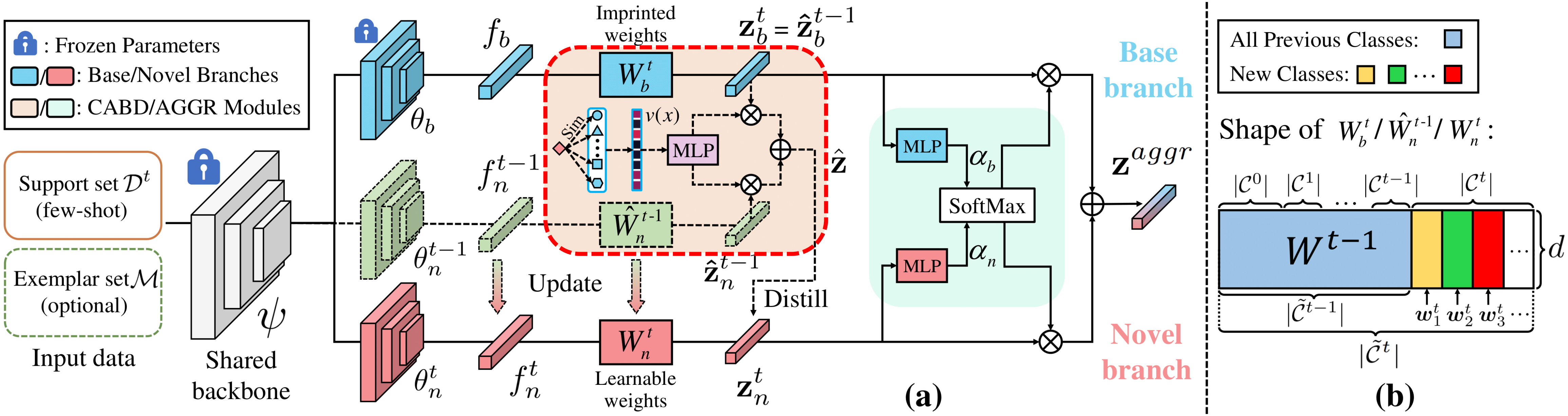
- Drawing knowledge from dual teachers:

$$\hat{\mathbf{z}} = \underbrace{\rho(\mathbf{x}) \cdot \hat{\mathbf{z}}_b^{t-1}}_{\text{Base model from session 0}} + \underbrace{(1-\rho(\mathbf{x})) \cdot \hat{\mathbf{z}}_n^{t-1}}_{\text{Previous model from session t-1}}$$

- Dynamic transfer based on semantic similarity:

$$\rho(\mathbf{x}) = \begin{cases} 1.0 & \text{if } y(\mathbf{x}) \in \mathcal{C}^0 \\ 1/(1+e^{-g_\vartheta(\boldsymbol{v}(\mathbf{x}))}) & \text{if } y(\mathbf{x}) \notin \mathcal{C}^0 \end{cases}$$

$$g_\vartheta(\boldsymbol{v}(\mathbf{x})) = \mathrm{MLP}\left(\left[cos(\boldsymbol{w}_c, \boldsymbol{w}_1), \ldots, cos(\boldsymbol{w}_c, \boldsymbol{w}_{|\mathcal{C}^0|})\right]; \vartheta\right)$$



Method

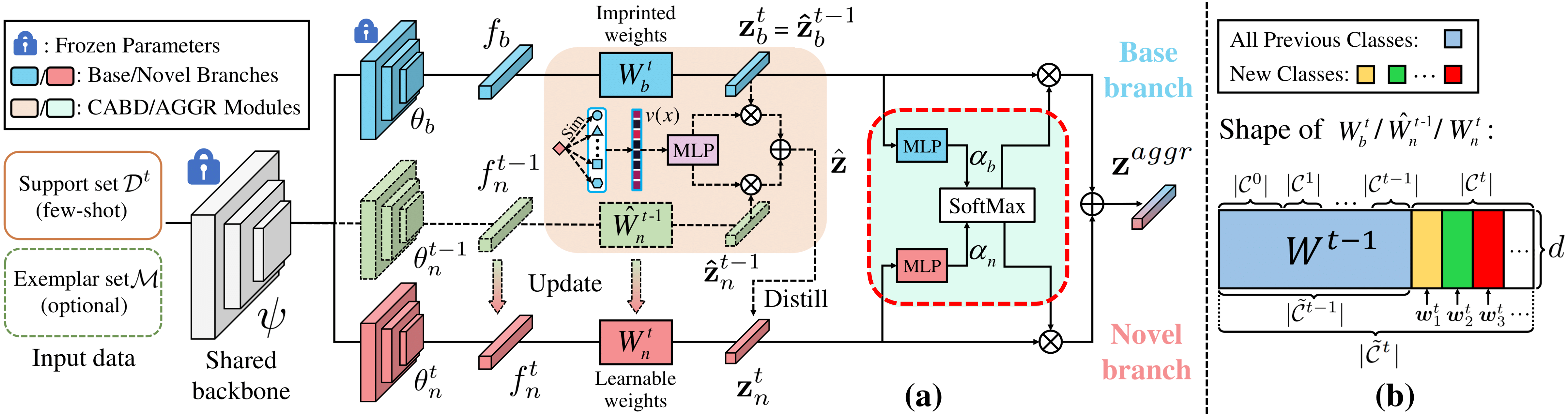
➢ Attention-based Prediction Aggregation (AGGR)

- Selectively merge predictions from the two branches:

$$\mathbf{z}^{aggr} = [\mathbf{z}_b^t, \mathbf{z}_n^t] * \mathrm{softmax}([\alpha_b, \alpha_n])^\top$$

- Binary classification loss to enhance discrimination:

$$\mathcal{L}_{bin} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^t\cup\mathcal{M}}\left[\mathrm{CE}([\alpha_b, \alpha_n], \mathbf{I}[y \in \mathcal{C}^0])\right]$$



Results

Performance on three FSCIL datasets

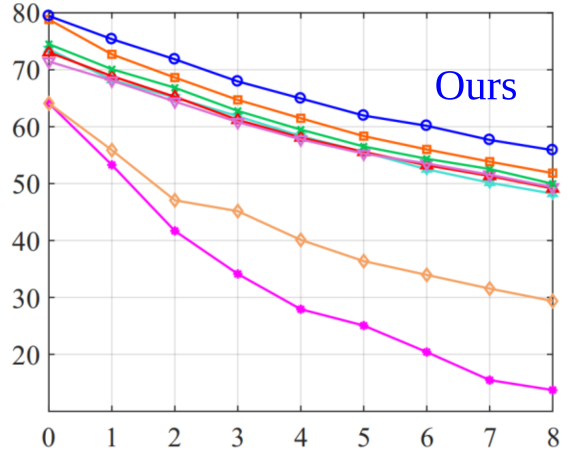
- Achieving state-of-the-art classification results.

(1) *mini*-ImageNet

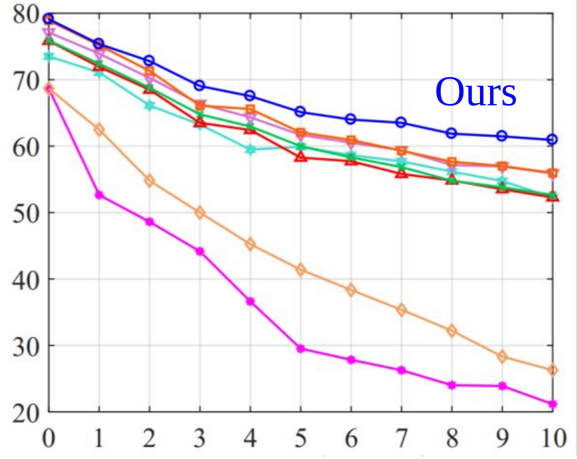
Method	Accuracy in each session (%)									Avg.	Final Impro.
	0	1	2	3	4	5	6	7	8		
iCaRL*◇ [26]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	33.29	+35.01
TOPIC [30]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	39.64	+27.80
ERL++** [8]	61.70	57.58	54.66	51.72	48.66	46.27	44.67	42.81	40.79	49.87	+11.43
IDLVQ* [3]	64.77	59.87	55.93	52.62	49.88	47.55	44.83	43.14	41.84	51.16	+10.38
CEC [39]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	57.75	+4.59
F2M** [28]	72.05	67.47	63.16	59.70	56.71	53.77	51.11	49.21	47.84	57.89	+4.38
CLOM [44]	73.08	68.09	64.16	60.41	57.41	54.29	51.54	49.37	48.00	58.48	+4.22
Replay* [21]	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	58.02	+4.01
MetaFSCIL [6]	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	58.85	+3.03
FACT♮ [41]	**75.32**	70.34	65.84	62.05	58.68	55.35	52.42	50.42	48.51	59.88	+3.71
Ours (0 exemplar)	74.65	69.89	65.44	61.76	59.49	56.11	53.28	51.74	50.49	60.32	
Ours (1 exemplar)[default]*	74.65	**70.43**	**66.29**	**62.77**	**60.75**	**57.24**	**54.79**	**53.65**	**52.22**	**61.42**	
Ours (5 exemplars)**	74.65	70.70	66.81	63.63	61.36	58.14	55.59	54.23	53.39	62.06	

*: method with 1 exemplar per class. **: method with 5 exemplars per class. ◇: results from [30]. ♮: results using the publicly available code from [41].

(2) CIFAR100



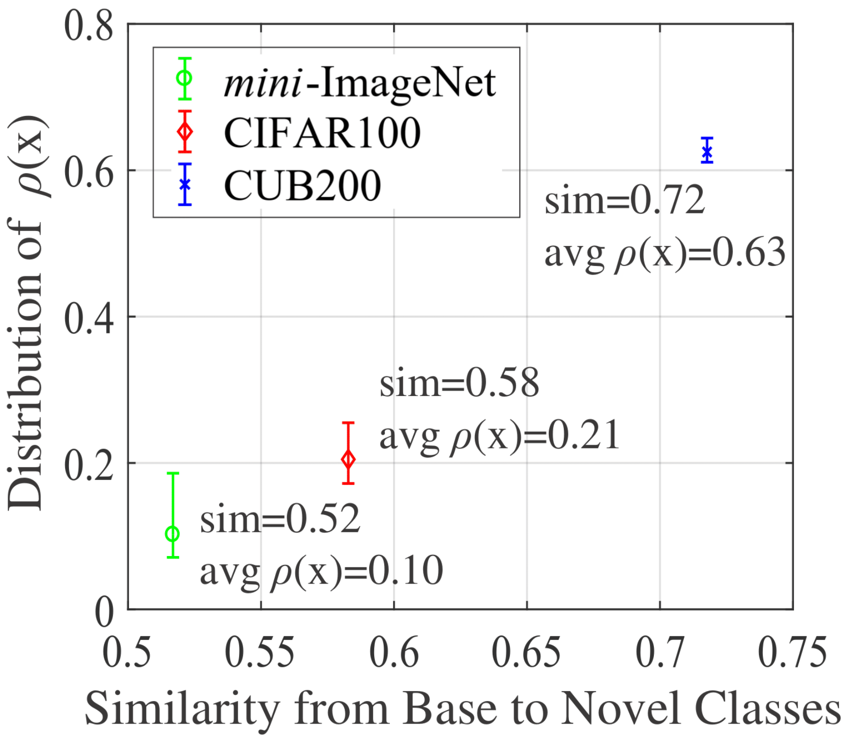
(3) CUB200



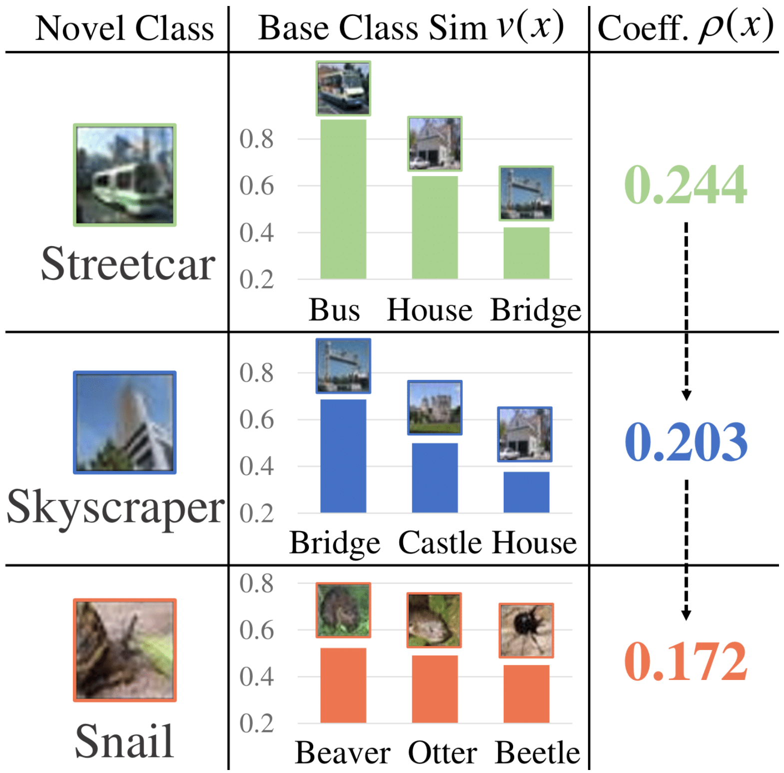
Results

Ablation on Class-Aware Bilateral Distillation (CABD)

- Distilling based on class-aware semantic similarity.
- More semantically related → more general knowledge to transfer.



(a) From a dataset-wise view

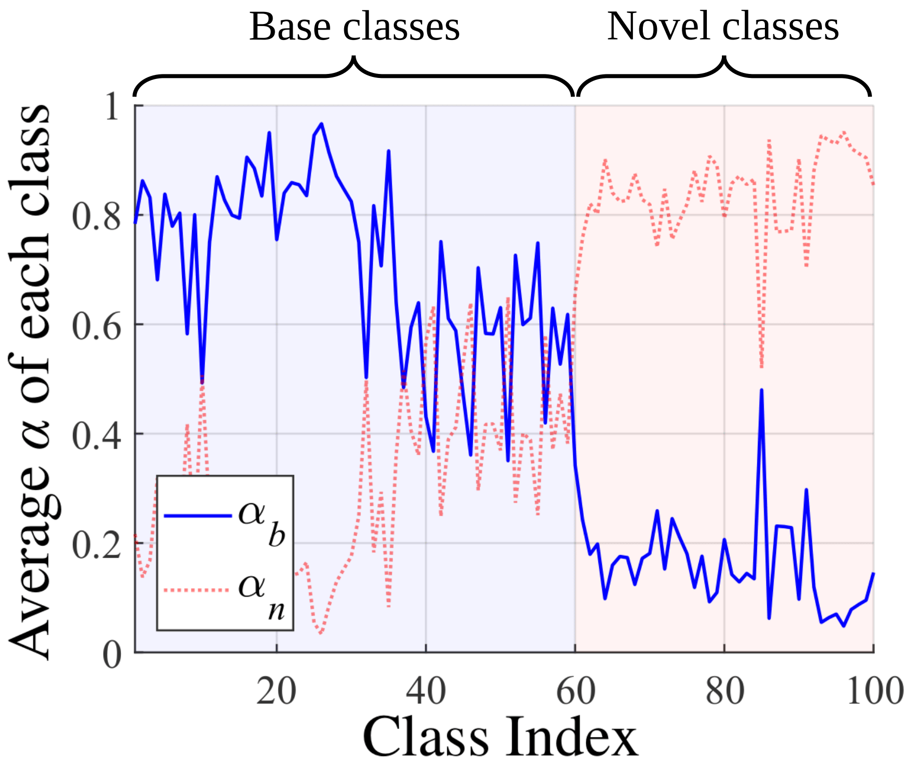


(b) From a class-wise view

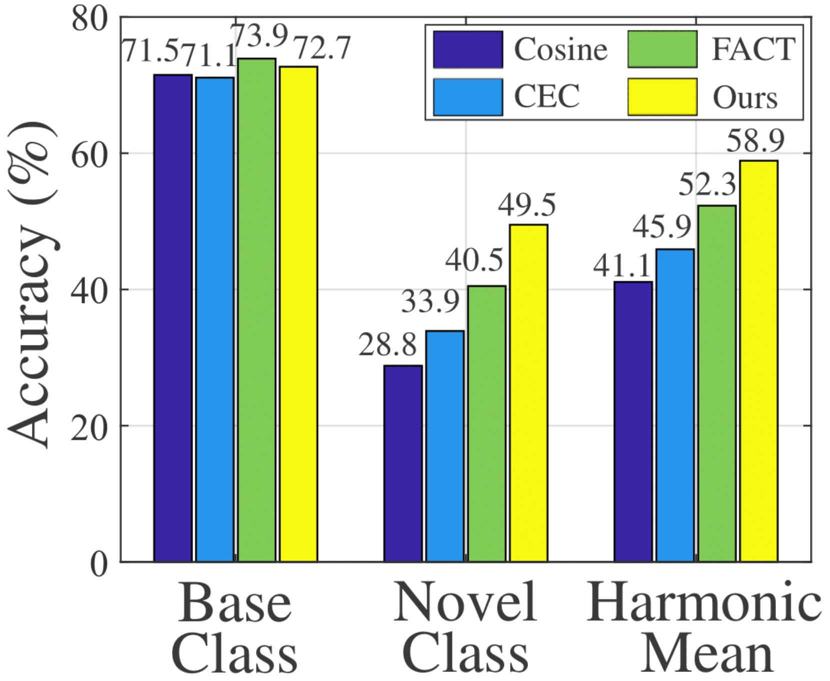
Results

Ablation on Attention-based Prediction Aggregation (AGGR)

- Reasonable aggregation weights for each instance.
- Better trade-off between base and novel classes.



(a) Attention weights for aggregation

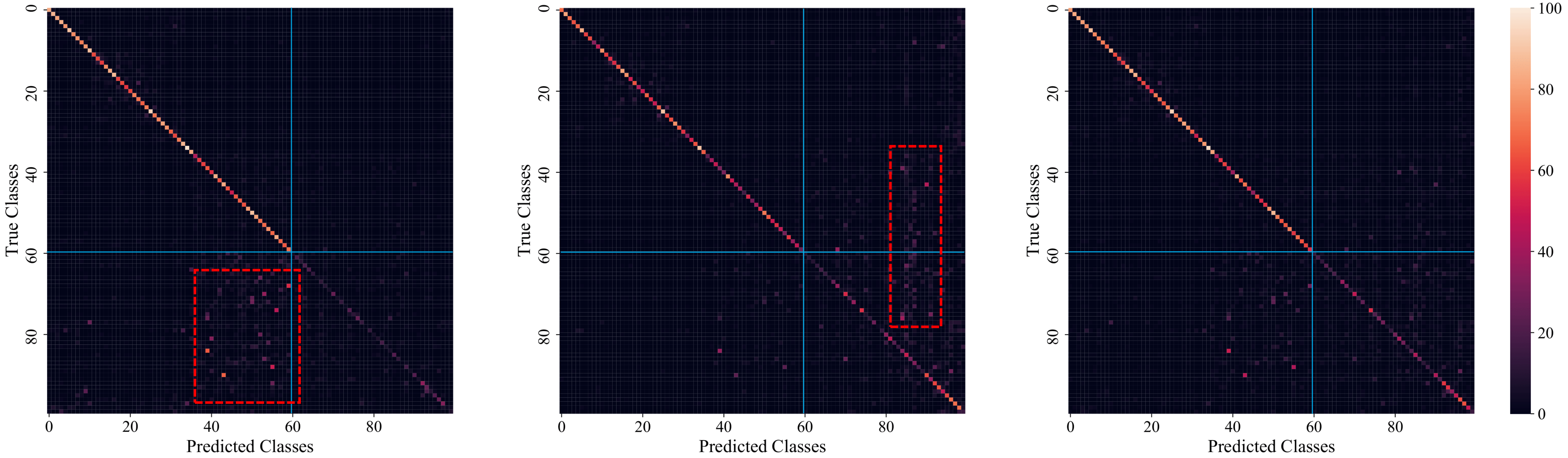


(b) Base & novel class trade-off

Results

Confusion Matrix

- Our method obtains a less scattered confusion matrix for better performance.



(a) Base branch
Acc. = 48.97%

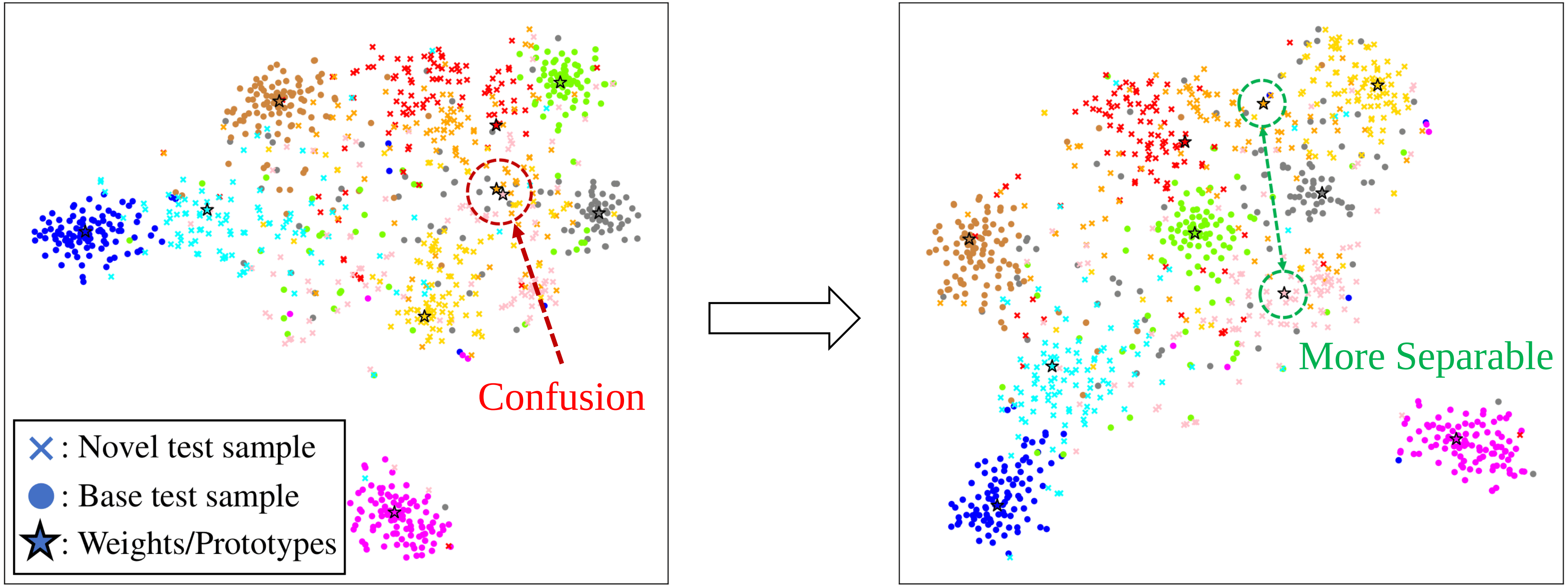
(b) Vanilla distillation
Acc. = 45.33%

(c) Our method
Acc. = 52.22%

Results

T-SNE Visualization

- Better adaptation to novel classes without forgetting.



(a) Base branch (before adaptation) (b) Novel branch (after adaptation)

Poster Session
WED-AM-344

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks!

Linglan Zhao[1,*], Jing Lu[2,*], Yunlu Xu[2], Zhanzhan Cheng[2,†], Dashan Guo[1], Yi Niu[2], Xiangzhong Fang[1]
[1]Department of Electronic Engineering, Shanghai Jiao Tong University [2]Hikvision Research Institute
{llzhao,dmlab_gds,xzfang}@sjtu.edu.cn, {lujing6,xuyunlu,chengzhanzhan,niuyi}@hikvision.com

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

HIKVISION®