Supplemental Material for 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
{llzhao,dmlab_gds,xzfang}@sjtu.edu.cn, {lujing6,xuyunlu,chengzhanzhan,niuyi}@hikvision.com

A. Visualizations of Adapted Features

In this part, we provide more analyses of our key component Class-Aware Bilateral Distillation (CABD). In Fig. 1, we visualize the feature embeddings and the corresponding classification weights (*i.e.*, prototypes) from the *mini*-ImageNet test set with (novel branch) or without (base branch) adaptation to novel classes using our proposed distillation module. For clarity, 5 base classes and 5 novel classes are randomly chosen and features of 100 per-class test samples are considered. As shown in the left part of Fig. 1 (highlighted with dark red circle), the orange novel class prototype (marked with star) is confused with the pink novel class without CABD, since the base branch is trained only on base categories which can not effectively adapt to novel concepts. By contrast, after adaptation to novel classes with our proposed CABD, the above two classes become more distinguishable in the novel branch shown in the dark green circles from the right part of Fig. 1. The above observation explains the improved performance of the novel branch in Table 2 of our main paper.

B. Analyses of Incremental Shot

For further validating the effectiveness of our proposed method, we vary the shot number (*i.e.*, the number of training samples in each incremental class) of the original N-way K-shot few-shot class-incremental learning task. We can see from Fig. 2 that our method can be applied to extreme cases where only a single training sample (1-shot) is provided, highlighting the robustness of the proposed approach. In addition, given more training samples from novel classes, improved performance is observed correspondingly. It is because our approach can better adapt to these incremental classes with the help of more training data, which proves the extendibility of our method.

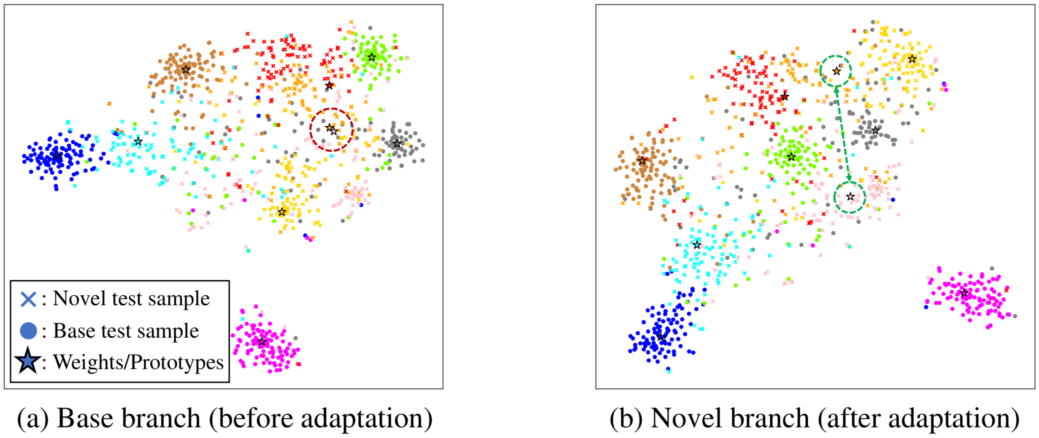
C. Datasets of Different Semantic Relations

For the readers to have a better understanding of the semantic characteristics of different benchmark datasets: *mini*-ImageNet [12], CIFAR100 [6] and CUB200 [13], we visualize each dataset by sampling one image for each class shown in Fig. 3. In addition, according to the setting of Few-Shot Class-Incremental Learning (FSCIL), we also split each dataset into the base classes and novel classes with colors blue and red, respectively.

As shown in Fig. 3, the fine-grained classification dataset CUB200 contains samples from only bird categories with similar appearance, which leads to strong semantic correlations between base and novel classes. In contrast, images from regular classification datasets *mini*-ImageNet and CIFAR100 show diversified visual looking, and the semantic similarities between base and novel classes are in a lower level compared to CUB200. The above observation verifies the experimental results in Fig. 5 (a) of the main paper, that is, quantitative semantic similarities between base and novel classes from *mini*-ImageNet and CIFAR100 datasets are lower than that of the fine-grained classification dataset CUB200. Furthermore, it also validates the empirical finding that more generalizable knowledge from base classes (*i.e.*, with a larger value of the coefficient $\rho(\mathbf{x})$ in Eq. 2 of the main paper) should be transferred for facilitating the learning of novel classes in CUB200 due to the strong semantic correlations between them.

We can see from Fig. 3 that the first half part of base classes from *mini*-ImageNet (indices 1-35) belong to animal classes (*e.g.*, "house finch", "robin" and "green mamba"), while the rest of base classes (indices 36-60) and the novel classes (indices 61-100) include a large proportion of inorganic objects. As a result, these base classes with indices 1-35 are relatively less similar to novel classes. Thus, for better handling these classes, the model should pay more attention to the predictions from the base branch by using a larger base branch attention weights α_b, which further confirms the results of Fig. 6 (a) in our main paper.

[*] Equal contribution. [†] Corresponding author.



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

Figure 1. T-SNE [11] plots of test samples and the corresponding classification weights/prototypes in the final session from *mini*-ImageNet with (novel branch) or without (base branch) our proposed Class-Aware Bilateral Distillation (CABD) module. Categories are represented by different colors. Best viewed in color.

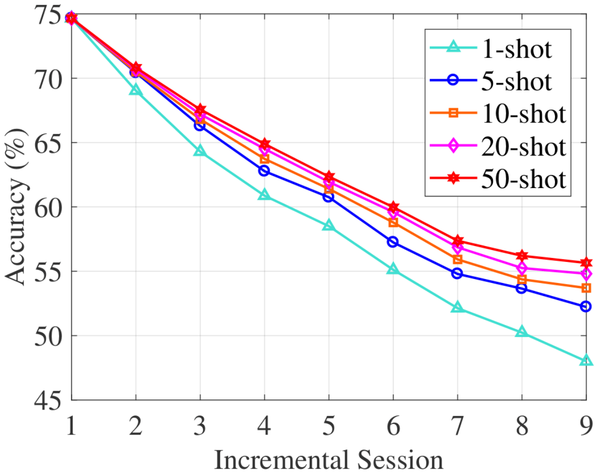


Figure 2. Experiments on the influence of incremental shot on *mini*-ImageNet dataset.

D. Analyses of Confusion Matrix

To better understand the unique difficulties of few-shot class-incremental learning task, we plot the confusion matrix generated by (a) our base branch, (b) vanilla knowledge distillation (*i.e.*, directly using the output of model t-1 for distillation) and (c) our full method in Fig. 4.

We can see from Fig. 4a that base branch specializes in classifying base classes with concentrated values on the diagonal of these categories. However, the base branch performs poorly on novel classes with much darker diagonal on them, since the base branch is only trained on the base training set without adaptation to novel classes.

By contrast, adapting with vanilla knowledge distillation in Fig. 4b can better handle novel class samples but fails to preserve base knowledge, resulting in darker diagonal on base classes compared to Fig. 4a and scattered prediction distribution. It is because the severe data scarcity of few-shot class-incremental learning not only causes the unique overfitting issue but also aggravates catastrophic forgetting.

As shown in Fig. 4c, with the proposed class-aware bilateral distillation module and attention-based aggregation module, our full method can address the above difficulties with concentrated values on the diagonal of both base and novel classes, confirming the observed performance gains in experiments.

E. Detailed Experimental Results

In Table 1 and Fig. 3 of our main paper, we provide comparisons to the state-of-the-arts. Here, we present more detailed results on CIFAR100 and CUB200. Table 1 and Table 2 show that our method with default setting (1 exemplar per class) significantly outperforms all previous approaches. Compared to the second-best results on both benchmark datasets, we achieve 4.04% and 4.97% increase in the accuracy from the final session, and 3.21% and 2.79% improvement in the average performance.

As discussed in Section 4 of the main paper, to trade off between memory cost and accuracy, our method is also flexible enough to address situations where none or more exemplars are available. It is observed from Table 1 and Table 2 that, we further improve the accuracy in the final session by 0.68% and 2.88% on CIFAR100 and CUB200 datasets when 5 exemplars are provided [4, 9]. Moreover, when exemplars are not provided in incremental sessions (*i.e.*, 0 exemplar), our method can still outperform all existing works, which validates the superiority of our proposed framework. Code is available at https://github.com/LinglanZhao/BiDistFSCIL.

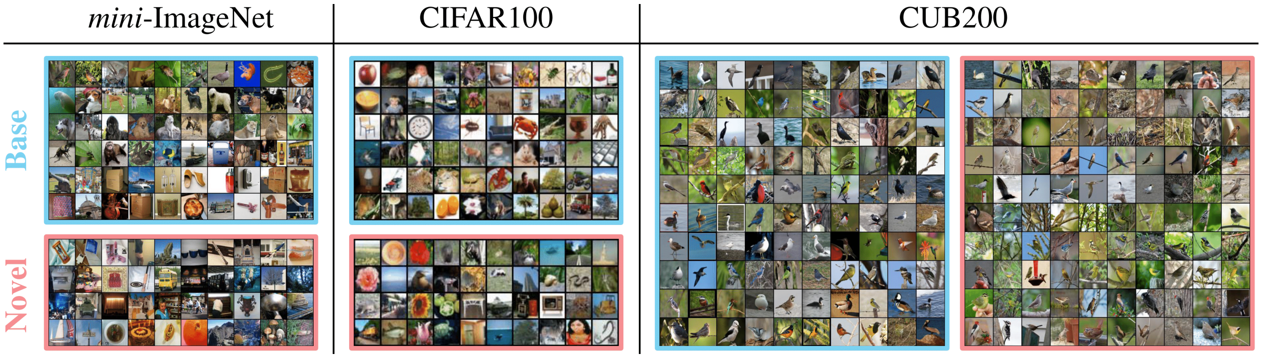
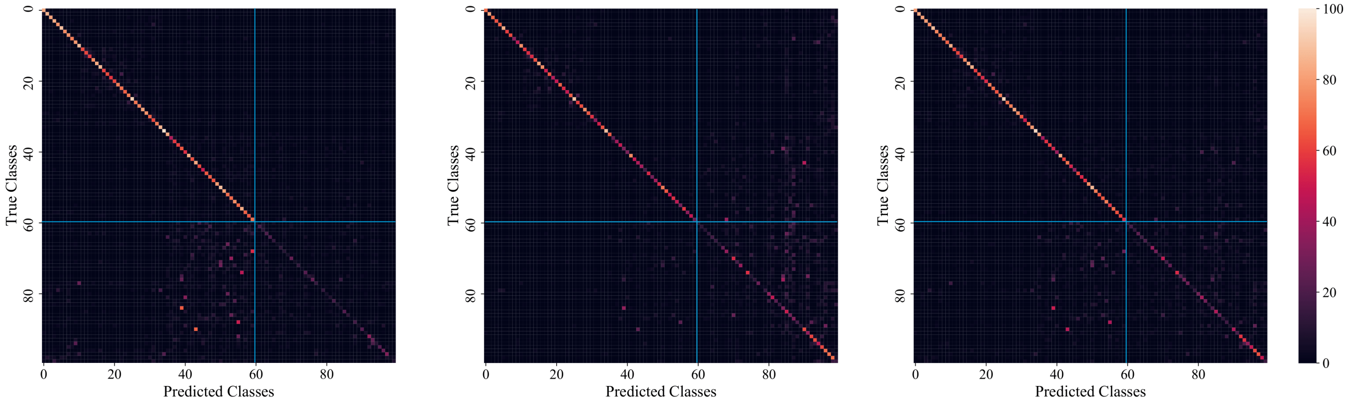


Figure 3. Visualizations of three FSCIL datasets *mini*-ImageNet [12], CIFAR100 [6] and CUB200 [13] with separated base (in blue) and novel (in red) classes. One example image for each class is sampled and the images are placed (from left to right, from top to bottom) by the order of class indices.



(a) Base branch (Accuracy = 48.97%)

(b) Vanilla distillation (Accuracy = 45.33%)

(c) Our full method (Accuracy = 52.22%)

Figure 4. Confusion matrices of baseline approaches and our full method on *mini*-ImageNet. Blue lines are used to separate base classes and novel classes. Our full method effectively improves the prediction in the final session resulting in a less scattered confusion matrix.

References

[1] Idan Achituve, Aviv Navon, Yochai Yemini, Gal Chechik, and Ethan Fetaya. Gp-tree: A gaussian process classifier for few-shot incremental learning. In *International conference on machine learning*, pages 54–65, 2021. 4

[2] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *European Conference on Computer Vision*, pages 233–248, 2018. 4

[3] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 14166–14175, 2022. 4

[4] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1255–1263, 2021. 2, 4

[5] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 831–839, 2019. 4

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 3

[7] Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *European Conference on Computer Vision*, 2022. 4

[8] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2001–2010, 2017. 4

[9] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in neural information processing systems*, pages 6747–6761, 2021. 2, 4

[10] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE conference*

Method	Accuracy in each session (%)									Avg.	Final Impro.
	0	1	2	3	4	5	6	7	8		
Ft-CNN$^{\diamond}$ [10]	64.10	36.91	15.37	9.80	6.67	3.80	3.70	3.14	2.65	16.24	+53.23
iCaRL$^{*\diamond}$ [8]	64.10	53.28	41.69	34.13	27.93	25.06	20.41	15.48	13.73	32.87	+42.15
EEIL$^{*\diamond}$ [2]	64.10	53.11	43.71	35.15	28.96	24.98	21.01	17.26	15.85	33.79	+40.03
LUCIR$^{*\diamond}$ [5]	64.10	53.05	43.96	36.97	31.61	26.73	21.23	16.78	13.54	34.22	+42.23
TOPIC [10]	64.10	55.88	47.07	45.16	40.11	36.38	33.96	31.55	29.37	42.62	+26.51
ERL++** [4]	73.62	68.22	65.14	61.84	58.35	55.54	52.51	50.16	48.23	59.29	+7.65
Cosine* [15]	74.55	67.43	63.63	59.55	56.11	53.80	51.68	49.67	47.68	58.23	+8.20
DeepEMD* [14]	69.75	65.06	61.20	57.21	53.88	51.40	48.80	46.84	44.41	55.39	+11.47
CEC [15]	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	59.53	+6.74
F2M** [9]	71.45	68.10	64.43	60.80	57.76	55.26	53.53	51.57	49.35	59.14	+6.53
CLOM [19]	74.20	69.83	66.17	62.39	59.26	56.48	54.36	52.16	50.25	60.57	+5.63
Replay* [7]	74.40	70.20	66.54	62.51	59.71	56.58	54.52	52.39	50.14	60.77	+5.74
MetaFSCIL [3]	74.50	70.10	66.84	62.77	59.48	56.52	54.36	52.56	49.97	60.79	+5.91
FACT$^{\natural}$ [17]	78.83	72.71	68.63	64.71	61.48	58.34	56.00	53.85	51.84	62.93	+4.04
Ours (0 exemplar)	79.45	75.20	71.34	67.40	64.50	61.05	58.73	56.73	54.31	65.42	
Ours (1 exemplar)[default]*	**79.45**	**75.38**	**71.84**	**67.95**	**64.96**	**61.95**	**60.16**	**57.67**	**55.88**	**66.14**	
Ours (5 exemplars)*	$\underline{79.45}$	$\underline{75.63}$	$\underline{72.00}$	$\underline{68.09}$	$\underline{65.54}$	$\underline{62.59}$	$\underline{60.76}$	$\underline{58.35}$	$\underline{56.56}$	$\underline{66.55}$	

*: methods with 1 exemplar per class. **: methods with 5 exemplars per class. $^{\diamond}$: results from [10]. *: results from [15]. $^{\natural}$: results using the code from [17].

Table 1. Comparisons to state-of-the-art FSCIL methods on CIFAR100. "Final Impro." highlights the improvement in the final session.

Method	Accuracy in each session (%)											Avg.	Final Impro.
	0	1	2	3	4	5	6	7	8	9	10		
Ft-CNN$^{\diamond}$ [10]	68.68	43.70	25.05	17.72	18.08	16.95	15.10	10.60	8.93	8.93	8.47	22.02	+52.46
iCaRL$^{*\diamond}$ [8]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	36.67	+39.77
EEIL$^{*\diamond}$ [2]	68.68	53.63	47.91	44.20	36.30	27.46	25.93	24.70	23.95	24.13	22.11	36.27	+38.82
LUCIR$^{*\diamond}$ [5]	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87	32.49	+41.06
TOPIC [10]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28	43.92	+34.65
SPPR [18]	68.68	61.85	57.43	52.68	50.19	46.88	44.65	43.07	40.17	39.63	37.33	49.32	+23.60
GP-Tree* [1]	72.84	67.00	62.98	58.19	54.84	51.77	49.40	47.57	45.47	44.05	42.72	54.26	+18.21
ERL++** [4]	73.52	71.09	66.13	63.25	59.49	59.89	58.64	57.72	56.15	54.75	52.28	61.17	+8.65
Cosine* [15]	75.52	70.95	66.46	61.20	60.86	56.88	55.40	53.49	51.94	50.93	49.31	59.36	+11.62
DeepEMD* [14]	75.35	70.69	66.68	62.34	59.76	56.54	54.61	52.52	50.73	49.20	47.60	58.73	+13.33
CEC [15]	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	61.33	+8.65
F2M** [9]	77.13	73.92	70.27	66.37	64.34	61.69	60.52	59.38	57.15	56.94	55.89	63.96	+5.04
Replay* [7]	75.90	72.14	68.64	63.76	62.58	59.11	57.82	55.89	54.92	53.58	52.39	61.52	+8.54
MgSvF [16]	72.29	70.53	67.00	64.92	62.67	61.89	59.63	59.15	57.73	55.92	54.33	62.37	+6.60
MetaFSCIL [3]	75.90	72.41	68.78	64.78	62.96	59.99	58.30	56.85	54.78	53.82	52.64	61.93	+8.29
FACT$^{\natural}$ [17]	78.91	75.19	71.34	66.09	65.59	62.06	60.92	59.31	57.65	57.01	55.96	64.55	+4.97
Ours (0 exemplar)	79.12	74.99	70.87	67.30	65.89	63.45	61.40	60.11	58.61	58.23	57.48	65.22	
Ours (1 exemplar)[default]*	**79.12**	**75.37**	**72.80**	**69.05**	**67.53**	**65.12**	**64.00**	**63.51**	**61.87**	**61.47**	**60.93**	**67.34**	
Ours (5 exemplar)*	$\underline{79.12}$	$\underline{75.63}$	$\underline{73.21}$	$\underline{69.93}$	$\underline{68.32}$	$\underline{66.30}$	$\underline{65.15}$	$\underline{64.96}$	$\underline{64.20}$	$\underline{64.03}$	$\underline{63.81}$	$\underline{68.61}$	

*: methods with 1 exemplar per class. **: methods with 5 exemplars per class. $^{\diamond}$: results from [10]. *: results from [15]. $^{\natural}$: results using the code from [17].

Table 2. Comparisons to state-of-the-art FSCIL methods on CUB200. "Final Impro." highlights the improvement in the final session.

on computer vision and pattern recognition, pages 12183–12192, 2020. 4

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[12] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, 2016. 1, 3

[13] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical report*, 2011. 1, 3

[14] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12203–12213, 2020. 4

[15] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan,

and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12455–12464, 2021. 4

[16] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[17] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9046–9056, 2022. 4

[18] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6801–6810, 2021. 4

[19] Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. In *Advances in neural information processing systems*, 2022. 4