

Supplementary Materials: Class Balance Matters to Active Class-Incremental Learning

Anonymous Authors

CONTENTS

The following items are included in the supplementary material:

- Details of Selected Benchmarks in Sec. A.
- More implementation details in Sec. B.
- More detailed experiments results, *e.g.*, comparison of CBS with other active learning methods and random selection under various labeling budget when applied them to L2P [15] and DualPrompt [14] in Sec. C.
- Further analysis of the effectiveness of CBS and the utilization of unlabeled data. in Sec. D.
- Limitation and future work in Sec. E.

A DETAILS OF BENCHMARKS

We conduct experiments on selected five publicly available image classification datasets, *i.e.*, CUB-200 [13], CIFAR-100 [6], *mini*-ImageNet [11], DTD [2] and Flowers102 [7], to evaluation our CBS. The first three datasets are commonly utilized for evaluation in CIL or FSCIL, while the latter two datasets are more challenging classification datasets usually adopted to evaluate for vision-language model [8]. We evenly divide each dataset into multiple subsets to construct incremental sessions, and the details are present in the supplementary materials.

- **CUB-200** is a dataset designed for fine-grained classification, comprises approximately 6,000 training images across 200 bird species. We evenly divide the 200 classes into 10 incremental sessions, with each session containing 20 classes and each class containing about 30 unlabeled images.
- **CIFAR-100** consists of 100 general classes, each of which contains 50,000 training images. We evenly divide the 100 classes into 5 incremental sessions, with each session containing 20 classes and each class containing about 500 unlabeled images.
- **mini-ImageNet** is a small set of ImageNet [11], which has 50,000 training images from 100 chosen classes. We evenly divide the 100 classes into 5 incremental sessions, with each session containing 20 classes and each class containing about 500 unlabeled images.
- **DTD** is a collection of 47 different texture with 2,820 training images. We evenly divide the first 40 classes into 2 incremental sessions, with each session containing 20 classes and each class containing about 60 unlabeled images.

- **Flowers102** is designed for fine-grained flower classification, consists of 102 flower classes, with a total of 4,093 training images. We evenly divide the first 100 classes into 5 incremental sessions, with each session containing 20 classes and each class containing about 40 unlabeled images.

In addition, we also evaluate the effectiveness of CBS on datasets that the unlabeled pool are inherently class-imbalanced (*e.g.*, CIFAR-100-LT). Specifically, we transform the unlabeled pool in each session of CIFAR-100 into a long-tail distribution [16] with an imbalance ratio of 10 to build the class-inherently imbalanced unlabeled pool, and the test set remains unchanged.

B IMPLEMENTATION DETAILS

All experiments are conducted with PyTorch on NVIDIA RTX 2080Ti GPU. We implement ACIL pipeline based on the PyTorch implementations of L2P, DualPrompt, and LP-DiF, respectively. For each CIL method, we incorporate our proposed CBS and compared active learning methods with it. On each dataset, we conduct experiments under the annotation budget size $B \in \{40, 60, 80, \dots, 200\}$ for each session, respectively. Note that our method selects B samples at once for each session, whereas some compared active learning algorithms are based on multiple rounds to selection, labeling, and training. Therefore, for these methods, we maintain their multi-round pipeline and make them select 20 samples in each round for labeling until the number of selected samples reaches B . For the optimizer and learning rate, we maintained consistency with the original implementations of L2P, DualPrompt, and LP-DiF when applying all the active learning methods. When applying CBS, all incremental learners train for 50 epochs in each session. When applying other active learning methods, we follow their multi-round training and labeling paradigm. To achieve both fairness and training efficiency, these methods train for 20 epochs in each of the first $R - 1$ rounds and 50 epochs in the R -th round, where $R = B/20$. Thus, we ensure that the methods we compare have sufficiently training epochs.

C MORE DETAILED EXPERIMENTS RESULTS

Comparison under various labeling budget. In main paper we have reported the Avg curves of our CBS and comparison with counterparts applied to **LP-DiF** [5] under various labeling budget B in Fig. 2. Here we report the corresponding results when apply CBS and comparison with counterparts to **L2P** [15] and **DualPrompt** [14], as shown in Fig. A and Fig. B. Generally, one can obtain the following observations: 1) For both L2P and DualPrompt, for each dataset, compared to existing SOTA active learning methods and random selection, our proposed CBS achieved the best or comparable performance under any specified labeling budget. Especially under lower labeling budget, *e.g.*, $B = 40$ or $B = 60$, the performance of CBS is significantly higher than other counterparts. 2) For both L2P and DualPrompt, our CBS achieves the highest Mean

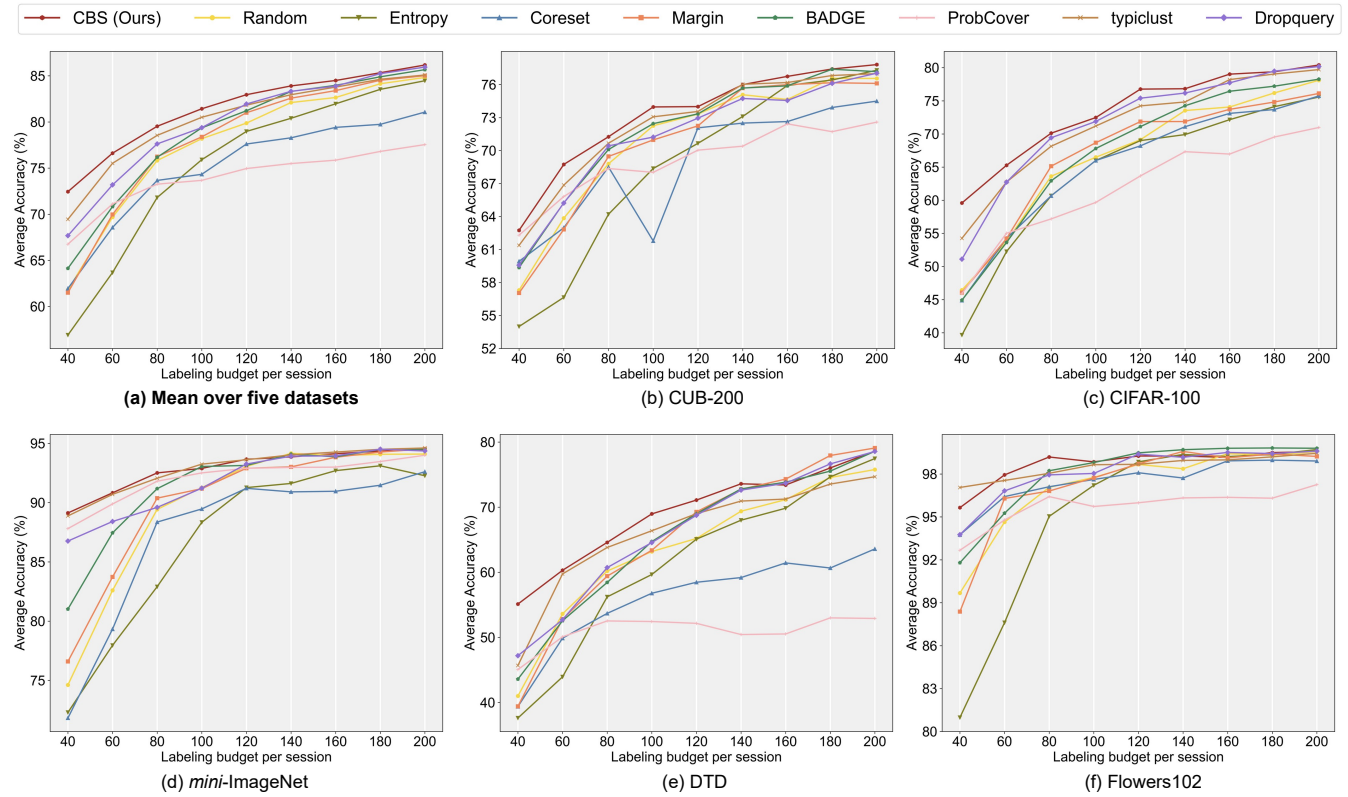


Figure A: Avg curves of our CBS and comparison with counterparts applied to L2P [15] on five datasets (i.e., (b) to (f)) under various labeling budget B . (a) shows the mean Avg curves over five datasets.

Avg over five datasets under each labeling budget compared to all the counterparts. The above results, along with those of LP-DiF in the main paper, fully demonstrate that our CBS can be plug-and-played with these methods which are based on pretrained models with prompt tuning techniques, and show its superiority for ACIL tasks compared to other active learning methods.

Further analysis the class balance of selected examples.

In main paper we have analyzed the class balance of selected examples by our CBS and other counterparts in terms of the “classes discovery ratio”. The “classes discovery ratio” reflects whether the samples selected by the active learning method can cover more categories. Here we report a more intuitive quantitative metric, i.e., the “class-imbalanced ratio”, to demonstrate the class balance of samples selected by different methods. The “class-imbalanced ratio” is calculated by dividing the number of samples of the class with the most samples selected by the active learning method in that session by the number of samples of the class with the least samples. The lower the class-imbalanced ratio, the more it indicates that the samples selected by the active learning method are more balanced across classes. Fig. C shows the comparison of CBS and other counterparts applied to LP-DiF in terms of “class-imbalanced ratio” on CUB-200 under various labeling budget. Each curve represents a specific active learning method, and each point on the curve indicates the class-imbalanced ratio of this method at the

corresponding session. Clearly, our CBS demonstrated the lowest class-imbalanced ratio in most sessions under various labeling budget settings. Specifically, when the labeling budget is low, our CBS outperforms other methods by a substantial margin, which explains why CBS achieves a higher Avg when the labeling budget is low compared to other methods in Fig. 2. In addition, we observed that many classic active learning methods exhibit very high imbalance rates compared to random selection, which also explains why the performance of these methods is lower than that of random selection.

Results on CIFAR-100-LT. Consider that the key idea of our CBS is to ensure the distribution of selected samples closely mirrors the distribution of the entire unlabeled pool, thereby achieving a class-balanced selection while also selecting samples that are representative and diverse. Hence, an unavoidable question is, if the unlabeled pool itself is severely class-imbalanced, can our CBS still choose out a balanced training set? To answer this question, we conduct experiments on CIFAR-100-LT, where the unlabeled pool of each session is a long-tailed distribution (a severe classes imbalance) to evaluate our CBS. Tab. A shows the comparison with other counterparts applied to LP-DiF on CIFAR-100-LT under $B = 100$, in terms of accuracy of each session and Avg, and Fig. E shows the comparison in terms of “class-imbalanced ratio”. To our surprise, our method still outperforms other methods in terms of

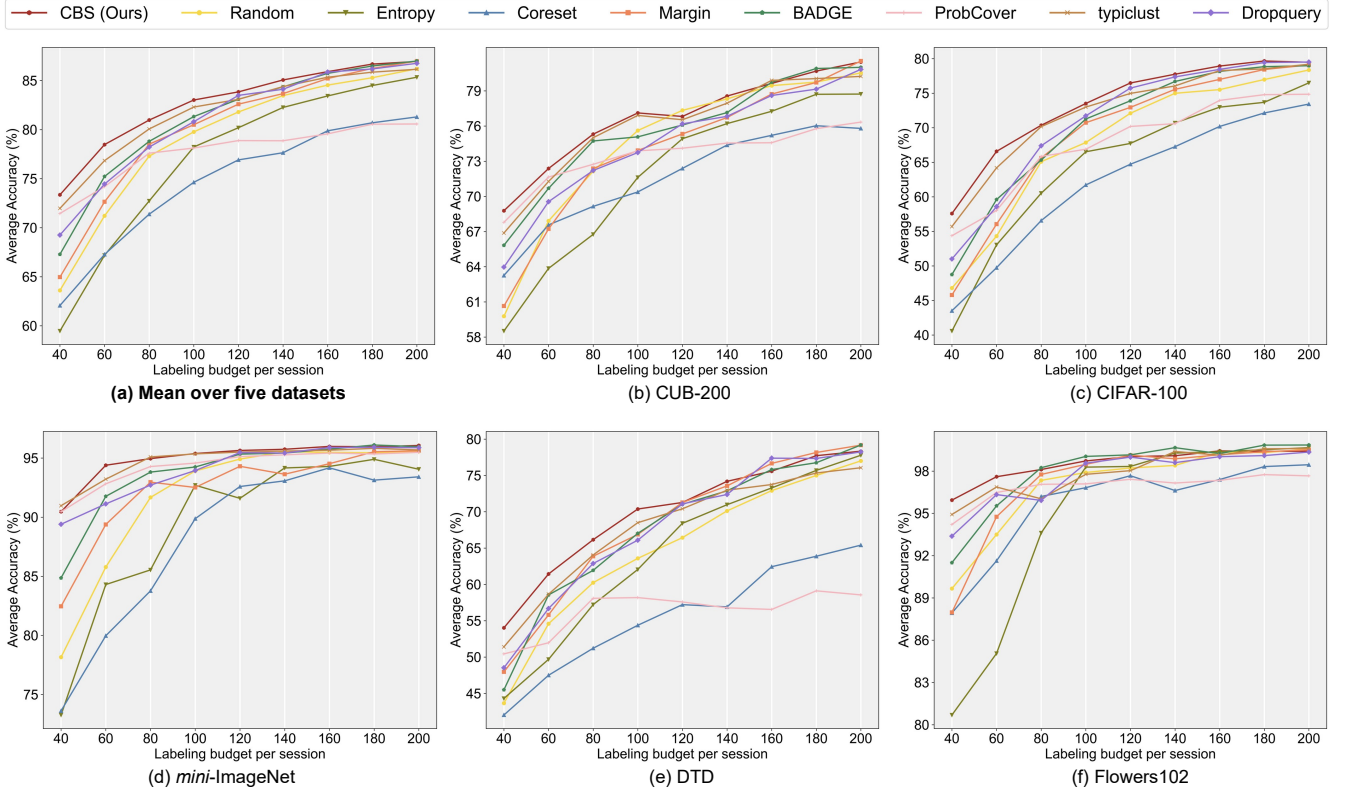


Figure B: Avg curves of our CBS and comparison with counterparts applied to DualPrompt [14] on five datasets (i.e., (b) to (f)) under various labeling budget B . (a) shows the mean Avg curves over five datasets.

performance although but the balance of the samples it selects does not have an advantage over other methods. We speculate that this is because other active learning methods adopt a multi-round train-label paradigm, making them more prone to overfitting on a very small number of imbalanced samples in the initial rounds. In contrast, our method can select B samples at once and then train the model, thereby better resisting overfitting. In future work, we will focus on exploring this issue further.

D MORE ANALYSIS

Further analysis the effect of CBS. The key idea of our CBS is to ensure the distribution of selected samples closely mirrors the distribution of the entire unlabeled pool. To more intuitively explain how CBS achieves this, we calculate the KL divergence between the Gaussian distribution of the selected samples for each class and the distribution of all samples of that class in the unlabeled pool, using the following formula:

$$D_{\text{KL}}(\mathcal{N}(\mu_j, \sigma_j^2) | \mathcal{N}(\hat{\mu}_j, \hat{\sigma}_j^2)) = \frac{1}{2} \sum_{d=1}^D \left(\frac{\sigma_{jd}^2}{\hat{\sigma}_{jd}^2} + \frac{(\hat{\mu}_{jd} - \mu_{jd})^2}{\hat{\sigma}_{jd}^2} + \ln \left(\frac{\hat{\sigma}_{jd}^2}{\sigma_{jd}^2} \right) - 1 \right), \quad (1)$$

where $\mathcal{N}(\mu_j, \sigma_j^2)$ represents the Gaussian distribution of all samples of class j and $\mathcal{N}(\hat{\mu}_j, \hat{\sigma}_j^2)$ represents the Gaussian distribution of samples selected by a active learning method of class j . Statistically, the smaller the D_{KL} , the closer the two Gaussian distributions are,

Table A: Comparison of our method with other active learning approaches when applying them to LP-DiF on CIFAR-100-LT, under $B = 100$. “Avg” represents the average accuracy across all incremental session.

Method.	Accuracy in each session (%) ↑					Avg ↑
	1	2	3	4	5	
LP-DiF [5]						
+ Random (Baseline)	49.50	55.80	56.15	46.74	44.49	50.53
+ Entropy [4]	54.40	51.15	43.08	36.99	27.70	42.66
+ Margin [10]	54.45	52.25	43.15	45.48	36.38	46.34
+ Coreset [12]	53.45	52.88	60.45	52.35	45.04	52.83
+ BADGE [1]	53.90	41.80	44.88	40.94	41.75	44.65
+ Typiclust [3]	58.55	55.50	57.10	49.27	46.19	53.32
+ ProbCover [17]	51.10	48.20	47.70	46.05	43.72	47.35
+ DropQuery [9]	55.50	55.52	51.53	44.92	45.26	50.54
+ CBS (Ours)	63.05	62.67	59.73	53.04	49.19	57.53

indicating that the selected samples are more representative of the entire sample distribution. We applied CBS and random selection to LP-DiF on CUB-200 under $B = 100$ to conduct the experiment, respectively. Fig. D shows the results, where each point in one curve represents the D_{KL} of the class j . Clearly, the samples selected by our CBS have a lower KL divergence with the entire sample set

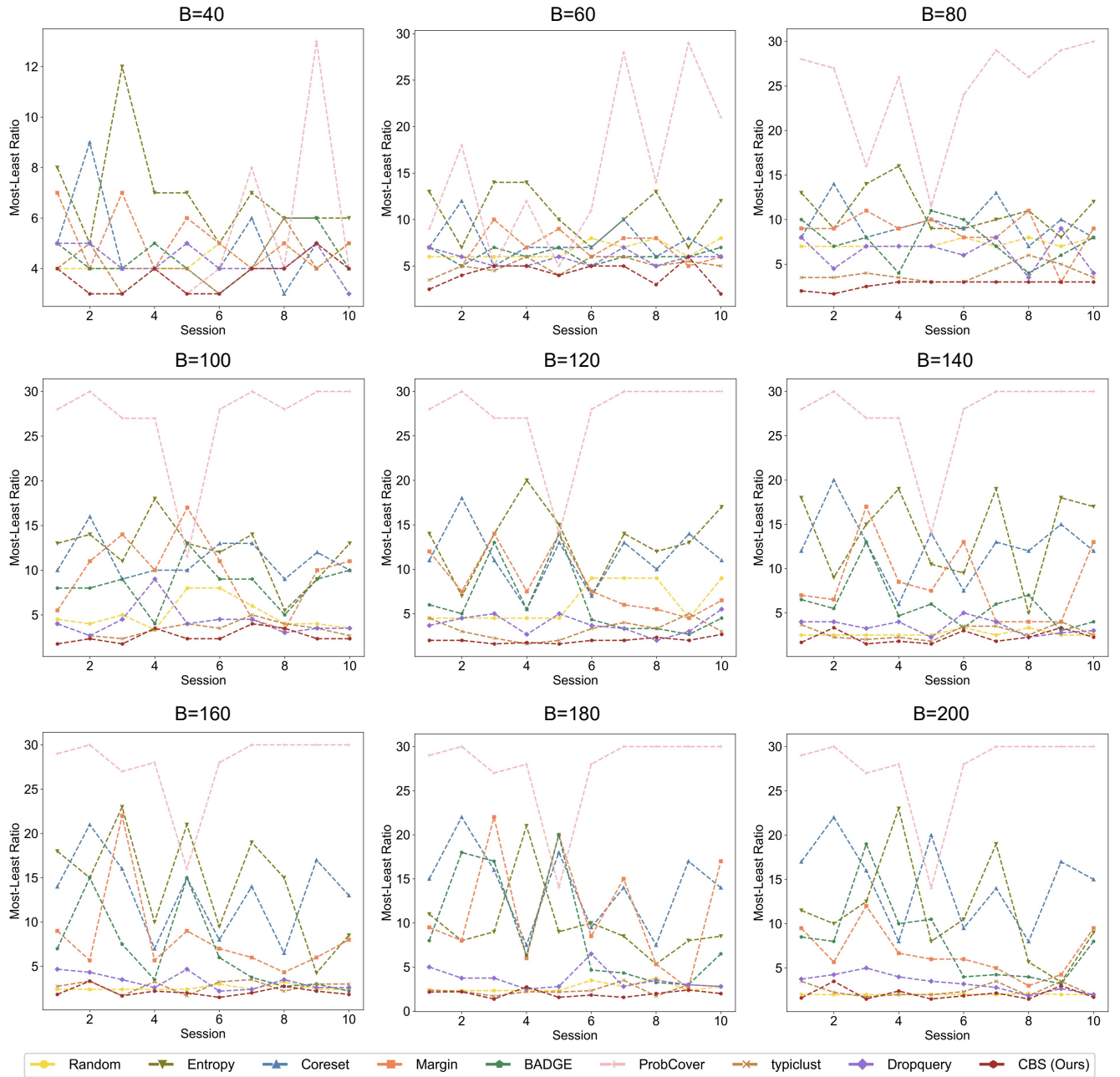


Figure C: Comparison of CBS and other counterparts applied to LP-DiF in terms of “class-imbalanced ratio” on CUB-200 under various labeling budget. Each curve represents a specific active learning method, and each point on the curve indicates the class-imbalanced ratio of this method at the corresponding session. The “class-imbalanced ratio” is calculated by dividing the number of samples of the class with the most samples selected by the active learning method in that session by the number of samples of the class with the fewest samples.

of most classes compared to those selected by random selection. This demonstrates that our method indeed ensures that the selected samples are more representative of the overall distribution.

The runtime cost of CBS. We compare the runtime cost of CBS and Dropquery [9]. Dropquery is a recent active learning method that focuses on performing active learning on pretrained models, achieving new SOTA of active learning problem. It first obtains

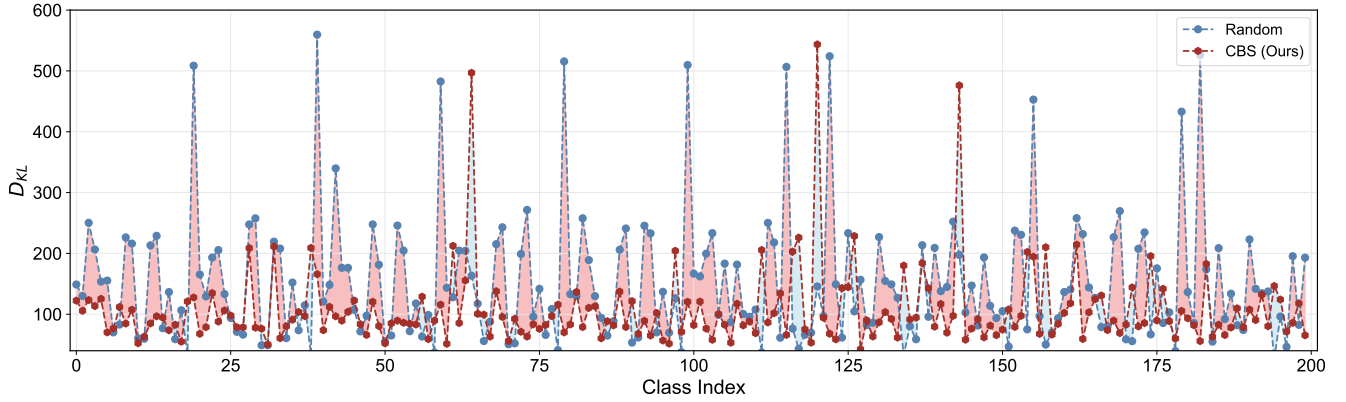


Figure D: The KL divergence between Gaussian distribution estimated by all samples and Gaussian distribution estimated by selected samples, on CUB-200 under $B = 100$. The blue curve and red curve represents applying random selection and CBS respectively. Each point in one curve represents the D_{KL} of a certain class.

Table B: Comparison with Dropquery in terms of runtime cost of each session and the Avg. Sec. represents second.

Method	Runtime cost (sec.) ↓	Avg. (%) ↑
Dropquery	149	72.07
CBS (Ours)	42	73.38

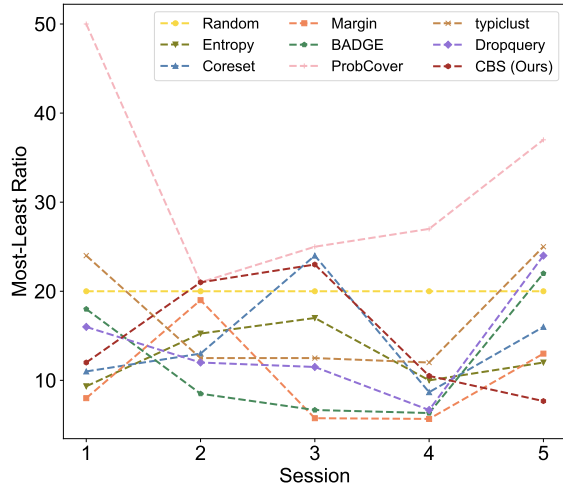


Figure E: Comparison of CBS and other counterparts applied to LP-DiF in terms of “class-imbalanced ratio” on CIFAR-100-LT under $B = 100$.

consistent predictions from the model for each input by using inputs from different views (by dropout the value of features), and retains samples with poor consistency for clustering. Next, it selects the samples closest to the center from each cluster. Unlike CBS, Dropquery still adopts a multi-round training-labeling paradigm, which may increase the computational cost of selecting samples.

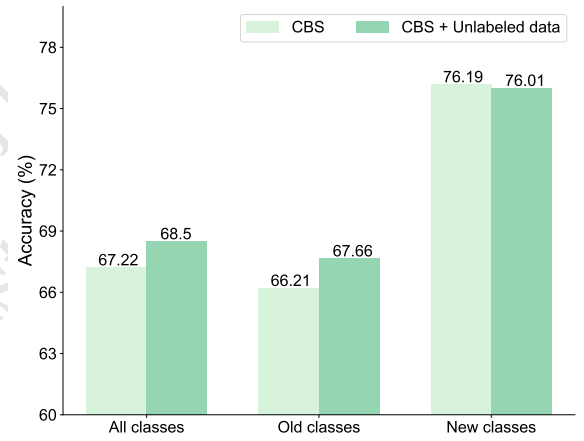


Figure F: Decoupling the performance of the last session to old classes and new classes respectively. The experiments are conducted on CUB-200 under $B = 100$.

Tab. ?? compares the runtime cost of samples selection of each session and Avg between our CBS and dropquery. Clearly, compared to Dropquery, our method has a lower runtime and achieves higher performance. This indicates that our method not only achieves high performance but is also more efficient.

Further analysis of utilizing unlabeled data for LP-DiF. When applying CBS to LP-DiF, we further exploit the unlabeled data not selected by CBS to improve the estimation method for the feature-level Gaussian distribution, which can generate higher-quality pseudo features for knowledge replay to enhance the model’s resistance to catastrophic forgetting. To more clearly demonstrate the effect of this design, we decouple the model’s accuracy in the last incremental session into accuracy on old classes and accuracy on new classes. Fig. F shows the decoupled results on the last session of CUB-200 under $B = 100$. “CBS + Unlabeled data” represent

utilizing unlabeled data to enhance the model’s resistance to catastrophic forgetting. Note that CBS + unlabeled performs better on all classes and old classes than pure CBS, *i.e.*, 67.22% vs. **68.5%**, and 66.21% vs. **67.66%** for old classes, while performance on new classes remains comparable. This fully reveals that utilizing unlabeled data can indeed enhance the model’s ability to resist catastrophic forgetting and improve overall performance.

E LIMITATION

In this paper, we introduce the task of active class incremental learning, which incorporates the idea of active sample selection into each incremental session of incremental learning to benefit incremental learner. In setting up the problem, we reference existing class incremental learning methods to establish the task of active class incremental learning, where the class space in each session has no overlap. However, in real-world applications, the requirement that new unlabeled data does not contain old classes is somewhat challenging to fulfill. Therefore, in future work, we may explore how to select the most informative samples from unlabeled data that may contain old classes.

Unpublished working draft.
Not for distribution.

REFERENCES

- [1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*.
- [2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3606–3613.
- [3] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets. In *International Conference on Machine Learning*. PMLR, 8175–8195.
- [4] Alex Holub, Pietro Perona, and Michael C Burl. 2008. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–8.
- [5] Zitong Huang, Ze Chen, Zhixing Chen, Erjin Zhou, Xinxing Xu, Rick Siow Mong Goh, Yong Liu, Chunmei Feng, and Wangmeng Zuo. 2024. Learning Prompt with Distribution-Based Feature Replay for Few-Shot Class-Incremental Learning. *arXiv preprint arXiv:2401.01598* (2024).
- [6] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [7] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 722–729.
- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [9] Sanket Rajan Gupta, Josiah Aklilu, Jeffrey J Nirschl, and Serena Yeung-Levy. 2024. Revisiting Active Learning in the Era of Vision Foundation Models. *arXiv e-prints* (2024), arXiv–2401.
- [10] Dan Roth and Kevin Small. 2006. Margin-based active learning for structured output spaces. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18–22, 2006 Proceedings* 17. Springer, 413–424.
- [11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.
- [12] Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.
- [13] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).
- [14] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. 2022. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*. Springer, 631–648.
- [15] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 139–149.
- [16] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. 2023. Learning Imbalanced Data with Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15793–15803.
- [17] Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. 2022. Active learning through a covering lens. *Advances in Neural Information Processing Systems* 35 (2022), 22354–22367.