Appendix of Bootstrap Your Own Prior: Towards Distribution-Agnostic Novel Class Discovery

This is the Appendix of the CVPR 2023 paper titled *Bootstrap Your Own Prior: Towards Distribution-Agnostic Novel Class Discovery*, which covers the following two parts: (1) implementation details, and (2) additional results.

A. Implementation Details

A.1. Pseudo-label Generation

As discussed in the main text, we use the Sinkhorn-Knopp algorithm [3] to solve the label assignment problem (Eqs. (1) and (2)). The solution can be written as

$$\mathbf{Y}^* = \mathrm{diag}(\boldsymbol{u})\mathbf{M}\mathrm{diag}(\boldsymbol{v})\,, \tag{A}$$

in which

$$\mathbf{M} = \exp\left(\frac{\mathbf{W}^\top\mathbf{X}}{\epsilon}\right)\,, \tag{B}$$

and $\boldsymbol{u} \in \mathbb{R}^{C^n}, \boldsymbol{v} \in \mathbb{R}^B$ are renormalization vectors to make $\mathbf{Y}^*$ a probability matrix, updated by simple matrix scaling iterations [3]:

$$\forall c : u_c \leftarrow [\mathbf{M}\boldsymbol{v}]_c^{-1}\,, \quad \forall b : v_b \leftarrow \left[\boldsymbol{u}^\top\mathbf{M}\right]_b^{-1}\,, \tag{C}$$

and we follow [2] to use a small iteration number of 3. In particular, we integrate the estimated class prior $\boldsymbol{p}$ into the iterations in Eq. (C) so that the generated label assignment $\mathbf{Y}^*$ can better reflect the true class distribution.

It is worth noting that, unlike base classes, the order of novel classes are undefined in NCD [9], which means that the generated $\mathbf{Y}^*$ in different iterations are not guaranteed to contain exactly same-order assignments. In other words, directly applying the estimated prior $\boldsymbol{p}$ can be an overstrict constraint since it implicitly forces the class order recorded in $\boldsymbol{p}$. Thus, before integrating $\boldsymbol{p}$ into Eq. (C), we reorder its elements such that it aligns with the predicted distributions in $\mathbf{M}$, *i.e.*, to ensure a consistent per-class constraint.

A.2. Network Details

We follow former works [4, 11] to use the same network architectures. In particular, besides the same image encoder (*i.e.*, ResNet-18 [7]), we also use the same base/novel head architecture. For the base head $h(\cdot)$, we use an ℓ_2-normalized linear layer with C^b output neurons; the novel head $g(\cdot)$ projects the 512-dimensional image features to the 256-dimensional ones with 2048 hidden units, followed by an ℓ_2-normalized linear layer with C^n output neurons.

We also follow [4, 11] to use multi-head clustering [1, 8] to smooth down possible clustering degeneration. We apply this strategy to the novel head $g(\cdot)$ with a head number of 4, *i.e.*, duplicating four novel heads and iterating over these heads in training. Note that we only report the averaged clustering accuracy of the four novel heads in our experiments. Similar to [4, 11], we also use the swapped prediction training strategy [2] to encourage the prediction consistency between different augmentations of a same input. Specifically, given an input image, we first generate its two augmentations $\boldsymbol{a}_1, \boldsymbol{a}_2$. The model will accordingly output two predictions whose labels are $\boldsymbol{y}_1, \boldsymbol{y}_2$, respectively. Then we train the model by minimizing $\mathcal{L}(\boldsymbol{a}_1, \boldsymbol{y}_2) + \mathcal{L}(\boldsymbol{a}_2, \boldsymbol{y}_1)$ (see Eq. (3)), in which the labels are swapped.

For network optimization, we also follow [4, 11], *i.e.*, using the SGD optimizer with momentum 0.9, linear warm-up in the first 10 epochs (200 epochs in total), cosine annealing learning rate (0.2 base, 0.001 min), a weight decay rate of 1.5×10^{-4}, and a batch size of 256, with standard data augmentations (moderate random crop, flip, jittering, and grey-scale).

B. Additional Results

B.1. Results on More Imbalance Ratios

We report in Tabs. A, B, D and E the results on more imbalance ratios, *i.e.*, 50 and 1 (balanced). Our proposed BYOP consistently improves the performance of original UNO [4] and ComEx [11], demonstrating its effectiveness across various distribution scenarios. It is surprising to see the non-trivial improvements under the conventionally balanced class distribution, validating the versatility of our proposed class prior estimation and dynamic temperature technique, which constantly encourage more accurate predictions under different class distributions.

Dataset →	CIFAR10 *(imbalance ratio: 50)*							CIFAR10 *(balanced)*						
Protocol →	Trad.	Task-aware			Task-agnostic			Trad.	Task-aware			Task-agnostic		
Method ↓	Nov.	Base	Nov.	All	Base	Nov.	All	Nov.	Base	Nov.	All	Base	Nov.	All
RS [5]	46.7	75.4	45.1	60.3	–	–	–	91.6	95.1	91.3	93.2	–	–	–
RS+ [5]	46.4	64.3	43.8	54.1	64.3	51.7	58.0	92.3	92.7	92.0	92.4	92.7	86.1	89.4
NCL [12]	55.0	73.5	48.2	60.9	–	–	–	93.2	94.6	92.5	93.6	–	–	–
UNO [4]	45.4	76.5	48.0	62.3	62.4	49.3	55.9	92.5	97.0	93.1	95.1	93.9	90.9	92.4
UNO + BYOP	59.2	76.8	49.9	63.4	62.8	51.6	57.2	94.8	**97.0**	**95.3**	**96.2**	94.0	**93.5**	93.8
ComEx [11]	49.1	77.0	51.9	64.5	65.6	53.0	59.3	95.8	96.5	92.7	94.6	95.1	92.3	93.7
ComEx + BYOP	**60.8**	**77.6**	**55.2**	**66.4**	**66.0**	**57.4**	**61.7**	**95.9**	96.8	93.0	94.9	**95.2**	92.5	**93.9**

Table A. Performance on CIFAR10 with different imbalance ratios. Results are reported in averaged top-1 classification/clustering accuracy (%). "Trad." means "Traditional", and "Nov." means "Novel". **Best** and second-best results are highlighted in each column.

Dataset →	CIFAR100-20 *(imbalance ratio: 50)*							CIFAR100-20 *(balanced)*						
Protocol →	Trad.	Task-aware			Task-agnostic			Trad.	Task-aware			Task-agnostic		
Method ↓	Nov.	Base	Nov.	All	Base	Nov.	All	Nov.	Base	Nov.	All	Base	Nov.	All
RS [5]	39.6	46.6	**39.1**	45.1	–	–	–	72.9	74.5	72.2	74.0	–	–	–
RS+ [5]	38.6	44.4	37.7	43.1	44.4	28.2	41.2	71.9	71.7	69.2	71.2	71.7	58.4	69.0
NCL [12]	40.4	46.2	35.3	44.0	–	–	–	**86.2**	73.9	**83.3**	75.8	–	–	–
UNO [4]	37.0	48.8	33.8	45.8	46.0	29.1	42.6	78.0	76.3	76.3	76.3	74.2	66.1	72.6
UNO + BYOP	47.0	49.6	36.9	47.1	46.6	33.6	44.0	79.8	77.0	78.3	77.3	74.6	68.1	73.3
ComEx [11]	40.2	49.6	35.8	46.8	47.2	34.2	44.6	79.1	78.1	76.3	77.7	76.6	74.1	76.1
ComEx + BYOP	**52.4**	**50.5**	36.2	**47.6**	**48.0**	**35.1**	**45.4**	79.7	**78.3**	77.3	**78.1**	**77.3**	**75.5**	**76.9**

Table B. Performance on CIFAR100-20 with different imbalance ratios. Results are reported in averaged top-1 classification/clustering accuracy (%). "Trad." means "Traditional", and "Nov." means "Novel". **Best** and second-best results are highlighted in each column.

Used C^n →	*Estimated*			*Truth*		
Imbalance ratio →	100	10	1	100	10	1
k-means	30.6	38.1	43.2	32.3	38.8	44.7
DTC [6]	39.9	43.9	46.9	43.0	44.7	48.5
UNO	34.3	45.7	76.9	35.2	46.9	78.0
UNO + BYOP	48.5	53.0	78.4	50.3	54.5	79.8

Table C. Clustering accuracy (%) on CIFAR100-20.

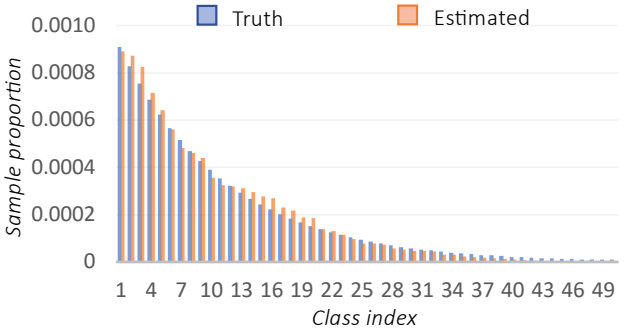

Figure A. True class distribution *vs.* estimated class distribution of the 50 novel classes in CIFAR100-50 with imbalance ratio 100.

B.2. Results with Unknown C^n

To isolate the effect of class distribution priors, and also for fair comparisons, we follow most of the recent works to assume the number of novel classes (C^n) to be known a priori. Given the fact that such knowledge is not always available in practical scenarios, we further experiment with an unknown C^n on CIFAR100-20.

Firstly, we estimate C^n using semi-supervised k-means introduced in [6], with 60 classes for feature pretraining, 20 classes in the probe set, and 20 classes in the novel set. We obtain the estimation $C^n = \{22, 22, 18\}$ for imbalance ratio in $\{100, 10, 1\}$, respectively. Then we rerun the experiments with the estimated C^n, and the results are as shown in Tab. C, where we also report the results using the true C^n for reference. Our method shows consistent improvement over the baselines. Although the estimated C^n seems reasonable even under high data imbalance, an accurate estimation on the large-scale imbalanced data remains an open challenge, which will be our future work.

B.3. More Qualitative Results

Estimated Class Prior. We show in Fig. A the comparison between the true class distribution and the estimated class distribution using BYOP in a random run. We can observe that BYOP achieves promising results in class prior estimation, providing a better clustering constraint compared with the conventionally uniform prior, which plays a critical role when data is highly imbalanced.

Dataset →	CIFAR100-50 *(imbalance ratio: 50)*							CIFAR100-50 *(balanced)*						
Protocol →	Trad.	Task-aware			Task-agnostic			Trad.	Task-aware			Task-agnostic		
Method ↓	Nov.	Base	Nov.	All	Base	Nov.	All	Nov.	Base	Nov.	All	Base	Nov.	All
RS [5]	32.0	48.7	25.4	37.1	–	–	–	46.8	79.2	46.5	62.9	–	–	–
RS+ [5]	28.0	38.9	23.2	31.1	38.9	22.3	30.6	**52.5**	70.7	**51.8**	61.3	70.7	49.1	59.9
NCL [12]	31.0	47.4	24.1	35.8	–	–	–	52.1	77.8	50.8	64.3	–	–	–
UNO [4]	26.8	50.3	25.1	37.7	41.9	24.7	33.3	48.5	79.2	47.8	63.5	71.2	45.2	58.2
UNO + BYOP	32.0	50.8	26.0	38.4	42.3	26.2	34.3	51.2	79.9	51.1	65.5	71.5	48.8	60.2
ComEx [11]	27.9	51.7	25.1	38.4	45.0	24.8	34.9	51.0	79.2	50.7	65.0	74.3	50.1	62.2
ComEx + BYOP	**33.6**	**52.0**	**26.3**	**39.2**	**45.9**	**26.5**	**36.2**	52.2	**80.3**	51.0	**65.7**	**76.8**	**50.7**	**63.8**

Table D. Performance on CIFAR100-50 with different imbalance ratios. Results are reported in averaged top-1 classification/clustering accuracy (%). "Trad." means "Traditional", and "Nov." means "Novel". **Best** and second-best results are highlighted in each column.

Dataset →	Tiny-ImageNet *(imbalance ratio: 50)*							Tiny-ImageNet *(balanced)*						
Protocol →	Trad.	Task-aware			Task-agnostic			Trad.	Task-aware			Task-agnostic		
Method ↓	Nov.	Base	Nov.	All	Base	Nov.	All	Nov.	Base	Nov.	All	Base	Nov.	All
RS [5]	24.4	**42.7**	15.1	28.9	–	–	–	17.3	64.7	17.8	41.3	–	–	–
RS+ [5]	23.7	30.3	15.7	23.0	30.3	16.3	23.3	20.9	52.9	21.1	37.0	52.9	22.4	37.7
NCL [12]	23.6	38.1	14.6	26.4	–	–	–	23.1	63.0	20.5	41.8	–	–	–
UNO [4]	21.6	39.5	17.8	28.7	29.5	17.1	23.3	34.0	68.3	34.3	51.3	55.7	32.8	44.3
UNO + BYOP	26.0	40.5	18.2	29.4	29.7	18.4	24.1	35.8	**68.7**	35.7	52.2	56.8	34.4	45.6
ComEx [11]	22.7	41.5	18.8	30.2	34.3	18.7	26.5	36.9	67.7	36.8	52.3	59.6	36.7	48.2
ComEx + BYOP	**26.1**	41.8	**18.8**	**30.3**	**35.3**	**19.1**	**27.2**	**37.4**	67.8	**37.6**	**52.7**	**61.0**	**37.6**	**49.3**

Table E. Performance on Tiny-ImageNet with different imbalance ratios. Results are reported in averaged top-1 classification/clustering accuracy (%). "Trad." means "Traditional", and "Nov." means "Novel". **Best** and second-best results are highlighted in each column.

t-SNE Visualizations. We provide visualizations in a more challenging scenario, *i.e.*, CIFAR10 with imbalance ratio 100. As shown in Fig. B(a), with a uniform class prior in training, UNO still tends to equally partition the samples, resulting in inferior performance, *e.g.*, wrongly partitioning majority class `dog` into different clusters. In contrast, benefiting from the estimated class prior, BYOP does a better job in preserving the true data structure. As shown in Fig. B(b), the majority class `dog` spans the largest space due to the significant data imbalance, which may, however, lead to indistinguishable decision boundaries between classes. This implies that distribution-agnostic NCD remains an open challenge in practice, especially under high data imbalance.

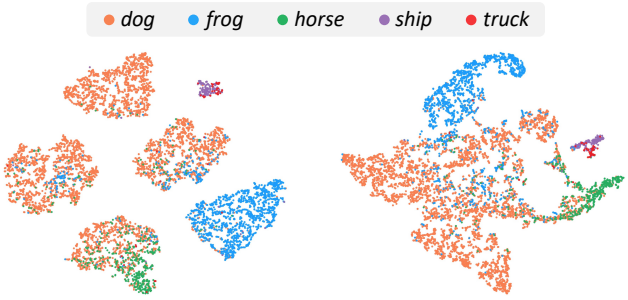

Figure B. t-SNE [10] visualizations of the five novel classes in the training split of CIFAR10 with imbalance ratio 100. **(a)** Output space of UNO [4]. **(b)** Output space of UNO + BYOP.

B.4. Parameter Analysis

Recall that we estimate the class distribution prior based on the model predictions stored in a first-in-first-out queue $\mathcal{K}$. Here we analyze the effect of different queue sizes. As shown in Fig. C, a small queue size may not be representative enough for the data distribution, yet a moderate queue size of 6000 is adequate to achieve promising results. It is interesting that larger queue sizes may not always lead to better performance. One possible reason can be the slower updating phenomenon—it takes more iterations to update the overall class distribution in a larger queue, which may delay the adaption of estimated class prior. Thus, we stick to the moderate queue size of 6000 in our experiments for both estimation accuracy and computation efficiency.

References

[1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proc. ECCV*, pages 139–156, 2018. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learn-

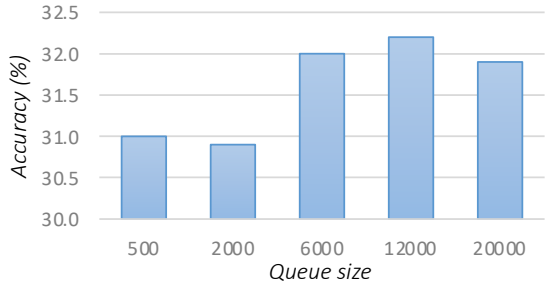

Figure C. Effect of queue size in BYOP. Results are reported in clustering accuracy (%) in the training split of CIFAR100-50 with imbalance ratio 50.

ing of visual features by contrasting cluster assignments. In *Proc. NeurIPS*, pages 9912–9924, 2020. 1

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. NeurIPS*, pages 2292–2300, 2013. 1

[4] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proc. ICCV*, pages 9284–9292, 2021. 1, 2, 3

[5] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proc. ICLR*, 2020. 2, 3

[6] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proc. ICCV*, pages 8401–8409, 2019. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 1

[8] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proc. ICCV*, pages 9864–9873, 2019. 1

[9] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *Proc. ECCV*, pages 437–455, 2022. 1

[10] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.*, 15(1):3221–3245, 2014. 3

[11] Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu, and Cheng Deng. Divide and conquer: Compositional experts for generalized novel class discovery. In *Proc. CVPR*, pages 14268–14277, 2022. 1, 2, 3

[12] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proc. CVPR*, pages 10867–10875, 2021. 2, 3