A Different self-supervised representations

The self-supervised representation provides us the information to partition the training data into k expert sub-datasets, so we analyze the performance of our method by fine-tuning different pretrained representations of other self-supervised ViT models, *i.e.* MoCo v3 [2] and MAE [4]. We initialize the ViT-B-16 model [3] with the parameters pretrained on ImageNet-1k by MoCo v3 for 300 epochs and by MAE for 800 epochs respectively. The result in Table 1 shows that with the self-supervised representation of DINO [1], our method performs $7.3-20.7\%$ better than the other two on CUB-200 and $1.8-22.4\%$ better on Stanford-Cars. We observe that DINO still shows the best performance on clustering the data based on class-irrelevant informations.

Table 1: Results with different self-supervised representations.

self-supervised ViT model	CUB-200			Stanford-Cars		
	All	Old	New	All	Old	New
DINO	**51.8**	**53.8**	**50.8**	**41.0**	**59.1**	**32.2**
MoCo v3	37.6	42.8	35.1	24.7	36.7	18.9
MAE	35.5	46.5	30.1	38.7	56.0	30.4

B Estimating the number of classes

As a more realistic scenario, the prior knowledge of the number of classes is unknown in the GCD. We follow the method in [5] to estimate the number of classes in the unlabeled dataset by leveraging the information of the labeled dataset. We compare our estimated number of classes in unlabeled data $\left|\hat{\mathcal{C}}^u\right|$ with the ground truth number of classes in unlabeled data $\left|\mathcal{C}^u\right|$ in Table 2. We find that on Stanford-Cars and FGVC-Aircraft, the number of classes estimated by our method is significantly closer to the ground truth compared with GCD [5]. Our method tends to show better performance on fine-grained datasets, given that the dataset partitioning can help the model learn more discriminative features when facing the more challenging datasets that have little obvious difference.

Table 2: Estimation of the number of classes in unlabeled data.

	CIFAR10	CIFAR100	ImageNet-100	CUB-200	Stanford-Cars	FGVC-Aircraft	Oxford-Pet
Ground truth	10	100	100	200	196	100	37
GCD [5]	9	100	109	231	230	80	34
XCon	8	97	109	236	206	101	34

C Performance with estimated class number

We use the class number estimated in Table 2 to evaluate our method, displaying the performance of our method when the unlabeled class number is unavailable. We report the results on generic image classification benchmarks in Table 3 and the results on fine-grained image classification benchmarks in Table 4. With our estimated class number $\left|\hat{\mathcal{C}}^u\right|$, our method performs better on Stanford-Cars and also reaches comparable results on the other five datasets except CIFAR10, which shows that our method is also promising under the more realistic condition.

D Ablation on contrastive fine-tuning

We further ablate the components of contrastive loss in Table 5. We find that only with unsupervised contrastive loss, *i.e.* $\lambda = 0$, the *ACC* drops $21.5-23.6\%$ on CUB-200 and $22.2-46.6\%$ on Stanford-Cars, which means the combination of supervised contrastive loss

Table 3: Results on generic datasets with our estimated class number.

known C^u	CIFAR10			CIFAR100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New
✓	**96.0**	97.3	**95.4**	**74.2**	**81.2**	**60.3**	**77.6**	**93.5**	**69.7**
✗	70.1	**97.4**	56.5	72.5	80.3	56.8	75.6	91.5	67.6

Table 4: Results on fine-grained datasets with our estimated class number.

known C^u	CUB-200			Stanford-Cars			FGVC-Aircraft			Oxford-Pet		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
✓	**52.1**	54.3	**51.0**	40.5	58.8	31.7	**47.7**	44.4	**49.4**	**86.7**	**91.5**	**84.1**
✗	51.0	**57.8**	47.6	**41.3**	**58.8**	**32.8**	46.1	**47.6**	45.3	82.1	81.7	82.4

and unsupervised contrastive loss with the balanced parmeter $\lambda = 0.35$ is necessary and can reach the best performance.

Table 5: Ablation study of contrastive loss.

λ	CUB-200			Stanford-Cars		
	All	Old	New	All	Old	New
0	29.6	30.2	29.3	10.8	12.5	10.0
0.35	**51.8**	**53.8**	**50.8**	**41.0**	**59.1**	**32.2**

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[5] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022.