Class-relation Knowledge Distillation for Novel Class Discovery

Peiyan Gu[1,*] Chuyu Zhang[1,2,*] Ruijie Xu[1] Xuming He[1,3]

[1]ShanghaiTech University, Shanghai, China [2]Lingang Laboratory, Shanghai, China

[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging, Shanghai, China

{zhangchy2,gupy,xurj2022,hexm}@shanghaitech.edu.cn

A. Implementation Details

For the CIFAR100 dataset, we adopt the SGD optimizer and employ a learning rate schedule that initially increases from 0.001 to 0.4 within the first 10 epochs and then decreases to 0.001 at 500 epochs using a cosine annealing schedule. Our batch size is 512. We are re-implementing NCL [5] on the CIFAR100-50 setting using their publicly available code, and we cite all other results from their published work.

As for the three fine-grained datasets, we use the Adamw optimizer, and our learning rate scheduling involves an initial increase from 0.0001 to 0.001 within ten epochs, followed by a decrease to 0.0001 at 100 epochs using a cosine annealing schedule. We utilize a batch size of 512 for all methods and reimplement the results of GCD [4] using the code they provided.

To enhance the performance of our clustering approach and for a fair comparison, we also employed a multi-head technique similar to UNO. We used four heads for the CIFAR100 dataset and two heads for the remaining three fine-grained datasets. Our novel class head includes a Multilayer Perceptron (MLP) and a cosine classifier.

Moreover, we determine the hyperparameter β through the validation set on the known class.

B. More visualization

To demonstrate the efficacy of our model, we selected the five representative novel classes and analyzed their relationships with known classes. The top four classes were selected based on their similarity to known classes, ordered from high to low, while the last class "wardrobe," was selected from special novel classes that will often appear in images together with some known classes. As shown in Fig.1, in many cases, the distributions generated by the supervised trained model have strong semantic information. Furthermore, the plot indicates that our model can better maintain relations between the novel and known classes than the baseline model(UNO). What's more, we analyze on the fine-grained datasets. As shown in figure, our model can inherit the relationship between the "Spitfire" class and known classes captured by the supervised trained model, while the baseline model loses this ability. Specifically, our model can capture the common characteristics of the Spitfire, C-47, C-130, and Cessna 208, such as propellers and forward wings. It also recognizes the unique color present in both the Spitfire and C-47. In contrast, the baseline model basically regards all known classes as the same. This shows that our model can well capture the potential relationship between novel classes and known classes on fine-grained datasets.

In addition, we present a comparative analysis of the accuracy of our model and the baseline model for each novel class. Fig.3 demonstrates that our model's predictions are more accurate than the baseline model in almost all classes.

C. More experiments without pre-trained model

We conduct experiments on fine-grained datasets with ResNet18 from scratch. As the Tab.1 shows, we still achieve sizeable improvement over existing methods, 2.1% on Stanford Cars, 1.8% on CUB, and 6.0% on Aircraft. This demonstrates that our method is also effective without pre-trained models.

Table 1. Pre-train ResNet18 on known classes, and then train on known+novel classes. Both stages are trained for 200 epochs.

Method	Stanford Cars	CUB	Aircraft
RankStats+	17.7	20.2	30.1
NCL	27.1	24.7	36.7
UNO	25.2	24.5	37.8
Ours	**29.2**	**26.5**	**43.8**

D. Temperature hyperparameter T analysis

We conducted an in-depth analysis of the temperature T as presented in S2. The results indicate that temperatures

*Both authors contributed equally.

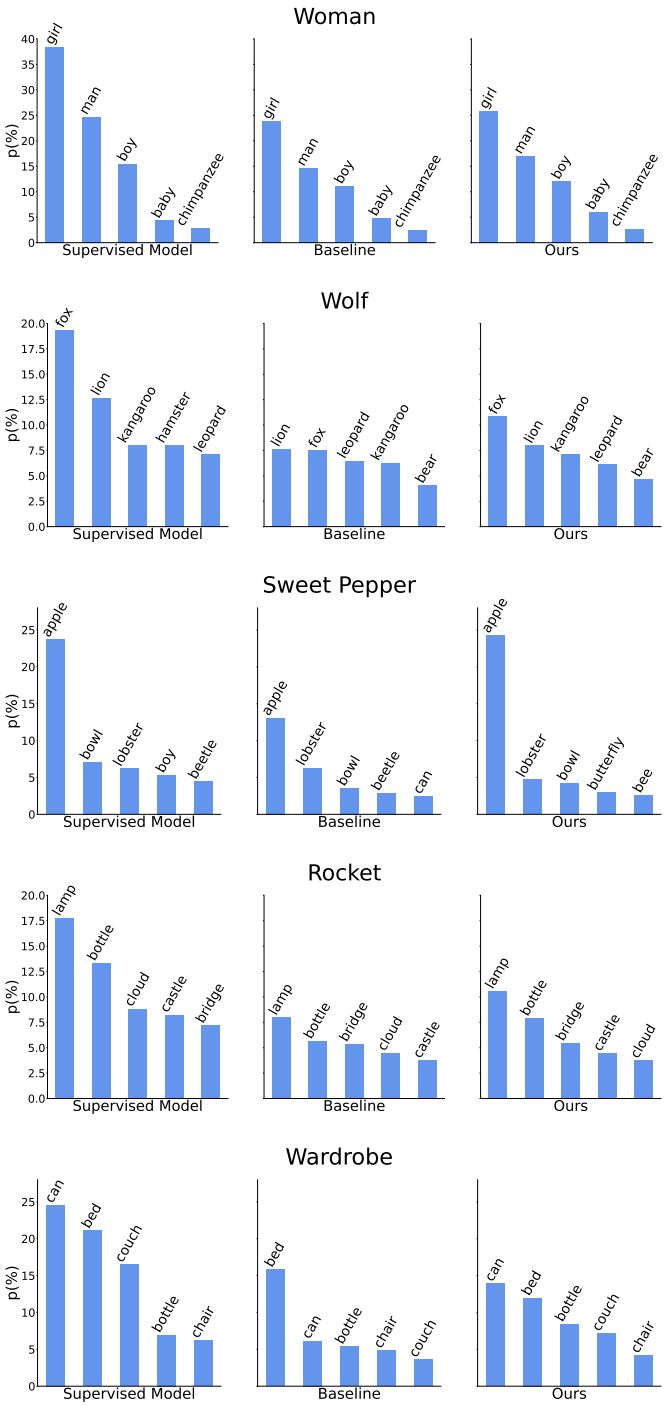

Figure 1. Visualization of quantified relative relationships. The bar labels represent the known classes in the CIFAR100-50 setting. Each plot shows average predictions for instances of a novel class on the known class head. In most cases, our model's predictions are more similar to the supervised trained model's predictions than the baseline model's predictions.

1, 2, 4, and 6 yield satisfactory performance, demonstrating the robustness of our model with respect to the temperature hyperparameter. Taking into consideration the established practices in knowledge distillation [1, 3], where the temperature is often set to 4, we have chosen this value for our model.

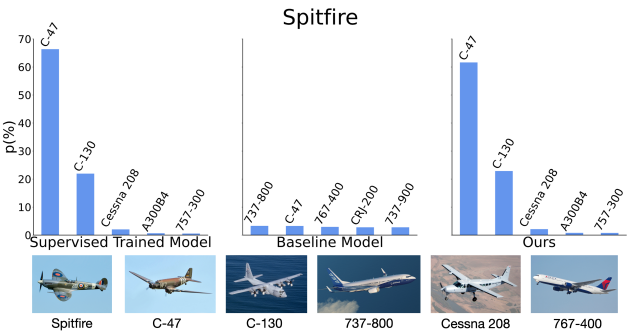

Figure 2. Visualization of relative relationships on Aircraft.

Table 2. Results on CIFAR100-50 with different temperature.

Temperature	1	2	4	6	8
Novel Acc	65.4	66.8	65.3	66.2	61.6

E. Learnable weight function

In the ablation study, we analyze various designs of our learnable weight function and demonstrate the superiority of our approach. However, the theoretically learnable weight function may have a degenerate solution, where the maximum weight is assigned to the sample with the smallest KL. Achieving this degenerate solution in practice is challenging due to the random selection of samples in each batch. Additionally, as shown in Fig.4, the mean statistics of eta remain relatively stable when the batch size is large.

F. Discussion with NCDwF [2]

In NCDwF, they focus on novel class discovery without forgetting, where known class data is not available in the discovery stage. To transfer knowledge, they introduce a mutual information regularization term for novel classes to couple the learning of labeled head to unlabeled head and expect to transfer semantic knowledge from known classes to novel classes. Meanwhile, known and novel classes still share a feature extractor. Differently, we transfer knowledge from a known classes pretrained model to a discovery trained model and expect the discovery trained model to maintain meaningful class relations. What's more, we also develop a simple and effective learnable weight function, which adaptively promotes knowledge transfer based on the semantic similarity between the novel and known classes. In addition, the outstanding results on a challenging dataset in Tab.3 show the superiority of our method. In conclusion, our method is totally different from NCDwF.

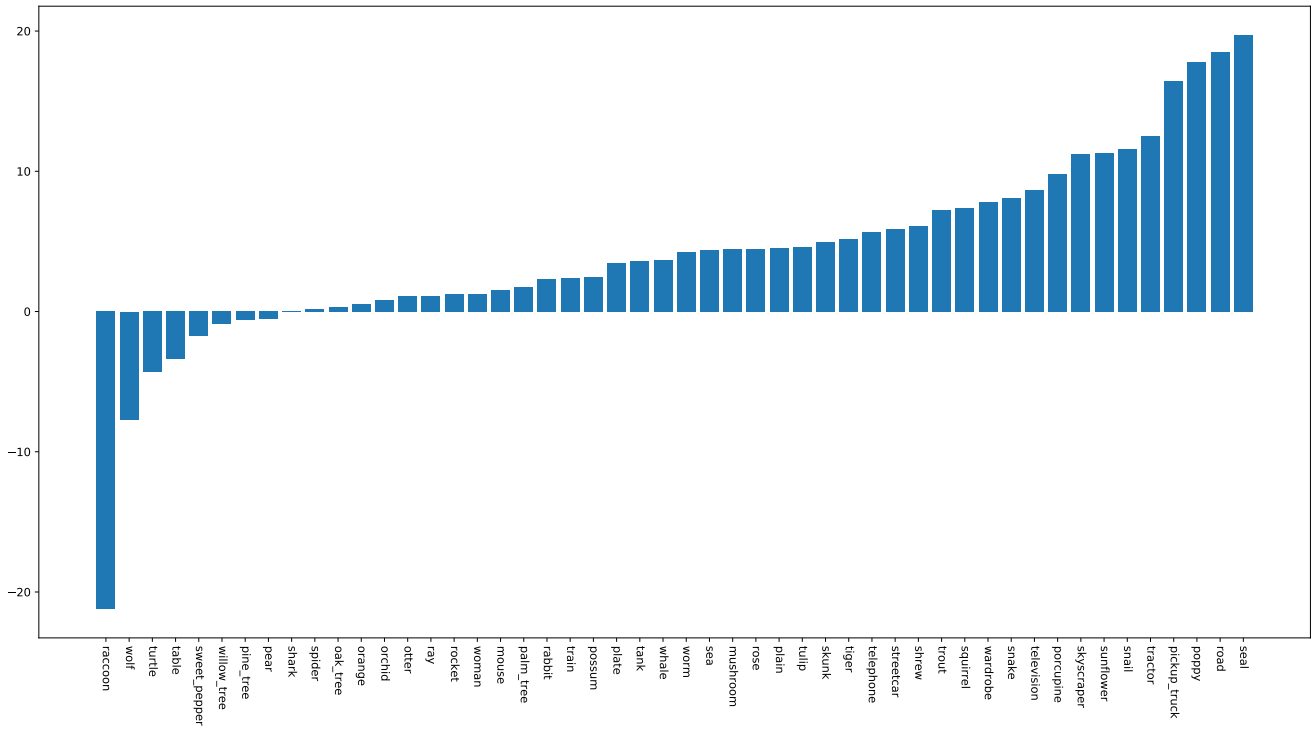

Figure 3. The difference between the accuracy of our model and the UNO model on each novel class.

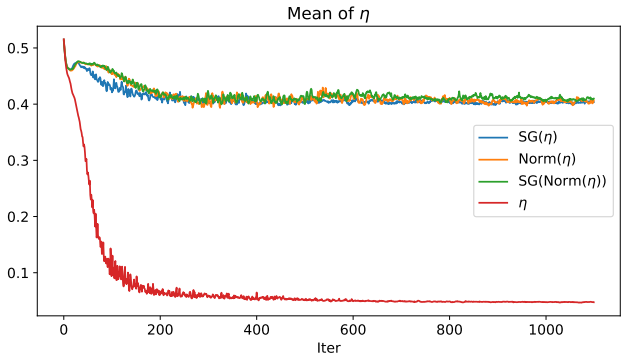

Figure 4. The mean of η for different weight function $g(\eta)$.

Method	CIFAR100-80	CIFAR100-50
UNO	90.4	60.4
NCDwF	91.3	61.2
Ours	91.2	65.3

Table 3. The results on CIFAR100 dataset under the same setting.

References

[1] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[2] KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 570–586. Springer, 2022.

[3] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *International Conference on Learning Representations*, 2020.

[4] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. *arXiv preprint arXiv:2201.02609*, 2022.

[5] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10875, 2021.