Improvement of Self-Organizing Classifier for Generalized Category Discovery

Rintaro Akashi
Advanced Engineering Course
National Institute of Technology, Kurume College
Fukuoka, Japan
a4101ra@kurume.kosen-ac.jp

Kousuke Matsushima
Department of Control and Information Systems Engineering
National Institute of Technology, Kurume College
Fukuoka, Japan
matsushima@kurume-nct.ac.jp

***Abstract*—Generalized Category Discovery is one of the most practical image classification tasks. It utilizes both labeled and unlabeled data for training, classifying both known and unknown data. However, previous research did not assume that the total number of classes within the dataset is unknown, despite this setting. This led to a contradiction: while unknown data is included, the number of classes must be known. To resolve this, we previously proposed a self-organizing classifier using prototype merging, achieving high accuracy even under conditions where the number of classes is unknown. Nevertheless, this approach suffered from reduced accuracy for known class samples. This paper proposes an improved method that utilizes a novel metric, prototype responsibility, in the merging computation. The result confirms improved accuracy for known classes compared to existing methods. We also discuss remaining challenges for this research.**

***Index Terms*—Image Recognition, Image Classification, Self-Supervised Learning, Driver-Assistance, Road Damage Detection, Self-Organizing Classifier**

I. Introduction

The growing demand for intelligent perception in autonomous driving and infrastructure management has accelerated advancements in computer vision. Deep learning, in particular, has enabled models to achieve near human-level accuracy in visual classification and recognition tasks. These achievements have facilitated real-world applications such as road condition monitoring and advanced driver-assistance systems. However, most of these successes rely heavily on supervised learning, which requires extensive amounts of annotated data. The high cost of data labeling limits the scalability of such approaches, while models trained in closed-set conditions struggle to recognize previously unseen categories, reducing their robustness in dynamic environments. To improve the efficiency of infrastructure maintenance, recent technologies have integrated on-board cameras and LiDAR sensors with 3D digital maps to update spatial information in real time. Mobile Mapping Systems (MMS) can capture accurate 3D data but are costly and impractical for frequent deployment. A more efficient strategy is to exploit pre-acquired point clouds and camera imagery using automatic classification techniques. Nevertheless, purely supervised approaches cannot adapt to the diversity of unlabeled data or the continuous emergence of new visual patterns typically encountered in real-world road scenes. These challenges have motivated the development of semi-supervised learning, which leverages both labeled and unlabeled data for training. Among such frameworks, Generalized Category Discovery (GCD) [1] has emerged as a promising task setting that simultaneously classifies known and unknown categories within a dataset. GCD assumes partial label availability and aims to discover novel classes while preserving recognition accuracy for known ones. Vaze et al. [1] first established a strong baseline by combining DINO-pretrained Vision Transformers [2] with feature-space clustering, while Wen et al. later proposed SimGCD [3], a parametric classification model incorporating contrastive learning and entropy regularization, achieving state-of-the-art performance with reduced computational complexity. Despite these advances, most existing GCD methods rely on the unrealistic assumption that the total number of categories is known in advance. Estimating this number often requires multiple clustering or retraining stages, which greatly increases computational overhead. Consequently, developing a model that can dynamically adapt to an unknown number of categories remains an open challenge. To address this issue, we proposed a self-organizing classifier in our previous work [4]. Our prior model automatically merges classifier prototypes during training. By optimizing the classifier's output dimension based on inter-prototype similarity, the proposed model maintains high classification accuracy without prior knowledge of the true number of categories. Experiments on benchmark datasets demonstrate that our method not only improves recognition of unknown categories but also enhances overall stability and efficiency compared with existing GCD approaches.

II. Related Works

A. Generalized Category Discovery

Generalized Category Discovery(GCD) is one of the semi-supervised learning tasks proposed by S. Vaze et al. [1]. Its key feature is the use of datasets containing both labeled and unlabeled images. Furthermore, the unlabeled images include not only those belonging to the same class as the

labeled samples but also those belonging to unknown classes. The goal is to classify all these samples. They employed a Vision Transformer pre-trained by DINO [2], fine-tuned via contrastive learning [5], as the backbone. Pre-training with DINO has been shown to help form distinct clusters in the feature space [2]. Applying clustering in this feature space prevents overfitting to labeled known samples while achieving high accuracy on unseen samples. S. Vaze et al. demonstrated a heuristic method is effective for deciding an optimal number of classes: they run clustering multiple times and adopted the number of classes that maximize the accuracy on labeled data [1]. However, performing clustering iterations repeatedly remains computationally challenging.

B. SimGCD

S. Vaze et al. avoided parametric classification, citing its tendency to overfit to known samples [1]. However, X. Wen et al. proposed SimGCD [3], a model employing parametric classification precisely because of its computational advantages over clustering. Their method also utilizes a backbone pre-trained by DINO and combines training with frameworks such as contrastive learning [3] and self-distillation. Furthermore, to mitigate prediction bias, they introduced a mean entropy term to the loss function. This term encourages the model to provide uniform predictions across all categories. As a result, despite not using clustering, they achieved accuracy comparable to S. Vaze et al.'s method. However, SimGCD shares a common limitation as other GCD method:the requirement of the true number of classes as prior knowledge.

III. Proposed Method

Our proposed model is based on SimGCD, applying its core techniques: backbone representation learning, self-distillation using data augmentation, and maximizing average entropy. The core of our contribution is the self-organizing classifier algorithm. Figure 1 illustrates the model structure and learning flow. Our model consists of a backbone that converts images into feature representations and a classifier that converts these feature representations into class-specific probabilities. During training, an MLP that reduces the dimensionality of the feature representations is inserted after the backbone and used for representation learning.

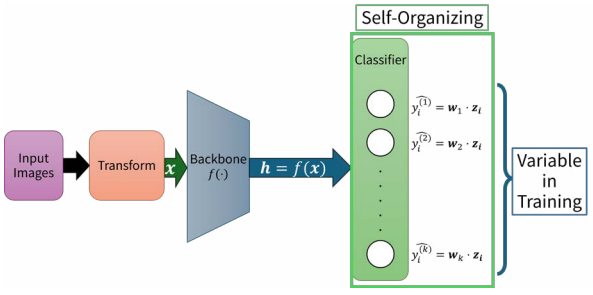

Fig. 1. Model structure of the proposed method

A. Representation Learning

Our method employs contrastive loss for representation learning, similar to SimGCD. We compute self-supervised contrastive loss for all samples and additionally apply supervised contrastive loss for labeled samples. Equation (1) denotes self-supervised contrastive loss, and Equation (2) denotes supervised contrastive loss.

$$L_{rep}^{u} = \frac{1}{|B|} \sum_{i \in B} -\log \frac{exp(\boldsymbol{z}_i^T \boldsymbol{z}_i'/\tau_u)}{\sum_i^{i \neq n} exp(\boldsymbol{z}_i^T \boldsymbol{z}_i'/\tau_u)} \tag{1}$$

$$L_{rep}^{s} = \frac{1}{|B^l|} \sum_{i \in B^l} \frac{1}{|N_i|} \sum_{q \in N_i} -\log \frac{exp(\boldsymbol{z}_i^T \boldsymbol{z}_i'/\tau_c)}{\sum_i^{i \neq n} exp(\boldsymbol{z}_i^T \boldsymbol{z}_i'/\tau_c)} \tag{2}$$

Here, $\boldsymbol{z}$ is the vector obtained by inputting the backbone feature representation into an MLP, and τ_u and τ_c are temperature parameters. Furthermore, B denotes the set of samples within a batch, and B_l denotes the subset of labeled samples within B. The former generates two distinct views $\boldsymbol{z}$ and $\boldsymbol{z}'$ from the same image and trains the model to make their feature representations similar. The latter utilizes labeled data and forces the model to obtain similar feature representations for data belonging to the same class.

$$L_{rep} = (1 - \lambda) L_{rep}^{u} + \lambda {L_{rep}}^{s} \tag{3}$$

B. Parametric Classification

The learning of parametric classification also follows the SimGCD approach, taking the form of self-distillation. For one of the two views, the logit generated from that view is used as a pseudo-label for training. Furthermore, for labeled samples, conventional supervised learning is also employed. Equation (4) shows the logit calculation formula for self-distillation.

$$\boldsymbol{p}_i^{(k)} = \frac{exp(\frac{1}{\tau_s}(\boldsymbol{h}_i/||\boldsymbol{h}_i||_2)^T)(\boldsymbol{c}_k/||\boldsymbol{c}_k||_2)}{\sum_{k'} exp(\frac{1}{\tau_s}(\boldsymbol{h}_i/||\boldsymbol{h}_i||_2)^T)(\boldsymbol{c}_{k'}/||\boldsymbol{c}_{k'}||_2)} \tag{4}$$

Here, $\boldsymbol{h}$ is the hidden feature output by the backbone, and $\boldsymbol{c} \in C$ represent the prototype vector and the set of the prototypes. Additionally, τ_s is the temperature parameter. Similarly, two logits $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ are generated from the two views derived from the same image. Since $\boldsymbol{q}_i$ is used as a soft pseudo-label for self-distillation, a smaller temperature parameter value than $\boldsymbol{p}_i$ is employed to encourage sharper predictions from the model. Furthermore, the temperature parameter is set smaller as learning progresses, starting from the early stages where the feature space is unstable and pseudo-label reliability is low. This ensures self-distillation yields more accurate predictions. Classification learning is performed using these logits via the standard cross-entropy loss:

$$l(\boldsymbol{q}_i^{(k)}, \boldsymbol{p}_i^{(k)}) = -\sum_k \boldsymbol{q}_i^{(k)} \log \boldsymbol{p}_i^{(k)} \tag{5}$$

For annotated data, common supervised learning is also performed using the one-hot labels $\boldsymbol{y}_i$, not just the pseudo

labels. Additionally, to mitigate model prediction bias, an average entropy term is introduced, encouraging the model to make uniform predictions even for unknown classes. Its form is shown in Equation (6). S. Vaze et al. demonstrated this loss enhances robustness against the number of unknown classes.

$$H(\bar{\boldsymbol{p}}) = -\sum_{k} \bar{\boldsymbol{p}}^{(k)} log \bar{\boldsymbol{p}}^{(k)} \tag{6}$$

The final classification objective function is expressed as follows.

$$L_{cls} = (1-\lambda) L_{cls}^{u} + \lambda {L_{cls}}^{s} \tag{7}$$

C. Output Dimension Optimization

The core contribution of our proposed method is achieved by this algorithm. The self-organizing classifier can become a robust model for unknown classes by merging the prototypes it holds based on their mutual similarity. When the classes contained within a dataset are unknown, we set an initial number of prototypes to be relatively large. We then optimize the model structure by performing a process that merges overly similar prototypes, which are a factor reducing accuracy. Our previous method [4] performed a merging process on pairs of prototypes whose similarity exceeded a threshold, combining them into their average vector. However, this approach risked merging prototype vectors of distinct known classes. Furthermore, it did not consider the relative influence each prototype has on predictions, potentially leading to inappropriate modifications of prototype vectors during merging. To address these limitations, we propose an enhanced algorithm that resolves these issues and improves accuracy. Our previous method [4] risked merging vectors representing known classes with each other or with other vectors. This could cause the classification capabilities acquired through supervised learning to be lost, potentially reducing accuracy. To prevent this, we imposed restrictions on the merging of prototypes representing known classes based on label information. Specifically, pairs of known classes are excluded from merging. When merging with prototypes outside known classes, the prototype not belonging to a known class is unilaterally invalidated. Furthermore, for merging unknown prototypes, a metric called responsibility is introduced, representing how many samples belong to each prototype. Responsibility η is calculated as follows:

$$\eta^{(k)} = \sum_{\boldsymbol{x} \in \mathcal{D}} \boldsymbol{p}(\boldsymbol{x})^{(k)} \tag{8}$$

The definition of responsibility follows a probabilistic interpretation similar to the E-step in the Expectation-Maximization (EM) algorithm. Summing over all samples yields a measure of prototype occupancy, which naturally reflects its representational strength within the feature space. Although our formulation does not perform a full EM optimization, it adopts this probabilistic intuition to stabilize prototype merging without additional iterations. To prevent numerical instability caused by scale differences among prototypes, the responsibility values are normalized within each merging operation as

$$\tilde{\eta}^{(k)} = \frac{\eta^{(k)}}{\sum_{j \in C_a} \eta^{(j)}} \tag{9}$$

where C_a denotes the set of prototypes under consideration. This normalization ensures consistent scaling of merging weights and prevents dominant prototypes from biasing the update. During merging, a new vector is calculated based on this using the following formula:

$$\boldsymbol{p}_{ij} = \begin{cases} \frac{\eta_i \boldsymbol{p}_i + \eta_j \boldsymbol{p}_j}{\eta_i + \eta j} & \text{if } \boldsymbol{x}_i \notin \mathcal{D}_l \text{ and } \boldsymbol{x}_j \notin \mathcal{D}_l, \\ \boldsymbol{p}_i & \text{if } \boldsymbol{x}_i \in \mathcal{D}_l \text{ and } \boldsymbol{x}_j \notin \mathcal{D}_l, \\ \boldsymbol{p}_j & \text{if } \boldsymbol{x}_i \notin \mathcal{D}_l \text{ and } \boldsymbol{x}_j \in \mathcal{D}_l. \end{cases} \tag{10}$$

This calculation ensures that prototypes with many samples—those possessing numerous feature vectors in their vicinity—are given greater weight during merging. This algorithm aims to improve accuracy while reducing computational load.

IV. Experiment

A. Experimental Setup

We conducted experiments on two datasets, CIFAR-10 and the Road Damage Dataset (RDD), to verify the effectiveness of the proposed method. CIFAR-10 is commonly used for performance evaluation in GCD. The Road Damage Dataset is primarily used for road damage detection tasks and consists of images depicting road cracks and manholes. This study conducted experiments on this dataset with the application of GCD in mind for fields such as optimizing infrastructure maintenance for roads and supporting safe driving. Details of the datasets are shown in Table I. For the experiments, the backbone used was RepViT [6] trained on ImageNet-1K [7]. Hyperparameters were set primarily following SimGCD. Training was conducted for 200 epochs with a batch size of 128. The learning rate decayed from 0.1 to 0 following a cosine annealing function. The pseudo-label temperature τ_s was linearly adjusted from 0.07 to 0.04 over 30 epochs. The parameter λ, determining the ratio of supervised to self-supervised learning loss, was set to 0.35. For the self-organizing classifier merging algorithm, the merging start epoch was set to 120, after sufficient learning progress. Prototype pairs with cosine similarity values exceeding 0.15 were

TABLE I
Details of the datasets

Dataset	Labelled		Unlabelled	
	Images	Classes	Images	Classes
CIFAR-10	12.5K	5	37.5K	10
RDD	21.3K	5	13.5K	7

targeted for merging. The merging threshold for cosine similarity was empirically set to 0.15. This value corresponds approximately to the inflection point of the similarity histogram (Fig. 2), where the inter-class and intra-class similarities begin to separate. Using this threshold provided stable merging behavior across multiple preliminary trials. Regarding the initial value for the output dimension, it was empirically set to 100 under the assumption that the dataset contents are completely unknown. This value provides sufficient capacity for the classifier to self-organize and has been validated as a stable initialization in preliminary experiments. However, the method itself does not depend on this particular number; the output dimension can be adjusted according to dataset complexity.

B. Evaluation

Model evaluation is primarily based on accuracy. As well as GCD and SimGCD, the Hungarian algorithm [8] solves the linear assignment problem between predicted labels and ground truth labels to calculate the accuracy. The accuracy is calculated by Equation 11.

$$ACC = \max_{p \in \mathcal{P}(\mathcal{Y}_\mathcal{U})} \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\{y_i = p(\hat{y}_i)\} \tag{11}$$

Additionally, to ensure reliable evaluation, we also calculated the F1 score, precision, and recall.

V. Results

The results of experiments for each dataset using this method are shown in Table II. To compare the effectiveness of the methods, Table II also shows the accuracy for SimGCD, our prior work [4]. Each represents scores: All for the entire dataset, Old for known samples, and New for unknown samples. Furthermore, the F1 score and mean precision are shown in Table III and Table IV. These results demonstrate that our proposed self-organizing classifier achieves higher accuracy than SimGCD when the number of classes in the dataset is unknown. Furthermore, Table II shows our method's significant improvement in accuracy compared to before introducing responsibilities and the deactivation of prototype merging for known classes in CIFAR-10. However, the accuracy of our novel method is lower than our prior method on RDD. Furthermore, comparing the F1 score and precision reveals that our proposed method yields lower values for unknown classes in Table III and Table IV. On the other hand, the high accuracy in Table II suggests that misclassifying known samples as unknown ones, or misclassifications between unknown classes, are having an impact. In other words, our method exhibits a greater tendency to predict unknown samples. Figure 3 shows the loss curve during training of the proposed method. Although occasional sharp increases occur even before prototype integration, the curve is generally decreasing.

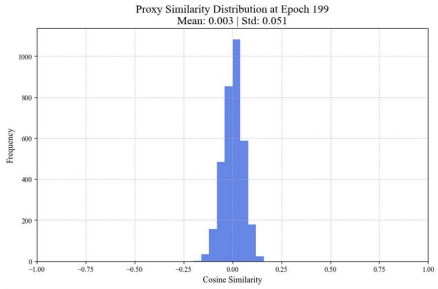

Fig. 2. Similarity histogram after learning

TABLE II
Accuracy on the datasets

	CIFAR-10			RDD		
	All	Old	New	All	Old	New
Proposed Method	82.16	87.20	77.12	84.20	86.88	74.90
Prior Method	74.99	75.00	74.98	85.68	88.02	77.55
SimGCD	73.04	76.22	69.86	78.59	79.19	76.50

TABLE III
F1 score on the datasets

	CIFAR-10			RDD		
	All	Old	New	All	Old	New
Proposed Method	79.09	91.75	66.43	81.43	88.52	63.71
Prior Method	74.08	82.75	65.40	83.09	88.94	68.47
SimGCD	78.95	84.59	73.31	78.84	83.11	68.14

TABLE IV
Mean precision on the datasets

	CIFAR-10			RDD		
	All	Old	New	All	Old	New
Proposed Method	78.72	97.43	60.01	83.57	92.72	60.71
Prior Method	77.82	95.83	59.80	84.08	91.72	65.01
SimGCD	91.17	98.49	83.84	84.42	92.04	65.36

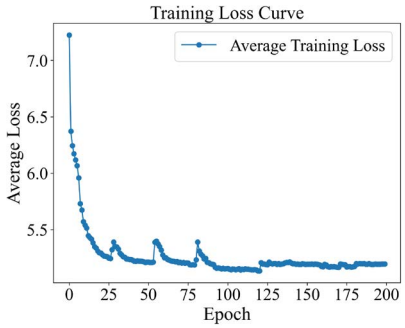

Fig. 3. Loss curve during training

VI. Discussion

A. Comparison with Conventional Methods

The proposed method achieved higher accuracy compared to existing methods on CIFAR-10. This is attributed to incorporating responsibility considerations into prototype merging and minimizing changes to prototypes of known classes. Experimental results demonstrate a significant improvement in accuracy for known classes. In particular, the restriction on merging prototypes within known classes appears to have been highly effective. This likely stems from the beneficial merging of prototypes distributed around the clusters of known classes, which would otherwise cause excessive fragmentation of the feature space. However, the new method did not improve accuracy compared with our previous approach for RDD. The different progress of prototype merging could cause it. Figure 4 shows the change in the number of prototypes in RDD for the three methods tested in this experiment. Even under identical conditions, our self-organizing classifier can produce varying learning results. This is likely because similarity is the only criterion used for merging. Intuitively, we hypothesize that overly similar prototype pairs may incorrectly split classes, thereby reducing accuracy. However, pairs with a significant impact accuracy are not necessarily more similar than the threshold. This mismatch is thought to cause the instability of the results.We plan to review the integration criteria to achieve more stable results in the future.

B. Limitations and Future Directions

Our proposed method achieved significant improvements in accuracy. However, at present, a detailed analysis of its algorithm has not been performed. In this study, the merging threshold was selected empirically based on the similarity distribution. However, the sensitivity of this hyperparameter has not yet been systematically analyzed. Future work will include a detailed investigation of the threshold's influence on model stability and class discovery accuracy, potentially leading to an adaptive thresholding strategy guided by similarity statistics. Through these efforts, we will deepen our understanding of merging criteria, build models adaptable to a wide range of benchmarks, and conduct more detailed analyses of their performance.

Another limitation is that the responsibility-based merging was heuristically defined. Although inspired by probabilistic clustering, the current formulation does not guarantee convergence to an optimal partition. Future research will explore a fully probabilistic framework, for example, by integrating an EM-like expectation step or Bayesian prototype updating scheme, to provide theoretical convergence guarantees.

VII. Conclusion

This paper proposes an improved method for self-organizing classifiers for GCD and verifies its performance under conditions where the number of classes is unknown. The results achieved an overall improvement in accuracy.

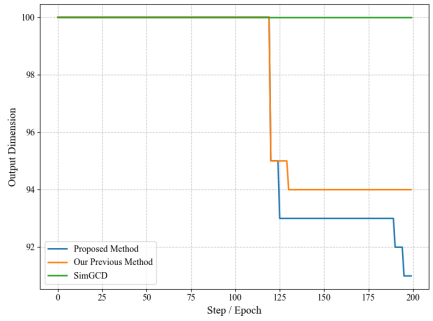

Fig. 4. The change of output dimension

Notably, significant improvement was observed for known classes, suggesting that the merging constraints for known classes and the introduction of responsibility scores were effective. In future research, we plan to explore adaptive determination of the merging threshold to further enhance the robustness of the self-organizing process across different datasets.

Acknowledgment

This work was supported by the Japan Construction Information Center and by JSPS KAKENHI Grant Number 25K15171.

References

[1] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Generalized category discovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7492–7501.

[2] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9650–9660.

[3] X. Wen, B. Zhao, and X. Qi, "Parametric classification for generalized category discovery: A baseline study," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 16 590–16 600.

[4] R. Akashi and K. Matsushima, "Generalized category discovery with a self-organizing classifier through prototype merging," in *The 17th International Conference on Information Technology and Electrical Engineering*, October 2025, pp. 484–489.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[6] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Repvit: Revisiting mobile cnn from vit perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 15 909–15 920.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[8] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109