# Supplementary Material

## A Evaluation Metrics

The **maximum forgetting metric** $\mathcal{M}_f$ and the **final discovery metric** $\mathcal{M}_d$ are proposed for evaluation. $\mathcal{M}_f$ measures the capability to maintain the performance on the known categories, which is the lower the better. $\mathcal{M}_d$ measures the ability to discover novel categories, which is the higher the better. To evaluate the performance of the clustering assignments, we follow the standard practise [1, 2] to adopt clustering accuracy. First, an optimal permutation $h^*$ that matches the cluster assignments $y_i^*$ with the ground truth label $y_i$ is obtained by solving the following optimization problem using Hungarian algorithm [3]:

$$h_t^* = \arg\min_h \frac{1}{M^t} \sum_{i=1}^{M^t} \mathbb{I}(y_i = h(y_i^*)), \tag{A.1}$$

where $M^t$ denotes the size of $\mathcal{D}_{test}^t$. The clustering accuracy on the known categories $\text{ACC}_{\text{known}}^t$ and novel categories $\text{ACC}_{\text{novel}}^t$ at time $t$ can be obtained as follows:

$$\text{ACC}_{\text{known}}^t = \frac{1}{M_{\text{known}}^t} \sum_{i=1}^{M_{\text{known}}^t} \mathbb{I}(y_{i,\text{known}} = h_t^*(y_i^*)), \tag{A.2}$$

$$\text{ACC}_{\text{novel}}^t = \frac{1}{M_{\text{novel}}^t} \sum_{i=1}^{M_{\text{novel}}^t} \mathbb{I}(y_{i,\text{novel}} = h_t^*(y_i^*)), \tag{A.3}$$

where $M_{\text{known}}^t$ and $M_{\text{novel}}^t$ denote the number of known category samples and novel category samples from $\mathcal{D}_{test}^t$, respectively. The $y_{i,\text{known}}$ and $y_{i,\text{novel}}$ denote the label of the known samples and category samples, respectively. Then, the maximum forgetting $\mathcal{M}_f$ and final discovery $\mathcal{M}_d$ can be obtained by

$$\mathcal{M}_f = \max_t \{\text{ACC}_{\text{known}}^0 - \text{ACC}_{\text{known}}^t\}, \tag{A.4}$$

$$\mathcal{M}_d = \text{ACC}_{\text{novel}}^T. \tag{A.5}$$

The importance of $\mathcal{M}_f$ and $\mathcal{M}_d$ are different. First, $\mathcal{M}_f$ should be sufficiently low, otherwise a model forgetting the previous learned tasks is not practically useful in the real world applications. Second, the model should improve $\mathcal{M}_d$ as much as possible on the condition of low $\mathcal{M}_f$.

## B Illustration of the GM Framework

During the growing phase, the novel data is detected by the novelty detection process, and the corresponding features are extracted by the dynamic branch. The cluster head is used to assign pseudo labels for the novel data. During the merging phase, the exemplar set is updated with the sifted novel data as well as the pseudo labels, and the static branch is unified with the dynamic branch. For testing, the features are extracted by the dynamic branch, and the predictions are provided according to distances to the prototypes.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

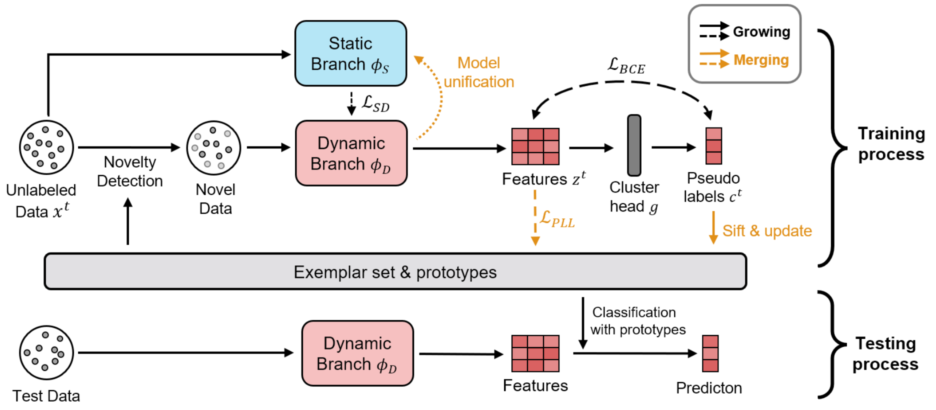

Figure B1: The illustration of the GM Framework.

## C Exemplar Set Construction

After the initial pre-training stage, the exemplar set is initialized, following iCaRL [4]. Denote $P_k = \{\boldsymbol{p}_{k,i}\}_i$ as the exemplar set of the $k$-th class. Then $\boldsymbol{p}_{k,i}$ is selected by solving the following optimization problem,

$$\boldsymbol{p}_{k,i} \longleftarrow \underset{\boldsymbol{x}_i^0}{\arg\min} \left\| \frac{1}{N_k^0} \sum_{y_i^0 = k} \phi_D^0(\boldsymbol{x}_i^0) - \frac{1}{i} [\phi_D^0(\boldsymbol{x}_i^0) + \sum_{s=1}^{i-1} \phi_D^t(\boldsymbol{p}_{k,s})] \right\|, \tag{C.1}$$

where $N_k^0$ denotes the number of samples belong to the $k$-th class.

## D Scenarios Details

We formulate four different scenarios for experiments as follows:

*Class Incremental Scenario (CI)*: the data are only drawn from the novel categories, *i.e.*, $\mathcal{C}^t \cap (\mathcal{C}^{t-1} \cup \cdots \cup \mathcal{C}^0) = \emptyset, \forall t \in \{1, \cdots, T\}$. CI is a standard scenario that only requires the models to learn from the novel samples and keep the performance of the known categories.

*Data Incremental Scenario (DI)*: the data are only drawn from the known categories, *i.e.*, $\mathcal{C}^0 = \mathcal{C}^t, \forall t \in \{1, \cdots, T\}$. DI is a simple scenario, which evaluates the models' capability to improve the feature representation with the continuous unlabeled data.

*Mixed Incremental Scenario (MI)*: the data are drawn from both novel categories and the known categories, *i.e.*, $\mathcal{C}^0 \subset \mathcal{C}^1 \subset \cdots \subset \mathcal{C}^T$. MI is more complicated than CI, where models are further required to identify whether the data comes from novel categories or not.

*Semi-supervised Mixed Incremental Scenario (SMI)*: the data are drawn from both novel categories and the known categories, *i.e.*, $\mathcal{C}^0 \subset \mathcal{C}^1 \subset \cdots \subset \mathcal{C}^T$, and a portion of the data are labeled. SMI is closer to the real-world application, where both labeled and unlabeled samples are provided in the incremental stages.

For CI, 70%/10%/10%/10% classes of the CIFAR-100, CUB-200, ImageNet-100, Stanford-Cars, and FGVC-Aircraft datasets are used in the initial stage and the following 3 time-steps of the continuous category discovery stage, respectively. For DI, 25% of the data are used for the initial stage, and 75% are used for the continuous category discovery stage, with 25% data at each time-step. For MI, 87% data from 0-70 classes in CIFAR-100 are used for initial stage. 7% of data from 0-70 classes, 70% data from 70-80 classes are used for $t = 1$ during the continuous category discovery stage. 2% data from 0-70 classes, 20% data from 70-80 classes and 90% data from 80-90 classes are used for $t = 2$. 3% data form 0-70 classes, 10% data from 70-80 classes, 10% data from 80-90 classes and all of the data from 90-100 classes are used during $t = 3$. For SMI, 20% of the data from MI are labeled during $t = 1, 2, 3$.

Table F1: Experimental results of 20/50 incremental classes at each time step on CI scenario and CIFAR-100 dataset.

| | Step=20 | | Step=50 | |
|---|---|---|---|---|
| | $\mathcal{M}_f$ | $\mathcal{M}_d$ | $\mathcal{M}_f$ | $\mathcal{M}_d$ |
| lower-bound of $\mathcal{M}_f$ (offline K-Means) | 7.18±0.57 | 17.07±0.51 | 11.69±0.79 | 15.46±0.31 |
| upper-bould of $\mathcal{M}_d$ (offline AutoNovel) | 5.58±0.37 | 32.79±1.07 | 6.20±0.62 | 30.26±1.41 |
| AutoNovel (online) | 63.02±0.48 | 27.03±3.16 | 62.51±0.22 | 12.32 ± 0.22 |
| DRNCD (online) | 65.67±1.68 | 7.12±2.15 | 68.53±0.89 | 7.25±1.05 |
| AutoNovel (online) + LwF | 16.41±2.64 | 27.18±1.28 | 26.57±1.47 | 10.10±4.90 |
| DRNCD (online) + LwF | 67.31±1.24 | 7.7±1.64 | 67.78±1.42 | 8.74±0.49 |
| iCaRL (fixed exemplars) + LwF | 31.02±1.06 | - | 38.83±0.75 | - |
| GM (Ours) | **10.86±0.33** | **34.37±0.80** | **13.36±0.49** | **13.69±0.32** |

Table F2: Additional ablation studies.

| | $\mathcal{M}_f$ | $\mathcal{M}_d$ |
|---|---|---|
| GM | 9.87±0.25 | 35.97±1.28 |
| Replace WTA with GCD | 7.59±0.18 | 31.43±1.35 |
| avg. sifting | 10.03±0.36 | 34.85±1.46 |
| Ramp-up weight for $\mathcal{L}_{\text{MSE}}$ | 8.69±0.17 | 33.20±0.96 |

# E Implementation Details

For all experiments, we use ResNet-18 [5] network as the encoder, and fully connected layers as the classifier and the cluster head. SGD optimizer [6] with learning rate 0.1, momentum 0.9 and weight decay $1e-4$ is used. The learning rate decays for every 60 epochs. The batch size is set to 128 for CIFAR-100 dataset, 64 for CUB-200, Stanford-Cars, and FGVC-Aircraft datasets, and 32 for ImageNet-100. All the experiments run on the platform with Intel Xeon CPU E5-2640 and Nvidia GeForce RTX 3080 GPUs.

# F Additional Experiments

## F.1 More Incremental Classes at each time step

In this section, we provide experimental results of 20/50 incremental classes at each time step on CI scenario and CIFAR-100 dataset, where GMNet can be well compatible with different number of incremental classes and obtain consistent performance improvement, shown in Table F1.

## F.2 Additional Ablation Studies

**Alternative of $\mathcal{L}_{\text{BCE}}$ with WTA hash.** GM is a general framework, supporting various methods to learn the representation of the samples. Here, we perform ablation study, which removes $\mathcal{L}_{\text{BCE}}$ +WTA and the cluster head, performs the supervised/unsupervised contrastive loss in GCD on the exemplar set/unlabeled samples respectively, and applies semi-supervised K-Means algorithm for label assignment. The results in CI scenario and CIFAR-100 dataset are shown in Table F2.

**Alternative of sample sifting strategy.** The maximum distance of top-$j$ NN is used for sifting in GM. However, the average of the $j$ distances could also be applied during the sifting process. The results of using average distance of top-$j$ nearest neighbor in CI scenario and CIFAR-100 dataset are shown in Table F2. Meanwhile, the influence of $j$ in GM under CI on CIFAR-100 is studied, shown in Table F3.

**Weights of each terms in the loss function**. The loss of GM framework contains four terms, *i.e.*, $\mathcal{L}_{\text{BCE}}$, $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{PLL}}$, $\mathcal{L}_{\text{SD}}$. The studies of the impact of the variation of the weight for each term is shown in Table F4. During the experiments, one of the weight varies while the other weights keep fixed. We can find that although four losses are involved, our method shows robust to different loss

Table F3: Influence of $j$ for sifting strategy under CI on CIFAR-100.

| $k$ | $\mathcal{M}_f$ | $\mathcal{M}_d$ |
|---|---|---|
| 5 | 10.06 | 34.10 |
| 10 | 10.11 | 35.03 |
| 15 | 9.79 | 35.77 |
| 20 | 9.64 | 34.13 |
| 25 | 9.54 | 33.87 |

Table F4: Impact of the variation of the weight for each term in the loss function. For each term, the corresponding weight varies while the other weights keep fixed.

| | $\mathcal{L}_{\text{BCE}}$ | | $\mathcal{L}_{\text{MSE}}$ | | $\mathcal{L}_{\text{PLL}}$ | | $\mathcal{L}_{\text{SD}}$ | |
|---|---|---|---|---|---|---|---|---|
| Weight | $\mathcal{M}_f$ | $\mathcal{M}_d$ | $\mathcal{M}_f$ | $\mathcal{M}_d$ | $\mathcal{M}_f$ | $\mathcal{M}_d$ | $\mathcal{M}_f$ | $\mathcal{M}_d$ |
| 0.1 | 9.61 | 21.00 | 10.21 | 32.47 | 9.11 | 37.03 | 10.20 | 36.67 |
| 0.25 | 8.94 | 27.93 | 10.61 | 34.63 | 9.84 | 37.17 | 10.29 | 37.13 |
| 0.5 | 9.41 | 33.77 | 9.81 | 35.07 | 9.45 | 36.23 | 10.41 | 35.93 |
| 0.75 | 10.36 | 35.17 | 10.11 | 35.80 | 9.94 | 35.87 | 10.49 | 36.9 |
| 1.0 | 9.79 | 35.77 | 9.79 | 35.77 | 9.79 | 35.77 | 9.79 | 35.77 |
| 2.5 | 9.94 | 37.00 | 10.37 | 39.87 | 20.43 | 35.23 | 10.49 | 36.9 |
| 5.0 | 9.84 | 36.6 | 9.10 | 35.57 | 30.59 | 35.13 | 9.81 | 33.00 |
| 7.5 | 10.27 | 33.33 | 9.72 | 34.2 | 71.07 | 1.23 | 9.87 | 33.4 |
| 10.0 | 10.07 | 36.3 | 8.44 | 31.63 | 71.07 | 0.66 | 10.01 | 32.70 |

weights. Except for using too small weight for $\mathcal{L}_{\text{BCE}}$ and too large weight for $\mathcal{L}_{\text{PLL}}$, in other cases, our model can obtain superior performance with a large range of loss weight (from 0.1 to 10). As the ramp-up function is widely used as the weight of MSE loss [1], experiments are conducted on to evaluate the performance of GM model with ramp-up weight for $\mathcal{L}_{\text{MSE}}$, shown in Table F2. The model with ramp-up weight performs slightly better on $\mathcal{M}_f$ and worse on $\mathcal{M}_d$.

**Study on novelty detection threshold** $\epsilon$. In our experiments, GM is not sensitive to this threshold and GM with the 0.6 theshhold will achieve relatively high performance. Here, we provide the ablation study on the novelty detection threshold, shown in Table F5. The results show that though GM achieves the best performance with threshold 0.6, it maintains relatively high performance with the values in [0.4, 0.7] and consistently outperforms existing methods. The experiments are conducted on the CIFAR-100 dataset in MI scenario.

**Discussion about the effectiveness of GMNet on** $\mathcal{M}_d$. For the compared methods, they can hardly solve the contradiction between the two tasks of classification and novel classes discovery in contiguous stages well, leading to the degraded performance . Labeled data is introduced in SMI, so that the model can fit the distribution of new data and learn the novel category information at each stage, reducing the difficulty of novelty detection and improving the performance of $\mathcal{M}_d$, but even so, existing methods still not well compatible with the two tasks, led to a more severe forgetting effect. For GMNet, through the double-branch and exemplar structure and $\mathcal{L}_{\text{SD}}$ and $\mathcal{L}_{\text{PLL}}$ on top of this, the model can be well compatible with two tasks in contiguous stages. we provide ablation study for them in Table F6 under MI scenario. We can find that, $\mathcal{L}_{\text{PLL}}$ could significantly decrease $\mathcal{M}_f$ and increase $\mathcal{M}_d$. EMA and $\mathcal{L}_{\text{SD}}$ could help the model leverage the effective representation learned

Table F5: Influence of $\epsilon$ for novelty detection under MI on CIFAR-100.

| $\epsilon$ | $\mathcal{M}_f$ | $\mathcal{M}_d$ |
|---|---|---|
| 0.4 | 9.79 | 29.27 |
| 0.5 | 10.29 | 32.33 |
| 0.6 | 9.79 | 35.77 |
| 0.7 | 10.16 | 30.37 |

Table F6: Ablation studies about the effectiveness of GMNet on $\mathcal{M}_d$ under MI on CIFAR-100.F

| | $\mathcal{M}_f$ | $\mathcal{M}_d$ |
|---|---|---|
| GM | 9.65±0.32 | 30.58±1.13 |
| GM w/o. $\mathcal{L}_{\text{SD}}$ | 22.03±0.49 | 22.47±2.62 |
| GM w/o. $\mathcal{L}_{\text{SD}}$, EMA | 26.20±0.67 | 24.35±1.09 |
| GM w/o. $\mathcal{L}_{\text{SD}}$, EMA, $\mathcal{L}_{\text{PLL}}$ | 46.46±0.81 | 18.02±1.77 |

Table F7: The alignment and uniformity of feature distribution under CI in CIFAR-100 dataset. Classification Model represents the model trained on known categories, Discovery Model represents the model trained for novel category discovery, and GM model represents the proposed method.

| | | Classification Model | Discovery Model | GM Model |
|---|---|---|---|---|
| Alignment | Known Categories | -1.008± 0.007 | -1.153±0.003 | -0.825±0.007 |
| | Novel Categories | -1.118±0.005 | -1.009±0.008 | -0.923±0.009 |
| Uniformity | Known Categories | -2.704±0.008 | -2.625±0.005 | -2.510±0.012 |
| | Novel Categories | -2.631±0.009 | -2.754±0.008 | -2.397±0.012 |

from the initial stage, and further improve the performance. With the help of proposed methods, GM obtains the best performance on both $\mathcal{M}_f$ and $\mathcal{M}_d$.

### F.3 Alignment and Uniformity

The visualization of the feature distribution of different models in Figure 2 is qualitative. In this section, the quantitative results of the distribution are provided. Following [7], the alignment and uniformity are calculated under CI on CIFAR-100 dataset. The alignment measures the distances between each positive pairs, while the uniformity represents the diversity of the feature. The results in Table F7 show that the changes of these two indicators are consistent with our assumption. Classification model shows smaller uniformity and larger alignment on on known categories than novel categories and discovery model show the opposite results, which indicates feature discriminative ability of the two model focus on different sample types (novel or known categories). Compared to the two models, GM model show better uniformity and Alignment on both novel and known categories.

## G Limitations and Future Works

Here we discuss the limitations of GM and the future works. First, although GM outperforms the compared baseline methods on both $\mathcal{M}_f$ and $\mathcal{M}_d$, the forgetting of the knowledge on the known categories is still not overcome completely. Second, the scenario in real-world applications is more complicated than the proposed four scenarios, which is worthy for further studying.

## H Potential Negative Social Impact

The proposed GM framework significantly reduces the impact of catastrophic forgetting and improves the ability of novel category discovery compared with other baseline methods, which is pratical in real-world applications with unlabeled data stream. However, specific settings of the model may lead to unreasonable usages with negative social impact, *i.e.*, identifying minorities and discriminating against them. Our method is not specifically designed for the imbalanced training set (*e.g.*, gender, age and race) against the potential social impact. It is recommended to clean, sift, and rebalance the imbalanced training set, especially the images for different gender and race. We also claim that the proposed GM framework is mainly for research purpose and could not be applied directly without the rational assessment towards the deployed environment.

## References

[1] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[2] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 22982–22994, 2021.

[3] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[4] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5533–5542, 2017.

[5] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[6] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147, 2013.

[7] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020. URL `http://proceedings.mlr.press/v119/wang20k.html`.