模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

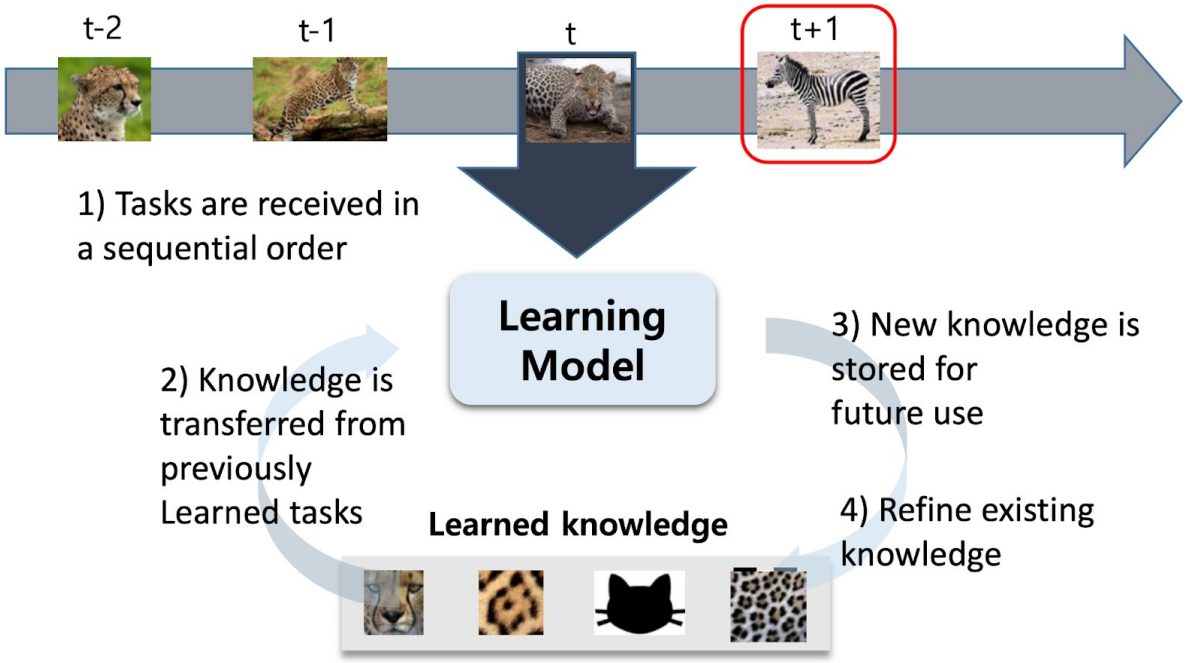
# Self-Sustaining Representation Expansion for Non-Exemplar Class-Incremental Learning

Kai Zhu[1] Wei Zhai[1] Yang Cao[1,3,†] Jiebo Luo[2] Zheng-Jun Zha[1]
[1] University of Science and Technology of China [2] University of Rochester
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{zkzy@mail., wzhai056@mail., forrest@}ustc.edu.cn jluo@cs.rochester.edu zhazj@ustc.edu.cn

CVPR 2022

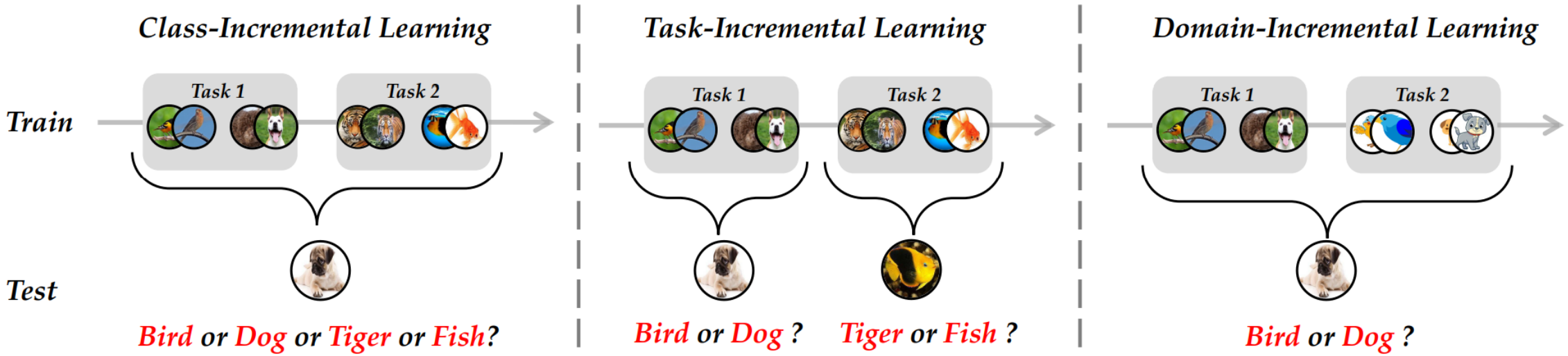
## Continual Learning

Continual learning (CL) tackles a learning scenario where a model continuously learns over a sequence of tasks.



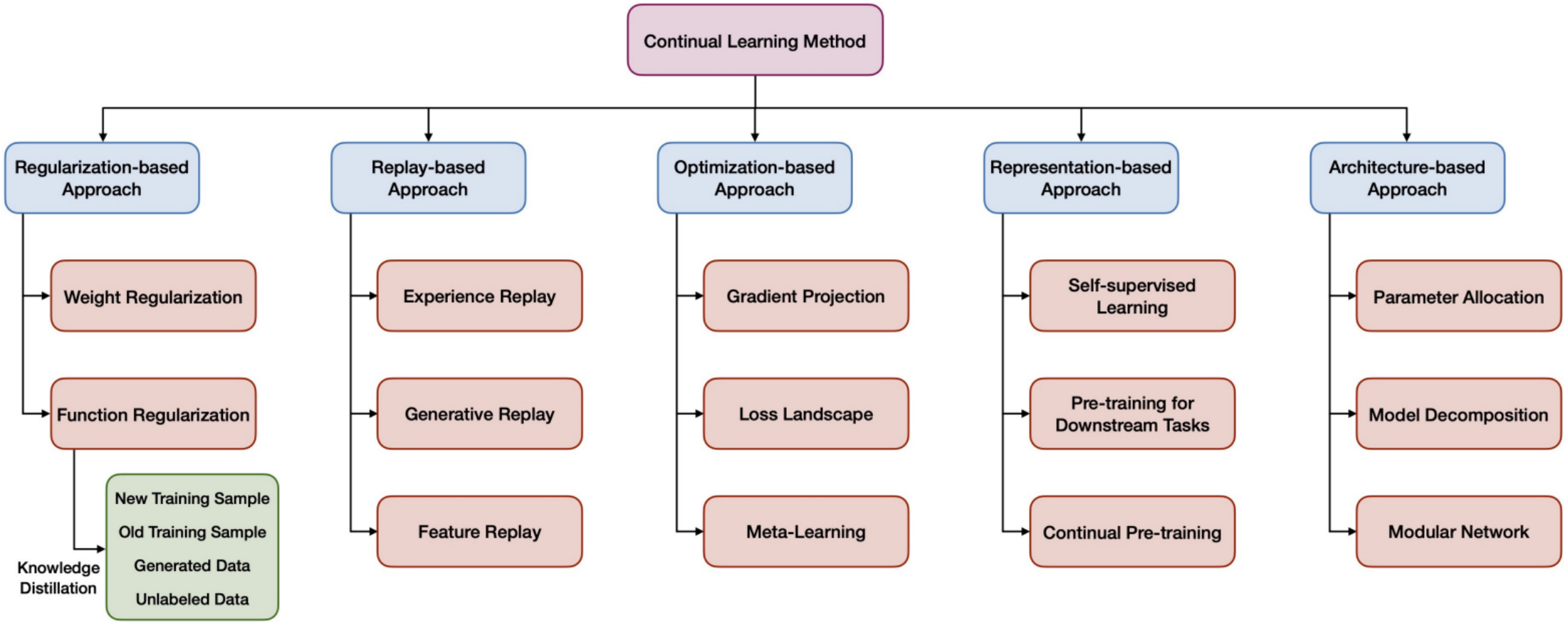

**Catastrophic Forgetting :**Phenomenon that ***forgets*** the knowledge of **previous classes** when the model trains with new incoming tasks.

## The typical setting of CIL



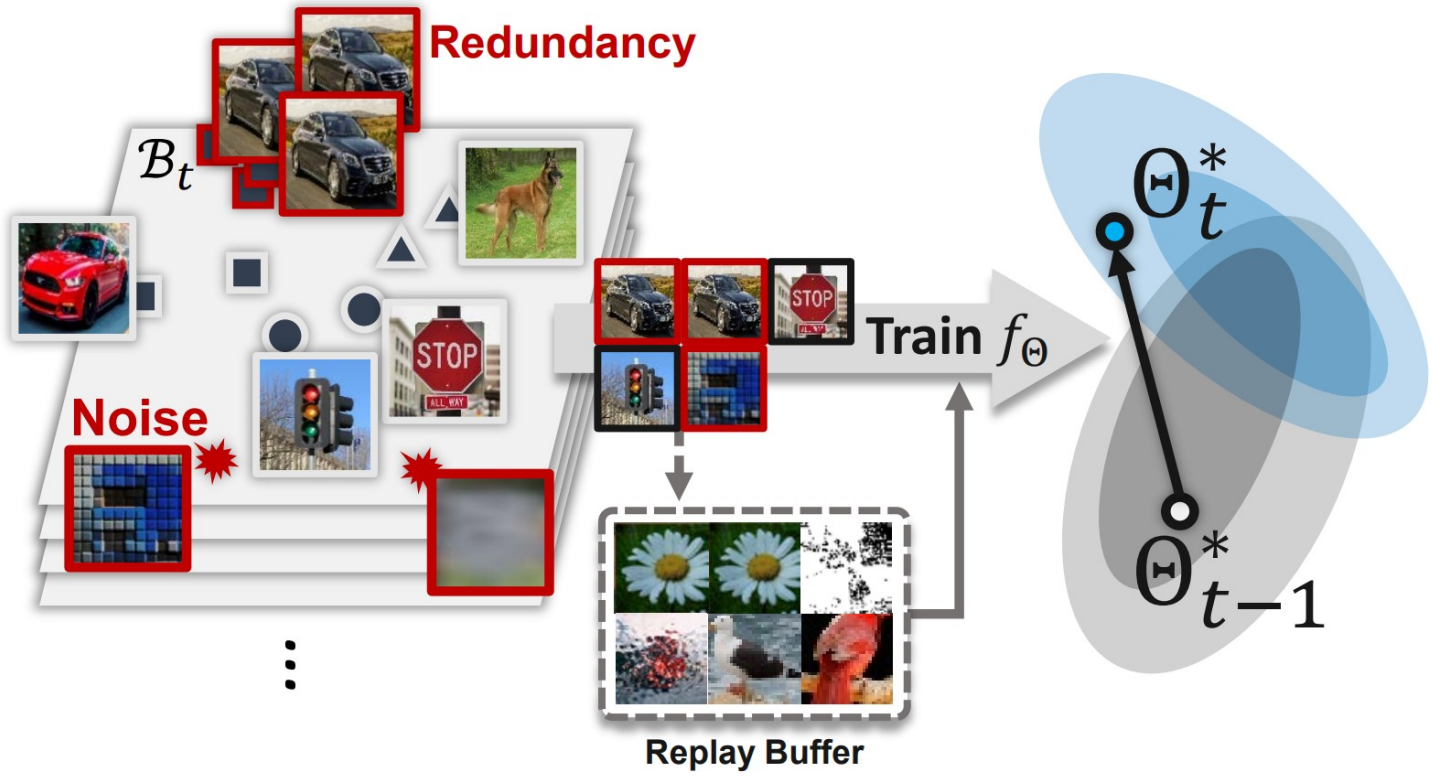

CIL and TIL share the same training protocol, while TIL is much easier during inference, i.e., only requiring classifying among corresponding label spaces.

# Overview of CIL Methods



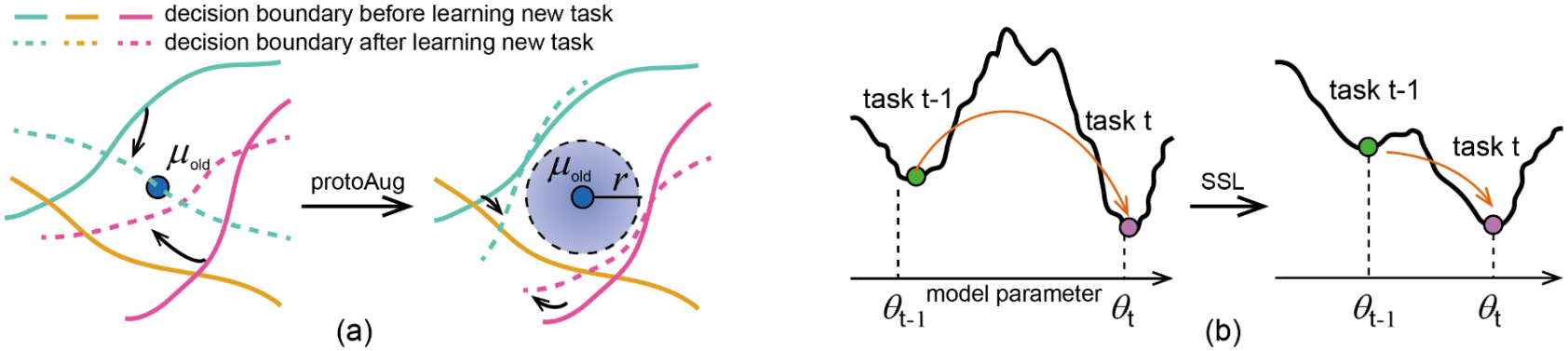
## Replay-based CL

Existing replay-based methods train on all the arrived instances and memorize a fraction of them in the replay buffer, resulting in a suboptimal perf.

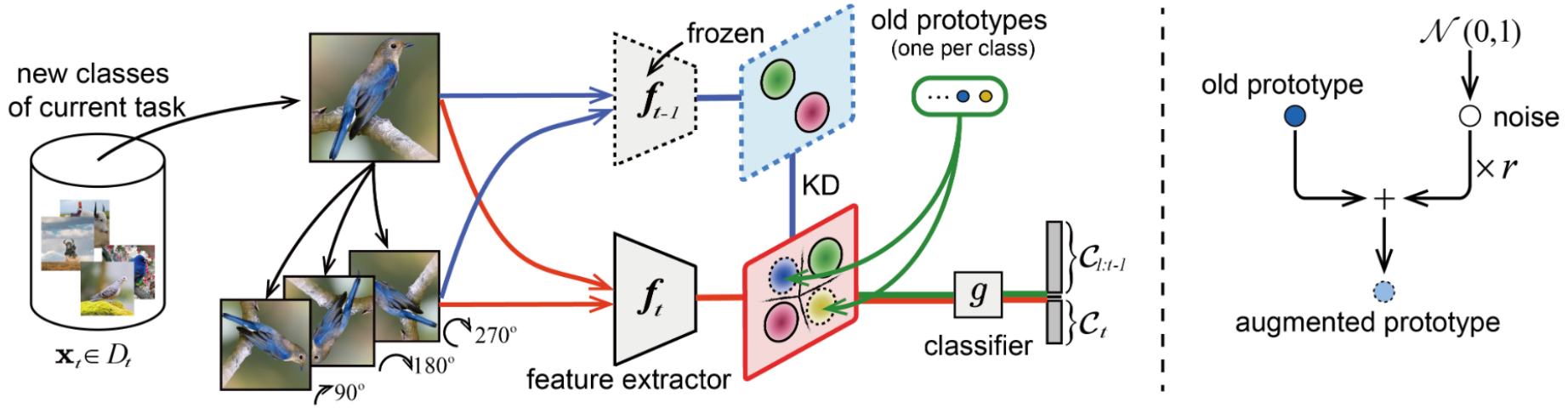


# Non-exemplar class-incremental learning (NECIL) PASS

## Motivation



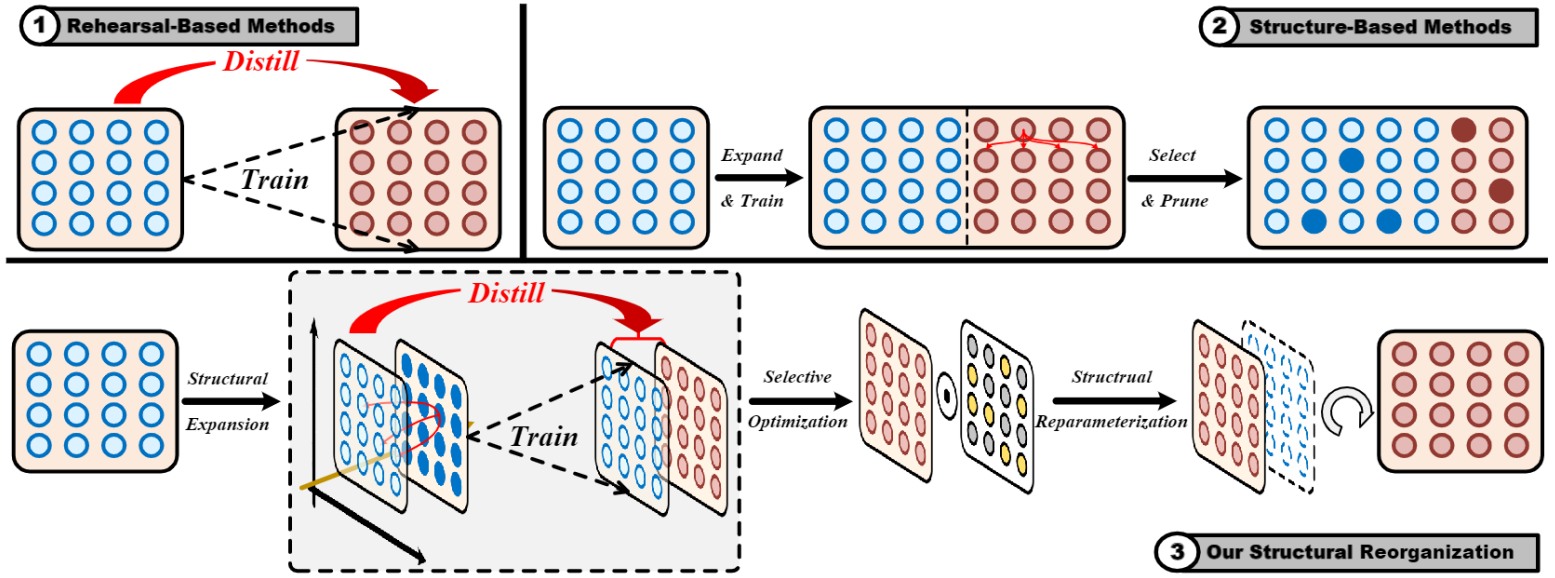

## Method

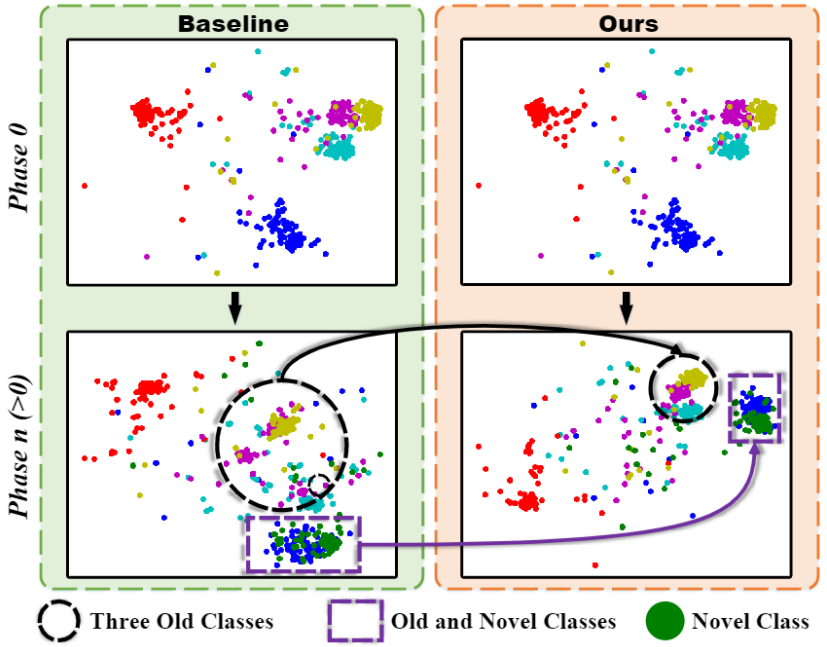



[1] Fei Zhu, et al. “Prototype Augmentation and Self-Supervision for Incremental Learning.” *CVPR 2021(Oral).*

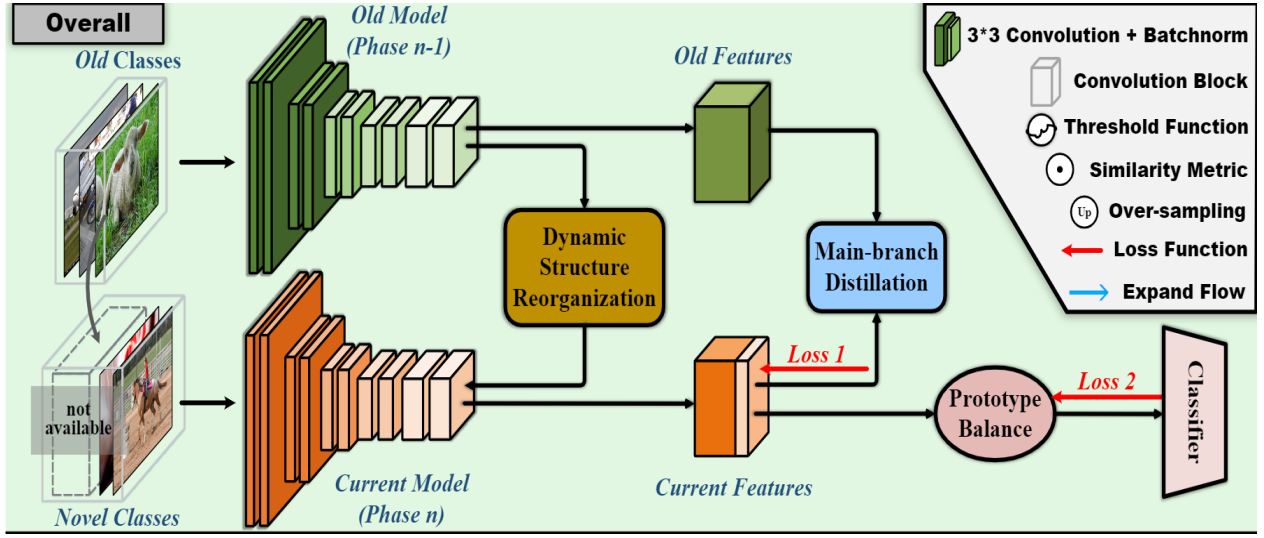
## Self-Sustaining Representation Expansion For NECIL

Replay-based and structure-based methods suffer from the unreliability of distillation in the absence of exemplars and continuously expanding structure, respectively.



# Standard NECIL Paradigm




## 1. Representation learning using knowledge distillation:

At task 1, model $f_\theta^1$ consisting of the feature extractor $f_e^1$ and classifier $g_c^1$

At task n:

$$r_q^n = f_e^n(Q;\theta_e^n). \qquad s_q^n = g_c^n(r_q^n;\theta_c^n)$$

$$\rightarrow \quad L_{ce} = F_{ce}(s_q^n, y_q^n),$$

$$r_q^{n-1} = f_e^{n-1}(Q;\theta_e^{n-1})$$

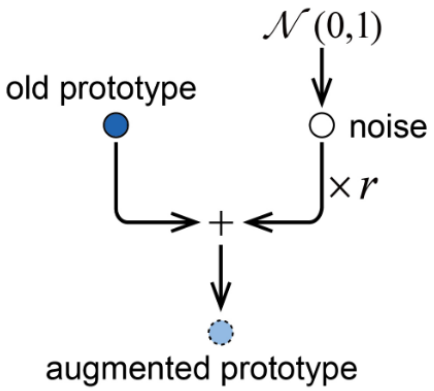
$$\rightarrow \quad L_{kd} = F_{kd}(r_q^n, r_q^{n-1}),$$

## 2. Incremental Classifier Calibration.

To overcome the imbalance between the exemplars and new samples in CIL

$$p_B = Up_B(Prototype),$$

$$L_{proto} = F_{ce}(p_B, y_B)$$




→ **Final Loss** $L = L_{ce} + \lambda L_{kd} + \gamma L_{proto},$

## Optimization

In the CIL, CE Loss can be turned into two parts:

$$L_{ce} = F_{ce}(s_q^n, y_q^n), \quad \rightarrow \quad L_{ce} = F_{ce}(s_q^n, y_q^n) + F_{ce}(s_e^n, y_e^n),$$

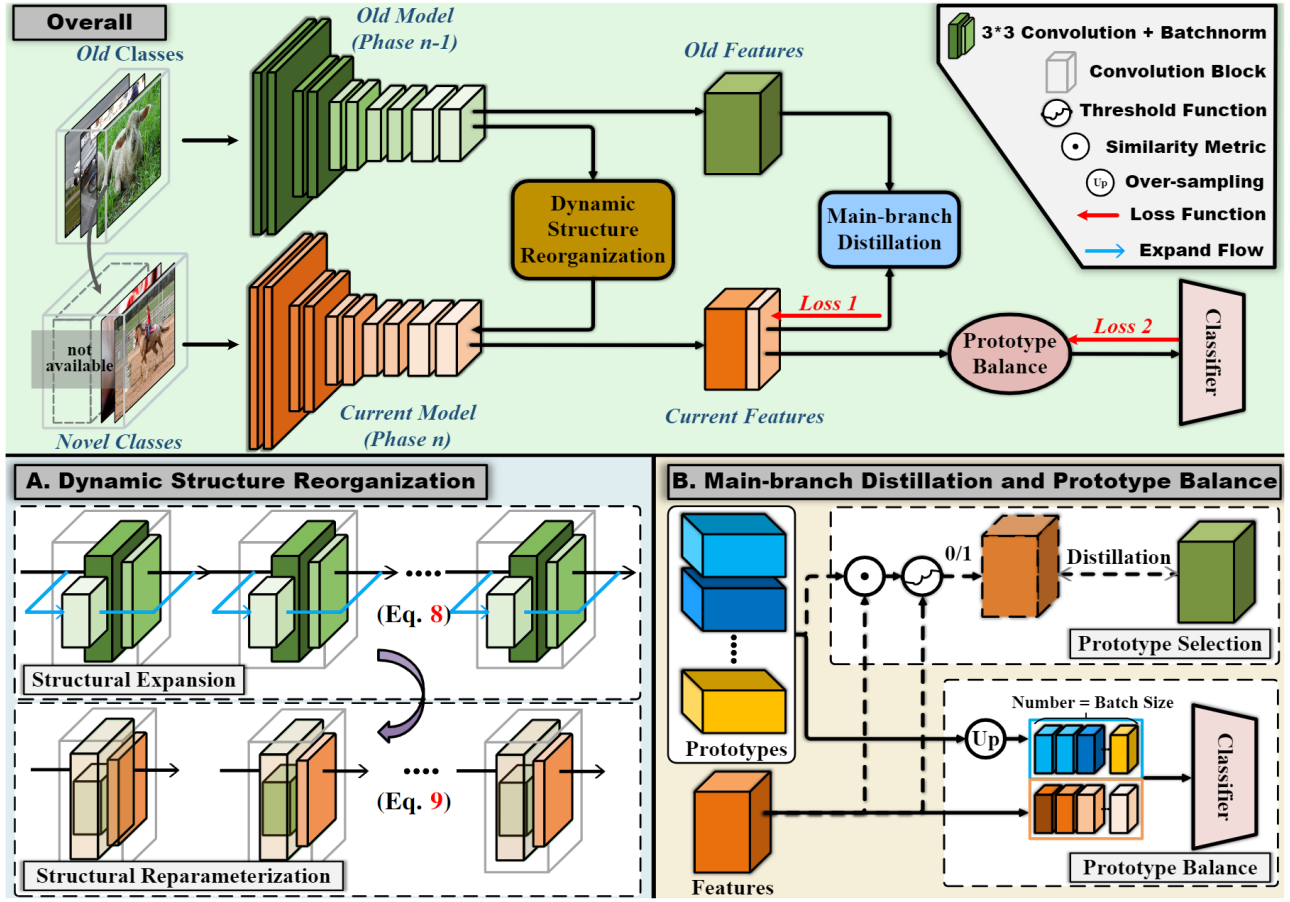
While this imbalance can bias the optimization process towards features that are more discriminative for the new class, the added distillation in KD Loss can alleviate this problem:

$$L_{kd} = F_{kd}(r_q^n, r_q^{n-1}), \rightarrow \quad L_{kd} = F_{kd}(r_q^n, r_q^{n-1}) + F_{kd}(r_e^n, r_e^{n-1}).$$

In NECIL settings, the joint optimization on the old and new class representations completely collapses into feature optimization.

# Self-Sustaining Representation Expansion

The proposed self-sustaining representation expansion scheme for NECIL:



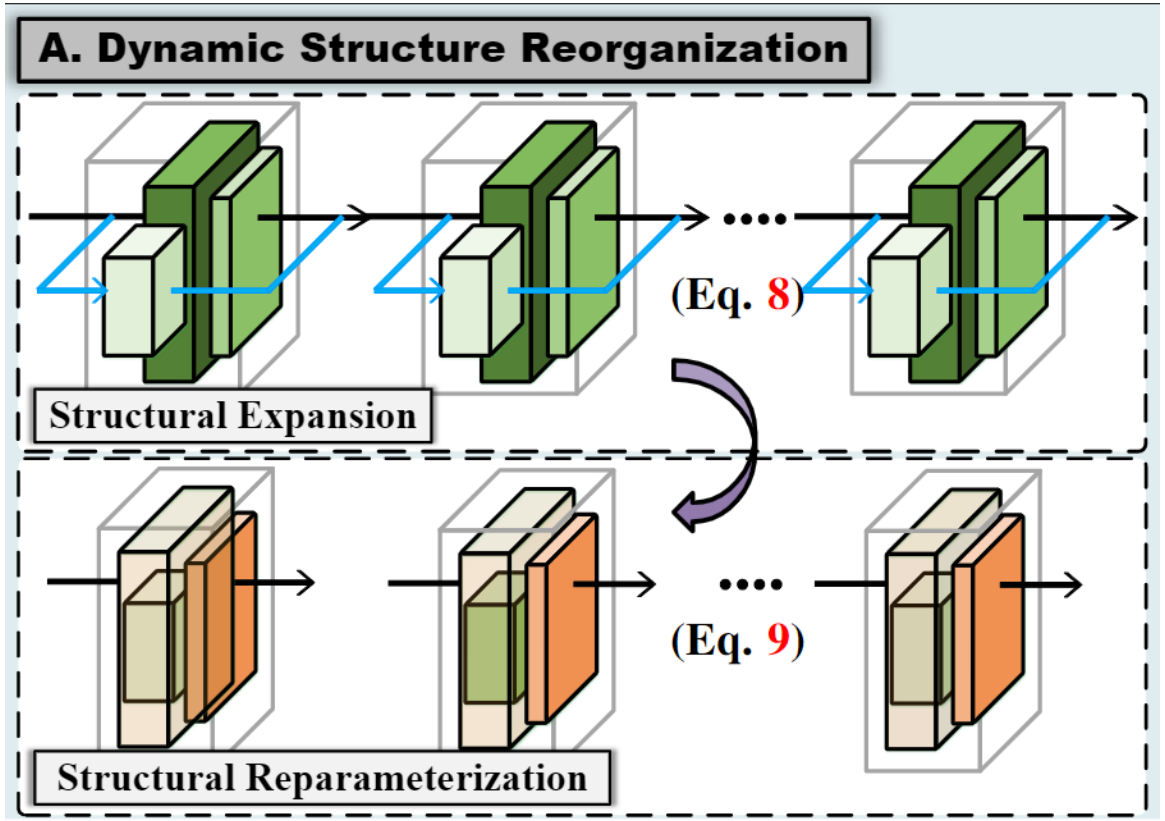

(1) Overview of the scheme,

(2) Dynamic Structure Reorganization

(3) Main-Branch Distillation

(4) Prototype Balance

# Dynamic Structure Reorganization.



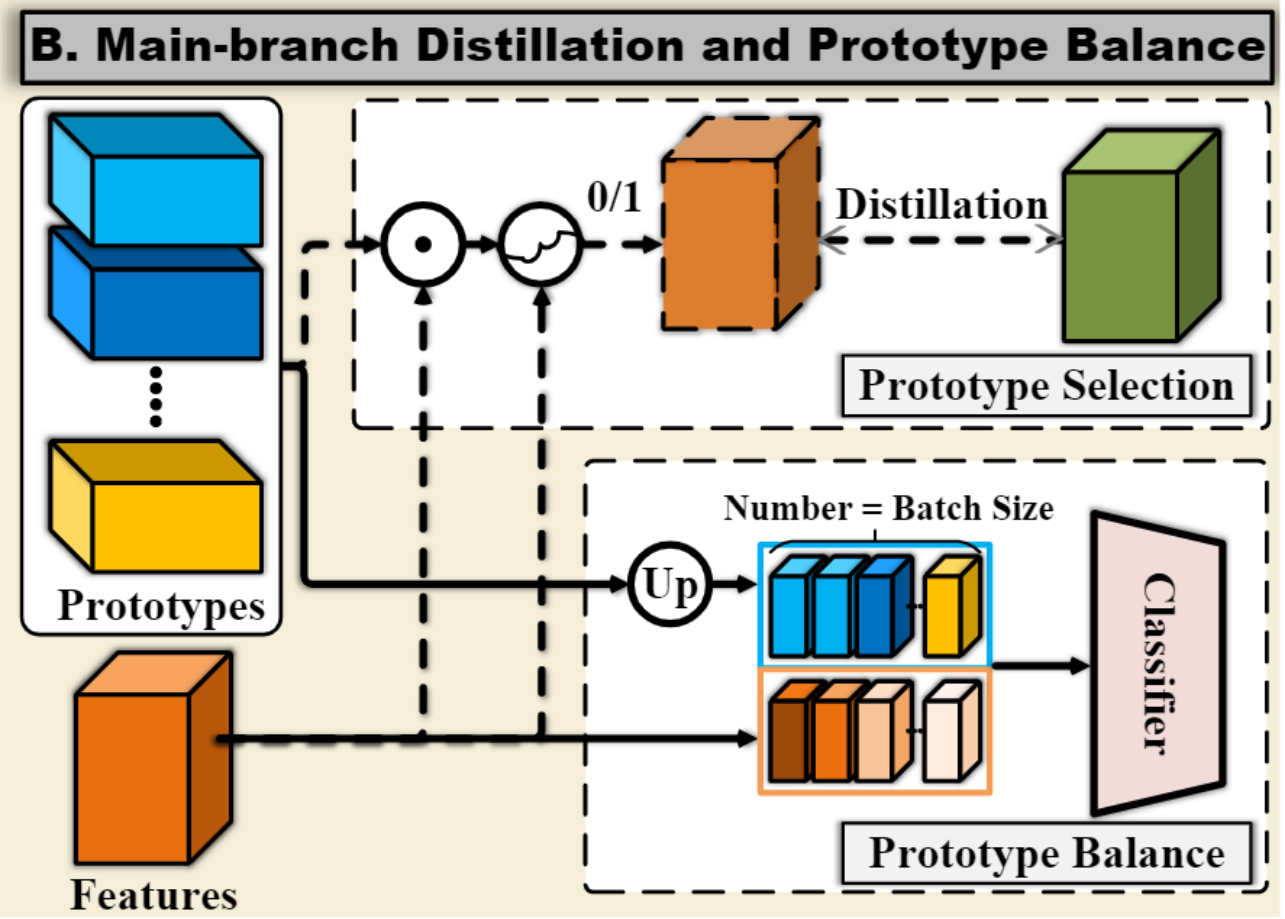

Insert a residual adapter to each convolution block of the fixed feature extractor from previous phase

$$f_e^n(Q;\theta_e^n) = F_{transform}(f_e^{n-1}(Q;\theta_e^{n-1})) = f_e^{n-1}(Q;\hat{\theta}_e^{n-1} \oplus \Delta\theta_e^n), \quad (8)$$

After training, the structural reparameterization to integrate the side branch information into the main branch losslessly

$$f_e^{n-1}(Q;\hat{\theta}_e^{n-1} \oplus \Delta\theta_e^n) = f_e^n(Q;\theta_e^{n\prime} \oplus 0) = f_e^n(Q;\theta_e^{n\prime}). \quad (9)$$

## Prototype Selection.




After mapping all new samples to the learned embedding space, compute the normalized cosine scores $Si$ between them and all prototypes.

$$Si = Cosine(N(r_q^n), Nor(Prototype)),$$

If $S_i > \tau$ :

add a mask for distillation loss to emphasize the distinction between the old and new classes

Else :

add a mask for cross-entropy loss to emphasize the learning of the new class features

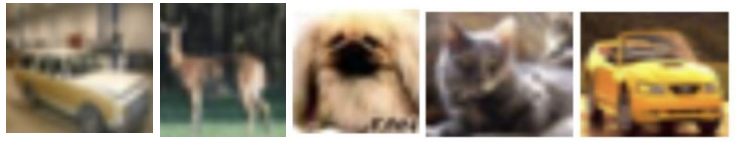
$$L = Mask_{ce}(L_{ce}) + \lambda Mask_{kd}(L_{kd}) + \gamma L_{proto}.$$

# Experiment Setup



## *Datasets*

1) **Split CIFAR-10 [Krizhevsky 2012]** A dataset with 60,000 images composed of ***five tasks*** from ***ten animal and vehicle classes***.



2) **Split CIFAR-100 [Krizhevsky 2012]** A dataset with 60,000 images composed of ***20 tasks*** from ***100 generic object classes***.

3) **Split Tiny-ImageNet [Russakovsky 2015]** A ***subset of ImageNet*** dataset. We construct ***20 tasks*** using ***100 classes***.

## *Metrics*

**1)** **Accuracy** is the average test accuracy of all the tasks completed until the continual learning of task $\mathcal{T}$

$$A_\tau = \frac{1}{\tau}\sum_{i=1}^{\tau} a_{\tau,i}$$

**2)** **Forgetting** is the average performance decrease of each task between its maximum accuracy and accuracy at the completion of training

$$F = \frac{1}{T-1}\sum_{i=1}^{T-1} \max_{\tau\in\{1,\ldots,T\}} \left(a_{\tau,i} - a_{T,i}\right)$$

$a_{\tau,i}$ is the test accuracy of task $i$ after learning task $\mathcal{T}_\tau$ using a KNN on frozen pre-trained representations on task $\mathcal{T}_\tau$

[Krizhevsky 2012] Krizhevsky, A. Learning multiple layer of features from tiny images. University of Toronto 2012
[Ruakovsk 2015] Imagenet large scale visual recognition challenge. International journal of computer vision, 2015

## Compared to the SOTA

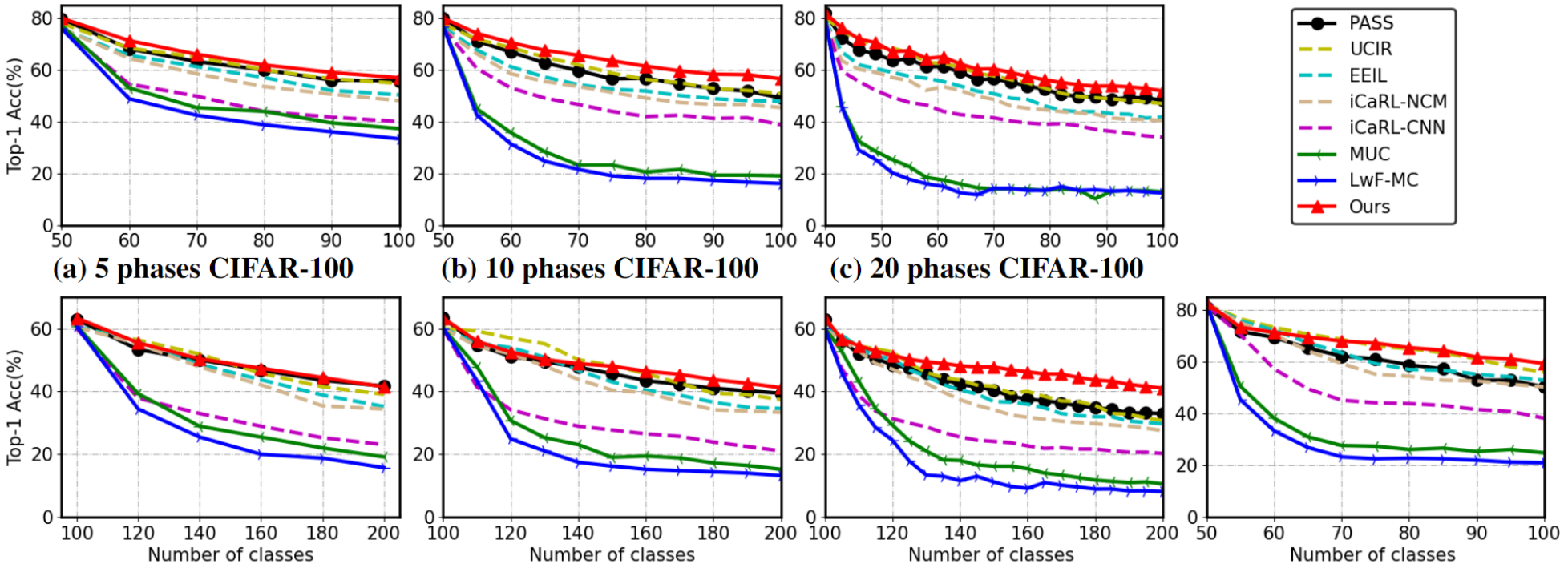
The proposed method achieves average improvement of 3, 3 and 6 points on CIFAR-100, TinyImageNet and ImageNetSubset, respectively.

| | Methods | CIFAR-100 | | | TinyImageNet | | | ImageNet-Subset |
|---|---|---|---|---|---|---|---|---|
| | | *P=5* | *P=10* | *P=20* | *P=5* | *P=10* | *P=20* | *P=10* |
| *(1) E=20* | iCaRL-CNN* | 51.07 | 48.66 | 44.43 | 34.64 | 31.15 | 27.90 | 50.53 |
| | iCaRL-NCM* [25] | 58.56 | 54.19 | 50.51 | 45.86 | 43.29 | 38.04 | 60.79 |
| | EEIL* [1] | 60.37 | 56.05 | 52.34 | 47.12 | 45.01 | 40.50 | 63.34 |
| | UCIR* [10] | 63.78 | 62.39 | 59.07 | 49.15 | 48.52 | 42.83 | 66.16 |
| *(2) E=0* | EWC* [14] | 24.48 | 21.20 | 15.89 | 18.80 | 15.77 | 12.39 | 20.40 |
| | LwF_MC* [25] | 45.93 | 27.43 | 20.07 | 29.12 | 23.10 | 17.43 | 31.18 |
| | MUC* [34] | 49.42 | 30.19 | 21.27 | 32.58 | 26.61 | 21.95 | 35.07 |
| | SDC [35] | 56.77 | 57.00 | 58.90 | - | - | - | 61.12 |
| | PASS [37] | 63.47 | 61.84 | 58.09 | 49.55 | 47.29 | 42.07 | 61.80 |
| | Ours | **65.88**+2.41 | **65.04**+3.20 | **61.70**+2.80 | **50.39**+0.84 | **48.93**+1.64 | **48.17**+6.10 | **67.69**+5.89 |

Table 3. Comparisons of the average incremental accuracy (%) with other methods on CIFAR-100, TinyImageNet, and ImageNet-Subset. P represents the number of phases and E represents the number of exemplars. Models with an asterisk * represent the reproduced results in [37]. The red footnotes in the last row represent the relative improvement compared with the results of SOTA.

## Compared to the SOTA

The proposed method achieves average improvement of 3, 3 and 6 points on CIFAR-100, TinyImageNet and ImageNetSubset, respectively.




(a) 5 phases CIFAR-100 (b) 10 phases CIFAR-100 (c) 20 phases CIFAR-100

(d) 5 phases TinyImagNet (e) 10 phases TinyImagNet (f) 20 phases TinyImagNet (g) 10 phases ImagNet-Subset

Figure 7. Classification accuracy on CIFAR-100, TinyImageNet and ImageNet-Subset, which contains the complete curves.

## Compared to the SOTA

The proposed method achieves average improvement of 3, 3 and 6 points on CIFAR-100, TinyImageNet and ImageNetSubset, respectively.

| | CIFAR-100 | | | TinyImageNet | | |
|---|---|---|---|---|---|---|
| Method | 5 | 10 | 20 | 5 | 10 | 20 |
| iCaRL-CNN | 42.13 | 45.69 | 43.54 | 36.89 | 36.70 | 45.12 |
| iCaRL-NCM | 24.90 | 28.32 | 35.53 | 27.15 | 28.89 | 37.40 |
| EEIL | 23.36 | 26.65 | 32.40 | 25.56 | 25.91 | 35.04 |
| UCIR | 21.00 | 25.12 | 28.65 | 20.61 | 22.25 | 33.74 |
| LwF_MC | 44.23 | 50.47 | 55.46 | 54.26 | 54.37 | 63.54 |
| MUC | 40.28 | 47.56 | 52.65 | 51.46 | 50.21 | 58.00 |
| PASS | 25.20 | 30.25 | 30.61 | 18.04 | 23.11 | 30.55 |
| Ours | **18.37** | **19.48** | **19.00** | **9.17** | **14.06** | **14.20** |

Table 4. Results of average forgetting on 5, 10 and 20 phases.

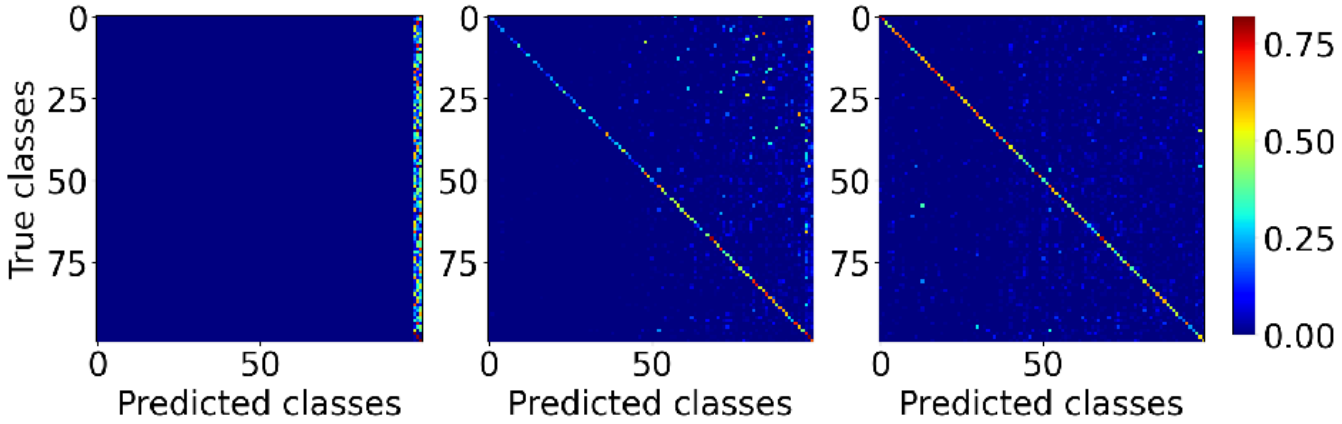



Figure 5. Confusion matrices of different methods on CIFAR-100.

# Experiment



## Analysis

The impact of the adapter structure.

| Method | CIFAR-100 | | |
|---|---|---|---|
| | 5 phases | 10 phases | 20 phases |
| 3×3 conv | 64.28 | 63.47 | 60.81 |
| 1×1 conv + bn | 65.88 | 64.84 | 60.72 |
| 1×1 conv | 65.87 | 65.12 | 61.60 |

Table 2. Performance under different expanding structures.

The impact of the threshold in prototype selection

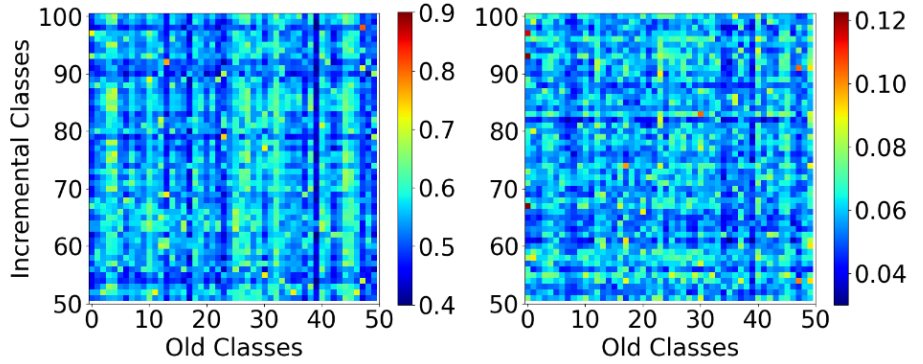



Figure 8. Statistics of similarity on the incremental samples.

## Ablation Study

| DSR | MBD | PSM | CIFAR-100 | | |
|---|---|---|---|---|---|
| | | | 5 phases | 10 phases | 20 phases |
| | | | 61.11 | 57.08 | 51.04 |
| √ | | | 64.86 | 63.25 | 54.09 |
| | √ | | 62.70 | 62.60 | 58.57 |
| √ | √ | | 65.10 | 63.87 | 60.60 |
| √ | √ | √ | 65.88 | 64.69 | 61.61 |

Table 1. Ablation study of our method on CIFAR-100.

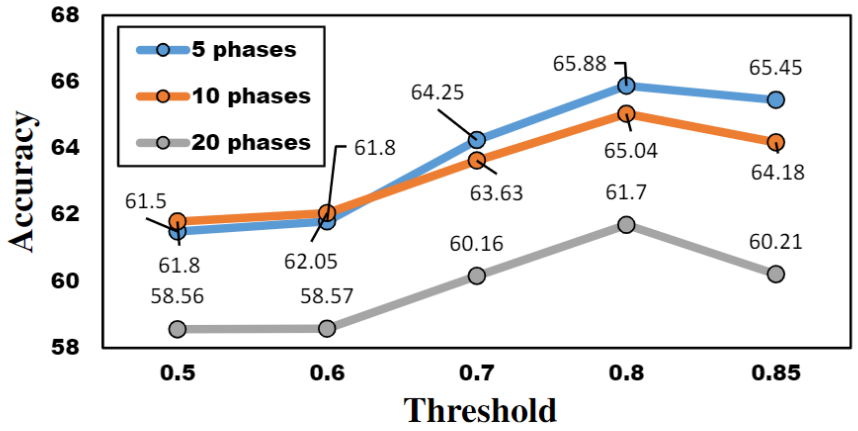



Figure 4. Illustration of the role of the selection mechanism.

## Visualization

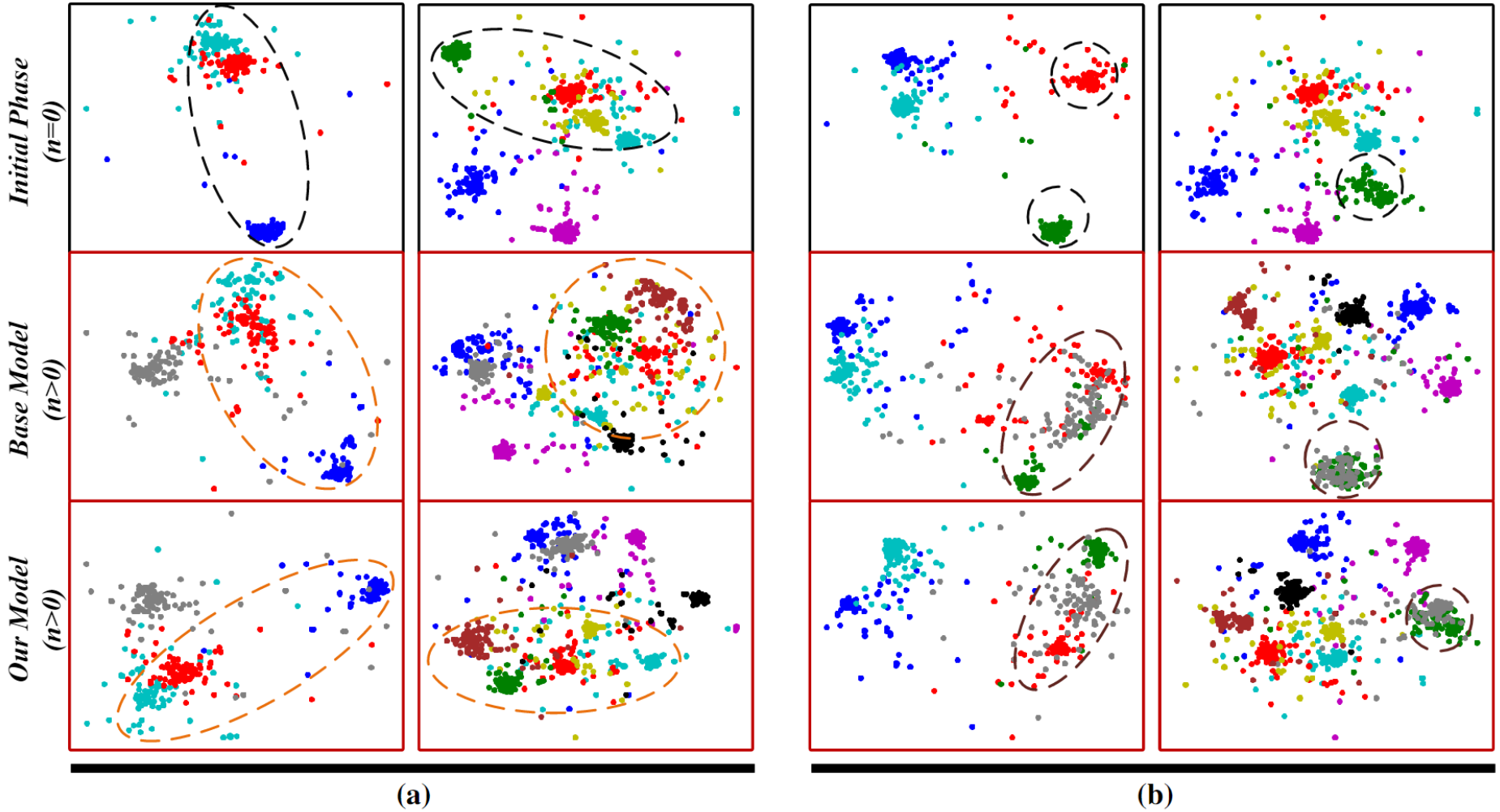



Figure 6. Effect of our scheme on the representation. (a) DSR maintains the discriminative features and inter-relations of old classes, thus enhancing the clustering and separation of the distribution of old classes. (b) MBD results in a better distinction between similar classes.

# Thanks