南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

L3A: Label-Augmented Analytic Adaptation for Multi-Label Class Incremental Learning

Xiang Zhang [1] **Run He** [1] **Jiao Chen** [1] **Di Fang** [2]
Ming Li [3] **Ziqian Zeng** [1] **Cen Chen** [2] **Huiping Zhuang** [1]

ICML 2025

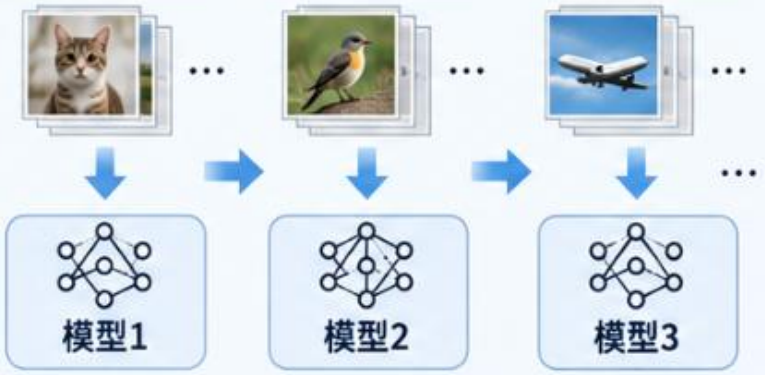
Background

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

持续学习 (Continual Learning)

数据分阶段到来，模型不断学习新知识

阶段1 阶段2 阶段3



难点：灾难性遗忘

多标签任务 (Multi-Label Learning)

一个样本可同时对应多个类别

标签集合：

{person, bicycle, car}

多热编码示例：

y = [1, 1, 0, 1, ⋯]

类增量学习 (Class-Incremental Learning)

新类别不断加入，测试时需识别所有历史类别与当前类别

阶段1 → 类别空间：{person}

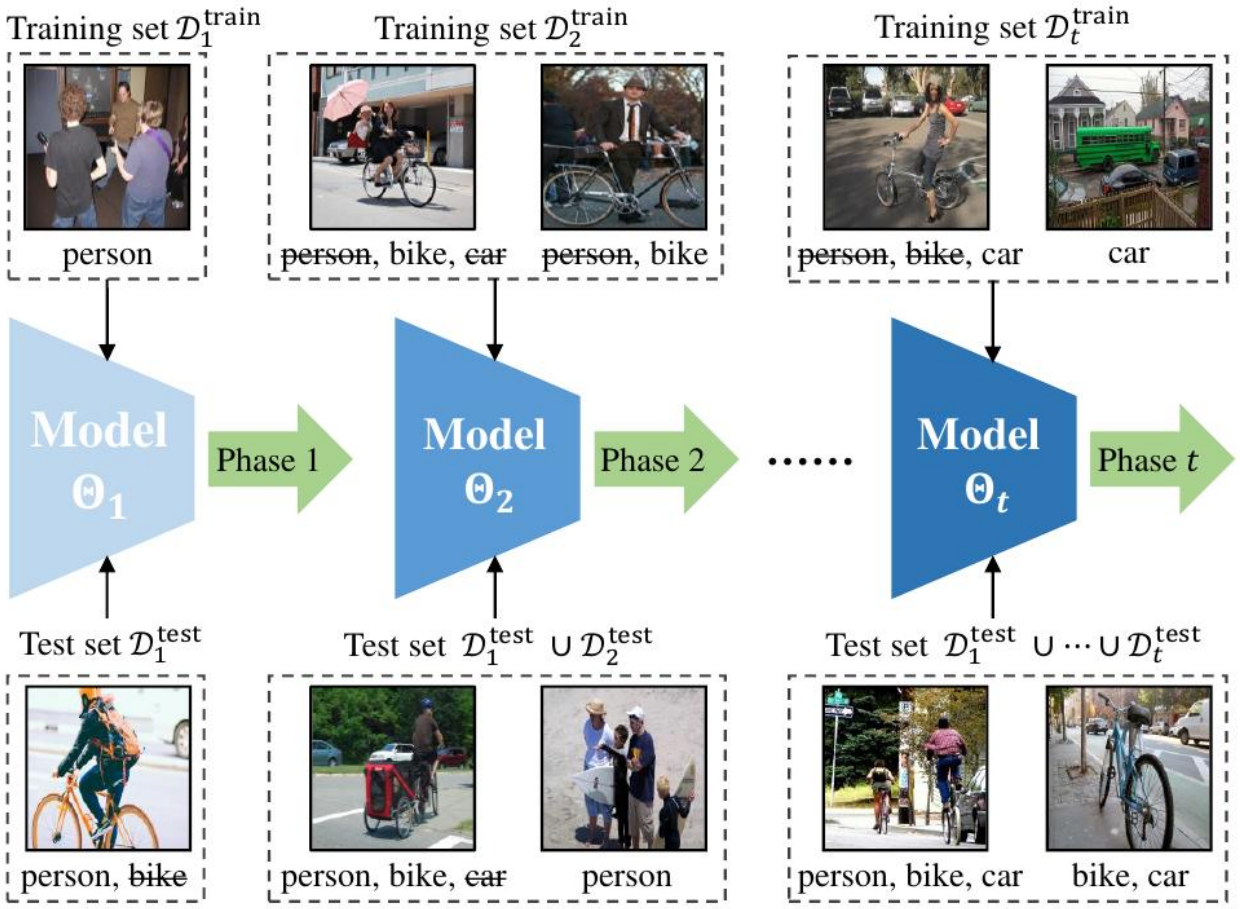
阶段2 → 类别空间：{person, bicycle}

阶段3 → 类别空间：{person, bicycle, car}

⋮

Background

Multi-Label Class-Incremental Learning (MLCIL)



I、label absence

每个阶段只提供当前类别标签，容易把真实存在的旧类当成负样本，从而加剧灾难性遗忘。

2、class imablance

多标签图像中类别经常共现，导致部分高频类出现很多、稀有类很少，模型容易偏向高频类别。

3、privacy protection

真实场景中往往不能长期保存历史样本，因此依赖 replay 的持续学习方法难以使用。

Background

现有MLCIL方法：

1、KRT：知识恢复与迁移

2、CSC：类增量 GCN 建模标签关系

→ 主要针对 MLCIL 的标签缺失
最优性能仍依赖 replay 历史样本

3、MULTI-LANE：预训练大模型 + Prompt Tuning、面向 replay-free MLCIL

→ 无需保存历史样本
未专门处理类别不平衡

现有方法仍难同时解决三项问题：label absence、class imbalance、privacy protection

许多的examplar-free CIL方法：$D_t \rightarrow Loss_t \rightarrow Backprop \rightarrow \theta_t$

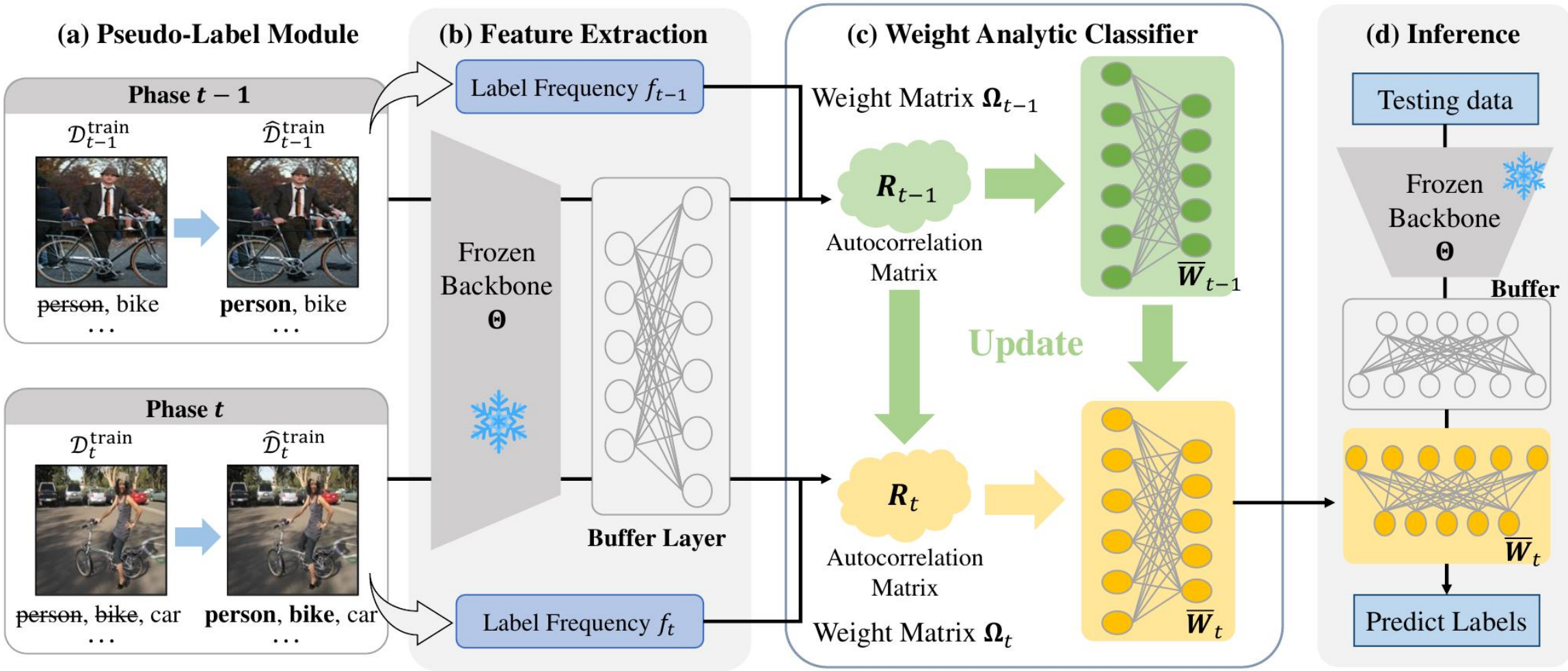
每阶段只有当前数据，所以梯度天然主要服务于：当前最新任务 → Task-Recency Bias

Analytic Continual Learning 不用这种 BP 式的逐步梯度优化，而是：闭式解+递归更新

$$\min_{W_t} \left\| X_{1:t} W_t - \hat{Y}_{1:t} \right\|_F^2 + \gamma \| W_t \|_F^2$$

L3A：Pseudo-Label 补齐历史标签 + Weighted Analytic Classifier 处理类别不均衡，并保持 exemplar-free。

Methods



$$\bar{W}_t = (X_{1:t}^\top \Omega_{1:t} X_{1:t} + \gamma I)^{-1} X_{1:t}^\top \Omega_{1:t} \hat{Y}_{1:t}. \quad (10)$$

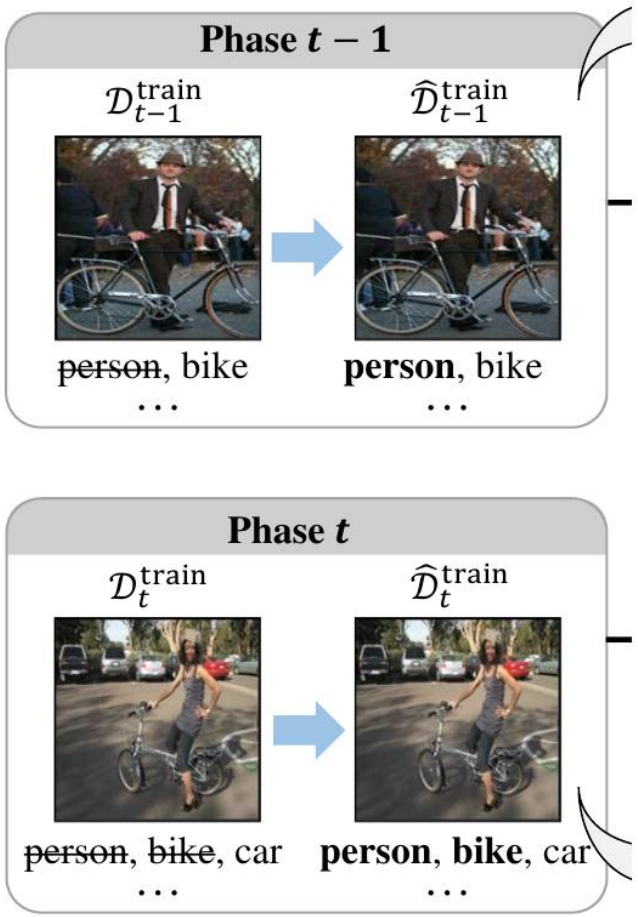
$$\bar{W}_t = \left[\bar{W}_{t-1} - R_t X_t^\top \Omega_t X_t \bar{W}_{t-1} \quad R_t X_t^\top \Omega_t \hat{Y}_t\right], \quad (12)$$

$$R_t = (X_{1:t}^\top \Omega_{1:t} X_{1:t} + \gamma I)^{-1}. \quad (11)$$

$$R_t = R_{t-1} - R_{t-1} X_t^\top \left(\Omega_t^{-1} + X_t R_{t-1} X_t^\top\right)^{-1} X_t R_{t-1}. \quad (13)$$

Methods

1、Pseudo-Label Module



$$\mathcal{D}_t^{\text{train}} = \{(\boldsymbol{\mathcal{X}}_t, \boldsymbol{Y}_t)\}$$

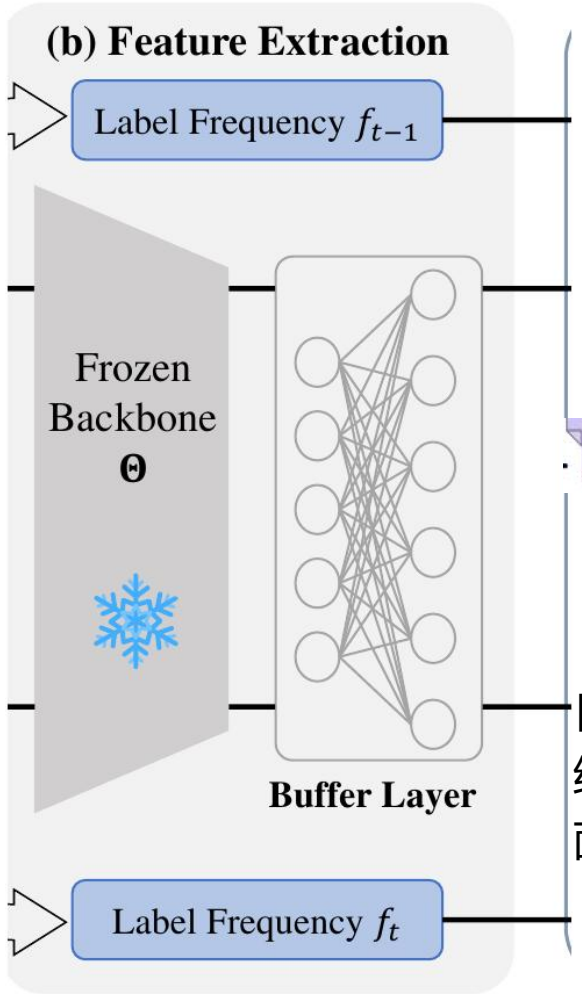
$$\tilde{y}_{t,i}^{(k)} = \begin{cases} 1 & \text{if } p_{t,i}^{(k)} \geq \eta \\ 0 & \text{otherwise,} \end{cases}$$

$$\hat{\boldsymbol{Y}}_t = \tilde{\boldsymbol{Y}}_t \cup \boldsymbol{Y}_t.$$

$$\hat{\mathcal{D}}_t^{\text{train}} = \{(\boldsymbol{\mathcal{X}}_t, \hat{\boldsymbol{Y}}_t)\}$$

$$\boldsymbol{X}_{1:t} = \begin{bmatrix} \boldsymbol{X}_{1:t-1} \\ \boldsymbol{X}_t \end{bmatrix}, \quad \hat{\boldsymbol{Y}}_{1:t} = \begin{bmatrix} \hat{\boldsymbol{Y}}_{1:t-1} & \boldsymbol{0} \\ \tilde{\boldsymbol{Y}}_t & \boldsymbol{Y}_t \end{bmatrix}.$$

2、Feature Extraction



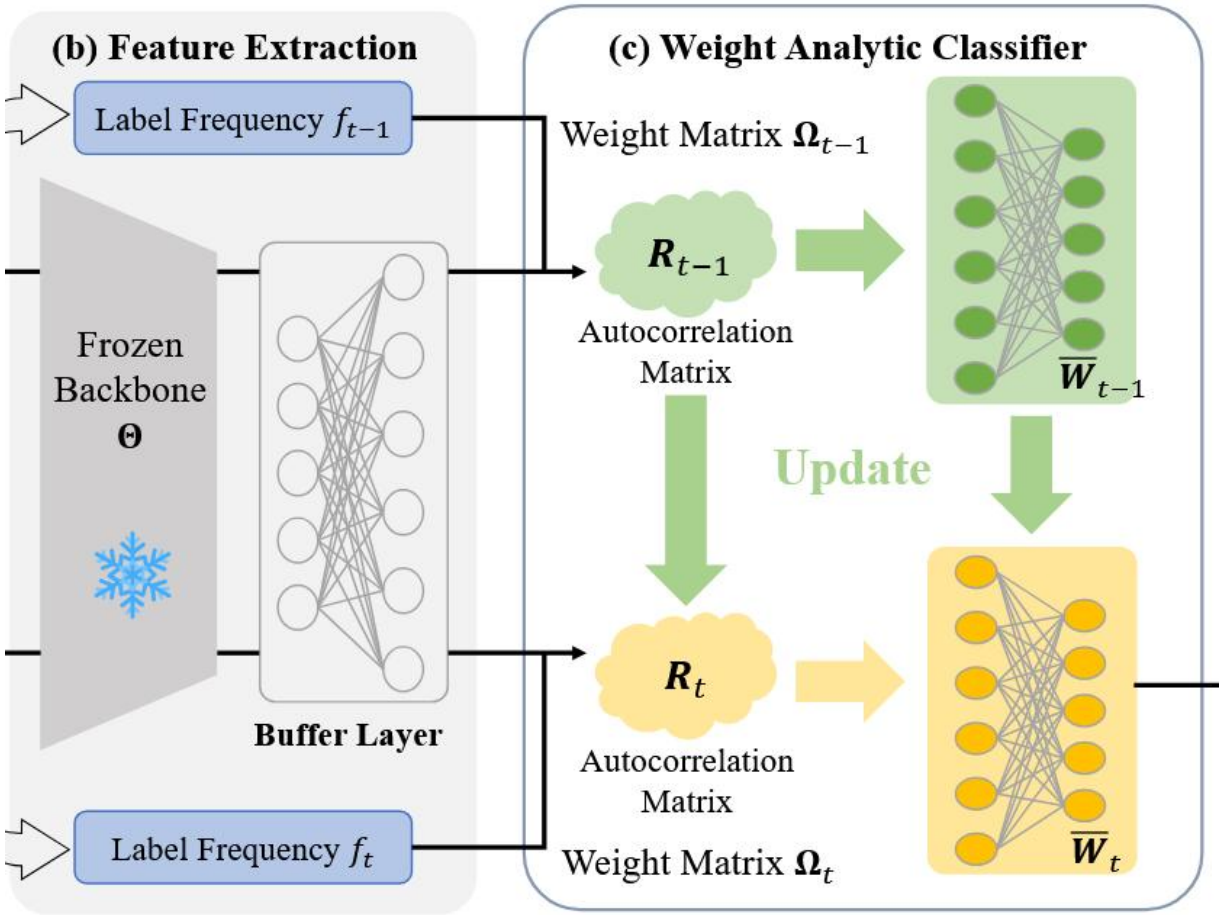
$$\boldsymbol{X}_t \leftarrow \phi(\boldsymbol{\mathcal{X}}_t, \boldsymbol{\Theta}).$$

$$f_{\text{buffer}}(\boldsymbol{X}) = \text{ReLU}(\boldsymbol{X}\boldsymbol{W}_{\text{rand}}).$$

目的：先把特征“掰到更容易线性分开”的空间，再交给后面的岭回归分类器。

Methods

3、Weighted Analytic Classifier (WAC)



$$\underset{\boldsymbol{W}_t}{\operatorname{argmin}} \|\boldsymbol{X}_{1:t}\boldsymbol{W}_t - \hat{\boldsymbol{Y}}_{1:t}\|_{\mathrm{F}}^2 + \gamma\|\boldsymbol{W}_t\|_{\mathrm{F}}^2.$$

$$v_t^{(k)} = 1/\sqrt{f^{(k)}}, \qquad \omega_{t,i} = \frac{\sum_{k=1}^{K} \hat{y}_{t,i}^{(k)} \cdot v_t^{(k)}}{\sum_{k=1}^{K} \hat{y}_{t,i}^{(k)}},$$

$$\begin{aligned}\boldsymbol{\Omega}_{1:t} &= \operatorname{diag}(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \cdots, \boldsymbol{\Omega}_t) \\ &= \operatorname{diag}(\omega_{1,1}, \cdots, \omega_{2,1}, \cdots, \omega_{t,N_t}).\end{aligned}$$

$$\underset{\boldsymbol{W}_t}{\operatorname{argmin}} \|\boldsymbol{\Omega}_{1:t}^{1/2}(\boldsymbol{X}_{1:t}\boldsymbol{W}_t - \hat{\boldsymbol{Y}}_{1:t})\|_{\mathrm{F}}^2 + \gamma\|\boldsymbol{W}_t\|_{\mathrm{F}}^2,$$

$$\underset{\boldsymbol{W}_t}{\operatorname{argmin}} \|\boldsymbol{\Omega}_{1:t}^{1/2}(\phi(\boldsymbol{\mathcal{X}}_{1:t}, \boldsymbol{\Theta})\boldsymbol{W}_t - f_{\mathrm{PL}}(\boldsymbol{Y}_{1:t}))\|_{\mathrm{F}}^2 + \gamma\|\boldsymbol{W}_t\|_{\mathrm{F}}^2, \tag{1}$$

Methods

To minimize the loss function of Equation (8), we compute its derivate with respect to $\boldsymbol{W}_t$:

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{W}_t}\left(\|\boldsymbol{\Omega}_{1:t}^{1/2}(\boldsymbol{X}_{1:t}\boldsymbol{W}_t-\hat{\boldsymbol{Y}}_{1:t})\|_{\mathrm{F}}^2+\gamma\|\boldsymbol{W}_t\|_{\mathrm{F}}^2\right)\\&=2\boldsymbol{X}_{1:t}^\top\boldsymbol{\Omega}_{1:t}\boldsymbol{X}_{1:t}\boldsymbol{W}_t-2\boldsymbol{X}_{1:t}^\top\boldsymbol{\Omega}_{1:t}\hat{\boldsymbol{Y}}_{1:t}+2\gamma\boldsymbol{W}_t.\end{aligned}\tag{9}$$

Let the derivate equal to zero, we can get the closed-form solution of $\bar{\boldsymbol{W}}_t$:

$$\bar{\boldsymbol{W}}_t=(\boldsymbol{X}_{1:t}^\top\boldsymbol{\Omega}_{1:t}\boldsymbol{X}_{1:t}+\gamma\boldsymbol{I})^{-1}\boldsymbol{X}_{1:t}^\top\boldsymbol{\Omega}_{1:t}\hat{\boldsymbol{Y}}_{1:t}.\tag{10}$$

The goal of L3A is to utilize $\bar{\boldsymbol{W}}_{t-1}$ to update $\bar{\boldsymbol{W}}_t$ recursively without other historical samples or data. We define the autocorrelation matrix in the t phase:

$$\boldsymbol{R}_t=(\boldsymbol{X}_{1:t}^\top\boldsymbol{\Omega}_{1:t}\boldsymbol{X}_{1:t}+\gamma\boldsymbol{I})^{-1}.\tag{11}$$

The iterative analytic solution process in WAC is shown in Theorem 3.1.

Theorem 3.1. *The optimal solution of $\bar{\boldsymbol{W}}_t$ can be calculated recursively by*

$$\bar{\boldsymbol{W}}_t=\begin{bmatrix}\bar{\boldsymbol{W}}_{t-1}-\boldsymbol{R}_t\boldsymbol{X}_t^\top\boldsymbol{\Omega}_t\boldsymbol{X}_t\bar{\boldsymbol{W}}_{t-1} & \boldsymbol{R}_t\boldsymbol{X}_t^\top\boldsymbol{\Omega}_t\hat{\boldsymbol{Y}}_t\end{bmatrix},\tag{12}$$

where the autocorrelation matrix can be calculated recursively by

$$\boldsymbol{R}_t=\boldsymbol{R}_{t-1}-\boldsymbol{R}_{t-1}\boldsymbol{X}_t^\top\left(\boldsymbol{\Omega}_t^{-1}+\boldsymbol{X}_t\boldsymbol{R}_{t-1}\boldsymbol{X}_t^\top\right)^{-1}\boldsymbol{X}_t\boldsymbol{R}_{t-1}.\tag{13}$$

Algorithm 1 Training process of L3A

Input: Training set $\mathcal{D}_1^{\text{train}},\cdots,\mathcal{D}_T^{\text{train}}$ with $\mathcal{D}_t^{\text{train}}\sim(\boldsymbol{\mathcal{X}}_t,\boldsymbol{Y}_t)$, the frozen backbone $\boldsymbol{\Theta}$.
Initialization: $\boldsymbol{R}_0\leftarrow\gamma\boldsymbol{I}$, $\bar{\boldsymbol{W}}_0\leftarrow 0$
for phase $t=1$ to T **do**
  ▷ *Pseudo-label module.* ◁
  $\hat{\mathcal{D}}_t^{\text{train}}\sim(\boldsymbol{\mathcal{X}}_t,\hat{\boldsymbol{Y}}_t)\leftarrow f_{\text{PL}}(\mathcal{D}_t^{\text{train}},\eta,\boldsymbol{\Theta},\bar{\boldsymbol{W}}_{t-1})$
  ▷ *Count the labels frequency.* ◁
  $\boldsymbol{f}_t\leftarrow\text{Count}(\boldsymbol{Y}_t,\boldsymbol{f}_{t-1})$
  ▷ *Calculate the sample-specific weight.* ◁
  for all $(\boldsymbol{\mathcal{X}}_{t,i},\hat{\boldsymbol{y}}_{t,i})\in\hat{\mathcal{D}}_t^{\text{train}}$ **do**
    $\omega_{t,i}\leftarrow\text{Weight}(\boldsymbol{f}_t,\hat{\boldsymbol{y}}_{t,i})$
  $\boldsymbol{\Omega}_t\leftarrow\text{diag}(\omega_{t,1},\omega_{t,2},\cdots,\omega_{t,N_t})$
  ▷ *Feature extraction.* ◁
  $\boldsymbol{X}_t\leftarrow\phi(\boldsymbol{\mathcal{X}}_t,\boldsymbol{\Theta})$
  Update $\boldsymbol{R}_t$ with Equation (13)
  Update $\bar{\boldsymbol{W}}_t$ with Equation (12)
  Return: $\bar{\boldsymbol{W}}_t$

Experiments

Method	Type	Memory	MS-COCO B0-C10				MS-COCO B40-C10			
			Avg.	Last			Avg.	Last		
			mAP	CF1	OF1	mAP	mAP	CF1	OF1	mAP
Upper-bound	Baseline	0	-	76.4	79.4	81.8	-	76.4	79.4	81.8
FT (Ridnik et al., 2021)	Baseline		38.3	6.1	13.4	16.9	35.1	6.0	13.6	17.0
TPCIL (Tao et al., 2020)	CIL	5/class	63.8	20.1	21.6	50.8	63.1	25.3	25.1	53.1
PODNet (Douillard et al., 2020)	CIL		65.7	13.6	17.3	53.4	65.4	24.2	23.4	57.8
DER++ (Buzzega et al., 2020)	CIL		68.1	33.3	36.7	54.6	69.6	41.9	43.7	59.0
KRT-R (Dong et al., 2023)	MLCIL		75.8	60.0	61.0	68.3	78.0	66.0	65.9	74.3
CSC-R (Du et al., 2025)	MLCIL		79.2	67.3	68.1	73.7	78.4	68.1	69.0	75.6
iCaRL (Rebuffi et al., 2017)	CIL	20/class	59.7	19.3	22.8	43.8	65.6	22.1	25.5	55.7
BiC (Wu et al., 2019)	CIL		65.0	31.0	38.1	51.1	65.5	38.1	40.7	55.9
ER (Rolnick et al., 2019)	CIL		60.3	40.6	43.6	47.2	68.9	58.6	61.1	61.6
TPCIL (Tao et al., 2020)	CIL		69.4	51.7	52.8	60.6	72.4	60.4	62.6	66.5
PODNet (Douillard et al., 2020)	CIL		70.0	45.2	48.7	58.8	71.0	46.6	42.1	64.2
DER++ (Buzzega et al., 2020)	CIL		72.7	45.2	48.7	63.1	73.6	51.5	53.5	66.3
KRT-R (Dong et al., 2023)	MLCIL		76.5	63.9	64.7	70.2	78.3	67.9	68.9	75.2
CSC-R (Du et al., 2025)	MLCIL		79.6	67.8	68.6	74.8	78.7	68.2	69.4	76.7
PRS (Kim et al., 2020)	MLCIL	1000	48.8	8.5	14.7	27.9	50.8	9.3	15.1	33.2
OCDM (Liang & Li, 2022)	MLCIL		49.5	8.6	14.9	28.5	51.3	9.5	15.5	34.0
KRT-R (Dong et al., 2023)	MLCIL		75.7	61.6	63.6	69.3	78.3	67.5	68.5	75.1
CSC-R (Du et al., 2025)	MLCIL		79.3	67.5	68.5	73.9	78.5	67.8	69.7	76.0
PODNet (Douillard et al., 2020)	CIL	0	43.7	7.2	14.1	25.6	44.3	6.8	13.9	24.7
oEWC (Schwarz et al., 2018)	CIL		46.9	6.7	13.4	24.3	44.8	11.1	16.5	27.3
LWF (Li & Hoiem, 2017)	CIL		47.9	9.0	15.1	28.9	48.6	9.5	15.8	29.9
KRT (Dong et al., 2023)	MLCIL		74.6	55.6	56.5	65.9	77.8	64.4	63.4	74.0
CSC (Du et al., 2025)	MLCIL		78.0	64.9	66.8	72.8	78.2	65.7	67.0	75.0
L3A	MLCIL	0	**81.5**	**69.7**	**72.5**	**77.6**	**79.9**	**69.5**	**72.9**	**78.8**

$$P_k = \frac{TP_k}{TP_k + FP_k}$$

$$R_k = \frac{TP_k}{TP_k + FN_k}$$

$$CP = \frac{1}{K}\sum_k P_k$$

$$CR = \frac{1}{K}\sum_k R_k$$

$$CF1 = \frac{2CP \cdot CR}{CP + CR}$$

$$OP = \frac{\sum_k TP_k}{\sum_k (TP_k + FP_k)}$$

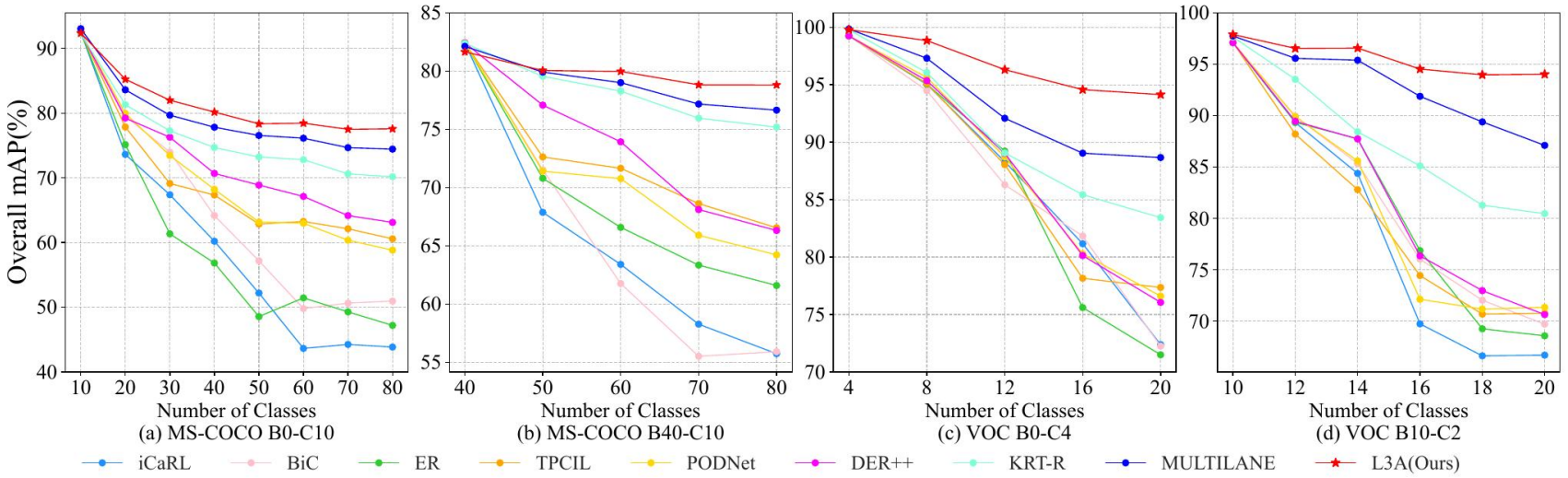
$$OR = \frac{\sum_k TP_k}{\sum_k (TP_k + FN_k)}$$

$$OF1 = \frac{2OP \cdot OR}{OP + OR}$$

Experiments

Table 2. Comparison of the results between L3A and prompt-based CIL methods on MS-COCO dataset. Data **in bold** represents the **best** results and data underline represents the second-best results.

Method	Backbone	Param.	MS-COCO B0-C10		MS-COCO B40-C10	
			Avg. mAP	Last mAP	Avg. mAP	Last mAP
Upper-bound	ViT-B/16	86.0M	-	83.2	-	83.2
L2P (Wang et al., 2022)			73.4	68.0	73.7	71.1
CODA-P (Smith et al., 2023)			74.0	65.4	73.9	67.5
MULTI-LANE (De Min et al., 2024)			79.1	74.5	78.8	76.6
L3A	TResNet-M	29.4M	**81.5**	**77.6**	**79.9**	**78.8**



即使使用更小的 TResNet-M，L3A 仍优于 ViT-B/16 的 MULTI-LANE：COCO 两个协议 Last mAP +3.1 / +2.2。

Experiments

Method	Memory	VOC B0-C4		VOC B10-C2	
		Avg. mAP	Last mAP	Avg. mAP	Last mAP
Upper-bound	0	-	94.7	-	94.7
FT		82.1	62.9	70.1	43.0
iCaRL	2/class	87.2	72.4	79.0	66.7
BiC		86.8	72.2	81.7	69.7
ER		86.1	71.5	81.5	68.6
TPCIL		87.6	77.3	80.7	70.8
PODNet		88.1	76.6	81.2	71.4
DER++		87.9	76.1	82.3	70.6
KRT-R		90.7	83.4	87.7	80.5
CSC-R		92.4	87.9	91.6	87.8
CSC	0	90.4	85.1	89.0	83.8
CODA-P		90.6	84.5	90.2	85.0
MULTI-LANE		<u>93.5</u>	<u>88.8</u>	<u>93.1</u>	<u>88.3</u>
L3A	0	**96.7**	**94.1**	**95.6**	**94.0**

Model	PL	WAC	COCO B0-C10		COCO B40-C10	
			Avg. mAP	Last mAP	Avg. mAP	Last mAP
Baseline	×	×	48.5	25.6	45.9	26.3
(1) w/o WAC	✓	×	70.9	56.7	77.7	72.1
(2) w/o PL	×	✓	81.4	77.3	79.6	78.4
L3A	✓	✓	**81.5**	**77.6**	**79.9**	**78.8**

Experiments

Ablation Study

Table 5. The ablation study on regularization term (γ).

γ	COCO B0-C10		COCO B40-C10	
	Avg. mAP	Last mAP	Avg. mAP	Last mAP
0.1	80.41	76.83	79.36	78.22
1	80.39	76.90	79.37	78.27
10	80.49	76.90	79.38	78.28
100	80.86	77.09	79.51	78.29
1000	**81.38**	77.56	**79.89**	**78.75**
10000	81.22	**77.61**	79.67	78.69

Table 6. The ablation study on buffer layer size.

Buffer layer size	COCO B0-C10		COCO B40-C10	
	Avg. mAP	Last mAP	Avg. mAP	Last mAP
512	79.33	75.24	77.78	76.43
1024	80.47	76.52	78.89	77.67
2048	81.08	77.20	79.58	78.44
4096	81.38	77.55	79.91	78.77
6144	**81.43**	**77.60**	79.94	78.75
8192	**81.43**	77.56	**79.96**	78.79
12288	81.40	77.56	79.93	**78.82**

Table 7. The ablation study on confidence threshold (η).

η	COCO B0-C10		COCO B40-C10	
	Avg. mAP	Last mAP	Avg. mAP	Last mAP
0.55	71.25	67.21	78.30	76.12
0.6	78.63	74.21	79.52	78.09
0.65	80.97	77.08	79.85	78.68
0.7	**81.45**	77.57	**79.91**	**78.77**
0.75	81.36	**77.58**	79.80	78.62
0.8	81.34	77.57	79.69	78.47
Dynamic	81.12	76.56	79.86	78.45

Table 8. Different weighting mechanism in the WAC module.

Weighting mechanism	COCO B0-C10		COCO B40-C10	
	Avg. mAP	Last mAP	Avg. mAP	Last mAP
$1/\sqrt{f^{(k)}}$	**81.45**	**77.57**	**79.91**	**78.77**
$1/f^{(k)}$	81.13	77.30	79.60	78.27
$1/(\log(f^{(k)})+1)$	81.36	77.41	79.87	78.27

Thanks