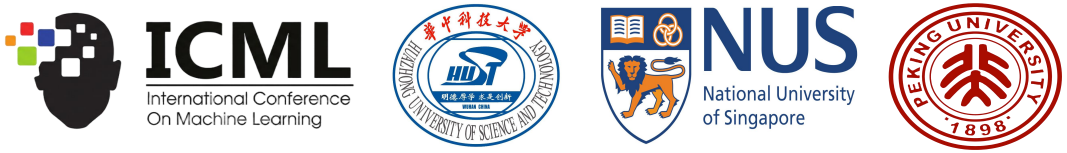
Learnability-Driven Knowledge Assimilation for Class-Incremental Semantic Segmentation

Low-margin regions bottleneck CISS — analysis and solution

Xinyue Zhang *[1], Xu Zou *[1], Wanjia Luo[2], Yanjie Wang[1], Jiahuan Zhou[3], Sheng Zhong[1], Luxin Yan[1]

[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
2Department of Physics, Faculty of Science, National University of Singapore, Singapore
3Wangxuan Institute of Computer Technology, Peking University, Beijing, China.

Contents OF Programme



01

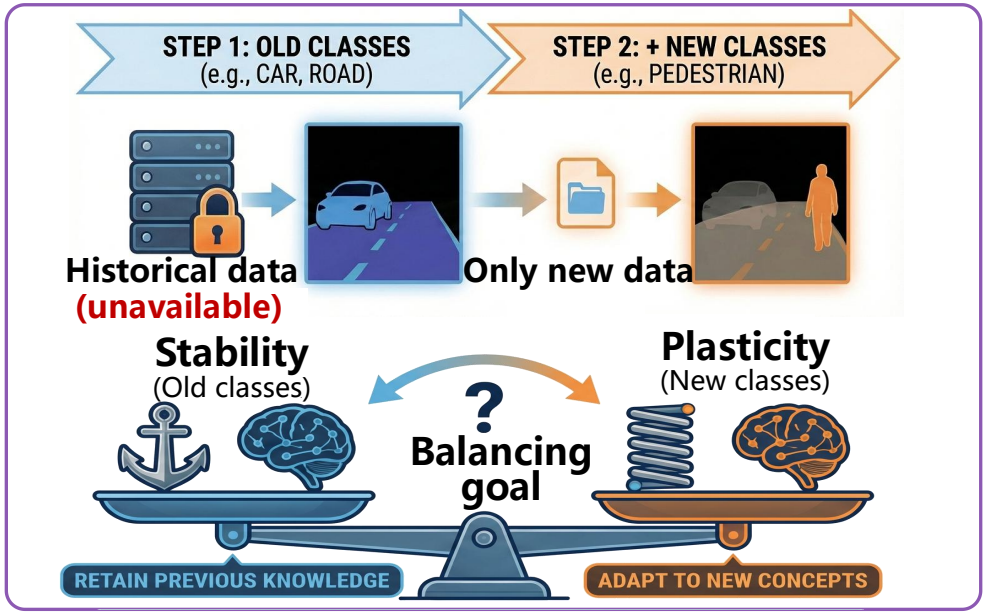
Task Definition and Related work

1 Task Definition and Related Work

ICML International Conference On Machine Learning

What is Class-Incremental Semantic Segmentation (CISS)?

Class-incremental semantic segmentation studies **how to incrementally learn new classes** over time **when historical training data are not fully accessible** due to privacy or storage constraints



Where can CISS be applied?

It is applied to fields such as **autonomous driving and robot perception**, where it plays a crucial role in dynamic environments and continual updates.

Autonomous driving

Robot perception

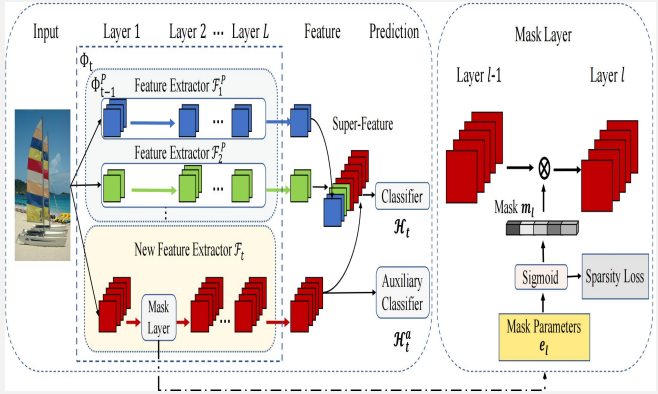
Crucial for dynamic environments & continuous updates.

1 Task Definition and Related Work

ICML International Conference On Machine Learning

Architecture Expansion

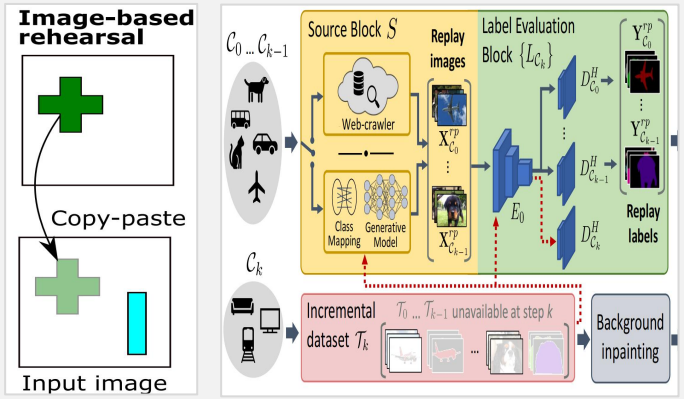
Dynamic network growth to accommodate new classes



Cons: Model size & computatioin cost increase with each steps

Replay-Based

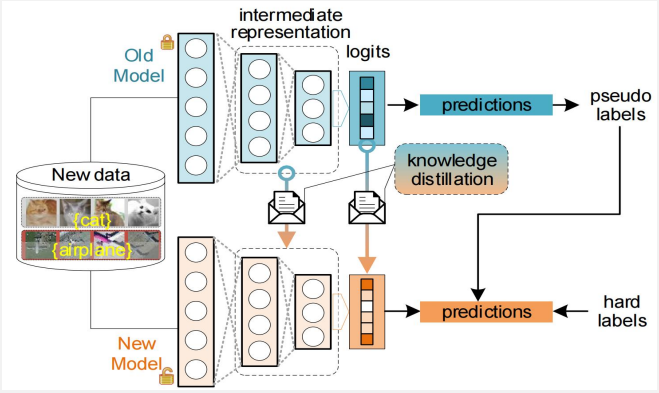
Store and generate samples from previous task.



Cons: Limited by privacy, memory, and generation-real mismatch.

Regularization-Based

Dominant approach: Adding constraints or penalties to the model (e.g., knowledge distillation, contrastive learning).



Cons: Bottleneck in low-margin regions during incremental learning.

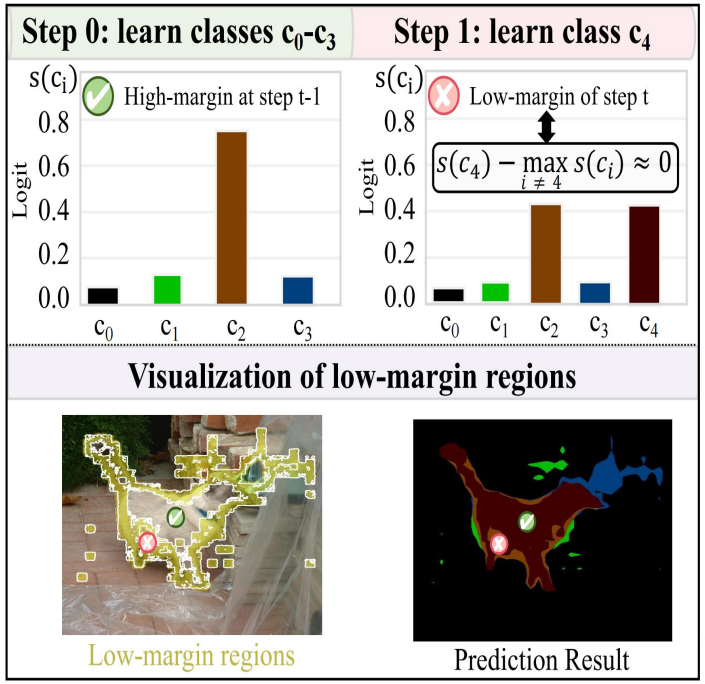
ICML International Conference On Machine Learning

NUS National University of Singapore

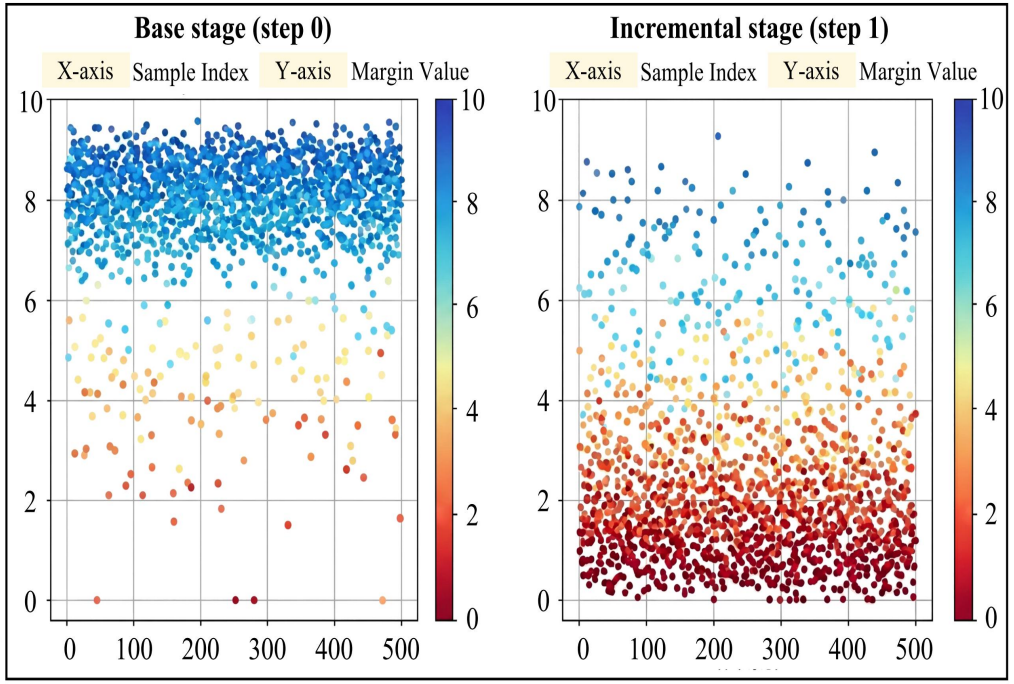
02

Problem Analysis

2 Problem Analysis



(a) Meaning of Low-Margin Regions.



(b) Distribution of Low-Margin Regions in Class-Incremental Semantic Learning.

Figure 1. **Low-margin regions bottleneck CISS.** (a) Pixels with a small logit margin (e.g., top-1 c4 vs. top-2 c2) form low-margin regions, where predictions are uncertain and easily flipped. (b) After incremental learning, more samples are distributed in the low-margin region, resulting in class confusion.

Definition of margion

Low-margin region definition. For an input x with training label $\tilde{y}$, let the most confident competing class be

$$c^*(x) = \arg\max_{j \neq y} s_j(x;\theta), \tag{2}$$

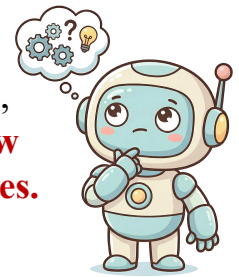
where y denotes the true class index. The logit margin is defined as:

$$m(x;\theta) = s_y(x;\theta) - s_{c^*(x)}(x;\theta). \tag{3}$$

Low-margin regions correspond to $m(x;\theta) \approx 0$, where the top-1 and top-2 logits are close, leading to high uncertainty and strong sensitivity to small perturbations.

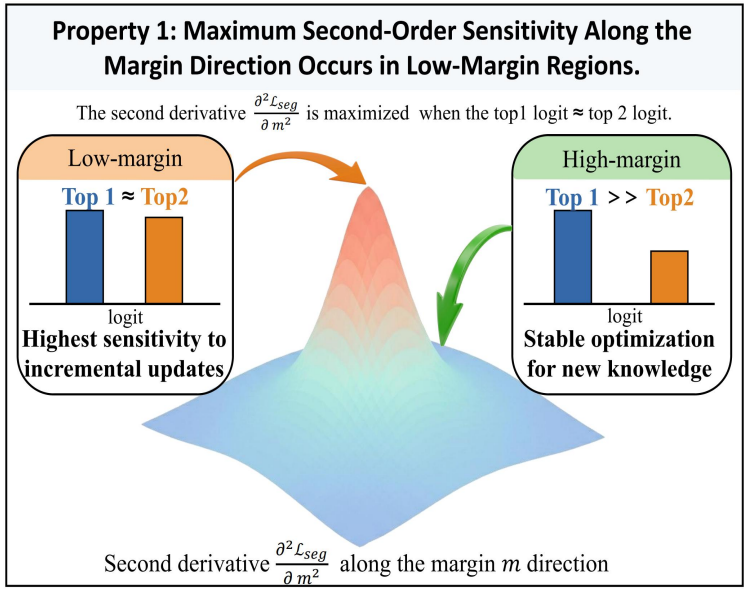
Impact of low-margion region

In these regions, the top-1 and top-2 logits are close, near an evolving decision boundary where predictions are uncertain. This **increases the likelihood of confusing** new classes with previously learned ones during incremental learning, **resulting in poor learning of new classes and forgetting of old ones.**



Although existing methods alleviate forgetting, new-class performance often remains limited. A key bottleneck arises from low-margin regions, where the top-1 and top-2 logits are highly competitive.

2 Problem Analysis



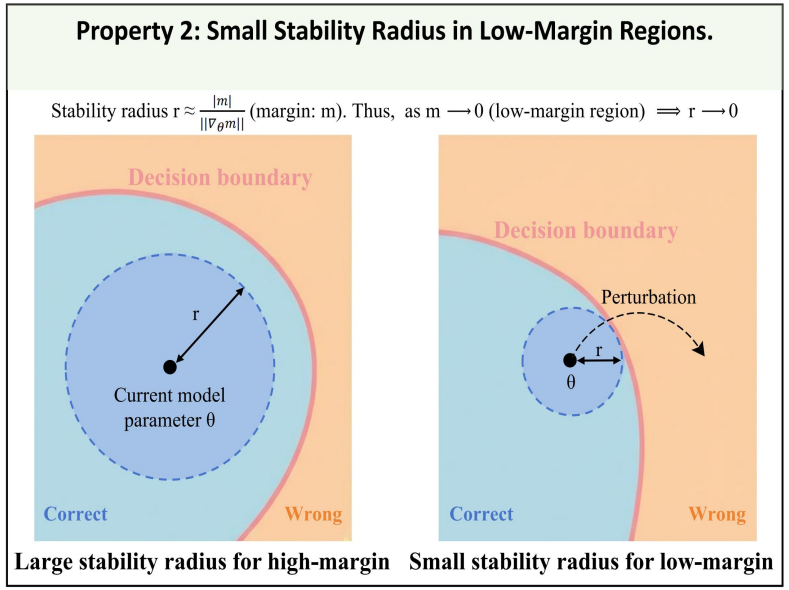


Figure 2. **Why low-margin regions bottleneck CISS.** Low margin regions exhibit high second-order sensitivity along the margin direction and a small stability radius, making them highly sensitive to parameter updates and thus prone to mis-segmentation during incremental learning of new classes.

Property 1: High directional second-order sensitivity

Property 1: Maximum Second-Order Sensitivity Along the Margin Direction Occurs in Low-Margin Regions. When learning new classes at step t, we adopt the softmax cross-entropy loss (Mao et al., 2023) as $\mathcal{L}_{seg}$. Considering the directional second-order sensitivity of $\mathcal{L}_{seg}$ along the margin direction $m(x, \theta)$, the directional second derivative under softmax is

$$\frac{\partial^2 \mathcal{L}_{seg}}{\partial m^2} = (p_y + p_{c^*}) - (p_y - p_{c^*})^2. \quad (3)$$

Here, p_y represents the predicted probability for the true class y, and p_{c^*} represents the predicted probability for the most confident competing class c^*, where c^* is the class with the second-highest logit. The upper bound of the second derivative is:

$$\frac{\partial^2 \mathcal{L}_{seg}}{\partial m^2} \lesssim 1, \quad (4)$$

and the Equation (3) is maximized near $m(x;\theta) \approx 0$. Consequently, low-margin regions exhibit the highest directional second-order sensitivity, such that small incremental updates can cause disproportionate shifts between competing logits, leading to performance degradation in CISS. Derivation details are provided in Appendix A.

Property 2: Small stabiliy radidus

Property 2: Small Stability Radius in Low-Margin Regions. A small perturbation $\Delta\theta$ changes the margin according to the first-order Taylor approximation (Pötzsche & Rasmussen, 2006):

$$m(x;\theta+\Delta\theta) = m(x;\theta) + \nabla_\theta m(x;\theta)^\top \Delta\theta + \mathcal{O}\left(\|\Delta\theta\|^2\right), \quad (5)$$

where $\nabla_\theta m(x;\theta)$ is the gradient of the logit margin with respect to the parameters. The stability radius r is the minimum perturbation required to reach the decision boundary:

$$r(x;\theta) \triangleq \inf_{\Delta\theta} \|\Delta\theta\| \quad \text{s.t.} \quad m(x;\theta)\, m(x;\theta+\Delta\theta) \le 0. \quad (6)$$

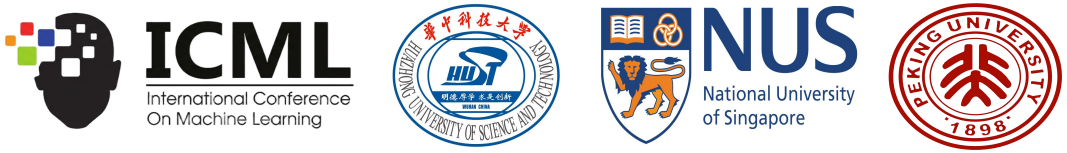
Any boundary-reaching perturbation satisfies

$$\|\Delta\theta\| \ge \frac{|m(x;\theta)|}{\|\nabla_\theta m(x;\theta)\|}, \quad (7)$$

and thus $r(x;\theta) \approx \frac{|m(x;\theta)|}{\|\nabla_\theta m(x;\theta)\|}$. When $m(x;\theta) \approx 0$, the stability radius is small. Consequently, during later incremental steps, learning new classes becomes highly sensitive to small parameter perturbations, which can flip predictions on old classes. Derivation details are provided in Appendix B.

Detailed introduction can be found in Section 3.2, with inference details in Appendix A and B.

We demonstrate and reveal that the difficulty of optimizing low-margin regions in CISS arises from two key properties: high directional second-order sensitivity and low stability radius. This insight inspires us to propose a method based on this critical finding.


ICML
International Conference
On Machine Learning
NUS
National University
of Singapore
PEKING UNIVERSITY
1898

03

Method

3 Method Learnability-Driven Knowledge Assimilation (LDKA)

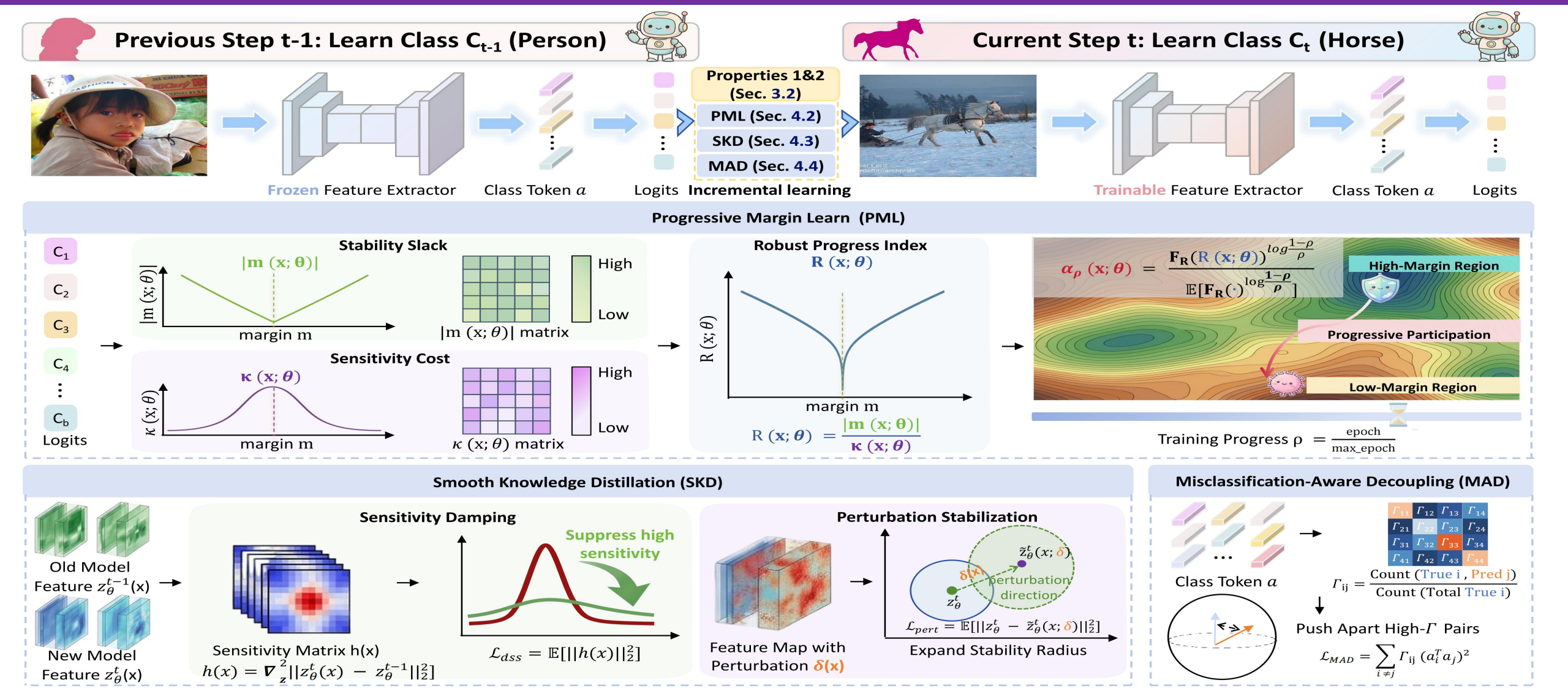
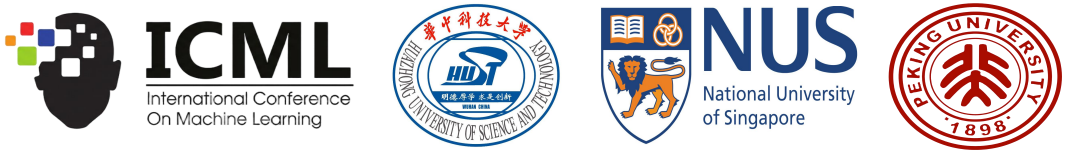


Figure 3. Overview of Learnability-Driven Knowledge Assimilation (LDKA). At step t, guided by the derived low-margin properties (Sec. 3.2), LDKA addresses low-margin optimization via three complementary strategies: PML determines what to learn by progressively allocating pixel-level optimization budget, SKD determines how to learn by stabilizing low-margin updates, and MAD determines where to separate by decoupling highly confused class pairs. Together, they provide an allocation–stabilization–decoupling optimization design.



04

Comprehensive Experiments

4 Comparison with the State-of-the-arts

ICML International Conference On Machine Learning

SOTA across 9 settings.

Comparison of the mean mIoU across all classes with recent methods shows that our method consistently achieves SOTA performance.

Table 1. Comparative experiments on VOC dataset. **Bold** and underlined values denote the best and second-best results, respectively. The † symbol indicates results reproduced using the same versioin of ViT. – indicates that the relevant experiment is not mentioned in the original papers. Across five incremental settings, our method achieves the highest overall mIoU, demonstrating consistent performance.

Method	Backbone	15-5 (2 steps)			19-1 (2 steps)			15-1 (6 steps)			2-2 (10 steps)			10-1 (11 steps)		
		0-15	16-20	All	0-19	20	All	0-15	16-20	All	0-2	3-20	All	0-10	11-20	All
MIB (Cermelli et al., 2020)	ResNet101	76.4	50.0	70.1	71.4	23.6	69.1	34.2	13.5	29.3	41.1	23.4	25.9	12.3	13.1	12.7
PLOP (Douillard et al., 2021)	ResNet101	75.7	51.7	70.0	75.4	37.4	73.6	65.1	21.1	54.6	24.1	11.9	13.6	44.0	15.5	30.4
SSUL (Cha et al., 2021)	ResNet101	78.4	55.8	73.0	77.8	49.8	76.5	78.4	49.0	71.4	–	–	–	74.0	53.2	64.1
MicroSeg (Zhang et al., 2022b)	ResNet101	82.0	59.2	76.6	79.3	62.9	78.5	81.3	52.5	74.4	60.0	50.9	52.2	77.2	57.2	67.7
RCIL (Zhang et al., 2022a)	ResNet101	78.8	52.0	72.4	68.5	12.1	65.8	70.6	23.7	59.4	28.3	19.0	20.3	55.4	15.1	36.2
LGKD (Yang et al., 2023)	ResNet101	79.5	54.8	73.6	77.3	42.9	75.7	70.6	30.9	61.1	–	–	–	–	–	–
CoMasTRe (Gong et al., 2024)	ResNet101	79.7	51.9	73.1	75.1	69.5	74.8	69.8	43.6	63.6	–	–	–	–	–	–
Adapter (Zhu et al., 2025b)	ResNet101	–	–	–	–	–	–	79.9	51.9	73.2	62.8	57.9	58.6	74.9	54.3	65.1
MIB† (Cermelli et al., 2020)	ViT	78.5	63.2	74.9	80.4	47.8	78.8	72.6	23.5	60.9	41.1	29.3	31.0	11.4	18.9	15.0
SSUL† (Cha et al., 2021)	ViT	79.7	55.3	73.9	80.8	31.5	78.5	78.1	33.4	67.5	60.3	40.6	43.4	74.3	51.0	63.2
MicroSeg† (Zhang et al., 2022b)	ViT	81.9	54.0	75.3	79.0	25.3	76.4	80.5	40.8	71.0	64.8	43.4	46.5	73.5	53.0	63.7
Incrementer (Shang et al., 2023a)	ViT	–	–	–	–	–	–	79.6	59.6	74.8	–	–	–	77.6	60.3	69.4
CoinSeg (Zhang et al., 2023)	ViT	82.1	63.2	77.6	81.5	44.8	79.8	82.7	52.5	75.5	70.1	63.3	64.3	80.1	60.0	70.5
MBS† (Park et al., 2024)	ViT	83.9	72.6	81.2	82.2	72.6	81.8	81.9	65.6	78.0	67.5	73.4	72.6	80.0	72.9	76.6
Nest (Xie et al., 2024)	ViT	81.2	67.4	77.9	79.7	60.0	78.8	77.0	53.3	71.4	–	–	–	65.2	35.8	51.2
CoGaMiD (Zhu et al., 2025a)	ViT	–	–	–	–	–	–	83.2	61.2	77.8	**73.4**	70.0	70.5	**81.1**	65.9	73.9
Ours	**ViT**	**84.6**	**75.3**	**82.4**	**82.9**	**73.9**	**82.5**	**83.3**	**69.9**	**80.1**	72.3	**75.8**	**75.3**	80.0	**73.1**	**76.7**
Joint (Upper bound)	ViT	85.5	80.3	84.3	84.4	79.6	84.2	83.9	79.1	82.8	77.3	85.5	84.3	85.0	84.7	84.9

SOTA across five incremental configurations on the VOC dataset

Table 2. Comparative experiments on ADE20K. Our method is capable of effectively learning new knowledge and resisting catastrophic forgetting without accessing old-class data for rehearsal. Notably, the overall performance of our method across the four incremental settings is very close to that of joint training, which is commonly regarded as the upper bound of performance in CISS.

Method	Backbone	100–50 (2 steps)			50–50 (3 steps)			100–10 (6 steps)			100–5 (11 steps)		
		0-100	101-150	All	0-50	51-150	All	0-100	101-150	All	0-100	101-150	All
SDR (Michieli & Zanuttigh, 2021b)	ResNet101	37.5	25.5	33.5	42.9	25.4	31.3	28.9	11.7	23.2	36.7	5.7	26.4
PLOP (Douillard et al., 2021)	ResNet101	41.9	14.9	33.0	48.8	21.0	30.4	40.5	13.6	31.6	39.1	7.8	28.7
SSUL (Cha et al., 2021)	ResNet101	41.3	18.0	33.6	48.4	20.2	29.7	40.2	18.8	33.1	39.9	17.4	32.4
REMINDER (Phan et al., 2022)	ResNet101	41.6	19.2	34.2	47.1	20.4	29.4	39.0	21.3	33.1	36.1	16.4	29.6
Microseg (Zhang et al., 2022b)	ResNet101	40.2	18.8	33.1	48.6	24.8	32.8	41.5	21.6	34.9	40.4	20.5	33.8
IDEC (Zhao et al., 2023b)	ResNet101	42.0	18.2	34.1	47.4	26.0	33.2	42.3	17.6	34.1	39.2	14.6	31.1
LGKD (Yang et al., 2023)	ResNet101	43.4	25.7	37.5	48.9	29.4	36.0	41.9	22.0	35.3	-	-	-
LAG (Yuan et al., 2024)	ResNet101	41.6	19.7	34.3	47.7	26.1	33.4	41.0	18.7	33.6	40.0	17.2	32.5
CoMasTRe (Gong et al., 2024)	ResNet101	45.7	26.0	39.2	-	-	-	42.3	18.4	34.4	40.8	15.8	32.5
Adapter (Zhu et al., 2025b)	ResNet101	43.1	23.6	36.6	49.3	27.3	34.7	42.9	19.9	35.3	42.6	18.0	34.5
MIB† (Cermelli et al., 2020)	ViT	46.4	35.0	42.6	52.2	35.6	41.2	43.0	30.8	39.0	40.2	26.6	35.7
SSUL† (Cha et al., 2021)	ViT	41.9	20.1	34.7	49.5	21.3	30.8	40.7	19.0	33.5	41.3	16.0	32.9
Microseg† (Zhang et al., 2022b)	ViT	41.1	24.1	35.5	49.8	23.9	32.6	41.0	22.6	34.9	41.2	21.0	34.5
Coinseg (Zhang et al., 2023)	ViT	41.6	26.7	36.7	49.0	28.9	35.7	42.1	24.5	36.3	43.1	24.1	36.8
CoMFormer (Cermelli et al., 2023)	ViT	44.7	26.2	38.6	-	-	-	40.6	15.6	32.3	39.5	13.6	30.9
MBS† (Park et al., 2024)	ViT	**49.4**	37.5	45.5	55.6	39.8	45.1	48.1	35.2	43.8	45.7	22.7	38.1
Nest (Xie et al., 2024)	ViT	42.8	27.8	37.8	49.7	29.3	36.2	41.8	23.8	35.8	40.5	19.9	33.7
CoGaMiD (Zhu et al., 2025a)	ViT	43.9	27.3	38.4	49.9	29.8	36.6	**49.9**	26.5	42.2	43.6	25.8	37.7
Ours	ViT	49.1	**40.7**	**46.3**	**55.8**	**42.3**	**46.9**	48.0	**38.5**	**44.9**	**46.1**	**28.1**	**40.1**
Joint (Upper bound)	ViT	52.9	42.6	49.5	58.9	44.7	49.5	52.7	42.4	49.3	52.6	42.6	49.3

SOTA across four incremental configurations on the ADE20K dataset

Quantitative analysis

4 Comparison with the State-of-the-arts

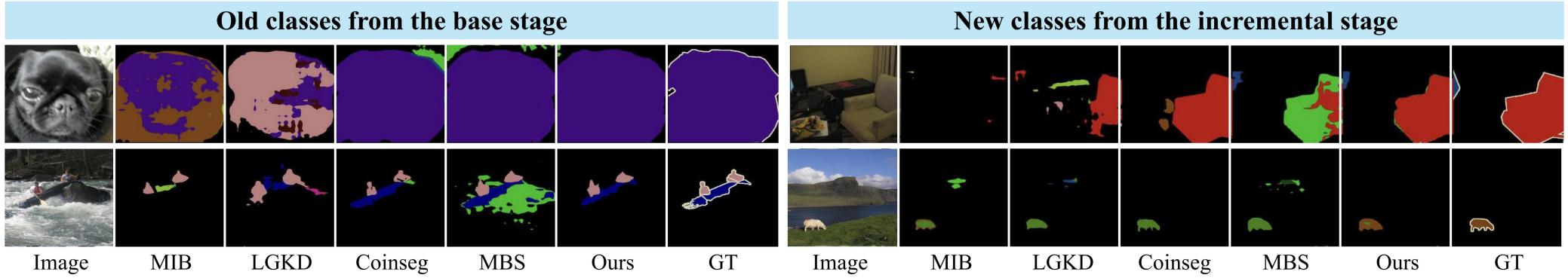


Figure 4. Qualitative results under the 15-1 setting. Results from the base stage and the incremental stage show that our method achieves **more accurate pixel-level segmentation of old classes with resistance to forgetting**, while **reducing misclassification of new-class pixels**.

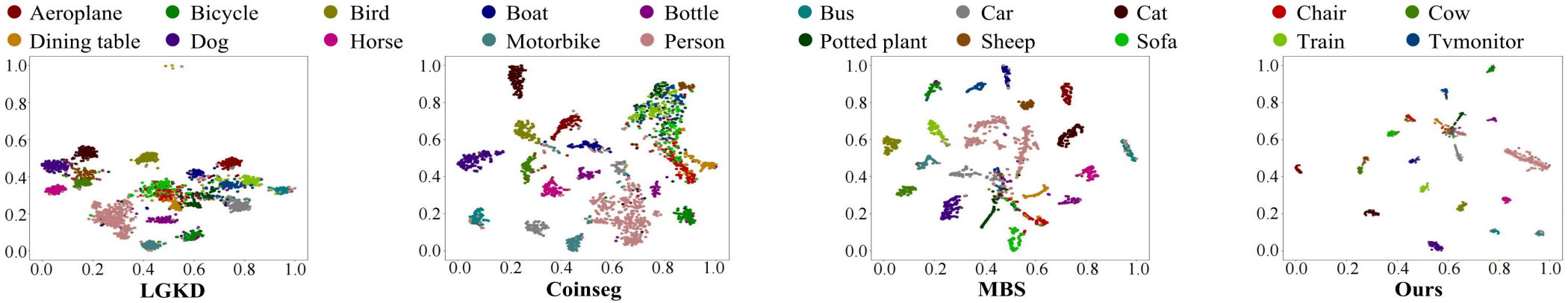


Figure 5. Qualitative analysis under the 15-1 configuration using the T-SNE plot. Our method shows **better intra-class clustering** than recent approaches, suggesting greater potential for accommodating new classes in the future, which is crucial for CISS.

Qualitative Analysis

4 Ablation studies

Effectiveness of Component Analysis in LDKA

Table 3. Ablation study of components on Pascal VOC 15-1. Integrating all components yields the best performance.

Num	Baseline	$\mathcal{L}_{\text{seg-PML}}$	$\mathcal{L}_{\text{SKD}}$	$\mathcal{L}_{\text{MAD}}$	15-1 (6 steps)		
					0–15	16–20	All
1	✓				71.8	45.0	65.4
2	✓	✓			82.2	61.4	77.2
3	✓	✓	✓		83.0	68.4	79.5
4	✓	✓	✓	✓	83.3	69.9	80.1

Each component plays an irreplaceable role, and the optimal performance is achieved when all components are involved.

Ablation about the design of $\mathcal{L}_{MAD}$

Table 4. Ablation study about the design of $\mathcal{L}_{\text{MAD}}$. $\mathcal{L}_{\text{MAD}}$ is more suitable for CISS than recent inter-class decoupling methods.

	15-1 (6 steps)		
	0–15	16–20	All
Contrastive learning	82.0	67.9	78.5
Contrastive learning with Γ_{ij}	82.2	68.6	78.8
Orthogonal	82.0	63.4	77.4
Ours	82.7	69.9	79.5

The ablation study with different class decoupling strategies shows that MAD outperforms all variants.

4 Ablation studies

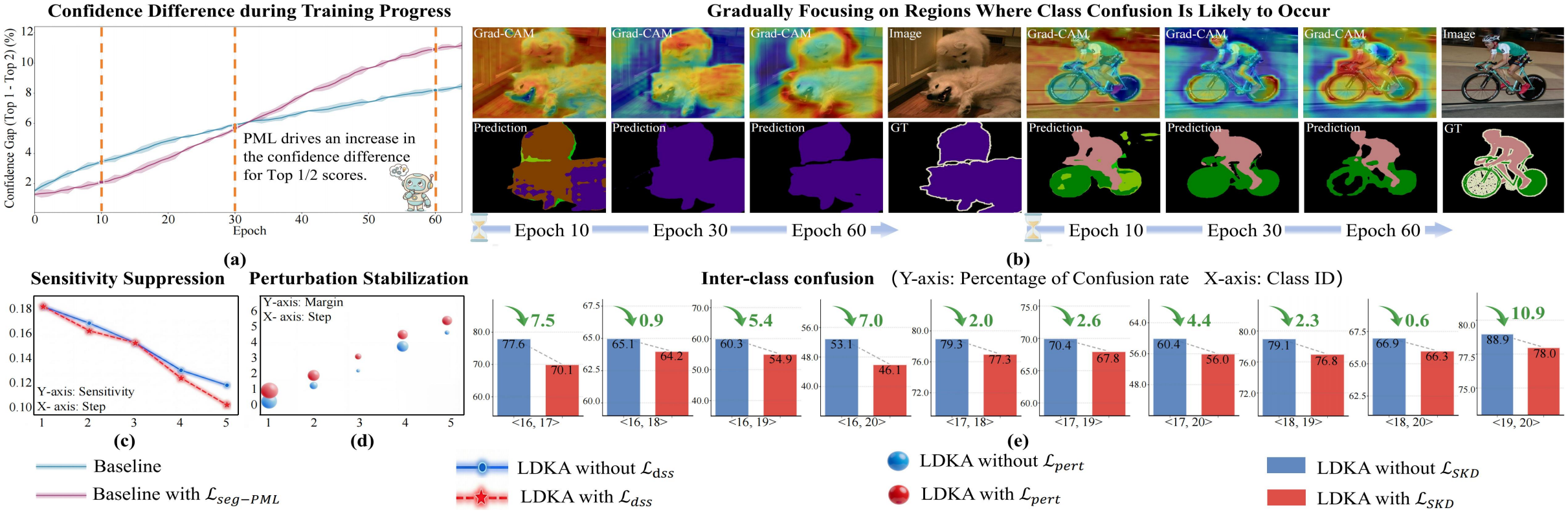


Figure 6. (a)–(b) analyze the effect of $\mathcal{L}_{seg-PML}$; (c)–(e) analyze $\mathcal{L}_{dss}$, $\mathcal{L}_{pert}$, and $\mathcal{L}_{SKD}$, respectively. $\mathcal{L}_{seg-PML}$ enlarges the top-1/top-2 confidence gap, while $\mathcal{L}_{dss}$ and $\mathcal{L}_{pert}$ suppress directional second-order sensitivity and increase the stability radius. $\mathcal{L}_{SKD}$ further mitigates class confusion.

Ablation studies

4 Ablation studies

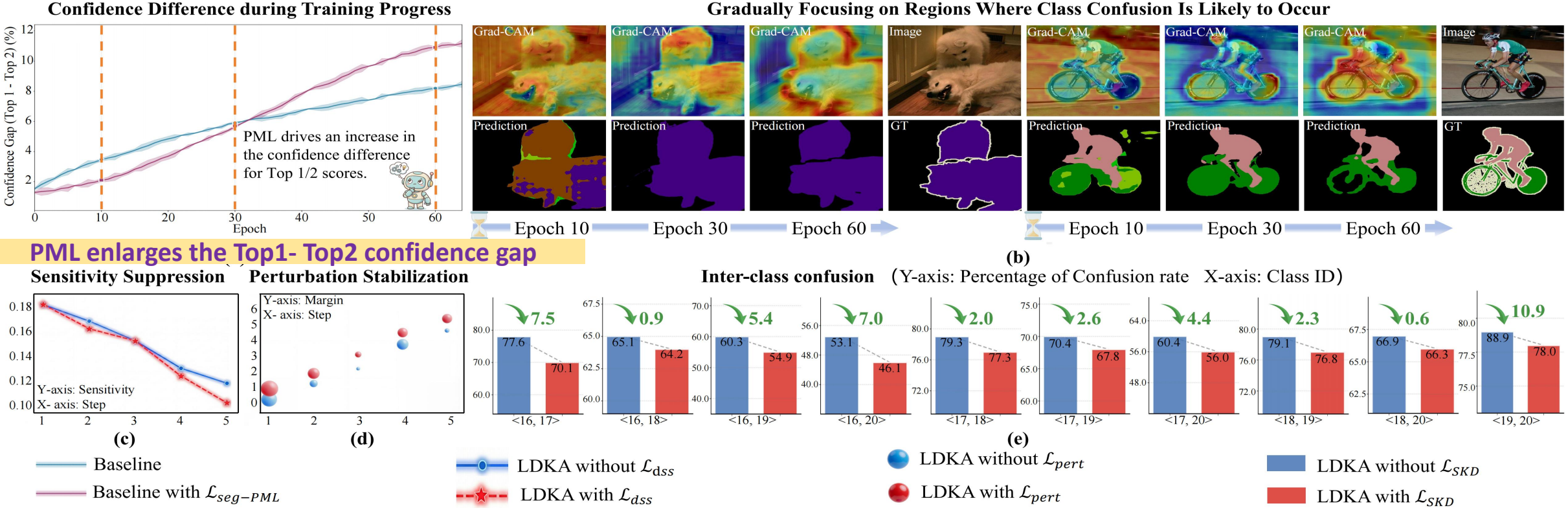


Figure 6. (a)–(b) analyze the effect of $\mathcal{L}_{seg-PML}$; (c)–(e) analyze $\mathcal{L}_{dss}$, $\mathcal{L}_{pert}$, and $\mathcal{L}_{SKD}$, respectively. $\mathcal{L}_{seg-PML}$ enlarges the top-1/top-2 confidence gap, while $\mathcal{L}_{dss}$ and $\mathcal{L}_{pert}$ suppress directional second-order sensitivity and increase the stability radius. $\mathcal{L}_{SKD}$ further mitigates class confusion.

4 Ablation studies

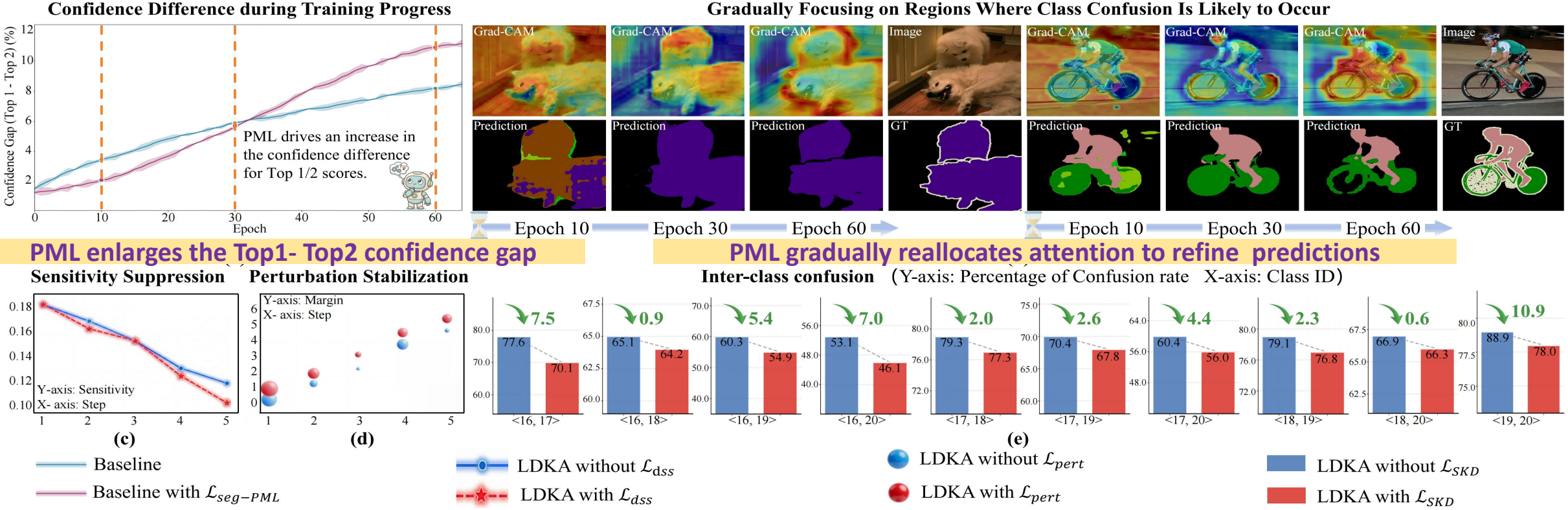
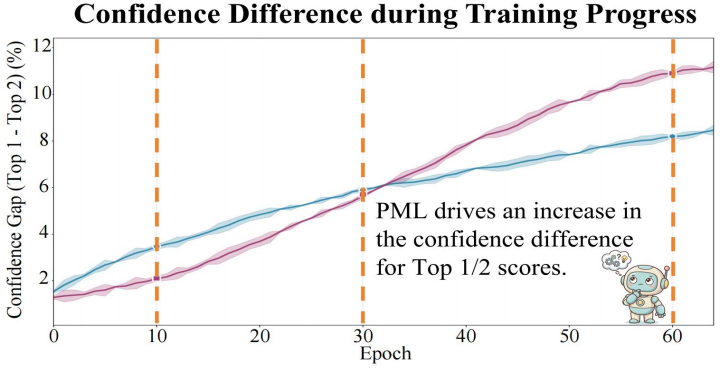
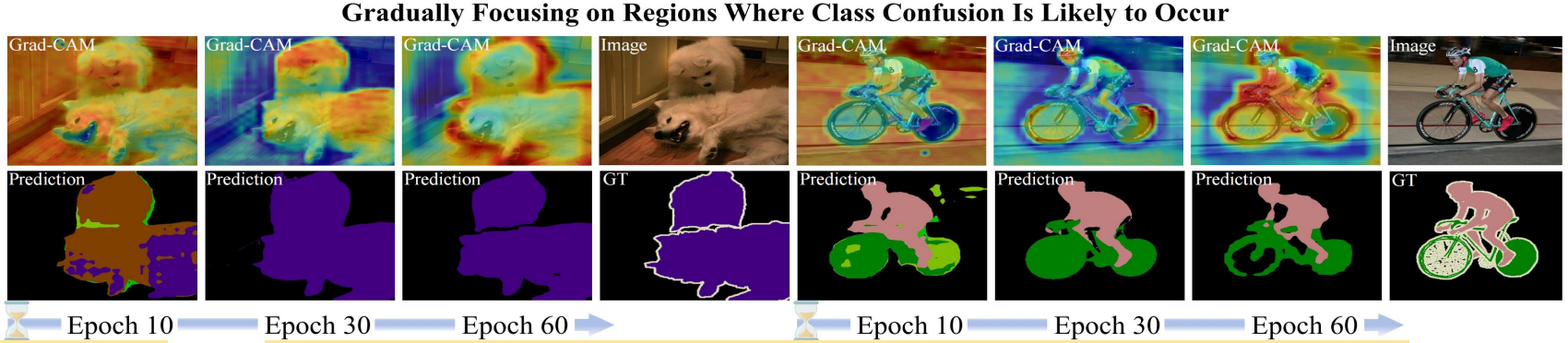


Figure 6. (a)–(b) analyze the effect of $\mathcal{L}_{seg-PML}$; (c)–(e) analyze $\mathcal{L}_{dss}$, $\mathcal{L}_{pert}$, and $\mathcal{L}_{SKD}$, respectively. $\mathcal{L}_{seg-PML}$ enlarges the top-1/top-2 confidence gap, while $\mathcal{L}_{dss}$ and $\mathcal{L}_{pert}$ suppress directional second-order sensitivity and increase the stability radius. $\mathcal{L}_{SKD}$ further mitigates class confusion.

Ablation studies

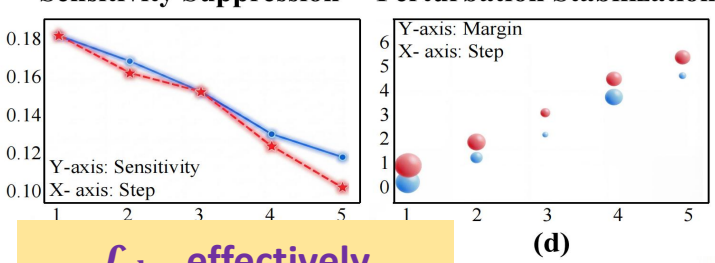
4 Ablation studies



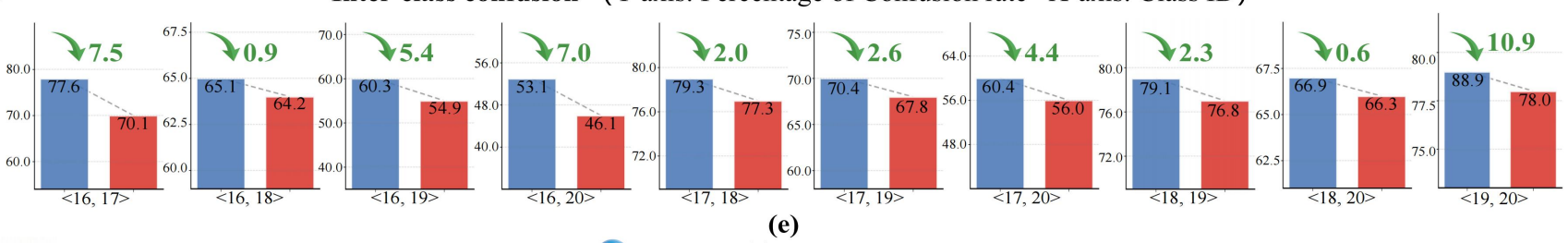


PML enlarges the Top1- Top2 confidence gap

PML gradually reallocates attention to refine predictions



$\mathcal{L}_{dss}$ effectively suppresses sensitivity.



LDKA without $\mathcal{L}_{dss}$ LDKA with $\mathcal{L}_{dss}$

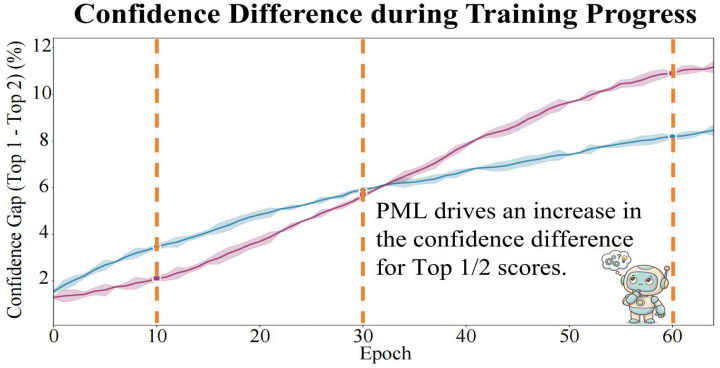
LDKA without $\mathcal{L}_{pert}$ LDKA with $\mathcal{L}_{pert}$

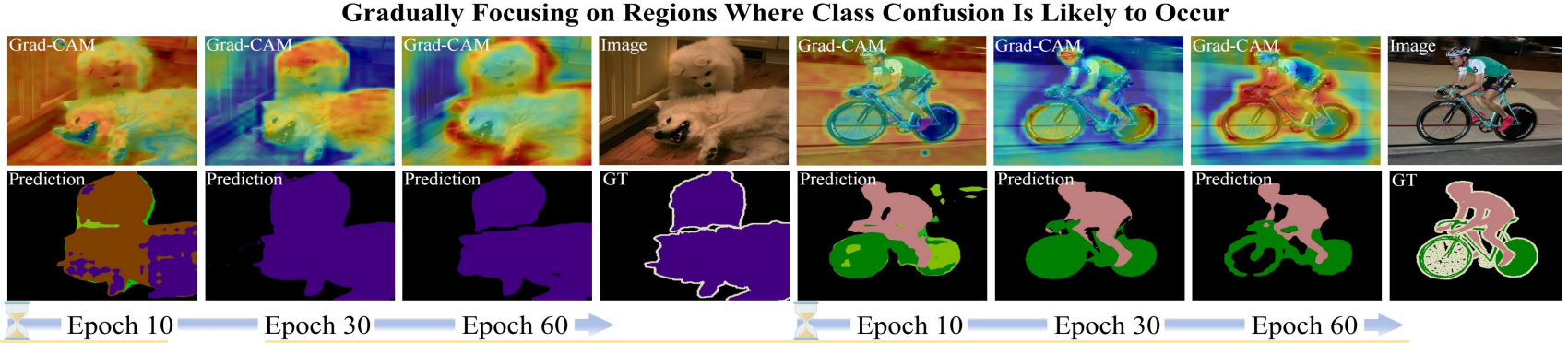
LDKA without $\mathcal{L}_{SKD}$ LDKA with $\mathcal{L}_{SKD}$

Figure 6. (a)–(b) analyze the effect of $\mathcal{L}_{seg-PML}$; (c)–(e) analyze $\mathcal{L}_{dss}$, $\mathcal{L}_{pert}$, and $\mathcal{L}_{SKD}$, respectively. $\mathcal{L}_{seg-PML}$ enlarges the top-1/top-2 confidence gap, while $\mathcal{L}_{dss}$ and $\mathcal{L}_{pert}$ suppress directional second-order sensitivity and increase the stability radius. $\mathcal{L}_{SKD}$ further mitigates class confusion.

Ablation studies

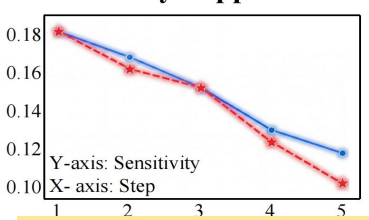
4 Ablation studies

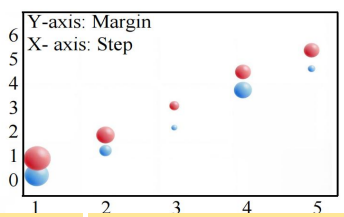


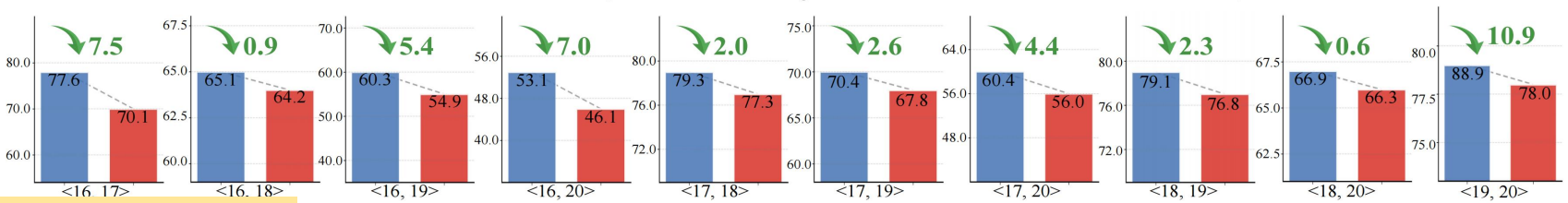


PML enlarges the Top1- Top2 confidence gap

PML gradually reallocates attention to refine predictions





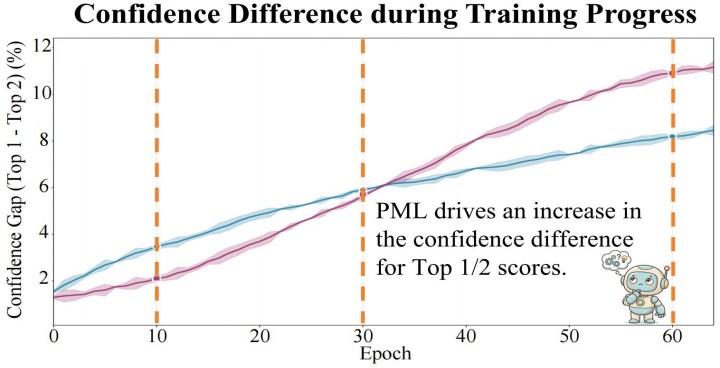


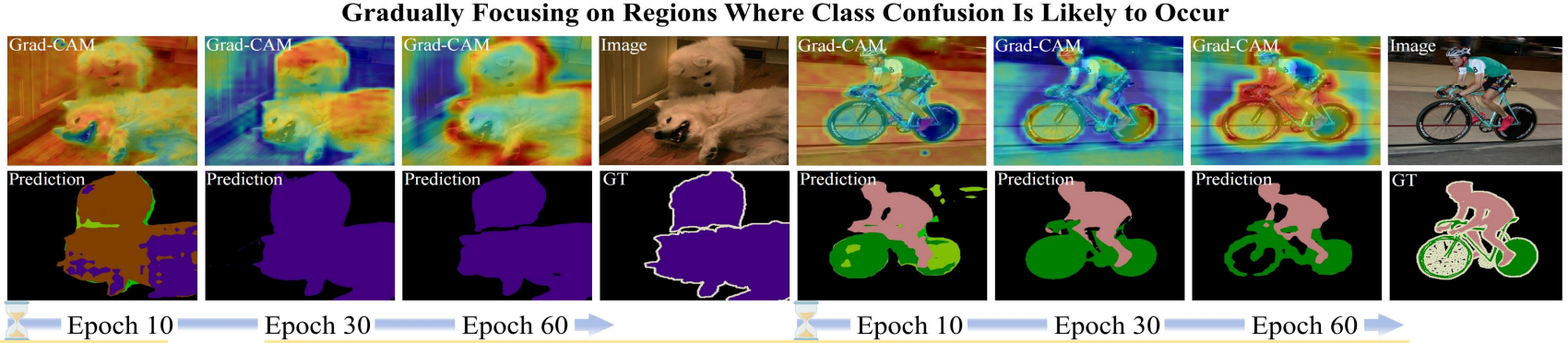
$\mathcal{L}_{dss}$ effectively suppresses sensitivity.

$\mathcal{L}_{pert}$ effectively increases the disturbance radius and margin

Figure 6. (a)–(b) analyze the effect of $\mathcal{L}_{seg-PML}$; (c)–(e) analyze $\mathcal{L}_{dss}$, $\mathcal{L}_{pert}$, and $\mathcal{L}_{SKD}$, respectively. $\mathcal{L}_{seg-PML}$ enlarges the top-1/top-2 confidence gap, while $\mathcal{L}_{dss}$ and $\mathcal{L}_{pert}$ suppress directional second-order sensitivity and increase the stability radius. $\mathcal{L}_{SKD}$ further mitigates class confusion.

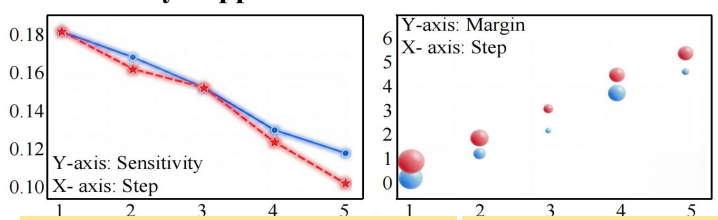
4 Ablation studies

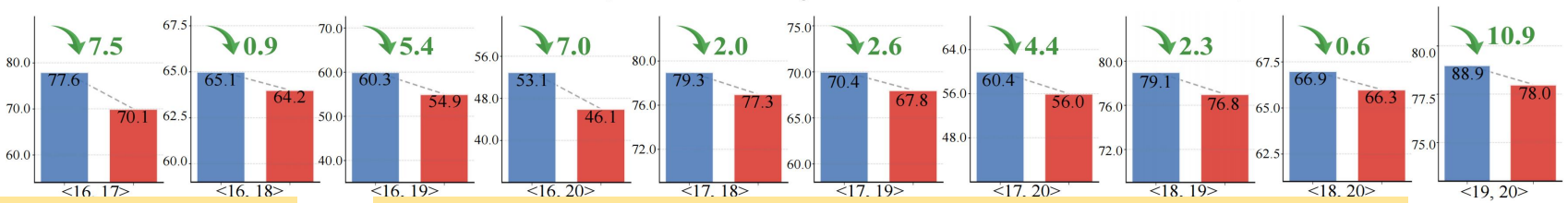




PML enlarges the Top1- Top2 confidence gap

PML gradually reallocates attention to refine predictions





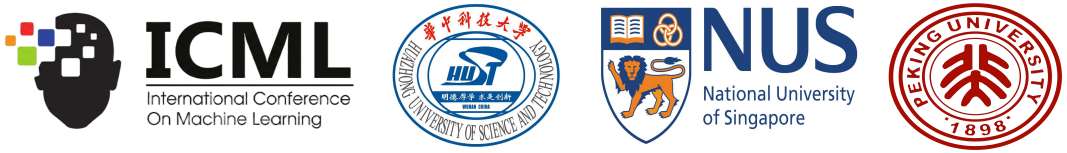
$\mathcal{L}_{\mathrm{dss}}$ effectively suppresses sensitivity.

$\mathcal{L}_{\mathrm{pert}}$ effectively increases the disturbance radius and margin

$\mathcal{L}_{\mathrm{MAD}}$ effectively reduces the confusion between new classes and previously learned classes in incremental learning

Figure 6. (a)–(b) analyze the effect of $\mathcal{L}_{seg-PML}$; (c)–(e) analyze $\mathcal{L}_{dss}$, $\mathcal{L}_{pert}$, and $\mathcal{L}_{SKD}$, respectively. $\mathcal{L}_{seg-PML}$ enlarges the top-1/top-2 confidence gap, while $\mathcal{L}_{dss}$ and $\mathcal{L}_{pert}$ suppress directional second-order sensitivity and increase the stability radius. $\mathcal{L}_{SKD}$ further mitigates class confusion.

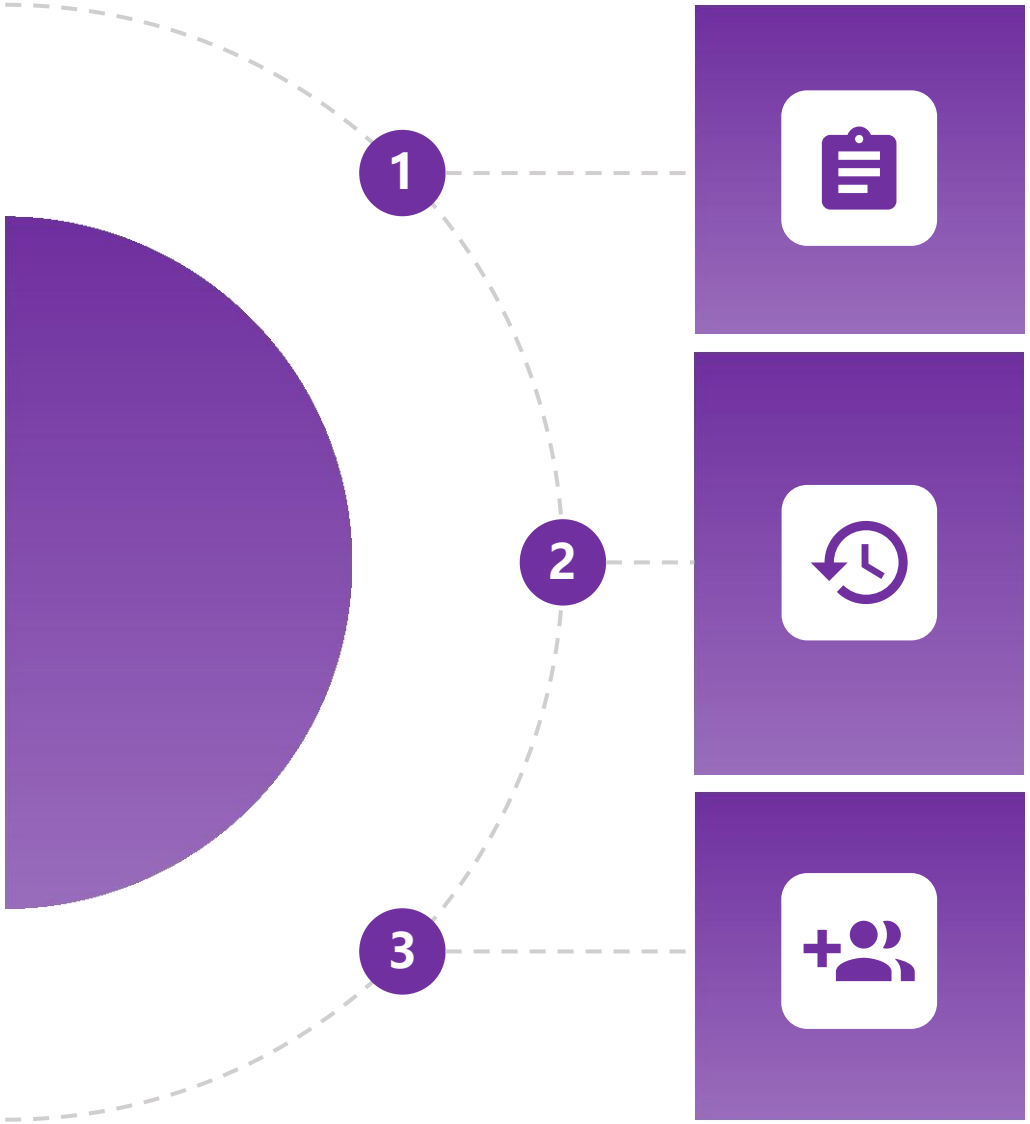
Ablation studies



05

Conclusion

5 Conclusion



Optimization diagnosis.

low-margin region

We provide a diagnosis of low-margin regions in CISS, demonstrating that these regions exhibit **high second-order margin sensitivityand a small stability radius**, which leads to ill-conditioned optimization.

LDKA

suppress sensitivity

Increase the stability radius

LDKA addresses low-margin optimization via three complementary strategies: PML determines **what to learn** by progressively allocating pixel-level optimization budget, SKD determines **how to learn** by stabilizing low-margin updates, and MAD determines **where to separate** by decoupling highly confused class pairs. Together, they provide an **allocation–stabilization–decoupling optimization** design.

Comprehensive evaluation.

Balance stability and plasticity

Experiments demonstrate that LDKA improves mIoU on newly emerging classes while preserving performance on old classes across 9 diverse incremental configurations

THANKS

Xinyue Zhang [*1], Xu Zou [*1], Wanjia Luo [2], Yanjie Wang [1], Jiahuan Zhou [3], Sheng Zhong [1], Luxin Yan [1]

ICML International Conference On Machine Learning

NUS National University of Singapore