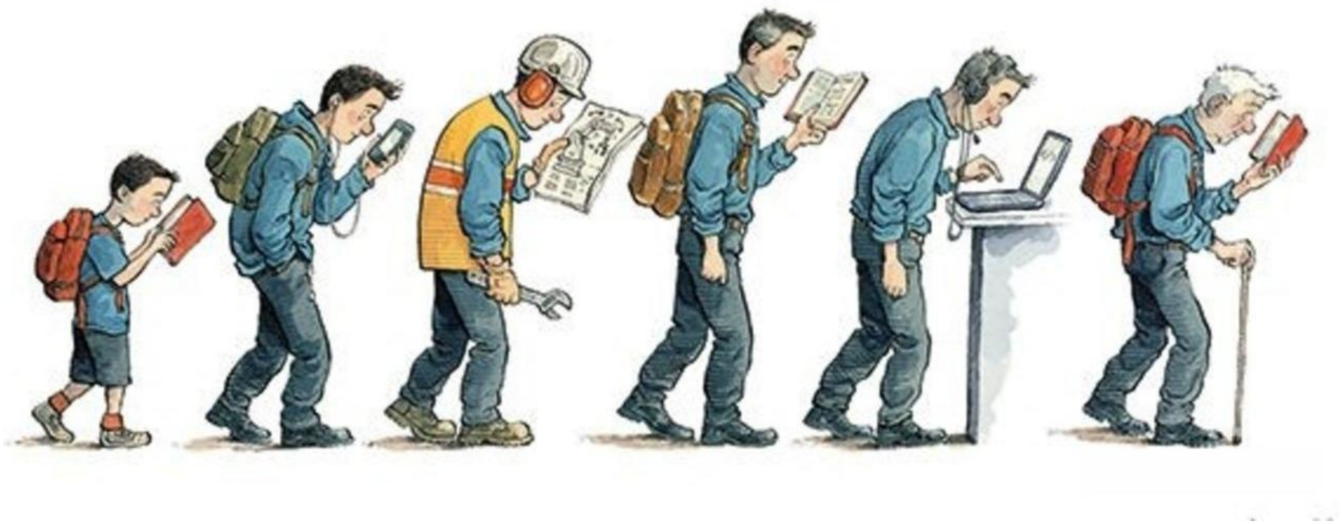
Semantic Shift Estimation via Dual-Projection and Classifier Reconstruction for Exemplar-Free Class-Incremental Learning

Run He[1], Di Fang[1], Yichen Xu[2], Yawen Cui[3],
Ming Li[4], Cen Chen[1], Ziqian Zeng[1], Huiping Zhuang[1*]

[1]South China University of Technology
[2]Institute of Science Tokyo
[3]The Hong Kong University of Science and Technology
[4]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)

❑ Incremental Learning

Incremental learning: Enables **continuous knowledge acquisition**, mimicking human behavior.

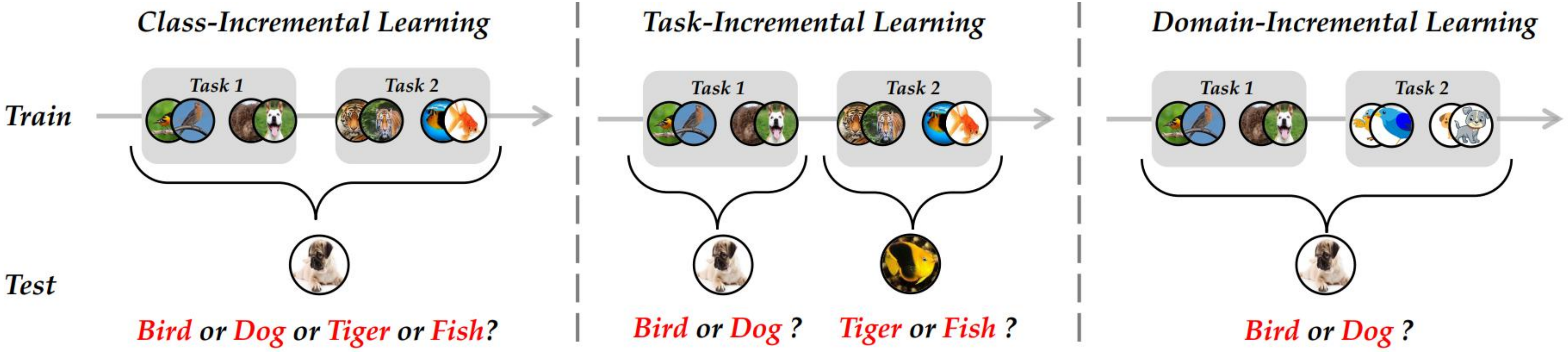


Practical Significance:

- No need for **retraining**;
- Adapt models to intricate usage.

❑ Three Settings of IL

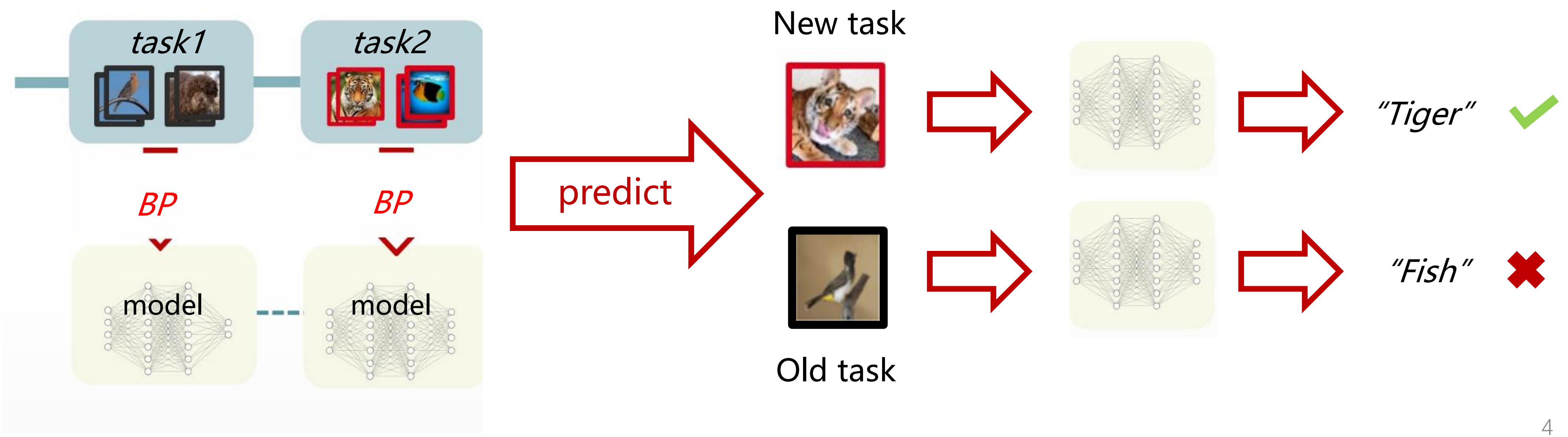
Class Incremental Learning (CIL) is one of the most difficult and most common setting in the field of IL.



Zhou, D. W., Wang, Q. W., Qi, Z. H., Ye, H. J., Zhan, D. C., & Liu, Z. (2023). Deep class-incremental learning: A survey. TPAMI, 2024.

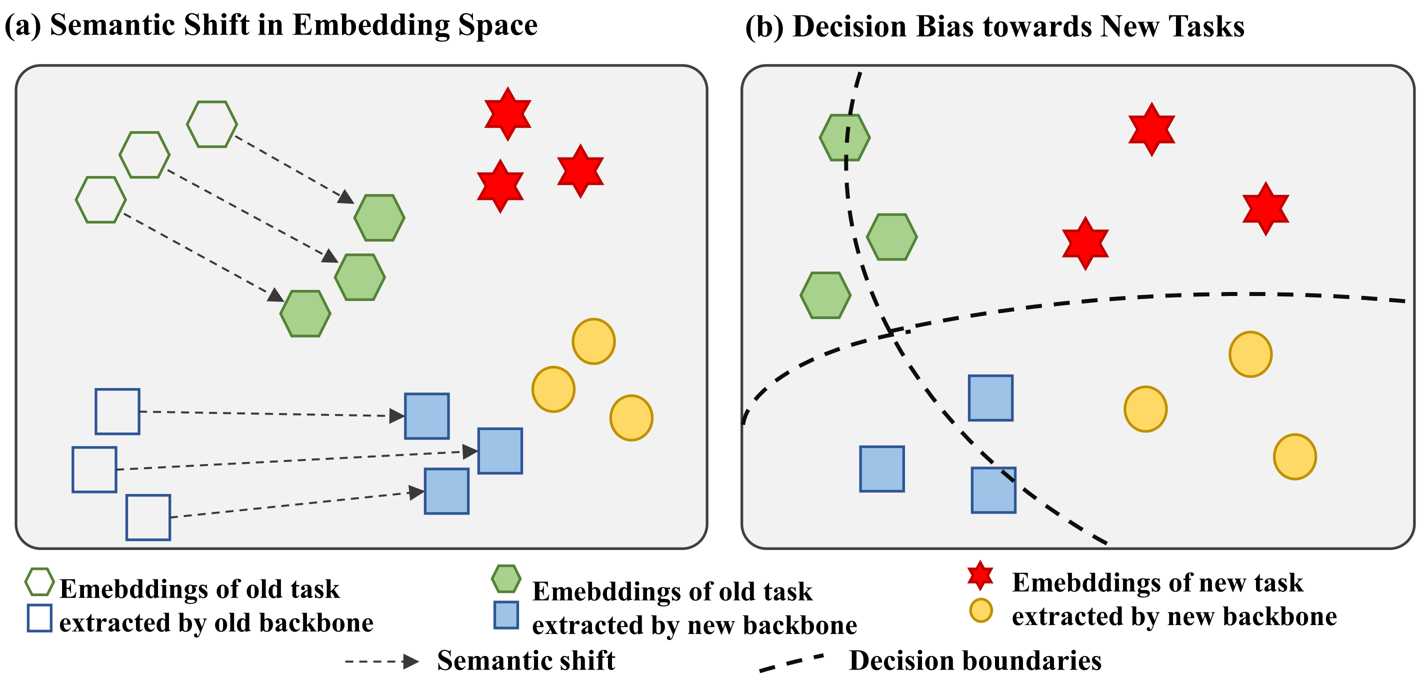
❑ Challenges: Catastrophic Forgetting

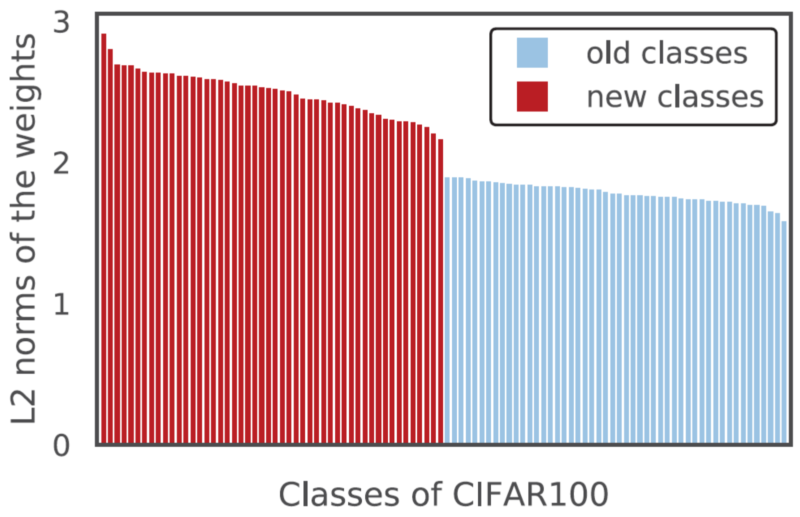
- ➢ Model Learns in **multiple stages** and different tasks;
- ➢ New model does well in **new tasks**;
- ➢ Performance **decrease for the previous** tasks.



❑ Challenges behind CF

- **Semantic Shift**: features of previous tasks shift after new task;
- **Decision bias**: biased decision boundaries since training on new data solely;

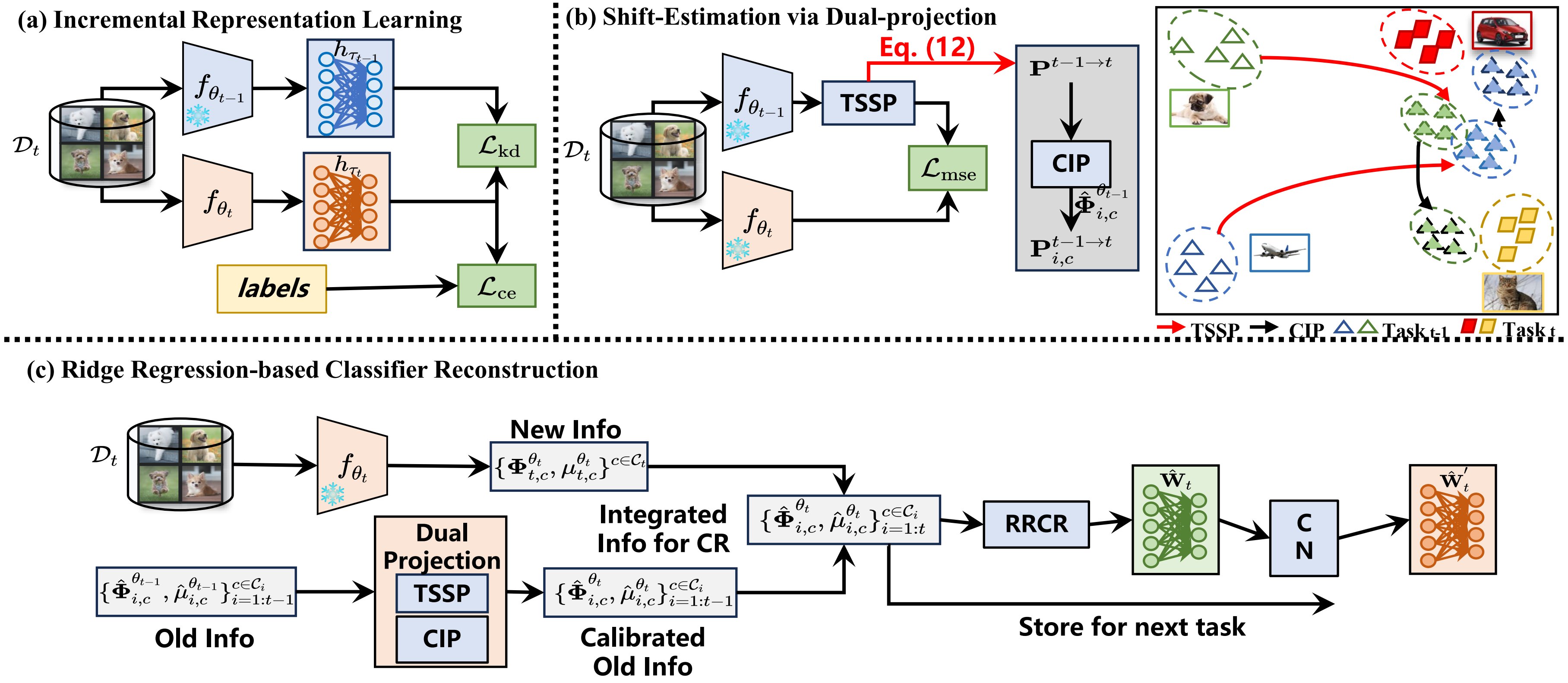




Hou, S., et.al. (2019). Learning a unified classifier incrementally via rebalancing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 831-839).

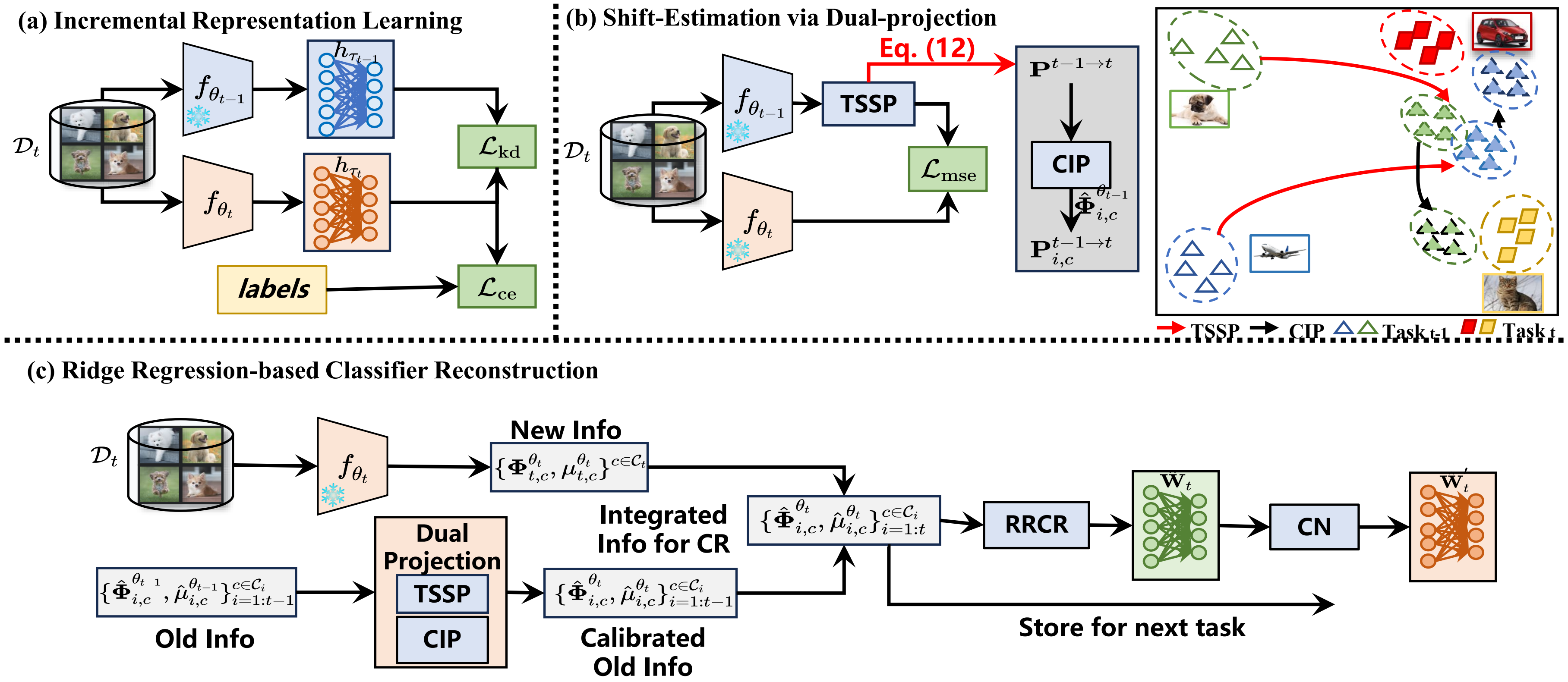
❑ Our Solution: DPCR

- **Dual Projection**: estimate the semantic shift across tasks;
- **Classifier Reconstruction**: reconstruct the classifier via ridge-regression;



Proposed Method

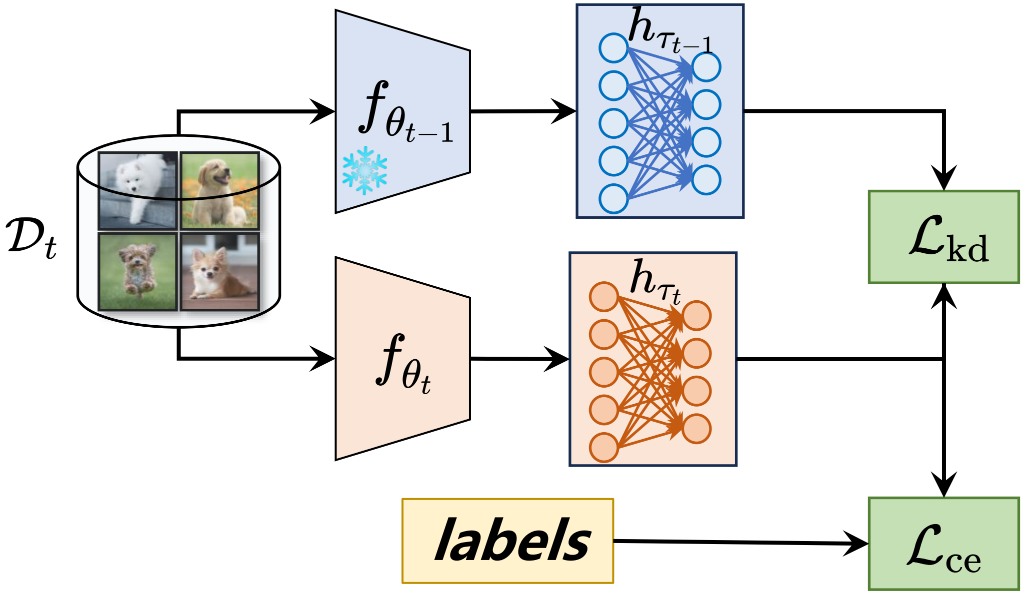
❑ DPCR: Dual Projection and Classifier Reconstruction



- **Dual Projection to estimate both task and class-specific shift**
- **Classifier Reconstruction addresses the decision bias**

Proposed Method

❑ Incremental Representation Learning

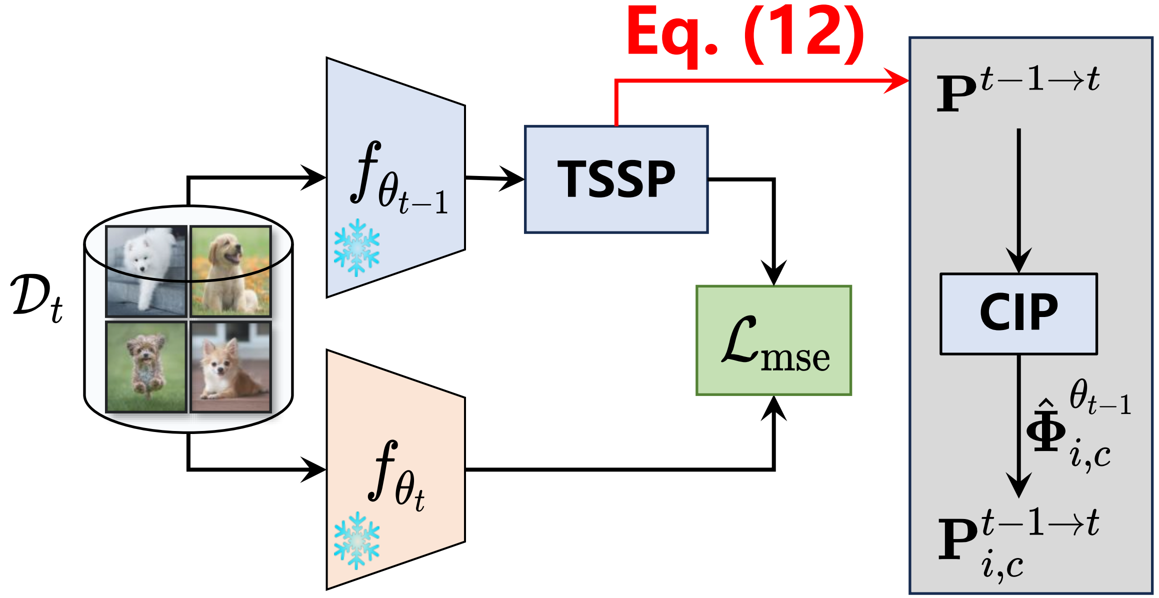


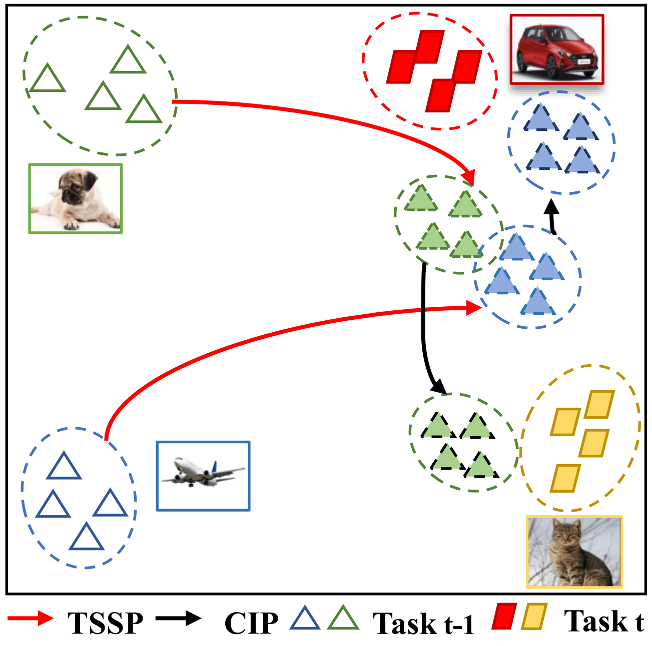
- **Utilize Knowledge Distillation to avoid CF**

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{ce}}(h^{\text{au}}_{\tau_t}(f_{\theta_t}(\mathcal{X}_t), y_t) + \alpha\mathcal{L}_{\text{kd}}(\mathcal{X}_t).$$

$$\mathcal{L}_{\text{kd}} = \mathcal{L}_{\text{ce}}(h^{\text{au}}_{\tau_{t-1}}(f_{\theta_{t-1}}(\mathcal{X}_t)), h^{\text{au}}_{\tau_t}(f_{\theta_t}(\mathcal{X}_t)))$$

❑ Shift Estimation via Dual-Projection (DP)





- **Obtain task shift via TSSP**

$$\underset{\boldsymbol{P}^{t-1\to t}}{\operatorname{argmin}}\ \mathcal{L}_{\text{mse}} = \|\boldsymbol{X}_t^{\theta_t} - \boldsymbol{X}_t^{\theta_{t-1}}\boldsymbol{P}^{t-1\to t}\|_{\text{F}}^2$$

$$\boldsymbol{P}^{t-1\to t} = (\boldsymbol{X}_t^{\theta_{t-1}\top}\boldsymbol{X}_t^{\theta_{t-1}} + \epsilon\boldsymbol{I})^{-1}\boldsymbol{X}_t^{\theta_{t-1}\top}\boldsymbol{X}_t^{\theta_t}$$

- **Inject class-specific information via CIP**

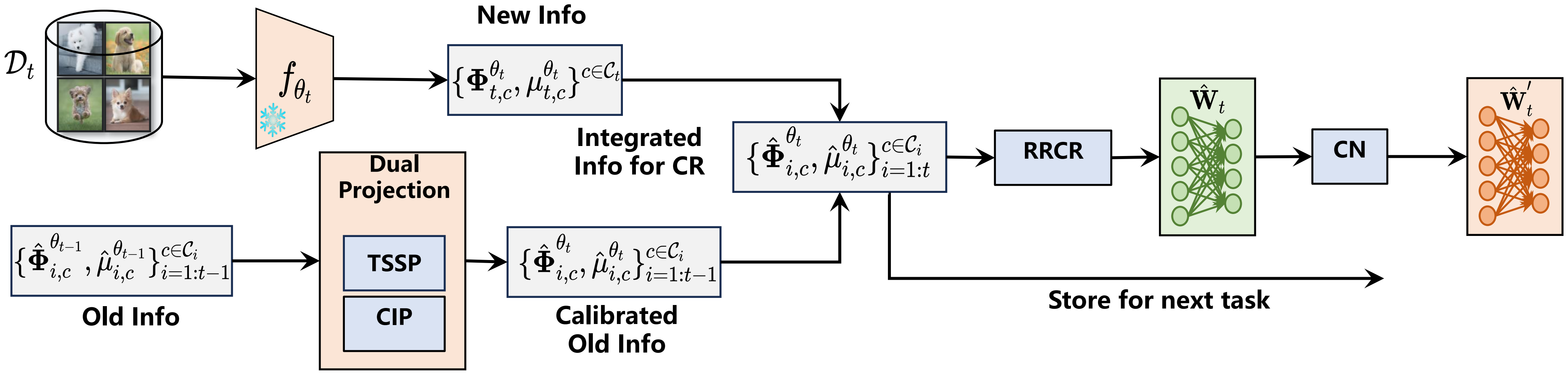
covariance $\Phi_{t-1,c}^{\theta_{t-1}} = \boldsymbol{X}_{t-1,c}^{\theta_{t-1}\top}\boldsymbol{X}_{t-1,c}^{\theta_{t-1}}$

Extract class information via SVD

$$\boldsymbol{U}_{t-1,c}, \boldsymbol{\Sigma}_{t-1,c}, \boldsymbol{U}_{t-1,c}^{\top} = \text{SVD}(\Phi_{t-1,c}^{\theta_{t-1}}) \quad \boldsymbol{U}_{t-1,c} = [\boldsymbol{U}_{t-1,c}^{r} \quad \boldsymbol{U}_{t-1,c}^{z}]$$

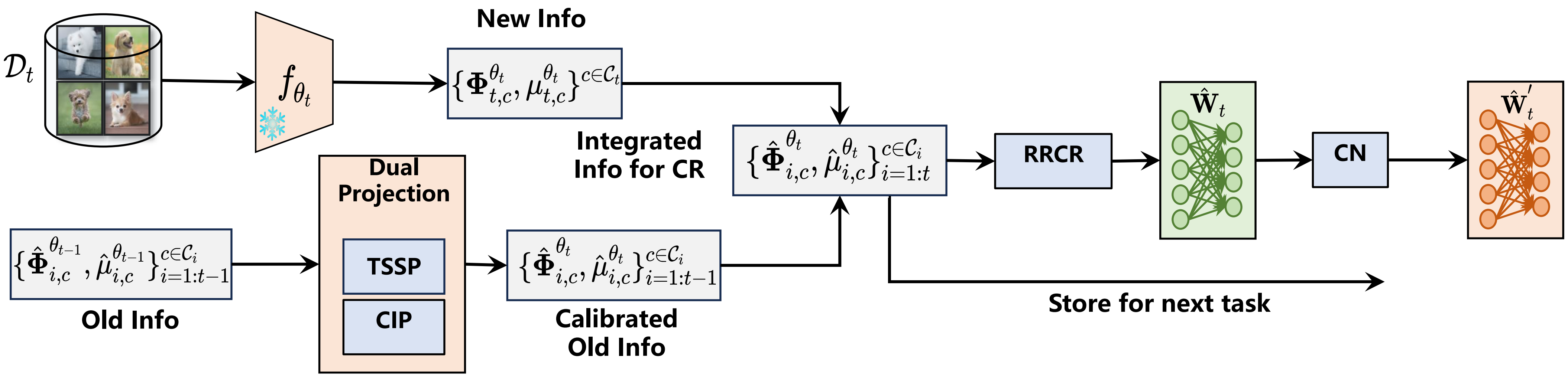
$$\boldsymbol{P}_{t-1,c}^{t-1\to t} = \boldsymbol{P}^{t-1\to t}\boldsymbol{U}_{t-1,c}^{r}\boldsymbol{U}_{t-1,c}^{r\top}$$

❑ Ridge Regression-based Classifier Reconstruction (RRCR)



- ➢ **Formulate the classifier training as a reconstruction process**
- ➢ **Calibrate the old information with DP analytically**

❑ Ridge Regression-based Classifier Reconstruction (RRCR)



➢ **Classifier training via ridge-regression**

$$\underset{\boldsymbol{W}_t}{\operatorname{argmin}} \quad \|\boldsymbol{Y}_{1:t} - \boldsymbol{X}^{\theta_1}_{1:t}\boldsymbol{W}_t\|_{\mathrm{F}}^2 + \gamma\,\|\boldsymbol{W}_t\|_{\mathrm{F}}^2$$

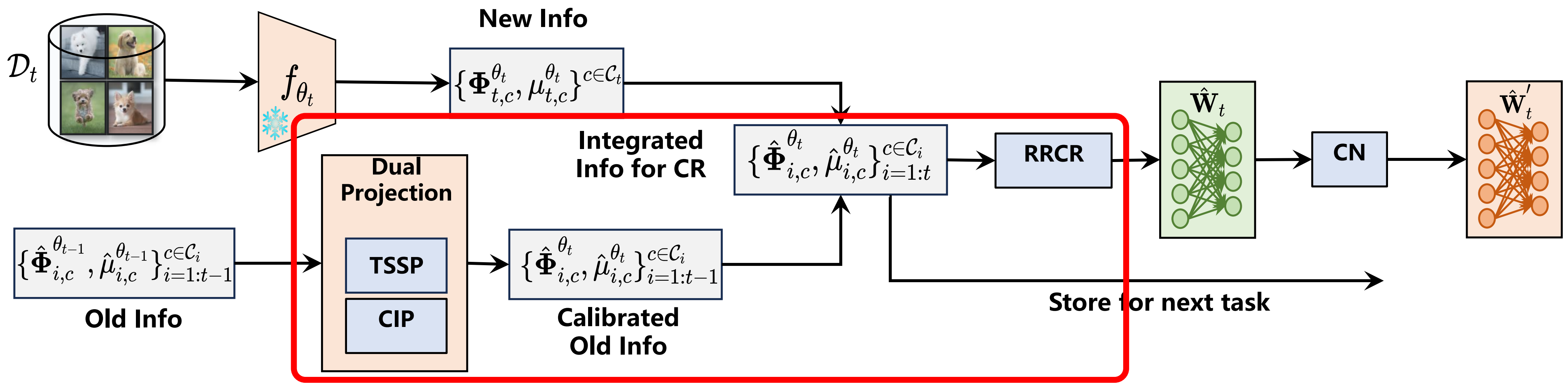
Solution

$$\hat{\mathbf{W}}_t = \Big(\sum_{i=1}^{t}\sum_{c\in\mathcal{C}_i} \boldsymbol{X}^{\theta_t\top}_{i,c}\boldsymbol{X}^{\theta_t}_{i,c} + \gamma\boldsymbol{I}\Big)^{-1}\sum_{i=1}^{t}\sum_{c\in\mathcal{C}_i}\boldsymbol{X}^{\theta_t\top}_{i,c}\boldsymbol{Y}_{i,c}$$
$$= \Big(\sum_{i=1}^{t}\sum_{c\in\mathcal{C}_i}\boldsymbol{\Phi}^{\theta_t}_{i,c} + \gamma\boldsymbol{I}\Big)^{-1}\sum_{i=1}^{t}\sum_{c\in\mathcal{C}_i}\boldsymbol{H}^{\theta_i}_{i,c}$$

$$\boldsymbol{H}^{\theta_t}_i = \sum_{c\in\mathcal{C}_i}\boldsymbol{X}^{\theta_t\top}_{i,c}\boldsymbol{Y}_{i,c} = \sum_{c\in\mathcal{C}_i} N_c\boldsymbol{\mu}^{\theta_t\top}_{i,c}\boldsymbol{y}_{i,c},$$

$$\boldsymbol{\Phi}^{\theta_t}_{i,c} = \boldsymbol{X}^{\theta_t\top}_{i,c}\boldsymbol{X}^{\theta_t}_{i,c}, \quad \boldsymbol{\mu}^{\theta_t}_{i,c} = \frac{1}{N_c}\sum_{j=1}^{N_c}\boldsymbol{x}^{\theta_t}_{i,c,j}.$$

❑ Ridge Regression-based Classifier Reconstruction (RRCR)



- ➢ **Calibrate semantic shift with DP**

$$\hat{\Phi}_{i,c}^{\theta_t} = \hat{X}_{i,c}^{\theta_t\top}\hat{X}_{i,c}^{\theta_t} = P_{i,c}^{t-1\to t\top}\Phi_{i,c}^{\theta_{t-1}}P_{i,c}^{t-1\to t}$$

$$\hat{\mu}_{i,c}^{\theta_t} = \frac{1}{N_c}\sum_{j=1}^{N_c}\hat{x}_{i,c,j}^{\theta_{t-1}} = \mu_{i,c}^{\theta_{t-1}}P_{i,c}^{t-1\to t},$$

$$\hat{H}_{i,c}^{\theta_t} = \sum_{c\in\mathcal{C}_i}^{C} N_c\hat{\mu}_{i,c}^{\theta_i\top}y_{i,c}.$$

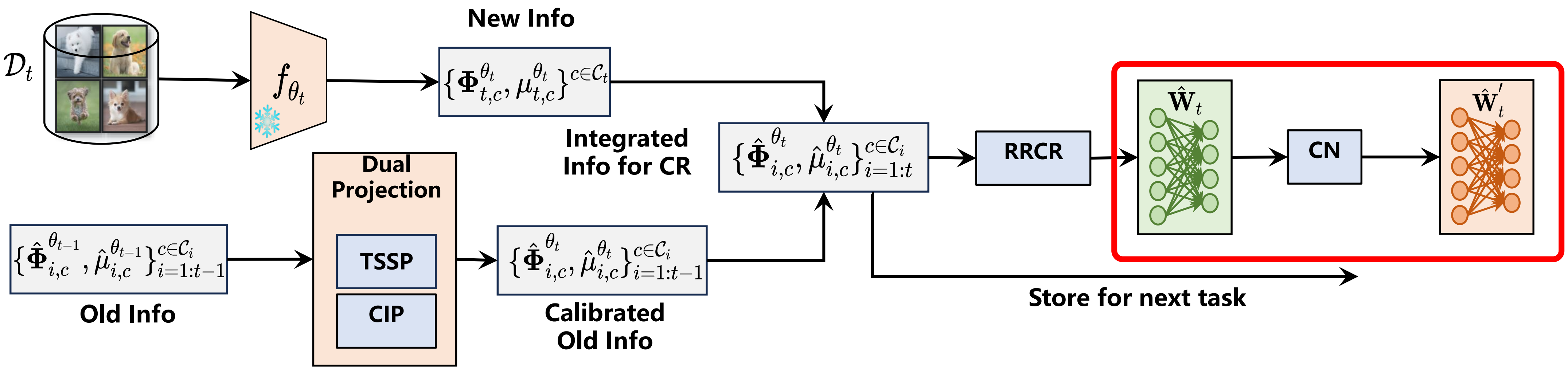
- ➢ **Integrate with new tasks**

$$\hat{W}_t = \left(\sum_{i=1}^{t-1}\hat{\Phi}_i^{\theta_t} + \Phi_t^{\theta_t}\right)^{-1}\left(\sum_{i=1}^{t-1}\hat{H}_i^{\theta_t} + H_t^{\theta_t}\right)$$

$$\hat{\Phi}_i^{\theta_t} = \sum_{c\in\mathcal{C}_i}\hat{\Phi}_{i,c}^{\theta_t}, \quad \hat{H}_i^{\theta_t} = \sum_{c\in\mathcal{C}_i}\hat{H}_{i,c}^{\theta_t}.$$

$$\Phi_t^{\theta_t} = \sum_{c\in\mathcal{C}_t}\Phi_{t,c}^{\theta_t}, \quad H_t^{\theta_t} = \sum_{c\in\mathcal{C}_t}H_{t,c}^{\theta_t}.$$

❑ Ridge Regression-based Classifier Reconstruction (RRCR)



➢ **Classifier Normalization**

$$\hat{W}'_t = [\frac{w_1}{\|w_j\|_1}, \frac{w_2}{\|w_2\|_2}, \dots, \frac{w_{tC}}{\|w_{tC}\|_2}]$$

❑ Compare with State-of-the-arts

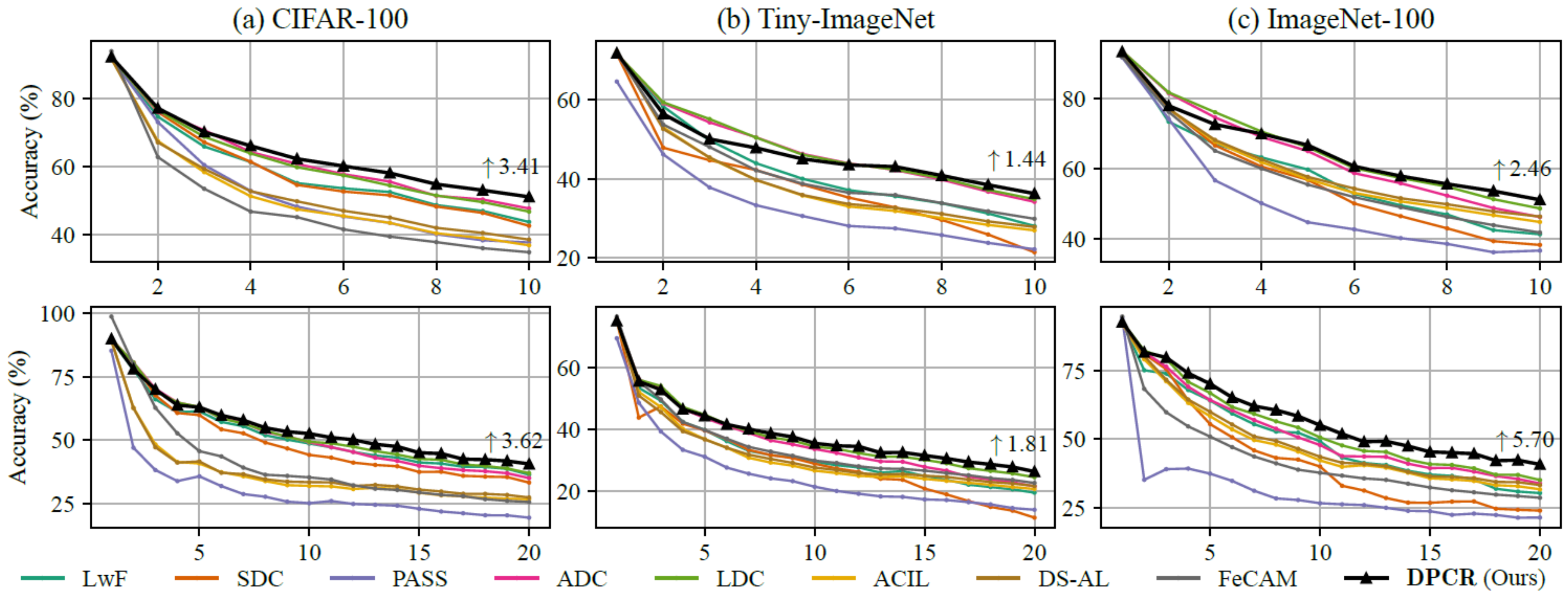
Methods	CIFAR-100				Tiny-ImageNet				ImageNet-100			
	T=10		T=20		T=10		T=20		T=10		T=20	
	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$
LwF (2017)	42.60	58.51	36.34	51.52	26.99	42.92	18.80	33.05	42.25	61.23	30.11	50.40
SDC (2020)	42.25	58.43	33.10	48.68	23.86	40.66	13.45	29.70	37.68	60.33	23.64	45.52
PASS (2021b)	44.47	55.88	28.48	42.65	23.89	36.82	12.50	25.38	36.52	52.02	19.59	31.55
ACIL (2022b)	35.53	50.53	27.22	39.58	26.10	41.86	21.40	33.60	44.61	59.77	33.05	48.58
FeCAM (2023)	34.82	49.14	25.77	41.21	29.83	42.19	22.69	34.48	41.92	58.21	28.64	43.04
DS-AL (2024b)	36.83	51.47	28.90	40.37	27.01	40.10	21.86	33.55	45.55	60.56	34.10	49.38
ADC (2024)	46.80	62.05	34.69	52.16	32.90	46.93	20.69	36.14	46.69	65.60	32.21	52.36
LDC (2024)	46.60	61.67	36.76	53.06	33.74	47.37	24.49	38.04	49.98	67.47	34.87	54.84
DPCR (Ours)	**50.24**$^{\uparrow 3.64}$	**63.21**$^{\uparrow 1.54}$	**38.98**$^{\uparrow 2.22}$	**54.42**$^{\uparrow 1.36}$	**35.20**$^{\uparrow 1.46}$	**47.55**$^{\uparrow 0.18}$	**26.54**$^{\uparrow 2.05}$	**38.09**$^{\uparrow 0.05}$	**52.16**$^{\uparrow 2.18}$	**67.51**$^{\uparrow 0.04}$	**38.35**$^{\uparrow 3.48}$	**57.22**$^{\uparrow 2.36}$

CUB200 (T=5)	$\mathcal{A}_f$ (%)	$\mathcal{A}_{avg}$ (%)
LwF (Li & Hoiem, 2017)	25.40	36.38
ACIL (Zhuang et al., 2022b)	21.14	33.14
DS-AL (Zhuang et al., 2024b)	21.28	32.36
SDC (Yu et al., 2020)	24.24	36.00
ADC (Goswami et al., 2024)	28.84	39.44
LDC (Gomez-Villa et al., 2024)	28.70	39.09
DPCR (Ours)	**29.51**	**39.44**

ImageNet-1k (T=10)	$\mathcal{A}_f$ (%)	$\mathcal{A}_{avg}$ (%)
LwF (Li & Hoiem, 2017)	22.01	42.40
ACIL (Zhuang et al., 2022b)	32.28	46.61
DS-AL (Zhuang et al., 2024b)	33.67	48.84
ADC (Goswami et al., 2024)	31.34	50.95
LDC (Gomez-Villa et al., 2024)	35.15	53.88
DPCR (Ours)	**35.49**	**54.22**

➢ **Outperform existing** EFCIL methods with considerable gap.

❑ Evolution Curve



➢ **Outperform existing** methods across the training tasks.

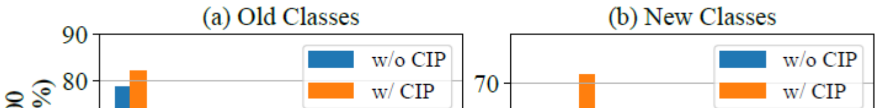
❑ Ablation Study

➢ The performance can **be further improved with TSSP, CIP and CN** on top of RRCR

Components	$\mathcal{A}_f$ (%)	$\mathcal{A}_{avg}$ (%)
RRCR	32.17	44.89
RRCR+TSSP	40.86	55.76
RRCR+TSSP+CIP	45.56	62.15
RRCR+TSSP+CIP+CN	**51.04**	**64.44**

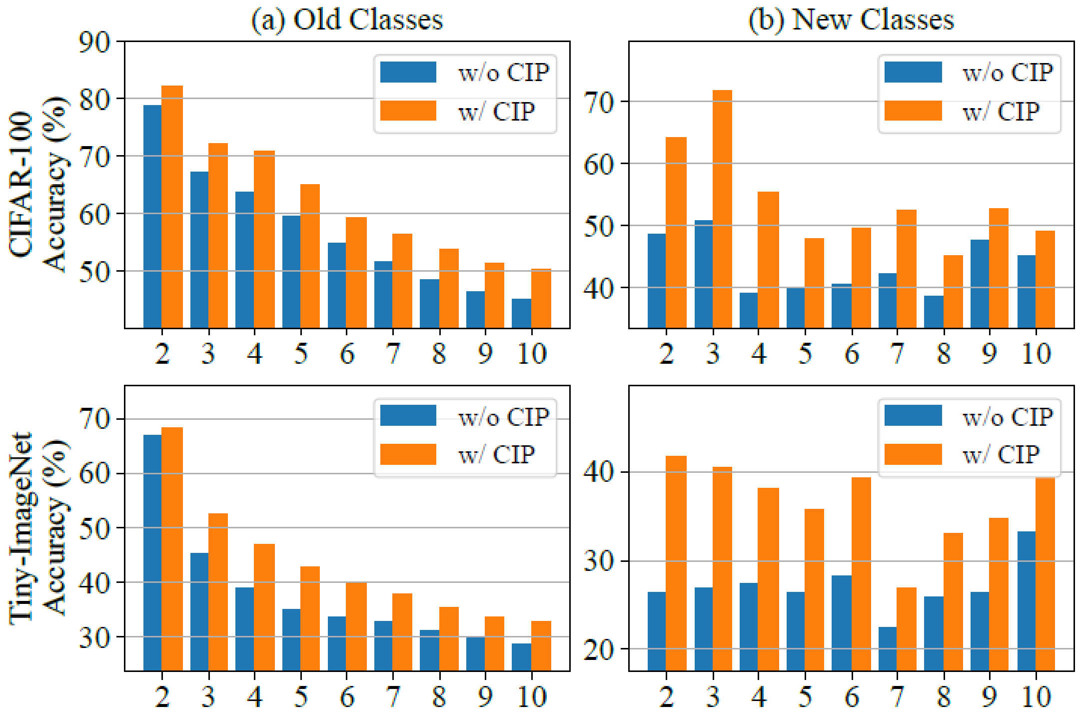
➢ DP **outperform existing methods that estimate semantic shift**

Methods	CIFAR-100				Tiny-ImageNet			
	T=10		T=20		T=10		T=20	
	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$	$\mathcal{A}_f$	$\mathcal{A}_{avg}$
ADC	47.65	62.63	35.17	52.16	30.71	41.81	18.63	31.55
LDC	47.40	62.39	37.10	53.28	32.90	43.67	23.57	34.08
DP-NCM	**49.19**	**63.47**	**37.64**	**53.86**	**33.47**	**43.86**	**24.90**	**35.22**

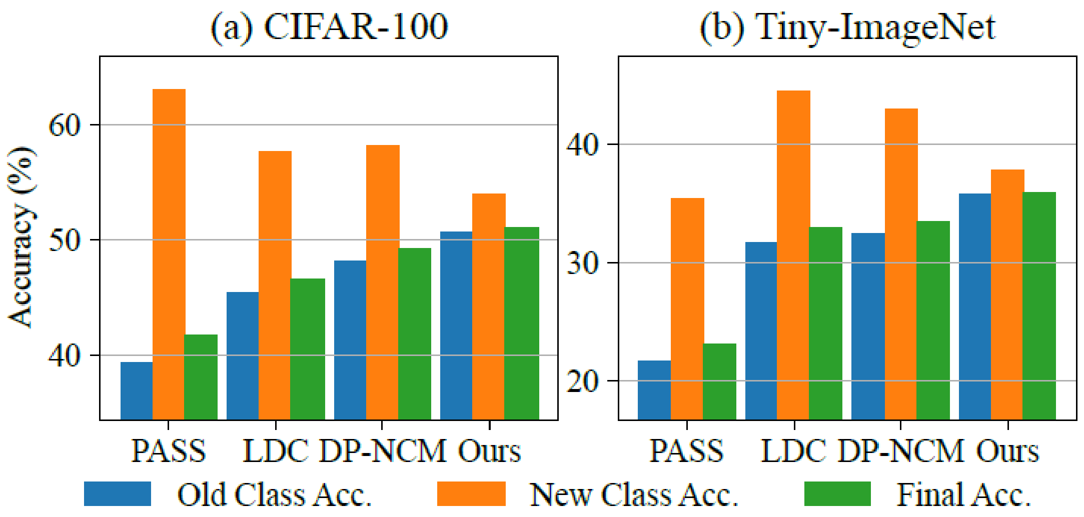


❑ Ablation Study

➢ CIP Enhances Both **the Stability and Plasticity**

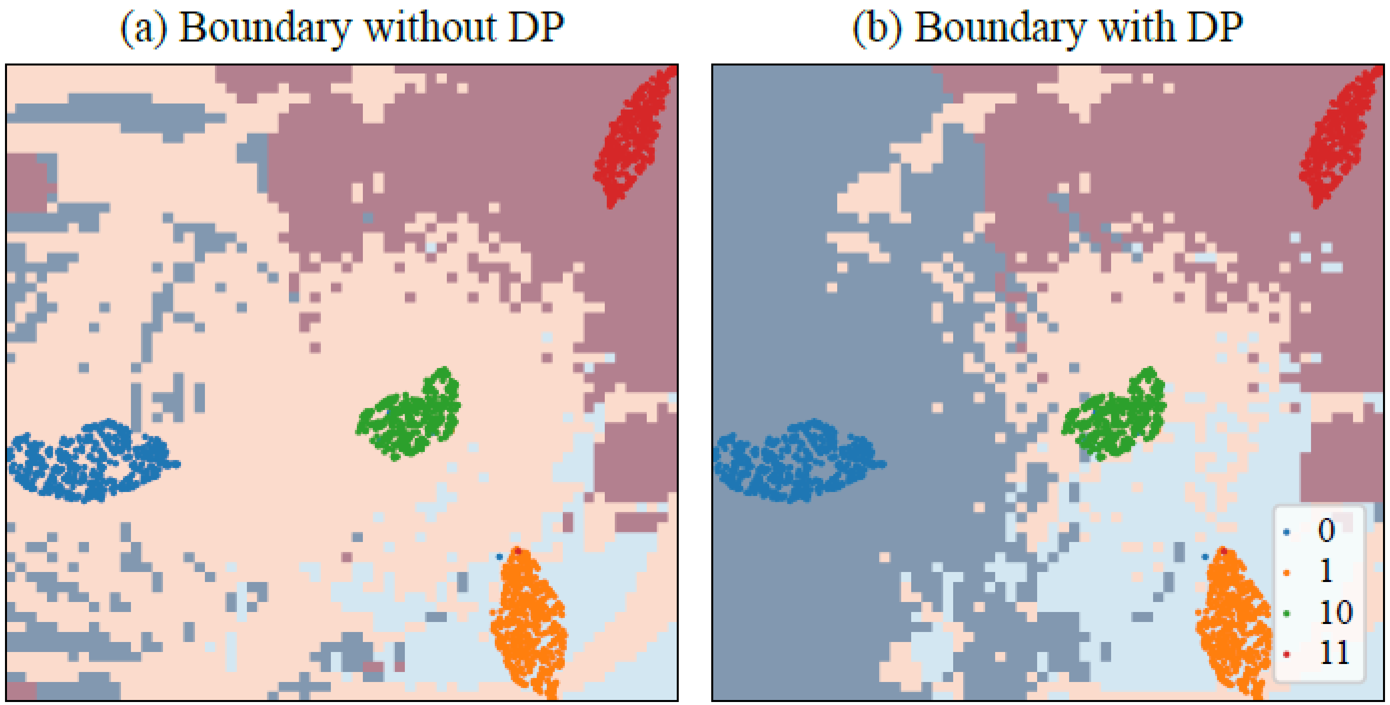


➢ RRCR reduces **decision bias**



Ablation Study

- Visualization of **effect of DP on the decision boundaries**



Our Codes

- Our codes are available at: https://github.com/RHe502/ICML25-DPCR.

- The corresponding QR code: