山东大学
SHANDONG UNIVERSITY

中国计算机学会《学科前沿讲习班》
The CCF Advanced Disciplines Lectures

具身感知:从瞬态理解走向持续进化

EMBODIED PERCEPTION : FROM TRANSIENT UNDERSTANDING TO CONTINUAL EVOLUTION

报告人：丛润民

山东大学控制科学与工程学院

机器智能与系统控制教育部重点实验室

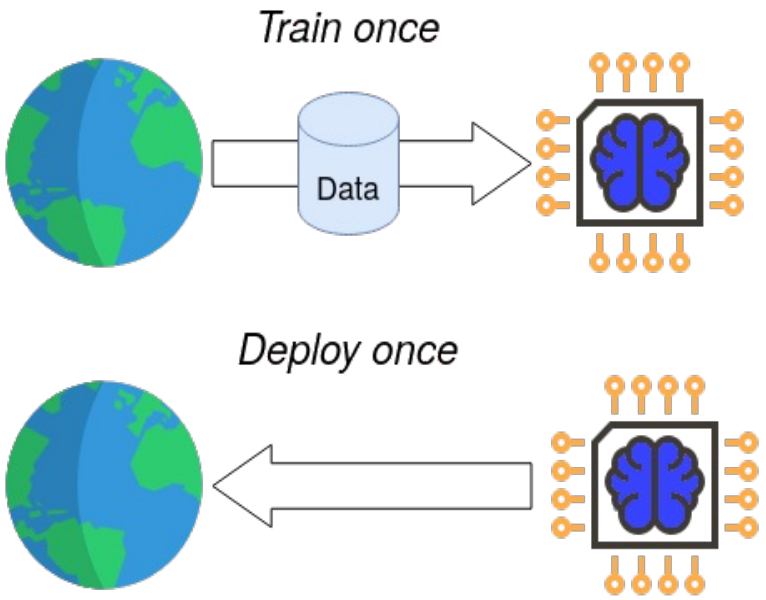
学无止境 气有浩然

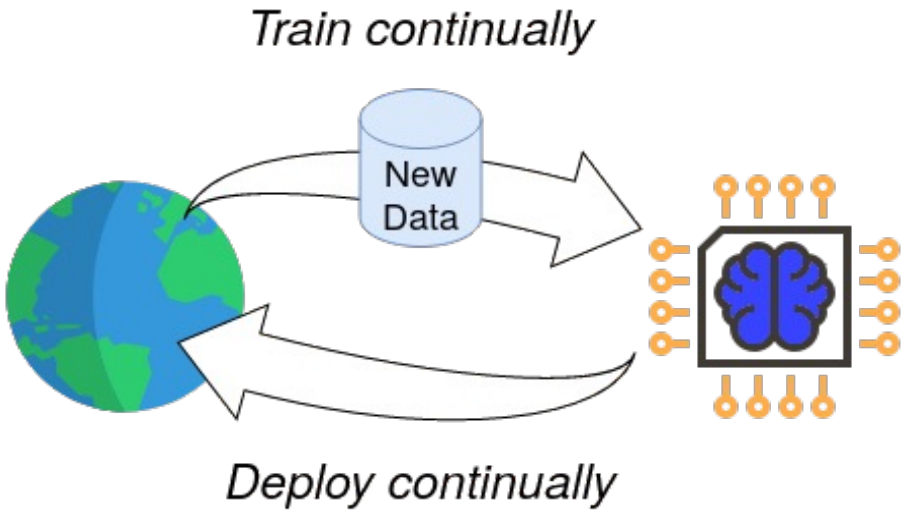
山东大学 SHANDONG UNIVERSITY

目录

CONTENTS

Introduction —— Transient → Continual Learning





Transient Perception → Continual Perception

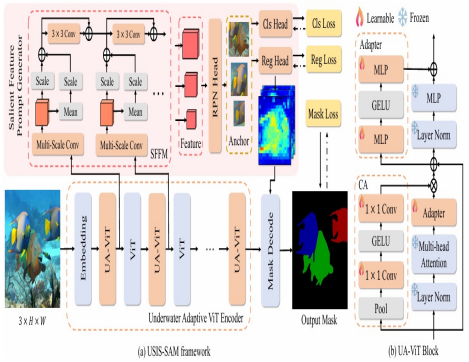
As shown in the above image, a conventional model can only be trained once and has fixed capabilities. In contrast, a model with continual learning abilities can continuously expand its capabilities to meet new requirements.

Introduction —— Transient → Continual Learning

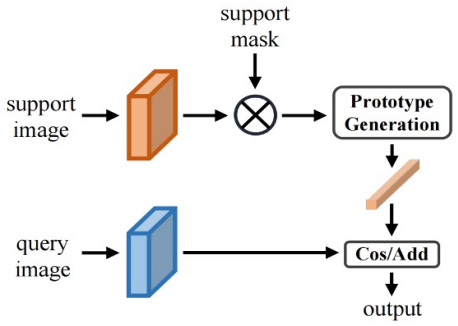
瞬态学习

持续学习

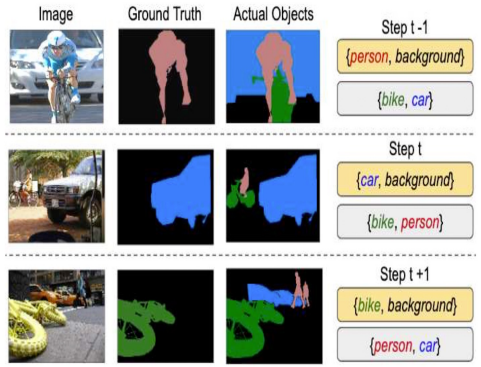
全监督 小样本 小模型 大模型



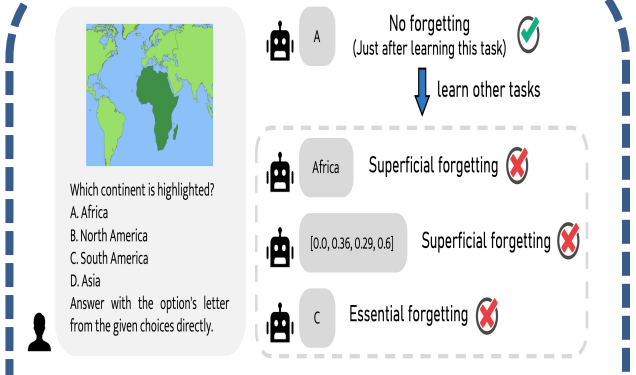
当前类别学习
全监督学习



新类别学习
小样本学习



小模型语义理解
持续学习



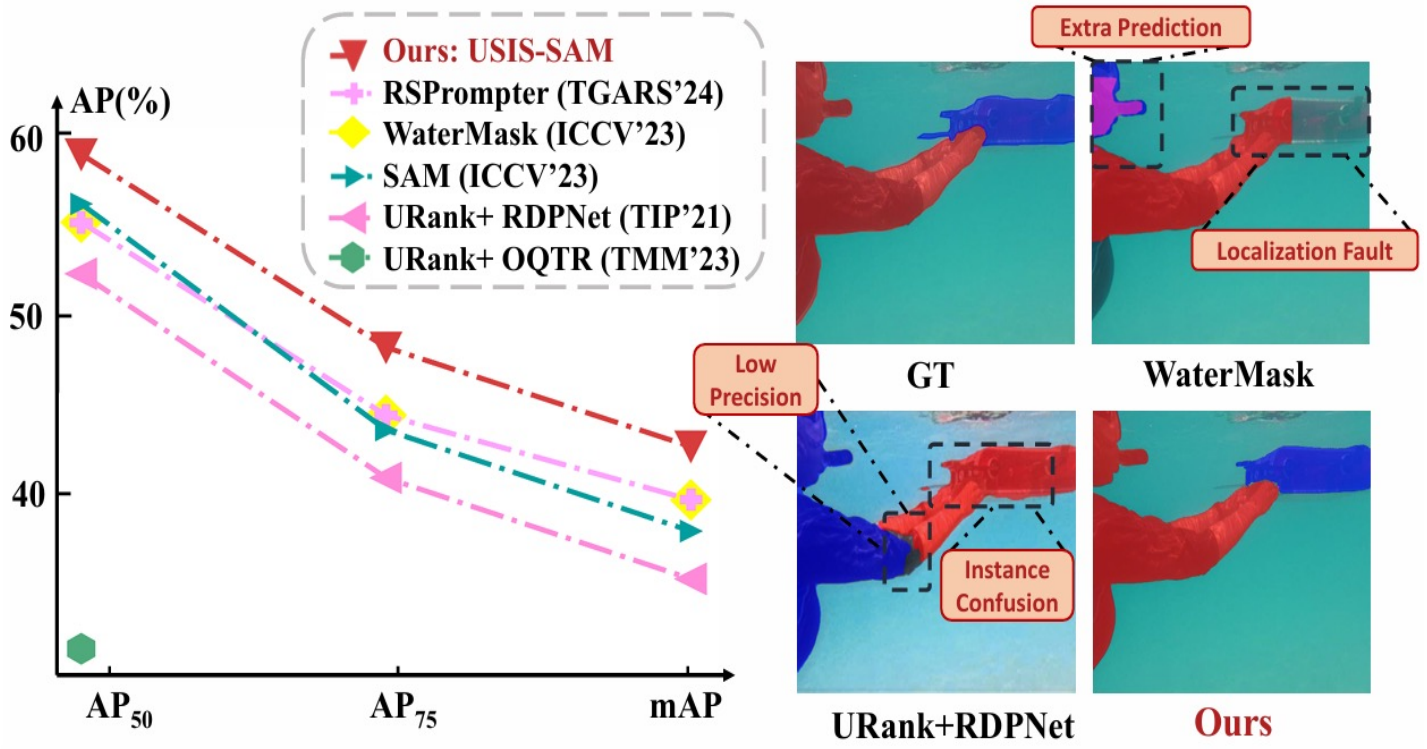
多模态大模型推理
持续学习

ICML 2024

Diving into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation and A Large-scale Dataset

Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Tianruo Yang, Sam Kwong, Runmin Cong

Introduction



- **Salient Instance Segmentation (SIS)**, an emerging and promising visual task, aims to segment out visually salient objects in a scene and distinguish individual salient instances, which is beneficial for vision tasks such as marine resource exploration and underwater human-computer interaction.

- However, directly transferring conventional SIS methods for land images to underwater scenes may struggle to achieve ideal performance attribute to the **domain gap** of **intrinsic characteristics and extrinsic circumstances between land and underwater living**.

Motivation

- On the one hand, there is **no general underwater salient image instance segmentation dataset** to promote training and evaluation of the underwater salient instance segmentation models.
- On the other hand, even state-of-the-art SIS models trained on large-scale land-based datasets coupled with the best underwater image enhancement algorithms **cannot achieve satisfactory performance in underwater environments**.
- To alleviate this issue, we construct the **first large-scale underwater salient instance segmentation (USIS) dataset**, **USIS10K**, aiming to promote the development of salient instance segmentation for underwater tasks.
- Simultaneously, we first attempt to apply Segment Anything Model (SAM) to underwater salient instance segmentation and propose **USIS-SAM**, aiming to **improve the segmentation accuracy in complex underwater scenes**.

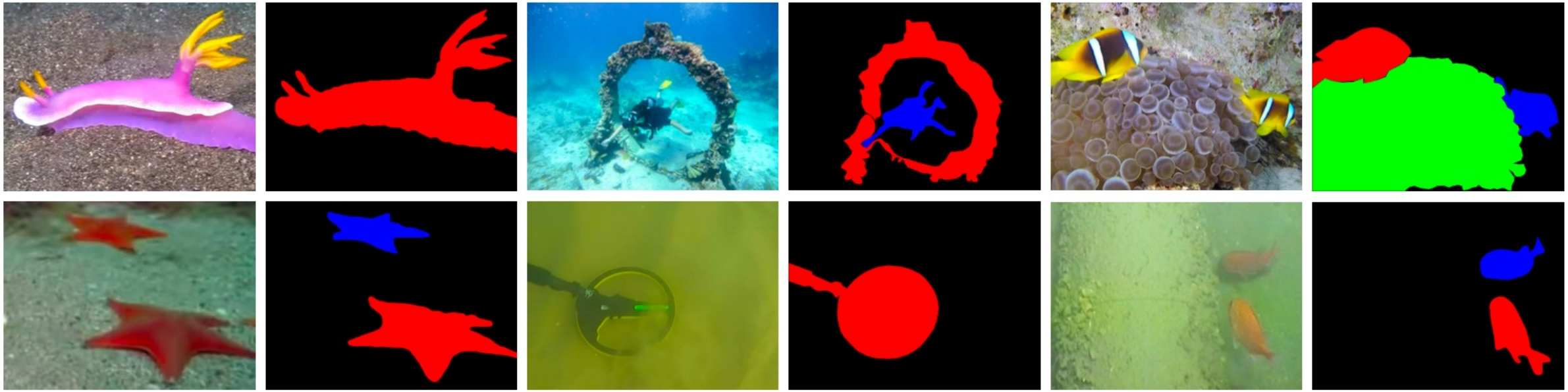
Contributions

a) We construct the **first large-scale dataset, USIS10K**, for the underwater salient instance segmentation task, which contains **10,632 images and pixel-level annotations of 7 categories**. As far as we know, this is **the largest salient instance segmentation dataset**, and includes Class-Agnostic and Multi-Class labels simultaneously.

b) We propose the **first underwater salient instance segmentation model, USIS-SAM**, as far as we know. In USIS-SAM, we design **Underwater Adaptive ViT Encoder** to incorporate underwater visual prompts into network via adapters, and **Salient Feature Prompter Generator** to automatically generate salient prompters, guiding an end-to-end segmentation network.

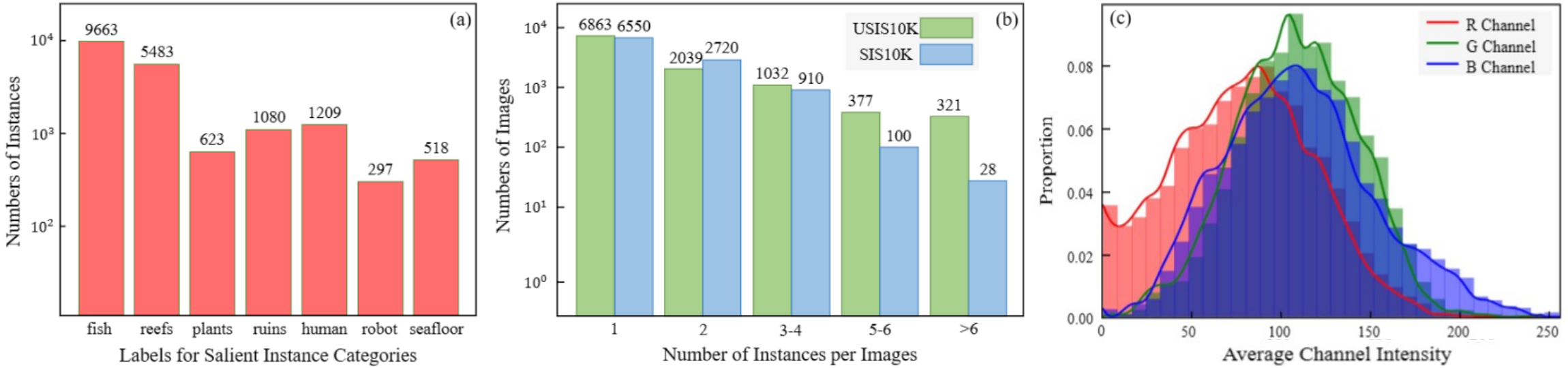
c) Extensive public evaluation criteria and large numbers of experiments verify the effectiveness of our USIS10K dataset and USIS-SAM.

USIS10K Dataset



Dataset	Year	Task	Label	Number	Max
ILSO	2017	SIS	×	2,000	8
SOC	2018	SIS	✓	3,000	8
SIS10K	2023	SIS	×	10,301	9
USIS10K	2024	USIS	✓	10,632	9

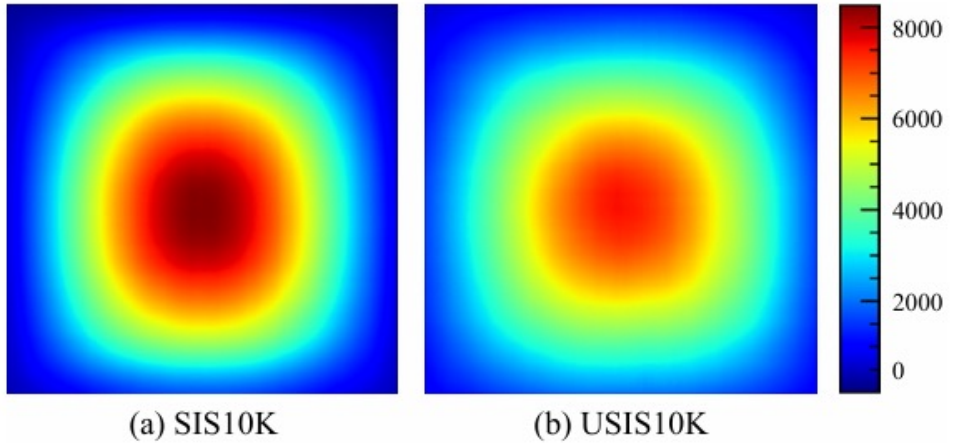
Dataset Statistic and Challenges



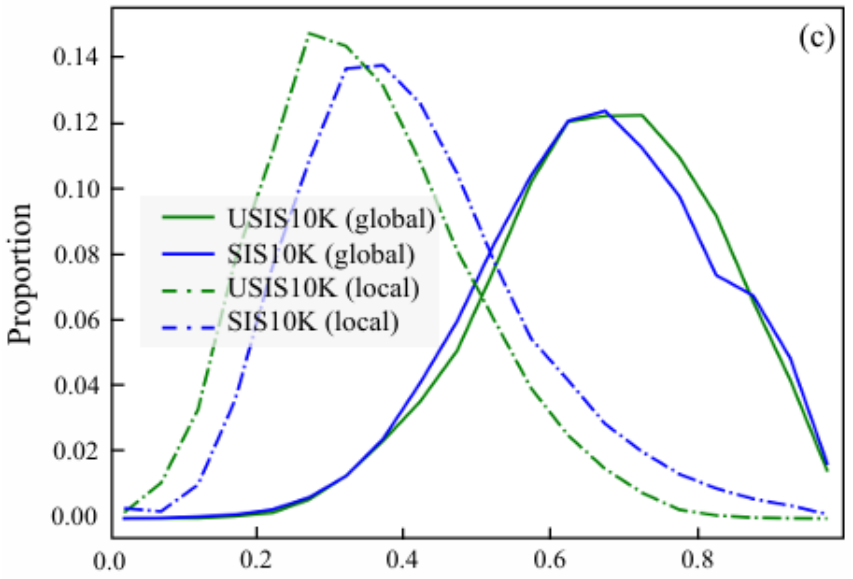
- **Challenge in the number of instance.**
  In USIS10K dataset, multiple salient instances may exist in a single image. There are 1731 images with more than 3 salient instances in our dataset, accounting for 16.3% of the total.
- **Challenges in small or large instances.**
  In USIS10K dataset, the average size of the salient instances is 34,336 pixels (approximately 185×185 pixels), which averaged 10.3% of the image size. There are 3053 salient instances smaller than 1% of the image area, (16.0% of the total), while there are 1733 instances larger than 30% of the image area, (9.1% of the total).
- **Challenges in channel intensity of underwater images.**
  Optical images inevitably suffer from color attenuation due to the selective absorption of water at different wavelengths. This poses an additional challenge for the network to properly understand and handle the image color distortion caused by this attenuation

Dataset Statistic and Challenges

- **Location of Salient Objects (Less central bias).**
  In the SIS10k dataset, approximately 13.5% of the locations have fewer than 1000 instances and 32% have fewer than 2,000 instances, while in our dataset, only 2.75% of the locations have fewer than 1,000 instances and 22.5% have fewer than 2,000 instances.



A set of salient maps from our dataset and SIS10K



Global/Local Color Contrast of Salient Instance

- **Color Contrast of Salient Instance.**
  Saliency is often related to the contrast between foreground and background, and it is critical to check whether salient instances are easy to detect. It can be seen that the global contrast of USIS10K is slightly higher than that of SIS10K. In addition, the local contrast of USIS10K at salient instances is lower than that of SIS10K. This poses a greater challenge in accurately segmenting the salient instance masks at the network boundary portion.

USIS-SAM

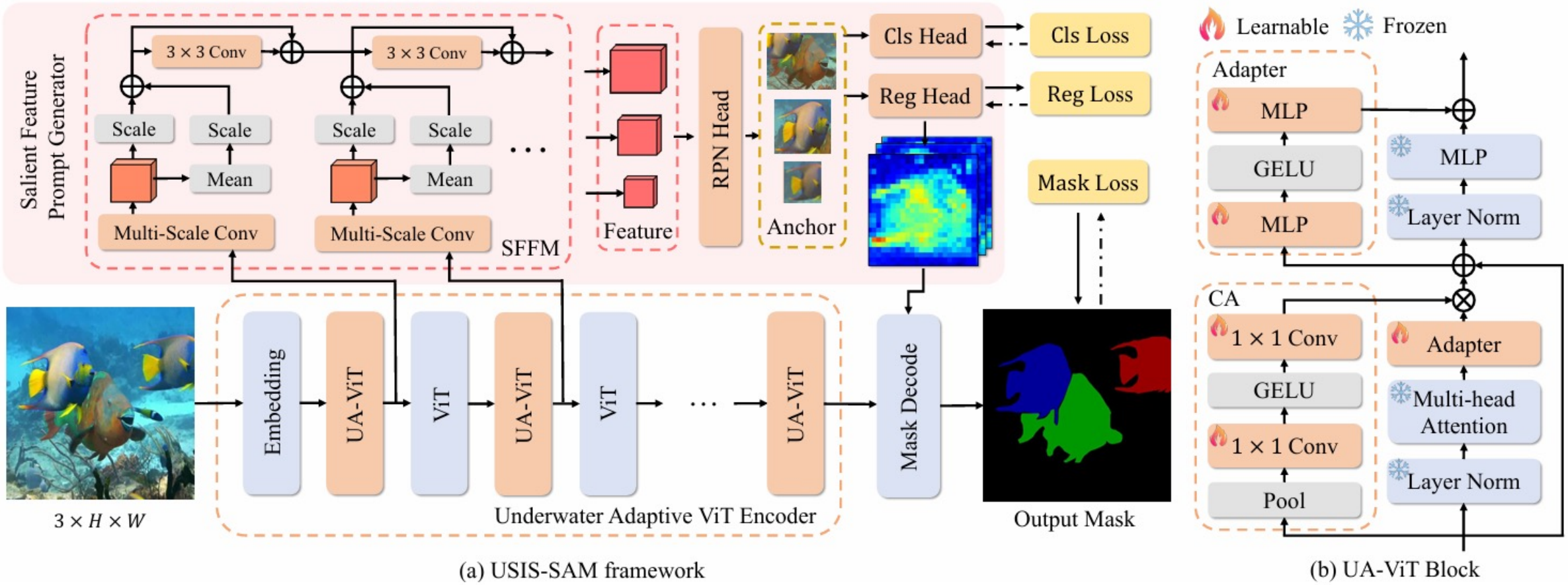


Figure 1. **(a) USIS-SAM framework.** The USIS-SAM framework modifies the SAM by adding the Underwater Adaptive ViT Encoder and the Salient Feature Prompt Generator. **(b) The structure of UA-ViT.** In the figure, SFFM stands for Salient Feature Fusion Module, CA stands for Channel Adapter.

Underwater Adaptive ViT Encoder

In USIS-SAM, we design the **Underwater Adaptive ViT (UA-ViT)** to **integrate underwater visual prompts into the network via adapter and channel adapter**. UA-ViT enables a more effective utilization of the SAM image encoder in underwater scenarios.

Adapter:

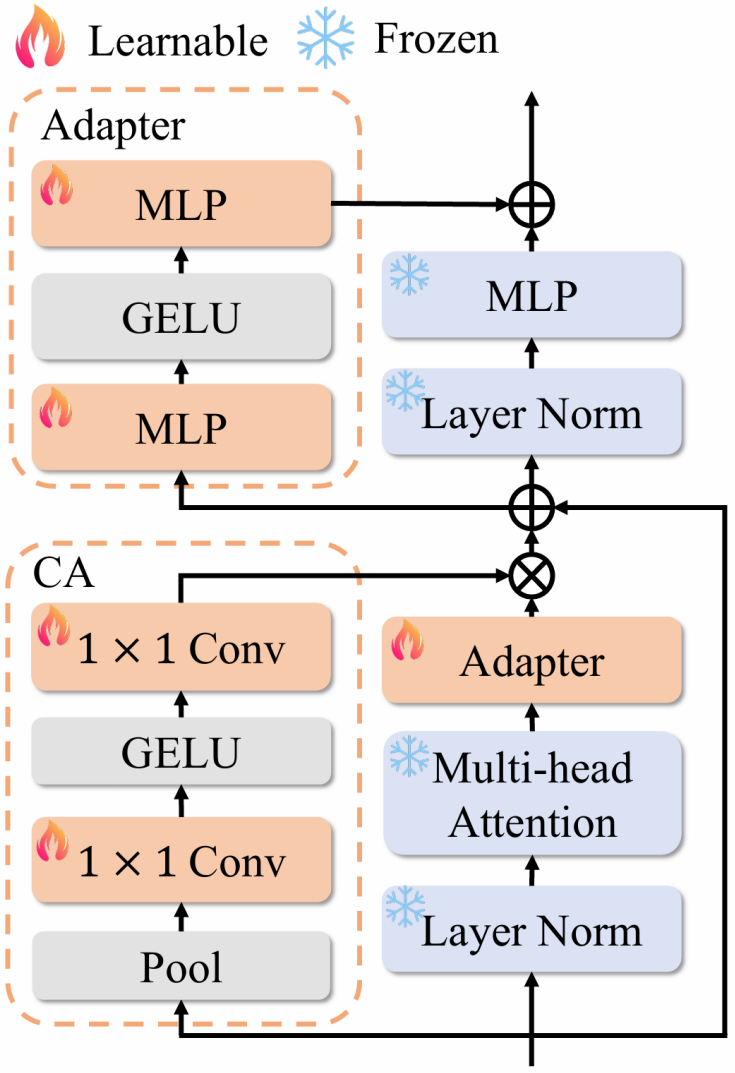
$$P = MLP_{out}\ (\sigma(MLP_{prompt}(F))),$$

where F is the input feature, and P is the output prompt for each adapter layer. σ is the activation function.

Channel Adapter:

$$C = F \times Conv_{up}(\sigma(Conv_{down}(Pool(F)))),$$

where C is the output feature after channel adapter, $Conv$ is a 1×1 convolutional layer, and $Pool$ is an average pooling layer.



Salient Feature Prompt Generator

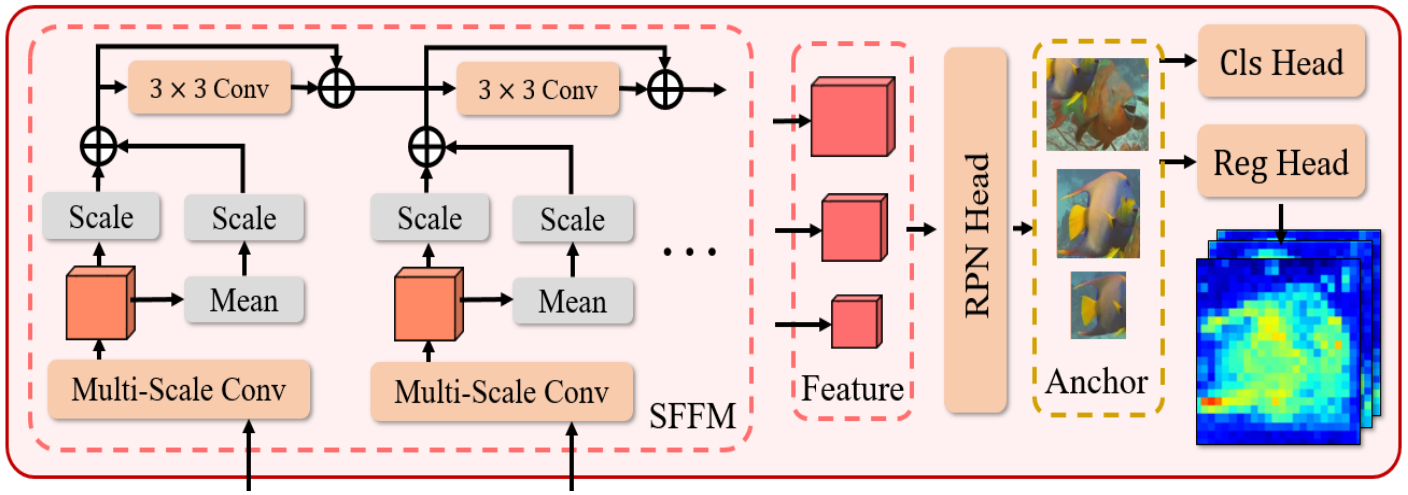


Figure 2. The structure of Salient Feature Prompt Generator (SFPG). The SFPG module efficiently filters out non-salient noise, allowing for robust feature aggregation of salient instances.

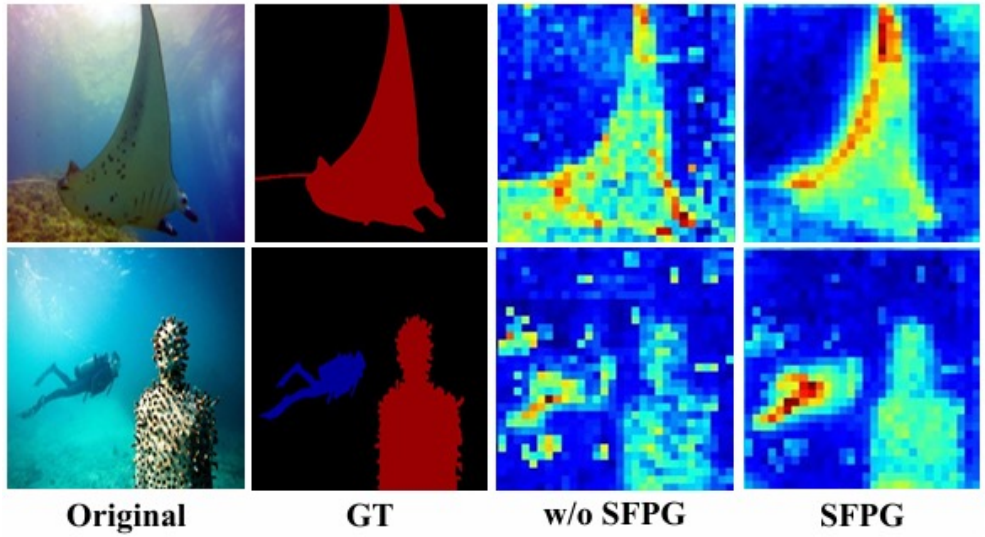


Figure 3. Visualize features generated by the Salient Feature Prompt Generator.

The USIS task needs the model to automatically recognize and segment each salient object in underwater images. However, SAM requires the user to explicitly provide foreground points, boxes, or texts as prompts to guide the model segmentation. Therefore, **we design the Salient Feature Prompt Generator to directly predict prompts embedding of salient instances, enabling end-to-end performing the USIS task**

Experiments

Method	Epoch	Backbone	Class-Agnostic			Multi-Class		
			mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
S4Net (Fan et al., 2019)	60	ResNet-50	32.8	64.1	27.3	23.9	43.5	24.4
RDPNet (Wu et al., 2021)	50	ResNet-50	53.8	77.8	61.9	37.9	55.3	42.7
RDPNet (Wu et al., 2021)	50	ResNet-101	54.7	78.3	63.0	39.3	55.9	45.4
OQTR (Pei et al., 2023)	120	ResNet-50	56.6	79.3	62.6	19.7	30.6	21.9
URank+RDPNet (Wu et al., 2021)	50	ResNet-101	52.0	80.7	62.0	35.9	52.5	41.4
URank+OQTR (Pei et al., 2023)	120	ResNet-50	49.3	74.3	56.2	20.8	32.1	23.3
WaterMask (Lian et al., 2023)	36	ResNet-50	58.3	80.2	66.5	37.7	54.0	42.5
WaterMask (Lian et al., 2023)	36	ResNet-101	59.0	80.6	67.2	38.7	54.9	43.2
SAM+BBox (Kirillov et al., 2023)	24	ViT-H	45.9	65.9	52.1	26.4	38.9	29.0
SAM+Mask (Kirillov et al., 2023)	24	ViT-H	55.1	80.2	62.8	38.5	56.3	44.0
RSPrompter (Chen et al., 2023a)	24	ViT-H	58.2	79.9	65.9	40.2	55.3	44.8
URank+RSPrompter (Chen et al., 2023a)	24	ViT-H	50.6	74.4	56.6	38.7	55.4	43.6
USIS-SAM	24	ViT-H	**59.7**	**81.6**	**67.7**	**43.1**	**59.0**	**48.5**

Table 1. **Quantitative comparisons with state-of-the-arts on the USIS10K datasets.** Urank stands for an underwater image enhancement method in UnderwaterRanker (AAAI 2023 oral), SAM+BBox uses inference results from Faster RCNN as prompts for prediction, SAM+Mask stands for Mask RCNN networks use SAM as backbone. The RSPrompter in the table is the RSPrompter-anchor framework.

Experiments

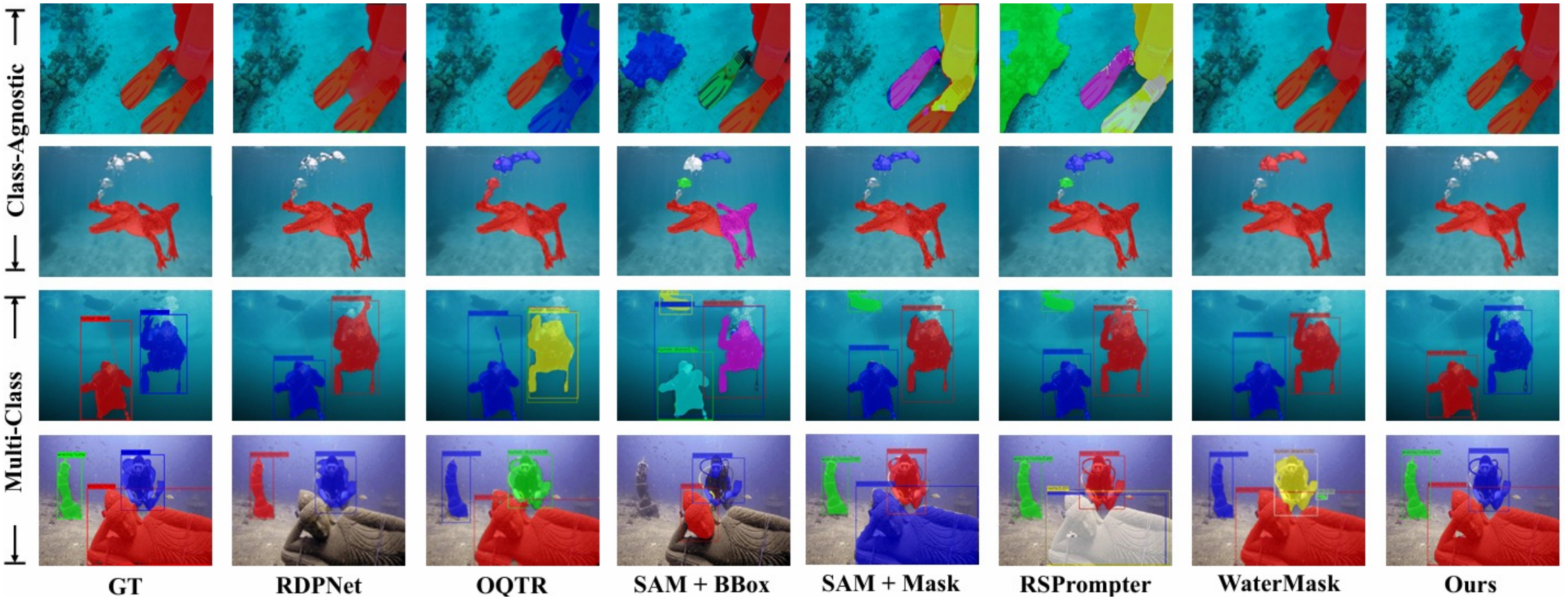


Figure 4. **Qualitative comparison on the USIS10K dataset.** Each salient instance is represented by a unique color, and the segmented mask is superimposed on the image.

Ablation Study

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o UA-ViT	41.5 (-1.6)	57.4 (-1.6)	47.0 (-1.5)
replace SFPG	42.2 (-0.9)	58.3 (-0.7)	47.5 (-1.0)

Table 2. Effectiveness of each component in USIS-SAM, replace SFPG means to use Multi-scale Feature Enhancer Module in RSPrompter (TGARS'24) instead of SFPG.

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o Adapter	41.7 (-1.4)	57.3 (-1.7)	47.3 (-1.2)
w/o CA	42.0 (-1.1)	57.7 (-1.3)	47.1 (-1.4)

Table 4. Effectiveness of each component in Underwater Adaptive ViT Encoder, w/o Adapter and w/o CA denote the removal of Adapters and Channel Adapter.

Methods	mAP	AP_{50}	AP_{75}
OQTR (Pei et al., 2023)	67.2	88.1	**81.7**
USIS-SAM	**70.1**	**89.0**	78.2

Table 3. Generalization Ability of USIS-SAM. Quantitative comparisons with state-of-the-art methods on SIS10K indicate that USIS-SAM did not overfit our dataset.

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o SFFM	42.3 (-0.8)	58.5 (-0.5)	47.2 (-1.3)
w/o Multi-Conv	42.5 (-0.6)	58.6 (-0.4)	47.7 (-0.8)

Table 5. Effectiveness of each component in Salient Feature Prompt Generator, w/o SFFM and w/o Multi-Conv denote the removal of the salient feature fusion module and multi-scale convolution module.

Conclusion and Future Work

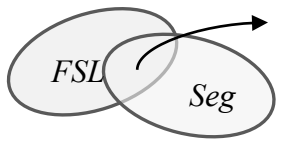
- We have constructed the **first general underwater salient image instance segmentation dataset** with pixel-level annotations, which enables us to **comprehensively explore the underwater salient instance segmentation task**.
- we first attempt to apply Segment Anything (SAM) model to underwater salient instance segmentation and propose **USIS-SAM**, aiming to improve the segmentation accuracy in complex underwater scenes. Extensive experiments have validated the **effectiveness** and **generalizability** of USIS-SAM.
- In future work, we plan to extend the USIS datasets to **broader and more challenging underwater images and underwater videos**.

IEEE TMM 2024

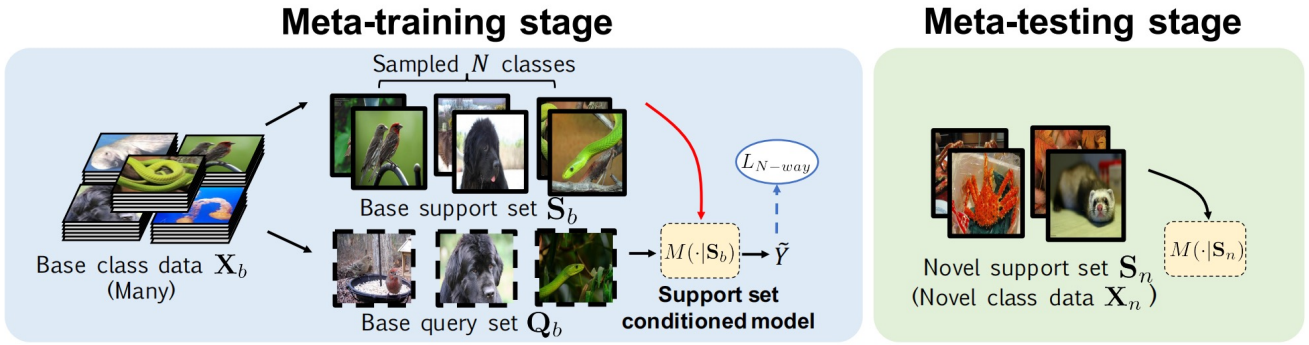
Query-guided Prototype Evolution Network for Few-Shot Segmentation

Runmin Cong, Hang Xiong, Jinpeng Chen, Wei Zhang, Qingming Huang, and Yao Zhao

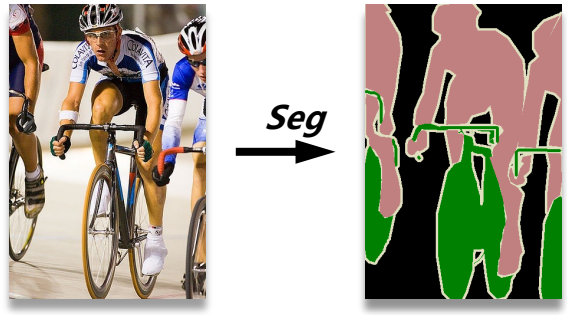
Problem Definition: Few-Shot Segmentation

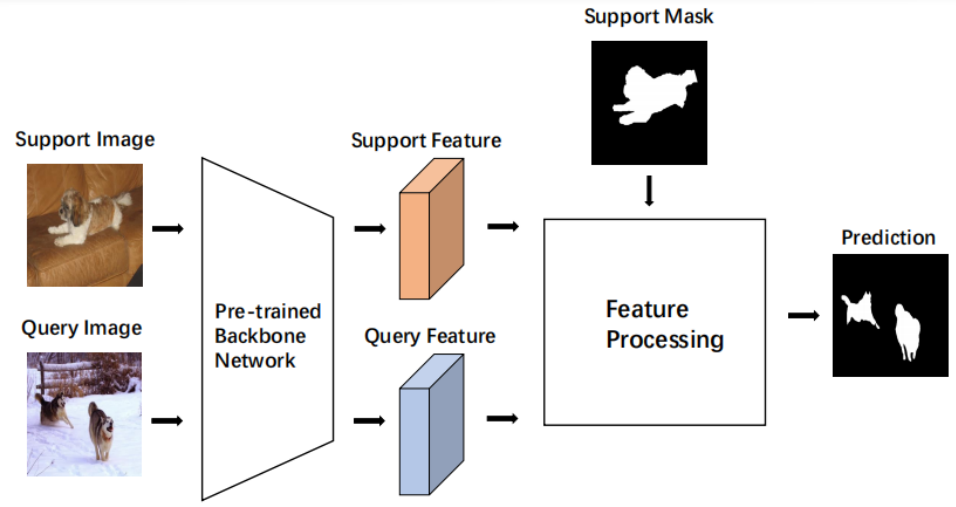


Few-shot learning[1]



Segmentation





Common network frameworks for FSS[2]

Few-Shot Segmentation

Main Purpose

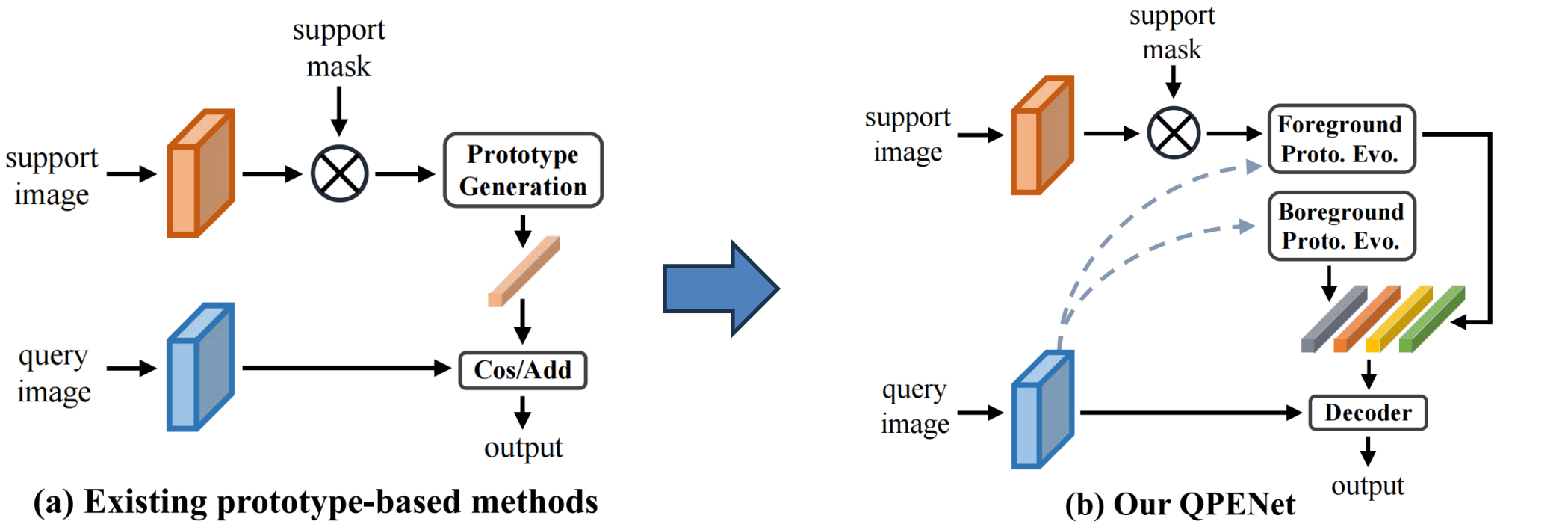
Segment the query image under the guidance of the support branch

Main Challenges

The query and support images differ in appearance, shape, and scale

[1] Chen et al, "A closer look at few-shot classification." ICLR 2019. [2] Tian et al, "Prior guided feature enrichment network for few-shot segmentation." TPAMI 2020.

Motivation



Previous FSS methods only use supporting features to generate prototypes, ignoring the **specific needs** of the query. The **large difference** between the query image and the supporting features can bring **negative impact** to the final prediction results.

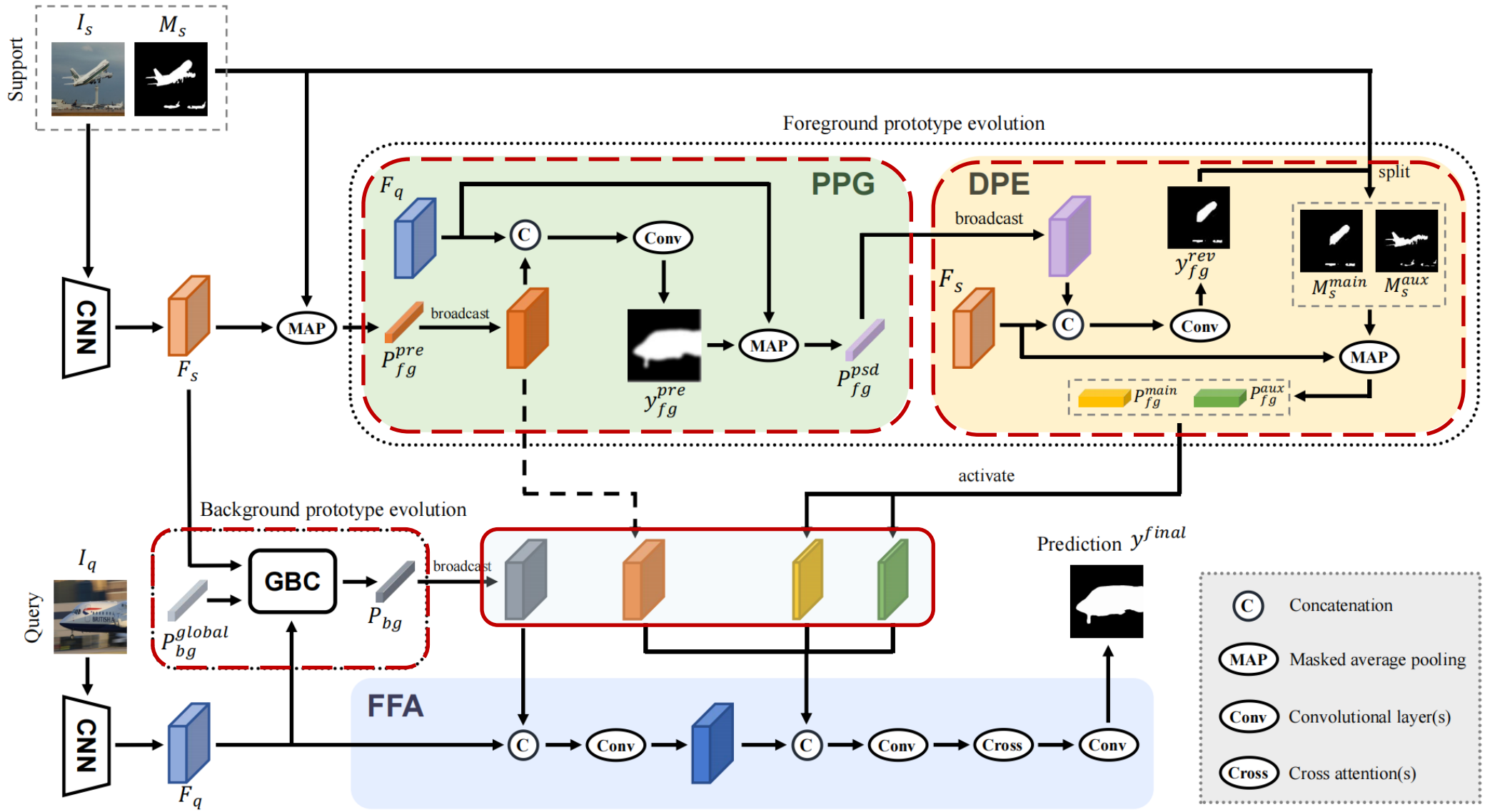
Therefore, we propose query-guided prototype evolution networks, which integrate query features into the generation process of foreground and background prototypes.

Contributions

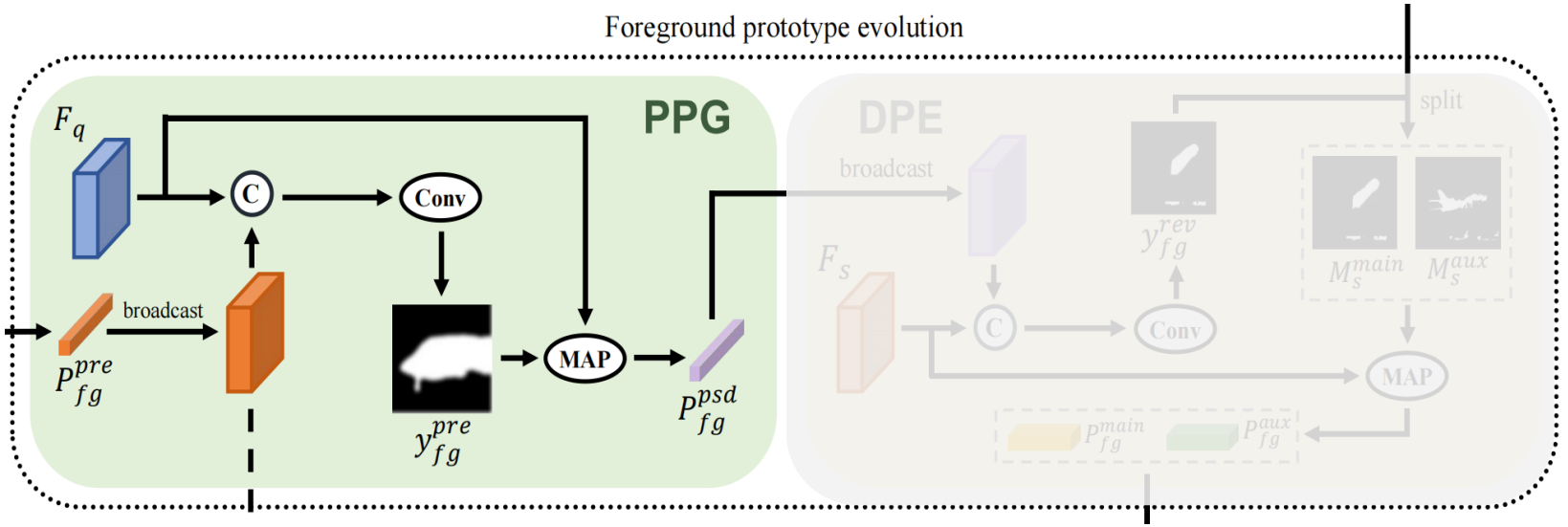
- We propose a novel FSS method named **QPENet**, which embodies the core idea of using the query image to guide the evolution of prototypes, thereby enhancing their efficacy for segmenting the specific query image.
- We introduce innovative **PPG** and **DPE** modules to facilitate the evolution of the foreground prototype following a **support-query-support** process, and a novel **GBC** module to eliminate components that might reflect the current foreground class from the global background prototype.
- Extensive experiments conducted on two widely adopted datasets demonstrate that our model excels in delivering **state-of-the-art** performance in the domain of FSS.

https://github.com/rmcong/QPENet_TMM24

Our QPENet


Support
I_s
M_s
CNN
F_s
MAP
Foreground prototype evolution
PPG
F_q
C
Conv
broadcast
P_{fg}^{pre}
y_{fg}^{pre}
MAP
P_{fg}^{psd}
DPE
broadcast
F_s
split
y_{fg}^{rev}
M_s^{main}
M_s^{aux}
C
Conv
MAP
P_{fg}^{main}
P_{fg}^{aux}
activate
Background prototype evolution
I_q
Query
GBC
P_{bg}^{global}
P_{bg}
broadcast
Prediction y^{final}
FFA
CNN
F_q
C
Conv
C
Conv
Cross
Conv
Concatenation
MAP
Masked average pooling
Conv
Convolutional layer(s)
Cross
Cross attention(s)

Evolution of Foreground Prototype



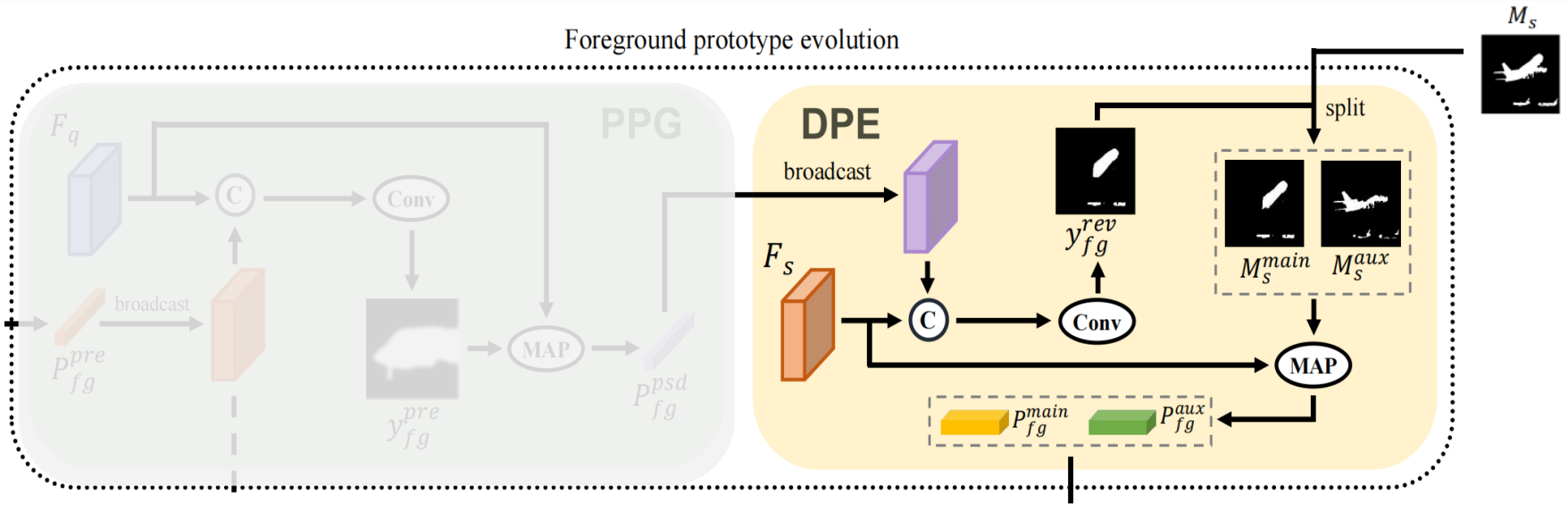
Pseudo-prototype Generation Module (PPG)

In order to obtain the **specific requirements** of the query image, the query image is initially predicted and the predicted foreground part is generated into a **pseudo-prototype**. The generated pseudo-prototype serves as a vehicle for the specific requirements of the query image.

$$y_{fg}^{pre} = Convs(Concat(F_q, BC(P_{fg}^{pre}))).$$

$$P_{fg}^{psd} = MAP(F_q, y_{fg}^{pre}) = \frac{\sum_{i=1}^{h\times w} F_q(i) \otimes y_{fg}^{pre}(i)}{\sum_{i=1}^{h\times w} y_{fg}^{pre}(i)}.$$

Evolution of Foreground Prototype



$$y_{fg}^{rev} = Convs(Concat(F_s, BC(P_{fg}^{psd}))).$$

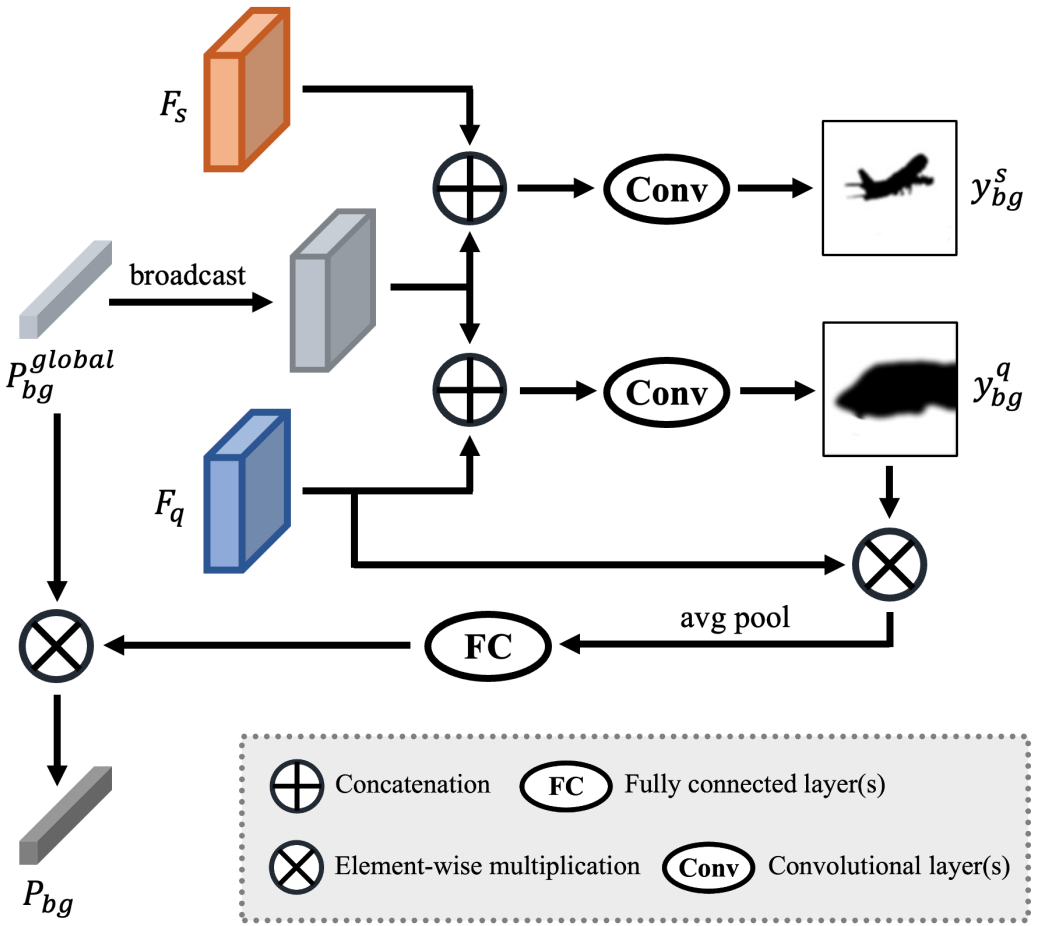
$$M_s^{main} = \mathbb{1}\{M_s = 1\} \otimes \mathbb{1}\{y_{fg}^{rev} = 1\},$$

$$M_s^{aux} = \mathbb{1}\{M_s = 1\} \otimes \mathbb{1}\{y_{fg}^{rev} \neq 1\}.$$

Dual Prototype Evolution Module (DPE)

The **pseudo-prototype** is used for guided segmentation of the support image. The segmentation result is compared with the support mask so that the support foreground is separated into two parts **with different intimacy** from the query image.

Evolution of Background Prototype



The background semantics of the query image is significantly different from that of the support image.

Initialize a **global background prototype**, and during training, predict the background maps of the query and support images to **update** it.

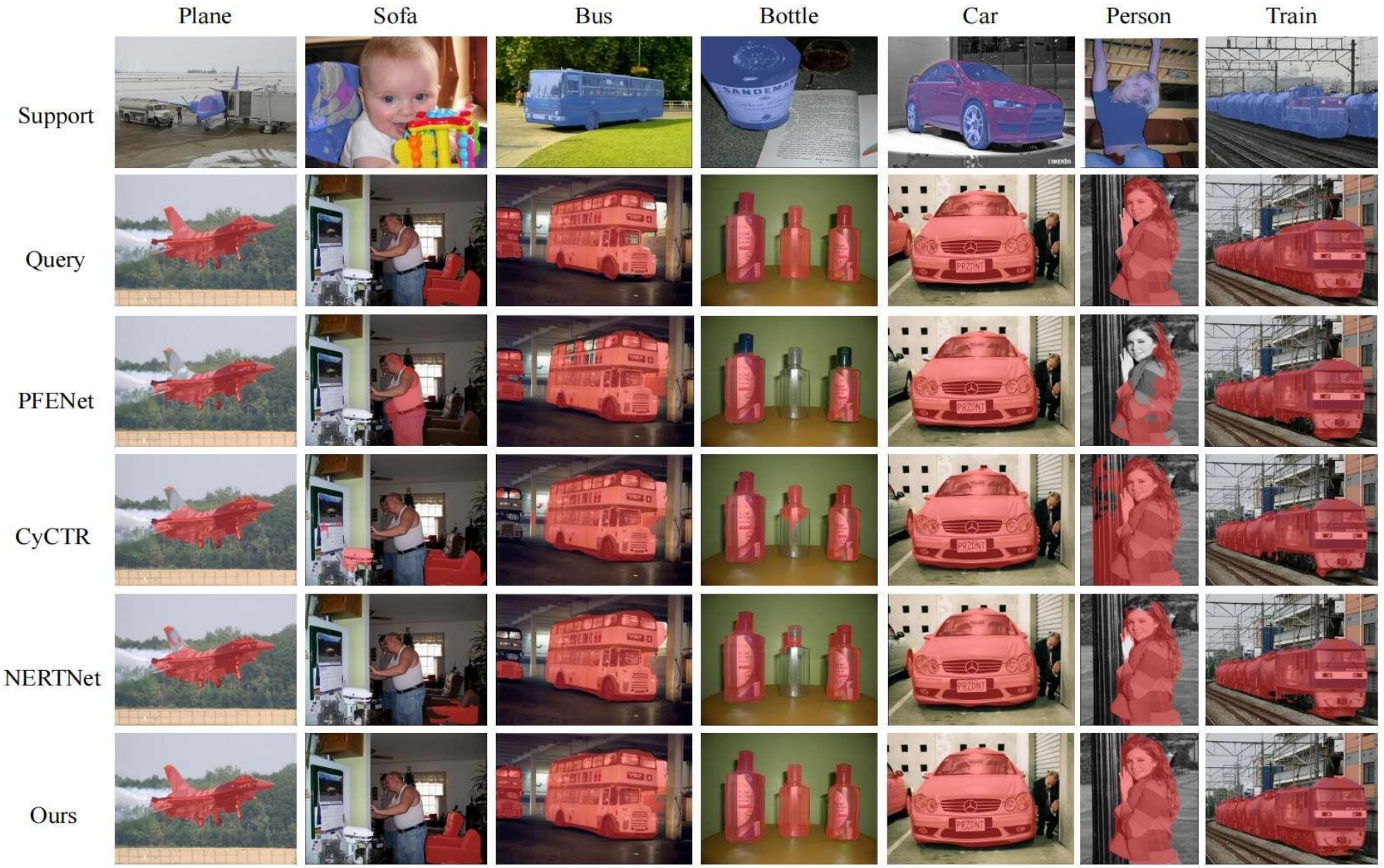
$$y_{bg}^{s} = Convs(Concat(F_s, BC(P_{bg}^{global}))),$$

$$y_{bg}^{q} = Convs(Concat(F_q, BC(P_{bg}^{global}))).$$

After obtaining the specific requirements of the query background, we **adapt** the global background prototype.

$$P_{bg} = P_{bg}^{global} \otimes FC(FC(GAP(F_q \otimes y_{bg}^{q}))),$$

Experiments



Experiments

PASCAL-5i dataset

Backbone	Methods	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
ResNet-50	CANet [61]	52.5	65.9	51.3	51.9	55.4	66.2	55.5	67.8	51.9	53.2	57.1	69.6
	PPNet [25]	48.6	60.6	55.7	46.4	52.8	-	58.9	68.3	66.8	58.0	63.0	-
	RPMMs [24]	55.2	66.9	52.6	50.7	56.3	-	56.3	67.3	54.5	51.0	57.3	-
	PFENet [1]	61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
	RePRI [62]	59.8	68.3	62.1	48.5	59.7	-	64.6	71.4	71.1	59.3	66.6	-
	ASGNet [26]	58.8	67.9	56.8	53.7	59.3	69.2	63.7	70.6	64.2	57.4	63.9	74.2
	CMN [30]	64.3	70.0	57.4	59.4	62.8	72.3	65.8	70.4	57.6	60.8	63.7	72.8
	SAGNN [31]	64.7	69.6	57.0	57.2	62.1	73.2	64.9	70.0	57.0	59.3	62.8	73.3
	CyCTR [63]	65.7	71.0	59.5	59.7	64.0	-	69.3	73.5	63.8	63.5	67.5	-
	DPNet [64]	60.7	69.5	62.8	58.0	62.7	-	64.7	70.8	69.0	60.1	66.2	-
	DCP [35]	63.8	70.5	61.2	55.7	62.8	75.6	67.2	73.1	66.4	64.5	67.8	79.7
	NERTNet [28]	65.4	72.3	59.4	59.8	64.2	**77.0**	66.2	72.8	61.7	62.2	65.7	78.4
	Ours	65.2	71.9	64.1	59.5	**65.2**	76.7	68.4	74.0	67.4	65.2	**68.8**	**80.0**
ResNet-101	DAN [65]	54.7	68.6	57.8	51.6	58.2	71.9	57.9	69.0	60.1	54.9	60.5	72.3
	PPNet [25]	52.7	62.8	57.4	47.7	55.2	70.9	60.3	70.0	69.4	60.7	65.1	77.5
	PFENet [1]	60.5	69.4	54.4	55.9	60.1	72.9	62.8	70.4	54.9	57.6	61.4	73.5
	RePRI [62]	59.6	68.6	62.2	47.2	59.4	-	66.2	71.4	67.0	57.7	65.6	-
	ASGNet [26]	59.8	67.4	55.6	54.4	59.3	71.7	64.6	71.3	64.2	57.3	64.4	75.2
	CyCTR [63]	69.3	72.7	56.5	58.6	64.3	72.9	73.5	73.2	60.1	66.8	67.0	75.0
	NERTNet [28]	65.5	71.8	59.1	58.3	63.7	75.3	67.9	73.2	60.1	66.8	67.0	78.2
	Ours	67.0	73.2	63.7	60.1	**66.0**	**77.1**	69.8	75.5	66.8	66.3	**69.6**	**81.1**

Experiments

CoCo-20i dataset

Backbone	Methods	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
ResNet-50	PPNet [25]	28.1	30.8	29.7	27.7	29.0	-	39.0	40.8	37.1	37.3	38.5	-
	RPMMs [24]	29.5	36.8	28.9	27.0	30.6	-	33.8	42.0	33.0	33.3	35.5	-
	ASGNet [26]	-	-	-	-	34.6	60.4	-	-	-	-	42.4	67.0
	RePRI [62]	31.2	38.1	33.3	33.0	34.0	-	38.5	46.2	40.0	43.6	42.1	-
	CyCTR [63]	38.9	43.0	39.6	39.8	40.3	-	41.1	48.9	45.2	47.0	45.6	-
	CMN [30]	37.9	44.8	38.7	35.6	39.3	61.7	42.0	50.5	41.0	38.9	43.1	63.3
	DPNet [64]	-	-	-	-	37.2	-	-	-	-	-	42.9	-
	DCP [35]	40.9	43.8	42.6	38.3	41.4	-	45.8	49.7	43.7	46.6	46.5	-
	NERTNet [28]	36.8	42.6	39.9	37.9	39.3	**68.5**	38.2	44.1	40.4	38.4	40.3	69.2
	Ours	41.5	47.3	40.9	39.4	**42.3**	67.4	47.3	52.4	44.3	44.9	**47.2**	**69.5**
ResNet-101	DAN [65]	-	-	-	-	24.4	62.3	-	-	-	-	29.6	63.9
	PFENet [1]	34.3	33.0	32.3	30.1	32.4	58.6	38.5	38.6	38.2	34.3	37.4	61.9
	SCL [27]	36.4	38.6	37.5	35.4	37.0	-	38.4	40.5	41.5	38.7	39.9	-
	SAGNN [31]	36.1	41.0	38.2	33.5	37.2	60.9	40.9	48.3	42.6	38.9	42.7	63.4
	NERTNet [28]	38.3	40.4	39.5	38.1	39.1	67.5	42.3	44.4	44.2	41.7	43.2	69.6
	Ours	39.8	45.4	40.5	40.0	**41.4**	**67.8**	47.2	54.9	43.4	45.4	**47.7**	**70.6**

Ablation Study

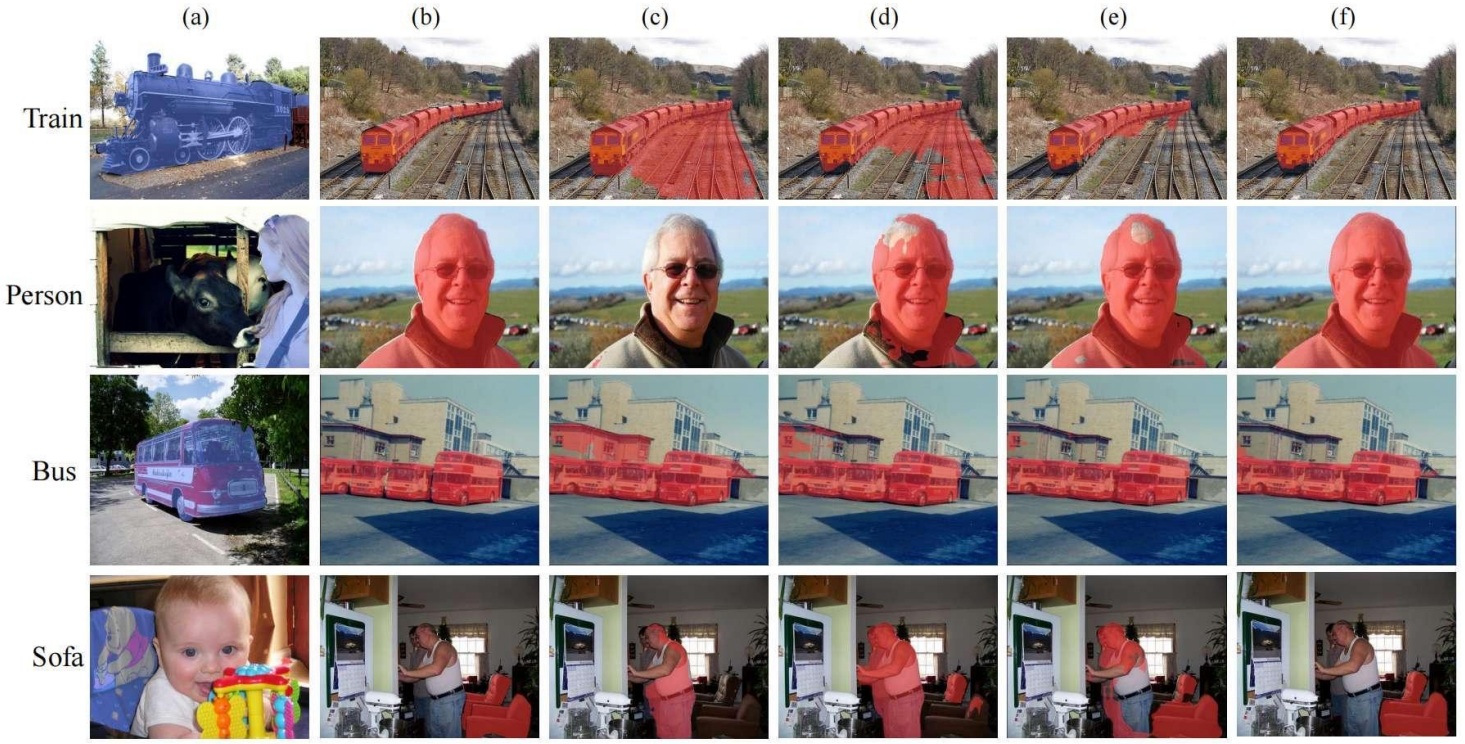


Fig. 5: Qualitative results for component analysis. (a) Annotated support image. (b) Annotated query image. (c) Predictions of the baseline model. (d) Predictions of the baseline model enhanced by *FGPE*. (e) Predictions of the baseline model enhanced by *FGPE* and *BGPE*. (f) Predictions of the full model.

TABLE IV: Comparison of class mIoU(%) and FB-IoU(%) across different foreground prototypes.

P_{fg}^{pre}	P_{fg}^{main}	P_{fg}^{aux}	mIoU	FB-IoU
✓			63.3	74.6
	✓		61.8	72.0
		✓	58.1	69.4
	✓	✓	64.1	74.7
✓	✓	✓	65.2	76.7

TABLE V: Ablation study of activation maps and DPE, assessed by class mIoU (%) and FB-IoU (%).

	mIoU	FB-IoU
w/o activation maps	61.6	73.0
w SGM	63.5	75.6
Full Model	65.2	76.7

Conclusion

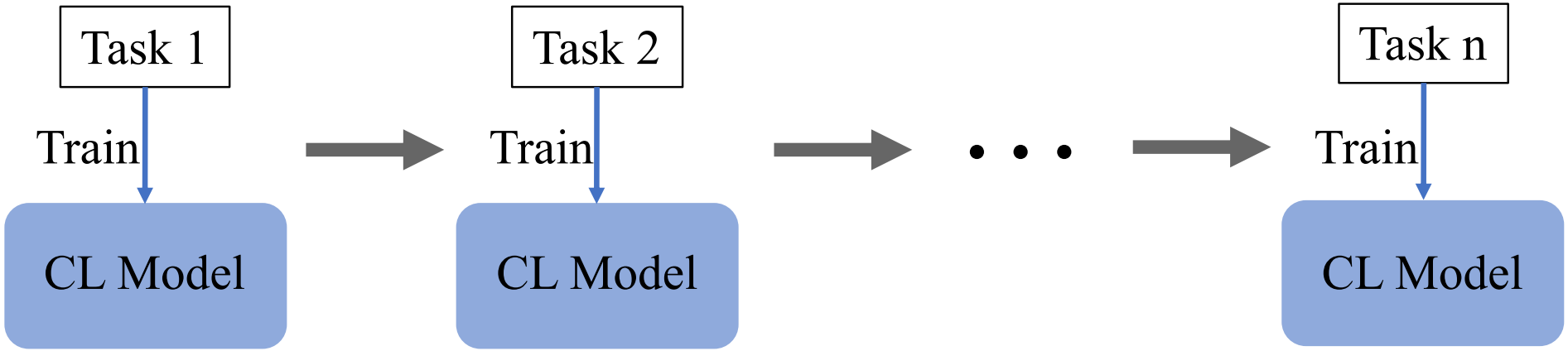
- In this paper, we propose **QPENet** to optimize prototypes to address the **specific characteristics** of the current query image.
- QPENet presents two types of prototypes, namely, **foreground and background prototypes**, both of which evolve under the guidance of the query features.
- Furthermore, we design the **FFA module** to maximize the utilization of these prototypes.
- In addition, extensive experiments on **PASCAL-5i** and **COCO-20i** datasets demonstrated the superiority of our proposed model.

IEEE TMM 2024

Modeling Inner- and Cross-Task Contrastive Relations for Continual Image Classification

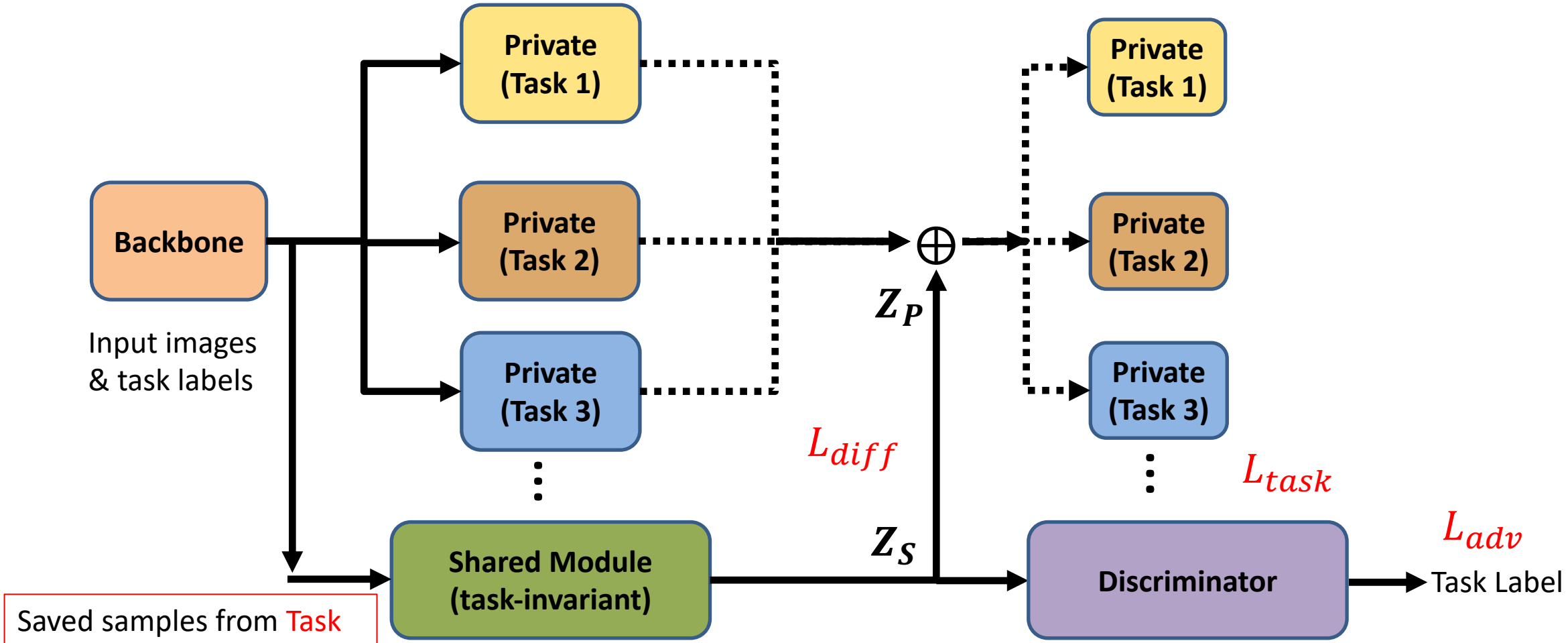
Yuxuan Luo, Runmin Cong, Xialei Liu, Horace Ho Shing Ip, and Sam Kwong

Introduction



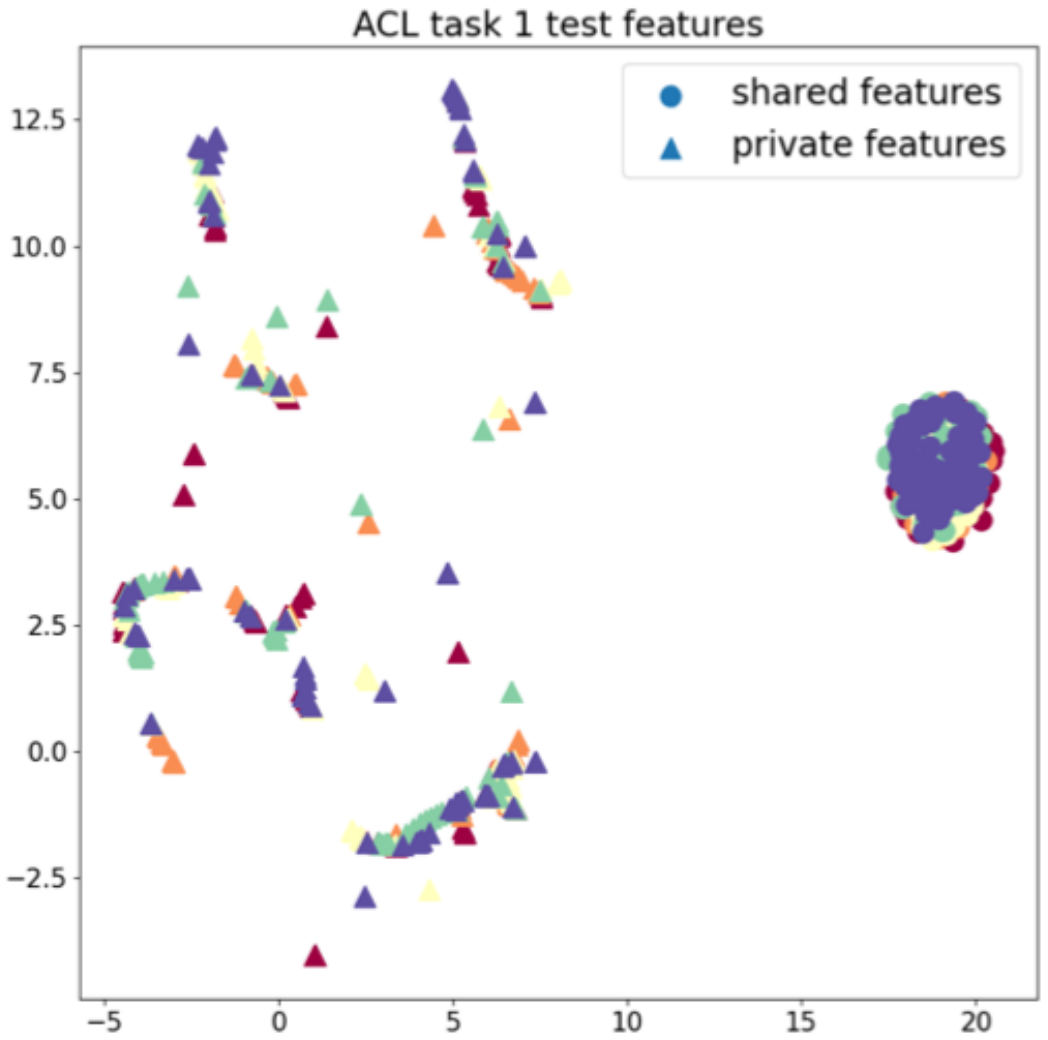
- Continual image classification, also known as continual learning (CL), aims to enable models to learn a number of classification tasks sequentially and predict new tasks(or categories) that have no overlap with previously learned tasks while maintaining good performance on previously learned tasks.
- **Catastrophic forgetting** refers to the problem that a model's performance on previously learned tasks significantly drops after learning a new task since no previous task samples are available in current task training, which is the key challenge that must be addressed in continual learning models.

Motivation —— ACL method



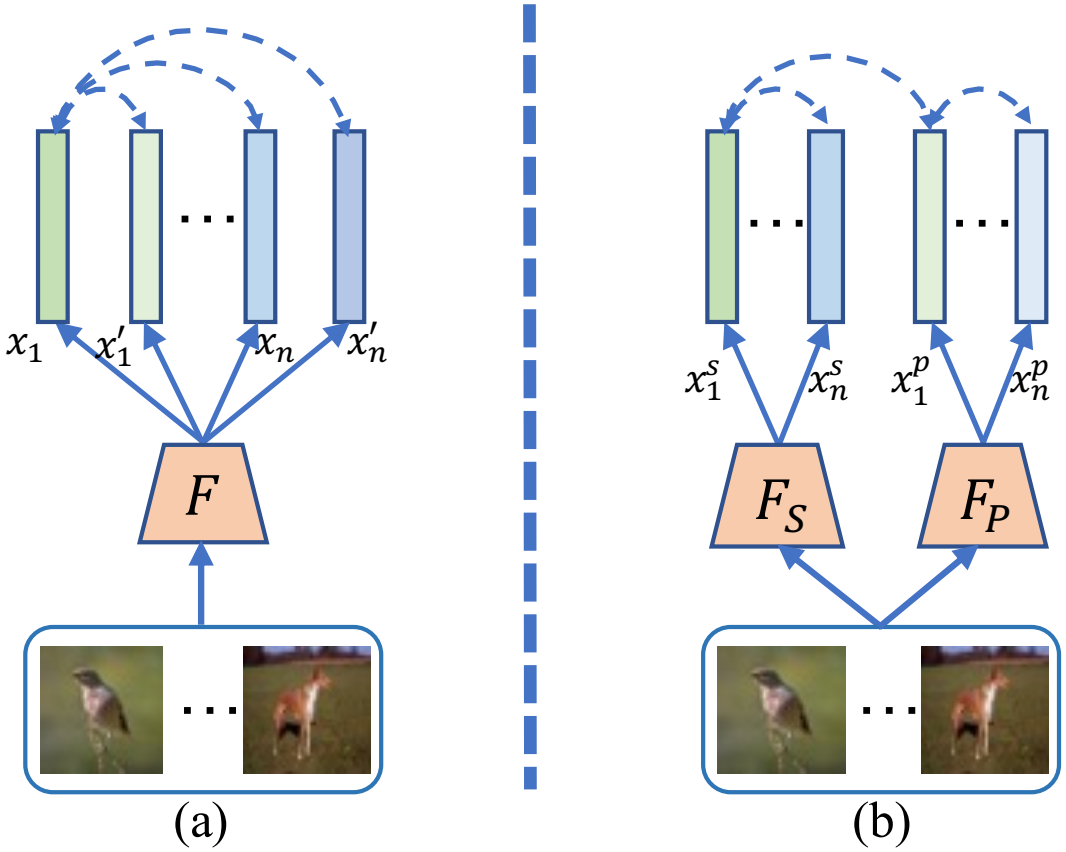
Adversarial Continual Learning, ECCV 2020

Motivation



- Current feature decomposition methods generally focus on increasing the distance between task-invariant and task-variant features without fully considering their relationship, which brings about **low-quality decomposition**. Moreover, these methods mainly focus on decomposing features within an individual task and **ignore the important information conveyed by the relationships between features across tasks**.

Motivation



- Most contrastive learning methods learn instance feature representations through **instance relation** by constructing learning pairs.
- By contrast, we design a comprehensive contrastive learning strategy under the feature decoupling framework to **enhance feature representation and model the relationship between task-invariant and task-variant features.**

Contributions

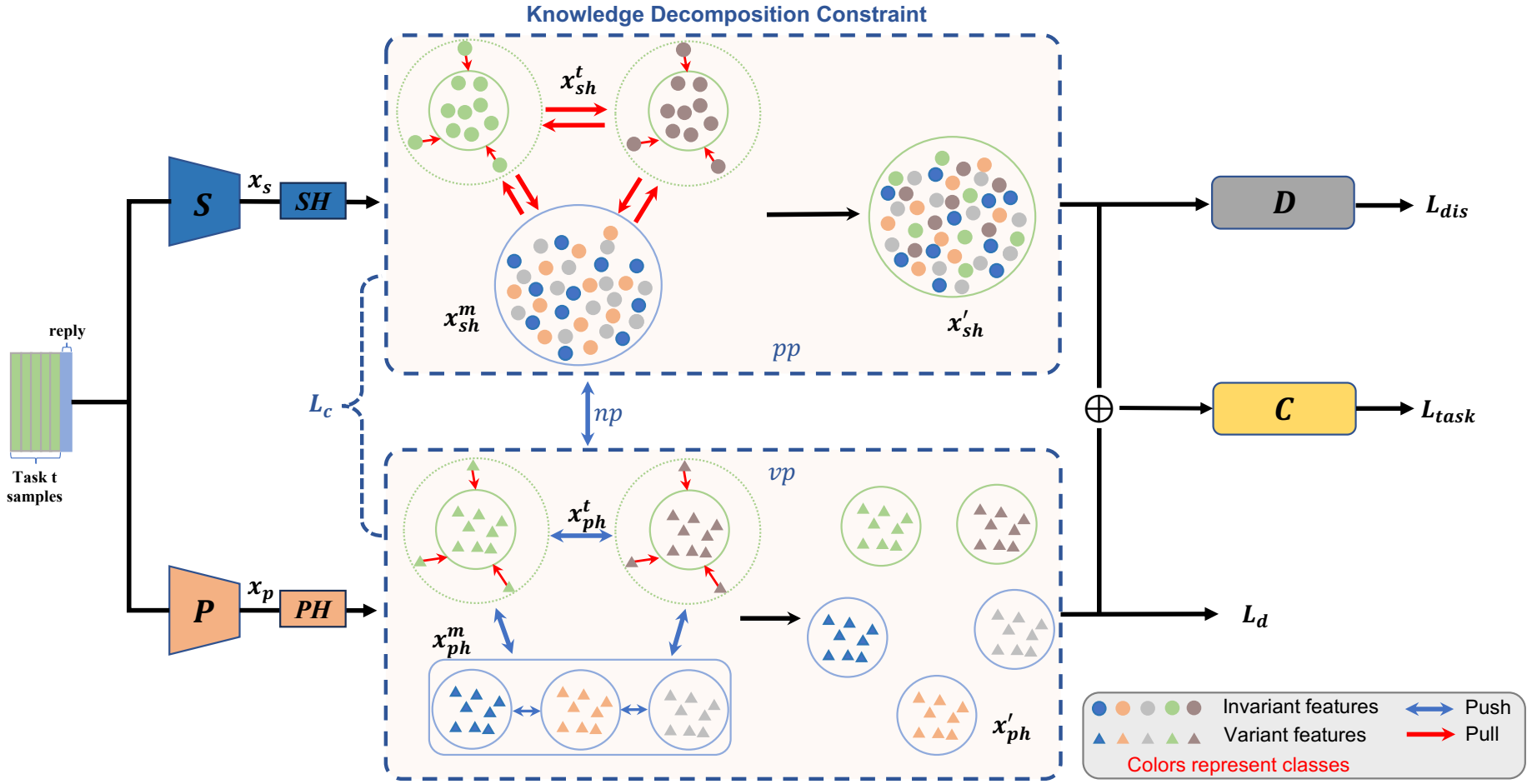
a) We propose an **ACCL** method for the continual classification task that effectively explores inner- and cross-task relations by constructing all-round contrastive learning pairs.

b) To capture the inner-task relations, we propose the KDC that utilizes **three types of contrast to decompose generalized task-invariant features and diverse variant features.**

c) Considering the relation between current and previous tasks, we incorporate **replay samples with contrastive constraints** to alleviate the forgetting problem and model the cross-task relations.

Our ACCL Method

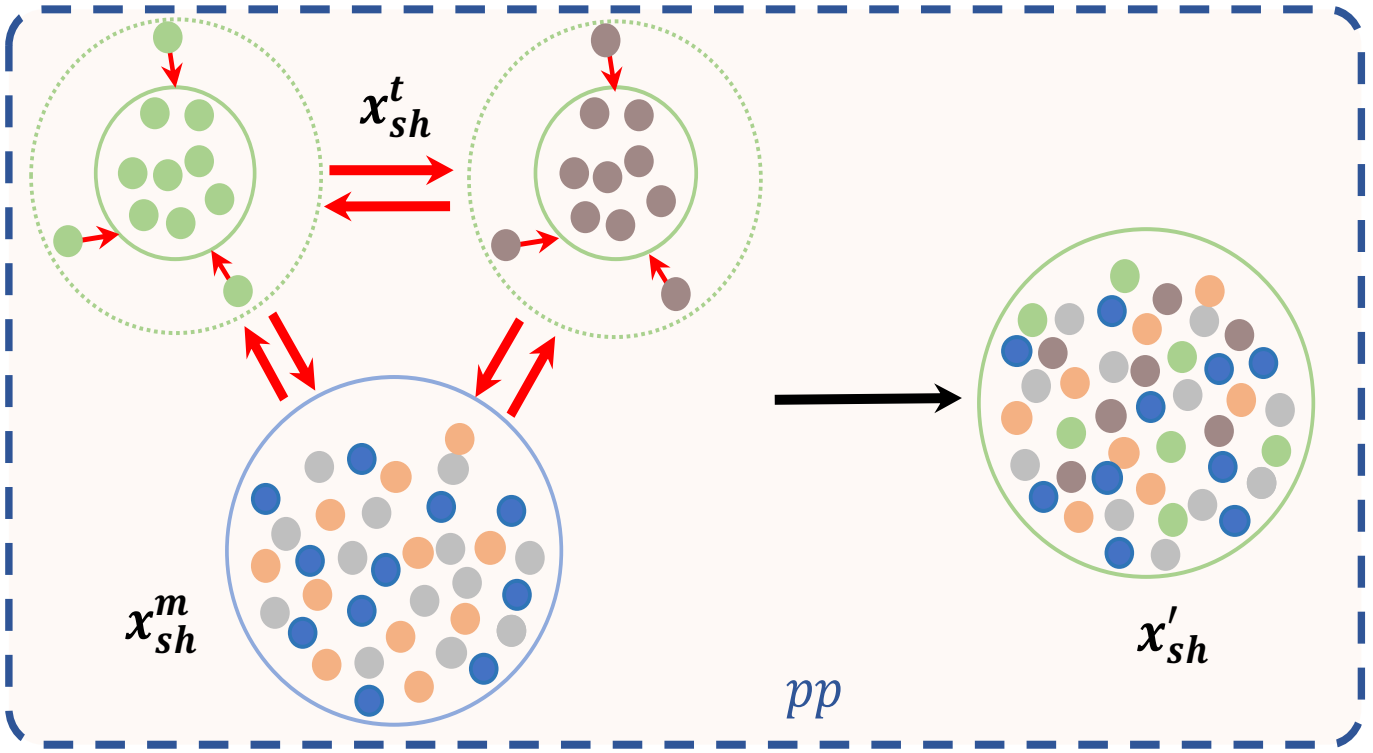
https://github.com/rmcong/ACCL TMM24



Our model comprises task-invariant and task-variant feature extractors, denoted as S and P, respectively, along with their respective projection heads, SH and PH. Additionally, it incorporates a knowledge decomposition module (KDC), a discriminator (D), and a classifier (C) specific to the task t.

Knowledge Decomposition Constraint

Improving Consistency of Task-Invariant Features.



Inner-task relation:

Learning pairs constructed by the task-invariant features (pp)：

$$pp = \{(x_{sh}^{c=i}, x_{sh}^{c=i}), (x_{sh}^{c=i}, x_{sh}^{c\neq i})\}.$$

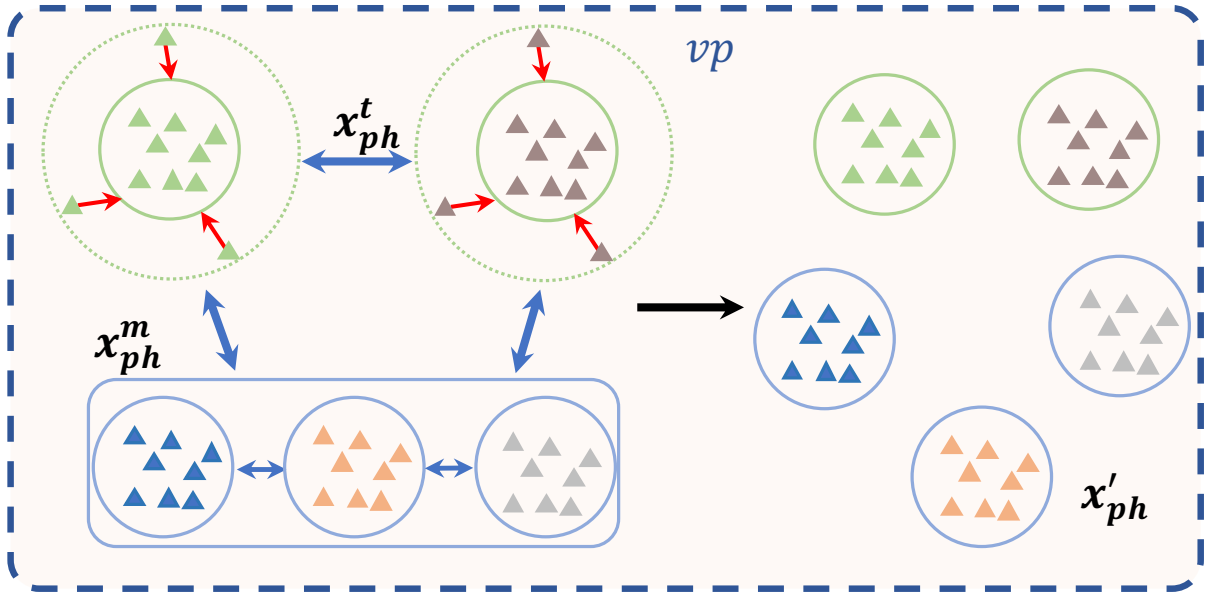
Cross-task relation:

Learning pairs constructed by the memory and current task:

pp ***includes positive pairs*** $\{x_{sh}^{c=i}, x_{sh}^{m}\}$

Knowledge Decomposition Constraint

Well-distributed Task-Variant Features.



Inner-task relation:

Learning pairs constructed by the task-variant features (vp):

positive pairs $\{(x_{ph}^{c=i}, x_{ph}^{c=i})\}$,

negative pairs $\{(x_{ph}^{c=i}, x_{ph}^{c\neq i})\}$.

Cross-task relation:

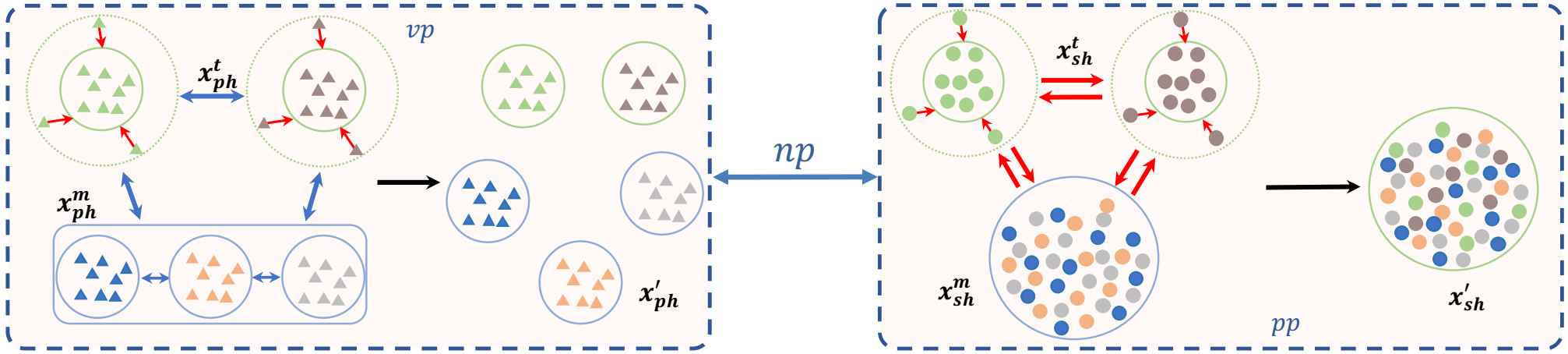
Learning pairs constructed by the memory and current task:

vp ***includes negative pairs*** $\{(x_{ph}^{c=i}, x_{ph}^{m})\}$

KDC combines vp to form the constraint L_p, which is similar to L_{sp}.

Knowledge Decomposition Constraint

Enhancing Discriminability between Invariant and Variant Features.



Inner-task relation:

Learning pairs construct by the task-invariant and task-variant features (np)：

$$np = \{(x_{sh}^{c=i}, x_{ph}^{c=i}), (x_{sh}^{c=i}, x_{ph}^{c\neq i})\}$$

Cross-task relation:

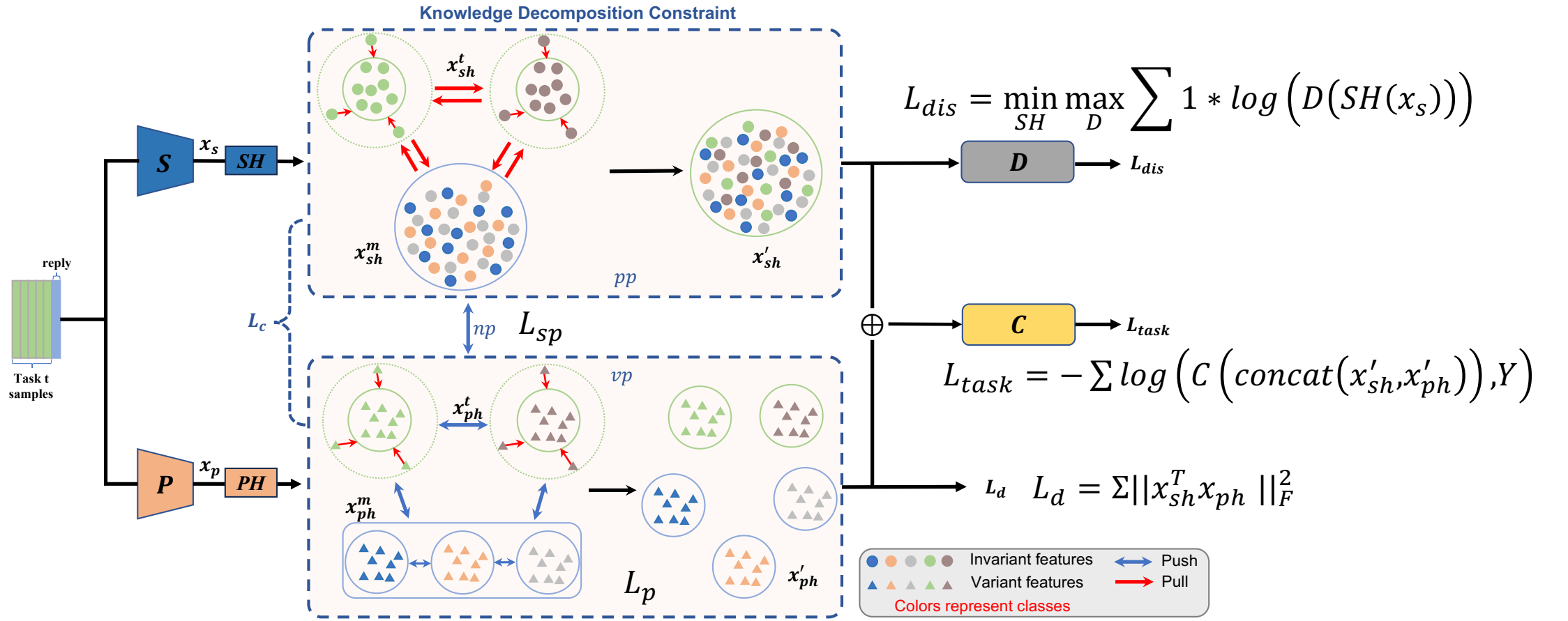
Learning pairs constructed by the memory and current task:

np ***includes negative pairs*** $\{(x_{sh}^{c=i}, x_{ph}^{m})\}$

As they use the same anchor, KDC combines np and pp to form the constraint $\boldsymbol{L_{sp}}$.

$$L_{sp} = \sum_{i\in I} \frac{-1}{|pp|} \sum_{ps\in\{pp\}} \frac{\exp(x_i \cdot x_{ps}/\tau)}{\sum_{a\in\{np,pp\}} \exp(x_i \cdot x_a/\tau)}$$

Our Method



$$L = \lambda_1 * L_{task} + \lambda_2 * L_d + \lambda_3 * L_{dis} + \lambda_4 * L_c$$

$L_c = L_{sp} + L_p + L_r$, where L_r is a regularization term.

Experiments

Quantitative comparisons with state-of-the-arts on the CIFAR-100 datasets.

Methods	Structure	Memory	ACC(%)	BWT(%)
EWC (NAS 2017)	AN	0	68.80	-0.02
GSS (NIPS 2019)	RN18	300	49.92	-
A-GEM (ICLR 2019)	RRN18	2000	63.98	-0.15
HAT (PMLR 2019)	AN	500	72.06	-0.00
ACL (ECCV 2020)	AN	200	78.08	-0.00
NCCL (NIPS 2021)	RRN18	50	74.39	-
HAL (AAAI 2021)	RRN18	200	47.88	-
GPM (ICLR 2021)	AN	0	72.48	-0.90
CCL-FP+ (BMVC 2022)	RN18	200	65.19	-
DeepCCG (NIPSW 2022)	RRN18	750	60.46	-
TRGP (ICLR 2022)	AN	0	74.46	-0.90
SGP (AAAI 2023)	AN	0	76.05	-0.01
ANCL (CVPR 2023)	RN32	2000	79.99	-
API (CVPR 2023)	AN	0	81.40	-0.08
SOI (CVPR 2023)	RN18	1000	74.27	-16.37
DFGP (ICCV 2023)	AN	0	74.59	-0.00
ACCL	AN	0	79.54	-0.26
ACCL	AN	200	80.55	0.05
ACCL	RRN18	0	80.88	0.19
ACCL	RRN18	200	**83.51**	0.13
ACCL	RRN18	500	<u>83.37</u>	0.69

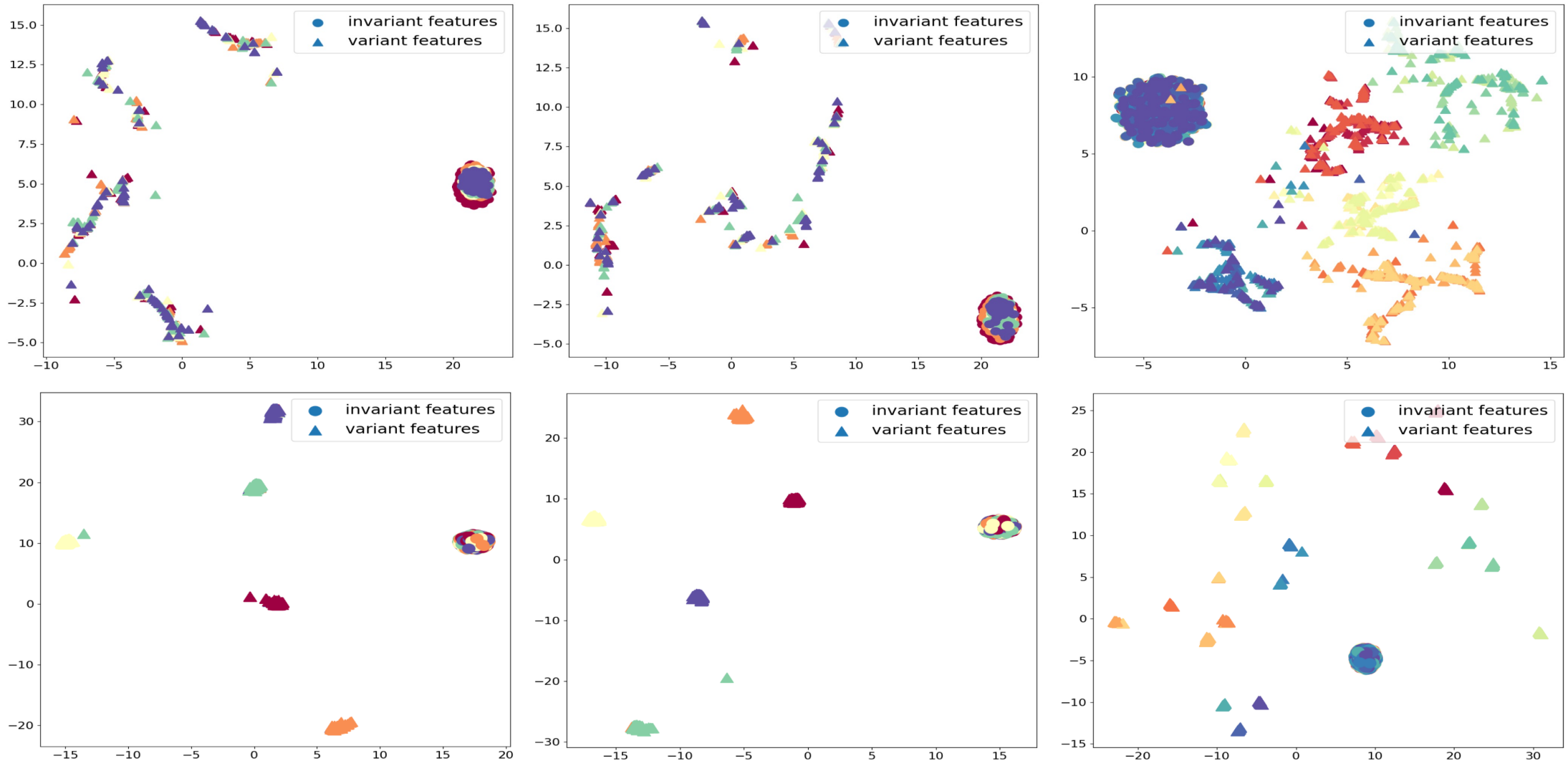
Quantitative comparisons with state-of-the-arts on the MiniImageNet datasets.

Methods	Structure	Memory	ACC(%)	BWT(%)
EWC (NAS 2017)	AN	0	52.01	-0.12
HAT (PMLR 2019)	AN	500	59.78	-0.03
GSS (NIPS 2019)	RN18	300	38.77	-
A-GEM (ICLR 2019)	RRN18	500	57.24	-0.12
ACL (ECCV 2020)	RRN18	200	62.07	0
NCCL (NIPS 2021)	RRN18	50	69.49	-
GPM (ICLR 2021)	RRN18	0	60.41	-0.70
DeepCCG (NIPSW 2022)	RRN18	750	43.04	-
TRGP (ICLR 2022)	RRN18	0	61.78	-0.50
SGP (AAAI 2023)	RRN18	0	62.83	-0.01
API (CVPR 2023)	RRN18	0	65.90	-0.3
SOI (CVPR 2023)	RN18	1000	39.61	-16.01
DFGP (ICCV 2023)	RRN18	0	69.92	-0.10
ACCL	RRN18	0	74.66	-0.51
ACCL	RRN18	200	**75.88**	0.12
ACCL	RRN18	500	<u>75.68</u>	0.45

Visualization

Baseline

ACCL



(c) (b) (a)

Ablation Study

#	pp	vp	np	r	ACC(%)	BWT(%)
1					54.60	-0.14
2	✓				60.51	-10.29
3		✓			57.97	-20.52
4	✓	✓			56.93	-21.54
5	✓		✓		72.03	-0.21
6	✓	✓		✓	70.16	-2.84
7	✓		✓	✓	72.70	-0.11
8	✓	✓	✓	✓	75.88	0.12

Ablation study of each component in KDC on MiniImageNet dataset.

λ_2		λ_3		λ_4	
value	ACC	value	ACC	value	ACC
0.06	75.40	0.001	75.29	0.10	75.51
0.08	75.72	0.003	75.43	0.25	75.88
0.10	75.88	0.005	75.88	0.50	75.68
0.12	75.80	0.007	75.54	0.75	75.70
0.14	75.68	0.009	75.33	1.00	75.67

Hyperparameter finetuning on the MiniImageNet dataset.

L_c	L_{dis}	L_d	ACC(%)	BWT(%)
	✓	✓	54.60	-0.14
✓	✓		66.69	-0.02
✓		✓	71.38	-0.04
✓	✓	✓	75.88	0.12

Ablation study of each loss term in KDC on MiniImageNet dataset.

	15-5(2 tasks)			**15-1**(6 tasks)			**19-1**(2 tasks)		
	old	new	all	old	new	all	old	new	all
Baseline	77.10	48.83	70.37	72.15	24.41	60.79	76.62	10.43	73.47
Ours	77.72	51.55	71.49	77.02	36.04	67.26	77.42	46.31	75.94

Quantitative comparison of continual semantic segmentation task with the baseline on the PASCAL VOC 2012 dataset.

Conclusion

- In this work, we design ACCL for continual classification tasks, integrating the KDC to **investigate the inner-task relations** when decoupling task-invariant and task-variant features. By incorporating an orthogonal constraint and an adversarial training strategy, our method enables the effective extraction of **well-distributed task-invariant and class-specific features within a single task**.
- Furthermore, we thoroughly investigate cross-task relations between memory samples and current task samples, which **supplement the inner-task relation, resulting in the effectiveness of preventing forgetting problems**.
- Extensive comparisons and ablation studies validate the superiority and effectiveness of our proposed method.

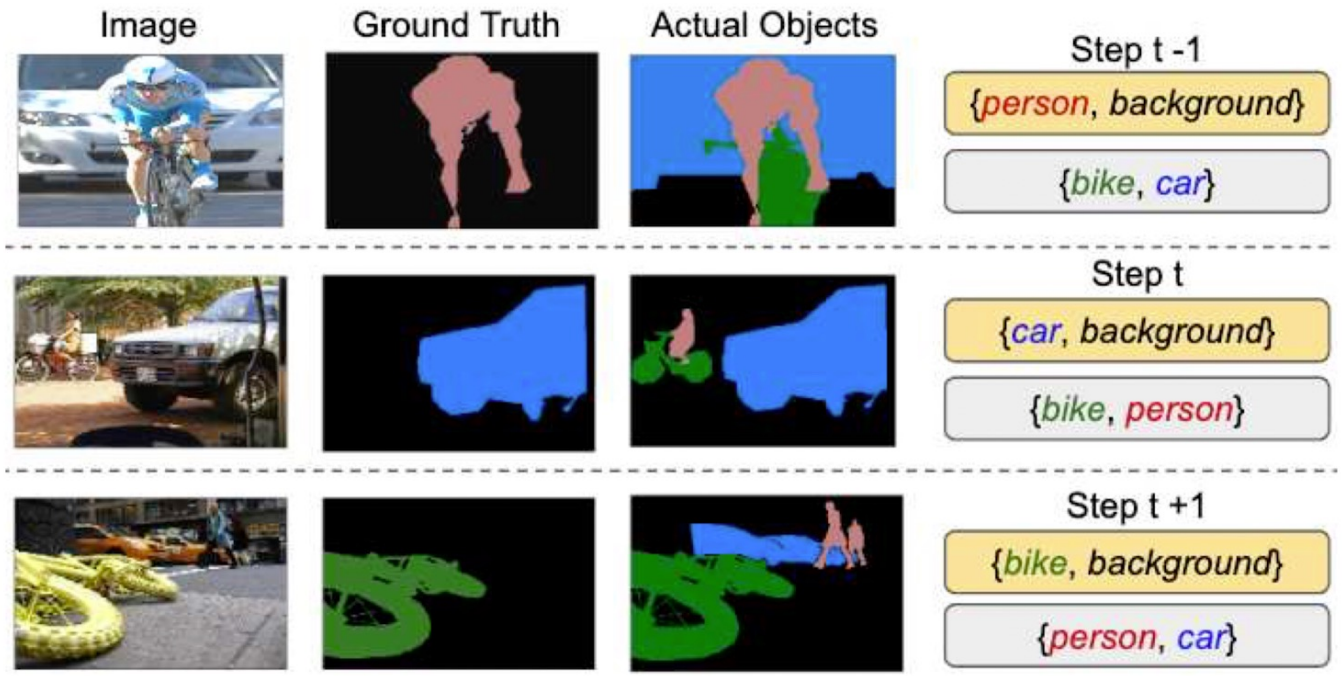
NeurIPS 2023 → TPAMI 2025

Saving 100x Storage: Prototype Replay for Reconstructing Training Sample Distribution in Class-Incremental Semantic Segmentation

Replay Without Saving: Prototype Derivation and Distribution Rebalance for Class-Incremental Semantic Segmentation

Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ip, and Sam Kwong

Task Definition



- In **class-incremental semantic segmentation (CISS)**, each step focuses on different classes, with its training set only annotates current classes, while previously learned classes and future classes are labeled as *background*.
- The images in each single-step training set contain at least one pixel from current classes, and images devoid of any current class are excluded.

Motivation

- There are **a lot of false positives for classes in incremental steps** (i.e., steps beyond the first).
- This is because **the proportion of current classes in the single-step training set is significantly higher than in the complete dataset, leading to classification bias**, which is especially pronounced in incremental steps with fewer classes.

To address this issue, the key is to augment past classes and background pixels in the training samples of the incremental steps, thereby reducing the proportion of the current class. At the same time, it is important to avoid triggering excessive storage requirements.

Solution

Prototype Replay	Background Repetition
At each task, the pixel occurrence count for each class is recorded. In subsequent tasks, pixel-level class prototypes are replayed based on these occurrence counts.	At each task, the cumulative pixel count of the background class is updated. In subsequent tasks, background features are duplicated according to this count.

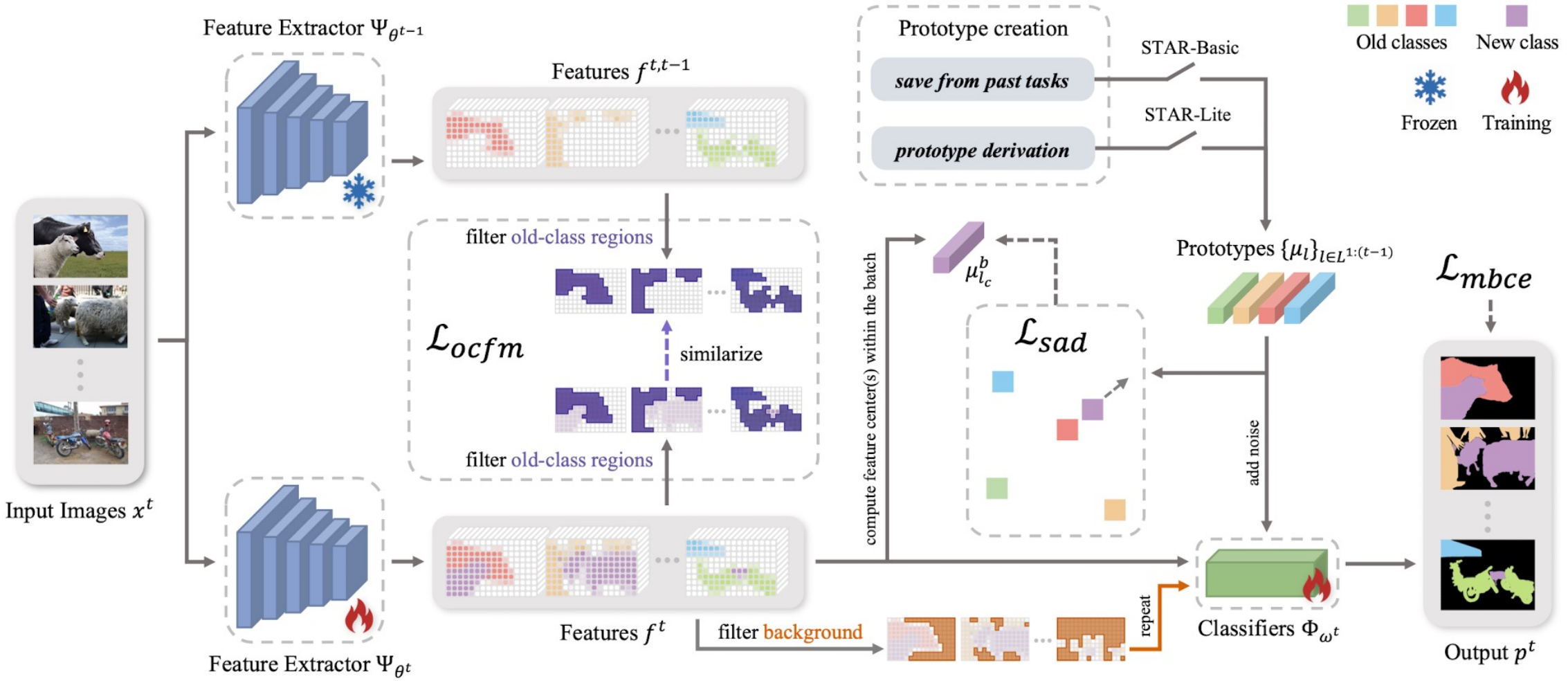
These two strategies respectively adjust the proportion of foreground and background classes within the single-step training samples to match the proportion in the "cumulative training set up to the current step", thus avoiding bias.

Contributions

- We propose a new CISS method named **STAR**. Its **basic version** stores compact prototypes and necessary statistics for each learned class. This enables a **comprehensive reconstruction of single-task training sample distributions**, aligning them with the complete dataset to mitigate classification bias.
- We develop a **prototype derivation method** that considers both the recognition and extraction patterns of the network. This empowers **prototype creation without the need for storage**, leading to a **lite version**.
- The **OCFM loss** is introduced to **retain learned knowledge in a spatially targeted manner**, maintaining old-class features while ensuring flexibility for learning new classes. Additionally, the **SAD loss** is designed to enhance the feature discriminability between similar old-new class pairs, facilitating the classification.

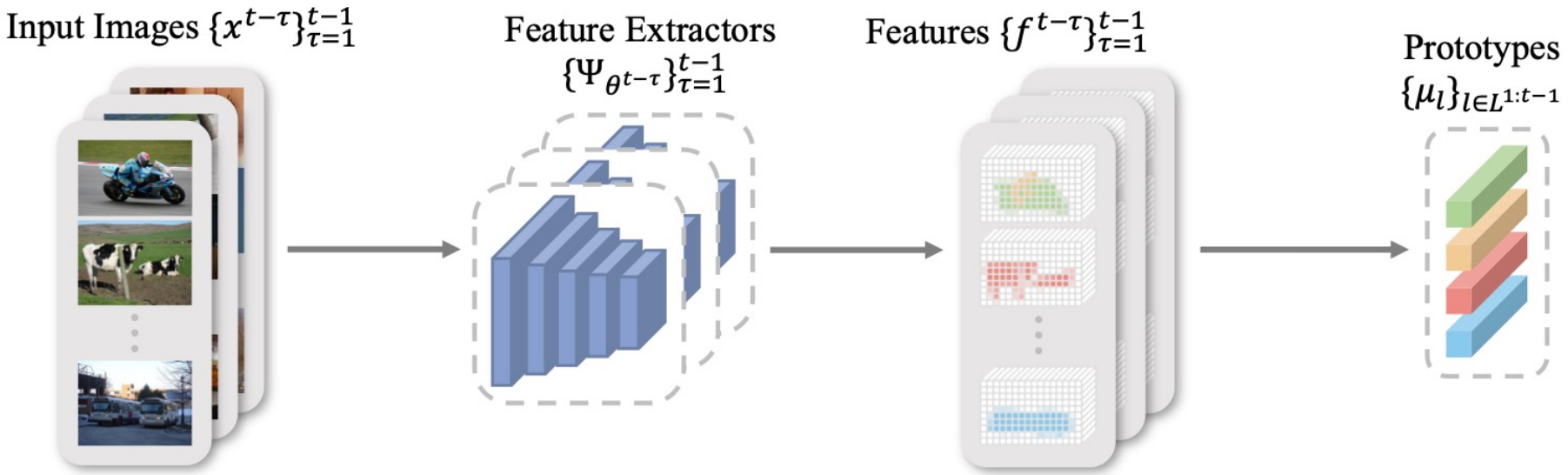
Our STAR Method

https://github.com/jinpeng0528/STAR



Prototype Replay – Basic Version

- After the learning of each task, all training samples in this task are passed through the frozen model to compute the features.
- The feature centers are stored as **prototypes** and replayed in subsequent tasks.
- Since prototypes are highly compact, they require only 1/100 of the storage compared to existing replay-based methods that storing raw images.



Prototype Replay – Lite Version

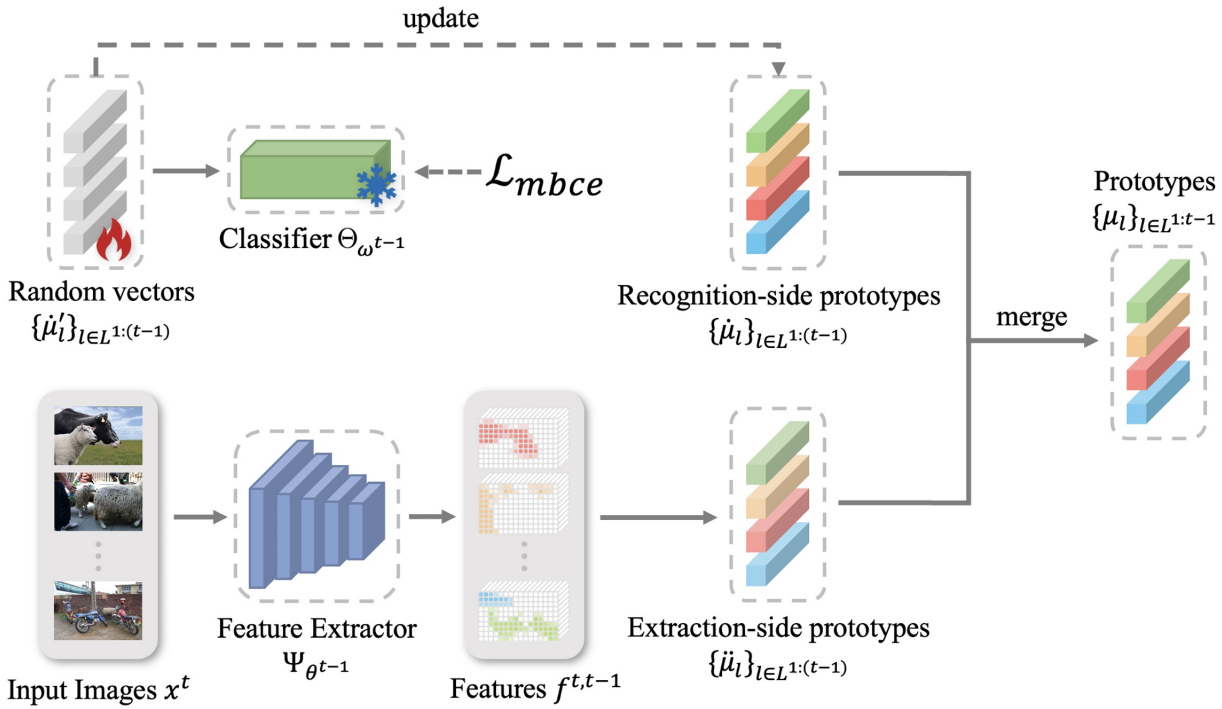
By leveraging the network's classification recognition and feature extraction patterns, prototypes are derived without the need for any storage.

Recognition Patterns

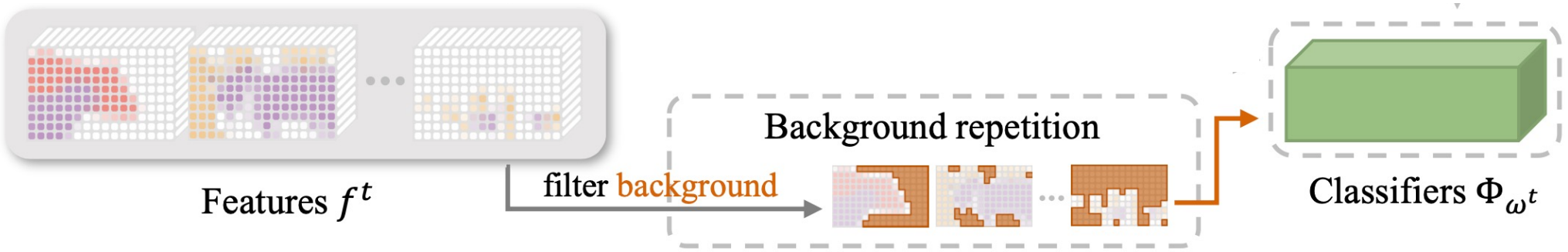
The classifier is isolated from the network. Then, it is used to infer representative features of previous classes, forming the **recognition-side prototype**.

Extraction Patterns

Images from the current task are fed into the network, and features from regions predicted as belonging to previous classes are aggregated to construct the **extraction-side prototypes**.



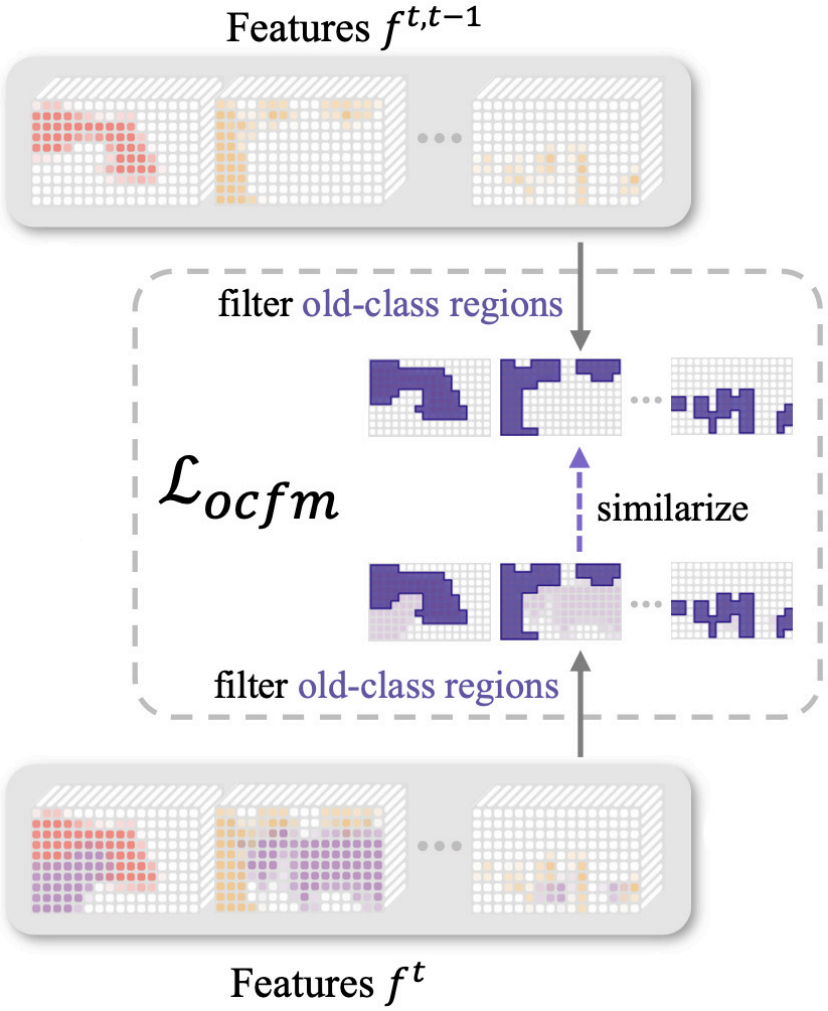
Background Repetition



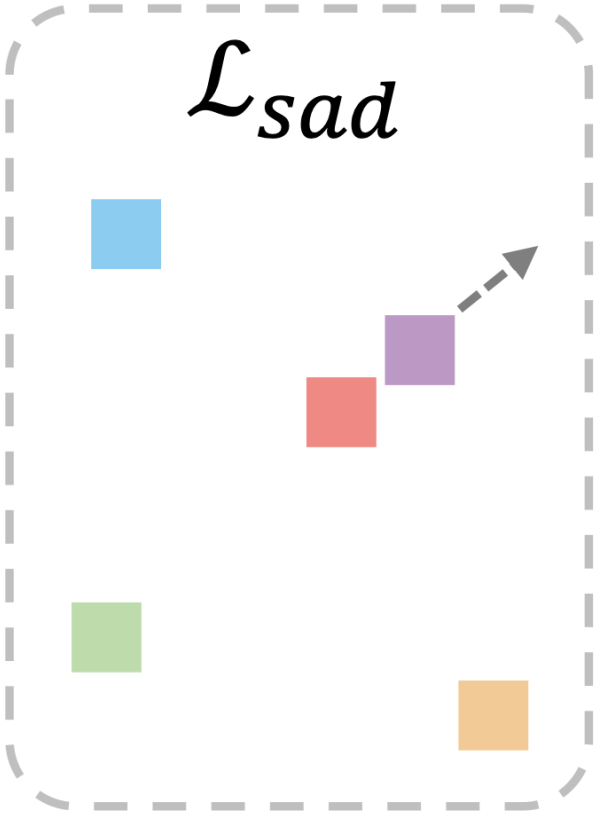
- Starting from the first step, we record and update the cumulative occurrence count of background pixels, η_{bg}.
- In subsequent steps, the background regions of input images are filtered out using current annotations and predictions from the previous model.
- **The features of these background regions are repeated multiple times and fed into the classifiers to add η_{bg} extra background pixels**, thus aligning the proportion of background in the single-step training samples with that of the ”cumulative training set up to the current step”.

Old-Class Feature Maintaining Loss

- A crucial prerequisite for effective prototype replay is the relative stability of old-class feature space.
- The old-class feature-maintaining loss utilizes current labels and predictions from the previous model to **locate old-class regions**. Within these regions, it **constrains the features extracted by the current model to be close to those extracted by the previous model**.



Similarity-Aware Discriminative Loss



- Some similar "new-old class pairs" are prone to confusion because they appear in different steps, making it challenging for the feature extractor to generate discriminative features.
- The most direct approach is to penalize the similarity of all "new-old class pairs" feature centers, increasing their distance.
- However, this method may lead to resource waste as some "new-old class pairs" are inherently dissimilar. Therefore, we **penalize the similarity between each new class feature center and its closest old class feature center**, focusing on the most challenging points.

Experiments

Method	19-1						15-5						15-1					
	Disjoint			Overlapped			Disjoint			Overlapped			Disjoint			Overlapped		
	base	*inc.*	*all*	*base*	*inc.*	*all*	*base*	*inc.*	*all*	*base*	*inc.*	*all*	*base*	*inc.*	*all*	*base*	*inc.*	*all*
MiB [10]	69.6	25.6	67.4	70.2	22.1	67.8	71.8	43.3	64.7	75.5	49.4	69.0	46.2	12.9	37.9	35.1	13.5	29.7
SDR [14]	69.9	37.3	68.4	69.1	32.6	67.4	73.5	47.3	67.2	75.4	52.6	69.9	59.2	12.9	48.1	44.7	21.8	39.2
PLOP [12]	75.1	38.2	73.2	75.4	37.4	73.5	66.5	39.6	59.8	75.7	51.7	70.1	49.0	13.8	40.2	65.7	17.3	54.2
SSUL [11]	77.4	22.4	74.8	77.7	29.7	75.4	76.4	45.6	69.1	77.8	50.1	71.2	74.0	32.2	64.0	77.3	36.6	67.6
STCISS [55]	76.6	36.0	75.4	76.1	43.4	74.5	76.9	54.3	71.3	76.7	54.3	71.1	70.1	34.3	61.2	71.4	40.0	63.6
RBC [58]	76.4	45.8	75.0	77.3	55.6	76.2	75.1	49.7	69.9	76.6	52.8	70.9	61.7	19.5	51.6	69.5	38.4	62.1
DKD [9]	77.4	43.6	75.8	77.8	41.5	76.0	77.6	54.1	72.0	78.8	58.2	73.9	76.3	39.4	67.5	78.2	44.3	70.1
UCD [56]	75.7	31.8	73.5	75.9	39.5	74.0	67.0	39.3	60.1	75.0	51.8	69.2	50.8	13.3	41.4	66.3	21.6	55.1
EWF [57]	78.2	3.2	74.6	77.9	6.7	74.5	79.3	38.2	69.5	79.4	38.2	69.5	75.3	22.5	62.7	78.5	31.6	67.3
STAR-Lite	77.9	46.4	**76.4**	78.1	49.1	**76.8**	78.5	58.3	**73.7**	79.7	59.4	**74.8**	78.5	45.9	**70.8**	80.0	51.2	**73.1**
RECALL [59]	65.0	47.1	65.4	68.1	55.3	68.6	69.2	52.9	66.3	67.7	54.3	65.6	67.6	49.2	64.3	67.8	50.9	64.8
PLOPLong [60]	-	-	-	74.8	39.7	73.1	-	-	-	76.0	48.3	69.4	-	-	-	72.0	26.7	61.2
SSUL-M [11]	77.6	43.9	76.0	77.8	49.8	76.5	76.5	48.6	69.8	78.4	55.8	73.0	76.5	43.4	68.6	78.4	49.0	71.4
DKD-M [9]	77.6	56.9	76.6	78.0	57.7	**77.0**	77.7	55.4	72.4	79.1	60.6	74.7	77.3	48.2	70.3	78.8	52.4	72.5
STAR-Basic	78.0	47.5	76.5	78.2	48.5	76.8	78.5	57.9	73.6	79.7	59.6	74.9	78.1	48.2	71.0	79.8	51.6	73.1
STAR-Basic†	77.9	53.6	**76.7**	78.1	56.3	**77.0**	78.6	58.4	**73.8**	80.1	62.2	**75.8**	77.8	50.4	**71.3**	79.8	55.5	**74.0**

Pascal VOC 2012 Dataset - 1

Method	10-1			5-3		
	base	*inc.*	*all*	*base*	*inc.*	*all*
MiB [10]	12.3	13.1	12.7	57.1	42.6	46.7
PLOP [12]	44.0	15.5	30.5	17.5	19.2	18.7
SSUL [11]	71.3	46.0	59.3	72.4	50.7	56.9
DKD [9]	73.1	46.5	60.4	69.6	53.5	58.1
EWF [57]	71.5	30.3	51.9	61.7	42.2	47.7
STAR-Lite	74.0	53.5	**64.3**	72.1	59.6	**63.2**
SSUL-M [11]	74.0	53.2	64.1	71.3	53.2	58.4
DKD-M [9]	74.0	56.7	65.8	69.8	60.2	62.9
STAR-Basic	72.6	55.4	64.4	70.7	61.8	64.3
STAR-Basic†	74.4	56.9	**66.1**	72.4	63.3	**65.9**

Pascal VOC 2012 Dataset - 2

Method	13-6			13-1		
	base	*inc.*	*all*	*base*	*inc.*	*all*
MiB [10]	52.8	17.9	41.8	51.6	22.9	42.5
PLOP [12]	53.2	10.1	39.6	52.4	15.1	40.6
DKD [9]	55.5	36.4	49.8	55.7	20.9	46.5
UCD [56]	53.0	18.6	42.1	52.2	23.4	43.1
STAR-Lite	56.6	50.5	54.7	55.7	31.2	**48.3**
STAR-Basic	56.4	50.9	**54.8**	55.7	31.1	**48.3**

CityScapes Dataset

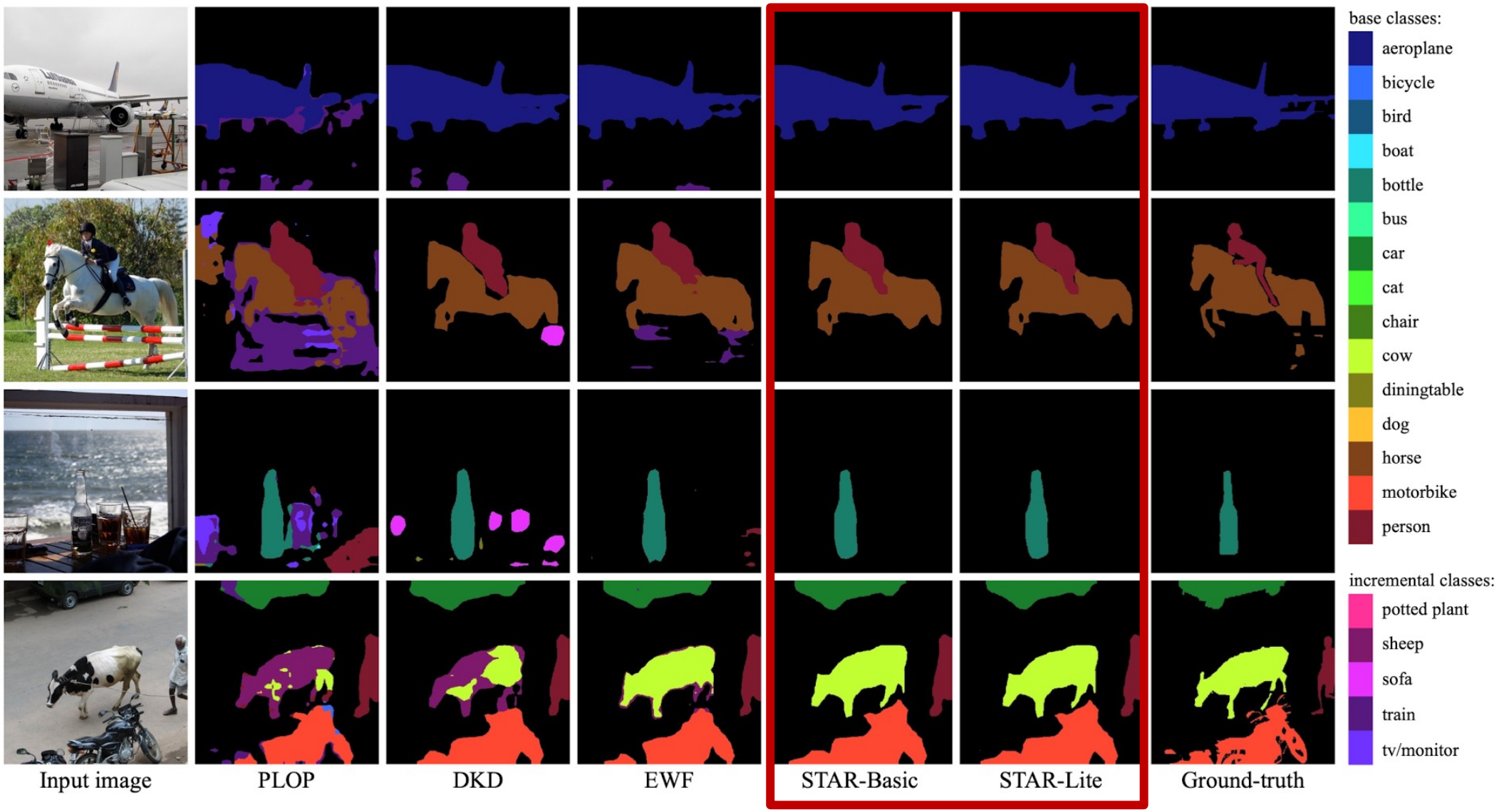
Method	100-50			100-10			50-50		
	base	*inc.*	*all*	*base*	*inc.*	*all*	*base*	*inc.*	*all*
MiB [10]	40.5	17.2	32.8	38.2	11.1	29.2	45.6	21.0	29.3
PLOP [12]	41.9	14.9	32.9	40.5	13.6	31.6	48.8	21.0	30.4
SSUL [11]	41.3	18.0	33.6	40.2	18.8	33.1	48.4	20.2	29.6
RCIL [54]	42.3	18.8	34.5	39.3	17.6	32.1	48.3	25.0	32.5
STCISS [55]	40.7	24.0	35.1	33.6	16.9	28.1	40.0	23.6	29.0
RBC [58]	42.9	21.5	35.8	39.0	21.7	33.3	49.6	26.3	34.2
DKD [9]	42.4	22.9	36.0	41.5	19.4	34.2	48.8	26.3	33.9
EWF [57]	41.2	21.3	34.6	41.5	16.3	33.2	46.1	19.8	28.5
STAR-Lite	42.4	24.3	**36.4**	42.0	20.4	**34.9**	48.7	26.9	**34.3**
PLOPLong [60]	41.9	14.9	32.9	40.5	13.6	31.6	48.8	21.0	30.4
SSUL-M [11]	42.8	17.5	34.4	42.9	17.7	34.5	49.1	20.1	29.8
DKD-M [9]	42.4	23.0	36.0	41.7	20.1	34.6	48.8	26.3	33.9
STAR-Basic	42.4	24.3	**36.4**	41.8	20.7	**34.8**	48.3	27.0	**34.2**

ADE20K Dataset

STAR-Basic: Save 100x Storage Cost

STAR-Lite: Replay Without Any Storage

Experiments



Ablation Study

	PR	BPR	OCFM	SAD	disjoint 15-1			overlapped 15-1		
					base	*inc.*	*all*	*base*	*inc.*	*all*
STAR-Lite	×	✓	✓	✓	77.9	41.8	69.3	79.1	40.4	69.9
	✓	×	✓	✓	76.4	30.8	65.5	78.1	32.3	67.2
	✓	✓	×	✓	76.9	37.7	67.6	79.5	45.5	71.4
	✓	✓	✓	×	75.7	33.9	65.7	79.5	47.4	71.9
	✓	✓	✓	✓	78.5	45.9	**70.8**	80.0	51.2	**73.1**
STAR-Basic	×	✓	✓	✓	77.9	41.8	69.3	79.1	40.4	69.9
	✓	×	✓	✓	77.5	33.6	67.0	78.6	36.0	68.5
	✓	✓	×	✓	76.8	36.3	67.1	79.3	46.1	71.4
	✓	✓	✓	×	75.8	34.8	66.0	79.6	48.9	72.3
	✓	✓	✓	✓	78.1	48.2	**71.0**	79.8	51.6	**73.1**

Ablation Study Results

PR: Prototype Replay

BPR: Background Pixel Repetition

OCFM: Old-Class Features Maintaining Loss

SAD: Similarity-Aware Discriminative Loss

Conclusion

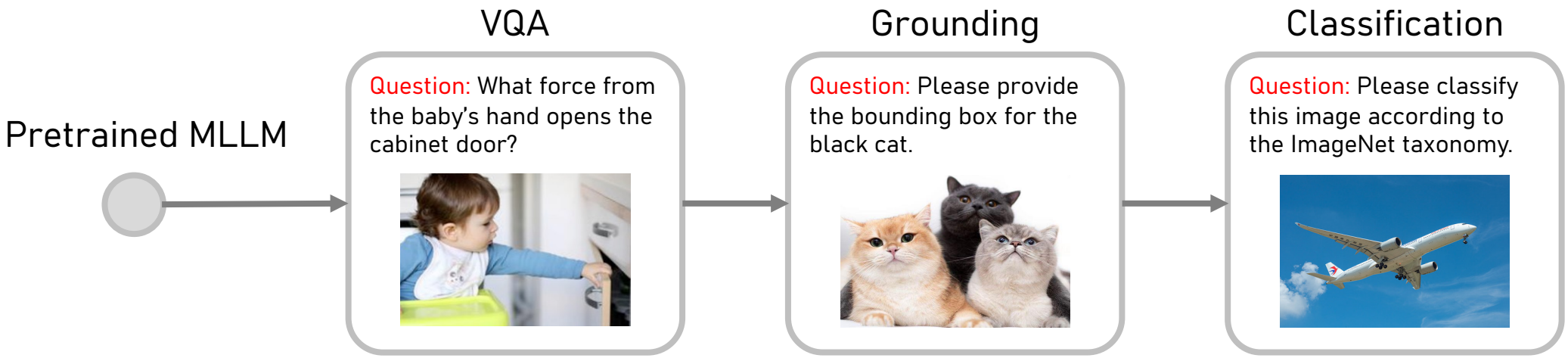
- This paper introduces **STAR**, a CISS method designed to **mitigate classification bias** arising from distribution variances between single-task training sets and the complete dataset.
- STAR employs two principal tactics: **prototype replay** and **background pixel repetition**. The former rectifies the distribution of foreground classes by replaying old-class prototypes, while the latter reintegrates missing background pixels by duplicating background pixels.
- Regarding the creation of prototypes, STAR diverges into two variants. **STAR-Basic** stores prototypes after learning each task for future replay, whereas **STAR-Lite** employs a novel prototype derivation method that considers the network's recognition and extraction patterns to deduce prototypes.
- The **OCFM loss** is introduced to maintain the features of old classes, ensuring the model's ability to learn new classes without losing prior knowledge. Additionally, the **SAD loss** is proposed to enhance feature differentiation between similar old and new class pairs, improving their distinguishability for the classifiers.

ICML 2025

SEFE: Superficial and Essential Forgetting Eliminator for Multimodal Continual Instruction Tuning

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ho Shing Ip, and Sam Kwong

Introduction



- In **Multimodal Continual Instruction Tuning (MCIT)**, a pretrained Multimodal Large Language Model (MLLM) is sequentially tuned on a series of multimodal tasks, aiming to learn new tasks while minimizing forgetting of previously learned ones.

Introduction

Does the forgetting problem become more severe or alleviated for large and small models under continual learning architectures?

Introduction

Does the forgetting problem become more severe or alleviated for large and small models under continual learning architectures?

Which continent is highlighted?
A. Africa
B. North America
C. South America
D. Asia
Answer with the option's letter from the given choices directly.

A — No forgetting (Just after learning this task)

learn other tasks

Africa — Superficial forgetting

[0.0, 0.36, 0.29, 0.6] — Superficial forgetting

C — Essential forgetting

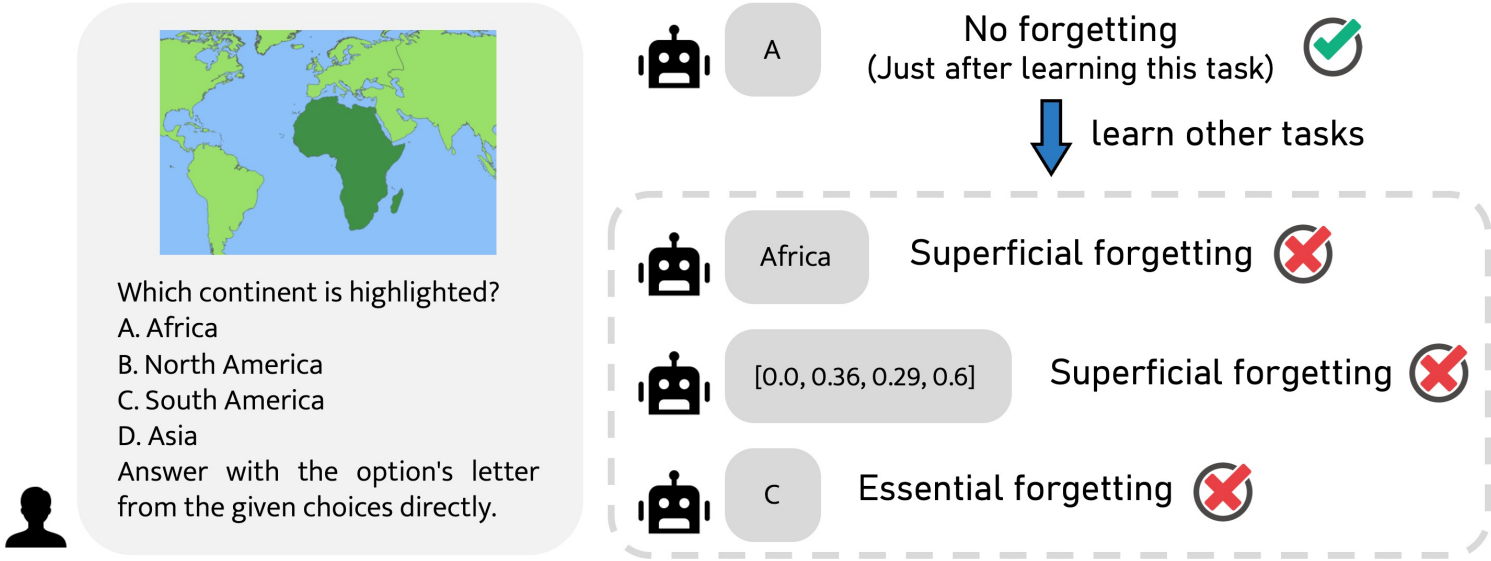
Contributions

a) We formally define *superficial forgetting* and *essential forgetting* in MCIT. Furthermore, our proposed method, SEFE, addresses these challenges and achieves state-of-the-art performance.

b) To mitigate *superficial forgetting*, we introduce the **Answer Style Diversification (ASD)** paradigm that unifies the answer domain across tasks by rephrasing questions, thereby reducing the model's bias toward specific response styles. Additionally, we create **CoIN-ASD**, an ASD-adjusted version of the CoIN benchmark, which can serve as a new benchmark for evaluating *essential forgetting* in future MCIT studies.

c) To address *essential forgetting*, we present **RegLoRA**. By identifying critical elements in the weight update matrices and applying regularization constraints, RegLoRA ensures that LoRA fine-tuning does not disrupt the model's existing knowledge.

Forgetting Types

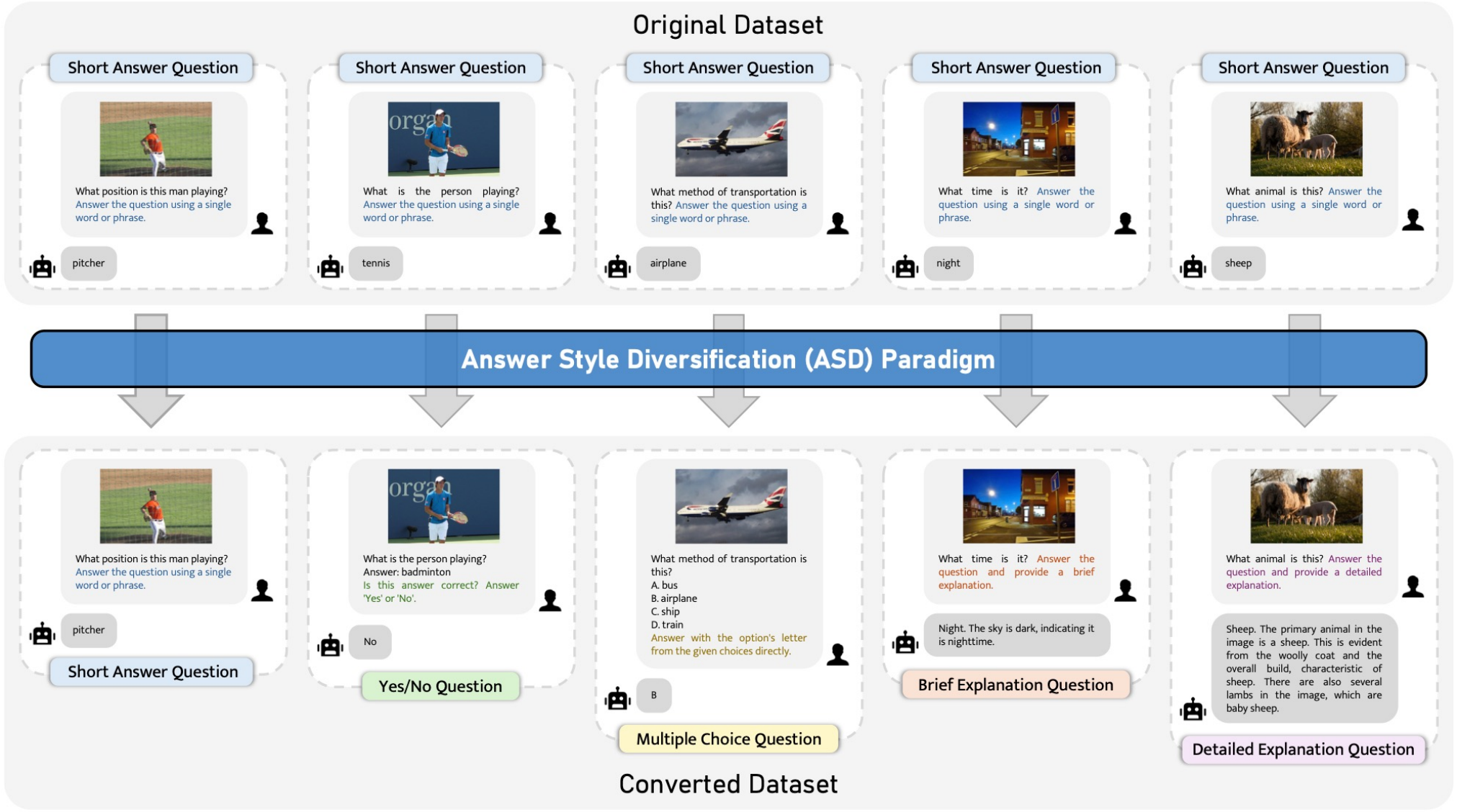


- **Superficial Forgetting:** task knowledge may be retained while the response style is forgotten.
- **Essential Forgetting:** task knowledge is forgotten.

Answer Style Diversification

- *Superficial forgetting* arises from the gap in answer space between tasks, as the model tends to respond in the answer style of the most recently learned task.
- To address this issue, the **Answer Style Diversification (ASD) paradigm** reformulate questions in each task into five unified formats, aligning the answer space across tasks.
- These five formats include Short Answer Question, Yes/No Question, Multiple Choice Question, Brief Explanation Question, and Detailed Explanation Question. After analyzing 15 mainstream benchmarks, we find that these formats sufficiently cover the requirements of all tasks.

Answer Style Diversification



Answer Style Diversification

Method	Accuracy on Each Task (%)								Aggregate Results (%)			
	SQA	*VQA*^{Text}	*ImgNet*	*GQA*	*VizWiz*	*Grd*	*VQA*^{v2}	*VQA*^{OCR}	MFT↑	MFN↑	MAA↑	BWT↓
FFT	2.95	36.38	52.35	46.40	33.90	0.00	61.65	50.00	65.87	35.45	36.73	-30.42
LoRA [20]	54.05	44.63	41.25	47.55	20.80	0.85	59.30	64.30	**70.21**	41.59	39.53	-28.62
O-LoRA [45]	75.40	52.89	71.85	47.30	37.35	7.10	61.85	61.20	69.30	51.87	49.56	-17.43
LoTA [38]	67.30	41.51	8.25	37.15	42.25	0.10	47.95	56.15	54.72	37.58	50.46	-17.14
FFT+ASD	74.50	50.12	65.40	54.35	45.50	0.00	64.40	68.50	68.28	52.85	57.18	-15.44
LoRA+ASD [20]	74.45	49.70	39.30	52.00	50.45	7.05	62.25	47.80	68.13	47.88	59.71	-20.26
O-LoRA+ASD [45]	75.20	55.36	67.50	54.70	52.90	15.40	64.45	35.05	65.59	52.57	61.63	-13.02
LoTA+ASD [38]	76.90	42.65	15.85	40.25	45.10	0.30	54.35	54.00	56.99	41.18	56.28	-15.82

MFT: Mean Fine-tune Accuracy

MFN: Mean Final Accuracy

MAA: Mean Average Accuracy

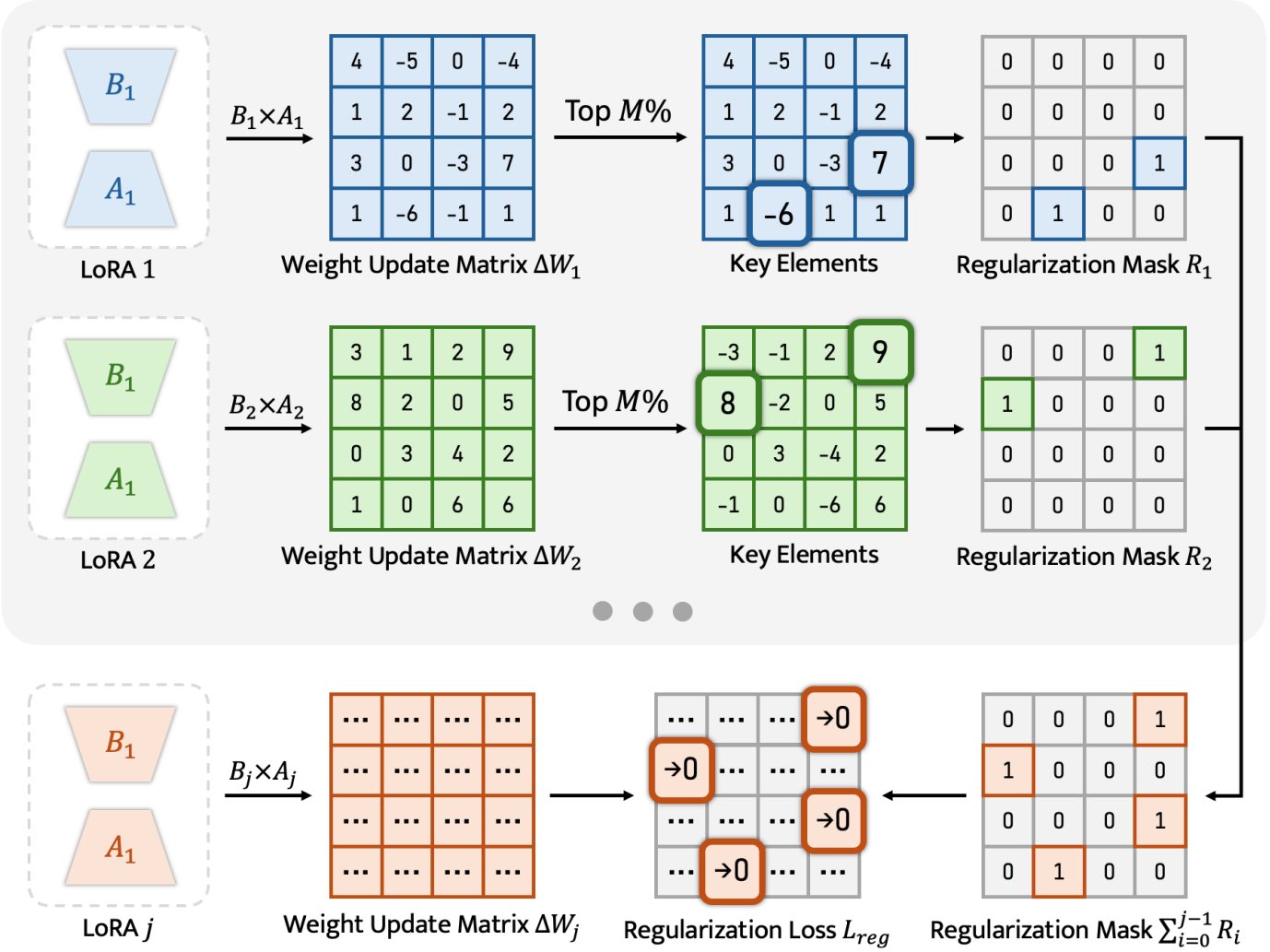
BWT: Backward Transfer

By adding ASD to existing methods, MFN, MAA, and BWT achieve average improvements of 7.00%, 14.63%, and 7.27%, respectively.

RegLoRA

- Although *superficial forgetting* is alleviated by ASD, *essential forgetting*—the true loss of past knowledge—still remains.
- Experiments reveal that only a small subset of parameters change significantly during task learning. These key parameters likely carry most of the task-specific knowledge.
- Therefore, we propose **RegLoRA**, which constrains updates to parameters significantly changed during previous tasks, thereby preserving knowledge of earlier tasks.

RegLoRA



- After each task, a **regularization mask** is saved to identify important elements for that task.
- During future training, updates to all previously identified elements are constrained.

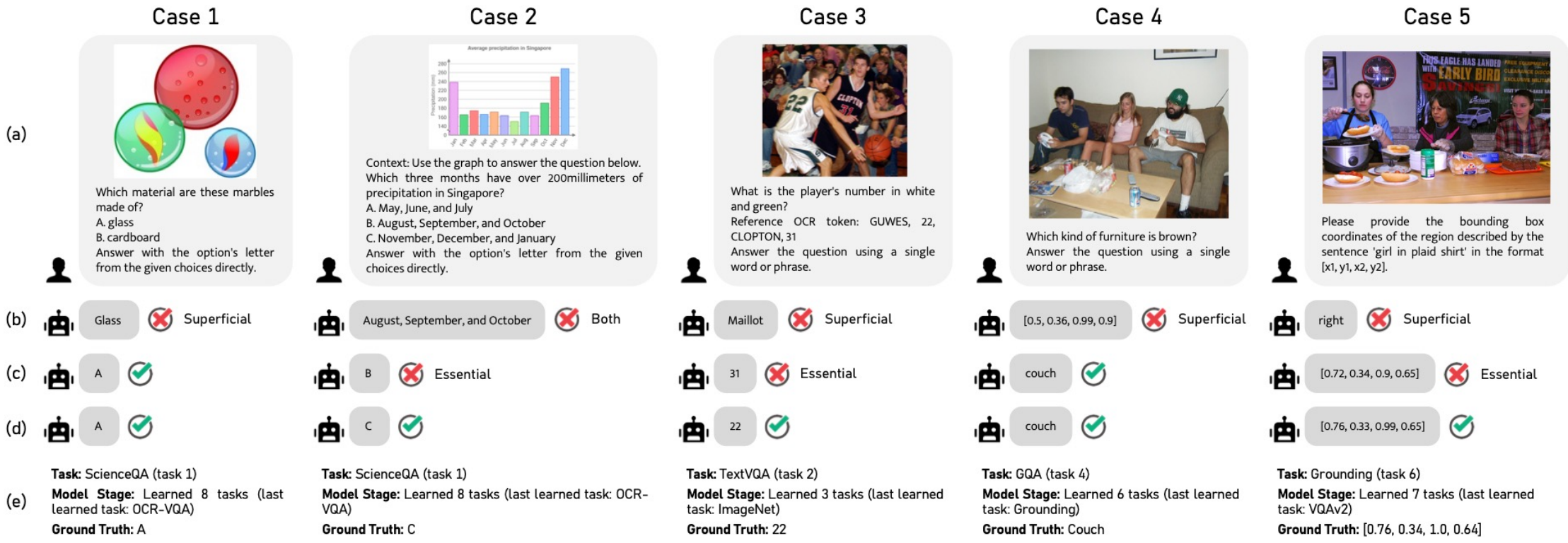
RegLoRA

Configuration	Aggregate Results (%)			
	MFT↑	MFN↑	MAA↑	BWT↑
Baseline (LoRA)	**70.21**	41.59	39.53	-28.62
+ ASD	68.13	<u>47.88</u>	<u>59.71</u>	<u>-20.26</u>
+ ASD + RegLoRA	<u>69.02</u>	**58.57**	**63.04**	**-10.45**

Quantitative Comparison

Method	Accuracy on Each Task (%)								Aggregate Results (%)			
	SQA	*VQA*^{Text}	*ImgNet*	*GQA*	*VizWiz*	*Grd*	*VQA*^{v2}	*VQA*^{OCR}	MFT↑	MFN↑	MAA↑	BWT↓
FFT	2.95	36.38	52.35	46.40	33.90	0.00	61.65	50.00	65.87	35.45	36.73	-30.42
LoRA [20]	54.05	44.63	41.25	47.55	20.80	0.85	59.30	64.30	**70.21**	41.59	39.53	-28.62
O-LoRA [45]	75.40	52.89	71.85	47.30	37.35	7.10	61.85	61.20	69.30	51.87	49.56	-17.43
LoTA [38]	67.30	41.51	8.25	37.15	42.25	0.10	47.95	56.15	54.72	37.58	50.46	-17.14
FFT+ASD	74.50	50.12	65.40	54.35	45.50	0.00	64.40	68.50	68.28	52.85	57.18	-15.44
LoRA+ASD [20]	74.45	49.70	39.30	52.00	50.45	7.05	62.25	47.80	68.13	47.88	59.71	-20.26
O-LoRA+ASD [45]	75.20	55.36	67.50	54.70	52.90	15.40	64.45	35.05	65.59	52.57	61.63	-13.02
LoTA+ASD [38]	76.90	42.65	15.85	40.25	45.10	0.30	54.35	54.00	56.99	41.18	56.28	-15.82
SEFE (Ours)	75.35	58.66	83.10	54.25	48.85	16.75	65.35	66.25	69.02	**58.57**	**63.04**	**-10.45**

Qualitative Comparison



(a) Instruction; (b) Response from the baseline model; (c) Response from the baseline model with ASD added; (d) Response from the baseline model with both ASD and RegLoRA added; (e) Basic information of the case.

Conclusion

- This paper identifies two forgetting types in MCIT—superficial forgetting, where the model's response style becomes biased, and essential forgetting, where factual knowledge is lost.
- To address these issues, we propose the SEFE method, which includes two components: the ASD paradigm and RegLoRA. ASD mitigates superficial forgetting by diversifying question types within tasks, improving response style robustness and knowledge assessment. RegLoRA combats essential forgetting by identifying and regularizing critical weight components across LoRAs to preserve knowledge.
- Experiments demonstrate that both ASD and RegLoRA are effective in tackling their respective forgetting types, and together in SEFE, they achieve state-of-the-art performance in mitigating catastrophic forgetting in MCIT.

山东大学
SHANDONG UNIVERSITY

THANKS FOR WATCHING


李华 副教授


熊航 硕士

陈锦芃 博士

罗宇轩 博士


Sam Kwong 教授（加拿大工程院院士）

学无止境 气有浩然