Deconfound Semantic Shift and Incompleteness in Incremental Few-shot Semantic Segmentation

Yirui Wu[1, 5], Yuhang Xia[1], Hao Li[1], Lixin Yuan[1], Junyang Chen[2*], Jun Liu[3], Tong Lu[4], Shaohua Wan[5]

[1]Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing, China
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[3]School of Computing and Communication, Lancaster University, Lancaster, UK
[4]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China
[5]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China
{wuyirui, xiayuhang, lihao, yuanlixin}@hhu.edu.cn, junyangchen@szu.edu.cn, j.liu81@lancaster.ac.uk, lutong@nju.edu.cn, shaohua.wan@ieee.org

Abstract

Incremental few-shot semantic segmentation (IFSS) expands segmentation capacity of the trained model to segment new-class images with few samples. However, semantic meanings may shift from background to object class or vice versa during incremental learning. Moreover, new-class samples often lack representative attribute features when the new class greatly differs from the pre-learned old class. In this paper, we propose a causal framework to discuss the cause of semantic shift and incompleteness in IFSS, and we deconfound the revealed causal effects from two aspects. First, we propose a Causal Intervention Module (CIM) to resist semantic shift. CIM progressively and adaptively updates prototypes of old class, and removes the confounder in an intervention manner. Second, a Prototype Refinement Module (PRM) is proposed to complete the missing semantics. In PRM, knowledge gained from the episode learning scheme assists in fusing features of new-class and old-class prototypes. Experiments on both PASCAL-VOC 2012 and ADE20k benchmarks demonstrate the outstanding performance of our method.

Introduction

The rise of pixel-level annotations in semantic segmentation has spurred the need for methods to incrementally expand model's capacity to learn new classes without retraining the model. Incremental few-shot semantic segmentation (IFSS) can continuously segment new classes with scarce incremental data while retaining to segment previously learned classes. (Cermelli et al. 2020b, 2021; Shi et al. 2022).

The main challenges of IFSS emerge with two aspects: semantic shift and semantic incompleteness. As shown in Fig. 1(a), the semantic shift is inherited from incremental semantic segmentation (ISS), where background classes from previous learning steps may shift to object classes at the current step or vice versa (Cermelli et al. 2020a; Douillard et al. 2021). Scarcity of new information and non-access to old information exacerbate shifts in IFSS, causing cognitive confusion about the model of old knowledge and exacerbating catastrophic forgetting. In Fig. 1(b), samples deviating from their ground-truth often lack some representative attribute features, especially when new classes with several samples are far away from compact clusters of old classes trained in previous steps, resulting in semantic incompleteness.

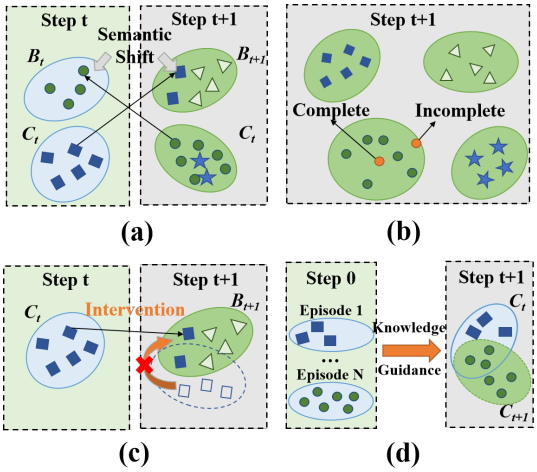

Figure 1: (a) In IFSS, old classes C_t may shift to background class B_{t+1} at current step or vice versa. (b) New-class samples might be far from the ground-truth center, due to semantic incompleteness caused by small and imbalanced new-class dataset. (c) We remove the confounding effect of semantic shift through intervention, obtaining unbiased distribution estimation of old classes. (d) To complete semantics, we fuse features of old and new classes with guidance of knowledge extracted from episode learning scheme.

Most previous incremental methods address semantic shift and representative features of new classes by using feature/label distillation with knowledge updating schemes (Cermelli et al. 2020a; Zhang et al. 2022; Phan et al. 2022), thus regulating new models to be predicable for either old or new classes. However, they do not always behave consistently in IFSS because they couple causal factors, i.e., data,

*Corresponding author.

features, knowledge, and labels within incremental learning. This inevitably hinders causality exploration among factors for model bias adjustment, especially when dealing with imbalanced new-class data in IFSS.

To avoid the coupling, we address problems of semantic shift and incompleteness with a causal inference framework that we explain and decouple causal relations among data, features, knowledge, and labels using logical graphs and formulas. Under the causal framework, we can answer that 1) the confounder from background to old classes and prediction cause semantic shift. As shown in Fig. 2(a), the confounder is defined as a common cause of other variables, rendering spurious correlations among them, even if they have no direct causal effects with each other (Yue et al. 2020). And 2) the difficulties to build convinced relationship between new classes and prediction result in semantic incompleteness (see Fig. 2(a)). Considering the extremely small sample set, it is reasonable to build knowledge-driven mappings rather than data-driven ones.

To this end, we propose to deconfound the causal effects that cause semantic shift and incompleteness in IFSS. Specifically, we remove the confounder of semantic shift with intervention, as shown in Fig. 1(c), where the causal effect between background and old class is removed to achieve an unbiased estimation of the old class. The reason to use intervention other than other deconfounders (Hu et al. 2021) is that intervention ranks higher with operative actions; the others would cause semantic bias by conditionally transferring shifted information from old to new class without considering the imbalanced setting. Following the intervention principle, we propose the Causal Intervention Module (CIM) as shown in Fig. 2(b), which involves operations of passively observing and actively adjusting. In each learning step, CIM observes the distribution of compact prototypes/clusters in pre-trained feature space and adaptively adjusts prototypes via a prototype-attention scheme. It indicates the directions toward a well-separated prototype distribution with less semantic bias.

In addition, knowledge proves to promote few-shot semantic segmentation by sharing semantic properties like "haired" and "quadruped" between old- and new-class samples. This inspires us to use knowledge for mitigation of semantic incompleteness. As shown in Fig. 1(d), a Prototype Refinement Module (PRM) is proposed to complete distinguish feature representations of new-class samples with knowledge guidance. PRM fills the missing attribute features of new-class samples via fusing features of old-class prototypes guided by knowledge, which is referred to as dependent relations from new class to old class and to prediction (see Fig. 2(d)). We extract such knowledge by the proposed episode learning scheme that creates variants of IFSS episodes by sampling from plenty of old-class prototypes to pretend "old" and "new" classes, learning meta knowledge about how to fuse features of new-class and old-class prototypes for semantic completion. To the best of our knowledge, our work is the first to introduce causal inference into IFSS, which explains the reason of semantic drift and incompleteness and offers causal based solutions to reasonably deconfound for unbiased predictions.

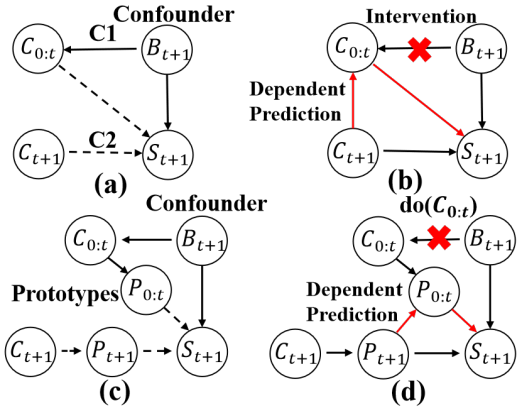

Figure 2: (a) The proposed structural causal model of IFSS to explain semantic shift (C1) and incompleteness (C2). (b) Removing the confounder of semantic shift with intervention and constructing dependent relations to complete semantics. (c) The prototype based structural causal model for IFSS. (d) Mitigating semantic shift and incompleteness through front door adjustment (do($C_{0:t}$)) and prototype based dependent prediction (red lines). Relevant notations are given in section *IFSS in Causal View*.

The contributions of this paper are as follows:

- We provide a causal analysis of semantic shift and incompleteness in IFSS, thus guaranteeing the superiority and reasonableness of our deconfounded method.
- We propose a causal intervention module to deconfond semantic shift, which uses the intervention operation to update prototypes of old class, thus mitigating the confounding bias.
- We propose a prototype refinement module to fill missing semantics, which guides feature fusion of new-class samples and old-class prototypes with the knowledge extracted from an episode learning scheme.

Related Work

IFSS

To retain learned knowledge, ISS methods (Douillard et al. 2021; Phan et al. 2022; Yan et al. 2021; Shang et al. 2023; Wang et al. 2024; Zhao, Yuan, and Shi 2023) utilize either knowledge distillation or regularization. MiB (Cermelli et al. 2020a) explicitly designs a novel distillation loss, reducing semantic biases during distillation. PLOP (Douillard et al. 2021) proposes a multi-scale distillation scheme, extracting inconsistent feature representations in scale to mitigate forgetting.

Existing methods design different learning paradigms for IFSS. PIFS (Cermelli et al. 2021) proposes prototype-based incremental few-Shot segmentation, coupling prototype learning and knowledge distillation. EHNet (Shi et al. 2022) represents class-based knowledge using category and hyper-class embedding. To address IFSS, they fail in providing a theoretical and causal solution by coupling data, features, knowledge and labels for data-driven modeling.

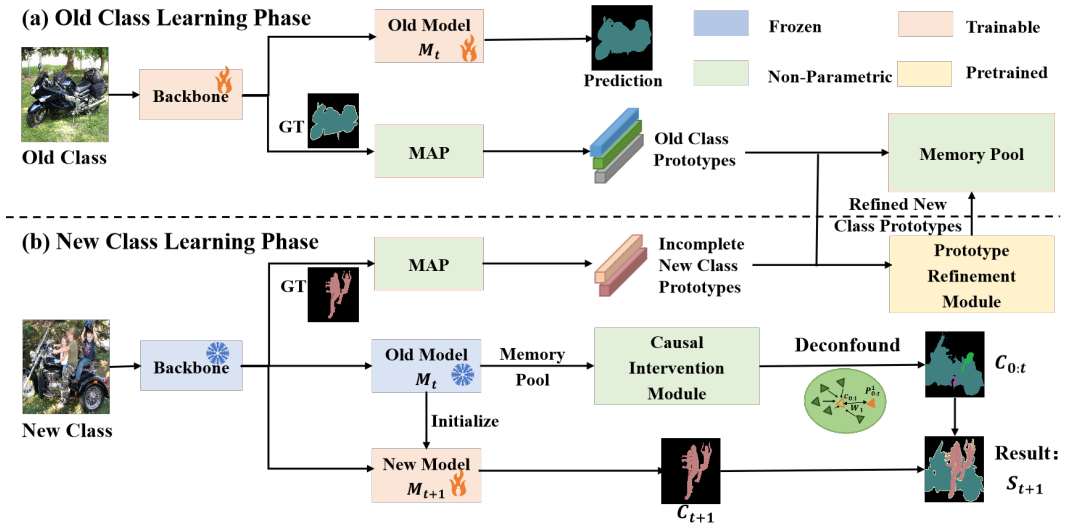

Figure 3: Method overview. (a) In old class learning phase, model learns to segment old classes and stores old-class prototypes in the memory pool. (b) During new class learning phase, model begins with PRM to refine incomplete new-class prototypes. Then, model deconfounds the confounder causing semantic shift by utilizing CIM after old model, obtaining unbiased old-class prototypes for unshifted predictions. We merge old-class and new-class predictions by NMS.

Causal Inference

Causal inference (Pearl, Glymour, and Jewell 2016; Pearl and Mackenzie 2018; Yang, Zhang, and Cai 2020) is widely used to mitigate spurious bias (Bareinboim and Pearl 2012; Liu et al. 2021) and disentangle effects(Besserve et al. 2020). They can be divided into intervention (Yue et al. 2020; Zhang et al. 2020a; Wang et al. 2021) and counterfactual (Niu et al. 2021; Yang et al. 2023). Intervention explicitly modifies SCM and eliminates spurious correlations by adjusting distribution of observed. Counterfactual extracts unbiased causal effects by observing the differences between the counterfactual/constrained and ground-truth SCM.

IFSS in Causal View

Task Definition

IFSS contains consecutive training steps, where step 0 requires quantity of samples to learn to segment old classes C_0. Subsequent steps from 1 to T gradually learn the new classes $C_1 \sim C_\mathrm{T}$, where T is the total number of learning steps. For t-th learning step ($t>0$), C_t contains k-shot samples (k is usually 1 or 5). The model learns to segment object classes $\{C_{0:t-1}, C_t\}$, and the background class B_t, and finally merges them as the prediction results S_t. Note that the object categories of different learning steps do not overlap, that is, $\forall i, j$ and $i \neq j$, $C_i \cap C_j = \emptyset$. After the t-th learning step, the model is tested on all object classes $C_{0:t}$.

Causal Effect Analysis

We use the structural causal model (SCM) (Pearl 2009) to explain the confound of semantic shift and semantic incompleteness. At $t+1$, we examine the causal relations of key variables and construct SCM, which include new classes C_{t+1}, old classes $C_{0:t}$, background classes B_{t+1} and prediction S_{t+1}. Thus, the relations in Fig. 2(a) can be formulated as $C_{0:t} \rightarrow S_{t+1}\&B_{t+1} \rightarrow S_{t+1}\&C_{t+1} \rightarrow S_{t+1}$. The prediction is obtained with $C_{0:t}$, B_{t+1} and C_{t+1} by

$$S_{t+1} = F(C_{0:t}, B_{t+1}, C_{t+1}), \tag{1}$$

where function $F(\cdot)$ denotes the merge operation (e.g., non-maximum suppression (NMS))(Neubeck and Gool 2006). Specifically, three relations about B_t, C_t, S_t are as follows. **1)** $B_{t+1} \rightarrow C_{0:t}$. In IFSS, the causal graph evolves to high complexity due to semantic shift. The old object class shifts to background in B_{t+1} interferes with the pre-learned old-class knowledge, possibly assign old-class pixels as background at current learning step. **2)** $C_{0:t} \dashrightarrow S_{t+1}$. B_{t+1} acts as a confounder between $C_{0:t}$ and S_{t+1}, which results in spurious correlations between old classes and prediction. **3)** $C_{t+1} \dashrightarrow S_{t+1}$. Due to the small new-class sample set, it is difficult to resolve a categorical mapping from C_{t+1} to S_{t+1} with unobserved and confusing factors. A possible solution to deconfound the above confounders represented as $\dashrightarrow$ is causal intervention, removing the correlations and building extra dependent mappings for prediction assistance. However, the above solution encounters two modeling difficulties. First, regarding the fact B_{t+1} satisfies the backdoor criterion in SCM, backdoor adjustment, the most popular mean to deconfond, is still not applicable to cut off $B_{t+1} \rightarrow C_{0:t}$ because B_{t+1} is unobserved for analysis. Second, directly intervening on $C_{0:t}$ would remove all causal relations that points to $C_{0:t}$, including dependent mapping $C_{t+1} \rightarrow C_{0:t} \rightarrow S_{t+1}$ that intended to be constructed to mitigate incompleteness.

We propose to adopt prototypes as intermediate components to intervene for prediction assistance in IFSS, which are not only observable, but also compact and robust se-

mantic representations for classes. In prototype based SCM for IFSS, we firstly insert old-class prototypes $P_{0:t}$ following $C_{0:t}$ and C_{t+1}(Fig. 2(c)), where $P_{0:t} \dashrightarrow S_{t+1}$ is confounded due to the confounder B_{t+1} interfering with the favoured causal relation from old-class prototypes to prediction, and $C_{t+1} \dashrightarrow P_{t+1} \dashrightarrow S_{t+1}$ is confounded due to few samples of new classes to obtain convinced prototypes. Then, we preform causal intervention on prototype-based SCM (Fig. 2(d)), which severs $B_{t+1} \rightarrow C_{0:t}$ to build $C_{0:t} \rightarrow P_{0:t} \rightarrow S_{t+1}$ through the front door adjustment $do(C_{0:t})$, since current condition satisfies its criterion, i.e., owning another observed causal route to skip the biased and confounded relation. Afterwards, we build dependent prediction (Fig. 2(d)), which completes the prediction capability with another old-class prototypes based mapping $P_{t+1} \rightarrow P_{0:t} \rightarrow S_{t+1}$, thus deconfounding $C_{t+1} \dashrightarrow P_{t+1} \dashrightarrow S_{t+1}$ with aid of knowledge gained from old-class prototypes. Finally, unbiased causal effects without semantic shift and incompleteness are resolved for robust prediction in IFSS.

Methodology

Model Overview

We introduce the proposed method with old class and new class learning phases. In the first phase (Fig. 3(a)), we train the model to segment old classes and store old-class prototypes in the memory pool. Specifically, we feed features extracted by ResNet-101 backbone, and ground-truth masks into masked average pooling (MAP) (Zhang et al. 2020b), computing old-class prototypes. Differing from conventional semantic segmentation methods, we combine the vectors of prototypes to segmentation head (Chen et al. 2017) for boosting performance. We update the segmentation model in current phase by cross-entropy loss between predicted and ground-truth masks.

In new class learning phase (Fig. 3(b)), the model learns to segment new classes resisting semantic shift and incompleteness. It is hard to construct convinced new-class prototypes with limited information due to the small sample set of new class. A Prototype Refinement Module (PRM) is thus designed to refine incomplete prototypes of new class, computing representative and compact feature clusters/prototypes with guidance of category-level knowledge embedded in PRM. Since the idea of meta learning is inherited for knowledge extraction, PRM only requires to be trained at the beginning of current phase. After refining, new-class prototypes are stored in memory pool for later prediction.

To obtain prediction S_{t+1}, we first initialize new model M_{t+1} with the frozen parameters of the old model M_t and prototypes stored in the memory pool, computing predictions for new classes C_{t+1}. To deconfound semantic shift with intervention, we then design a CIM following M_t and memory pools, intervening to obtain a unbiased distribution estimation of old classes for unshifted predictions $C_{0:t}$. Finally, we merge $C_{0:t}$ and C_{t+1} to compute S_{t+1} by NMS.

We update new model with total loss L^{t+1}, which consists of cross-entropy loss L_{CE}^{t+1} between new-class masks and ground truth, and the distillation loss L_{KD}^{t+1} (Cermelli et al. 2020a) between old class masks and predictions:

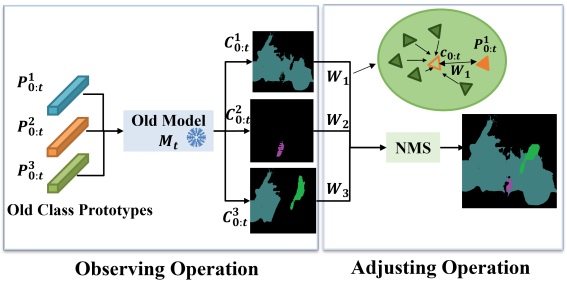

Figure 4: Causal intervention module (CIM) involves operations of passively observing and actively adjusting to deconfond. In observing, CIM predicts old-class segmentation masks with the gathered information about all old-class prototypes. In adjusting, we adaptively adjust old-class prototypes guided by the attention weights achieved by the propose prototype-attention scheme, intervening the confounder for unshifted predictions.

$$\begin{cases} L^{t+1} = \frac{1}{|C_{t+1}|} \sum_{(x,y)\in C_{t+1}} L_{CE}^{t+1}(x,y) + \lambda L_{KD}^{t+1}(x,y) \\ L_{CE}^{t+1}(x,y) = -\frac{1}{|I|} \sum_{i\in I} \log p_x^{t+1}(i, y_i) \\ L_{KD}^{t+1}(x,y) = -\frac{1}{|I|} \sum_{i\in I} \sum_{c\in C_{0:t}} p_x^t(i,c) \log \hat{p}_x^{t+1}(i,c) \end{cases} \tag{2}$$

where λ is a hyper-parameter, and $\hat{p}_x^{t+1}(i,c)$ is the unbiased probability that M_{t+1} classifies pixel i as old class c:

$$\hat{p}_x^{t+1}(i,c) = \begin{cases} p_x^{t+1}(i,c), & \text{if } c \notin B_{t+1} \\ \sum_{k\in C_{t+1}} p_x^{t+1}(i,k), & \text{if } c \in B_{t+1}. \end{cases} \tag{3}$$

Causal Intervention Module

Based on causal analysis of semantic shift in IFSS, we summarize that old-class prototypes $P_{0:t}$ satisfies the front door criterion in an ordered and firmed relation $C_{0:t} \rightarrow P_{0:t} \rightarrow S_{t+1}$. Following principles of front door adjustment, we first observe distributions of old classes for information gathering and then remove the confounder causing semantic shift with intervention operation $do(\cdot)$, thus predicting unbiased segmentation masks:

$$\begin{aligned} \mathcal{P}(S_{t+1}|do(C_{0:t}=c)) &= \sum\nolimits_p \mathcal{P}(P_{0:t}=p|C_{0:t}=c) \\ &\times \sum\nolimits_{c'} \mathcal{P}(S_{t+1}|C_{0:t}=c', P_{0:t}=p)\mathcal{P}(C_{0:t}=c'), \end{aligned} \tag{4}$$

where function $\mathcal{P}(\cdot)$ calculates probability. More precisely, $\mathcal{P}(S_{t+1}|do(C_{0:t}=c))$ predicts the unbiased masks of old class c after intervention. Similarly, $\mathcal{P}(S_{t+1}|C_{0:t}=c', P_{0:t}=p)$ predicts mask of one old class c' with aid of one prototype p. $\mathcal{P}(P_{0:t}=p|C_{0:t}=c)$ is assumed as weight about one prototype p for the resolving old class c, which adaptively adjusts the importance of different prototypes for c during incremental learning. We calculate such weight by negatively correlating it with the distance between the center of c and feature vectors of prototype p. Note that we ignore the class ratio weight $P(C_{0:t}=c')$, i.e., the prior probability of old classes, which is included in parameters of old model.

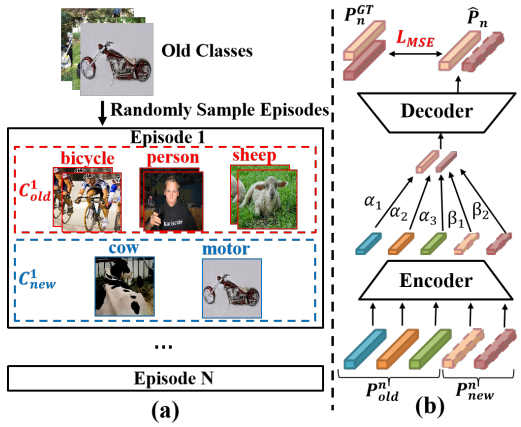

Figure 5: Structure design of prototype refinement module (PRM). (a) The proposed episode learning scheme learns knowledge on how to fuse features for semantic completing. (b) An encoder-decoder is utilized to fuse features of new-class prototypes P^n_{new} and old-class prototypes P^n_{old} guided by the pre-learned and transferred knowledge, predicting with completeness semantics.

Based on the above modeling, we construct CIM with observing and adjusting operation (Fig. 4). During observation, CIM predicts old-class segmentation masks with the gathered information about all old-class prototypes p clustered by the old model M_t, i.e., $C^p_{0:t} = \sum_{c' \in C_{0:t}} \mathcal{P}(S_{t+1}|C_{0:t} = c', P_{0:t} = p) = M_t(p)$. During adjusting, we firstly propose prototype-attention scheme to compute prototype weight $W_{p,t}$ based on the distance between p and old classes at t-th learning step:

$$W_{p,t} = \mathcal{P}(P_{0:t} = p|C_{0:t} = c) = \text{Softmax}(-\|p - \phi_{0:t}\|^2_2) \quad (5)$$

where $\phi_{0:t}$ denotes the center of old classes, and $\|\cdot\|_2$ refers to Euclidean distance. By calculating prototype-wise attention values, CIM successfully indicates directions towards a well-separated prototype distribution with less semantic bias to deconfound shift. Then, we adaptively adjust prototypes by multiplying them with their corresponding weights $W_{p,t}$. Finally, we use NMS to merge the prototype-based old-class prediction masks $C^p_{0:t}$, computing a unbiased old-class mask $m_{C_{0:t}}$:

$$m_{C_{0:t}} = \arg\max_{c \in C_{0:t}} H(\sum\nolimits_p W_{p,t} C^p_{0:t}, c), \quad (6)$$

where function $H(\cdot, c)$ denotes the probability that each pixel in a mask belongs to the class c.

Prototype Refinement Module

Since PRM is designed to refine incomplete new-class prototypes guided by the knowledge pre-learned from old-class prototypes, we propose an episode learning scheme to train PRM following rules of episodic training (Vinyals et al. 2016; Wang et al. 2022).

As shown in Fig. 5(a), the proposed episode learning scheme consists of N episodes, each episode drawing from old classes $C_{0:t}$ and formulating as an IFSS task with only two incremental learning steps, i.e., initialization at $t = 0$ and the consecutive learning step at $t = 1$. During $t = 1$, C_0 is pretended as source of "old" and "new" classes for the first incremental learning step, where the number of new classes is defined as $|C_1|$ and old classes are the remaining ones $C_0 - C_1$. Note that $|C_1|$ is far smaller than that of old classes $|C_0 - C_1|$, acting as simulation of few-shot learning scenario. We further extract prototype features of each episode for training. Specifically, we directly obtain old-class prototypes P_b from the memory pool, and we get incomplete new-class prototypes P_n by feeding k-shot samples into backbone and MAP, being similar with the process of old-class learning phase.

After preparation, all prototypes are fed into an encoder-decoder model (Fig. 5(b)) and learn knowledge to fuse features of new-class and old-class prototypes for semantic completing. Specifically, we first encode the input prototypes into a unified low-dimensional embedding space, and then refine new-class prototypes inside space via a decoder:

$$\hat{P}_n = D[\sum\nolimits_i \alpha_i E(P_b) + \sum\nolimits_j \beta_j E(P_n)]. \quad (7)$$

Here, $E(\cdot)$ and $D(\cdot)$ denote encoder and decoder structure, respectively. i and j denote index of old and new classes, respectively, and α_i denotes the coefficient of i-th old-class prototype P_i. Note that the coefficients of j-th new-class prototypes β_j should be fixed as 1 because the predicted new class prototypes should be regulated not exceeding too much from the input new-class prototypes.

Experiments

Experimental Setup

Datasets. We conduct experiments on two widely used ISS datasets, i.e., Pascal-VOC 2012 (Everingham et al. 2015) and ADE20k (Zhou et al. 2017).

Metrics. We consider four mIoU metrics. We use the average mIoU on old classes C_0 (full) and new classes $C_{1:\mathrm{T}}$ (full) to evaluate the anti-forgetting capability of old classes and learning ability of new classes. We use the average mIoU on new classes under 1-shot and 5-shot, respectively (denoted as $C_{1:\mathrm{T}}$ (1/5-shot)) to evaluate the performance of our method.

IFSS settings. 1) Tasks. For Pascal-VOC 2012, we consider three tasks, including 19-1 (T=2), 15-5 (T=2), and 15-1 (T=6). For ADE20k, we similarly consider three tasks, including 100-50 (T=2), 50-50 (T=3), and 100-10 (T=6). **2) Disjoint or overlapped.** As described in (Cermelli et al. 2020a), images only contain pixels of the previous or current classes in the disjoint settings. In the overlapped settings, images may contain pixels of future classes, which presents a more realistic and challenging scenario, leading us to conduct experiments under such setting.

Baselines. The simple fine-tuning (**FT**) on each C_t are reported as the baseline. We compare our method with three state-of-the-arts IFSS methods, namely **GIFS** (Cermelli et al. 2020b), **PIFS** (Cermelli et al. 2021) and **EHNet** (Shi et al. 2022)). In addition, four popular ISS methods:

Method	**19-1** (T=2)				**15-5** (T=2)				**15-1** (T=6)			
	0-19 (full)	20 (full)	20 (5-shot)	20 (1-shot)	0-15 (full)	16-20 (full)	16-20 (5-shot)	16-20 (1-shot)	0-15 (full)	16-20 (full)	16-20 (5-shot)	16-20 (1-shot)
FT	6.80	12.90	0.08	0.00	2.10	33.10	1.88	0.04	0.20	1.80	1.30	0.00
GIFS	64.31	33.72	25.90	14.39	64.70	37.41	32.00	17.40	40.77	14.93	8.32	4.25
PIFS	63.75	34.47	26.41	14.68	60.90	38.92	33.40	18.62	40.34	14.81	9.36	4.33
EHNet	56.71	36.23	30.23	13.21	57.10	45.70	33.42	19.73	39.38	17.29	10.71	6.05
ILT	67.10	12.30	12.70	0.25	66.30	40.60	10.29	5.96	4.90	7.80	3.64	0.12
MiB	70.20	22.10	17.13	4.10	75.50	49.40	16.10	6.25	35.10	13.50	2.75	2.81
EM (replay)	73.76	**43.42**	23.50	6.34	75.56	49.89	17.11	7.87	**75.77**	**40.34**	4.52	2.47
PLOP	75.35	37.35	15.42	0.86	75.73	51.71	11.23	2.52	65.12	21.11	2.44	0.68
PIFS+CIM	70.89	35.63	26.87	14.65	67.11	39.21	33.78	18.91	52.76	18.95	9.81	5.62
EHNet+CIM	64.36	36.88	30.75	13.74	65.25	46.33	33.63	20.03	50.21	18.17	10.98	6.83
MiB+PRM	71.04	27.45	22.59	12.29	75.54	49.34	24.51	15.36	38.14	17.29	9.35	5.06
PLOP+PRM	74.73	38.29	23.85	10.29	74.98	**52.14**	24.81	13.94	65.47	22.89	10.71	5.91
Ours	**76.04**	38.77	**31.22**	**15.32**	**76.30**	50.67	**34.20**	**21.37**	66.49	24.19	**15.44**	**7.42**

Table 1: mIoU results of comparative methods on Pascal-VOC 2012 dataset under different IFSS settings.

Method	**100-50** (T=2)				**50-50** (T=3)				**100-10** (T=6)			
	0-100 (full)	101-150 (full)	101-150 (5-shot)	101-150 (1-shot)	0-50 (full)	51-150 (full)	51-150 (5-shot)	51-150 (1-shot)	0-100 (full)	101-150 (full)	101-150 (5-shot)	101-150 (1-shot)
ILT	18.29	14.40	3.82	1.42	3.53	12.85	3.95	1.35	0.11	3.06	1.25	0.07
MiB	40.52	17.17	6.85	3.22	45.57	21.01	7.12	3.83	38.21	11.12	2.63	1.71
PLOP	41.87	14.89	5.44	2.93	**47.33**	20.27	5.68	3.47	38.59	14.21	1.97	0.61
MiB+PRM	40.67	17.83	8.24	4.87	45.32	21.27	8.34	5.77	36.18	13.81	4.07	2.16
PLOP+PRM	41.78	17.03	8.61	4.53	46.81	20.94	8.17	5.08	38.77	**14.47**	3.88	2.14
Ours	**41.89**	**17.92**	**10.41**	**6.84**	46.57	**21.54**	**10.89**	**8.07**	**38.94**	13.93	**6.96**	**4.19**

Table 2: mIoU results of comparative methods on ADE20K dataset under different IFSS settings.

ILT (Michieli and Zanuttigh 2019), **MiB** (Cermelli et al. 2020a), **EM** (Yan et al. 2021) and **PLOP** (Douillard et al. 2021) are also compared.

Comparison with SOTAs

Comparison with IFSS methods on Pascal-VOC 2012. Table 1 shows the experimental results with 19-1 (T=2), 15-5 (T=2) and 15-1 (T=6) compared with IFSS methods. Compared with the latest EHNet method, our method improves mIoU of old classes (0-15) and new classes (16-20) by 33.63% and 10.88% with 15-5 (T=2) setting, respectively. In both 1-shot and 5-shot segmentation tasks, our method outperforms EHNet in all cases. This proves that our method can strike the balance between anti-forgetting of old classes and learn new ones.

Comparison with ISS methods on Pascal-VOC 2012. In Table 1, our method outperforms all other methods on C_0 and $C_{1:T}$ (1/5 shot) mIoU except EM, which uses replay technique that partially preserves old class data to resist forgetting, might causing severe privacy concerns. With the short 15-5 (T=2) setting, our method outperforms PLOP by 0.75% on mIoU of old classes (0-15), which indicates the capability of our method in resisting forgetting. With 5-shot and 1-shot, the mIoU scores on new classes (16-20) is greatly increased to 34.20% and 21.37%, respectively. This proves that our method can effectively fuse features between new-class and old-class prototype for semantic completing. With the long 15-1 (T=6) setting, the mIoU result of new classes (16-20) is greatly increased to 15.44% and 7.42% under 5-shot and 1-shot, respectively. It proves that our method can resist forgetting in long-term incremental learning.

Comparison with ISS methods on ADE20k. Table 2 shows results with 100-50 (T=2), 50-50 (T=3) and 100-10 (T=6) settings. With the short 100-50 (T=2) setting, our method outperforms PLOP by 0.02% and 20.35% on old (0-100) and new (101-150) classes, respectively. Moreover, the mIoU result achieved by our method increases to 6.84% and 10.41% with 1-shot and 5-shot setting, respectively, which is superior to the comparative methods. Though the mIOU result of our method is slightly lower than that achieved by PLOP on old classes (0-50), our method is more robust and outperforms the others in most cases, especially With the long 100-10 (T=6) setting, mIoU of new classes (101-150) by our method is greatly increased to 6.96% and 4.19% under 5-shot and 1-shot settings, respectively.

Results of comparative methods embedded with CIM and PRM. Considering CIM and PRM are model-agnostic, we embed CIM into IFSS methods for unbiased predictions of old classes and PRM into ISS methods. As shown in Table 1, on Pascal-VOC 2012, the embedded EHNet+CIM with 19-1 (T=2), 15-5 (T=2), and 15-1 (T=6) settings improve mIoU on old classes by 13.49%, 14.27%, and 27.50%, respectively. By embedding PRM into MiB and PLOP, the segmentation performance on new classes under few-shot setting can be greatly improved. By embedding CIM into PIFS and EHNet, the anti-forgetting capability is improved. As shown in Table 2, on ADE20K, PLOP+PRM improves mIoU on the new class by 46.40%(1-shot) and 43.84%(5-

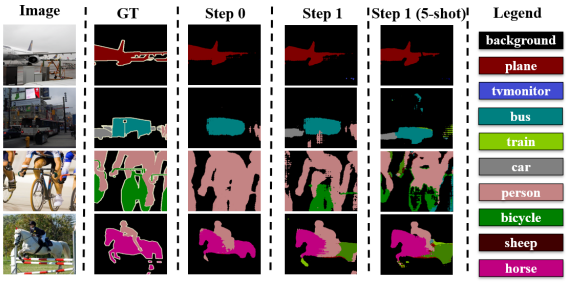

Figure 6: Qualitative results of our method on Pascal-VOC 2012 dataset with 15-5 (T=2) setting.

Module		**15-5** (T=2)			
PRM	CIM	0-15 (full)	16-20 (full)	16-20 (5-shot)	16-20 (1-shot)
✗	✗	60.81	35.91	16.10	6.25
✓	✗	64.35	48.30	31.15	19.27
✗	✓	74.11	42.83	18.17	7.93
✓	✓	**76.30**	**50.67**	**34.20**	**21.37**

Table 3: Ablation study of PRM and CIM on Pascal-VOC 2012 dataset with 15-5 (T=2) setting.

Distillation Loss	0-15 (full)	16-20 (full)	16-20 (5-shot)	16-20 (1-shot)
KD	61.41	38.71	21.66	12.94
POD	66.72	41.29	24.87	15.54
Local POD	75.81	**50.89**	31.19	20.65
UNKD (Ours)	**76.30**	50.67	**34.20**	**21.37**

Table 4: Results with different distillation loss on Pascal-VOC 2012 dataset with 15-5 (T=2) setting.

shot) with 50-50 (T=3) setting. Combing the results of Table 1 and Table 2, we observe that the proposed CIM and PRM are practical and efficient for various IFSS and ISS methods.

Ablation Study

Table 3 shows the ablation results of PRM and CIM on Pascal-VOC 2012 dataset. When we remove PRM, the segmentation performance on new classes decreases sharply. With full-shot, 1-shot and 5-shot settings, mIOU on new classes decrease to 15.47%, 46.87%, and 62.89%, respectively. This proves that PRM enhances feature representability successfully by introducing the transferred knowledge and old-class prototypes. When we remove CIM, mIOU on old classes drops 15.66%, which proves CIM can effectively solve the semantic shift problem and prevent forgetting on old-class information.

Based on Eq.(2), the comparative results of different distillation losses, including KD, POD (Douillard et al. 2020), local POD (Douillard et al. 2021) and UNKD we used (Cermelli et al. 2020a) are shown in Table 4. As observed, UNKD performs better than other distillation losses in most cases UNKD is inferior to the second best local POD by only 0.22% under 16-20 (full) setting. This indicates the inconsistency of such distillation strategy during training. Besides, we can analyse that POD fails to improve with the new classes (16-20) and would be easily to be overfitting with new classes due to plasticity property.

		0-15 (full)	16-20 (full)	16-20 (5-shot)	16-20 (1-shot)
λ	5	75.70	50.32	34.21	21.33
	10	76.30	50.67	34.20	21.37
	20	76.33	48.92	33.76	21.25
Lr	0.05	75.32	50.60	33.42	21.51
	0.01	76.30	50.67	34.20	21.37
	0.001	75.58	49.33	33.12	20.89

Table 5: Results with different value of hyper-parameter on Pascal-VOC 2012 dataset with 15-5 (T=2) setting.

Episodes	15-5(T=2)	15-1 (T=6)				
	Step1	Step1	Step2	Step3	Step4	Step5
1000	320.7s	317.2s	298.4s	301.7s	314.3s	325.4s
2000	634.1s	632.1s	613.7s	613.0s	625.6s	649.0s

Table 6: Results of pre-training time for PRM.

Table 5 shows the ablation results with different λ and Lr values, where λ is the weight of the distillation loss, Lr is the learning rate for Step 0, and the learning rate for further steps is 0.1*Lr. We use $\lambda = 10$ for all experiments for a fair comparison. Table 6 shows results of PRM pre-training time on Pascal-VOC 2012. PRM is pre-trained before each Step to learn how to use all previous class knowledge to supplement the semantics of new classes. Note that our CIM module does not require training.

Qualitative Experiments

Visualization results on Pascal-VOC 2012 dataset with 15-5 (T=2) setting are shown in Fig. 6. Specifically, our method can maintain the segmentation of old classes with only one "plane" class. Dealing with the complex scene shown in the second image, our method can accurately segment "bus" at step 0, while the segmentation of person is incorrect due to its small size and edge location. At step 1, our method can learn new class "car" while retaining the knowledge on old class "bus" and "person". Even in few-shot setting, our method can segment the car well. In the third image, our method confuses the future class "bicycle" with the current class "person" at step 0, but learns to segment "bicycle" at step 1. The fourth image shows a failure, where our method forgets old class "horse" after incremental learning. This is because the small and imbalanced new-class dataset could cause serious cognition confusion in few-shot setting.

Conclusion

This paper proposes a causal framework to deconfound semantic shift and incompleteness. CIM is proposed to resist semantic shift by removing the confounder with intervention. PRM is proposed to complete semantics with the transferred knowledge for feature fusion. Experimental results demonstrate that our method outperforms the existing IFSS and ISS methods. An extensive study shows the rationality and efficiency of the proposed causal framework for IFSS.

Acknowledgments

This work was supported by the National Key R&D Program of China (2023YFC3006501), the Natural Science Foundation of Jiangsu Province of China (BK20242050), High Performance Computing Platform, Hohai University, the National Natural Science Foundation of China (Grant No. 62372223, U24A20330, and No. 62172438), Stable Support Project of Shenzhen (20231120161634002), Shenzhen Science and Technology Program (20240806132029001 and JCYJ20220818103200002), Guangdong Provincial Department of Education (2024KTSCX060), and the Natural Science Foundation of Guangdong Province (Grant No. 2024A1515011155).

References

Bareinboim, E.; and Pearl, J. 2012. Controlling Selection Bias in Causal Inference. 100–108.

Besserve, M.; Mehrjou, A.; Sun, R.; and Schölkopf, B. 2020. Counterfactuals uncover the modular structure of deep generative models. In *Proceedings of International Conference on Learning Representations*.

Cermelli, F.; Mancini, M.; Bulò, S. R.; Ricci, E.; and Caputo, B. 2020a. Modeling the Background for Incremental Learning in Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9230–9239.

Cermelli, F.; Mancini, M.; Xian, Y.; Akata, Z.; and Caputo, B. 2020b. A Few Guidelines for Incremental Few-Shot Segmentation. *CoRR*, abs/2012.01415.

Cermelli, F.; Mancini, M.; Xian, Y.; Akata, Z.; and Caputo, B. 2021. Prototype-based incremental few-shot semantic segmentation. In *Proceedings of the 32nd British Machine Vision Conference*.

Chen, L.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR*, abs/1706.05587.

Douillard, A.; Chen, Y.; Dapogny, A.; and Cord, M. 2021. PLOP: Learning Without Forgetting for Continual Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4040–4050.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Proceedings of European Conference on Computer Vision*, 86–102.

Everingham, M.; Eslami, S. M. A.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.*, 111(1): 98–136.

Hu, X.; Tang, K.; Miao, C.; Hua, X.; and Zhang, H. 2021. Distilling Causal Effect of Data in Class-Incremental Learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 3957–3966.

Liu, Y.; Chen, J.; Chen, Z.; Deng, B.; Huang, J.; and Zhang, H. 2021. The Blessings of Unlabeled Background in Untrimmed Videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 6176–6185.

Michieli, U.; and Zanuttigh, P. 2019. Incremental Learning Techniques for Semantic Segmentation. In *Proceedings of International Conference on Computer Vision Workshops*, 3205–3212.

Neubeck, A.; and Gool, L. V. 2006. Efficient Non-Maximum Suppression. In *Proceedings of International Conference on Pattern Recognition*, 850–855.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.; and Wen, J. 2021. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 12700–12710.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; Glymour, M.; and Jewell, N. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.

Phan, M.; Ta, T.; Phung, S. L.; Tran-Thanh, L.; and Bouzerdoum, A. 2022. Class Similarity Weighted Knowledge Distillation for Continual Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16845–16854.

Shang, C.; Li, H.; Meng, F.; Wu, Q.; Qiu, H.; and Wang, L. 2023. Incrementer: Transformer for Class-Incremental Semantic Segmentation with Knowledge Distillation Focusing on Old Class. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 7214–7224.

Shi, G.; Wu, Y.; Liu, J.; Wan, S.; Wang, W.; and Lu, T. 2022. Incremental Few-Shot Semantic Segmentation via Embedding Adaptive-Update and Hyper-class Representation. In *Proceedings of ACM Conference on Multimedia*, 5547–5556.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *Proceedings of Annual Conference on Neural Information Processing Systems*, 3630–3638.

Wang, K.; Liu, X.; Bagdanov, A.; Herranz, L.; Jui, S.; and van de Weijer, J. 2022. Incremental Meta-Learning via Episodic Replay Distillation for Few-Shot Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 3728–3738.

Wang, Q.; Wu, Y.; Yang, L.; Zuo, W.; and Hu, Q. 2024. Layer-Specific Knowledge Distillation for Class Incremental Semantic Segmentation. *IEEE Trans. Image Process.*, 33: 1977–1989.

Wang, T.; Zhou, C.; Sun, Q.; and Zhang, H. 2021. Causal Attention for Unbiased Visual Recognition. In *Proceedings of International Conference on Computer Vision*, 3071–3080.

Yan, S.; Zhou, J.; Xie, J.; Zhang, S.; and He, X. 2021. An EM Framework for Online Incremental Learning of Semantic Segmentation. In *Proceedings of ACM Multimedia Conference*, 3052–3060.

Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P.; and Zhang, L. 2023. Context De-Confounded Emotion Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 19005–19015.

Yang, X.; Zhang, H.; and Cai, J. 2020. Deconfounded Image Captioning: A Causal Retrospect. *CoRR*, abs/2003.03923.

Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X. 2020. Interventional Few-Shot Learning. In *Proceedings of Annual Conference on Neural Information Processing Systems*.

Zhang, C.; Xiao, J.; Liu, X.; Chen, Y.; and Cheng, M. 2022. Representation Compensation Networks for Continual Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7043–7054.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.; and Sun, Q. 2020a. Causal Intervention for Weakly-Supervised Semantic Segmentation. In *Proceedings of Annual Conference on Neural Information Processing Systems*.

Zhang, X.; Wei, Y.; Yang, Y.; and Huang, T. S. 2020b. SG-One: Similarity Guidance Network for One-Shot Semantic Segmentation. *IEEE Trans. Cybern.*, 50(9): 3855–3865.

Zhao, D.; Yuan, B.; and Shi, Z. 2023. Inherit With Distillation and Evolve With Contrast: Exploring Class Incremental Semantic Segmentation Without Exemplar Memory. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45: 11932–11947.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 5122–5130.