Memory-Based Contrastive Learning with Optimized Sampling for Incremental Few-Shot Semantic Segmentation

Yuxuan Zhang*†, Miaojing Shi‡, Taiyi Su*† and Hanli Wang*†§
*Department of Computer Science & Technology, Tongji University, Shanghai, P. R. China
†Key Laboratory of Embedded System & Service Computing, Ministry of Education, Tongji University, Shanghai, P. R. China
‡Department of Control Science & Engineering, Tongji University, Shanghai, P. R. China

***Abstract*—Incremental few-shot semantic segmentation (IFSS) aspires to incrementally expand a semantic segmentation model's proficiency to identify new classes based on few samples. However, it grapples with the dual challenges of catastrophic forgetting due to feature drift in old classes, and overfitting triggered by inadequate samples for pixel-level segmentation in new classes. To address these issues, a novel approach integrating pixel-wise and region-wise contrastive learning, complemented by a strategy for optimized example and anchor sampling is proposed. The proposed method incorporates a region memory and pixel memory designed to explore the high-dimensional visual embedding space more effectively. The memory, retaining the feature embeddings of known classes, facilitates the calibration and alignment of the old and new class features during the learning process of new classes. This process considerably reduces feature drift and improves the model's adaptability. To mitigate overfitting, the proposed approach implements optimized example and anchor sampling strategies, which together increase the model's stability during incremental few-shot learning. The proposed model is validated on the PASCAL VOC 2012 and the COCO dataset, showing competitive performance and demonstrating the effectiveness of proposed method.**

***Index Terms*—incremental learning, few-shot learning, semantic segmentation, contrastive learning, dynamic memory**

I. Introduction

Semantic segmentation is a fundamental task in computer vision, aiming to predict the class label for each pixel of an image. Traditional methods [1]–[4] require substantial labeled data, which is costly and time-consuming to obtain. Few-shot learning [5]–[7] has been proposed to learn from a small number of examples, but it still faces challenges, such as catastrophic forgetting [8], where a model loses prior knowledge when learning new classes.

In the context of few-shot semantic segmentation, catastrophic forgetting occurs when a model is trained on new classes and forgets the features of old classes. Overfitting, on the other hand, occurs when a model is trained on too few examples and fails to generalize to new examples. In few-shot semantic segmentation, overfitting is particularly challenging because of the need to segment images at the pixel level. To address these challenges, Cermelli et al. [9] proposed the Incremental Few-shot Semantic Segmentation (IFSS) task, using the prototype-based knowledge distillation. It relieves the catastrophic forgetting issue by constraining the invariance of old class segmentation scores. Moreover, the overfitting to novel categories is suppressed by boosting the consistency between old and updated models. Shi et al. [10] proposed to build hyper-class feature representations, thereby helping to relieve the representation drift during the incremental learning. Such enhancement in flexibility and stability is crucial for various industrial applications like autonomous driving and video surveillance.

§Corresponding author: H. Wang, Email: hanliwang@tongji.edu.cn.

More specifically, in the IFSS task, a base set with sufficient training samples is firstly provided to initialize the learnable parameters of a segmentation model. Next, a few pixel-level annotated training samples of novel categories are given to help the model incrementally expand its segmentation ability to novel categories. Our method offers potential solutions to the prevalent issues of catastrophic forgetting and overfitting. The motivation behind our method is to address the unavailability of old class features when learning new classes, a common issue with existing methods. The proposed approach utilizes a dynamic memory mechanism to preserve the features of old classes. In the process of learning new classes, features of both old and new classes are aligned, thereby resulting in a structured embedding space. This alignment mitigates the effects of catastrophic forgetting and overfitting and leads to a more robust and adaptive model.

The proposed approach integrates the pixel-wise and region-wise contrastive learning with a dedicated optimized positive/negative example and anchor sampling strategy for shaping a well-structured embedding space. We first design region and pixel memory modules to better explore the high-dimensional visual embedding space. The memory, retaining the pixel-level and region-level feature embeddings of known classes, facilitates the calibration and alignment of the old and new class features during the learning process of new classes. This process considerably reduces feature drift and improves the model's adaptability. On the other hand, to mitigate model overfitting with a small amount of training data, we implement optimized example and anchor sampling strategy, focusing on the selection of informative samples and directing the segmentation model to concentrate more on hard-segmented pixels. These strategies enhance the model's stability during incremental few-shot learning.

In brief, the contributions of this paper are:

- A novel approach integrating pixel-wise and region-wise contrastive learning with memory mechanisms, is proposed to effectively explore the high-dimensional visual embedding space while addressing catastrophic forgetting and overfitting.
- Optimized example and anchor sampling strategies are implemented to increase the model's stability during incremental few-shot learning.
- Experimental results on the PASCAL VOC 2012 [11] and the COCO [12] datasets show competitive performance of proposed method compared to the state of the art.

II. Method

A. Preliminaries

In the IFSS model, the semantic space expands as the learning progresses. At step t, the categories C_t are encountered. After this learning step, the model's semantic space expands to $T_t = T_{t-1} \cup C_t$, where $T_{t-1} = \bigcup_{i=0}^{t-1} C_i$ represents the semantic embedding space

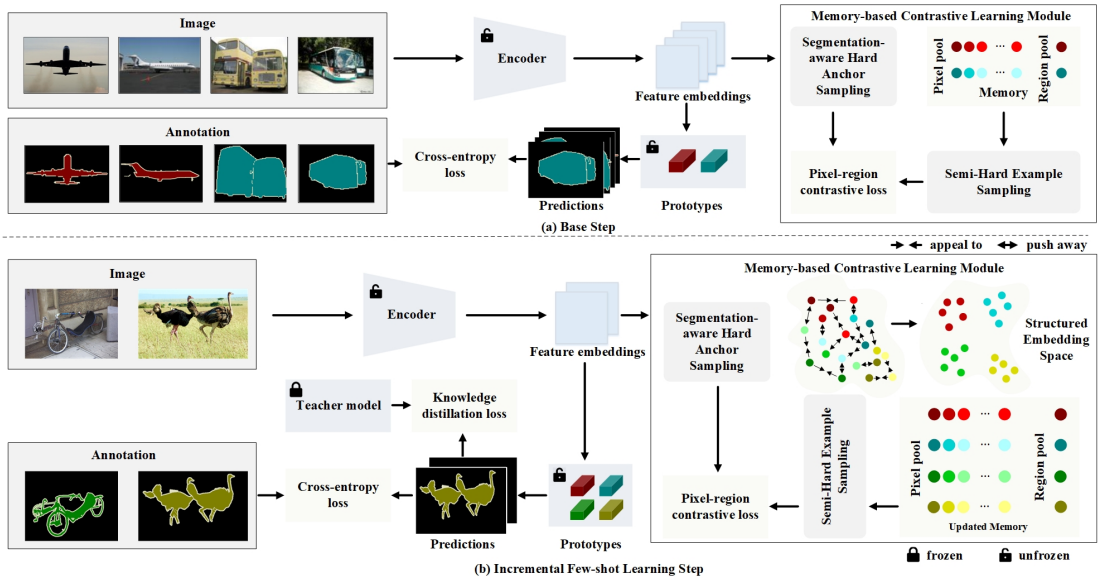

Fig. 1. Overall schematics of the proposed incremental few-shot segmentation method. Dynamic memory aims to preserve features at two levels of granularity, reducing the model's tendency to forget base classes. Optimized sampling focuses on selecting valuable anchors and positive/negative samples. Combined with pixel-region contrastive learning of old class embeddings, it aids in calibrating and aligning old and new class features during the learning of new classes.

learned up to step $t-1$. At each step, a dataset $D_t = (X_t^n, Y_t^n)_{n=1}^{N_t}$ is provided to update the learnable parameters. Here, X_t^n denotes the n-th training image, and Y_t^n is its corresponding label map. For the IFSS task, the base dataset D_0 is given in the initial step (i.e., $t=0$) to initialize the model parameters, containing a larger number of training samples compared to later steps. After the initial step, the dataset D_t is in a few-shot setting, meaning each category has only one or a few labeled training instances. This satisfies the condition $N_t \ll N_0$ for all $t \geq 1$. For simplicity, the categories provided in the initial step are referred to as base categories, and those encountered during the incremental learning stage are called novel categories. In step t, only the dataset D_t can be accessed by the model, and datasets from previous steps are not available.

The feature extractor is denoted as $f(\cdot \mid \Theta)$, where Θ symbolizes the parameters of the network. This extractor then maps the input into an $\mathbb{R}^D$ dimensional embedding space. For classification, each class c is equipped with a prototype-based classifier p_c within this $\mathbb{R}^D$ space. The classifiers for all individual classes can be shown as $P = \{p_i \mid i \in C\}$. The overall schematics of the proposed incremental few-shot segmentation method are depicted in Fig. 1.

B. Region-Wise Memory-Based Contrastive Learning

Dynamic Memory. A crucial element of the proposed methodology is the dynamically updated memory, which is a dynamic representation of the semantic embedding space. The memory is updated at each learning step t with the new categories C_t encountered, expanding T_t to $T_t = T_{t-1} \cup C_t$.

The memory stores pixel-wise embeddings for known categories, capturing fine-grained information, and region-wise embeddings, which provide a more holistic view of each semantic class, thereby preserving coarse-grained information. This combination effectively captures both detailed and broader features, enriching the representation of the visual embedding space. The pixel-level memory has a size of $|T_t| \times M \times D$, where $|T_t|$ is the number of classes learned so far, M is the number of pixel embeddings stored per class, and D is the dimension of the pixel embeddings. The region-level memory is of size $|T_t| \times N \times D$, where N is the number of region embeddings stored. During each training iteration, the elements of region memory are updated by enqueuing and dequeuing region-wise embeddings, obtained by average pooling all pixel embeddings of a particular category in a given image. Then the elements of pixel memory are updated by enqueuing and dequeuing random selected pixel-wise embeddings. Additionally, the pixel memory and region memory are both round-robin queues, which provide a comprehensive and evolving representation of the semantic space as learning progresses.

Pixel-region Contrastive Learning. The proposed approach leverages pixel-region contrastive learning by employing both pixel-wise and region-wise contrastive loss. This is based on the pixel-to-pixel contrastive and pixel-to-region contrastive methods, which aim to regularize the embedding space by pulling together pixel samples of the same class and pushing apart pixel samples of different classes.

For a pixel i with its ground truth semantic label $\bar{c}$, the positive samples are other pixels also belonging to the class $\bar{c}$, while the negatives are pixels belonging to other classes $C \setminus \bar{c}$. Our supervised, pixel-wise contrastive loss is defined as:

$$L_{cl}^{i} = \frac{1}{|\mathcal{P}_i|} \sum_{i^+ \in \mathcal{P}_i} -\log \frac{\exp(i \cdot i^+/\tau)}{\exp(i \cdot i^+/\tau) + \sum_{i^- \in \mathcal{N}_i} \exp(i \cdot i^-/\tau)}, \tag{1}$$

where $\mathcal{P}_i$ and $\mathcal{N}_i$ denote pixel embedding collections of the positive and negative samples, respectively; for anchor pixel embedding i, '$\cdot$' denotes the inner product, and τ is a temperature hyper-parameter.

Furthermore, our pixel-region contrastive loss allows us to explore pixel-to-region relationships. When computing the pixel-wise contrastive loss for an anchor pixel embedding i belonging to class $\bar{c}$, stored region embeddings with the same class $\bar{c}$ are viewed as positives, while region embeddings with other classes $C \setminus \bar{c}$ are negatives. And $\mathcal{P}_i$ and $\mathcal{N}_i$, shown in Eq. 1, denote the positive and negative elements from joint pixel and region pool.

C. Optimized Sampling

Segmentation-Aware Hard Anchor Sampling. To improve the effectiveness of contrastive learning, an anchor sampling strategy called segmentation-aware hard anchor sampling is introduced. The categorization ability of an anchor embedding is treated as its importance during contrastive learning, thus the pixels with incorrect predictions, i.e., $c \neq \bar{c}$, are treated as hard anchors. For the contrastive loss computation (Eq. 1), half of the anchors are randomly sampled

and half are the hard ones. This strategy promotes contrastive learning to concentrate more on pixels that are challenging for segmentation, shaping more segmentation-aware embeddings.

Semi-hard Example Sampling. In addition to hard anchor sampling, a semi-hard example sampling strategy is also employed for selecting the positive and negative samples used in contrastive learning. The gradient of the contrastive loss (Eq. 1) w.r.t. the anchor embedding i can be given as:

$$\frac{\partial L_{cl}^{i}}{\partial i} = -\frac{1}{\tau \left|\mathcal{P}_i\right|} \sum_{i^+ \in \mathcal{P}_i} \left((1 - p_{i^+}) \cdot i^+ - \sum_{i^- \in \mathcal{N}_i} p_{i^-} \cdot i^- \right), \quad (2)$$

$$p_{i^{+/-}} = \frac{\exp\left(i \cdot i^{+/-}/\tau\right)}{\sum_{i' \in \mathcal{P}_i \cup \mathcal{N}_i} \exp\left(i \cdot i'/\tau\right)}, \quad (3)$$

where $p_{i^{+/-}} \in [0, 1]$ denotes the matching probability between a positive/negative $i^{+/-}$ and the anchor i. Negatives with dot products (i.e., $i \cdot i^-$) closer to 1 are viewed as harder negatives, which are similar to the anchor i. Similarly, the positives with dot products (i.e., $i \cdot i^+$) closer to -1 are considered as harder positives, which are dissimilar to i. It can be found that harder negatives bring more gradient contributions, i.e., p_{i^-} , than easier negatives. This principle also holds true for positives, whose gradient contributions are $1 - p_{i^+}$. However, optimizing with the hardest negatives for metric learning is likely to result in bad local minima [13]. Thus a "semi-hard example sampling" strategy is further designed: for each anchor embedding i, the top 10% nearest negatives (resp. top 10% farthest positives) from the joint memory are first collected, from which the remaining ones are randomly sampled to U negatives (resp. V positives) for the contrastive loss computation. By adopting this sampling strategy, the proposed method can learn more robust and discriminating embeddings.

D. Loss Function

As shown in Fig. 1, the framework includes the base step and incremental few-shot learning step. Besides the previously mentioned pixel-region contrastive loss L_{cl}, the cross-entropy loss L_{ce} is also used on both stages and the prototype-based knowledge-distillation loss L_{kd} [9] is applied only on the incremental few-shot learning step. Thus, the total loss is:

$$L_{tot} = \begin{cases} L_{ce} + \lambda_1 L_{cl}, & \text{on base step} \\ L_{ce} + \lambda_1 L_{cl} + \lambda_2 L_{kd}, & \text{on incremental step} \end{cases} \quad (4)$$

III. Experiments

A. Experimental Setup

Dataset and Metric. The proposed method is evaluated on two popular semantic segmentation datasets: PASCAL VOC 2012 with 20 categories (10,582 training and 1,449 test images) and COCO with 80 categories (around 110k training and 5k test images). Following the same protocol as in [9], both datasets are divided into four folds for cross-validation. Three folds form the base set, and the remaining fold is used for testing.

Implementation Details. The codes are implemented using PyTorch and run on two tesla V100 GPU cards. Training details vary slightly for the PASCAL VOC 2012 and COCO datasets. For PASCAL VOC 2012, the base step includes 30 epochs and the incremental learning stage uses 1000 iterations, with initial learning rates of 0.01 and 0.001, respectively. For the COCO dataset, the base set training lasts for 20 epochs with a poly learning rate of 0.01. The incremental learning phase sets the iterations at 2000 and the initial learning rate at 0.001. According to protocol [9], the proposed method is evaluated in both single and multiple few-shot step settings, using cross-validation. The single few-shot step setting introduces all new categories at once during an incremental step, while the multiple few-shot step setting introduces new categories across multiple stages. The effectiveness is measured using mean Intersection-over-Union (mIoU). The sizes of the pixel pool and region pool are set at 8000 and 2000, respectively, with 1024 positives and 2048 negatives. In line with [9], Deeplab-v3 [21] with ResNet-101 [22] is employed.

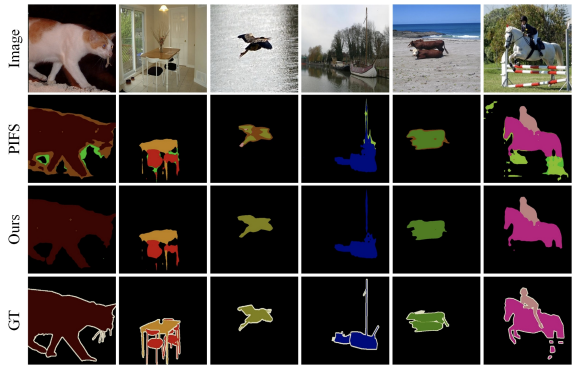

Fig. 2. The qualitative comparison between PIFS and ours. All prediction results are from the last step on the 1-shot single few-shot step setting of the PASCAL VOC 2012 dataset.

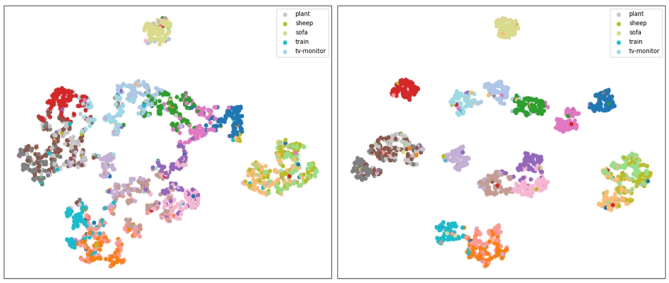

Fig. 3. T-SNE visualization of features learned with (left) PIFS and (right) our memory-based contrastive learning method on Pascal VOC 2012. The five categories in the legend represent novel classes.

B. Main Results

Experimental results on both PASCAL VOC 2012 and COCO datasets reveal the proposed method's effectiveness, particularly on novel categories, even under challenging few-shot step settings. Table I presents the significant improvements on the PASCAL VOC 2012 dataset, surpassing FT, RT, AMP, SPN, and PIFS in a single few-shot step setting (SS). 'FT' indicates directly finetuning the model on novel classes using the cross-entropy loss, and 'HM', as the principal metric, indicates the harmonic mean of the mIoU on base and novel classes. The method outperforms MIB, ILT, LWF, DWI, and WI on novel categories, despite slightly inferior performance on base categories. For instance, on a 2-shot task, the HM mIoU increases by 17.3%, 4.4%, 9.2%, 8.0%, and 8.9% compared to these respective methods. The method also exhibits superiority under multiple few-shot step settings (MS), outperforming PIFS on all 1-shot, 2-shot, and 5-shot tasks. Furthermore, it achieves state-of-the-art performance across these tasks, marking a significant advancement in the field. Table II demonstrates similar results on the COCO dataset. In both single and multiple few-shot step settings, our method's HM mIoU consistently surpasses all other methods, thereby leading in performance across 1-shot, 2-shot, and 5-shot tasks. The results show the proposed method's high effectiveness in segmenting categories learned in previous few-shot learning steps.

TABLE I

THE EXPERIMENTAL RESULTS ON THE PASCAL VOC 2012 DATASET.

Method	SS									MS								
	1-shot			2-shot			5-shot			1-shot			2-shot			5-shot		
	mIoU (%)			mIoU(%)			mIoU(%)			mIoU (%)			mIoU (%)			mIoU (%)		
	base	novel	HM	base	novel	HM	base	novel	HM	base	novel	HM	base	novel	HM	base	novel	HM
FT	58.3	9.7	16.7	59.1	19.7	29.5	55.8	29.6	38.7	47.2	3.9	7.2	53.5	4.4	8.1	58.7	7.7	13.6
WI [14]	62.7	15.5	24.8	63.3	19.2	29.5	63.3	21.7	32.3	66.6	16.1	25.9	66.6	19.8	30.5	66.6	21.9	33.0
DWI [15]	64.3	15.4	24.8	64.8	19.8	30.4	64.9	23.5	34.5	67.2	16.3	26.2	67.5	21.6	32.7	67.6	25.4	36.9
RT [16]	59.1	12.1	20.1	60.9	21.6	31.9	60.4	27.5	37.8	49.2	5.8	10.4	36.0	4.9	8.6	45.1	10.0	16.4
AMP [17]	57.5	16.7	25.8	54.4	18.8	27.9	51.9	18.9	27.7	58.6	14.5	23.2	58.4	16.3	25.5	57.1	17.2	26.4
SPN [18]	59.8	16.3	25.6	60.8	26.3	36.7	58.4	33.4	42.5	49.8	8.1	13.9	56.4	10.4	17.6	61.6	16.3	25.8
LWF [8]	61.5	10.7	18.2	63.6	18.9	29.2	59.7	30.9	40.8	42.1	3.3	6.2	51.6	3.9	7.3	59.8	7.5	13.4
ILT [19]	64.3	13.6	22.5	64.2	23.1	34.0	61.4	32.0	42.1	43.7	3.3	6.1	52.2	4.4	8.1	59.0	7.9	13.9
MIB [20]	61.0	5.2	9.7	63.5	12.7	21.1	65.0	28.1	39.3	43.9	2.6	4.9	51.9	2.1	4.0	60.9	5.8	10.5
PIFS [9]	60.9	18.6	28.4	60.5	26.4	36.8	60.0	33.4	42.8	64.1	16.9	26.7	65.2	23.7	34.8	64.5	27.5	38.6
Ours	61.8	19.5	**29.6**	63.3	27.6	**38.4**	62.2	35.5	**45.2**	65.0	19.2	**29.6**	65.6	25.9	**37.1**	66.1	28.5	**40.0**

TABLE II

THE EXPERIMENTAL RESULTS ON THE COCO DATASET.

Method	SS									MS								
	1-shot			2-shot			5-shot			1-shot			2-shot			5-shot		
	mIoU (%)			mIoU(%)			mIoU(%)			mIoU (%)			mIoU (%)			mIoU (%)		
	base	novel	HM	base	novel	HM	base	novel	HM	base	novel	HM	base	novel	HM	base	novel	HM
FT	41.2	4.1	7.5	41.5	7.3	12.4	41.6	12.3	19.0	38.5	4.8	8.5	40.3	6.8	11.6	39.5	11.5	17.8
WI [14]	43.8	6.9	11.9	44.2	7.9	13.5	43.6	8.7	14.6	46.3	8.3	14.1	46.5	9.3	15.5	46.3	10.3	16.9
DWI [15]	44.5	7.5	12.8	45.0	9.4	15.6	44.9	12.1	19.1	46.2	9.2	15.3	46.5	11.4	18.3	46.6	14.5	22.1
RT [16]	46.2	5.8	10.2	46.7	8.8	14.8	46.9	13.7	21.2	38.4	5.2	9.2	43.8	10.1	16.4	44.1	16.0	23.5
AMP [17]	37.5	7.4	12.4	35.7	8.8	14.2	34.6	11.0	16.7	36.6	7.9	13.0	36.0	9.2	14.7	33.2	11.0	16.5
SPN [18]	43.5	6.7	11.7	43.7	10.2	16.5	43.7	15.6	22.9	40.3	8.7	14.3	41.7	12.5	19.2	41.4	18.2	25.3
LWF [8]	43.9	3.8	7.0	44.3	7.1	12.3	44.6	12.9	20.1	41.0	4.1	7.5	42.7	6.5	11.3	42.3	12.6	19.4
ILT [19]	46.2	4.4	8.0	46.3	6.5	11.5	47.0	11.0	17.8	43.7	6.2	10.9	47.1	10.0	16.5	45.3	15.3	22.9
MIB [20]	43.8	3.5	6.5	44.4	6.0	10.6	44.7	11.9	18.8	40.4	3.1	5.8	42.7	5.2	9.3	43.8	11.5	18.2
PIFS [9]	40.8	8.2	13.7	40.9	11.1	17.5	42.8	15.7	23.0	40.4	10.4	16.5	40.1	13.1	19.8	41.1	18.3	25.3
Ours	43.1	9.4	**15.4**	42.0	12.0	**18.7**	44.2	16.8	**24.3**	41.2	11.1	**17.5**	40.9	14.5	**21.4**	42.5	19.4	**26.6**

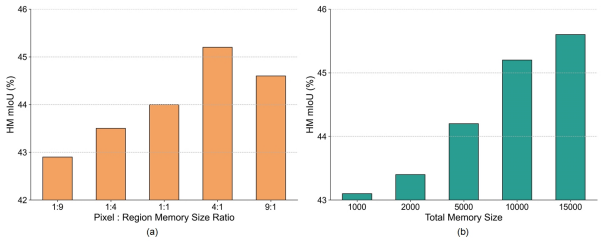

Fig. 4. Parameter selection on 5-shot SS of the PASCAL VOC 2012 dataset to investigate the effectiveness of (a) pixel-region ratio and (b) total memory size.

Some qualitative results of 1-shot segmentation are shown in Fig. 2, where each row represents the image, prediction of PIFS, our prediction, and ground-truth, respectively. As the figure shows, the proposed method provides more precise segmentation masks than PIFS. As shown in Fig. 3, the learned pixel embeddings by ours become more compact and well separated, suggesting that our method shapes a well-structured semantic feature space by employing pixel-region contrastive learning.

TABLE III

COMPARISON OF DIFFERENT DYNAMIC MEMORY DESIGNS ON 1-SHOT SS OF THE PASCAL VOC 2012 DATASET.

Memory	VOC-SS 1-shot		
	mIoU-B	mIoU-N	HM
Baseline (w/o contrast)	60.9	18.6	28.4
Mini-Batch (w/o memory)	62.7	18.4	28.5
Pixel Memory	61.2	18.9	28.9
Pixel + Region Memory	61.8	19.5	**29.6**

TABLE IV

COMPARISON OF DIFFERENT HARD EXAMPLE SAMPLING STRATEGIES ON 1-SHOT SS OF THE PASCAL VOC 2012 DATASET.

Sampling		VOC-SS 1-shot		
Anchor	Pos./Neg.	mIoU-B	mIoU-N	HM
Baseline (w/o contrast)		60.9	18.6	28.4
Random	Random	61.2	18.6	28.5
	Semi-Hard	61.0	18.8	28.7
Seg.-aware hard	Random	61.4	19.2	29.3
	Semi-Hard	61.8	19.5	**29.6**

C. Ablation Study

Dynamic Memory Design. The design of the dynamic memory is firstly validated. The results are summarized in Table III. Based on "Baseline (w/o contrast)", a variant, "Mini-Batch w/o memory", is first derived, in which we only compute pixel contrast loss within each mini-batch, without extra memory. It achieves 28.5% HM mIoU. This variant is then provided with pixel and region memories separately, leading to consistent performance gains ($28.5\% \rightarrow 28.9\%$ for pixel memory and $28.5\% \rightarrow 29.6\%$ for joint memory). This reveals i) the effectiveness of the dynamic memory design; and ii) necessity of comprehensively considering both pixel-to-pixel contrast and pixel-to-region contrast.

Sampling Strategy. Table IV presents a comprehensive examination of sampling strategies proposed in II-C. Our main observations are the following: i) For positive/negative sampling, mining hard-predicted pixels rather than "random" sampling, is indeed useful; ii) semi-hard sampling is more favored, as it improves the robustness of training by avoiding overfitting outliers in the training set; and iii) For anchor sampling, "seg.-aware hard" strategy further improves the performance over "random" sampling only. This suggests that exploiting task-related signals in supervised metric learning can help develop better segmentation models.

Memory Ratio and Size. In Fig. 4, we investigate the optimal selection of two key parameters: the pixel-region ratio and the overall memory size. In experiment (a), the total memory size is fixed at 10,000, while experiment (b) maintains a 4:1 pixel-region ratio, with the number of positive/negative samples scaling proportionally to changes in the total memory size. The results demonstrate that allocating more capacity to pixel embeddings rather than region embeddings is advantageous. Meanwhile, increasing the overall memory size consistently improves the performance, affirming the effectiveness of preserving more knowledge. However, an excessively large memory leads to high computational costs during training. Therefore, after comprehensive consideration of computational efficiency and effectiveness, we chose 10,000 as the total memory size.

IV. CONCLUSION

In this paper, we present a novel approach for incremental few-shot semantic segmentation, integrating pixel-wise and region-wise memory-based contrastive learning with optimized sampling strategies. Our key contributions include the design of a region

memory and pixel memory, effectively capturing the dynamically evolving semantic embedding space during incremental learning. We introduce pixel-region contrastive learning as a novel regularization strategy, along with optimized anchor and example sampling, to ensure robustness. The proposed method achieves new state-of-the-art performance, as demonstrated by comprehensive experiments.

V. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61976159, in part by the Shanghai Innovation Action Project of Science and Technology under Grant 20511100700, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[3] C. Wang, Y. Zhang, M. Cui, P. Ren, Y. Yang, X. Xie, X.-S. Hua, H. Bao, and W. Xu, "Active boundary loss for semantic segmentation," in *Proc. AAAI*, Feb. 2022, pp. 2397–2405.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NIPS*, Dec. 2021, pp. 12 077–12 090.

[5] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. NIPS*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., Dec. 2016, pp. 3630–3638.

[6] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. CVPR*, Jun. 2018, pp. 1199–1208.

[7] T. Zhang and W. Huang, "Few-shot classification with shrinkage exemplars," *CoRR*, vol. abs/2305.18970, May 2023.

[8] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[9] F. Cermelli, M. Mancini, Y. Xian, Z. Akata, and B. Caputo, "Prototype-based incremental few-shot segmentation," in *Proc. BMVC*, Nov. 2021, pp. 155–164.

[10] G. Shi, Y. Wu, J. Liu, S. Wan, W. Wang, and T. Lu, "Incremental few-shot semantic segmentation via embedding adaptive-update and hyper-class representation," in *Proc. ACM MM*, Oct. 2022, pp. 5547–5556.

[11] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[12] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. ECCV*, vol. 8693, Sep. 2014, pp. 740–755.

[13] T. T. Cai, J. Frankle, D. J. Schwab, and A. S. Morcos, "Are all negatives created equal in contrastive instance discrimination?" *CoRR*, vol. abs/2010.06682, Oct. 2020.

[14] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *CoRR*, vol. abs/1803.02999, Mar. 2018.

[15] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proc. CVPR*, Jun. 2018, pp. 4367–4375.

[16] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Proc. ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Aug. 2020, pp. 266–282.

[17] M. Siam and B. N. Oreshkin, "Amp: Adaptive masked proxies for few-shot segmentation," in *Proc. ICCV*, Oct. 2019, pp. 5248–5257.

[18] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero and few-label semantic segmentation," in *Proc. CVPR*, Jun. 2019, pp. 8256–8265.

[19] U. Michieli and P. Zanuttigh, "Incremental learning techniques for semantic segmentation," in *Proc. ICCV Workshops*, Oct. 2019, pp. 3205–3212.

[20] F. Cermelli, M. Mancini, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *Proc. CVPR*, Jun. 2020, pp. 9230–9239.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, Jun. 2017.

[22] R. Ferjaoui, M. A. Cherni, F. Abidi, and A. Zidi, "Deep residual learning based on resnet50 for covid-19 recognition in lung ct images," in *Proc. CoDIT*, May 2022, pp. 407–412.