

Incremental Few-Shot Semantic Segmentation via Embedding Adaptive-Update and Hyper-class Representation

Guangchen Shi
shiguangchen@hhu.edu.cn
Collage of Computer and Information
Hohai University
Nanjing, China

Yirui Wu*
wuyirui@hhu.edu.cn
Collage of Computer and Information
Hohai University
Nanjing, China

Jun Liu
junliu@sutd.edu.sg
Information Systems Technology and
Design Pillar
Singapore University of Technology
and Design
Singapore

Shaohua Wan
shaohua.wan@uestc.edu.cn
Shenzhen Institute for Advanced
Study
University of Electronic Science and
Technology of China
Shenzhen, China

Wenhai Wang
wangwenhai@pjlab.org.cn
Shanghai AI Laboratory
Shanghai, China

Tong Lu
lutong@nju.edu.cn
National Key Lab for Novel Software
Technology
Nanjing University
Nanjing, China

ABSTRACT

Incremental few-shot semantic segmentation (IFSS) targets at incrementally expanding model’s capacity to segment new class of images supervised by only a few samples. However, features learned on old classes could significantly drift, causing catastrophic forgetting. Moreover, few samples for pixel-level segmentation on new classes lead to notorious overfitting issues in each learning session. In this paper, we explicitly represent class-based knowledge for semantic segmentation as a category embedding and a hyper-class embedding, where the former describes exclusive semantical properties, and the latter expresses hyper-class knowledge as class-shared semantic properties. Aiming to solve IFSS problems, we present EHNet, i.e., Embedding adaptive-update and Hyper-class representation Network from two aspects. First, we propose an embedding adaptive-update strategy to avoid feature drift, which maintains old knowledge by hyper-class representation, and adaptively update category embeddings with a class-attention scheme to involve new classes learned in individual sessions. Second, to resist overfitting issues caused by few training samples, a hyper-class embedding is learned by clustering all category embeddings for initialization and aligned with category embedding of the new class for enhancement, where learned knowledge assists to learn new knowledge, thus alleviating performance dependence on training data scale. Significantly, these two designs provide representation capability for classes with sufficient semantics and limited biases, enabling to perform segmentation tasks requiring high semantic dependence. Experiments on PASCAL-5ⁱ and COCO datasets show that EHNet achieves new state-of-the-art performance with remarkable advantages.

KEYWORDS

incremental learning, few-shot learning, semantic segmentation, adaptive update, hyper-class representation

1 INTRODUCTION

Few-shot semantic segmentation [21, 28, 40] addresses to segment a new category of images with few samples, decreasing the cost of expensive pixel-level annotations. In a real-world scenario, we expect the trained model to segment new classes without forgetting knowledge learned from old classes, which is a natural task for human beings. However, fine-tuning a deployed model with few samples of new classes leads to a severe catastrophic forgetting problem [31], since models tend to forget knowledge about old classes when facing representation conflict between old and new classes as shown in Fig. 1(a). The gap between humans and machine learning models inspires researchers to facilitate incremental few-shot segmentation (IFSS), which aims to learn a segmentation model for both old and new classes with only few new samples.

The main challenges in IFSS are catastrophic forgetting of already acquired knowledge and overfitting networks to few samples of new classes. Most current incremental methods [7, 11, 24, 49] use parameters or embedding vectors to represent category knowledge from a cognitive-inspired perspective, which addresses catastrophic forgetting by knowledge representation with new-class updating scheme. However, representation error for old classes would accumulate iteratively, since they couple knowledge learning and representation in each updating iteration as shown in Fig. 1(b), thus inevitably hindering to maintain useful and consistent knowledge learned from old classes.

In this paper, we propose EHNet, i.e., Embedding adaptive-update and Hyper-class representation Network for IFSS, addressing problems of catastrophic forgetting and overfitting. We learn two kinds of embedding vectors, i.e., category and hyper-class embedding, from few samples of a new class, where the former describes exclusive semantical properties, and the latter expresses hyper-class knowledge as class-shared semantic properties. One key benefit of such knowledge representation is to replace requirement of new

*corresponding author.

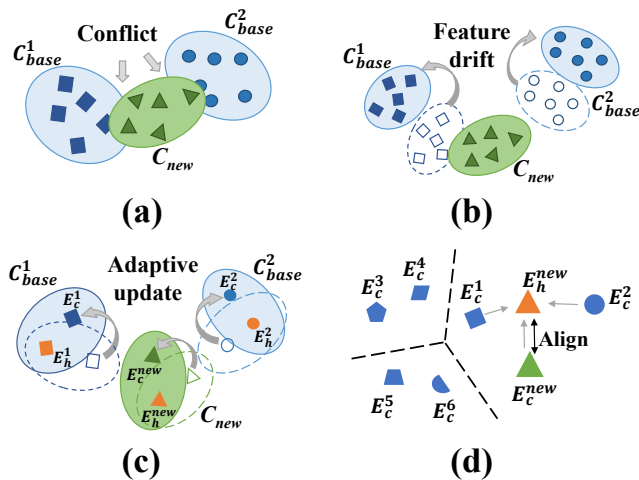


Figure 1: (a) In few-shot semantic segmentation, knowledge representation of base classes C_{base} and new classes C_{new} often conflicts, resulting in catastrophic forgetting problem. (b) Due to coupling between knowledge learning and representation, feature embeddings would drift to mismatch the true distribution of base classes. (c) To resist feature drift, EAUS maintains hyper-class embeddings E_h to store old knowledge with representation and adaptively updates category embeddings E_c to combine new class. (d) Hyper-class embedding E_h^{new} is generated by firstly clustering category embeddings corresponding to base classes (in blue) and then aligning with category embedding of new class E_c^{new} .

parameters training with prediction on fixed-length semantic embeddings, thus preventing training from scratch when learning new classes.

To mitigate catastrophic forgetting, we propose an Embedding Adaptive-Update Strategy (EAUS) as shown in Fig. 1(c), where category embeddings are adaptively adjusted with an attention scheme and hyper-class embeddings remain unchanged. In this manner, a well-separated representation of classes is built, where old knowledge is well maintained in memory functional design, i.e., hyper-class. EAUS decouples knowledge learning and representation to solve feature drift via selectively update, and alleviates the requirement of new-class sample number by keeping old knowledge. The core of category embeddings updating is a class-attention scheme, which computes a safe displacement vector for each class by contextualizing individual class weights over the representations of all classes. This adapted class-attention scheme not only highlights the discriminative representations between base and new classes to generate better decision boundaries over all involved classes, but also indicates directions towards a well-separated representation with less semantic biases during incremental learning sessions.

Observation for semantical segmentation proves to segment unseen classes with knowledge, where newly discovered samples may share semantic properties like 'haired' and 'quadruped' with the classes that have been learned. The hyper-class is thus formulated as an abstract representation that contains semantic properties of

similar classes, which enables to reduce data-scale dependence and overfitting by sharing semantic knowledge during learning sessions. A hyper-class embedding is learned by clustering category embeddings of all classes for initialization and aligning with the new-class category embedding for enhancement, which is shown in Fig. 1(d). On the one hand, the clustering algorithm is applied to the set of all category embeddings to generate a raw hyper-class embedding, thus extracting a similar semantic representation as new hyper-class knowledge. On the other hand, we align the generated hyper-class embedding with new-class category embedding to enhance correlated semantic information and eliminate uncorrelated information.

Significantly, EAUS builds a well-separated representation with few semantic biases. Meanwhile, the hyper-class knowledge complements and enhances semantic information. These two designs provide representations for class with sufficient semantics but limited biases, thus enabling to well perform image segmentation tasks requiring high semantic dependence.

In summary, the contributions of this paper are:

- We propose an embedding adaptive-update strategy to avoid catastrophic forgetting, where hyper-class embeddings remain fixed to maintain old knowledge, and category embeddings are adaptively updated with a class-attention scheme, combining new classes learned in incremental sessions.
- To resist overfitting caused by few training samples, a hyper-class is firstly learned by clustering category embeddings and then aligned with new-class category embedding for correlation enhancement, thus alleviating performance dependence on training data scale.
- Experimental results show EHNNet achieves state-of-the-art performance with remarkable advantages.

2 RELATED WORK

2.1 Semantic Segmentation

Semantic segmentation aims to classify each pixel of an image into a set of preset categories. According to different network structures, current methods can be roughly divided into three categories, i.e., CNN-based, RNN-based and GNN-based methods. CNN-based methods [23, 29, 50] utilize convolution operations to extract semantic information from feature maps for pixel-level label prediction. Considering dependence of context information, RNN-based methods [16, 34, 44] use recurrent layers to capture local and global spatial structure information of images. Using topological structure of graphs, GNN-based methods [30, 32, 42] transform task of image segmentation into the classification task of graph nodes.

Among them, CNN-based methods have received more popularity. For instance, to reconstruct high-resolution prediction images, UNet [38] and its variant [1] use an encoder-decoder structure to segment images. The encoder downsamples the feature map to obtain a large field of view and capture abstract feature representations, while the decoder gradually restores fine-grained information. Considering information loss caused by pooling, Chen et al. [5] propose DeepLab, where dilated convolution is used to enlarge receptive fields while maintaining the resolution of feature maps.

2.2 Few-Shot learning

Few-shot learning aims to learn a model, which can be easily transferred to new tasks with limited training data. We roughly divide current few-shot learning methods into two categories, i.e. initialization based and metric learning based methods. The former methods [6, 10, 20, 36] generally define few-shot learning as "learning to fine-tune", which aims to learn proper model initialization or predict network parameters. For example, MAML [10] explicitly trains the parameters of model to produce good generalization performance, which is easy to perform another task with few training samples and gradient updating steps. To get quick convergence within a few updates, Ravi et al. [36] propose a LSTM-based meta-learner model with general initialization, specially designing for a few-shot learning scenario.

Metric learning based methods [12, 41, 43] firstly learn a projection function that projects the inputs to an embedding space, and then define a certain distance metric that measures the distance between any two embeddings. For example, Siamese Network [17] compares samples in query set and support set by calculating similarity between their extracted feature vectors, thus performing few-shot classification based on known category labels.

Facing the problem of domain shifts between training and test datasets, Yuan et al. [47] propose a novel forget-update module, which could improve the discrimination by learning to forget and generate new features based on each task. Different from current few-shot methods, EHNet utilizes old knowledge to help learn new knowledge, where new classes could share the semantic features of base classes, thus alleviating the dependence on data scale when learning new classes.

2.3 Incremental learning

Incremental learning [22, 35] studies how to extend the knowledge of a model without a performance drop on old knowledge. Previous works can be roughly grouped in two categories [18], i.e., replay-based and regularization-based methods.

In replay-based methods, samples of previous tasks are either stored or generated at first and then replayed when learning the new task. For example, Li et al. [25] propose to jointly learn new labels and base classes from the outputs of a pre-trained teacher model. Since it's not capable to properly distinguish between old and new classes, later approaches [2, 19] utilize the memory of old classes for further training by considering distribution relationship between base and new classes.

Regularization-based methods [8, 11, 25] protect old knowledge from being covered by imposing constraints on new tasks. For example, to regularize the learning of new classes, Ren et al. [37] propose Attention Attractor Network, which utilizes old weights to train a set of new weights that could recognize new classes. To make classifiers learned on individual sessions suitable for all classes, Zhang et al. [49] propose a continually evolved classifier, which utilizes a graph model to propagate context information between classifiers for progressively adaptation.

Recently, incremental learning has been extensively studied for image segmentation task [13, 33, 45]. However, previous works in incremental segmentation focused on new classes that come with

rich samples. In this paper, we explore the incremental segmentation task with few-shot setting.

3 TASK DESCRIPTION

Incremental few-shot segmentation (IFSS) aims to generate a model that learns to segment new classes from few new samples without forgetting knowledge about old classes. IFSS has several learning sessions that come in sequence. Once the learning of the model steps into the next session, the training datasets in previous sessions are no longer available. Meanwhile, evaluation in each session involves classes of all previous sessions and the current session.

Specifically, let $\{D_s^0, D_s^1, \dots, D_s^n\}$ denote the support sets of different learning sessions, and the corresponding label space of dataset D_s^i is denoted by C^i . Different datasets have no overlapped classes, i.e. $\forall i, j$ and $i \neq j$, $C^i \cap C^j = \emptyset$. At the i -th learning session, only D_s^i can be used for network training. For evaluation, the query set D_q^i at session i includes test data from all previous and current classes, i.e., the label space is $C^0 \cup C^1 \dots \cup C^i$. Usually, the support set D_s^0 in the first session is a large dataset, where sufficient samples are available for training. On the contrary, support sets in all following sessions only include few training samples.

4 METHODOLOGY

4.1 Overview

Facing problems of catastrophic forgetting and overfitting in IFSS, we propose EHNet, mainly including representation of hyper-class knowledge and an Embedding Adaptive-Update Strategy (EAUS). Hyper-class knowledge could provide additional semantic information, which alleviates the dependence on data scale during learning new classes, thus avoiding overfitting caused by few training samples. EAUS keeps hyper-classes fixed as memory to maintain old knowledge, and adaptively updates category embeddings with a class-attention scheme to obtain a well-separated representation with few semantic biases. EHNet can be divided into two stages, i.e., incremental few-shot learning stage on support sets and segmentation stage on query sets, as shown in Figure 2.

In incremental few-shot learning stage, we first learn two embeddings, i.e., raw category embedding E_c^{raw} and raw hyper-class embedding E_h^{raw} , from a training sample. E_c^{raw} describes exclusive semantic properties, and E_h^{raw} expresses hyper-class knowledge as class-shared semantic properties. Since E_h^{raw} is obtained by clustering on base classes, there are semantic biases between E_c^{raw} and E_h^{raw} . To eliminate the semantic biases, E_c^{raw} and E_h^{raw} are semantically aligned through Cross Information Module (CIM), thus generating a category embedding E_c and a hyper-class embedding E_h . Finally, E_c and E_h are stored in memory pool, saving semantic embeddings to keep memory of the learned classes, where category embeddings of both base and new classes are updated via EAUS to obtain a well-separated representation with few semantic biases.

In segmentation stage, to match the aforementioned approach that keeps memory of base classes by saving embeddings, we propose Class-Agnostic Segmentation Module (CASM) as shown in Figure. 2 (b), which segments each class based on the corresponding semantic embeddings in memory pool.

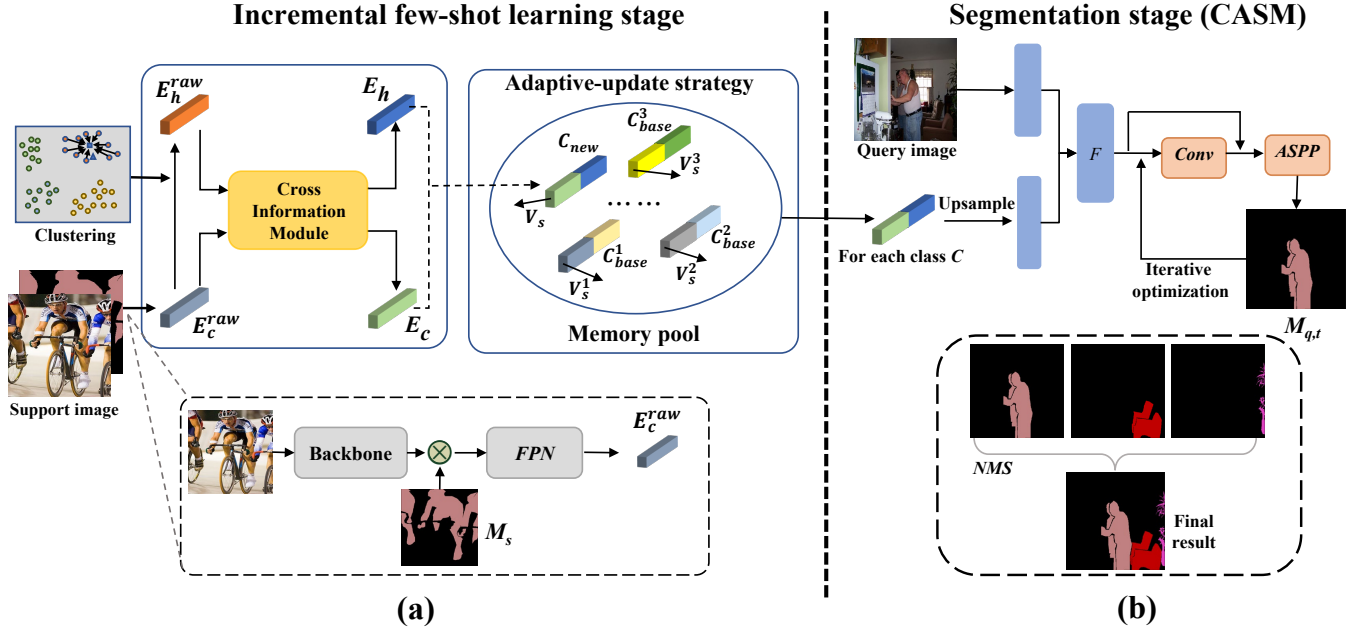


Figure 2: The design of EHNet. (a) In incremental few-shot learning stage, a support image is represented as a hyper-class embedding E_h and a category embedding E_c . E_h and E_c are stored in memory pool, where category embeddings of all classes are adaptively updated to obtain a well-separated representation. (b) In segmentation stage, objects of each class are segmented, based on corresponding E_h and E_c . Those segmentation results are integrated by non-maximum suppression (NMS) to generate the final result.

4.2 Semantic Embedding

To reduce the overfitting issues caused by few training samples, we use category embedding and hyper-class embedding to jointly describe feature representation of a class. The prediction network has not seen a new class, while the new class may share semantic properties with base classes it has seen. For instance, the network has never seen tigers before. However, many typical attributes of tigers can be found from base classes (e.g., cat, leopard, lion). We use hyper-class embedding to represent the similar attributes of base classes, thus assisting the network to understand tigers by reducing the dependence on training data scale.

The generation of category embedding and hyper-class embedding is shown in Figure 2 (a). First, a support image is input into backbone network to generate feature maps. Then, the feature maps are multiplied by a binary mask M_s to remove irrelevant background, which outputs feature representation only containing target objects. Afterwards, the feature representation is input into Feature Pyramid Network (FPN) [26], which extracts high-level semantic information and generates a raw category embedding E_c^{raw} . Hereafter, we perform clustering on base classes and obtain a raw hyper-class embedding E_h^{raw} based on E_c^{raw} . Finally, Cross Information Module (CIM) semantically aligns E_h^{raw} and E_c^{raw} to generate a hyper-class embedding E_h and a category embedding E_c , which could enhance category-relevant information and eliminate irrelevant information in hyper-class embeddings.

In k -shot learning, embeddings of new classes are updated via our EAUS during learning multiple samples. Owing to multiple updates,

features of the k samples are fused, which generates an enhanced embedding representation with more semantic information. Details of fusing multiple samples are described in Section 4.3.

Category Embedding. Category embeddings describe exclusive semantical property of classes. Considering variances of target objects in size, we use feature pyramid network (FPN) [26] to extract semantic information in different levels. By global pooling, the output of FPN is compressed into a 512-dimensional feature vector as a raw category embedding.

Hyper-class Embedding. Hyper-class embedding expresses hyper-class knowledge as shared semantic properties, which is generated by clustering on base classes and aligning semantic information with the corresponding new-class category embedding. Specifically, for a new class, our network searches for similar base classes through K -means clustering, which computes the center of all category embeddings as a raw hyper-class embedding for the new class. The raw hyper-class embedding is then aligned in semantic information with the new-class category embedding, thus enhancing correlated semantic information and eliminating irrelevant information. Owing to hyper-class embedding, new classes can share the semantic information of base classes, where old knowledge could help learn new knowledge, thus reducing the dependence on training data scale.

It's noted the number of selected similar base classes has an impact on hyper-class knowledge. A small number of similar classes can't guarantee to provide sufficient shared semantic information,

meanwhile a large number of similar classes could bring noise by introducing irrelevant classes.

Cross Information Module. Since hyper-class embeddings generated by a untrainable clustering algorithm may not semantically match the new-class category embedding, we propose cross information module (CIM) to achieve semantic alignment between hyper-class and category embedding, thus enhancing category-relevant information and eliminating irrelevant information in hyper-class embeddings.

The design of CIM is shown in Figure 3. Firstly, raw hyper-class embedding E_h^{raw} and raw category embedding E_c^{raw} are sent to a pair of two-layer fully-connected (FC) layers, respectively. The Sigmoid activation function attached after the FC layer transforms vector values into importance weights of channels. Afterwards, the embedding vectors in two branches are fused by element-wise multiplication. Intuitively, only similar semantic features would own a high activation value in the fused vector. Finally, we adopt the fused vector to weight the raw embedding vectors, thus generating enhanced embedding representations. In comparison to the raw embeddings, the enhanced embeddings focus on the correlated semantic information, thus obtaining semantic alignment between the new-class hyper-class and category embeddings.

4.3 Embedding Adaptive-Update Strategy

To resist catastrophic forgetting, we design Embedding Adaptive-Update Strategy (EAUS), which selectively and adaptively updates embeddings of base and new classes, thus mitigating feature drift and generating well-separated embedding representation for all classes.

Specifically, we keep hyper-class embeddings consistent as memory to maintain old knowledge in learning sessions, thus reducing feature drift effect. First, we compute a relation coefficient $e_{i,j}$ between class i and class j based on their category embeddings E_c^i and E_c^j :

$$e_{i,j} = \langle \Phi(E_c^i), \Psi(E_c^j) \rangle \quad (1)$$

where $\Phi(\cdot)$ and $\Psi(\cdot)$ are linear transformation functions that project the original representations to a new metric space, and $\langle \cdot, \cdot \rangle$ is a similarity function that computes the inner product between two embedding vectors.

Then, we normalize all the coefficients with a softmax function to obtain the attention weight $a_{i,j}$ of two classes:

$$a_{i,j} = \text{softmax}(e_{i,j}) = \frac{\exp(e_{i,j})}{\sum_{l=1}^{|P|} \exp(e_{i,l})} \quad (2)$$

where $|P|$ represents the number of classes in memory pool.

Afterwards, we perform subtraction operation on the embedding vectors of class i and j to obtain a subtraction vector, which locally indicates the updating direction. Finally, Based on the normalized attention weight and the corresponding subtraction vector, we calculate a safe displacement vector V_s for each category embedding. Considering information of all classes, V_s indicates directions towards a well-separated representation with less semantic biases during learning sessions. Therefore, the category embedding E_c^i can be adaptively updated as $E_c^{i'}$ by guidance of V_s :

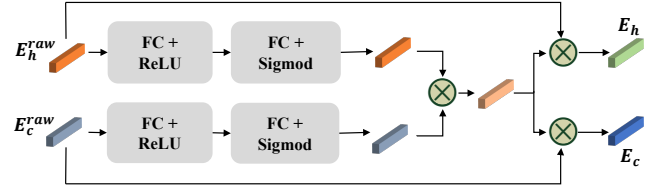


Figure 3: The design of CIM. Given raw hyper-class embedding E_h^{raw} and raw category embedding E_c^{raw} , CIM generates updated category embedding E_c and hyper-class embedding E_h , which achieves semantic alignment between these two embeddings.

$$E_c^{i'} = E_c^i + \sum_{l=1}^{|P|} a_{j,l} W(E_c^i - E_c^l) \quad (3)$$

where $W(\cdot)$ is a linear transformation. It's noted that we repeat the operations above to update category embeddings of all classes.

In k -shot learning, we directly apply embeddings of new samples to update the learned embeddings in EAUS, thus fusing semantic information from multiple samples without an additional fusion module. Specifically, for a class i , we define subtraction vectors in k -shot learning as $E_c^i - E_c^j$ with different class j , and $E_c^i - E_c^l$ with the same class l . It's noted that we force $E_c^i - E_c^j$ and $E_c^i - E_c^l$ to separate and fuse embeddings, thus forming well-separated feature representation and fuse enhancement of same-class semantic information, respectively. In fact, embeddings of the same class generally obtain a larger relation coefficient due to similar values, thus achieving an effective fusion effect in EAUS.

4.4 Class-Agnostic Segmentation Module

Class-Agnostic Segmentation Module (CASM) segments each class based on corresponding embeddings in memory pool, without considering what class an embedding represents. The design of CASM is shown in Figure 2 (b). First, a query image is processed by backbone network to generate feature maps, which contain target objects of different classes. For each target class, we perform an up-sampling operation on the semantic embeddings, and concatenate them with the feature maps to obtain feature maps F for dense comparison. After concatenation, we process feature maps via a 3×3 convolution block with a residual connection. Afterwards, we adopt an atrous spatial pyramid pooling module proposed in [5] to capture multi-scaled information and output the segmentation result, which is iteratively optimized to obtain compact boundary predictions. The process can be formulated as:

$$M_{q,t+1} = f_A(F \oplus \text{conv}(F + M_{q,t})) \quad (4)$$

where \oplus represents element-wise addition, $+$ represents concatenation operation, function $\text{conv}(\cdot)$ refers to convolution operation, and $f_A(\cdot)$ represents the process of atrous spatial pyramid pooling module. $M_{q,t}$ and $M_{q,t+1}$ refer to the predicted masks from current iteration step and the next iteration step, respectively.

Finally, for each pixel p , class with the highest confidence is selected as output class label by non-maximum suppression (NMS) algorithm:

$$NMS(p) = \arg \max_{c \in C} H(p, c) \quad (5)$$

where C represents the set of all learned classes, and $H(p, c)$ represents the probability that the pixel p belongs to the class c .

It's noted that backbone networks in learning stage and segmentation stage are the same in structure and parameters, since it's necessary to represent support and query images in the same feature space for few biases. Specifically, we use the first four layers of Resnet-50 [15] pre-trained on base dataset D_s^0 as backbone network.

5 EXPERIMENTS

5.1 Dataset and Metric

PASCAL-5ⁱ [39] is a widely used few-shot segmentation dataset, which consists of PASCAL VOC 2012 [9] and additional annotations in SDS [14]. The dataset includes 20 categories, which are divided into 4 splits, and each split contains 5 categories.

COCO 2014 [27] is a challenging large-scale dataset containing 80 categories. COCO is designed for natural scene understanding by acquiring data from complex daily scenes, where targets in images are segmented by precise pixel-level labels. The original dataset contains 82783 training images and 40504 validation images.

Based on the task description of incremental few-shot segmentation, we make special settings on datasets. On *PASCAL-5ⁱ*, the first split is considered as a base dataset, each class thus containing sufficient samples. Other splits are used for incremental learning in different sessions, each class only containing few samples for new-class training. On *COCO 2014*, we divide the 80 classes into 4 splits, and each split contains 20 classes. Similarly, the first split is a base dataset, and other splits are incremental few-shot datasets.

Following [39], we measure the per-class foreground Intersection-over-Union (IoU) and use the mean IoU over all classes (mIoU) to report the results.

5.2 Implementation Details

We implement EHNet using Pytorch library, and train it for 200 epochs on four Nvidia 1080Ti GPUs. We adopt cross-entropy loss function to evaluate the segmentation loss, and use StepLR scheduler in training, which reduces the learning rate (initial value as 0.0001) to 0.9 times for every 20 epochs of training. We evaluate performance of EHNet by setting $k = 1$ and 5 shots per new class. To ensure reliable results with random sample selection in k -shots learning, we run all tests 10 times and report the mean result for comparisons.

5.3 Comparisons with Other Methods

Since there exists few incremental few-shot segmentation methods for comparisons, we modify current few-shot segmentation and incremental few-shot learning methods for IFSS task. For few-shot segmentation methods, i.e., CANet [48] and ARNet [40], we save embeddings of classes for new-class segmentation, meanwhile we

either non-update or apply EAUS to update embeddings for comparisons. For incremental few-shot learning method, i.e., AAN [37] and XtarNet [46], we apply CASM to them for goal of performing segmentation. Besides, we compare EHNet with several incremental segmentation methods, e.g., MiB [3] and EMNet [45].

PASCAL-5ⁱ. Comparison results on *PASCAL-5ⁱ* are shown in Table 1, where "INC" means the embeddings of classes are saved for incremental learning and keep unchanged without updating. Table 1 shows EHNet is superior to other methods, achieving new state-of-the-art performance.

As the number of learned classes increases, all methods face a great challenge in learning new classes without forgetting base classes. It's noted that the performance of few-shot segmentation methods (CANet[48] and ARNet[40]) without updating drops significantly, due to the conflicts between feature representations of new and base classes. Equipped with EAUS, few-shot segmentation methods greatly improve their performance by comparing performance without updating, which proves that EAUS significantly mitigates catastrophic forgetting by maintaining hyper-class embeddings as old knowledge and adaptively updating category embeddings as source of new-class knowledge. The incremental segmentation methods (MiB[3] and EMNet[45]) perform poorly in few-shot settings, since they generally require large number of samples for new-class learning.

The improvement in segmentation performance with different learning sessions proves that incremental few-shot learning methods (AAN[37] and XtarNet[46]) could reduce catastrophic forgetting by iteratively learning new-class knowledge. However, due to feature drift in knowledge representation and non-separated embeddings with insufficient semantics, they have limited ability in segmentation task that requires high semantics for accurate pixel-level labeling. PIFS couples prototype learning and knowledge distillation for IFSS, while its representation error for old classes would accumulate iteratively, thus inevitably hindering to maintain useful and consistent knowledge learned from old classes.

Moreover, comparison methods fail to model commonalities between old and new classes. In other words, they can't utilize old knowledge to help better learn new-class embedded knowledge, which is proved by poor segmentation performance on new classes. On the contrary, EHNet introduces hyper-class embeddings as old knowledge to share semantic properties of base classes with new-class learning, which alleviates the dependence on training samples scale of new classes and thereby avoids overfitting issues.

COCO. Comparison results on *COCO* dataset are shown in Table 2, where EHNet achieves the best performance on *COCO* dataset as well. Regarding that *PASCAL-5ⁱ* only contains 20 categories and *COCO* contains 80 categories, increasing in category number requires a high distinguish capability of models to compute pixel-level predictions, which can be proved by general decrease in performance when comparing results on *PASCAL-5ⁱ* and *COCO* achieved by the same method. As shown in Table 2, performance of those methods without strategies to update embeddings drops significantly when learning session increases. Such phenomenon can be explained by the fact that more classes involved in *COCO* result in more obvious conflicts in embedding representation, thus causing a significant catastrophic forgetting effect. Meanwhile, experimental

Table 1: Comparison results of 1-shot and 5-shot segmentation on PASCAL-5ⁱ dataset. Best in bold.

Method	1-shot							5-shot								
	session-0		session-1		session-2		session-3		session-0		session-1		session-2		session-3	
	New	Base	New	Base	New	Base	New	New	Base	New	Base	New	Base	New		
MiB [3]	48.3	25.4	27.2	19.4	20.3	16.2	12.9	52.7	30.8	31.3	29.6	25.7	21.6	19.6		
EMNet [45]	50.8	21.8	18.0	16.8	13.5	10.7	10.9	53.5	36.3	32.3	25.4	24.9	19.6	15.8		
CANet [48] +INC	51.5	33.1	38.8	26.2	28.4	19.5	23.9	55.5	37.2	39.6	31.3	34.6	21.6	25.2		
CANet [48] +EAUS	52.1	46.6	43.1	37.2	39.6	32.6	34.6	54.5	49.3	46.3	39.4	40.9	35.6	36.4		
ARNet [40] +INC	54.0	33.4	40.4	27.7	31.3	22.6	25.8	55.9	40.3	38.8	30.9	33.1	24.4	22.1		
ARNet [40] +EAUS	54.8	47.9	48.2	41.1	42.9	37.0	36.9	56.4	49.0	47.2	43.1	43.8	38.5	40.2		
AAN [37] +CASM	48.2	44.6	42.6	38.5	35.8	33.8	34.2	49.1	45.3	44.3	39.4	37.6	34.6	35.5		
XtarNet [46] +CASM	47.4	43.7	41.8	38.7	37.2	34.7	32.0	50.7	45.1	44.9	40.9	41.2	37.8	36.8		
PIFS [4]	53.9	48.1	46.2	43.6	41.2	38.2	37.4	54.2	49.2	47.5	43.9	42.6	40.1	39.4		
EHNet (ours)	56.7	51.4	53.2	46.3	46.8	40.9	41.8	57.1	53.4	55.2	50.5	51.2	44.6	45.7		

Table 2: Comparison results of 1-shot and 5-shot segmentation on COCO dataset. Best in bold.

Method	1-shot		5-shot	
	Base	New	Base	New
MiB [3]	15.1	17.0	18.4	19.4
EMNet [45]	10.7	11.8	15.8	18.9
CANet [48] +INC	15.6	13.2	19.7	14.9
CANet [48] +EAUS	23.3	28.1	24.8	30.1
ARNet [40] +INC	13.7	11.6	10.6	11.4
ARNet [40] +EAUS	22.8	26.1	27.7	30.1
AAN [37] +CASM	27.2	29.5	28.4	29.7
XtarNet [46] +CASM	26.1	28.6	25.7	26.8
PIFS [4]	28.8	31.4	29.7	31.8
EHNet (ours)	29.7	33.1	33.4	36.6

results show that our EAUS can effectively mitigate catastrophic forgetting by a smaller drop in performance.

Qualitative results. Some qualitative results of 1-shot segmentation are shown in Figure 4, where each row represents the support set, query set, prediction, and ground-truth, respectively. From left to right, we show the segmentation results with four different sessions and some failure cases, respectively. With the increase of sessions, more classes are required to be segmented by EHNet, which brings difficulties in segmentation with new-class learning. From the failure examples, we can observe that segmentation of small objects or in dim environments is still challenging.

5.4 Ablation Study

To prove effectiveness of the proposed modules, we implement ablation experiments on PASCAL-5ⁱ dataset to report average mIOU performance of 4 splits with 1-shot setting.

Semantic embedding. To validate the effectiveness of constructing hyper-class embeddings (E_h), category embeddings (E_c), and the strategy to keep E_h and update E_c in EAUS, we compare experiment results by either removing or updating certain embeddings. Results are shown in Table 3, where “×”, “√” and “○” represent the embedding is removed, updated, and unchanged, respectively.

Table 3: Results achieved with different settings on semantic embeddings. Best in bold.

E_h	E_c	Base	New	Mean
√	×	22.7	25.3	24.0
×	√	35.3	30.6	33.0
√	√	41.6	46.4	44.0
√	○	45.4	44.9	45.2
○	√	46.2	49.6	47.9

Table 3 shows the elimination of certain embeddings would reduce the performance of EHNet, which means that each semantic embedding plays a positive role in segmentation. When both embeddings are updated, segmentation performance on base classes drops significantly, which proves that updating both embeddings leads to semantic feature drift and poor performance.

Embedding adaptively-update strategy. To validate the effectiveness of updating strategy in EAUS, we apply different embedding update strategies for comparisons, e.g., non-update strategy and trainable linear transformation (LT) strategy. To explore whether embeddings of base classes should be updated in EAUS, we adaptively update embeddings of either base class (C_{base}) or new class (C_{new}).

Results are shown in Table 4, where “√” presents corresponding embeddings are updated, and “Non-update” presents only storing embeddings without updates. Table 4 shows EHNet without embedding updating is unsatisfied in performance, indicating catastrophic forgetting problem occurs in incremental learning without intervention. Meanwhile, LT alleviates forgetting problem to a certain extent proved by relatively better performance, but it’s still limited in building well-separated representation. When EAUS is only applied to base or new classes, performance is degraded, indicating that globally modeling of embedding space for incremental learning is hardly achieved by only updating base or new classes.

Number of iterations in CASM. To validate the effectiveness of iterative optimization in CASM, we compare segmentation results and speed with different iteration numbers in Table 5. It’s noted the segmentation speed is measured by the number of processed frames per second (FPS). Table 5 shows segmentation performance gets

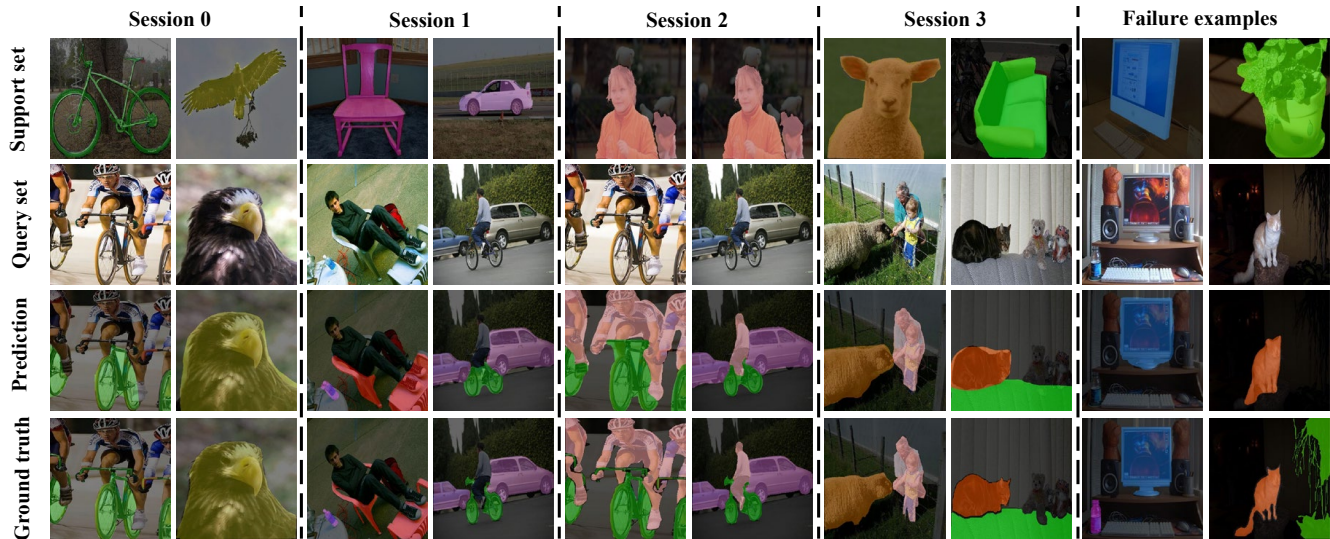


Figure 4: Qualitative results of EHNet, where we show the segmentation results in different sessions and some failure examples. The new classes in the current session will become base classes in subsequent sessions.

Table 4: Results with different strategies on updating embeddings. Best in bold.

Method	C_{base}	C_{new}	Base	New	Mean
Non-update	✓	✓	15.6	18.7	17.2
LT	✓		24.1	23.7	23.5
LT		✓	20.2	25.6	22.9
LT	✓	✓	25.2	27.3	26.3
EAUS (ours)	✓		44.1	34.8	39.5
EAUS (ours)		✓	36.5	47.9	42.2
EAUS (ours)	✓	✓	46.2	49.6	47.9

Table 5: Results of EHNet with different numbers of iterations. Best in bold.

Iterations	Speed	Base	New	Mean
0	24.7	42.7	45.7	44.2
1	16.4	44.1	46.9	45.5
2	12.4	45.0	48.6	46.8
3	9.8	45.7	49.5	47.6
4	8.2	46.2	49.6	47.9
5	6.9	46.4	48.6	47.5

better and speed gets slower with the increasing number of iterations. Being larger than 4 iterations, raising the number of iterations doesn't contribute to segmentation performance, which proves that 4 iterations in CASM keep a balance between performance and computation cost. Essentially, excessive iterations would lead our model to favor obvious objects with clear boundaries, thus harming segmentation of objects with less salience in query images.

6 CONCLUSION

In this paper, we focus on IFSS task and present EHNet. To avoid catastrophic forgetting, we propose an embedding adaptive-update strategy, where hyper-class embedding keeps unchanged as memory to maintain old knowledge, and category embeddings are adaptively updated to combine new classes learned on incremental sessions. To resist overfitting when learning with few samples of new classes, we learn hyper-class embeddings by clustering and aligning with category embeddings, where new classes could share semantic features of base classes, thus alleviating the dependence on training data scale. Comprehensive experiments show that EHNet achieves new state-of-the-art performance.

ACKNOWLEDGMENTS

This work was supported by Key Laboratory of Water Big Data Technology of Ministry of Water Resources, in part by a grant from National Key R&D Program of China under Grant No. 2021YFB3900601, the Fundamental Research Funds for the Central Universities under Grant B220202074, and National Research Foundation, Singapore underits AI Singapore Programme (AISG Award No: AISG-100E-2020-065).

REFERENCES

- [1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 12 (2017), 2481–2495.
- [2] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-End Incremental Learning. In *Proceedings of 15th European Conference on Computer Vision (ECCV)*, Vol. 11216. 241–257.
- [3] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. 2020. Modeling the Background for Incremental Learning in Semantic Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9230–9239.
- [4] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. 2021. Prototype-based Incremental Few-Shot Segmentation. In *32nd British Machine Vision Conference 2021, BMVC*. BMVA Press, 155.

- [5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. PAMI* 40, 4 (2018), 834–848.
- [6] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-shot Classification. In *Proceedings of 7th International Conference on Learning Representations (ICLR)*.
- [7] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtaash Harandi. 2021. Semantic-aware Knowledge Distillation for Few-Shot Class-Incremental Learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2534–2543.
- [8] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyang Wu, and Rama Chellappa. 2019. Learning Without Memorizing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5138–5146.
- [9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* 88, 2 (2010), 303–338.
- [10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. 1126–1135.
- [11] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. 2021. Incremental Few-Shot Instance Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1185–1194.
- [12] Xiuwen Gong, Dong Yuan, and Wei Bao. 2020. Online Metric Learning for Multi-Label Classification. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*. 4012–4019.
- [13] Yanan Gu, Cheng Deng, and Kun Wei. 2021. Class-Incremental Instance Segmentation via Multi-Teacher Networks. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*. 1478–1486.
- [14] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. 2014. Simultaneous Detection and Segmentation. In *Proceedings of 13th European Conference on Computer Vision (ECCV)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8695. 297–312.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
- [16] Li Kang, Ziqi Zhou, Jianjun Huang, and Wenzhong Han. 2022. Renal tumors segmentation in abdomen CT Images using 3D-CNN and ConvLSTM. *Biomed. Signal Process. Control* 72, Part (2022), 103334.
- [17] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. In *Proceedings of International Conference on Machine Learning Workshop*.
- [18] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars. 2019. Continual learning: A comparative study on how to defy forgetting in classification tasks. *CoRR abs/1909.08383* (2019).
- [19] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. 2019. Incremental Learning with Unlabeled Data in the Wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 29–32.
- [20] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. 2019. Meta-Learning With Differentiable Convex Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*. 10657–10665.
- [21] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. 2021. Adaptive Prototype Learning and Allocation for Few-Shot Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8334–8343.
- [22] Tianjiao Li, QiuHong Ke, Hossein Rahmani, Rui En Ho, Henghui Ding, and Jun Liu. 2021. Else-Net: Elastic Semantic Network for Continual Action Recognition from Skeleton Data. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 13414–13423.
- [23] Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, and Jian Sun. 2020. Learning Dynamic Routing for Semantic Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8550–8559.
- [24] Zhizhong Li and Derek Hoiem. 2018. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2018), 2935–2947.
- [25] Zhizhong Li and Derek Hoiem. 2018. Learning without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 12 (2018), 2935–2947.
- [26] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2017. Feature Pyramid Networks for Object Detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 936–944.
- [27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of 13th European Conference Computer Vision (ECCV)*, Vol. 8693. 740–755.
- [28] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. 2020. CRNet: Cross-Reference Networks for Few-Shot Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4164–4172.
- [29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3431–3440.
- [30] Yi Lu, Yaran Chen, Dongbin Zhao, and Jianxin Chen. 2020. Graph-FCN for image semantic segmentation. *CoRR abs/2001.00335* (2020).
- [31] Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, Vol. 24. Academic Press, 109–165.
- [32] Yanda Meng, Meng Wei, Dongxu Gao, Yitian Zhao, Xiaoyun Yang, Xiaowei Huang, and Yalin Zheng. 2020. CNN-GCN Aggregation Enabled Boundary Regression for Biomedical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention*, Vol. 12264. Springer, 352–362.
- [33] Umberto Michieli and Pietro Zanuttigh. 2021. Knowledge distillation for incremental learning in semantic segmentation. *Comput. Vis. Image Underst.* 205 (2021), 103167.
- [34] Fausto Milletari, Nicola Rieke, Maximilian Baust, Marco Esposito, and Nassir Navab. 2018. CFCM: Segmentation via Coarse to Fine Context Memory. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV*, Vol. 11073. Springer, 667–674.
- [35] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan M. Williams, and Jun Liu. 2021. Recent Advances of Continual Learning in Computer Vision: An Overview. *CoRR abs/2109.11369* (2021).
- [36] Sachin Ravi and Hugo Larochelle. 2017. Optimization as a Model for Few-Shot Learning. In *Proceedings of 5th International Conference on Learning Representations*.
- [37] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard S. Zemel. 2019. Incremental Few-Shot Learning with Attention Attractor Networks. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. 5276–5286.
- [38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, Vol. 9351. 234–241.
- [39] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. 2017. One-Shot Learning for Semantic Segmentation. In *Proceedings of BMVC*.
- [40] Guangchen Shi, Yirui Wu, Shivakumara Palaiahnakote, Umappa Pal, and Tong Lu. 2021. ARNet: Active-Reference Network for Few-Shot Image Semantic Segmentation. In *Proceedings of International Conference on Multimedia and Expo (ICME)*. 1–6.
- [41] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. 4077–4087.
- [42] Savannah Thais and Gage DeZoort. 2021. Instance Segmentation GNNs for One-Shot Conformal Tracking at the LHC. *CoRR abs/2103.06509* (2021).
- [43] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. 3630–3638.
- [44] Francesco Visin, Adriana Romero, Kyunghyun Cho, Matteo Matteucci, Marco Ciccone, Kyle Kastner, Yoshua Bengio, and Aaron C. Courville. 2016. ReSeg: A Recurrent Neural Network-Based Model for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR*. 426–433.
- [45] Shipeng Yan, Jiale Zhou, Jiangwei Xie, Songyang Zhang, and Xuming He. 2021. An EM Framework for Online Incremental Learning of Semantic Segmentation. In *ACM Multimedia*. 3052–3060.
- [46] Sung Whan Yoon, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. 2020. XtarNet: Learning to Extract Task-Adaptive Representation for Incremental Few-Shot Learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 119)*. 10852–10860.
- [47] Minglei Yuan, Chunhao Cai, Tong Lu, Yirui Wu, Qian Xu, and Shijie Zhou. 2022. A novel forget-update module for few-shot domain generalization. *Pattern Recognit.* 129 (2022), 108704.
- [48] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. 2019. CANet: Class-Agnostic Segmentation Networks With Iterative Refinement and Attentive Few-Shot Learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5217–5226.
- [49] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. 2021. Few-Shot Incremental Learning With Continually Evolved Classifiers. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12455–12464.
- [50] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaoqiang Wang, Amrith Tyagi, and Amit Agrawal. 2018. Context Encoding for Semantic Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7151–7160.