Incremental Few Shot Semantic Segmentation via Class-agnostic Mask Proposal and Language-driven Classifier

Leo Shan
University of Chinese Academy of Sciences
Beijing, China
shanlianlei18@mails.ucas.edu.cn

Wenzhang Zhou
University of Chinese Academy of Sciences
Beijing, China
zhouwenzhang19@mails.ucas.ac.cn

Grace Zhao
University of Chinese Academy of Sciences
Beijing, China
zhaoguiqin20@mails.ucas.edu.cn

ABSTRACT

Incremental Few-Shot Semantic Segmentation (IFSS) aims to extend pre-trained segmentation models to new classes with limited annotated images without accessing old training data. During incrementally learning novel classes, the data distribution of old classes will be corrupted, leading to catastrophic forgetting. Meanwhile, the samples of the new class are limited, making it impossible for the model to learn a satisfactory representation of the new class. Previous IFSS methods are mainly based on distillation or storing old data. In this paper, we propose a new IFSS framework called CaLNet, i.e., **C**lass-**a**gnostic mask proposal and **L**anguage-driven classifier incremental few-shot semantic segmentation network. Specifically, CaLNet employs a class-agnostic mask proposal, and due to its class-agnostic nature, the capabilities of mask proposals can be easily extended from base classes to novel classes. As a result, incremental learning is only needed in the classifier part. Meanwhile, when incrementally learning novel classes, it is challenging for the classifier to learn a complete representation of the new classes due to the limited number of samples. Based on this, we combine the language embedding into the visual features, making the expression of the new class complete. Results on Pascal-VOC and COCO show that CaLNet achieves a new SOTA.

CCS CONCEPTS

• **Computing methodologies** → **Image segmentation**.

KEYWORDS

Semantic segmentation, incremental learning, few-shot learning

ACM Reference Format:
Leo Shan, Wenzhang Zhou, and Grace Zhao. 2023. Incremental Few Shot Semantic Segmentation via Class-agnostic Mask Proposal and Language-driven Classifier. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611783

1 INTRODUCTION

Incremental Semantic Segmentation (ISS) [3, 12] can continuously learn new classes without losing the ability to segment old classes, thus improving the adaptability of models. In the real world, we hope the model can learn new classes with only a few samples due to the cost of expensive pixel-level annotations for semantic segmentation tasks. The task of incrementally learning new classes via few-shots is called Incremental Few-Shot Semantic Segmentation (IFSS). IFSS needs to solve the problems of overfitting and insufficient representation of novel classes in FSS, as well as the catastrophic forgetting [34] in ISS.

The current IFSS methods are mainly based on knowledge distillation, pseudo label, and generating or storing old data to avoid feature drift and overfitting to new classes. PIFS [5], the first method in the IFSS field, uses a prototype-based distillation method, which has a significant impact on the feature distribution of previously learned classes when learning new classes, so it is not an excellent solution. EHNet [37] needs to save the data of the old category when incrementally learning new categories and requires higher

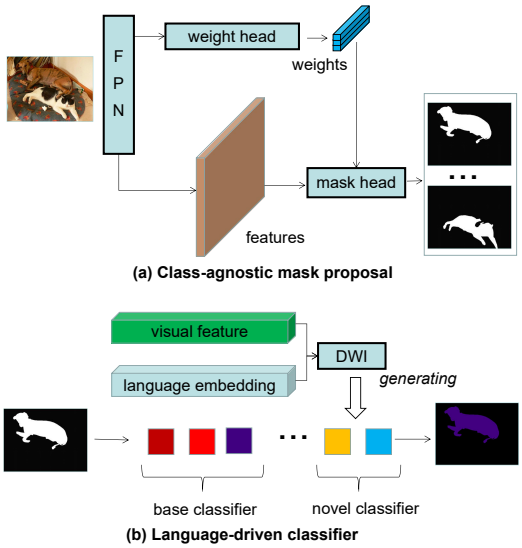

Figure 1: Schematic illustration of our proposed method. (a) represents the class-agnostic mask proposal. The overall structure is based on CondInst [42], and uses dynamically generated convolution kernels to output class-agnostic masks. (b) represents the language-driven classifier. DWI stands for dynamic weight imprinting. In addition to limited visual features, Language embedding is also added and combined via dynamic weights to obtain a complete expression of the novel classes.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611783

conditions such as support sets. GAPS [32] uses guided copy-and-paste synthesis, but it also needs to store data and relies heavily on the synthetic model. SRAA [58] proposes the semantic-guided relation, alignment, and adaptation method, but it requires very complex operations on the feature level, and the improvement is not apparent, which is not conducive to practical use.

On the contrary, in Incremental Few-shot Instance Segmentation (IFIS) and Incremental Few-shot Object Detection (IFOD), due to the existence of class-agnostic Region Proposal Network (RPN), the problem of few-shot and incremental learning only occurs on the classifier head, which greatly simplifies the training process. This paradigm transforms IFIS and IFOD tasks into incremental classification tasks. Since they are all classification tasks, the idea of mature Incremental Few-shot Learning (IFL) can be easily transferred. Inspired by the success of IFIS and IFOD, we propose a Class-agnostic Mask Proposal and Language-driven Classifier Incremental Few-shot Semantic Segmentation Network (CaLNet). The overall structure is shown in Figure 1. First, the class-agnostic mask proposal will extract various masks. This procedure does not involve classes, so incremental learning is not required. Then, the extracted masks are assigned classes. To solve the problems of overfitting and insufficient representation ability brought about by few-shot learning of novel classes, a language-driven classifier is proposed, as shown in Figure 1 (b). We introduce the class-agnostic mask proposal and language-driven classifier, respectively.

Class-agnostic mask proposal: Current instance segmentation proposes the mask through the convolutional center-based representation, i.e., using features from a single point in the feature map [43, 44]. However, unlike instance segmentation, there are both countable things and uncountable stuff in semantic segmentation. And stuff is difficult to extract based on convolutional center-based representation. In semantic segmentation (requiring output classes) tasks, in order to classify regions accurately, models must take into account the semantics of other distant neighboring regions (both things and stuff), which requires a larger receptive field and a more global context. Due to the central representational locality, it performs poorly when larger receptive fields and more global context are required, thereby degrading class-specific task performance.

However, it is demonstrated in Entity Segmentation [31] (ES), which is a segmentation task that does not require output classes, that although the center-based representation provides a smaller receptive field and more local context, it achieves great success. The Segment Anything Model (SAM) [18] that has received a lot of attention recently does not output classes, but it can accurately extract all masks of the image. These successes are due to all sub-regions (even spatially distant sub-regions) within a filled region that tends to share similar colors and textures with others around their central location. For example, a sky and a forest often share the same color and texture within their respective regions. Moreover, due to the smaller receptive field and more local context, the generalization of the class-agnostic mask proposal is very strong. The class-agnostic network is trained on the COCO dataset [2, 22] and can achieve incredible performance on the ADE20K [55] dataset [31]. This cross-dataset evaluation fully demonstrates the generalization ability of class-agnostic segmentation models. In a nutshell, under the condition of the class-agnostic setting, the mask proposal can achieve the same generalization performance as the region proposal in IFIS and IFOD, which provides the prerequisite for the class-agnostic proposal to be used in IFSS.

Language-driven classifier: Since we decouple the mask proposal and the classifier, and the class agnostic mask proposal has strong generalization, incremental learning only appears in the classification part. In the previous IFIS and IFOD, the average feature of the novel class is directly used as the classifier [13], which is effective for instance segmentation and object detection but is suboptimal for semantic segmentation. Semantic segmentation contains stuff, and the stuff of the same class varies greatly, which makes it challenging for limited images to obtain a complete and comprehensive class expression. In addition, the success of CLIP [33], referring expression segmentation [16, 23, 45], and open vocabulary segmentation [38, 47] shows that language embedding and visual features contain strong semantic relationships. Based on this, we employ the very easy-to-obtain language-driven embedding as a compensation input to generate a novel classifier, making the representation of the novel class more complete.

In summary, our contributions are as follows:

- We abandon the previous IFSS paradigm and propose a class-agnostic mask proposal and language-driven classifier incremental framework, which achieves the current state-of-the-art results on the benchmark datasets.
- The class-agnostic mask proposal has strong generalization and is perfectly suitable for IFSS tasks, which makes the mask proposal not require incremental learning to simplify the IFSS network structure and the training procedure. Moreover, various strategies are proposed to make the extracted mask accurate, complete, and without overlap.
- To solve the insufficient expressive ability of the classifier brought about by few-shot learning, we use the easy-to-obtain language-driven word embedding as a supplementary input to generate a classifier, thereby improving the coverage of different parts of the stuff.

2 RELATED WORK

2.1 Incremental Few Shot Segmentation

Incremental few-shot segmentation (IFS) can be divided into Incremental Few-shot Semantic Segmentation (IFSS) and Incremental Few-shot Instance Segmentation (IFIS).

At present, IFSS is gaining more and more attention. PIFS [5] is the first work of IFSS, and it has established a very strong baseline. PIFS is based on prototypes combined with distillation and re-normalization methods and achieves outstanding performance compared to other few-shot semantic segmentation or incremental learning methods. [37] maintains old knowledge through a superclass, but it needs to store old knowledge and support images, and its setup is quite different from ours and PIFS. It can be seen that the previous works employ various methods to solve the insufficient representation ability of few-shot learning and catastrophic forgetting of incremental learning, i.e., they use distillation or storing old class data to solve catastrophic forgetting and use weight imprint to solve the insufficient representation ability of limited samples. We discard the previous distillation-based paradigm and

propose a framework that decouples mask proposal and classification. Meanwhile, with the super generalization characteristics of the class-agnostic mask proposal, the incremental learning is only in the classifier part, which makes the training very simple.

IFIS adds a mask head on the basis of IFOD. The IFOD method [20, 28] typically employs Faster RCNN [36] as its backbone network. In [28], knowledge distillation ensures that the prediction of the base class matches its pre-training prediction after fine-tuning on the new class. In [29], weights for box heads are dynamically generated based on class-specific codes extracted from target class examples. In this way, each class has a different box header for object detection. iFS-RCNN [27] also uses a different set of Mask-RCNN classification, bounding box, and segmentation heads for each class, where the differences between the heads are fine-tuned for each new class separately in the last layer. ONCE [29] proposes a class code generator to generate classification headers for new classes. Fine-tuning-based works also have been proposed to achieve more accurate performance. They exploit knowledge distillation loss to overcome catastrophic forgetting. However, their network structure and training method did not preserve the performance of the base class. Recent approaches replace the standard fully-connected classifier in Mask-RCNN with a cosine similarity classifier to implement IFIS. It can be seen that the basis of IFIS is RPN, which is not applicable to IFSS. We extend the idea of class-agnostic proposals to the field of IFSS and overcome the problem of stuff in semantic segmentation.

2.2 Incremental Few-Shot Learning

Incremental Few-shot Learning (IFL) is recently proposed to address the few-shot input problem in the class-incremental setting [56], [53]. TOPIC [40] defines a benchmark setting for IFL, using neural gas structures to preserve feature topology between old and new classes to resist forgetting. Exemplar relation graph [11] maintains a graph representing class relationships for knowledge distillation. FSLL [24] adopts the idea of a class-incremental learning algorithm, which selects some parameters to update in each incremental session to resist overfitting. Noting the characteristics of few-shot instances, semantic-aware knowledge distillation [8] considers using auxiliary word embeddings of few-shot instances to facilitate model update. The self-boosting prototype improvement [59] considers the single-stage scaling capability and adapts the feature representation to various generated incremental episodes. To avoid overfitting on few-shot inputs, FSL [24] selects some parameters to update in each incremental session. The current state-of-the-art method is CEC [51], which separates the training process into embedding learning and classifier learning. LIMIT [57] allows for multi-stage scaling capabilities. [14, 30, 49] propose a similar setting, namely generalized few-shot learning (GFSL). It also addresses scenarios where a pre-trained model will learn new classes with limited instances. Typical GFSL algorithms try to solve the problem by classifier weight generalization [48], subspace regularization [1], and attention-based regularization [35].

However, IFL focuses on single-class classification tasks, while IFSS has multiple classes in one image and also contains the background, so IFSS requires a special design.

2.3 Mask Proposal

Mask proposals can be classified into class-specific and class-agnostic. Class-specific mask proposals are often used in one-stage instance segmentation. Among them, CondInst [42] uses the convolutional center-based representation directly generated from regression to generate masks. Maskformer [7] combines this idea with the transformer, generates a mask, and then assigns the corresponding category to the mask. However, due to the locality of the center-based representation, this strategy works poorly for semantic segmentation containing stuff, which often results in incomplete masks.

The class-agnostic mask proposal appears in Entity Segmentation (ES) [31]. Open-world entity segmentation is of great significance to our work. In order to realize open-world image editing, entity segmentation is proposed based on the CondInst network structure. We take inspiration from ES [31] to extract accurate and complete masks. In general, both [31] and our work take advantage of the superior generalization of center-based representation on class-agnostic mask proposals.

2.4 Language-driven Classifier

Recently, using text as supervised information for segmentation has achieved great success in referring segmentation [16, 23, 45], and open vocabulary segmentation [38, 47]. The core idea is to use language embedding as a visual classifier, which shows that the same class of visual features and text embedding are in close relationships in high-dimensional space. And language embedding can be easily obtained through self-supervision methods [10]. This inspires us to complement the classifier with language embedding. The previous weight imprinting methods [14, 30] only use limited novel class features to generate new classifiers, which makes the obtained classifier overfit to the only samples. Language embedding can significantly alleviate the problem and improve accuracy.

3 TASK DEFINITION

Formally, we denote the set of semantic classes already learned after step t as $C^{0:t}$, where a learning step denotes an update of the model's output space. During training, the network receives a sequence of data $\left\{\mathcal{D}^0, \ldots, \mathcal{D}^T\right\}$, where $\mathcal{D}^t = \{(x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}^t\}$. In each $\mathcal{D}^t$, x is an image in the space $\mathcal{X} \in \mathbb{R}^{|I| \times 3}$, with I the set of pixels, and y its corresponding label mask in $\mathcal{Y}^t \subset \left(C^t\right)^{|I|}$. Note that $\mathcal{D}^0$ is a large dataset while $\mathcal{D}^t$ are few-shot ones, i.e. $\left|\mathcal{D}^0\right| \gg \left|\mathcal{D}^t\right|, \forall t \geq 1$. The model is first trained on the large dataset $\mathcal{D}^0$ and incrementally updated with few-shot datasets. We name the first learning step on $\mathcal{D}^0$ as the base step. Note that at step t, the model has access only to $\mathcal{D}^t$. For $\mathcal{D}^t$, if an image contains the class C^t, the image is selected. The difference is that except C^t, the other classes on the label are masked as the background.

Our goal is to learn a model ϕ^t that maps each pixel to a probability distribution over the set of classes, i.e., $\phi^t : \mathcal{X} \rightarrow \mathbb{R}^{|I| \times |C^{0:t}|}$, where t denotes the last learning step. We assume ϕ^t composed of a feature extractor $f^t : \mathcal{X} \rightarrow \mathbb{R}^{|I| \times d}$ and a classifier (i.e., MaskHead) $g^t : \mathbb{R}^{|I| \times d} \rightarrow \mathbb{R}^{|I| \times |C^{0:t}|}$, such that $\phi^t = g^t \circ f^t$. Here, d is the feature dimension and g^t is a classifier with parameters $\boldsymbol{W}^t = \left[\boldsymbol{w}_0^t, \ldots, \boldsymbol{w}_{|C^t|}^t\right] \in \mathbb{R}^{d \times |C^{0:t}|}$.

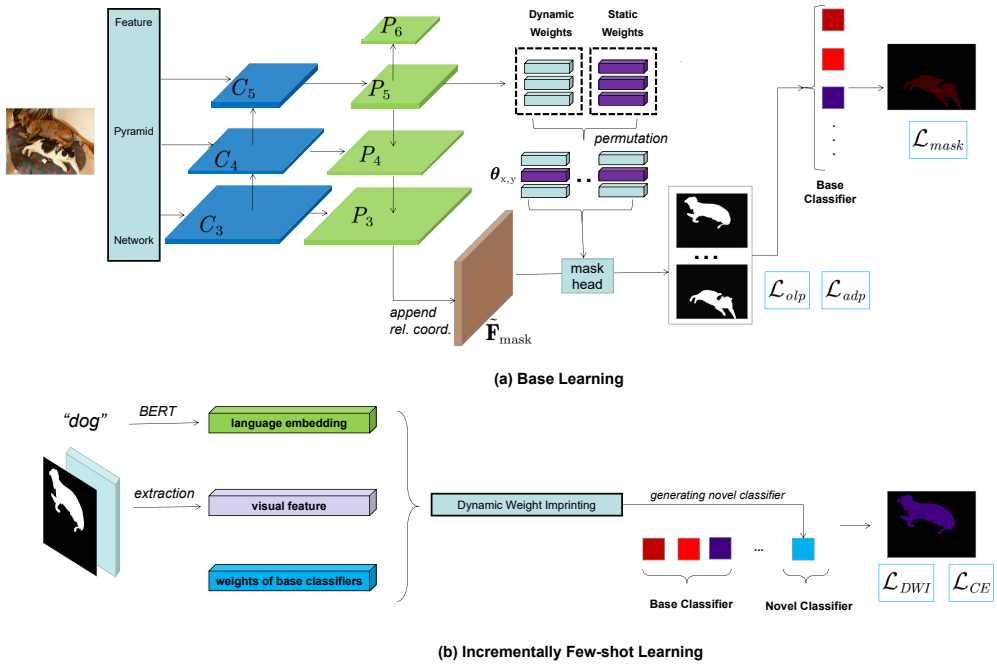

Figure 2: The overall architecture of the proposed framework. (a) is base learning, i.e, the process of learning base classes, and (b) is the process of incrementally few shot learning novel classes. The main structure of base learning is borrowed from CondInst [42], but the output is class-agnostic masks, and the classification part only contains the classifier of the base classes. During incrementally few shot learning, we use dynamic weight imprinting so that language embedding can supplement the classifier of novel classes.

4 METHOD

4.1 Overview

The overall procedure is shown in Figure 2. (a) represents base learning, i.e., the process of learning base classes. The class-agnostic mask proposal is built on the basis of CondInst [42]. Consistent with CondInst, the image is subjected to Feature Pyramid Network (FPN) to obtain features of different scales. The feature $\tilde{\mathbf{F}}_{\text{mask}}$ is obtained by appending the relative coordinate to the P_3 layer feature of the network (the feature of the largest spatial resolution layer). Then, inspired by [31], we get the dynamic filter of the mask proposal from P_5 through several layers of convolution and then form the comprehensive kernel bank with the static filters. Different kernels can get different masks, and these masks are post-processed to calculate the loss with the ground truth. The mask proposal part only updates the parameters during base learning. In Section 4.2, we introduce the training of class-agnostic mask proposal, including more specific structure, class-agnostic segmentation loss $\mathcal{L}_{mask}$, overlap limitation loss $\mathcal{L}_{ovl}$, and proposal adaptation loss $\mathcal{L}_{adp}$.

Figure 4.3 (b) shows the process of incremental few-shot learning, which only occurs in the classifier part. Different from the previous, we add a language embedding that can be obtained very easily, thus increasing the expressive power of the novel classifier. We introduce the process and loss of dynamic weight imprinting in Section 4.3. Losses of all stages are introduced in Section 4.4.

4.2 Class-agnostic Mask Proposal

Inspired by entity segmentation [31], we use not only dynamic convolutions but also static convolutions to generate masks. Among them, the static weight is consistent with that in [31]. As shown in Figure 2, there are three layers of convolution, and each layer can be dynamic or static, so there are a total of 8 permutations. Consistent with [31], we delete the combination with only static convolution so that for each position (x, y), seven kinds of features can be obtained. Note that we only use the branch with full dynamic convolutions during testing, and the rest are used to assist network training to increase the integrity and non-conflict of extracted masks. In order to get a complete and accurate mask without overlapping each other, class-agnostic segmentation optimization and overlap limitation are required, which will be introduced next.

Class-agnostic segmentation optimization: After obtaining the predicted masks, it is necessary to assign a label for each mask to calculate losses. The first is to modify the original Ground Truth (GT) with multi-classes into binary masks. Meanwhile, when matching GT and predictions, consistent with CondInst [42], each location on the feature map P_i of FPN is either associated with an instance and thus a positive sample or considered a negative sample. Different from CondInst, we also conduct certain operations on the background part. We use $P_i \in \mathbb{R}^{H\times W\times C}$ to denote the feature map. As shown in previous works [36, 42, 44], a location (x, y) on a feature map can be mapped back to the input image, and a mapped location

is considered responsible for an instance if it falls in the central region of the instance. Any location outside the central region is marked as a negative sample. The central region is defined as the box $(c_x - rs, c_y - rs, c_x + rs, c_y + rs)$, where (c_x, c_y) represents the centroid of the instance, s is the downsampling rate of P_i, and r is a constant scalar of 1.5. After obtaining the mask proposal and the corresponding GT, we use Dice loss for optimization, and Dice loss is shown as follows,

$$L_{\text{mask}}\left(\{\boldsymbol{\theta}_{x,y}\}\right) = \frac{1}{N_{\text{pos}}} \sum_{x,y} \mathbb{1}_{\{c^*_{x,y}>0\}} L_{dice}\left(\text{MaskHead}\left(\tilde{\mathbf{F}}; \boldsymbol{\theta}_{x,y}\right), \mathbf{M}^*_{x,y}\right), \quad (1)$$

where $c^*_{x,y}$ is the classification label of location (x, y), which is the class of the objects associated with the location or 0 (i.e., background) if the location is not associated with any objects. N_{pos} is the number of locations where $c^*_{x,y} > 0$. $\mathbb{1}_{\{c^*_{x,y}>0\}}$ is the indicator function, being 1 if $c^*_{x,y} > 0$ and 0 otherwise. Not operating on the background area is to keep the background features scattered so as to facilitate the learning (registration) of the novel class. $\boldsymbol{\theta}_{x,y}$ is the generated filters' parameters at location (x, y). $\tilde{\mathbf{F}}$ is the combination of features and coordinates. MaskHead denotes the mask head, which consists of a stack of convolutions with dynamic and static parameters. $\mathbf{M}^*_{x,y} \in \{0, 1\}^{H\times W\times C}$ is the mask of the objects associated with location (x, y). L_{dice} is the dice loss as in [26], which is used to overcome the foreground-background sample imbalance. Note that, in order to compute the loss between the predicted mask and the ground-truth mask $\mathbf{M}^*_{x,y}$, they are required to have the same size.

Overlap Limitation: In addition, each mask is prohibited from containing overlapping regions. However, dynamic kernels tend to produce overlapping masks with high confidence due to concept overlap between neighboring entities and independent losses of different kernels. While post-processing strategies [17, 19] can address mask overlap, they are driven by handcrafted heuristics and are less efficient. Instead, inspired by [31], we introduce an optimization that encourages the model to learn to suppress overlap between predicted entity masks. We obtain the representative mask $\mathbf{Q^n}$ of the n-th cluster by averaging all its masks generated via θ_7, i.e.,

$$\mathbf{Q^n} = \frac{\sum_{i\in\Omega(n)} \text{MaskHead}\left(\tilde{\mathbf{F}}; \theta^{7|i}\right)}{|\Omega(n)|}, \quad (2)$$

where $\Omega(n)$ returns the set of kernel indices belonging to the n-th regions and $\theta^{7|i}$ indicates the i-th kernel in θ_7. θ_7 is the bank with all dynamic kernels. In order to simplify the representation, we omit the subscript (x, y) of kernels θ and $\mathbf{M}^*$. Given the whole representative entity masks Q of all regions, we apply the pixel-wise softmax function to induce a strong suppression of non-maximal entities in the pixel-wise mask prediction, which is inspired by [31]. To achieve overlap suppression, we adopt a separate training loss $\mathcal{L}_{olp}$, i.e.,

$$\mathcal{L}_{olp} = L_{dice}(\text{Softmax}(\mathbf{Q}), \mathbf{M}^*). \quad (3)$$

$\mathcal{L}_{olp}$ calculates the loss with comprehensive possible masks, which is a strict penalty so that the masks do not overlap each other.

Adaption of mask proposal to background: The Class-agnostic setting allows the mask to be potentially extracted during base training so that a mask belonging to the same class can be extracted even if the class has no relevant supervision. This operation is similar to unsupervised clustering. Region Proposal Network (RPN) [36] in object detection has similar properties. Since the RPN is class-agnostic, during the learning process, the common features of the detected object can be learned, which allows the RPN to extract objects even if it never encounters corresponding classes. On the other hand, $\mathcal{L}_{mask}$ only involves the base classes, and $\mathcal{L}_{olp}$ covers all classes, including the background, thus can extend the ability of class-agnostic mask proposal to novel classes when learning base classes. Furthermore, we specifically design a loss for proposal generalization, and the process of $\mathcal{L}_{adp}$ in the background is as follows,

$$\mathcal{L}_{adp} = -\sum_{\substack{i,j\in\Omega(b)\\ i\neq j}} L_{dice}\left(\text{MaskHead}\left(\tilde{\mathbf{F}}; \theta^{7|i}\right), \text{MaskHead}\left(\tilde{\mathbf{F}}; \theta^{7|j}\right)\right) \quad (4)$$

where $\Omega(b)$ denotes the background regions. L_{adp} makes no overlap between the disjoint background masks, i.e., each kernel corresponds to a fixed background mask, and different kernels do not interfere with each other. This operation is similar to a clustering operation so that the ability of the mask proposal can be easily expanded, as demonstrated in SAM [18] and others [54].

4.3 Language-driven Classifier

The great success of segmentation models based on natural language and visual relations [33] naturally inspired us to use language embedding as a supplement to generate classifiers. The overall architecture is shown in Figure 2 (b). First, the text words of the classes need to be converted into language embedding. We use well-pre-trained BERT [10] to generate embedding. Compared with other works, due to the huge amount of data, BERT can generate embedding that covers more contexts, which can largely make up for the lack of expressiveness of classifiers brought about by few-shot learning. After obtaining the language-driven embedding, the visual feature filtered by the mask is merged with it. Meanwhile, in order to make the generated novel classifier and the classifier of the base classes in the same space, we also add the classifier weights of the base class to the weight generation. Thus, the novel class classification head $\boldsymbol{w}_n$ is calculated by

$$\boldsymbol{w}_n = \boldsymbol{w}_f \odot \overline{\boldsymbol{f}}_n + \boldsymbol{w}_{att} \odot \boldsymbol{z} + \boldsymbol{w}_p \odot \boldsymbol{p}_{emb}, \quad (5)$$

where $\boldsymbol{z}$ is the output of the base class classification heads after the same attention as in [14], and $\overline{\boldsymbol{f}}_n$ denotes the mean visual features of the current learning novel class. Different from [14], the input newly adds the language embedding $\boldsymbol{p}_{emb}$. $\boldsymbol{w}_f, \boldsymbol{w}_{att}$, and $\boldsymbol{w}_p \in \mathbb{R}^d$ are learnable weight vectors. $\odot$ is the Hadamard product. In this way, the classifier weights of the novel class can be obtained completely and accurately.

4.4 Optimization Goal

The whole training process is divided into base training and incremental few-shot training.

4.4.1 Base Learning: During base training, the model needs to learn: 1) the ability of class-agnostic mask proposal, 2) the ability to generalize potential classes to be learned in the background, and 3) the ability to generate a classifier for the base classes. In order to meet these three requirements, the loss of base learning is as follows,

$$\mathcal{L}_{\text{Base}} = \mathcal{L}_{mask} + \lambda_1 \mathcal{L}_{olp} + \lambda_2 \mathcal{L}_{adp}. \tag{6}$$

Unlike CondInst [42], since we do not have the annotation of instance segmentation, we only use the loss of semantic segmentation. λ_1 and λ_2 are set to 1 by default.

4.4.2 Incremental Few-shot Learning: During the incremental learning process, the feature extraction part, the mask proposal part, and the classifier part of the base class will not be updated. The only part that needs to be trained is dynamic weight imprinting (DWI), i.e., the classifier of the novel class is obtained from DWI, so only the cross-entropy loss and the loss in DWI [14] are needed for the incremental few-shot learning, i.e.,

$$\mathcal{L}_{\text{Novel}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{DWI}}. \tag{7}$$

5 EXPERIMENTS

5.1 Datasets

We use Pascal-VOC 2012 (VOC) [52] containing 20 classes, and COCO [2, 22] where, as in previous works [5, 50], we use the 80 thing classes. We consider 15 and 60 of the classes as base (C^0) and 5 and 20 as new (novel) $(C^t \backslash C^0)$, for VOC and COCO respectively. The protocols start with pretraining on base classes and multiple steps on novel classes, i.e., 5 steps of 1 novel class on VOC and 4 steps of 5 novel classes on COCO. In line with PIFS, we divide VOC in 4 folds of 5 classes and COCO in 4 folds of 20 classes, running experiments 4 times by considering each fold in turn as the set of new classes. For each setting, we consider 1, 2, or 5 images in an incremental learning step, and we average the results over multiple trials, each using a different set of images. The images are randomly sampled from the set of images containing at least one pixel of the new classes without imposing any restriction on the existence of the old class. During incrementally learning novel classes, we only rely on the few images provided. In learning the base classes, consistent with previous incremental semantic segmentation works [3, 12], labels except base classes are marked as background, which best fits the actual world situation. We use cross-validation, i.e., 20% of the data in the training set is regarded as the real validation set. Finally, we report results on each dataset's entire validation set (for the test), considering all visible classes, which is consistent with previous works [5, 21].

5.2 Implementation Details

Our code is based on entity segmentation [31] and CondInst [42]. Compared with baseline, we keep all components except that we set the number of class to 1. In the inference phase, we first rank all entities according to their (aggregated) confidence scores: C_i (centrality) $\times$ E_i (entity probability), and derive their corresponding masks. Each entity is encoded into an entity-specific dynamic kernel, resulting in N kernels for N entities. With the resulting combined kernel bank and encoded features, the segmentation mask of N entities is directly produced by a series of convolutions. Finally, the final non-overlapping prediction map is obtained by selecting the entity ID with the largest confidence score at each pixel.

We follow [3], using SGD as optimizer with momentum 0.9, weight decay 10^{-4} and a polynomial learning rate policy, i.e. $\text{lr} = \text{lr}_{\text{init}} \left(1 - \frac{\text{iter}}{\max \text{iter}}\right)^{0.9}$. During training, we apply the same data augmentation of [6], performing random scaling and horizontal flipping with a crop size of 512×512. In line with PIFS, we use a different learning rate and training iterations depending on the dataset, the number of shots, and the learning steps. Specifically, during learning base classes, networks are trained for 30 epochs on Pascal-VOC and 20 epochs on COCO with a learning rate of 10^{-2} and batch size of 24. During incrementally learning step t, we set the batch size to min $(10, |\mathcal{D}_t|)$. In the incremental training of VOC, the network is trained for 200 iterations per step with a learning rate of 10^{-4}. In COCO, the network is trained for 100 iterations per step with a learning rate of 10^{-4}. These hyperparameters are shared by all methods. We compute the results via single-scale full-resolution images without any post-processing.

5.3 Comparison with State-of-the-art Methods

Following the protocol GFSS work [4] and PIFS, we evaluate the method's performances via three metrics based on the mean Intersection over Union (mIoU): mIoU on base classes (mIoU-Base), mIoU on novel classes (mIoU-Novel), and the Harmonic Mean (HM) of the two . We report the results after the last step as in [3, 25].

The comparison methods include fine-tuning (FT), few-shot classification (FSC), few-shot segmentation (FSS), incremental learning (IL), and the PIFS method for IFSS. FSC includes weight imprinting (WI) [30], dynamic imprinting (DWI) [14], and RT [41]. FSS includes adaptive masked proxies (AMP) [39] and semantic projection networks (SPN) [46]. IL methods include learning without forgetting (LwF) [21], incremental learning techniques (ILT) [25], and modeling the background (MiB) [3]. In addition, we make a comparison with Incremental Few-shot Learning (IFL) methods. We convert these methods from classification tasks to segmentation task and keep all contribution points. For IFL, SubReg [1] proposes a new subspace regularization scheme, which encourages the weight vectors of new classes to lie in the subspace spanned by the weights of existing classes. Const [15] exploits hyper-dimensional embeddings, allowing continuous representation of more classes than fixed dimensions in a vector space. MetaFSCIL [9] proposes a meta-learning-based two-layer optimization. FACT [56] contains the same motivation as ours, which reserves the embedding space for future new classes but needs to generate virtual prototypes. The three jobs are currently the most representative and most effective jobs of the IFL, and we compare with them.

5.3.1 Evaluation on PASCAL VOC Dataset. For novel classes, 1-shot, 2-shot, and 5-shot experiments are performed, respectively. The results are shown in Table 1. It can be seen that the FSC methods, especially WI and DWI, show very good results. And the effects on the base and novel classes are both competitive, which shows that the imprint method is very suitable for IFSS tasks. However, for the methods of FSS and IL, the performances on both base classes and novel classes are very poor. These manifestations are explainable. FSS methods do not have any measure for keeping

	Method	1-shot			2-shot			5-shot		
		mIoU-Base	mIoU-Novel	HM	mIoU-Base	mIoU-Novel	HM	mIoU-Base	mIoU-Novel	HM
	FT	47.2	3.9	7.2	53.5	4.4	8.1	58.7	7.7	13.6
FSC	WI [30]	66.6	16.1	25.9	66.6	19.8	30.5	66.6	21.9	33.0
FSC	DWI [14]	67.2	16.3	26.2	67.5	21.6	32.7	67.6	25.4	36.9
FSC	RT [41]	49.2	5.8	10.4	36.0	4.9	8.6	45.1	10.0	16.4
FSS	AMP [39]	58.6	14.5	23.2	58.4	16.3	25.5	57.1	17.2	26.4
FSS	SPN [46]	49.8	8.1	13.9	56.4	10.4	17.6	61.6	16.3	25.8
IL	LwF [21]	42.1	3.3	6.2	51.6	3.9	7.3	59.8	7.5	13.4
IL	ILT [25]	43.7	3.3	6.1	52.2	4.4	8.1	59.0	7.9	13.9
IL	MiB [3]	43.9	2.6	4.9	51.9	2.1	4.0	60.9	5.8	10.5
IFL	SubReg [1]	55.4	13.2	21.3	56.7	12.7	20.8	59.7	13.5	22.0
IFL	Const [15]	58.4	12.1	20.0	61.3	13.4	22.0	62.2	17.2	27.0
IFL	FACT [56]	57.0	14.6	23.2	57.4	15.1	23.9	58.8	15.2	24.2
	PIFS [5]	64.1	16.9	26.7	65.2	23.7	34.8	64.5	27.5	38.6
	Ours	**74.2**	**17.4**	**28.2**	**74.4**	**26.1**	**38.6**	**74.7**	**30.1**	**42.9**

Table 1: Results on VOC.

	Method	1-shot			2-shot			5-shot		
		mIoU-Base	mIoU-Novel	HM	mIoU-Base	mIoU-Novel	HM	mIoU-Base	mIoU-Novel	HM
	FT	38.5	4.8	8.6	40.3	6.8	11.7	39.5	11.5	17.8
FSC	WI [30]	46.3	8.3	14.0	46.5	9.3	15.4	46.3	10.3	16.8
FSC	DWI [14]	46.2	9.2	15.3	46.5	11.4	18.3	46.6	14.5	22.1
FSC	RT [41]	38.4	5.2	9.1	43.8	10.1	16.4	44.1	16.0	23.5
FSS	AMP [39]	36.6	7.9	13.1	36.0	9.2	14.6	33.2	11.0	16.5
FSS	SPN [46]	40.3	8.7	14.3	41.7	12.5	19.2	41.4	18.2	25.3
IL	LwF [21]	41.0	4.1	7.4	42.7	6.5	11.3	42.3	12.6	19.4
IL	ILT [25]	43.7	6.2	10.8	47.1	10.0	16.5	45.3	15.3	22.8
IL	MiB [3]	40.4	3.1	5.8	42.7	5.2	9.3	43.8	11.5	18.2
IFL	SubReg [1]	38.4	8.0	13.2	39.5	10.1	16.0	40.0	10.3	16.4
IFL	Const [15]	39.0	8.2	13.6	40.6	11.4	17.8	41.1	11.3	17.7
IFL	FACT [56]	37.9	8.6	14.0	38.9	11.7	18.0	39.4	12.3	18.7
	PIFS [5]	40.4	10.4	16.6	40.1	13.1	19.8	41.1	18.3	25.3
	Ours	**48.4**	**10.6**	**17.4**	**48.5**	**13.4**	**21.0**	**48.6**	**18.6**	**26.9**

Table 2: Results on COCO.

base class memory. While IL methods need large samples for novel classes to learn them, the limited setting of few-shot makes the its role unable to play. The IFL methods achieve a good balance between keeping the old class memory and learning new classes, but the overall effect is worse than FSC and FSS. Our method has a very remarkable advantage in keep the memory of the old classes, which is nearly 10 points higher than the previous method (PIFS). Similarly, on the premise of keep a strong memory of old classes, the proposed method is also very competitive in learning new classes. Therefore, our HM results are significantly improved over the all previous best methods.

5.3.2 Evaluation on COCO Dataset. For novel classes, 1-shot, 2-shot, and 5-shot experiments are performed, respectively. The results are shown in Table 1. It can be seen that the FSC methods, especially WI and DWI, show very good results. And the effects on the base and novel classes are both competitive, which shows that the imprint method is very suitable for IFSS tasks. However, for the methods of FSS and IL, the performances on both base classes and novel classes are very poor. These manifestations are explainable. FSS methods do not have any measure for keeping base class memory. While IL methods need large samples for novel classes to learn them, the limited setting of few-shot makes the its role unable to play. IFL methods also not show advantages in segmentation tasks. Compared with PIFS, our method has significant improvements

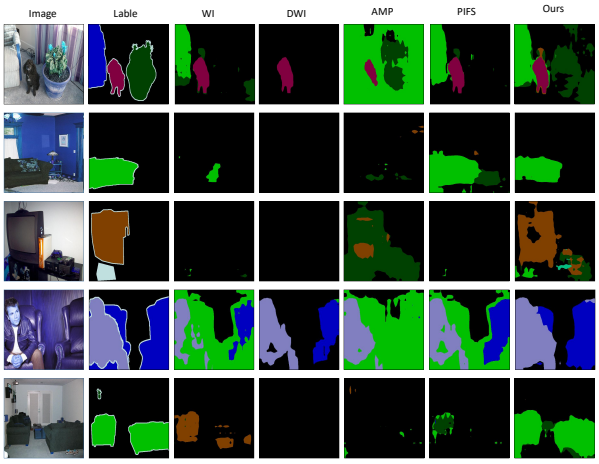

Figure 3: Visual results on the VOC 1-shot setting.

Agno.	$\mathcal{L}_{olp}$	$\mathcal{L}_{adp}$	Emb.	Base	Novel	HM
✓	✓	✓	✓	74.2	17.4	28.2
	✓	✓	✓	66.7	16.4	26.3
✓		✓	✓	73.1	16.9	27.5
✓	✓		✓	73.2	16.7	27.2
✓	✓	✓		74.3	16.4	26.9

Table 3: Ablation of the different components.

both in the memory of base classes and the learning of novel classes, demonstrating our method's effectiveness.

5.3.3 Visual Analysis: The visual results are shown in Figure 3. We show the results when the novel classes are the last five, i,e, potted plant, sheep, sofa, train, and tv monitor. As shown in the figure, WI and DWI with fixed representations have excellent memory for base classes. But they are very prone to assign pixels of novel classes to similar base classes or even to the background. Our method has a huge advantage in learning novel classes. Firstly, our method can extract new classes (as in the first line), and secondly, our method can segment the pixels of the new classes into the correct ones (as in line 2, 3, and 5). Meanwhile, our method also shows an advantage in keeping the memory of base classes like the fourth line of the picture. With segmentation, there are only base categories, and without overlap, everything is done.

5.4 Ablation study

In order to verify the effectiveness of each module, we conduct sufficient ablation experiments, including whether the mask proposal is class-agnostic (i.e., Agno. in Table 3), penalty on overlap $\mathcal{L}_{olp}$, adaption of proposal ability to the background $\mathcal{L}_{adp}$, and the embedding of the classifier (i.e., Emb.). Experiments are conducted on VOC dataset.

5.4.1 Class-agnostic **vs** *class-specific.* : The extracted mask is classified after the proposal, which differs from the structure in CondInst [42]. Without using class-agnostic, the decoder part of the network is the same as CondInst [42]. During the experiment, except for generating the base classifier during training, the rest of the settings, such as the generation of the novel classifier during incremental learning, remained unchanged. The results of the experiment are shown in the first and second rows in Table 3. It can be seen that without class-agnostic, the effect drops significantly. This result is consistent with our previous analysis, i.e., the class-specific setting makes the ability of the mask proposal very poor, thus making the effect of both the base class and the novel class poor.

5.4.2 The effect of overlap limitation. : The most common error in class-agnostic mask proposals is overlap [31, 42, 44]. To solve this problem, we use a way to make a limitation of it. The results With and without $\mathcal{L}_{olp}$ are shown in row 1 and row 3 of Table 3. After losing $\mathcal{L}_{olp}$, the base class accuracy drops due to the confusion between the masks. More importantly, the accuracy of the novel class has dropped greatly, which shows that $\mathcal{L}_{olp}$ also has a hidden effect on the mask extraction of the background class.

5.4.3 The effect of adaption to background. : During base training, although we cannot obtain the annotations of novel classes, we can obtain data that may contain novel classes. The novel class exists in the background, so we need to pre-extract the information in the background. The overlap operation on the mask obtained by the background-centered kernel is similar to clustering the background information, thereby enhancing the generalization of the mask proposal. The results of the experiment are shown in row 1 and row 4 of Table 3. The accuracy of base and novel classes has decreased to a certain extent, especially the novel class, which is mainly caused by the insufficient accuracy of the extracted mask.

5.4.4 The effect of Language Embedding. : Due to the limited training samples when learning a novel class, it is difficult for the learned classifier to cover the entire novel classes, so we use language embedding to supplement the classifier. The results of the experiment are shown in row 1 and row 5 of Table 3. After losing the language embedding, the accuracy of the novel class drops by one point.

6 CONCLUSION

In this work, according to the characteristics of IFSS, we decouple the mask extraction and classifier and propose a Class-agnostic mask proposal and Language-driven classifier incremental few-shot semantic segmentation network. The mask proposal has strong generalization characteristics under the premise of class agnostic, and our work is based on this. During base training, we propose multiple optimization functions to make the extracted masks complete, accurate, and non-conflicting. More importantly, the ability of the mask proposal can be transferred easily from the base class to the novel class. When learning novel classes, in order to make up for the lack of training samples for novel classes, we use language-driven embedding as a supplement and use dynamic weight imprinting for combination. Our idea of decoupling mask proposal and classification is worth learning. In future work, how to add the pre-trained model to the network to make the extracted mask more accurate is a research direction.

REFERENCES

[1] Afra Feyza Akyürek, Ekin Akyürek, Derry Tanti Wijaya, and Jacob Andreas. 2021. Subspace regularizers for few-shot class incremental learning. *arXiv preprint arXiv:2110.07059* (2021).

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2016. COCO-Stuff: Thing and Stuff Classes in Context. *computer vision and pattern recognition* (2016).

[3] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. 2020. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9233–9242.

[4] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. 2020. A few guidelines for incremental few-shot segmentation. *arXiv preprint arXiv:2012.01415* 2 (2020).

[5] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. 2020. Prototype-based Incremental Few-Shot Semantic Segmentation. *arXiv preprint arXiv:2012.01415* (2020).

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *european conference on computer vision* (2018).

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems* 34 (2021), 17864–17875.

[8] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. 2021. Semantic-aware Knowledge Distillation for Few-Shot Class-Incremental Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr46437.2021.00256

[9] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. 2022. MetaFSCIL: a meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14166–14175.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. 2022. Few-Shot Class-Incremental Learning via Relation Knowledge Distillation. *Proceedings of the AAAI Conference on Artificial Intelligence* (Sep 2022), 1255–1263. https://doi.org/10.1609/aaai.v35i2.16213

[12] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. 2020. PLOP: Learning without Forgetting for Continual Semantic Segmentation. *computer vision and pattern recognition* (2020).

[13] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. 2021. Incremental few-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1185–1194.

[14] Spyros Gidaris and Nikos Komodakis. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4367–4375.

[15] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. 2022. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9057–9067.

[16] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. 2022. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18145–18154.

[17] AlexanderM. Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. 2019. Panoptic Feature Pyramid Networks.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[19] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. 2021. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 214–223.

[20] Yiting Li, Haiyue Zhu, Jun Ma, ChekSing Teo, Cheng Xiang, Prahlad Vadakkepat, and TongHeng Lee. 2021. Towards Generalized and Incremental Few-Shot Object Detection.

[21] Zhizhong Li and Derek Hoiem. 2016. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016).

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *european conference on computer vision* (2014).

[23] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. 2023. PolyFormer: Referring Image Segmentation as Sequential Polygon Generation. *arXiv e-prints* (2023), arXiv–2302.

[24] Pratik Mazumder, Pravendra Singh, and Piyush Rai. 2022. Few-Shot Lifelong Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* (Sep 2022), 2337–2345. https://doi.org/10.1609/aaai.v35i3.16334

[25] Umberto Michieli and Pietro Zanuttigh. 2019. Incremental Learning Techniques for Semantic Segmentation. *international conference on computer vision* (2019).

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 565–571.

[27] Khoi Nguyen and Sinisa Todorovic. 2022. iFS-RCNN: An Incremental Few-shot Instance Segmenter.

[28] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. 2020. Incremental Few-Shot Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.01386

[29] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. 2020. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13846–13855.

[30] Hang Qi, Matthew Brown, and David G. Lowe. 2017. Low-Shot Learning with Imprinted Weights. *computer vision and pattern recognition* (2017).

[31] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. 2022. Open World Entity Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[32] Ri-Zhao Qiu, Peiyi Chen, Wangzhe Sun, Yu-Xiong Wang, and Kris Hauser. 2022. GAPS: Few-Shot Incremental Semantic Segmentation via Guided Copy-Paste Synthesis. (2022).

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[34] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2016. iCaRL: Incremental Classifier and Representation Learning. *computer vision and pattern recognition* (2016).

[35] Mengye Ren, Renjie Liao, Ethan Fetaya, and RichardS. Zemel. 2018. Incremental Few-Shot Learning with Attention Attractor Networks.

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[37] Guangchen Shi, Yirui Wu, Jun Liu, Shaohua Wan, Wenhai Wang, and Tong Lu. 2022. Incremental Few-Shot Semantic Segmentation via Embedding Adaptive-Update and Hyper-class Representation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5547–5556.

[38] Gyungin Shin, Weidi Xie, and Samuel Albanie. [n. d.]. ReCo: Retrieve and Co-segment for Zero-shot Transfer Supplementary Material. ([n. d.]).

[39] Mennatullah Siam, Boris Oreshkin, and Martin Jagersand. 2019. Adaptive masked proxies for few-shot segmentation. *arXiv preprint arXiv:1902.11123* (2019).

[40] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. 2020. Few-Shot Class-Incremental Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.01220

[41] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. 2020. Rethinking few-shot image classification: a good embedding is all you need?. In *European Conference on Computer Vision*. Springer, 266–282.

[42] Zhi Tian, Chunhua Shen, and Hao Chen. 2020. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 282–298.

[43] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9627–9636.

[44] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. 2020. Solo: Segmenting objects by locations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 649–665.

[45] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11686–11695.

[46] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. 2019. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8256–8265.

[47] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. 2022. A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-Language Model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*. Springer, 736–753.

[48] Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. 2021. Learning Adaptive Classifiers Synthesis for Generalized Few-Shot Learning. *International Journal of Computer Vision* 129, 6 (Apr 2021), 1930–1953. https://doi.org/10.1007/s11263-020-01381-4

[49] SungWhan Yoon, Doyeon Kim, Jun Seo, and Jaekyun Moon. 2020. XtarNet: Learning to Extract Task-Adaptive Representation for Incremental Few-Shot

Learning.

[50] 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. *computer vision and pattern recognition* (2019).

[51] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. 2021. Few-Shot Incremental Learning with Continually Evolved Classifiers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr46437.2021.01227

[52] Jianguo Zhang. 2006. The PASCAL Visual Object Classes Challenge.

[53] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. 2020. MgSvF: Multi-Grained Slow vs. Fast Framework for Few-Shot Class-Incremental Learning.

[54] Yuzhong Zhao, Qixiang Ye, Weijia Wu, Chunhua Shen, and Fang Wan. 2023. Generative Prompt Model for Weakly Supervised Object Localization. *arXiv preprint arXiv:2307.09756* (2023).

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.

[56] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. 2022. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9046–9056.

[57] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. 2022. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[58] Yuan Zhou, Xin Chen, Yanrong Guo, Shijie Hao, Richang Hong, and Qi Tian. 2023. Advancing Incremental Few-shot Semantic Segmentation via Semantic-guided Relation Alignment and Adaptation. *arXiv preprint arXiv:2305.10868* (2023).

[59] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. 2021. Self-Promoted Prototype Refinement for Few-Shot Class-Incremental Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr46437.2021.00673