Prototypical Kernel Learning and Open-set Foreground Perception for Generalized Few-shot Semantic Segmentation

Kai Huang, Feigege Wang, Ye Xi, Yutao Gao

Alibaba Group

zhouwan.hk, feigege.wfgg, yx150449, yutao.gao@alibaba-inc.com

Abstract

Generalized Few-shot Semantic Segmentation (GFSS) extends Few-shot Semantic Segmentation (FSS) to simultaneously segment unseen classes and seen classes during evaluation. Previous works leverage additional branch or prototypical aggregation to eliminate the constrained setting of FSS. However, representation division and embedding prejudice, which heavily results in poor performance of GFSS, have not been synthetical considered. We address the aforementioned problems by jointing the prototypical kernel learning and open-set foreground perception. Specifically, a group of learnable kernels is proposed to perform segmentation with each kernel in charge of a stuff class. Then, we explore to merge the prototypical learning to the update of base-class kernels, which is consistent with the prototype knowledge aggregation of few-shot novel classes. In addition, a foreground contextual perception module cooperating with conditional bias based inference is adopted to perform class-agnostic as well as open-set foreground detection, thus to mitigate the embedding prejudice and prevent novel targets from being misclassified as background. Moreover, we also adjust our method to the Class Incremental Few-shot Semantic Segmentation (CIFSS) which takes the knowledge of novel classes in a incremental stream. Extensive experiments on PASCAL-5^i and COCO-20^i datasets demonstrate that our method performs better than previous state-of-the-art.

1. Introduction

Few-shot semantic segmentation [25, 39, 2, 15] aims to handle the situation of only sparse annotated data available, which perplexes most normal semantic segmentation methods. Inherited from the superiority of few-shot learning, the FSS methods can quickly be expended to unseen domains with only few labeled samples. Typically, most FSS methods are built on episode-based meta-learning [27], each episode data contains a support set and a query set shared
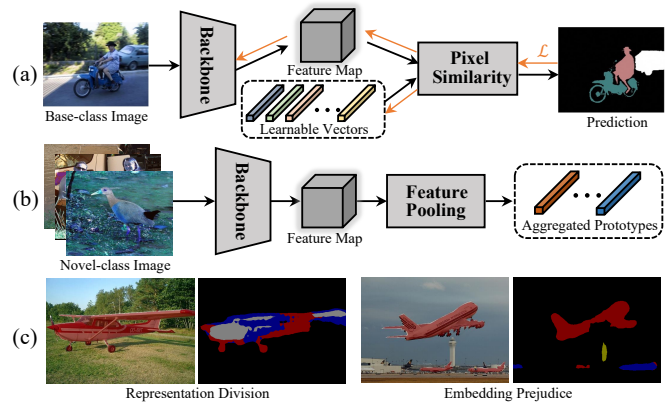

Figure 1. Illustration of our motivation. (a) shows the generation of base-class representation vectors, it learns from abundant base-class images. (b) is the process of novel-class prototypes, which are aggregated by the lightweight feature pooling since few samples available. (c) demonstrates the representation division and embedding prejudice problem existed in current GFSS.

with same classes. In general, the support set consists of few support images with pixel-wise annotations. The models are supported to learn a generic meta-learner with abundant episodic data sampled from the base classes, and adopt to novel classes with the guidance of the support set. However, the majority of current FSS methods can not escape from the limitation of support set, in which specific as well as single novel class annotations are needed corresponding to query images during the inference. Another weakness is that models can only segment the specific novel class of support set while lack the identification of base classes or multi-novel classes. The recent work [15] leverages the prediction of base classes with a base learner, and thus gives a mask region, which instructs the area does not need to be segmented, to guide the recognition of novel concepts. Despite the significant improvement has been made in base classes, the issue of heavily relying on the support samples with the same classes for each query image still exists.

To this end, Generalized Few-shot Semantic Segmentation (GFSS) was proposed [28] where novel classes are

aforehand registered to the vanilla model learned from the base classes, making it possible to segment the base and novel simultaneously without extra information in the inference phase. The typical similarity metric enables a more flexible way to expand novel classes with few annotated samples above the base classes. It can directly perform the inference the same as normal segmentation models, with no need of the prerequisite which informs the target classes contained in query images. Regardless, there are still two challenges in GFSS as shown in Fig.1(c): 1) The representation division between base-class and novel-class blocks the accurate predictions, where the former are learned from sufficient data and the latter are aggregated from limited samples with feature pooling approaches; 2) the prejudice of feature embedding makes the novel-class pixels have the tendency to be identified as varied background, since the feature extractor is trained in base classes.

In this work, we propose to joint prototypical kernel learning and open-set foreground perception to address the above-mentioned problems. It exploits a base-class kernel update schema by applying the pixel assembled feature of each input image to extra supervise the update of base-class kernels. Apart from the normal update with abundant base-class samples, those kernels are endowed with the ability of prototypical representation by the prototypical kernel update, which is consistent with the knowledge aggregation of novel-class with few labeled data. To mitigate the feature embedding preconception, we leverage the common feature patterns cross multi-class targets by an open-set foreground perception, which serves the purpose to offer the class-agnostic as well as open-set foreground prediction. Finally, a conditional bias based inference is proposed to assemble the class-wise prediction and the binary foreground prediction. Furthermore, we expand our method from the GFSS to class incremental few-shot segmentation which registers novel classes with an incremental stream.

In summary, our primary contributions can be summarized as: 1) we present a prototypical kernel learning for maintaining the representation consistency between base-class and novel-class by the notion of kernel; 2) we propose to learn the class-agnostic foreground perception cross the multi-class targets to mitigate the feature embedding preconception, and devise a conditional bias based inference for ensemble prediction; 3) extensive experiments on GFSS and CIFSS demonstrate that the proposed method achieves higher performance compared with other approaches.

2. Related Work

Few-shot Segmentation aims to give a dense pixel prediction for the query images with only a few annotated support images available. The most recent methods adopt two-branch paradigm, *i.e.*, a support branch for target classes prototypical representation and a query branch for feature pairs comparison. Typically, the support branch is performed by the mask-involved feature pooling, such as masked average pooling [26], dynamic kernel generation [18]. The query branch tries to strengthen the metric ability between support prototypes and query feature, with directly pixel similarity [5], convolutional correlation [18, 19], attention-based interaction [36, 40] and so on. However, limited by the instance-based feature comparison of query branch, those methods can not handle on the GFSS scene, in which both base classes and novel classes are included to validation. BAM [15] introduces a base learner module for highly reliable base classes segmentation, and further incorporates into the meta prediction with an ensemble module. Nevertheless, it still needs specific support data and only works with single novel class for each query task. CAPL [28] proposes a novel class registration phase to register the novel information to base classifier, making it possible to be qualified for the segmentation of both base and multi-novel classes without additional support samples.

Prototypical Learning is meant to excavate or utilize the representative examples directly with the exemplar-driven pattern. Owing to the needs of representation with limited data, the prototypical learning methods have been widely used in many few-shot tasks [27, 14, 5, 41], especially few-shot segmentation [32, 48]. A typical delegate is the prototype-based framework, in which annotated support data is class-wise aggregated as prototypes and then guide the segmentation of query samples. The prototypical learning of FSS can be roughly divided into implicit and explicit. The implicit methods mainly focus on the embedding learning and tend to use the feature pooling approaches, such as masked average pooling [26], attention-based aggregation [36] and kernel generation [18]. While the explicit way is to construct learnable prototypes, so that the labeled instances [48] or meta episode data [37] can react it directly.

Foreground Modeling has gained rapid development with the effectiveness of targets detection. The most representative methods are class activation mapping and its variants [47, 24, 31], which are early used for feature visualization [47] and further as the rough foreground mask to initialize the progressive inception of many weak-supervised approaches [6, 4]. Later, salient detection [23, 46, 33] has been raised for binary segmentation tasks, with the strong ability of flexible saliency perception. The mainstream salient works only consider the intra interaction on each single image separately, which suffer from obscure and coarse results. Co-salient [42, 9] extends the separate mode to co-exploration within the image group, while extra information, *e.g.*, edge [8], class-wise gradient [44], is always needed for discriminative feature learning. In this work, we propose to learn the common foreground patterns cross multi-class group instances, so as to directly promote to the detection of novel targets without class limitation.
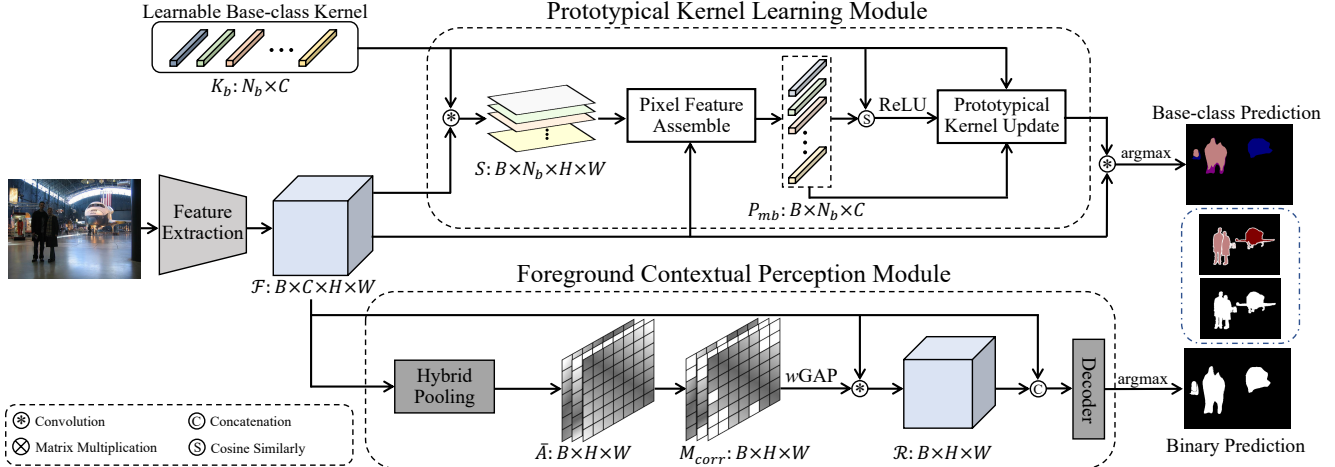
Figure 2. Overview of training framework. For the upper branch, the learnable base-class kernels are optimized by the Prototypical Kernel Learning (PKL). The down branch, named as Foreground Contextual Perception (FCP), is constructed to learn the open-set foreground detection by leveraging the contextual feature patterns cross multi-class targets. Batch is omitted in PKL for better display.

3. Method

3.1. Task Formulation

Generalized few-shot segmentation is designed to quickly integrate unseen classes into a segmentation model with only a few annotated data. It aims to actualize segmentation tasks on novel classes with limited labeled samples and base classes with sufficient labeled data simultaneously. GFSS model learns the basal knowledge from the base classes set C_{base}, and novel classes set C_{novel} is registered to this base model with a few labeled data. There are no overlap between the base and novel classes set, *i.e.*, $C_{base} \cap C_{novel} = \emptyset$. In the inference process, the model should be qualified to the segmentation of the union set $C_{base} \cup C_{novel}$. Different from the normal FSS which are specific to the modeling of few-shot learner, GFSS is proposed to explore the possibility of empowering the few-shot learning ability to a vanilla segmentation model.

3.2. Overview

The overall architectures for training and inference are shown in Fig. 2 and Fig. 4, which mainly consists of three components: Prototypical Kernel Learning (PKL), Foreground Contextual Perception (FCP) and Conditional Bias Based Inference (CBBI). The feature of input is respectively fed to two branches. For the upper branch, we employ the PKL to dynamically get the refined base-class kernels specific to the assembled feature of input image, and the class-wise predictions are produced by applying those kernels to the pixel features. For the down branch, the binary foreground prediction is decoded by the FCP which incorporates relevant contextual patterns cross multi-classes. The final prediction masks are obtained by combining the predictions of these two branch with the CBBI.

3.3. Prototypical Kernel Learning Module

Inspired by the dynamic kernel [11, 34, 43] in semantic segmentation, which possess better representation ability than convolution kernels with flexible structures [35] or roles [16], we extend this to GFSS task by adopting prototypical kernel learning to maintain the representation consistency of base-class and novel-class.

Specifically, N_b learnable kernels $K_b \in \mathbb{R}^{N_b \times C}$ are constructed to represent the knowledge of N_b base classes, in which the convolution between these kernels and feature maps is taken on the segmentation basis. Concretely, given the input feature maps $\mathcal{F} \in \mathbb{R}^{B \times C \times H \times W}$ of B input images, the semantic segmentation output is formulated as:

$$S = \text{softmax}(K_b * \mathcal{F}), S \in \mathbb{R}^{B \times N_b \times H \times W}, \quad (1)$$

where C, H, and W indicate the number of channels, height, and width of the feature map respectively, and N_b is the number of classes in base-class set.

Pixel Feature Assembling. As the segmentation output of each kernel can reflect the class-wise pixel-to-prototype mapping, the pixel feature assembling is performed by aggregating the input feature with the pseudo semantic segmentation masks as:

$$P_{mb} = \sum_x^H \sum_y^W S(x, y) \cdot \mathcal{F}(x, y), P_{mb} \in \mathbb{R}^{B \times N_b \times C}. \quad (2)$$

The generation of P_{mb} are analogous to the novel-class feature aggregation with few labeled samples, which takes the ability of prototypical representation.

Prototypical Kernel Update. Ordinarily, the kernel of base-class learn the general perception from abundant annotated samples, it would be broader and more inclusive

than the novel-class kernels learned from few data, which exists representation division. Those ambiguous instances, *e.g.*, the novel classes samples, are more predisposed to be caught by the base-class kernels. Thus, the prototypical kernel update is proposed to adapt the base-class kernels to few-shot task with extra one-off update specific to each input:

$$\tilde{K}_b = K_b - \alpha \nabla_{K_b} \mathcal{L}(K_b) \qquad (3)$$

where $\nabla_{K_b} \mathcal{L}(K_b) = \nabla_{K_b}(K_b - P_{mb}^i)^2$ is the gradient descent of the mean squared error loss between K_b and the assembled feature of i-th images P_{mb}^i, α is the adaptation learning rate, which denotes as kernel-to-prototype similarity with the restraint of ReLU function that insures constant positive:

$$\alpha = \text{ReLU}(\text{cosine}(K_b, P_{mb}^i)), \alpha \in \mathbb{R}^{1 \times N_b}. \qquad (4)$$

The adaptation learning rate is designed to adaptive filter out ineffective gradient signals caused by those classes do not exist in this image. Since the classes existed in input image are a subset of base-class set, the pixel feature assembling of those missing classes is invalid and will lead to incorrect kernel update. The adaptation learning rate will decay to near zero to block the ineffective update as the kernel-to-prototype similarities are low. Finally, The predicted segmentation mask is obtained by:

$$M(x,y) = \arg\max_i \tilde{S}^i(x,y), \qquad (5)$$

where $\tilde{S}^i(x,y)$ is the score map between the updated base-class kernels and the feature maps followed by Eq. (1).

3.4. Foreground Contextual Perception Module

The aforementioned PKL module aims to adjust the general base-class kernels to few-shot scenario with assembled prototypical feature, so that the representation divergence between base and novel classes is mitigated. While the perception ability of novel foreground targets remains poor with the limitation of backbone training, where only the base-class data is sufficiently accessible. Inspired by the episode-based paradigm in various few-shot learning [30, 25, 14], which can boost the model to adapt to unknown scenes, we propose a foreground contextual perception module to perform the open-set foreground perception. The batch group of images sampled from the base categories set can be considered as a pseudo episode, in which each batch group is randomly mixed with the samples of different base-class targets. The FCP module aims to mine common foreground patterns cross multi-class samples with abundant pseudo episode, thus possesses the ability of class-agnostic perception and prevents novel targets from being misclassified as background.

Hybrid Pooling. As shown in Fig.3, the feature maps of pseudo episode are first passed by two independent lin-
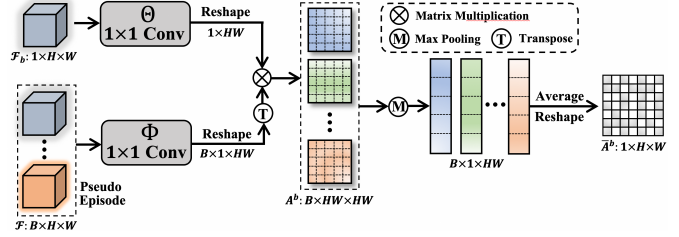

Figure 3. Illustration of the Hybrid Pooling in FCP module.

ear mappings, and the pixel-wise correlation is calculated within the feature map groups:

$$A^b = \Phi(\mathcal{F})^T \Theta(\mathcal{F}_b) \in \mathbb{R}^{B \times HW \times HW}, \qquad (6)$$

where Φ and Θ represent 1×1 convolution layers, $\mathcal{F}_b$ is the b-th feature map, and A^b denotes the pixel-wise correlation between the b-th feature map and whole pseudo episode. The first dimension of A^b indicates all images of pseudo episode, while the second and third dimensions mean the total feature pixels of each image in pseudo episode and the total feature pixels of query. The reshaping and transposing operations of matrix dimension are omitted for the simplification of presentation. A max pooling and average pooling are separately acted on the second and first dimensions to obtain a more representative correlation $\bar{A} \in \mathbb{R}^{B \times H \times W}$. It is worth mentioning that performing max pooling on pixel dimension can help to find out discriminative correlation pairs, and average pooling on image dimension is supported to learn the common foreground feature patterns.

Foreground Correlation Responses. The correlation normalized by a softmax layer is truncated as the correlation mask M_{corr}:

$$M_{corr}(h,w) = \mathbb{1}(\bar{A}(h,w) > 0.5), \qquad (7)$$

where $\mathbb{1}(*)$ is the binary indicator that outputs 1 when $*$ is true 0 otherwise, h and w are the height index and width index with the limitation of H and W. Although the correlation mask can embody the most discriminative feature pixel position, the relation response of other positions are discarded with the consideration of computational complexity. A sharp correlation information may carries more hard noisy, which blocks the learning of FCP module. Therefore, the spacial weighted Global Average Pooling (wGAP) [39, 38] of the correlation mask together with the normalized original feature maps is introduced to obtain a more robust correlation prototypes as:

$$p_{wGAP} = \frac{\sum_{h,w}\|\mathcal{F}(h,w)\|_2 M_{corr}(h,w)}{\sum_{h,w} M_{corr}(h,w)} \in \mathbb{R}^{B \times C}, \qquad (8)$$

where $\|*\|_2$ is the L2 normalization. The correlation prototypes reflect the positive relations between the pixel and whole image pixels within the pseudo episode. The refined

correlation responses can be obtained by applying the correlation prototypes to the original feature maps as:

$$\mathcal{R}(h, w) = \frac{1}{B} \sum_{i=1}^{B} p_{wGPA} \|\mathcal{F}(h, w)\|_2. \qquad (9)$$

Compared to the pixel-wise correlation referred to Eq. (6), the refined correlation responses are more robust with fewer sharp noises, as well as more accurate foreground perception. The final binary foreground masks are predicted by decoding the residual features:

$$M_f = \mathbb{D}(\mathcal{F} + \mathcal{F} \times \mathcal{R}), \qquad (10)$$

where $\mathbb{D}$ represents a correlation decoder built with the four cascaded convolutional layers.

3.5. Training Loss and Novel Classes Registration

Training Loss. The loss for training our model consists of two parts corresponding to the PKL module and FCP module respectively. The loss function for PKL module is illustrated as:

$$\mathcal{L}_{ce} = \frac{1}{HW} \sum_{x,y} \sum_{p^i \in P_{ts}} \mathbb{1} \left[M^{x,y} = i \right] \log \hat{S}_c^{x,y}, \qquad (11)$$

where $\mathbb{1}(*)$ is the binary indicator same as Eq. (7). We adopt the IoU loss [23] to train the FCP module as:

$$\mathcal{L}_{iou} = \frac{1}{B} \sum_{b=1}^{B} (1 - \frac{\hat{Y} \cap Y}{\hat{Y} \cup Y}), \qquad (12)$$

where $\hat{Y}$ is the binary predictions and Y is the ground truth. Finally, the total end-to-end loss over the batch is:

$$\mathcal{L}_{total} = \lambda \sum_{b \in B} \mathcal{L}_{ce}^b + (1 - \lambda)\mathcal{L}_{iou}, \qquad (13)$$

where λ is utilized to balance the loss functions.

Novel Classes Registration. After training, the novel-class kernels $K_n \in \mathbb{R}^{N_n \times C}$ can be obtained by passing the support images of novel classes to the frozen feature extraction, following with the class-wise averaged wGAP referred to Eq. (8). Then, the novel classes knowledge is registered to the model by concatenating the novel-class kernels to the updated base-class kernels:

$$K_w = Concat([K_b, K_n]), K_w \in \mathbb{R}^{(N_b + N_n) \times C}. \qquad (14)$$

3.6. Conditional Bias Based Inference

Given the input test image I, the final decision of position $\mathbf{p}$ is defined as follows:

$$O_I(\mathbf{p}) = \begin{cases} \max(c_{1st}, c_{2nd} + b), & O_s(\mathbf{p}) == 1 \\ c_{1st}, & \text{otherwise} \end{cases}, \qquad (15)$$
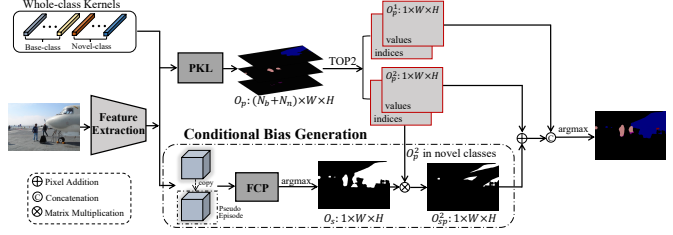

Figure 4. Framework of the CBBI. We perform the FAH module with only itself contained in the pseudo episode.

where c_{1st} and c_{2nd} are the first and second candidates according to the class-wise similarity, O_s indicates the output of FCP module and b is the constant of bias.

As shown in Fig. 4, the refined meta specific prototypes are obtained by feeding the feature map of test image together with the class-wise prototypes to the PKL module. With the concerns of consistency and noise accumulation, we only update the base-class prototypes. Then, the class-wise predictions O_p are calculated by the cosine similarity between the updated whole-class prototypes and query feature maps. The first and second candidate masks are the value of Top2 maps along with the class dimension of O_p.

For the branch of conditional bias generation, the open-set foreground mask O_s can be obtained by the FCP module. A pointwise multiplication between it and the second candidate mask is adopted to obtain the conditional bias mask O_{sp}^2. Then, the second candidate mask is recalculated as:

$$\hat{O}_p^2(\mathbf{p}) = O_p^2(\mathbf{p}) + b \cdot O_{sp}^2(\mathbf{p}). \qquad (16)$$

The final output can be obtained by the maximum argument between the first and second candidate masks.

3.7. Extension to Class Incremental FSS

Following the incremental learning in other few-shot tasks [10, 22], we give the definition of class incremental setting in few-shot segmentation. The Class Incremental Few-shot Semantic Segmentation (CIFSS) shares the same classes set with GFSS, in which there are base classes set C_{base} with adequate data available and novel classes set C_{novel} with a small amount of annotated data. Different from the one-off task of novel classes in GFSS, CIFSS focuses on the novel classes in a session stream. The model learns from the base session S_b with the base classes and adapts to novel classes with insufficient samples in sequential, *i.e.*, the sequence session $\{S_n^1, \cdots, S_n^T\}$ with the novel classes. It is critical to mention that we can only access to the data of t-th session when attaching this session, and there is no overlap within these novel sessions. The performance of model with current session S_n^t is evaluated on the test samples of S_n^t and the previous sessions $\{S_n^1, \cdots, S_n^{t-1}\}$. The GFSS can be viewed as a exception constructed with only base session and one novel session.

Owing to the extensible kernel and the mechanism of novel classes registration, the proposed model can be easy expanded to CIFSS. Similar to the training of GFSS, we learn the PKL and FCP module in the base session and all parameters are fixed during the sequence novel sessions. There are a series of novel classes registration processes corresponding to incremental classes learning with serialized novel sessions. To be specific, assume K_w^t is the whole-class kernels of t-th session, the extended kernels of next session K_w^{t+1} is obtained by concatenating the novel session kernels K_n^{t+1}:

$$K_w^{t+1} = Concat([K_w^t, K_n^{t+1}]), \in \mathbb{R}^{(N_w^t + N_n^{t+1}) \times C}, \quad (17)$$

where the calculation of K_n^{t+1} is the same with novel classes registration of GFSS referred to 3.5. In the inference process, only the kernels of base session are passed to the PKL module, and all the kernels of few-shot sessions are treated as a novel whole in the CBBI.

4. Experiments

4.1. Implementation Details

Dataset. Following [28], we evaluate the proposed method on two public datasets, *i.e.*, PASCAL-5^i and COCO-20^i. PASCAL-5^i is modified from PASCAL VOC [7] with extra mask annotations from Semantic Boundaries Dataset (SBD) [12] dataset, it has 20 object classes and one background class. COCO-20^i consists of MS COCO [17] with 80 object classes and one background class for segmentation. In terms of GFSS, all the object classes are evenly divided into four folds: $\{5^i, i \in \{0, 1, 2, 3\}\}$ for PASCAL-5^i and $\{20^i, i \in \{0, 1, 2, 3\}\}$ for COCO-20^i. The rest object classes as well as background are treated as base classes, where the object classes of specific fold are conducted as novel classes. Models are directly evaluated on the validation set with both base and novel classes simultaneously. In terms of CIFSS, it shares the same fold division with GFSS. The object classes of specific fold are evenly divided into five novel sessions, while the rest object classes and the background class are included into the base session. On PASCAL-5^i, there are 16 object classes in base session and 1 novel object class for each incremental session. As for COCO-20^i, the number of object classes in base session is 61 and novel object classes in each incremental session is 5. We evaluate the CIFSS models on the classes belonging to base session, current and previous incremental novel sessions with the validation set.

Evaluation metrics. Similar to [28], we use mean Intersection over Union (mIoU) for base classes as well as base session mIoU$_\mathcal{B}$ and novel classes as well as incremental novel sessions mIoU$_\mathcal{N}$. As for GFSS setting, we also provide the overall mIoU with all classes denoted by mIoU$_\mathcal{O}$. Since the base classes are the vast majority, the arithmetic

Table 1. Quantitative results with GFSS setting on PASCAL-5^i and COCO-20^i. Best-performing results are highlighted in bold.

Methods	1-shot				5-shot			
	mIoU$_\mathcal{B}$	mIoU$_\mathcal{N}$	mIoU$_\mathcal{O}$	hIoU	mIoU$_\mathcal{B}$	mIoU$_\mathcal{N}$	mIoU$_\mathcal{O}$	hIoU
PASCAL-5^i								
CANet [39]	8.73	2.42	7.23	3.79	9.05	1.52	7.26	2.60
PFENet [29]	8.32	2.67	6.97	4.04	8.83	1.89	7.18	3.11
SCL [38]	8.88	2.44	7.35	3.83	9.11	1.83	7.38	3.05
PANet [32]	31.88	11.25	26.97	16.63	32.95	15.25	28.74	20.85
BAM [15]	64.65	17.06	53.32	27.00	65.28	19.99	54.50	30.61
CAPL [28]	65.48	18.85	54.38	29.27	66.14	22.41	55.72	31.96
Ours	**68.84**	**26.90**	**58.86**	**37.83**	**69.22**	**34.40**	**61.18**	**45.42**
COCO-20^i								
Baseline [28]	36.68	5.84	29.06	10.01	36.91	7.26	29.59	12.13
BAM [15]	44.08	6.74	34.86	11.69	46.10	11.21	37.49	18.03
CAPL [28]	44.61	7.05	35.46	12.18	45.24	11.05	36.80	17.76
Ours	**46.36**	**11.04**	**37.71**	**17.83**	**46.77**	**14.91**	**38.90**	**22.61**

mean might be dominated by mIoU$_\mathcal{B}$ [1]. We adopt the harmonic mean hIoU of mIoU$_\mathcal{B}$ and mIoU$_\mathcal{N}$ which can emphasize the impact of novel classes.

Training details. For fair comparison, the PSPNet [45] with RestNet50 [13] is adopted as our feature extractor, which is the same as most FSS and GFSS methods, and we train our model for 50 epochs both on PASCAL-5^i and COCO-20^i. The model is optimized by the SGD with an initial learning rate of 2.5e-3, where momentum is 0.9 and the weight decay is set to 1e-4. All images together with the masks are all resized to 473×473 for training and tested with their original sizes. The λ of loss function is set to 0.6 and the bias value b of CBBI is 0.5. We obtain the specific split results by averaging over five different random seeds, and the final reported results are averaged with all splits.

4.2. Comparisons

GFSS Setting. Table 1 shows the 1-shot and 5-shot quantitative results of different methods on PASCAL-5^i and COCO-20^i. It can be found that the proposed method shows substantial gains compared with the advanced approaches and achieves the state-of-the-art performance of all GFSS settings and datasets. As for the overall mIoU on PASCAL-5^i, the proposed method exceeds the CAPL [28] by 4.48% and 5.46% for 1-shot setting and 5-shot setting respectively. With the significant improvement on novel classes, our method surpasses the CAPL [28] by 8.56% and 13.46% hIoU, demonstrating the superiority of the proposed method. Meanwhile, better performance on novel classes can also weaken the confusion between base classes and novel classes and improve the mIoU$_\mathcal{B}$.

CIFSS Setting. We compare our method with the following modified approaches as: 1) the relevant incremental few-shot learning methods, such as PIFS [3] and iFS-RCNN [20]; 2) the current FSS and GFSS methods, *i.e.*, PFENet [29], BAM [15] and CAPL [28]. From Table 2, our method shows substantial gains compared with existing approaches.

Table 2. 1-shot Quantitative results with CIFSS setting on the PASCAL-5^i and COCO-20^i. Best-performing results are highlighted in bold. * We use the weights of PSPNet [45] backbone pretrained on base classes as the same as the BAM [15].

Datasets	Methods	sesson 0 $m\text{IoU}_\mathcal{B}$	session 1 $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$	session 2 $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$	session 3 $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$	session 4 $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$	session 5 $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$
PASCAL-5^i	PFENet* [29]	74.43	67.21	10.34	17.92	66.62	14.12	23.30	64.97	11.12	18.99	65.77	12.08	20.41	62.50	11.08	18.82
	iFS-RCNN [20]	72.43	67.71	11.08	19.04	67.17	16.63	26.66	65.51	12.94	21.61	65.48	11.75	19.92	63.36	12.39	20.73
	CAPL [28]	74.86	70.51	17.94	28.60	69.07	21.11	32.34	67.89	18.03	28.49	67.14	19.57	30.31	65.53	19.08	29.55
	PIFS [3]	75.04	71.16	19.49	30.60	69.80	24.33	36.08	68.05	19.92	30.82	68.44	21.08	32.23	65.83	20.99	31.83
	Our	**75.49**	**72.93**	**24.60**	**36.79**	**71.21**	**29.32**	**41.54**	**70.76**	**25.75**	**37.76**	**69.98**	**26.11**	**37.88**	**68.40**	**27.28**	**39.00**
COCO-20^i	PFENet* [29]	54.11	45.79	4.33	7.91	44.58	5.18	9.28	43.07	3.36	6.23	42.40	5.64	9.96	41.99	5.05	9.02
	iFS-RCNN [20]	53.42	50.43	5.03	9.15	45.52	6.60	11.53	43.86	5.07	9.09	43.37	7.11	12.22	42.83	6.63	11.48
	CAPL [28]	54.43	50.25	6.69	11.81	46.67	7.04	12.23	45.92	5.53	9.87	44.08	7.77	13.21	43.85	7.51	12.82
	BAM [15]	**54.80**	**52.04**	8.83	15.10	**49.01**	8.94	15.12	47.03	7.03	12.23	45.56	9.19	15.29	44.48	8.83	14.73
	PIFS [3]	54.27	51.59	8.90	15.18	48.52	9.53	15.93	46.96	6.94	12.09	45.69	9.93	16.31	44.27	8.80	14.68
	Our	54.39	51.78	**11.47**	**18.55**	48.33	**11.22**	**18.08**	**47.28**	**9.61**	**15.89**	**46.25**	**11.47**	**18.33**	**45.24**	**10.74**	**17.31**

Table 3. Ablation Study of the effect with different components.

PKL	FCP	CBBI	1-shot $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$	5-shot $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$
			61.22	14.50	23.45	62.07	19.73	29.94
✓			66.82	22.95	34.17	67.73	29.18	40.79
	✓		64.49	16.73	26.57	65.28	23.39	34.44
✓	✓		67.73	23.58	34.98	68.05	31.45	43.02
✓	✓	✓	**68.84**	**26.90**	**37.83**	**69.22**	**34.40**	**45.42**

Table 4. Ablation performances of the proposed method. *fixed*ALR means using fixed learning rate referred to Eq. (3). $\text{pixel}_{AP \to MP}$ indicates replacing average pooling with max pooling in pixel dimension referred to Eq. (6), and $\text{image}_{MP \to AP}$ is the similar replacement in image dimension. *u*CBBI and *c*CBBI denote the unconditional form and learnable convolution form for CBBI.

Settings	1-shot $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$	5-shot $m\text{IoU}_\mathcal{B}$	$m\text{IoU}_\mathcal{N}$	$h\text{IoU}$
w/o PKL	64.49	16.73	26.57	65.28	23.39	34.44
*fixed*ALR	66.90	20.48	31.36	67.72	26.08	37.66
concat+AP	65.91	19.41	30.00	66.86	25.51	36.93
concat+LP	66.27	21.19	32.11	67.71	26.89	38.49
w/o FCP	66.82	22.95	34.17	67.73	29.18	40.79
$\text{pixel}_{AP \to MP}$	65.94	22.31	33.34	66.42	28.91	40.28
$\text{image}_{MP \to AP}$	65.05	21.94	32.81	66.18	28.30	39.65
GradCAM [24]	66.94	23.36	34.74	68.08	29.80	41.46
PAGENet [33]	67.57	24.80	36.28	68.83	31.70	43.41
MINet [21]	67.89	24.99	36.53	68.65	31.47	43.16
w/o CBBI	67.73	23.58	34.98	68.05	31.45	43.02
*u*CBBI	66.29	20.83	31.70	67.13	25.38	36.83
*c*CBBI	65.53	18.86	29.29	66.90	23.39	34.66
Ours	68.84	26.90	37.83	69.22	34.40	45.42

Since majority approaches adopt analogous backbone, we get similar performance in the session 0, which only construct with base classes. The superiority of our method both in base and novel classes continues to emerge as the novel session increased. It partly embodies the significant benefit of our old and novel classes knowledge ensemble mechanism, and partly verifies that the improvement on novel classes can result better performance of base classes as well.

4.3. Ablation Studies

Component-wise Effectiveness. We ablate the three major components of proposed method in Table 3: prototypical kernel learning module, foreground contextual perception module and conditional bias based inference. We use the PSPNet [45] modified with learnable kernels as the baseline, in which simple concat operation is conducted between base-class and novel-class kernels. We can see that adding either one of PKL module or FCP module brings substantial improvement compared with the baseline. It means the proposed two modules are indeed beneficial to the novel classes learning in few-shot scenarios, and can mitigate the knowledge division with the extremely class-imbalanced GFSS setting. Moreover, it worth noting that the FCP module explicit plays a vital role in providing the general foreground information, the implicit effect is to boost the feature expression of unknown targets. Fig. 5 shows the qualitative results, which straightforward embodies the effectiveness of proposed method.

Prototypical Kernel Learning Module. The motivation of PKL module is to empower the prototypical ability for base-class kernels, so as to maintain the consistency with the registered novel classes, which is also prototypical aggregated from limited data. We replace the PKL with directly adopting the base-class kernels to the input feature maps as our baseline. As shown in the upside of Table 4, we first construct the ablation studies on the adaptation learning rate, it is clear that the fixed learning rate brings worse results with the distraction of noisy gradient signals during the prototypical kernel update. Meanwhile, we introduce several alternative prototype fusion approaches which are performed by the adaptive pooling (AP) or liner projection (LP) after the channel-wise *concat* operation between K_b and P_{mb}. The results show that our method with the proposed prototypical kernel update remarkably outperforms the aforementioned approaches. It can strike a better balance between generality and specificity, especially for novel targets in generalized inference process, thus is more suitable to the GFSS.

Foreground Contextual Perception module. The FCP aims to offer a general foreground detection which helps to the mining of novel targets. Similarly, the model with-
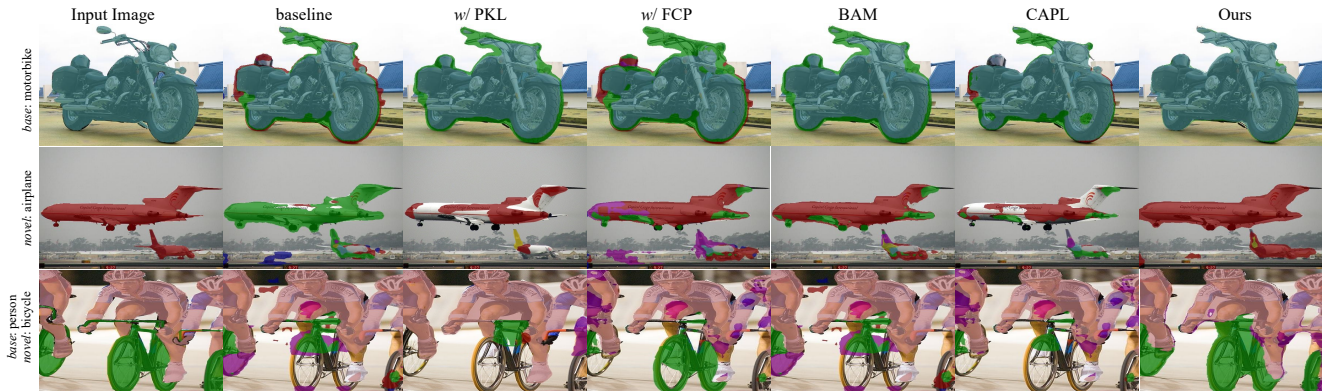
Figure 5. Qualitative results of our method, ablation of each components, BAM [15] and CAPL [28] on PASCAL-5^i benchmark.

out FCP is conducted as the baseline, as well as replacing the CBBI with normal inference. The middle part of Table 4 demonstrates the ablation results of the pixel-wise group correlation aggregation referred to Eq. (6), which is the key of FCP, as well as the comparison with other alternative foreground detection approaches. Using average pooling on pixel dimension of the pixel-wise group correlation tends to introduce more background noisy, leading to worse performance. We also compare our FCP with CAM-based and well-designed salient object detection approaches. Since the CAM-based methods are heavily relied on the extra classes information, such as class-wise gradients [24] or activation values [31], it lacks the ability of novel classes detection. Similar deficiency also lies in other well-designed salient detection methods, *e.g.*, PA-GENet [33] and MINet [21], which only consider the interaction within specific classes and lose the generality of foreground detection. The FCP pays more attention to the robust and universal feature patterns with abundant pseudo episode, thus is more easy to generalize to novel targets.

Conditional Bias Based Inference. In our method, the CBBI plays a role in combining the outputs of the PKL module and the FCP module. We also experiment on the way of fusing those outputs in the bottom part of Table 4. When the class-agnostic foreground mask predicted by the FCP module is utilized with unconditional form, the performance gets worse with the interference of background prediction. We also replace the strategy based inference with learnable convolution structure with end-to-end training, the underperform results indicate that introducing parameterized inference module may heavy the optimization of model and even block the learning of other modules.

Performance distinction between PASCAL-5^i and COCO-20^i. We further investigate the improvement gap between PASCAL-5i and COCO-20i referred to Table 1, *e.g.*, 8.05% v.s. 3.99% exceed CAPL [28] on mIoU of novel classes. The most obvious difference between these two datasets is that coco-20i contains larger classes set than

Table 5. Results on the subset of COCO. *a+b/c* indicates the combination of classes, where *a* is the number of base classes, *b*=1 represents the background class and *c* is the number of novel classes.

Methods	15+1/5		30+1/10		45+1/15		60+1/20	
	mIoU$_\mathcal{B}$	mIoU$_\mathcal{N}$	mIoU$_\mathcal{B}$	mIoU$_\mathcal{N}$	mIoU$_\mathcal{B}$	mIoU$_\mathcal{N}$	mIoU$_\mathcal{B}$	mIoU$_\mathcal{N}$
Baseline [28]	51.19	10.08	47.74	8.22	41.37	7.10	36.68	5.84
CAPL [28]	63.39	16.08	57.22	13.39	50.08	10.16	44.61	7.05
Ours	66.70	23.76	59.48	18.91	52.01	14.27	46.36	11.04
Improvement	+3.31	+7.76	+2.26	+5.52	+1.92	+4.11	+1.75	+3.99

PASCAL-5^i. Thus, we design the ablation experiments with gradually enlarging the subset of COCO, which is started from a similar scale with PASCAL-5^i. As shown in Table 5, we find parallel improvements are raised when the scale and distribution of classes are similar to PASCAL-5^i, and as the number of classes increasing, the gaps are reduced. Comparable phenomenon also happens on the contrast between the CAPL [28] and baseline, indicating the common issue that the effect of GFSS model may get weakened as the scale of classes set increases.

5. Conclusion

In this work, we have proposed to leverage the prototypical kernel learning and open-set foreground perception for generalized few-shot semantic segmentation. Special to two of major issues existed in current GFSS, *i.e.*, representation division and embedding prejudice. We offer an effective scheme with three components, which are PKL, FCP and CBBI. To maintain the prototypical representation consistency and prevent novel targets from being misclassified as background, the PKL and FCP are proposed to perform adaptive prototypical kernel update and open-set targets detection, respectively. Moreover, the CBBI is designed for the ensemble of the aforementioned two modules and offers the final prediction. Besides, we further extend the GFSS to CIFSS which is more challenging with incremental stream of novel classes. Extensive experiments demonstrate that the proposed method achieves state-of-the-art results.

References

[1] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9536–9545, 2021. 6

[2] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13979–13988, 2021. 1

[3] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot semantic segmentation. In *British Machine Vision Conference*, 2021. 6, 7

[4] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4288–4298, 2022. 2

[5] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *British Machine Vision Conference*, volume 3, 2018. 2

[6] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 2

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[8] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4339–4354, 2021. 2

[9] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12288–12298, 2021. 2

[10] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1185–1194, 2021. 5

[11] Hang Gao, Xizhou Zhu, Stephen Lin, and Jifeng Dai. Deformable kernels: Adapting effective receptive fields for object deformation. In *International Conference on Learning Representations*, 2019. 3

[12] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 991–998, 2011. 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[14] Kai Huang, Jie Geng, Wen Jiang, Xinyang Deng, and Zhe Xu. Pseudo-loss confidence metric for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8671–8680, 2021. 2, 4

[15] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8057–8067, 2022. 1, 2, 6, 7, 8

[16] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18857, 2022. 3

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014. 6

[18] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2022. 2

[19] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6952, 2021. 2

[20] Khoi Nguyen and Sinisa Todorovic. ifs-rcnn: An incremental few-shot instance segmenter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7010–7019, 2022. 6, 7

[21] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, 2020. 7, 8

[22] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13855, 2020. 5

[23] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 2, 5

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 618–626, 2017. 2, 7, 8

[25] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference*, 2017. 1, 4

[26] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5249–5258, 2019. 2

[27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1, 2

[28] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2022. 1, 2, 6, 7, 8

[29] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1050–1065, 2020. 6, 7

[30] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016. 4

[31] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020. 2, 8

[32] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019. 2, 6

[33] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1448–1457, 2019. 2, 7, 8

[34] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 3

[35] Jialin Wu, Dai Li, Yu Yang, Chandrajit Bajaj, and Xiangyang Ji. Dynamic filtering with large sampling field for convnets. In *Proceedings of the European Conference on Computer Vision*, pages 185–200, 2018. 3

[36] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 517–526, 2021. 2

[37] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8721–8730, 2021. 2

[38] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8312–8321, 2021. 4, 6

[39] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019. 1, 4, 6

[40] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34:21984–21996, 2021. 2

[41] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6974–6983, 2021. 2

[42] Kaihua Zhang, Tengpeng Li, Bo Liu, and Qingshan Liu. Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3095–3104, 2019. 2

[43] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 3

[44] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Proceedings of the European Conference on Computer Vision*, pages 455–472. Springer, 2020. 2

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017. 6, 7

[46] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8779–8788, 2019. 2

[47] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2

[48] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022. 2