FGNet: Towards Filling the Intra-class and Inter-class Gaps for Few-shot Segmentation

Yuxuan Zhang[1], **Wei Yang**[1,2,3,*], **Shaowei Wang**[4]
[1] School of Computer Science and Technology, University of Science and Technology of China
[2]Suzhou Institute for Advanced Research, University of Science and Technology of China
[3]Hefei National Laboratory
[4] Institute of Artificial Intelligence and Blockchain, Guangzhou University
yxzhang123@mail.ustc.edu.cn, qubit@ustc.edu.cn, wangsw@gzhu.edu.cn

Abstract

Current few-shot segmentation (FSS) approaches have made tremendous achievements based on prototypical learning techniques. However, due to the scarcity of the support data provided, FSS methods still suffer from the intra-class and inter-class gaps. In this paper, we propose a uniform network to fill both the gaps, termed FGNet. It consists of the novel design of a Self-Adaptive Module (SAM) to emphasize the query feature to generate an enhanced prototype for self-alignment. Such a prototype caters to each query sample itself since it contains the underlying intra-instance information, which gets around the intra-class appearance gap. Moreover, we design an Inter-class Feature Separation Module (IFSM) to separate the feature space of the target class from other classes, which contributes to bridging the inter-class gap. In addition, we present several new losses and a method termed B-SLIC, which help to further enhance the separation performance of FGNet. Experimental results show that FGNet reduces both the gaps for FSS by SAM and IFSM respectively, and achieves state-of-the-art performances on both PASCAL-5^i and COCO-20^i datasets compared with previous top-performing approaches.

1 Introduction

Brilliant efforts have been made in image semantic segmentation [Long *et al.*, 2015; Szegedy *et al.*, 2017; Badrinarayanan *et al.*, 2017; Chen *et al.*, 2017; Yu *et al.*, 2020], achieving excellent performance in several large-scale labeled datasets [Silberman *et al.*, 2012; Zhou *et al.*, 2017; Cordts *et al.*, 2016]. However, current top-performing approaches rely heavily on extensive pixel-wise annotations, which is time-consuming and labour-intensive. To handle this issue, few-shot segmentation (FSS) [Shaban *et al.*, 2017] has received lots of attention in recent years.

FSS is an extension task of few-shot learning, aiming to learn the generalization ability from the given classes and adapt it to the arbitrary novel classes with only a handful of support samples. The mainstream strategy of FSS follows the pattern of metric learning [Dong and Xing, 2018; Wang *et al.*, 2019; Liu *et al.*, 2020] based on a global descriptor, named prototype. In particular, the prototype denotes a representation vector of a specific category, and the prototype-based methods generate a representative prototype for each category from the limited support samples. Then the prototype is leveraged to activate the query feature for predicting the mask of the query image.

However, FSS suffers from a dilemma, which lies in two aspects, i.e., intra-class gap and inter-class gap. On the one hand, the given support images are limited while the query images are various, resulting in the intra-class appearance gap between support data and query data. On the other hand, the feature space of support data is also sparse and in low coverage, leading to the problem of the inter-class classification gap. As shown in Figure 1, such an issue results in an ambiguity to the distinction between the target class and the non-target class with the similar representation.

Current FSS approaches mainly focus on refining the prototype quality, enhancing query features [Yang *et al.*, 2020; Li *et al.*, 2021] or seeking for appropriate matching mechanisms [Wang *et al.*, 2020; Siam *et al.*, 2021]. In spite of their high performances, those methods fail to eliminate the intra-class appearance gap and inter-class classification gap essentially. Especially, no matter how the prototype is refined, the intersection between support images and query images remains inadequate. Moreover, the issue of the inter-class gap is rarely discussed, which makes the generation of prototypes

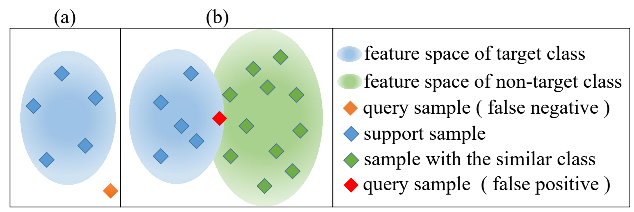

Figure 1: Intra-class and inter-class gaps of few-shot segmentation. (a) The information intersection of support data and query data is not adequate, causing the intra-class gap. (b) The target class may share the similar feature with the given non-target classes, resulting in an ambiguity to predict the data near the decision boundary.

* Corresponding author
Code is available at: github.com/YXZhang979/FGNet

very difficult to distinguish the different classes with similar representations [Okazawa, 2022]. In this paper, we concentrate on tackling the above two problems in one single uniform framework.

Due to the scarcity of the support data, the pattern in the query image may not be contained in the prior knowledge. Accordingly, we propose a Self-Adaptive Module (SAM) to reduce the intra-class gap, which is motivated by [Liu and Qin, 2020; Fan *et al.*, 2022]. We get around the issue and propose a self-adaptive mechanism to establish an enhanced prototype for further prediction. Such a prototype contains the underlying information of the query sample itself, which caters to each query data by self-alignment, as the intra-instance similarity is higher than the cross-instance similarity. Moreover, we propose an inter-class loss to increase the similarity of support prototype and query prototype, aiming to guide the network to extract the intrinsic feature of each specific category.

As to the inter-class gap, the instance of the target class may share a similar feature space with the non-target class, due to the inadequacy of the support data. Therefore, we propose an Inter-class Feature Separation Module (IFSM) to distance the inter-class representations. Specifically, we reduce the prototype similarity between different categories to make the prototype discriminating. Moreover, due to the setting of FSS, the background area may contain the latent non-target classes. To distinguish the foreground with the latent instance [Yang *et al.*, 2021] of the non-target class in the background region, we leverage superpixel-guided clustering [Li *et al.*, 2021] and propose a background SLIC (B-SLIC) method to divide the background into several sub-areas. Then we present a novel loss to enlarge the distance between the support prototype and the background prototypes of each sub-area. In this way, the separation performance is improved to secernate the different categories, especially those with highly analogous representations.

Combining the above building blocks, we propose a uniform network to fill both the intra-class and inter-class gaps, named FGNet. To evaluate the performance of FGNet, we conduct extensive experiments and ablation studies. Experimental results show that FGNet surpasses previous SOTAs on both PASCAL-5^i [Everingham *et al.*, 2010] and COCO-20^i [Lin *et al.*, 2014] datasets.

In summary, our main contributions are as follows:

- We propose FGNet, a uniform prototypical learning network to fill both the intra-class and inter-class gaps for few-shot segmentation.
- We introduce two modules, i.e., SAM and IFSM, to get around the intra-class appearance discrepancy and separate the prototype of different classes, respectively. We also present several new losses and B-SLIC to further improve the separation performance of FGNet.
- Extensive experiments show that FGNet surpasses other prevalent FSS approaches and achieves state-of-the-art performances on both PASCAL-5^i and COCO-20^i on the metric of mean intersection over union (MIoU).

2 Related Work

Semantic segmentation Semantic segmentation is a fundamental task in computer vision, which classifies each pixel into a pre-defined category. The mainstream paradigm is based on the fully convolutional network (FCN) [Long *et al.*, 2015], which replaces all the linear layers with the convolutional layers. Recent breakthroughs in semantic segmentation leverage the encoder-decoder structure for better feature extraction [Chen *et al.*, 2018], the dilated convolution to enlarge the receptive field [Mehta *et al.*, 2018] and the attention mechanism to model long-range dependency [Strudel *et al.*, 2021; Xie *et al.*, 2021a]. However, these approaches rely heavily on large-scale annotated datasets, resulting in a poor adaptation ability to the unseen classes with only a handful of the annotated samples.

Few-shot learning Few-shot learning aims to learn the generalization ability to conduct classification on unseen categories with only a handful of training samples available. Existing methods can be roughly divided into two branches, i.e., meta-learning based approaches [Baik *et al.*, 2021; Xu *et al.*, 2021; Ding *et al.*, 2021] and metric-learning based approaches [Doersch *et al.*, 2020; Chen *et al.*, 2022]. The main idea of the former is to improve the capability of fast adaptation to the novel classes. In the latter approaches, the distance of similarity measurement is employed to seek for the relevance of the support-query pair. Different from few-shot classification, few-shot segmentation predicts the mask in the pixel-level, which is different and more challenging than the classification task.

Few-shot segmentation FSS is a challenging task that extends semantic segmentation to the few-shot scenario. It requires conducting pixel-wise prediction of the unseen categories with only a small number of annotated samples. OSLSM [Shaban *et al.*, 2017] first introduced the task of FSS, which proposes a two-branch network based on meta-learning strategy. Recently, metric-learning based approaches [Wang *et al.*, 2019; Liu *et al.*, 2020] are proposed for FSS, which constrcuts a global descriptor for each specific category, named prototype. The later works mainly focus on improving the quality of prototypes [Li *et al.*, 2021; Yang *et al.*, 2020; Liu *et al.*, 2022], enhancing the matching mechanism [Wang *et al.*, 2020; Zhuge and Shen, 2021], leveraging background information [Dong *et al.*, 2021; Tang *et al.*, 2021] and introducing memory networks [Xie *et al.*, 2021b; Wu *et al.*, 2021].

Despite their success, existing approaches can hardly eliminate the intra-class and inter-class gaps. Although the prototype is refined to be comprehensive, the intra-class appearance difference remains inevitable. Such a gap causes an obstacle for predicting the mask of query images whose features are not overlapped with the limited support data. Moreover, there is little attention to the inter-class classification gap, leading to the ambiguous decision boundary to classify the non-target data with a similar representation to the target class. Therefore, this gives the motivation of this paper: *can we fill both the gaps of FSS in a uniform framework to enhance the performance*?

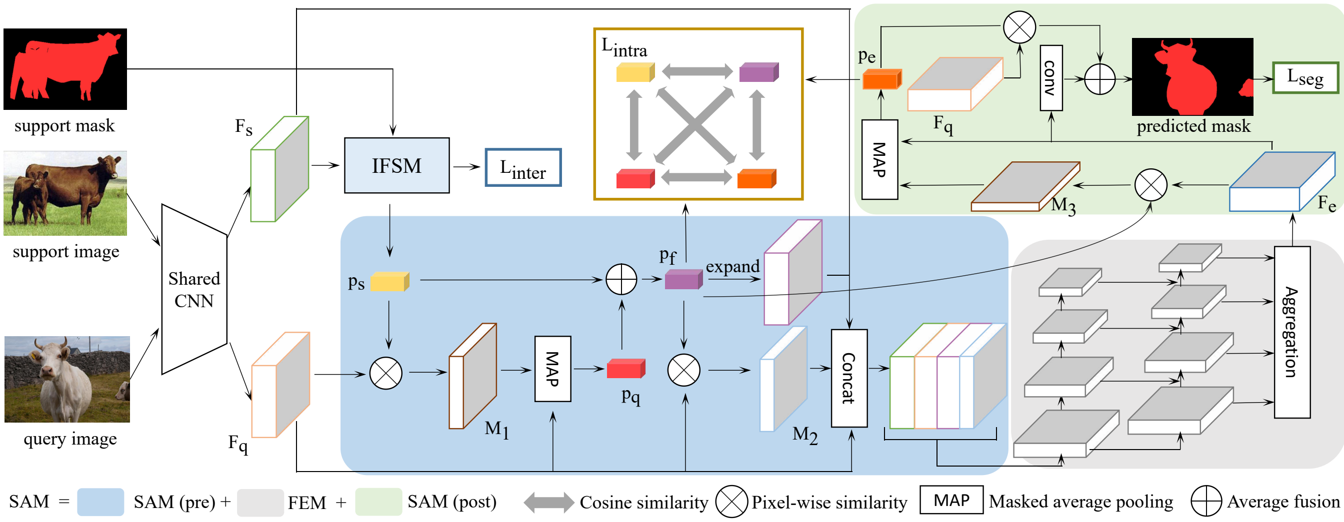

Figure 2: Overall architecture of FGNet. The Self-Adaptive Module (SAM) consists of SAM (pre), FEM and SAM (post). SAM and IFSM are the core modules of FGNet, aiming at reducing the intra-class gap and the inter-class gap, respectively.

3 Method

3.1 Problem Definition

Different from the classic semantic segmentation, FSS aims at learning the generalization ability to adapt to unseen categories. Specifically, models are trained on the set of categories C_{train} and tested on the set of novel categories C_{test}, where $C_{train} \cap C_{test} = \phi$. Both the training set $D_{train} = \{I_{S/Q}, M_{S/Q}\}$ and test set $D_{test} = \{I_{S/Q}, M_{S/Q}\}$ are composed of several episodes, where $I \in R^{H\times W\times 3}$ denotes the RGB image and $M \in R^{H\times W}$ represents the binary mask. The subscripts S and Q stand for support and query, respectively. We follow the episodic fashion [Shaban *et al.*, 2017] to train and test our model. Each episode is composed of k support samples $\{(I_S^i, M_S^i)\}_{i=1}^k$ and a query sample (I_Q, M_Q), which share the same category c. With a batch size of B, the model predicts the query mask $\tilde{M}_Q$ to approximate the corresponding ground-truth mask M_Q.

3.2 Overview of FGNet

The overall architecture of FGNet is illustrated in Figure 2. Such a network is composed of two core modules, i.e., a Self-Adaptive Module (SAM) and an Inter-class Feature Separation Module (IFSM), focusing on handling the intra-class variation and separating the target class with the classes that share similar features, respectively.

The overview data flow of FGNet is as follows. First, the support image and the query image are fed into a shared convolutional neural network (CNN) [LeCun *et al.*, 1989] for feature extraction. Through IFSM, we obtain an inter-class loss L_{inter} for distancing the representations of different classes. As shown in Figure 2, SAM consists of three parts, i.e., SAM (pre), feature enhancement module (FEM) and SAM (post). To narrow the intra-class gap, SAM exploits the query feature and calculates a query prototype for self-alignment. Subsequently, features and similarity maps are concatenated and fed into an FPN-like network [Tian *et al.*, 2020] for feature enhancement. This module aims to rectify the scale inconsistency and refine the feature in a multi-scale manner. Leveraging the enhanced feature, we generate an enhanced prototype, which activates the query feature to predict the mask. Moreover, the enhanced feature passes through three 1×1 convolutional blocks, followed by a softmax operation to predict another mask. The average fusion of the two predicted masks forms the final prediction $\tilde{M}_Q$. Therefore, the segmentation loss L_{seg} is calculated by the binary cross entropy loss of predicted mask $\tilde{M}_Q$ and ground-truth M_Q. In addition, we calculate the intra-class loss L_{intra} based on these prototypes to improve the descriptor similarity and compact the feature space of the same class. Accordingly, the total loss function L is formulated as:

$$L = \alpha_1 L_{inter} + \alpha_2 L_{intra} + \alpha_3 L_{seg} \qquad (1)$$

where α_1, α_2 and α_3 are balanced factors, and we empirically set $\alpha_1 = 0.25$, $\alpha_2 = 0.25$ and $\alpha_3 = 0.5$, respectively. We dive into the details of SAM and IFSM below.

3.3 Self-Adaptive Module

Despite great efforts to refine the prototype [Li *et al.*, 2021; Liu *et al.*, 2020], the huge intra-class variation remains inevitable due to the scarcity of support data and the diversity of query data. Therefore, we design SAM to exploit the query feature and establish a query prototype to match the query feature itself. The query prototype is more effective to predict the query mask, since the intra-instance similarity is much higher than the traditional cross-instance similarity. Such a query prototype is homologous to the corresponding query feature. That is, it is suitable to fill the intra-class gap and resolve the issue from the intra-instance perspective. Moreover, we propose an intra-class loss to improve the similarity of the support prototype and the query prototype, which guides the

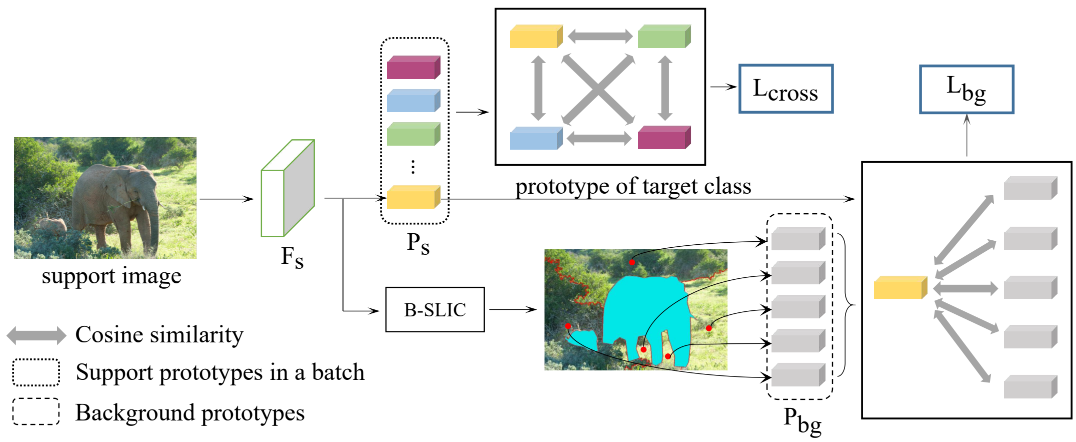

Figure 3: Overall pipeline of IFSM.

network to extract more intrinsic features of each specific category.

Given the query feature F_q and the support prototype p_s, we first calculate the similarity map M_1 through pixel-wise cosine similarity, and then generate a query prototype p_q by a masked average pooling operation, formulated as:

$$p_q = \frac{\sum_{(x,y)} F_q^{(x,y)} \mathbb{1}[M_1^{(x,y)} > \mu]}{\sum_{(x,y)} \mathbb{1}[M_1^{(x,y)} > \mu]} \tag{2}$$

where (x, y) denotes the coordinate, μ represents a threshold to activate M_1, and $\mathbb{1}$ is an indicator function. Each pixel $M_1^{(x,y)} \in [0, 1]$ stands for the confidence of the foreground. The selection of μ is significant for establishing a query prototype for self-alignment. Here we set $\mu = 0.7$ empirically. As the foreground region is of high similarity and sensitive to the noise, we need to select high-confidence features to obtain the prototypes, which are further used for self-alignment. Note that, each activation threshold of the similarity map in this paper is 0.5 unless explicitly stated.

Subsequently, we establish a fused prototype p_f through the aggregation of p_s and p_q, computed by:

$$p_f = \beta_1 p_s + \beta_2 p_q \tag{3}$$

where β_1 and β_2 represent the weights for prototype fusion and we set $\beta_1 = \beta_2 = 0.5$ empirically. Then we activate the query feature by p_f. Such a self-matching process generates a high-quality similarity map M_2, as the intra-instance information is fully exploited for self-adaptation.

Motivated by [Tian *et al.*, 2020], we employ an FPN-like network (FEM) for feature enhancement. The concatenation of F_s, F_p, M_2 and the expansion of p_f forms the input of FEM. Such a network outputs the enhanced feature F_e with comprehensive and multi-scale information of the specific class. Then we activate F_e by p_f to compute the similarity map M_3 that is further utilized to generate the enhanced prototype p_e, which is formulated by:

$$p_e = \frac{\sum_{(x,y)} F_e^{(x,y)} \mathbb{1}[M_3^{(x,y)} > \tau]}{\sum_{(x,y)} \mathbb{1}[M_3^{(x,y)} > \tau]} \tag{4}$$

where τ is a threshold of confidence to establish the enhanced prototype p_e, and $\mathbb{1}$ is an indicator function. Here we set $\tau = 0.5$ empirically.

Employing the above four prototypes, we define the intra-class loss function L_{intra} as:

$$L_{intra} = 1 - \sum_{p_i \in \mathfrak{S}P} \sum_{p_j \in \mathfrak{S}P} \frac{cos(p_i, p_j)\mathbb{1}[p_i \neq p_j]}{\mathbb{1}[p_i \neq p_j]} \tag{5}$$

where cos is the cosine similarity and $\mathfrak{S}P$ denotes the permutation of the prototype set $P = \{p_s, p_q, p_f, p_e\}$. We compute the cosine similarity between prototype pairs to narrow their discrepancy. This intra-class loss aims to guide the network to extract more intrinsic features of a specific category, despite the appearance gap between the support image and the query image.

Following the procedure described in Section 3.1, we obtain the final mask $\tilde{M}_Q$. Thus the segmentation loss L_{seg} is formulated as:

$$L_{seg} = BCE(\tilde{M}_Q, M_Q) \tag{6}$$

where BCE stands for the binary cross entropy loss of the predicted query mask $\tilde{M}_Q$ and its ground-truth M_Q.

3.4 Inter-class Feature Separation Module

IFSM is divided into two branches, as shown in Figure 3. On the one hand, reducing the similarity of prototypes with different classes helps to improve the separation performance [Okazawa, 2022] on differnet category. On the other hand, background regions usually contain latent classes [Yang *et al.*, 2021], while they are ignored due to the irrelevance of the target class. Therefore, we take both scenarios into account to enlarge the representation distance of the target class and non-target class.

Given a batch size B, we first calculate the support prototype for each episode and obtain the support prototype set $P_s = \{p_s^{c_1}, p_s^{c_2}, ..., p_s^{c_B}\}$, where the superscript c_i represents the category of the i-th prototype. Accordingly, the cross-class loss L_{cross} is formulated by:

$$L_{cross} = \sum_{i=1}^{B}\sum_{j=1}^{B} \frac{cos(p_s^{c_i}, p_s^{c_j})\mathbb{1}[c_i \neq c_j]}{\mathbb{1}[c_i \neq c_j]} \tag{7}$$

where $\mathbb{1}$ and cos are defined the same as those in Eqs. (4) and (5), respectively. The average similarity of the pairwise prototypes with different classes forms the cross-class loss. Such a loss is introduced to separate the representation space of distinct categories, resulting in the improvement of predicting the ambiguous data near the decision boundary.

Furthermore, we fully exploit the background information, as the ignored latent-class instance may hide in the background region due to the special setting of FSS. Compared with the foreground instance, the information of the background is extremely complicated. Specifically, the background area contains not only the noncontinuous stuff, e.g., sky, but also other continuous things of the non-target classes. Hence, we divide the background region into N_{sub} sub-areas, motivated by the superpixel method SLIC [Achanta *et al.*, 2012]. Such a strategy, which is called B-SLIC by us, aims to cluster the pixels with similar representations to form several small sub-regions in the background area. Our B-SLIC operation takes the pixel-level feature and the coordinate into account to calculate the distance for clustering, inspired by [Irving, 2016]. Thus the distance D between two different pixels is calculated by:

$$D = \sqrt{(d_f)^2 + (d_c/m)^2} \tag{8}$$

where d_f and d_c denote the Euclidean distance of feature and coordinate spaces of the two pixels. The balanced factor m is set to be $m = \sqrt{N_{bg}/N_{sub}}$ [Achanta *et al.*, 2012], where N_{bg} represents the total number of background pixels. Note that, the setting of N_{sub} follows the strategy in [Li *et al.*, 2021]. Accordingly, we obtain N_{sub} background regions and establish the background prototype set $P_{bg} = \{P_{bg}^1, P_{bg}^2, ..., P_{bg}^{N_{sub}}\}$. Therefore, the background-class loss L_{bg} is formulated by:

$$L_{bg} = \frac{1}{N_{sub}} \sum_{i=1}^{N_{sub}} cos(p_s, p_{bg}^i) \tag{9}$$

Finally, according to Eqs. (7) and (9), the total inter-class loss L_{inter} is calculated by:

$$L_{inter} = \gamma_1 L_{cross} + \gamma_2 L_{bg} \tag{10}$$

where γ_1 and γ_2 are the balanced factors and we set $\gamma_1 = \gamma_2 = 0.5$ empirically. Such a loss targets to reduces the representation similarity of different categories, which enlarges the distance between classes and refines the ambiguous decision boundary.

4 Experiments

Datasets and Metric. To evaluate the performance of FGNet, we conduct experiments on two widely-used FSS datasets, i.e., PASCAL-5^i [Shaban *et al.*, 2017] and COCO-20^i [Lin *et al.*, 2014]. PASCAL-5^i and COCO-20^i are derived from the traditional segmentation datasets Pascal VOC 2012 [Everingham *et al.*, 2010] and MS COCO [Lin *et al.*, 2014], with the extra annotations in SDS [Hariharan *et al.*, 2014] and FWB [Nguyen and Todorovic, 2019], respectively. The categories are partitioned into four equal splits for cross-validation. Specifically, three splits are selected for training, while the rest is for evaluation. During inference, 1k support-query pairs in PASCAL-5^i and 20k support-query pairs in COCO-20^i are randomly selected for evaluation. Besides, we use MIoU as our primary metric to evaluate FGNet under both 1-shot and 5-shot settings.

Implementation Details. The prevalent backbone ResNet [He *et al.*, 2016] pretrained on ImageNet [Deng *et al.*, 2009] is employed as our feature extractor. The features in block2 and block3 are concatenated to produce the feature map. We use SGD optimizer to train FGNet, with 0.9 momentum and 5e-3 initial learning rate. To separate different classes, we set a large batch size of 16. All images are cropped to 473×473 resolution and augmented by random horizontal flipping. Moreover, we remove the last ReLU for better generalization [Yang *et al.*, 2021].

4.1 Comparison with State-of-the-art

To evaluate the effectiveness of FGNet, we report our main results on PASCAL-5^i and COCO-20^i datasets. Employing the commonly-used backbone ResNet-101, our method achieves the best mean performances in both 1-shot and 5-shot scenarios on both datasets, compared with several previous state-of-the-art approaches.

PASCAL-5^i We list our results of PASCAL-5^i in Table 1. Our method is superior to other top-performing approaches with respect to MIoU under both 1-shot and 5-shot settings. Specifically, in the 1-shot task, FGNet reaches 68.6% MIoU, which improves the previous SOTA [Okazawa, 2022] by 1.1% MIoU. In the 5-shot task, our method achieves 73.3% MIoU, outperforming the previous best performance [Fan *et al.*, 2022] by 0.8%. Moreover, FGNet obtains high performances on split-0 and split-2 in both 1-shot and 5-shot scenarios. Note that, the improvements of the 1-shot scenario are higher than the 5-shot scenario for most models.

COCO-20^i We present our results of the challenging COCO-20^i dataset in Table 2. As shown in the table, our method outperforms the previous methods by a large margin and achieves separately 48.1% MIoU and 54.1% MIoU under 1-shot and 5-shot settings. In particular, FGNet exceeds the previous best performance [Okazawa, 2022] by 1.2% MIoU and 0.8% MIoU in 1-shot and 5-shot scenarios, respectively. Interestingly, the improvement of the 1-shot task is also greater than the 5-shot task for most models. We believe that the situation is not a coincidence and we will further discuss the possible reason in Section 4.3.

	1-shot					5-shot				
Method	split-0	split-1	split-2	split-3	mean	split-0	split-1	split-2	split-3	mean
PPNet [Liu *et al.*, 2020]	52.7	62.8	57.4	47.7	55.2	60.3	70.0	69.4	60.7	65.1
PFENet [Tian *et al.*, 2020]	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
ASGNet [Li *et al.*, 2021]	59.8	67.4	55.6	54.4	59.3	64.6	71.3	64.2	57.3	64.4
MLC [Yang *et al.*, 2021]	60.8	71.3	61.5	56.9	62.6	65.8	74.9	71.4	63.1	68.8
HSNet [Min *et al.*, 2021]	67.3	72.3	62.0	**63.1**	66.2	<u>71.8</u>	74.4	67.0	**68.3**	70.4
SSP [Fan *et al.*, 2022]	63.7	70.1	<u>66.7</u>	55.4	64.0	70.3	**76.3**	<u>77.8</u>	65.5	<u>72.5</u>
IPRNet [Okazawa, 2022]	<u>67.8</u>	**74.6**	65.7	62.2	<u>67.5</u>	70.0	<u>75.9</u>	71.8	<u>65.8</u>	70.9
FGNet (Ours)	**69.4**	<u>73.8</u>	**68.3**	<u>62.8</u>	**68.6**	**72.8**	75.7	**79.4**	65.3	**73.3**

Table 1: Performance on PASCAL-5^i in MIoU with per-split results under 1-shot and 5-shot settings, using the backbone of ResNet-101. The best and second-best results are in bold and underlined, respectively.

	1-shot					5-shot				
Method	split-0	split-1	split-2	split-3	mean	split-0	split-1	split-2	split-3	mean
PFENet [Tian *et al.*, 2020]	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	27.4
MLC [Yang *et al.*, 2021]	**51.1**	38.7	28.5	31.6	37.5	**57.8**	47.1	37.8	37.6	45.1
HSNet [Min *et al.*, 2021]	37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5
SSP [Fan *et al.*, 2022]	39.1	45.1	42.7	41.2	42.0	47.4	54.5	50.4	49.6	50.2
IPRNet [Okazawa, 2022]	42.9	<u>50.6</u>	<u>46.8</u>	**47.4**	46.9	<u>50.7</u>	<u>58.3</u>	<u>52.8</u>	<u>51.3</u>	<u>53.3</u>
FGNet (Ours)	<u>44.2</u>	**51.9**	**49.4**	<u>47.0</u>	**48.1**	49.8	**58.8**	**55.6**	**52.3**	**54.1**

Table 2: Performance on COCO-20^i in MIoU with per-split results under 1-shot and 5-shot settings, using the backbone of ResNet-101. The best and second-best results are in bold and underlined, respectively.

SAM	IFSM	split-0	split-1	split-2	split-3	mean
		59.8	66.5	55.3	57.1	59.7
✓		**73.1**	74.8	64.3	**68.5**	70.2
	✓	69.1	67.6	71.1	57.8	66.4
✓	✓	72.8	**75.7**	**79.4**	65.3	**73.3**

Table 3: Ablation results of the 5-shot setting on PASCAL-5^i for investigating the influence of Self-Adaptive Module (SAM) and Interclass Feature Separation Module (IFSM) for FGNet.

	split-0	split-1	split-2	split-3	mean
w/o L_{intra}	67.8	74.1	**72.0**	62.2	69.0
w/ L_{intra}	**73.1**	**74.8**	64.3	**68.6**	**70.2**

Table 4: Ablation results of the 5-shot setting on PASCAL-5^i for investigating the influence of using the intra-class loss in SAM.

4.2 Ablation Study

An ablation experiment is conducted to verify the necessity of SAM and IFSM, which are the core modules of FGNet. The results are presented in Table 3. The performance of the vanilla model (using similarity map M_1 as the final prediction similar to [Wang *et al.*, 2019]) without SAM and IFSM is 59.7% MIoU. With the incorporation of SAM, the model obtains an improvement of 10.5% MIoU. Besides, the introduction of IFSM increases the MIoU by 6.7%. Each module gains a significant improvement, compared with the vanilla approach. Furthermore, combined with both SAM and IFSM, our method achieves 73.3% MIoU, which is 13.6% MIoU higher than the vanilla network. Therefore, we dive into investigating how SAM and IFSM narrow the intra-class and inter-class gaps separately.

4.3 Intra-class Gap Reduction

To make our self-adaptive method more easily understood, we conduct experiments and analyses to demonstrate how SAM narrows the intra-class appearance gap.

Self-Adaptive process. The visualization results of the self-adaptive procedure of SAM are illustrated in Figure 4. First, we employ the support prototype to activate the query feature to obtain the activation of the similarity map M_1. Notice that, M_1 is unsatisfactory due to the inadequate intersection of support data and query data. Taking the first row in Figure 4 for an example, the support image contains only the pattern of cat head and claws, which brings difficulty in recognizing cat body in the query image. To fill the intra-class gap, we calculate the query prototype (by Eq. (2) and Eq. (3)) and leverage it for self-alignment. Specifically, we generate a fused prototype by Eq. (4). Such a fused prototype contains the intra-instance information of the query sample, which benefits the activation of the query feature itself. With the activation of the fused prototype, we obtain a relatively high-quality mask M_2. After the enrichment of FEM, the enhanced feature contains more comprehensive information of the discriminating category, contributing to generating the enhanced prototype, and this process assists to match with the query feature for the final prediction.

Thresholds of similarity maps. The thresholds of μ in M_1 and τ in M_3 are significant for feature selection and prototype establishment. We explore the best combination of the

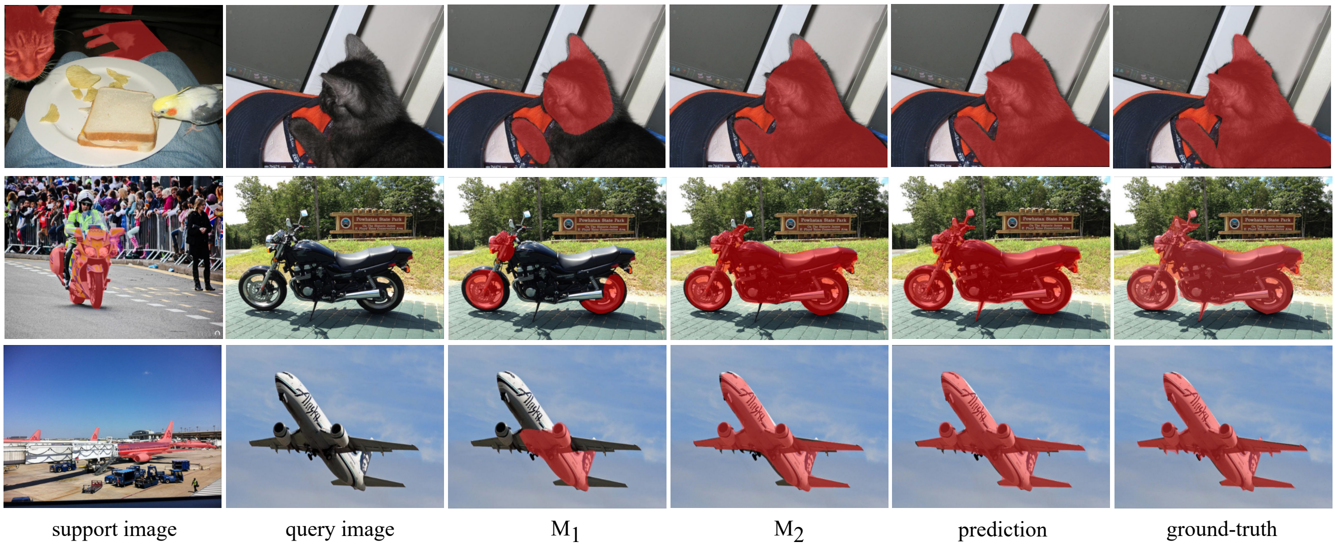

Figure 4: Qualitative results of the 1-shot setting in COCO-20^i. The sequence of M_1, M_2 and prediction illustrates the process that SAM conducts self-adaptation for high-quality mask generation. Note that, M_1 and M_2 are the activation results of the similarity maps with the thresholds μ and τ, respectively.

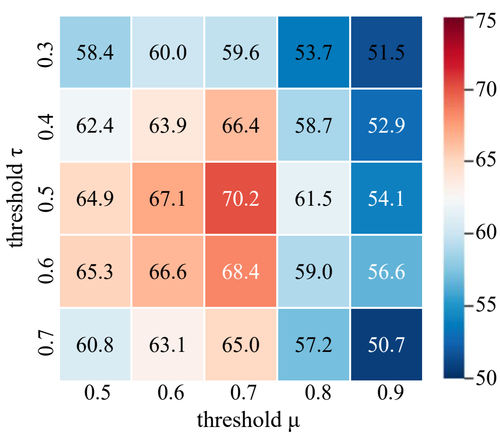

Figure 5: Results of different combination choices for the similarity map thresholds μ and τ.

L_{cross}	L_{bg}	split-0	split-1	split-2	split-3	mean
✓		65.7	68.6	70.2	55.3	65.0
	✓	61.9	65.3	68.7	**58.9**	63.7
✓	✓	**69.1**	**67.6**	**71.1**	57.8	**66.4**

Table 5: Ablation results of the 5-shot setting on PASCAL-5^i in exploring the effectiveness of the cross-class loss and the background-loss in IFSM.

two thresholds and the results are shown in Figure 5. The combination of $\mu = 0.7$ and $\tau = 0.5$ achieves the best MIoU performance. We think that the high-confidence feature in M_1 is important to capture the underlying characteristics of the query sample. Thus $\mu = 0.7$ is suitable to activate the query feature itself and generate a reasonable query prototype. Moreover, an intermediate threshold $\tau = 0.5$ in M_3 is appropriate, as the fused prototype and the enhanced feature require taking more acceptable representations into account, contributing to generating a comprehensive prototype for the final prediction.

Intra-class loss. Besides the self-adaptive process, SAM also employs an extra intra-class loss L_{intra} to reduce the intra-class gap. Such a loss aims to guide the backbone network to extract more intrinsic features of the discriminating category rather than the superficial appearance features. As shown in Table 4, the removal of L_{intra} decreases the performance by 1.2% MIoU. Therefore, we think that minimizing the intra-class loss is beneficial to the target class. Despite the apparent difference between the support data and query data, the network digs out the underlying and discriminating representations of each category. The intra-class loss results in the compaction of the intra-class feature space of each specific category, which eliminates the variation from another perspective.

Advantages of SAM. We summarize the advantages of SAM to narrow the intra-class appearance gap in two aspects: 1) Leveraging the self-adaptive process, we generate a prototype that caters to the query sample for self-alignment; 2) With the intra-class loss L_{intra}, the backbone tends to extract underlying representations of each specific class, which compacts the intra-class prototype space for an accurate prediction. Furthermore, as mentioned earlier, the improvement of the 1-shot task is higher than the 5-shot task. We think that the 1-shot task benefits more from the self-adaptation process, as the prototype under the 1-shot setting is more unreliable. With the feature enhancement and self-adaptive mechanism, the 1-shot task owns ample room for improvement compared with the 5-shot scenario.

split-0	↑	split-2	↑
0 aeroplane	9.7	10 dining-table	22.0
1 bicycle	17.3	11 dog	11.3
2 bird	4.1	12 horse	5.4
3 boat	11.6	13 motorbike	15.9
4 bottle	2.4	14 person	23.8

Table 6: MIoU results of the improvement using our method IFSM compared with the vanilla model on each specific class of split-0 and split-2 of PASCAL-5^i. Note that, ↑ represents (% MIoU) improvement.

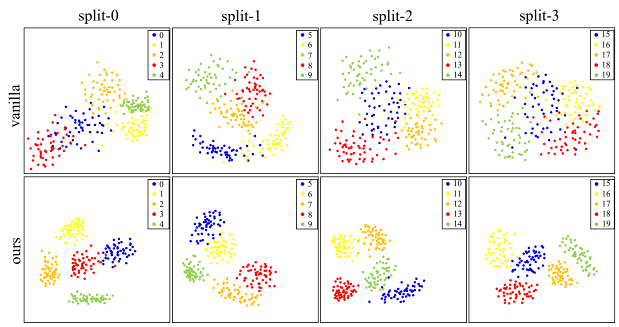

Figure 6: Visualization results of the prototypes for prediction by t-SNE on the different four splits of PASCAL-5^i. The first row and the second row demonstrate the results of the vanilla model and our approach, respectively.

4.4 Inter-class Gap Reduction

We carry out experiments and analyses to demonstrate how IFSM overcomes the inter-class classification gap.

Inter-class loss. We investigate the influence of the cross-class loss L_{cross} and the background separation loss L_{bg}, which are the two parts of the inter-class loss L_{inter}. As shown in Table 5, the removal of L_{cross} and L_{bg} decreases the result by 2.7% MIoU and 1.4% MIoU, respectively. Therefore, both L_{cross} and L_{bg} are significant to the total inter-class loss, since L_{cross} distances the feature spaces between different categories and L_{bg} differentiates the foreground with the latent instances in the background region, which reduces the similarity of the foreground prototype with the latent non-target prototypes and rectifies the decision boundary.

Performance on similar classes. To evaluate the effectiveness of different classes that are difficult to distinguish, we conduct sufficient experiments to obtain the improvement of each category. For a fair comparison, we select 1k support-query pairs for each category in split-0 and split-2 (the significantly improved splits) under the 5-shot setting. As shown in Table 6, the improvement of 10 dining-table and 14 person are 22.0% MIoU and 23.8% MIoU, respectively, which are much higher than other classes. The features of such categories are likely to be ambiguous with other classes since they usually appear in complex scenarios. Specifically, 10 dining-table is usually covered with cluttered tablewares and 14 person appears in diverse scenes with various poses and decorations. Moreover, the performance of 1 bicycle and 13 motorbike are increased by 17.3% MIoU and 15.9% MIoU, respectively. The representations of these two classes are similar to each other, leading to the challenge of accurate prediction without the consideration of the inter-class relation. Therefore, with the incorporation of IFSM, the network is guided to extract intrinsic representation for each specific category, resulting in distancing the feature space of similar classes that are difficult to classify.

Advantages of IFSM. IFSM closes the inter-class gap by separating the prototype spaces of the different categories. As shown in Figure 6, the prototypes (visualized by t-SNE [Van der Maaten and Hinton, 2008]) of each category are adjacent and mixed with respect to the vanilla model. However, the prototype distance is enlarged with IFSM, which reduces the ambiguity of classification and rectifies the decision boundary. On the one hand, with the cross-class loss, the network distances the representations of similar classes. On the other hand, the background separation loss further reduces the prototype similarity of the target class with other latent non-target classes. Therefore, IFSM enlarges the margin of prototypes of different categories to improve the separation performance, which fills the inter-class gap.

5 Conclusion

In this paper, we proposed FGNet to fill the intra-class and inter-class gaps for few-shot segmentation. To narrow the intra-class gap, we introduced a Self-Adaptive Module (SAM) to fully exploit the query representations for self-alignment. Moreover, we proposed an Inter-class Feature Separation Module (IFSM) to separate the prototype spaces of different classes, which bridges the inter-class gap. In addition, we put forward B-SLIC to take the latent classes in the background region into account, and designed several new losses to further improve the separation performance of FGNet. Experimental results show that FGNet effectively fills both the gaps, and meanwhile achieves SOTA performances on multiple datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62172385), the Anhui Initiative in Quantum Information Technologies (No. AHY150300), and the Innovation Program for Quantum Science and Technology (No. 2021ZD0302900).

References

[Achanta *et al.*, 2012] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012.

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017.

[Baik *et al.*, 2021] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee.

Meta-learning with task-adaptive loss function for few-shot learning. In *ICCV*, pages 9465–9474, 2021.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[Chen *et al.*, 2022] Haoxing Chen, Huaxiong Li, Yaohui Li, and Chunlin Chen. Multi-level metric learning for few-shot image recognition. In *ICANN*, pages 243–254. Springer, 2022.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[Ding *et al.*, 2021] Nan Ding, Xi Chen, Tomer Levinboim, Sebastian Goodman, and Radu Soricut. Bridging the gap between practice and pac-bayes theory in few-shot meta-learning. *NeurIPS*, 34:29506–29516, 2021.

[Doersch *et al.*, 2020] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *NeurIPS*, 33:21981–21993, 2020.

[Dong and Xing, 2018] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018.

[Dong *et al.*, 2021] Kaiqi Dong, Wei Yang, Zhenbo Xu, Liusheng Huang, and Zhidong Yu. Abpnet: Adaptive background modeling for generalized few shot segmentation. In *ACM MM*, pages 2271–2280, 2021.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[Fan *et al.*, 2022] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *ECCV*, pages 701–719. Springer, 2022.

[Hariharan *et al.*, 2014] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312. Springer, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Irving, 2016] Benjamin Irving. maskslic: Regional superpixel generation with application to local pathology characterisation in medical images. *arXiv preprint arXiv:1606.09518v2*, 2016.

[LeCun *et al.*, 1989] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[Li *et al.*, 2021] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *CVPR*, pages 8334–8343, 2021.

[Lin *et al.*, 2014] 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 *ECCV*, pages 740–755. Springer, 2014.

[Liu and Qin, 2020] Jinlu Liu and Yongqiang Qin. Prototype refinement network for few-shot segmentation. *arXiv preprint arXiv:2002.03579*, 2020.

[Liu *et al.*, 2020] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *ECCV*, pages 142–158. Springer, 2020.

[Liu *et al.*, 2022] 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 *CVPR*, pages 11553–11562, 2022.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Mehta *et al.*, 2018] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *ECCV*, pages 552–568, 2018.

[Min *et al.*, 2021] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*, pages 6941–6952, 2021.

[Nguyen and Todorovic, 2019] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, pages 622–631, 2019.

[Okazawa, 2022] Atsuro Okazawa. Interclass prototype relation for few-shot segmentation. In *ECCV*, pages 362–378. Springer, 2022.

[Shaban *et al.*, 2017] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017.

[Siam *et al.*, 2021] Mennatullah Siam, Naren Doraiswamy, Boris N Oreshkin, Hengshuai Yao, and Martin Jagersand. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings. In *IJCAI*, pages 860–867, 2021.

[Silberman *et al.*, 2012] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

[Strudel *et al.*, 2021] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, pages 7262–7272, 2021.

[Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[Tang *et al.*, 2021] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *CVPR*, pages 3918–3928, 2021.

[Tian *et al.*, 2020] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *TPAMI*, 2020.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[Wang *et al.*, 2019] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, pages 9197–9206, 2019.

[Wang *et al.*, 2020] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, pages 730–746. Springer, 2020.

[Wu *et al.*, 2021] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *ICCV*, pages 517–526, 2021.

[Xie *et al.*, 2021a] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021.

[Xie *et al.*, 2021b] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *ICCV*, pages 7293–7302, 2021.

[Xu *et al.*, 2021] Hui Xu, Jiaxing Wang, Hao Li, Deqiang Ouyang, and Jie Shao. Unsupervised meta-learning for few-shot learning. *PR*, 116:107951, 2021.

[Yang *et al.*, 2020] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *ECCV*, pages 763–778. Springer, 2020.

[Yang *et al.*, 2021] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *ICCV*, pages 8721–8730, 2021.

[Yu *et al.*, 2020] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12416–12425, 2020.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.

[Zhuge and Shen, 2021] Yunzhi Zhuge and Chunhua Shen. Deep reasoning network for few-shot semantic segmentation. In *ACM MM*, pages 5344–5352, 2021.