Interclass Prototype Relation for Few-Shot Segmentation

Atsuro Okazawa

R&D Promotion Office AI Strategy Office,
SoftBank Corp., Tokyo, Japan
atsuro.okazawa@g.softbank.co.jp

Abstract. Traditional semantic segmentation requires a large labeled image dataset and can only be predicted within predefined classes. Solving this problem of few-shot segmentation, which requires only a handful of annotations for the new target class, is important. However, with few-shot segmentation, the target class data distribution in the feature space is sparse and has low coverage because of the slight variations in the sample data. Setting the classification boundary that properly separates the target class from other classes is an impossible task. In particular, it is difficult to classify classes that are similar to the target class near the boundary. This study proposes the Interclass Prototype Relation Network (IPRNet), which improves the separation performance by reducing the similarity between other classes. We conducted extensive experiments with Pascal-5^i and COCO-20^i and showed that IPRNet provides the best segmentation performance compared with previous research.

Keywords: Semantic Segmentation, Few-shot Segmentation, Few-shot Learning, Metric Learning

1 Introduction

Recent advances in semantic segmentation have been brought about by advanced convolutional neural networks (CNNs) [15] and large labeled image datasets [9], [18], [6], [50]. However, semantic segmentation with fully supervised learning requires a substantial number of annotations per pixel and can be time-consuming to create. To solve this problem, few-shot segmentation that requires only a handful of annotations for a new target class is important. Few-shot segmentation aims to obtain generalization ability from known classes and adapt them to new target classes via a few shots, namely, support data. However, few-shot segmentation is not under the condition in which features can be extracted from a large amount of data with all variations (Fig.1(a)), and the target class data distribution in the feature space is sparse and has low coverage (Fig.1(b)). Therefore, there is an essential problem in that it is not possible to set the classification boundary that separates the target class from other classes properly. In particular, it is difficult to classify classes that have features like those of the target class near the boundary. To tackle this important problem without increasing the

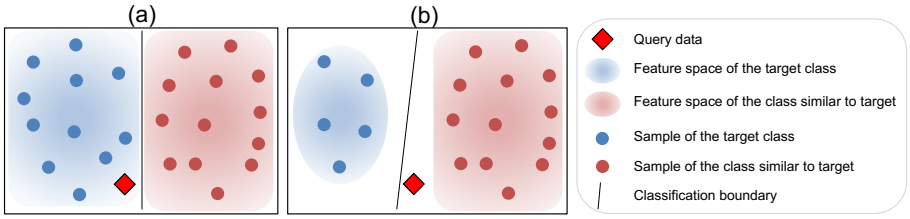

Fig. 1. Each shade is the area to which most of the samples in that class are mapped. The black line is the classification boundary calculated from the samples. (a) shows the area to be mapped from sufficient samples is given, not in the condition of few-shot. (b) shows the area mapped in the few-shot problem with a few samples. If query data is plotted near the boundary between the two classes you want to classify, we cannot detect the class that originally belonged to the target class in the case of few-shot with a narrow-mapped area.

number of shots for the target class, it is important to differentiate the features between each class when learning generalization abilities from known classes.

Few-shot segmentation is an extension of the technology based on few-shot learning [36], [32], [31], [39], [11], [28], and it tackles the more difficult task of predicting the label for each pixel, instead of predicting a single label for the entire image in few-shot learning. Few-shot learning has meta-learning [39], [11] and metric learning [36], [32], [31] as the mainstream methods. Meta-learning was introduced by Shaban et al. [29] and metric learning by Snell et al. [8] for segmentation problems. In particular, the metric learning approach to the problem of few-shot segmentation has been actively studied in recent years and has been successful. Our research is related to few-shot segmentation using metric learning. The method of few-shot segmentation using metric learning has been pushed into a global descriptor called a prototype using supporting data [30]. The support data are a few samples with a target class. The prototype is a vector representation of the features for the target class and there are many studies based on the method of inference by comparing the prototype with the features of the query image. For better prototyping and proper comparison of support and query, there are several earlier studies. For example, there are studies that have introduced a mechanism to separate the foreground and background [38], [19], [42], and studies that have introduced a multi-scale architecture [45], [34], [16]. In these studies, the prototype extracted from the support data was appropriately compared with the query data, and training was performed based on the loss function between the query data and its ground truth.

However, in few-shot segmentation with only a few shot samples, as mentioned above, there are few variations of support data using the target class, and prototype generation is performed from a sparse feature space. Therefore, it is particularly difficult to obtain a prototype that can classify classes with features similar to those of the target class. However, owing to the problem setting of the few-shot, it is not possible to increase the amount of data for the target class and make the feature space dense.

Accordingly, we propose the Interclass Prototype Relation Network (IPRNet) that improves separation performance by reducing the similarity between types to highlight the differences between the prototypes of similar classes. IPRNet has 1) the Interclass Prototype Relation Module, which aims to improve the separation performance between similar classes by reducing the similarity between prototypes of each class, and 2) the Respective Classifier Module, which aims to improve the separation performance by integrating respective estimations of the target class and background. We hypothesized that these modules could improve the separation performance of the target and other classes. In this study, we verified this hypothesis using two experiments. First, we evaluated whether the performance would improve compared to earlier research with the best performance. Second, we conducted an ablation study to verify that the modules proposed in IPRNet are effective in classifying similar classes. The contributions of this study are as follows.

- We propose a novel few-shot segmentation method called IPRNet, which improves the separation performance between the target class and other classes that are especially similar to the target.
- We evaluate Pascal-5^i [29] and COCO-20^i [18] and show that the proposed method improves the mean intersection over union (mIoU) over the existing best-performing method.
- Through an ablation study, we verified whether the proposed method is effective for classifying similar classes.

Our code is at:
https://sb-biz.primedrive.jp/v2/access?key=PGru7XxrXU-tQe1ZzB2EpQ

2 Related Work

Few-shot segmentation mostly consists of few-shot learning-based technology that improves model generalization ability and semantic segmentation technology that solves pixel-level classification problems. We describe the existing research related to these constituent requirements and issues.

Semantic Segmentation In semantic segmentation, deep neural networks based on convolutional neural networks (CNNs) [15] have been successful. Starting from fully convolutional networks [22], especially encoder-decoder structure proposed by Segnet [1] has become the basic network structure in recent semantic segmentation. Recently, a faster method Enet [27], and encoder-decoder structures that ensemble multi-scale features to express all frequency information have been proposed [43], [3], [4], [5], [48], [47]. The latest research also proposed a convolution-free and resolution deterioration-free method [49] based on the transformer [35], which is a model that uses only the attention mechanism instead of CNNs [15].

Few-shot Learning Few-shot learning focuses on the generalization ability of the model and enables learning for new class predictions using a few annotated samples. The mainstream existing methods are metric learning [36], [32], [31] and meta-learning [39], [11], [28]. The core idea of metric learning is the distance measurement, which is formulated as an optimization of the distance or similarity between the images and regions. Meta-learning focuses on achieving a high-speed learning ability by defining specific optimization functions and loss functions. Among these methods, the concept of a prototypical network [31] is widely adopted for few-shot segmentation, and it is possible to reduce the calculation cost significantly while maintaining high performance. Many methods focus on image classification, but recently few-shot segmentation has attracted attention.

Few-shot Segmentation Few-shot segmentation is an extension of few-shot learning that addresses the more difficult task of predicting a label for each pixel rather than predicting a single label for the entire image. Few-shot meta-learning was introduced into the segmentation problem by Shaban et al. [29], and there is a lot of research on its enlargement [33], [14], [2], [23]. Few-shot metric learning was successfully introduced by Snell et al. [8]. Many of the methods so far drop the problem into a 1-way classification problem in order to apply it to episodic learning [36] to acquire generalization ability. In previous research, they pushed the support data into a global descriptor to obtain a prototype that is the features of the class in the first step [30]. In the next step, the target object and background are separated by comparing the prototype with the query image [26], [46], [38], [45], [21], [41], [34], [16], [42]. In addition to these, research to absorb the difference in size, position, and orientation of the target object on the support image and the query image, and to compare them correctly have been conducted [20], [10], [44], [37], [40]. There is also a method to infer from the correlation between all positions of query data and support data [25]. However, most of them are solved by general 1-way classification problems, so only the relationship between the target class and the background can be considered. ASR [19] is a method that uses multiple latent class vectors, but the feature map channel is divided and assigned to each class. Therefore, if many classes are included, the number of channels assigned to one class will decrease and it will not work effectively. Existing methods do not fully consider the relationships between different classes, making proper classification difficult. In particular, it is the most difficult to separate from similar classes near the discriminant boundary, and to the best of our knowledge, no research has been conducted on this problem. This research proposes a novel IPRNet that focus on improvement between similar classes, which are particularly difficult to classify.

3 Problem Definition

The major difference between few-shot segmentation and general semantic segmentation is that the training and test set categories do not intersect. Particu-

larly, at the inference stage, the test set had classes that were not found during the training. Specifically, given the train set $S_{train} = \{(I^{S/Q}, M^{S/Q})\}$ and the test set $S_{test} = \{(I^{S/Q}, M^{S/Q})\}$, the categories of the two sets do not intersect ($S_{train} \cap S_{test} = \phi$). Here, $I \in R_{H \times W \times 3}$ shows an RGB image and $M \in R_{H \times W}$ shows a segmentation mask. The subscripts S and Q represent support and query, respectively. We mimicked the first one-shot segmentation study [29] and applied training and testing to an episodic learning framework. In each episode, the input to the model consists of the query image I^Q and k samples (I_i^S, M_i^S), $i \in \{1, 2, \ldots, k\}$ from the support set. All support image and query image have the same class c. Training selects (I^Q, M^Q, I_i^S, M_i^S) of batch size b set from the train set $S_{train} = \{(I^{S/Q}, M^{S/Q})\}$ and estimates the query mask $\tilde{M^Q}$ to approximate ground truth mask M^Q.

4 Proposed Method

4.1 Design guideline of network structure

As mentioned in the introduction, the target class, which has a few shots in few-shot segmentation, is difficult to classify similar classes because the data plotted in the feature space is sparse and has low coverage. To address this problem, we propose the Interclass Prototype Relation Network (IPRNet). IPRNet has two modules: the Interclass Prototype Relation Module (IPRM) and the Respective Classifier Module (RCM). These modules aim to improve the identification performance of similar classes by reducing the similarity between prototypes and extracting the differences between classes. An overview of the network is shown in Figure 2.

4.2 Interclass Prototype Relation Network

This section describes the overall flow of IPRNet. First, the support and query images were fed into pretrained shared CNNs (pretrained by ImageNet [7]) to extract features. Next, we passed the support features with the support masks through the IPRM. Through IRPM, we obtain prototypes that represent feature vectors for each class, and the value L_r that indicates the similarity between each prototype. Then, for more accurate pixel-by-pixel matching, the matching between prototypes and the query feature is performed by calculating the cosine similarity in map process. The two similarity maps can be obtained by matching with the query feature for the target class prototype and the background prototype, respectively. This process was inspired by earlier research on MLC [42]. The input to the multi-scale network is the concatenation of the support features, the query feature, and two similarity maps. The output is a relation feature valid for classification and L_m which is a multi-scale loss introduced in PFENet [34]. The multi-scale network sets up the top-down structure of FPN-like [17] by using the feature enrichment module introduced in PFENet [34] to obtain multi-scale information. This structure enables fast multi-scale aggregation, by transferring

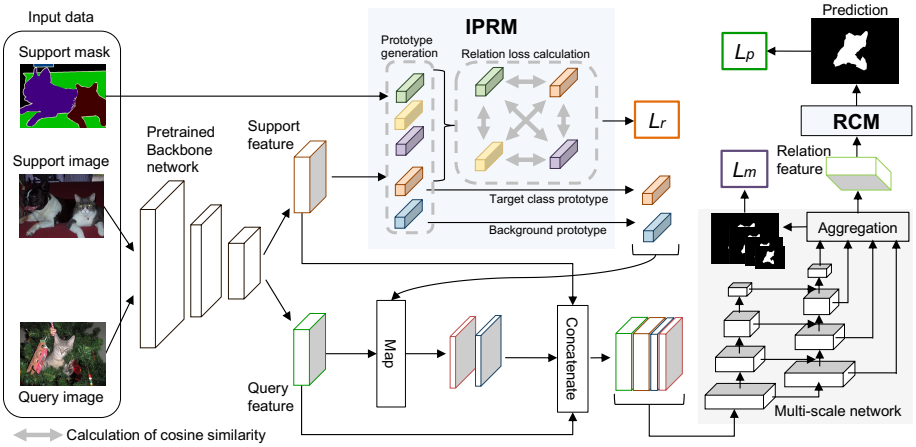

Fig. 2. The Overall architecture of the proposed method, Interclass Prototype Relation Network (IPRNet).

features from finer to coarser and by easing feature interaction. Each scale yielded segmentation results for calculating the loss. The average loss value for each scale was L_m. Finally, the relation feature, a multi-scale information-intensive feature map, is generated by fusing all the different scales into a concatenated feature map by convolution.

The relation feature is the input to the RCM. The RCM performs a discriminative process to classify the foreground and background more and obtains the final inference result $\tilde{M^Q}$. Then, the loss L_p between the inference result and ground-truth mask M^Q is calculated. The loss function, which is the cost function of training, is given by equation (1).

$$Loss = w_1 L_r + w_2 L_m + w_3 L_p \tag{1}$$

w_1, w_2, and w_3 are the weight coefficients, which were trained with $w_1 = 0.4$, $w_2 = 0.2$, and $w_3 = 0.4$, respectively. Here, we describe the details of the proposed IPRM and RCM.

4.3 Interclass Prototype Relation Module

The Interclass Prototype Relation Module (IPRM) was proposed to reduce the similarity between classes. The IPRM has a prototype generation process and relation loss calculation process. The prototype generation process calculates the prototypes of all classes present in the batch and the support images. This process obtains a prototype, which is a global descriptor of a particular class in an image by taking as input the feature map extracted from the support image and the segmentation mask paired with it. We employed the masked average pooling strategy [30] to compute the prototype for each class. The prototype P_i^c

of the c th class in the i th support image is computed using equation (2).

$$P_i^c = \frac{\sum_{x,y} F_i^{x,y} \mathbb{1}[M_i^{x,y} = c]}{\sum_{x,y} \mathbb{1}[M_i^{x,y} = c]} \tag{2}$$

F_i is the feature map extracted via the backbone network with the support image I_i^S as the input. The subscripts x and y indicate the horizontal and vertical spatial positions of the feature map, respectively. $M_i^{x,y}$ is the segmentation mask. By applying this, we can eliminate the regions of the feature map other than the specified class. Equation (2) was calculated for all the support images in the batch. The maximum number of prototypes n obtained is the number of all classes in which S_{train} has $c \in \{0, 1, 2, \ldots, n\}$.

Next, the relation loss calculation process was performed using the obtained prototypes. This process calculates the average value L_r of the cosine similarity between different classes. The relation loss L_r is calculated using equation (3,4).

$$L_r = \frac{\sum_{c_s}^{n} \sum_{c_t}^{n} Sim(P^{c_s}, P^{c_t}) \mathbb{1}[c_s \neq c_t]}{\sum_{c_s}^{n} \sum_{c_t}^{n} \mathbb{1}[c_s \neq c_t]} \tag{3}$$

$$Sim(P^{c_s}, P^{c_t}) = \frac{P^{c_s} \cdot P^{c_t}}{\|P^{c_s}\| \cdot \|P^{c_t}\|} \tag{4}$$

The c_s and c_t refer to class numbers and the difference between the prototypes of two different classes is measured by the cosine similarity expressed in the equation (4). The similarity between the prototypes of two different classes c_s,c_t computed in equation (4) is calculated for all combinations switching between pairs of prototypes to be measured using the equation (3). The average of these prototype similarity values are the relation loss L_r. The relation loss L_r is designed to improve the separation performance between each class by training the network such that the similarity between each class is low.

Further, among the prototypes calculated by the equation (2), the prototype of the target class and the prototype extracted from the background region is selected. The prototypes were compared with the query feature and their respective similarity maps were computed. These similarity maps were used as input to the multi-scale network.

4.4 Respective Classifier Module

An overview of the Respective Classifier Module (RCM) is shown in Figure 3. The RCM takes as input the relation feature, which is the output of the multi-scale network shown in Figure 2. It was designed to improve the separation performance between the target class of objects and the background by estimating each of them independently, reintegrating the results. The relation feature F_r is branched for foreground target class prediction and background prediction. Then, it is transformed into a probability distribution representation V_1

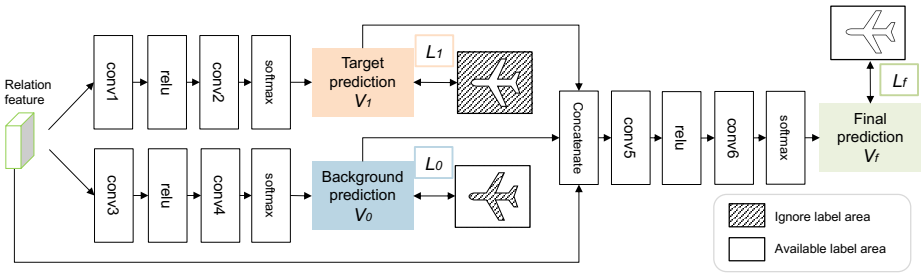

Fig. 3. Configuration diagram of the Respective Classifier Module (RCM).

and V_0 through two convolutional layers, the activation layer, and the softmax layer respectively. V_1 indicates the probability distribution of a single object in the target class, and V_0 indicates the probability distribution of the background region only. The loss calculation between the probability distribution representation V and the ground-truth mask M^Q can be expressed by equation (5).

$$L_c = \frac{-\sum_{x,y} D_c(M^Q) log(V_c^{x,y})}{\sum_{x,y} D_c(M^Q)} \tag{5}$$

$$D_c(M^Q) = \begin{cases} c & \text{if } M^Q = c, \\ 255 & \text{if } M^Q \neq c. \end{cases} \tag{6}$$

The loss is calculated by cross-entropy error. V_1, the calculated loss value L_1 is the loss function of a single object of the target class, and given probability distribution V_0, the calculated loss value L_0 is the loss function of the background region only. Equation (6) is a function that, according to the given class ID c, returns 255 pixel positions other than those corresponding to the class ID in the ground-truth mask M^Q. Two hundred and fifty-five means ignore the label, an area that is not used when calculating loss, thus eliminating the relationship between the background and foreground. This mechanism allows the RCM to acquire a discriminator through training that can make inferences about its pixels based on information about the object itself only.

The final estimation result V_f can be obtained by fusing the information from the concatenation of the foreground probability distribution V_1, background probability distribution V_0, and relation feature F_r with the two convolution layers, the activation layer, and the softmax layer. The cross-entropy error L_f was also obtained between the prediction result V_f and the ground-truth mask M^Q.

The final loss value L_p output by the RCM is obtained by the weighted addition of the loss value L_1 to the prediction result of the foreground object alone, the loss value L_0 to the prediction result of the background region only,

and the loss value L_f to the final prediction result. This can be defined by the following equation (7).

$$L_p = \alpha L_1 + \beta L_0 + \gamma L_f \tag{7}$$

α, β, and γ are the weight coefficients, which were trained by setting $\alpha = 0.15$, $\beta = 0.15$, and $\gamma = 0.7$, respectively.

5 Experiments

5.1 Datasets and Evaluation Metric

To analyze the performance of IPRNet, we selected two datasets that are widely used for few-shot segmentation, Pascal-5^i [29] and COCO-20^i [18]. Pascal-5^i [29] contains images from PASCAL VOC 2012 [9] and additional annotations from SBD [12]. For training, a total of twenty class categories were evenly divided into four splits, and model training was performed using a cross-validation method. Specifically, three splits were selected as the training data during the training process, and the remaining splits were used for testing. During testing, one thousand support-query pairs were randomly sampled and evaluated [29]. COCO-20^i [18], unlike Pascal-5^i [29], is an incredibly challenging dataset. This is because it is a large dataset having 82,081 images, and many objects are included in the images of realistic scenes. Following the FWB [26], the eighty classes of COCO-20^i [18] were evenly divided into four splits, and the same cross-validation scheme was used. To obtain more stable results, we randomly sampled 20,000 pairs of during the testing [34]. As an evaluation metric, we used the mean Intersection-over-Union (mIoU), which is commonly used in semantic segmentation.

An ablation study was conducted to verify the influence of the proposed IPRM and RCM. To verify the separation performance between similar classes, we compared the per-class IoU results of IPRNet and the baseline without our IPRM and RCM.

5.2 Implementation details

ResNet [13] is employed as the backbone network, and block2 and block3 are concatenated to generate the feature map [46]. The input support and query images were cropped to an image size of 400×400 pixels and fed into the backbone network. The initial learning rate was set to 0.05, momentum and weight decay to 0.9 and 0.0001, respectively, and the optimizer was trained with a batch size of thirty-two using the SGD optimizer. We also adapted the poly method [3], where the decay of the learning rate is achieved by multiplying by $(1 - current\_iter)^{power}$, and the *power* is set to 0.9. The pretrained backbone network is frozen so that it does not learn class-specific representations of the training data. We implemented it using Pytorch experimented with an NVIDIA A10G GPU.

In the ablation study, for the IPRM deletion, we also modified a mechanism that uses only the conventional masked average pooling strategy [30] to acquire the target class prototype and background class prototype. The RCM is completely removed and replaced with a mechanism that calculates the loss between the output of the multi-scale network and the ground truth.

5.3 Experimental results

The effectiveness of our method was evaluated on two benchmark datasets Pascal-5^i [29] and COCO-20^i [18]. We extensively experimented with the 1-shot and 5-shot on 1-way problems in various few-shot split settings using the widely used encoder networks ResNet-50 and ResNet-101. Here, an n-way k-shot implies that k samples are given for each class among the n classes. Extensive experiments show that the mIoU is improved over the conventional method in all the cases. The ablation study confirmed that removing the IPRM and RCM reduced the mIoU.

Pascal-5^i First, we describe the results for Pascal-5^i in Table 1. Our method has an improved mIoU in all conditions compared with HSNet [25], which has the highest mIoU among the conventional methods. In order of performance improvement, we first see a 1.7% mIoU improvement for ResNet-50 and 1.3% mIoU improvement for ResNet-101 in the 1-shot setting. These are followed by 0.7% mIoU improvement for ResNet-50 and 0.5% mIoU improvement for ResNet-101 in the 5-shot setting.

COCO-20^i Next, we describe the results for COCO-20^i in Table 2. Our method has improved mIoU in all conditions compared with HSNet [25], which has the highest mIoU among the conventional methods. In order of performance improvement, we first see a 6.1% mIoU improvement for ResNet-50 and 5.7% mIoU improvement for ResNet-101 in the 1-shot setting. This is followed by a 4.2% mIoU improvement for ResNet-50 and 3.8% mIoU improvement for ResNet-101 in the 5-shot setting.

Ablation study We conducted an ablation study to investigate the influence of the IPRM and RCM, which are the main components of our model. All ablation study experiments were the result of a 5-shot setup performed on the COCO-20^i dataset using the ResNet50 backbone. The results are presented in Table 3. The largest decrease in mIoU is seen in the case where both IPRM and RCM are removed by 4.4%, followed by the case where only IPRM is removed by 2.5%, and finally, the case where only RCM is removed 1.8%. The mIoU decreased by more than 0.7% when the IPRM was removed then when the RCM was removed.

Table 1. Performance on Pascal-5^i in IoU with per-splits results. Some results are from [16, 21, 25, 34, 38, 42].

Backbone	Method	1shot					5shot				
		s-0	s-1	s-2	s-3	mean	s-0	s-1	s-2	s-3	mean
ResNet-50	PANet [38]	44.0	57.5	50.8	44.0	49.1	55.3	67.2	61.3	53.2	59.3
	PPNet [21]	48.6	60.6	55.7	46.5	52.8	58.9	68.3	66.8	58.0	63.0
	PFENet [34]	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
	ASGNet [16]	58.8	67.9	56.8	53.7	59.3	63.7	70.6	64.2	57.4	63.9
	MLC [42]	59.2	71.2	**65.6**	52.5	62.1	63.5	71.6	**71.2**	58.1	66.1
	HSNet [25]	64.3	70.7	60.3	60.5	64.0	**70.3**	73.2	67.4	**67.1**	69.5
	Ours	**65.2**	**72.9**	63.3	**61.3**	**65.7**	70.2	**75.6**	68.9	66.2	**70.2**
ResNet-100	FWB [26]	51.3	64.5	56.7	52.2	56.2	54.8	67.4	62.2	55.3	59.9
	PPNet [21]	52.7	62.8	57.4	47.7	55.2	60.3	70.0	69.4	60.7	65.1
	PFENet [34]	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
	ASGNet [16]	59.8	67.4	55.6	54.4	59.3	64.6	71.3	64.2	57.3	64.4
	MLC [42]	60.8	71.3	61.5	56.9	62.6	65.8	74.9	71.4	63.1	68.8
	HSNet [25]	67.3	72.3	62.0	**63.1**	66.2	**71.8**	74.4	67.0	**68.3**	70.4
	Ours	**67.8**	**74.6**	**65.7**	62.2	**67.5**	70.0	**75.9**	**71.8**	65.8	**70.9**

Table 2. Performance on COCO-20^i in IoU with per-splits results. Some results are from [21, 25, 34, 42].

Backbone	Method	1shot					5shot				
		s-0	s-1	s-2	s-3	mean	s-0	s-1	s-2	s-3	mean
ResNet-50	PPNet [21]	28.1	30.8	29.5	27.7	29.0	39.0	40.8	37.1	37.3	38.5
	PFENet [34]	36.5	38.6	34.5	33.8	35.8	36.5	43.3	37.8	38.4	39.0
	MLC [42]	**48.0**	36.6	27.4	28.2	35.1	**54.9**	42.1	34.9	33.6	41.4
	HSNet [25]	36.3	43.1	38.7	39.2	39.2	43.3	51.3	48.2	45.0	46.9
	Ours	42.2	**48.9**	**45.5**	**44.6**	**45.3**	48.0	**55.7**	**50.7**	**50.1**	**51.1**
ResNet-100	FWB [26]	17.0	18.0	21.0	28.9	21.2	19.1	21.5	23.9	30.1	23.7
	PFENet [34]	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	27.4
	MLC [42]	**51.1**	38.7	28.5	31.6	37.5	**57.8**	47.1	37.8	37.6	45.1
	HSNet [25]	37.2	44.1	42.4	41.3	41.2	45.9	53.0	51.8	47.1	49.5
	Ours	42.9	**50.6**	**46.8**	**47.4**	**46.9**	50.7	**58.3**	**52.8**	**51.3**	**53.3**

IOU of each class To verify whether our proposed method is effective for objects with similar features that are difficult to classify, we evaluated it using COCO-20^i is an incredibly challenging dataset that has many objects in the images of realistic scenes. The evaluation is perfomed by comparing IoUs per class with the baseline, which eliminates the IPRM and RCM, and our proposed IRPNet. The results are presented in Table 4. Significant improvements in IoU were as follows 1 person 20.7%, 61 dining table 20.2%, 73 fridge 15.4%, 66 remote 13.4%, 67 keyboard 12.3%, 27 handbag 11.6%, 29 suitcase 11.5%, 25 backpack 11.3%, 74 book 11.3%, 44 knife 10.3%; and an increase of more than 10% in IoU. The classes with a lower IoU were 34 kite -8.9%, 65 mouse -8.2%, 48 apple -4.6%, 59 potted plant -3.8%, 13 park meter -3.5%, 4 motorcycle -2.8%, 10 traffic light -1.8%, 26 umbrella -1.5%, 69 microwave -1.5%, 80 toothbrush -1.0%; and a decrease of over 1% in IoU. The following is a discussion.

Table 3. Influence of the IPRM and RCM on COCO-20^i in IoU with 5-shot ResNet-50 backbone condition.

IPRM	RCM	s-0	s-1	s-2	s-3	Mean
		42.0	52.1	46.4	46.4	46.7
✔		46.8	53.6	47.9	49.0	49.3
	✔	44.9	53.2	47.4	48.9	48.6
✔	✔	**48.0**	**55.7**	**50.7**	**50.1**	**51.1**

Table 4. COCO-20^i performance in all classes of IoU experimented with 5shot ResNet-50. Baseline means the result of eliminating the IPRM and RCM.

s-0	Baseline	**Ours**	s-1	Baseline	**Ours**	s-2	Baseline	**Ours**	s-3	Baseline	**Ours**
1 Person	30.3	**51.0**	2 Bicycle	52.9	**55.8**	3 Car	35.7	**38.5**	4 Motorcycle	**56.0**	53.2
5 Airplane	73.0	**76.0**	6 Bus	69.0	**72.2**	7 Train	72.1	**72.8**	8 Truck	**35.1**	34.7
9 Boat	40.9	**50.5**	10 T.light	**40.8**	39.0	11 Fire H.	77.4	**82.7**	12 Stop sign	76.5	**81.4**
13 Park meter	**60.1**	56.6	14 Bench	35.7	**38.3**	15 Bird	64.5	**69.0**	16 Cat	77.4	**82.0**
17 Dog	65.0	**73.6**	18 Horse	70.6	**74.7**	19 Sheep	75.2	**76.8**	20 Cow	73.0	**78.7**
21 Elephant	79.7	**83.0**	22 Bear	83.6	**85.6**	23 Zebra	75.9	**76.2**	24 Giraffe	72.6	**75.6**
25 Backpack	18.5	**29.8**	26 Umbrella	**60.0**	58.5	27 Handbag	21.2	**32.8**	28 Tie	17.8	**18.6**
29 Suitcase	42.7	**54.2**	30 Frisbee	69.6	**75.9**	31 Skis	31.3	**38.7**	32 Snowboard	37.4	**46.3**
33 Sports ball	41.3	**48.4**	34 Kite	**51.3**	42.4	35 B. bat	31.0	**35.1**	36 B. glove	48.3	**50.4**
37 Skateboard	**42.4**	**42.4**	38 Surfboard	64.7	**68.8**	39 T.racket	58.4	**65.3**	40 Bottle	28.4	**32.6**
41 W. glass	34.0	**37.6**	42 Cup	49.8	**56.6**	43 Fork	19.7	**22.7**	44 Knife	34.8	**45.1**
45 Spoon	14.2	**16.8**	46 Bowl	31.1	**31.9**	47 Banana	41.6	**45.5**	48 Apple	**39.4**	34.8
49 Sandwich	50.5	**52.8**	50 Orange	47.5	**49.5**	51 Broccoli	33.1	**36.8**	52 Carrot	23.1	**27.3**
53 Hot dog	61.7	**67.5**	54 Pizza	84.7	**87.5**	55 Donut	67.6	**70.8**	56 Cake	44.1	**51.9**
57 Chair	7.2	**14.2**	58 Couch	34.5	**37.0**	59 P. plant	**11.9**	8.1	60 Bed	53.7	**56.8**
61 D.table	27.5	**47.7**	62 Toilet	57.2	**63.6**	63 TV	44.5	**52.2**	64 Laptop	50.9	**55.4**
65 Mouse	**50.1**	41.9	66 Remote	45.4	**58.8**	67 Keyboard	21.0	**33.3**	68 Cellphone	56.2	**65.8**
69 Microwave	**34.9**	33.4	70 Oven	23.6	**29.7**	71 Toaster	**42.4**	41.8	72 Sink	30.8	**33.0**
73 Fridge	23.9	**39.3**	74 Book	10.2	**21.5**	75 Clock	59.6	**64.6**	76 Vase	41.7	**46.1**
77 Scissors	43.4	**43.8**	78 Teddy	60.6	**64.2**	79 Hairdrier	47.9	**49.4**	80 Toothbrush	**34.3**	33.3

5.4 Discussion

Comparison with State-of-the-Arts We discuss the mIoU of our IPRNet and the latest and best performing HSNet [25]. By comparing the results of Pascal-5^i and COCO-20^i described in Table 1 and 2, we can see that the performance of the COCO-20^i condition is better than that of the Pascal-5^i condition, regardless of the backbone network and number of shots. We discuss that our mechanism of training to avoid similarity between prototypes is effective for this difficult problem; COCO-20^i, contains many more difficult objects to classify. Comparing the experimental results of 1-shot and 5-shot, the mIoU of 1-shot is higher regardless of the backbone network and dataset. This is because, as mentioned in the introduction, when the number of shots is smaller, the feature space covered by the support data is sparser and the classification performance was worse for classes near the classification boundary. Therefore, we can conclude that the proposed method is more effective.

Performance of similar classes that are difficult to classify For IPRNet and the baseline without the IPRM and RCM, we compare and discuss IoU for each of the classes shown in Table 4. Two main cases of classes that are difficult to classify are considered because of the existence of similar classes.

The first is an object that is likely to be similar to many other classes with a wide range of variations, specifically morphological changes, pictorial changes through decoration, and combinations with complex backgrounds. Specifically, there are 1 person with IoU increased by 20.7% and 61 dining table with IoU increased by 20.2%. 1 Person has all types of shape changes due to a deformable body and pictorial changes due to decoration (Fig.4(a),(b)). 61 Dining table is a complex combination of objects, most of which are placed on top of each other; therefore, the pixel boundaries of the objects are always shared with various occlusions (Fig.4(c),(d)).

Second, the class is an object with a simple shape and few features. Specifically, 73 fridge IoU increased by 15.4%, 66 remote by 13.4%, 67 keyboard by 12.3%, 27 handbag by 11.6%, 29 suitcase by 11.5%, 25 backpack by 11.3%, 74 book by 11.3%, and 44 knife by 10.3%. For example, 73 fridge is a symmetrical rectangular object with a few prominent patterns or protrusions on its surface. Hence, it was difficult to classify them based on similar rectangles and backgrounds (Fig.4(e)). 66 Remote is also a small rectangular body (Fig.4(f)).

For all other classes with an improved IoU, there are many classes where the IoU has increased because the RCM effect is considered to have improved the separation performance from the background. However, compared to the classes where these performances have increased by over 10%, the degree of conformity between the two cases are considered to be low. Specifically, 45 spoon and 43 fork are considered not to have improved IoU by as much as 44 knife because the shape of the tip is more non-graphical and distinctive compared to 44 knife.

Performance of characteristic objects Further, we consider the classes with the lowered IoU as proper nouns that have a complex shape or structure unique to that object that is unparalleled. Specifically, there are 80 toothbrush, 69 microwave, 26 umbrella, 10 traffic light, 4 motorcycle, 13 park meter, 59 potted plant, 48 apple, 65 mouse and 34 kite. For example, 34 kite is a discriminative proper nouns with no other similar concepts and limited use, so it is not often combined with several backgrounds (Fig.4(g)). 9 Potted plants had complex shapes that could not be represented by rectangles or spheres (Fig.4(h)). For these objects, we consider that how to extract the unique features of the object to be more important than acquiring the differences from other classes, and the training to reduce the similarity of prototypes between different classes, which is the aim of the IPRM, does not work effectively.

Qualitative Evaluation Fig.5 shows the difference between the baseline and our IPRNet using t-SNE [24] for the prototype of the target class extracted from the query image. At the baseline, prototypes between different classes are adjacent or overlapping, and there is no clear classification. However, in IPRNet, the distance between each prototype is increased, and it can be observed that the prototypes are more separated.

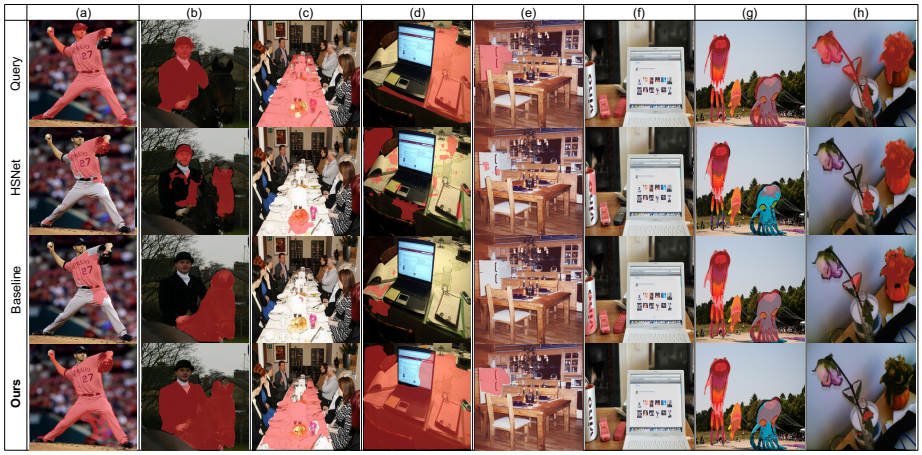

Fig. 4. Examples of the recognition result using our proposed method, baseline and HSNet [25]. The red shade is the area of the target class. It is the ground truth in the case of query, the recognition result in the cases of the baseline, HSNet [25] and Ours.

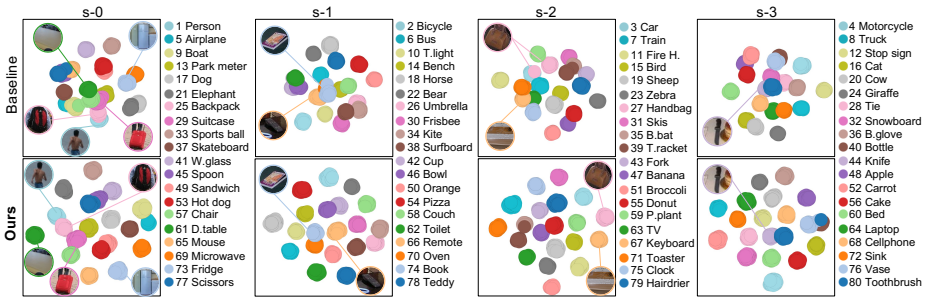

Fig. 5. Visualization results by t-SNE [24] of prototype obtained from query image of the baseline and IPRNet. They experiment with ResNet-50 5shot setting on COCO-20^i.

6 Conclusion

This study proposes a novel IPRNet that introduces a mechanism to improve separation performance by reducing the similarity between different classes. Experiments show that mIoU is improved over the existing few-shot segmentation methods [21] [34] [42] [25]. In addition, through an ablation study, we verified the separation performance between similar classes that are difficult to classify.

Acknowledgments We would like to thank Mr.Katsushi Yamashita, Director, from R&D Promotion Office, SoftBank Corp. and all of R&D Promotion Office members. Mr.Yamashita's witty and helpful support helped us, especially. We would like to thank Mr.Shogo Hamano from Department of Mechano-Informatics, the University of Tokyo for his informative discussion.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI) pp. 2481–2495 (2017)
2. Boudiaf, B., Kervadec, H., Masud, Z.I., Piantanida, P., Ayed, I.B., Dolz, J.: Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: CVPR. pp. 13979–13988 (2021)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI) **40**(4), 834–848 (2018)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Stefan Roth, B.S.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
8. Dong, N., Xing, E.P.: Few-shot semantic segmentation with prototype learning. In: The British Machine Vision Conference (BMVC) (2017)
9. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**, 98–136 (2014)
10. Fan, Z., Yu, J.G., Liang, Z., Ou, J., Gao, C., Xia, G.S., Li, Y.: Fgn: Fully guided network for few-shot instance segmentation. In: CVPR. pp. 9172–9181 (2020)
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning(ICML). vol. 70, pp. 1126–1135 (2017)
12. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: ECCV. pp. 297–312 (2014)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Hendryx, S.M., Leach, A.B., Hein, P.D., Morrison, C.T.: Meta-learning initializations for image segmentation. arXiv preprint arXiv:1912.06290 (2020)
15. LeCun, Y., Bernhard.E.Boser, Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation **1**(4), 541–551 (1989)
16. Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., Kim, J.: Adaptive prototype learning and allocation for few-shot segmentation. In: CVPR. pp. 8334–8343 (2021)
17. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
18. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
19. Liu, B., Ding, Y., Jiao, J., Ji, X., Ye, Q.: Anti-aliasing semantic reconstruction for few-shot semantic segmentation. In: CVPR. pp. 9747–9756 (2021)

20. Liu, W., Zhang, C., Lin, G., Liu, F.: Crnet: Cross-reference networks for few-shot segmentation. In: CVPR. pp. 4165–4173 (2020)
21. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: ECCV. pp. 142–158 (2020)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. p. 3431–3440 (2015)
23. Lu, Z., He, S., Zhu, X., Zhang, L., Song, Y.Z., Xiang, T.: Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In: ICCV. pp. 8741–8750 (2021)
24. Maaten, L.V.D., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
25. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: ICCV. pp. 6941–6952 (2021)
26. Nguyen, K.D.M., Todorovic, S.: Feature weighting and boosting for few-shot segmentation. In: ICCV. pp. 622–631 (2019)
27. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)
28. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: International Conference on Learning Representations (ICLR) (2017)
29. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. In: The British Machine Vision Conference (BMVC) (2017)
30. Siam, M., Oreshkin, B.N., Jagersand, M.: Amp: Adaptive masked proxies for few-shot segmentation. In: ICCV. pp. 5249–5258 (2019)
31. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Neural Information Processing Systems(NeurIPS). vol. 30 (2017)
32. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR. pp. 1199–1208 (2018)
33. Tian, P., Wu, Z., Qi, L., Wang, L., Shi, Y., Gao, Y.: Differentiable meta-learning model for few-shot semantic segmentation. The AAAI Conference on Artificial Intelligence **34**(07), 12087–12094 (2020)
34. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jiaa, J.: Prior guided feature enrichment network for few-shot segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI) (2020)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Neural Information Processing Systems(NeurIPS). vol. 30 (2017)
36. Vinyals, O., Blundell, C., Lillicrap, T., koray kavukcuoglu, Wierstra, D.: Matching networks for one shot learning. In: Neural Information Processing Systems(NeurIPS). vol. 29 (2016)
37. Wang, H., Zhang, X., Hu, Y., Yang, Y., Cao, X., Zhen, X.: Few-shot semantic segmentation with democratic attention networks. In: ECCV. pp. 730–746 (2020)
38. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: ICCV. pp. 9197–9206 (2019)
39. Wang, Y., Hebert, M.: Learning to learn: Model regression networks for easy small sample learning. In: ECCV. pp. 616 – 634 (2016)
40. Xie, G.S., Liu, J., Xiong, H., Shao, L.: Scale-aware graph neural network for few-shot semantic segmentation. In: CVPR. pp. 5475–5484 (2021)
41. Yang, B., Liu, C., Li, B., Jiao, J., Ye, Q.: Prototype mixture models for few-shot semantic segmentation. In: ECCV. pp. 763–778 (2020)

42. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: Mining latent classes for few-shot segmentation. In: ICCV. pp. 8721–8730 (2021)
43. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (ICLR) (2016)
44. Zhang, B., Xiao, J., Qin, T.: Self-guided and cross-guided learning for few-shot segmentation. In: CVPR. pp. 8312–8321 (2021)
45. Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q., Yao, R.: Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In: ICCV. pp. 9587–9595 (2019)
46. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: CVPR. pp. 5217–5226 (2019)
47. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: ECCV. pp. 418–434 (2018)
48. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 6230–6239 (2017)
49. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021)
50. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)