https://doi.org/10.1007/s43684-023-00058-2

Autonomous Intelligent Systems

ORIGINAL ARTICLE **Open Access**

Distilling base-and-meta network with contrastive learning for few-shot semantic segmentation

Xinyue Chen[1†], Yueyi Wang[1†], Yingyue Xu[1] and Miaojing Shi[2*]

Abstract

Current studies in few-shot semantic segmentation mostly utilize meta-learning frameworks to obtain models that can be generalized to new categories. However, these models trained on base classes with sufficient annotated samples are biased towards these base classes, which results in semantic confusion and ambiguity between base classes and new classes. A strategy is to use an additional base learner to recognize the objects of base classes and then refine the prediction results output by the meta learner. In this way, the interaction between these two learners and the way of combining results from the two learners are important. This paper proposes a new model, namely Distilling Base and Meta (DBAM) network by using self-attention mechanism and contrastive learning to enhance the few-shot segmentation performance. First, the self-attention-based ensemble module (SEM) is proposed to produce a more accurate adjustment factor for improving the fusion of two predictions of the two learners. Second, the prototype feature optimization module (PFOM) is proposed to provide an interaction between the two learners, which enhances the ability to distinguish the base classes from the target class by introducing contrastive learning loss. Extensive experiments have demonstrated that our method improves on the PASCAL-5^i under 1-shot and 5-shot settings, respectively.

Keywords: Semantic segmentation, Few-shot learning, Meta learning, Contrastive learning, Self-attention

1 Introduction

Semantic segmentation is to perform per-pixel classification on an image, which partitions the image into sections according to categories. Thanks to the availability of large amounts of data, deep neural networks have been developed rapidly. Various computer vision tasks based on deep neural networks have made great progress. However, collecting and labeling large amounts of data are time-consuming and laborious. Moreover, the neural network trained by the fully supervised learning relying on a large number of data cannot be extended to new classes. To deal with the above problems, weakly supervised learning [1], few-shot learning [2], and zero-shot learning [3] have emerged. Few-shot learning is proposed to learn the deep learning model on seen categories that can be generalized to unseen classes with a few annotated samples.

The research topic of this paper is few-shot semantic segmentation (FSS), where the model utilizes only a small number of annotated samples to segment the objects of new categories from the image. Most existing FSS methods [2, 4–10] achieve the generalization using the meta-learning framework which enables a model to transfer previous knowledge to unseen categories. During the meta-training phase, the model is trained with a batch of training episodes sampled from the set of base classes. Then, the knowledge learned from base categories is used to segment objects of novel categories during the meta-testing. Most previous approaches have attempted to segment new

*Correspondence: mshi@tongji.edu.cn
[2]College of Electronic and Information Engineering, Tongji University, Shanghai, 201804, China
Full list of author information is available at the end of the article †Equal contributors

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

classes by using parameters trained on base classes without fine-tuning, such as PFENet [4], PANet [5], CANet [6], ASGNet [7], and so on. However, the parameter-sharing mechanism in the meta-learning framework inevitably results in the network being biased toward base classes, which leads to semantic confusion and ambiguity between the new classes and the base classes. For instance, the objects of base classes would be segmented in meta-testing when the new classes have similar semantic concepts to base classes. Recently, BAM [11] has introduced an extra branch (base learner) to the meta-learning framework (meta learner) to identify the regions of base classes. Then, the coarse outputs from the two learners are fused into more precise predictions through an ensemble module. In this ensemble module, an adjustment factor has been introduced to estimate the scene differences between the input image pairs. This approach can distinguish base classes from novel classes, providing a new perspective for future work in the FSS field, but there is still potential room for improvement. Firstly, the base learner and the meta learner are independent of each other without any interactions, so the knowledge of the base learner is unable to affect the meta-training. Secondly, the ensemble module uses features extracted from the low-level block of the backbone network to calculate the adjustment factor without attention, which hinders the key features from contributing to the adjustment factor.

Based on BAM [11], the model proposed in this paper introduces contrastive learning to connect the two learners. Also, our model utilizes the self-attention module to weigh the low-level features. To sum up, the primary contributions can be summarised as follows:

- We propose to use contrastive learning loss to enable the base learner and the meta learner to interact with each other, supervising meta learner to learn better image representations to distinguish novel classes from base classes.
- We apply the self-attention module to low-level features extracted from the backbone network to obtain a more accurate adjustment factor in the ensemble module.
- Sufficient experiments on PASCAL-5^i verify the effectiveness of our methods, and our performance exceeds the original BAM [11] model.

2 Related works

2.1 Semantic segmentation

Semantic segmentation is an essential computer vision task that aims to classify each pixel in the given image according to predefined categories. Benefiting from fully convolutional networks (FCNs) [12], there is great progress in the field of semantic segmentation. Recently, numerous FCN-based models have been designed to accomplish semantic segmentation. For instance, [13] proposed a symmetrical encoder-decoder structure based on FCN, termed U-net, to reconstruct segmentation step by step. Yu and Koltun [14] proposed the dilated convolution to enlarge the receptive field without resolution loss, thereby improving segmentation by using contextual information. The pyramid pooling module (PPM) proposed in PSPNet [15] aggregates multi-scale information by pooling in different sizes. DeepLab V2 [16] developed atrous spatial pyramid pooling (ASPP) to obtain and fuse multi-scale information by using filters with different expansion rates. However, these approaches rely on large-scale annotated samples and cannot work well on novel classes, thereby hampering segmentation in real-world applications.

2.2 Few-shot learning

Many tasks are researching how models trained on base classes can be devised to recognize new classes, such as works in image classification [17–19], object detection [20, 21], and semantic segmentation [2, 5, 6, 10]. These works are classified into the field of few-shot learning or zero-shot learning. Few-shot learning (FSL) aims to recognize objects of novel classes given a small number of annotated samples. Most recent works in few-shot learning employ the meta-learning framework proposed by Vinyals et al. [22]. In meta-learning approaches, there is a batch of learning episodes during the training stage. Each learning episode consists of several images sampled from the dataset of a base class, which simulates the few-shot scenarios of novel classes. The FSL approaches under the meta-learning framework can be subdivided into three categories: (a) model-based, (b) metric-based, and (c) optimization-based. Santoro et al. [23] proposed a model-based approach to access cross-task knowledge using an external memory network. [22, 24, 25] exploited the metric-based idea of transforming data into embedding vectors in high-dimensional space, thus converting the classification problem into the nearest neighbour problem in embedding space. [26–29] explored the optimization-based idea in order to design an update strategy of model parameters that can converge with a few samples, thus enabling the model to generalize quickly to unseen categories. Our work introduces metric-based FSL to address the few-shot semantic segmentation.

2.3 Few-shot semantic segmentation

Few-shot segmentation aims to make dense pixel-level predictions for novel classes given only a few annotated samples. Since the OSLSM for FSS was proposed by Shaban et al. [30], many excellent models have emerged. Most approaches to solving FSS use the metric-based meta-learning framework. Specifically, this kind of framework usually employs two branches to generate a foreground prototype of a support image from the support branch first, and then obtain predicted segmentation of the query

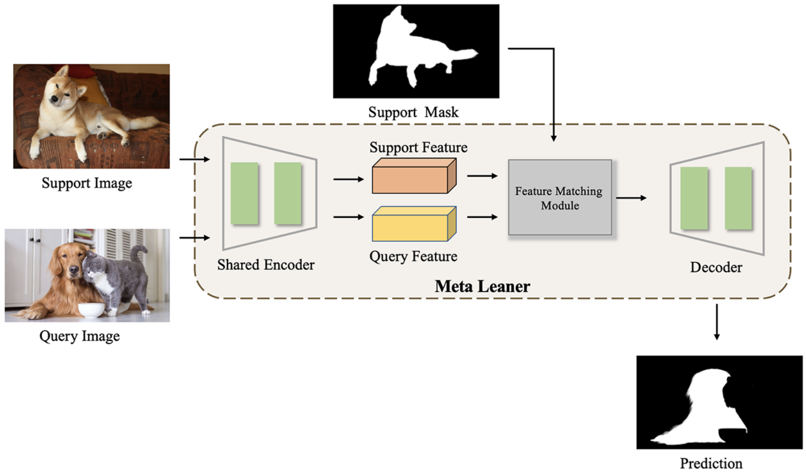

Figure 1 Summary of recent parametric-based FSS models under the meta-learning framework. The grey box indicates the Meta learner section

image by pixel-level matching between the support prototype and the query feature. Depending on the metric tools, metric-based meta-learning methods are divided into parameter-based and prototype-based models. Parameter-based framework shown in Fig. 1 usually uses convolution to build a feature matching block for exploring the relation between support features and query features. Following this framework, CANet [6] firstly utilized convolution to refine the segmentation result. Inspired by CANet [6], PFENet [4] designed the feature enrichment module using convolutions instead of cosine similarity to fuse support and query features.

The prototype-based framework shown in Fig. 2 uses a non-parameter metric way such as cosine similarity to measure the similarity between the extracted support prototypes and the query features. PANet [5] firstly utilized the pseudo-label of the query image represented by the distance between the prototypes and the query image to segment the support images. ASGNet [7] employed a superpixel-guided clustering strategy to produce some part-aware prototypes for support images and then allocate these prototypes to each pixel according to the similarity between each prototype and the query features. NTRENet [8] proposed background and distracting object prototypes to explicitly mine and eliminate the background and distracting regions in the query image.

These approaches mentioned above are based on the meta-learning framework, but their trained models are usually biased towards seen categories, which leads to semantic confusion between seen classes and similar unseen classes. Therefore, the preference of these models results in a generalization problem hindering the recognition of new categories. To address this problem, Lang et al. [11] proposed BAM which introduces an extra semantic segmentation model (base learner) trained on the base dataset to segment objects of base classes in the image. In addition, BAM [11] designed an ensemble module to obtain the final prediction by integrating the coarse results from the base learner and the meta learner. Although BAM [11] has achieved the state-of-the-art performance, we observe the two learners are independent without interaction. This paper focuses on generating the interaction between these two learners for achieving more accurate segmentation results.

2.4 Contrastive learning

Contrastive learning aims to learn better feature representations by automatically constructing positive and negative samples, by which positive pairs are made closer together in the projection space. In contrast, negative pairs are forced away from each other. SimCLR [31] used various data argumentation methods to construct positive and negative sample pairs of each image for learning a robust image representation space. Wang et al. [32] considered the global semantic similarity of all pixels in the whole training set, reducing the distances between positive pairs and enlarging the distances between negative pairs. Liu et al. [8] first introduced contrastive learning to FSS, so as to learn more precise prototypes that help the model distinguish target objects from distracting objects. Inspired by this work, we introduce contrastive learning to generate the interaction between the two learners of BAM [11].

3 Method

In this section, we first give the definition of FSS in Subsect. 3.1. Then, we describe the details of our proposed model in Subsects. 3.2, 3.3, 3.4, and 3.5. Figure 3 gives an overview of our Distilling Base and Meta (DBAM) model which consists of the base learner, the meta learner, the self-attention-based ensemble module (SEM), and the prototype feature optimization module (PFOM).

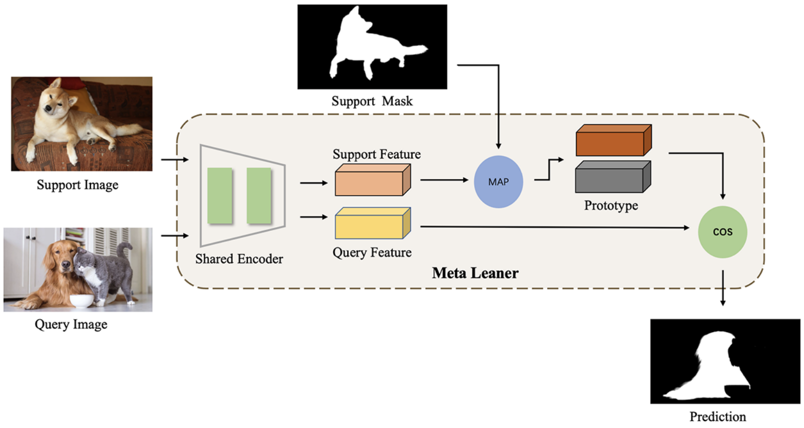

Figure 2 Summary of recent prototype-based FSS models under the meta-learning framework. The grey box indicates the Meta learner section

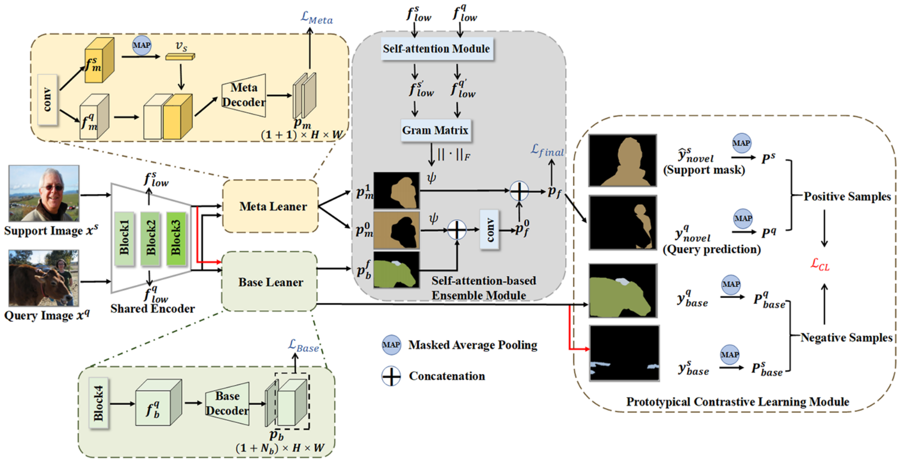

Figure 3 Overall framework of our DBAM model. (x^s, x^q) are input image pairs to the shared encoder. The shared encoder extracts different levels of feature maps. The two learners use the extracted features to generate the coarse predictions. The coarse predictions from the two learners are then fused by the adjustment factor in the self-attention-based ensemble module to get the final prediction p_f which is utilized for generating the prototype of target P. In addition, the prediction from the base learner is used to generate the base prototype P_{base}. Finally, base prototypes and target prototypes are regarded as negative and positive samples, optimized by the prototypical contrastive learning module. The function of "red lines" is to provide base prototypes P_{base}^s from support images

3.1 Problem definition

For the FSL task, the whole dataset is divided into a base dataset D_{base} and a novel dataset D_{novel} by categories, where the D_{base} with base classes C_{base} contains sufficient annotated images for the meta training phase only, and the D_{novel} with novel classes C_{novel} has scarce annotated samples for meta testing phase only. These two sets are disjoint ($C_{\text{base}} \cap C_{\text{novel}} = \emptyset$). Current methods use the episodic paradigm [30] during the meta training and testing. For 1-way K-shot segmentation, in each episode, $K+1$ image-mask pairs randomly sampled from the D_{novel} are divided into the support set S and the query set Q. After completing the episodic training, we evaluate our model on all test episodes sampled from the D_{novel}.

3.2 Base learner

To address the problem of FSS models being biased towards the seen classes, the BAM [11] model introduces a base learner using PSPNet [15] to predict the regions of base classes in the query images. It initially uses the encoder network $\mathcal{E}$ and a following convolutional block $\mathcal{F}_{\text{conv}}$ to extract feature maps f_b^q of the query image x^q, which the following equation can conclude:

$$f_b^q = \mathcal{F}_{\text{conv}}\left(\mathcal{E}\left(x^q\right)\right) \in \mathbb{R}^{c \times H \times W}, \tag{1}$$

where $\mathcal{F}_{\text{conv}}$ denotes Block4 shown in Fig. 3. Then, the decoder network $\mathcal{D}_b$ are applied to get the prediction result p_b, which can be formulated as:

$$p_b = \text{softmax}\left(\mathcal{D}_b\left(f_b^q\right)\right) \in \mathbb{R}^{(1+N_b)\times H\times W}, \tag{2}$$

where softmax(•) represents the operation of generating probability maps p_b along the channel dimension. N_b denotes the number of base classes, and $N_b + 1$ represents the number of base classes and a background class. Base learner measures the difference between the prediction p_b and the ground-truth m_b^q using standard cross-entropy (CE) loss.

$$\mathcal{L}_{\text{base}} = \frac{1}{n_{bs}}\sum_{i=1}^{n_{bs}} CE\left(p_{bi}, m_{bi}^q\right). \tag{3}$$

3.3 Meta learner

Given a set of support images and the corresponding masks $\mathcal{S} = (x^s, m^s)$ and a query image x^q, the meta learner aims to segment objects in the query image that belong to the same class as provided support mask. Following BAM [11], we use the meta learner to produce the class-related prototype v_s, which can be formulated as:

$$f_m^s = \mathcal{F}_{1\times 1}\left(\mathcal{E}\left(x^s\right)\right), \tag{4}$$

$$f_m^q = \mathcal{F}_{1\times 1}\left(\mathcal{E}\left(x^q\right)\right), \tag{5}$$

$$v_s = \mathcal{F}_{\text{pool}}\left(f_m^s \odot \mathcal{R}\left(m^s\right)\right), \tag{6}$$

where $\mathcal{F}_{1\times 1}$ is a 1×1 convolution operation for reducing the dimensionality. f_m^s, f_m^q denote the intermediate feature maps for support images and the query image. $\mathcal{R}$ is the operation for reshaping support mask. $\odot$ represents Hadamard product. $\mathcal{F}_{\text{pool}}$ is the average-pooling for generating v_s of $c \times 1 \times 1$. Afterwards, we feed the combination of expanded v_s and f_m^q into the decoder network $\mathcal{D}_m$ to get the final prediction result p_m, which can be formulated as:

$$p_m = \text{softmax}\left(\mathcal{D}_m\left(\mathcal{C}\left(\mathcal{P}(v_s), f_m^q\right)\right)\right) \in \mathbb{R}^{2\times H\times W}, \tag{7}$$

where $\mathcal{P}$ and $\mathcal{C}$ represent the dimensional expansion and the concatenating operation, respectively. Meta leaner adopts binary cross-entropy loss (BCE) to evaluate the difference between the prediction p_m and ground truth m^q:

$$\mathcal{L}_{\text{meta}} = \frac{1}{n_e}\sum_{i=1}^{n_e} BCE\left(p_{mi}, m_i^q\right). \tag{8}$$

3.4 Self-attention-based ensemble module

Since the meta learner receives features from both support and query images, the meta learner is susceptible to the large difference between the input image pairs, causing some regions in the query image to be incorrectly activated [33]. The ensemble module proposed in BAM [11] leverages the adjustment factor to suppress the incorrectly activated region of the meta learner output and then fuse the output of the two learners. Firstly, the adjustment factor is obtained from the difference in scenes between the query and supports. Specifically, the Gram matrices G^s and G^q of the low-level features f_{low}^s and f_{low}^q of the query and support images are computed respectively. Then, the adjustment factor ψ can be obtained by Frobenius norm $\mathcal{F}$ of the difference between two Gram matrices G^s and G^q.

$$\psi = \left\| G^s - G^q \right\|_{\mathcal{F}}. \tag{9}$$

In the next step, the adjustment factor is used to refine the predictions of the meta learner. Specifically, ψ is expanded to the same dimension as meta output to obtain an adjustment map $\mathcal{M}_\psi$, and then the foreground p_m^1 and background p_m^0 obtained from the meta learner are concatenated with adjustment map $\mathcal{M}_\psi$ respectively. Refined results $p_m^{1'}$ and $p_m^{0'}$ are obtained after 1×1 convolution operation respectively.

$$p_m^{0'} = \mathcal{F}_{\mathcal{M}_\psi}\left(p_m^0\right), \tag{10}$$

$$p_m^{1'} = \mathcal{F}_{\mathcal{M}_\psi}\left(p_m^1\right), \tag{11}$$

where "0" and "1" denote the background and foreground respectively. Finally, the fine-grained results of the meta learner are fused with the predictions of the base learner. The foreground in the base learner prediction result is objects of base classes in the query image p_b^f, which also belongs to the background in the meta learner prediction result. Therefore, p_b^f and $p_m^{0'}$ are concatenated and fused by a 1×1 convolution operation $\mathcal{F}_{\text{ensemble}}$ to obtain the background of the final prediction result p_f^0.

$$p_f^0 = \mathcal{F}_{\text{ensemble}}\left(p_b^f, p_m^{0'}\right). \tag{12}$$

Then, the final prediction result p_f is generated by concatenating the background p_f^0 and the foreground $p_m^{1'}$ as follow:

$$p_f = p_f^0 \bigoplus p_m^{1'}. \tag{13}$$

We propose the improved ensemble module, namely the self-attention-based ensemble module by applying the self-attention mechanism to low-level features to obtain a more semantically explicit re-weighted feature map. Because self-attention can capture the semantic relationship between any two positions in the feature map, so the obtained re-weighted feature's semantic information is clearer. Therefore, the adjustment factor (see Fig. 4) produced by our self-attention based ensemble module guides

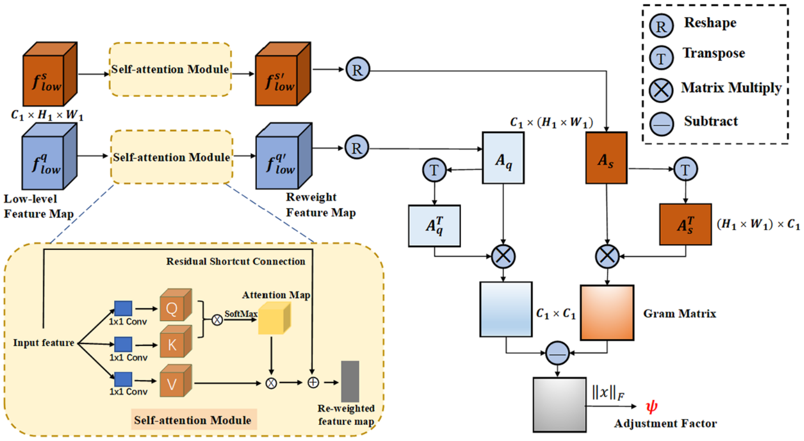

Figure 4 The workflow of calculating the adjustment factor ψ in the self-attention-based ensemble module

the fusion of two images more exactly than before. The low-level features $f^{q}_{\text{low}}, f^{s}_{\text{low}} \in \mathbb{R}^{C1\times H1\times W1}$ extracted from the shared encoder are fed into the self-attention module to obtain re-weighted features $f^{q'}_{\text{low}}, f^{s'}_{\text{low}} \in \mathbb{R}^{C1\times H1\times W1}$ with more clearer and explicit semantic information. The self-attention operations of the two input features are the same, and that of the query one can be expressed as:

$$Q_q = \mathcal{F}_{1\times 1}\left(f^{q}_{\text{low}}\right) \in \mathbb{R}^{\frac{C}{8}\times W_1\times H_1}, \tag{14}$$

$$K_q = \mathcal{F}_{1\times 1'}\left(f^{q}_{\text{low}}\right) \in \mathbb{R}^{\frac{C}{8}\times W_1\times H_1}, \tag{15}$$

$$V_q = \mathcal{F}_{1\times 1''}\left(f^{q}_{\text{low}}\right) \in \mathbb{R}^{\frac{C}{8}\times W_1\times H_1}, \tag{16}$$

$$f^{q'}_{\text{low}} = \text{softmax}\left(\frac{Q_q K_q^T}{\sqrt{d_k}}\right) V_q \oplus f^{q}_{\text{low}} \in \mathbb{R}^{C_1\times W_1\times H_1}, \tag{17}$$

where $\mathcal{F}_{1\times 1}$, $\mathcal{F}_{1\times 1'}$ and $\mathcal{F}_{1\times 1''}$ denote three 1 × 1 convolution operations which can project input features into a high dimension space. $\oplus$ indicates the concatenation operation along the channel dimension. Re-weighted feature $f^{q'}_{\text{low}}$ can be obtained by residual concatenating original input f^{q}_{low} with attention map.

3.5 Prototype feature optimization module

The base learner and the meta learner are independent of each other in BAM [11]. In other words, the meta learner is still likely to confuse the features of base classes and the target class with the help of the base learner. We aim at enabling the meta learner to distinguish base classes from the target class by distilling the knowledge of the base learner into the meta learner, in order to obtain better segmentation results. A recent work of prototypical contrastive learning proposed by Liu et al. [8] regards the region which is complementary to the union of the background region and target region as the distracting object region. For the query prototype P^q, the corresponding support prototype P^s is the positive sample, while the distracting object prototypes in both query and support are negative samples. Inspired by this approach, we propose the prototype feature optimization module to make the prototype feature of the target object in the query image different from that of the objects of base classes and the prototype of the target object in the query image close to that of the support image. Different from [8], we propose to treat query prediction prototype and corresponding support prototype (P^q, P^s) as the positive pair, while the query prediction prototype and the prediction prototypes of base classes in support and query images, predicted by the base learner, (P^q, P^s_{base}) and (P^q, P^q_{base}) as negative pairs. We use the masked average pooling (MAP) to extract the prototypes for positive and negative pairs respectively, which is shown in Eq. (18) and Eq. (19).

$$\text{positive} = \begin{cases} P^s = \mathcal{F}_{MAP}(\hat{y}^s_{\text{novel}}), \\ P^q = \mathcal{F}_{MAP}(y^q_{\text{novel}}), \end{cases} \tag{18}$$

$$\text{negative} = \begin{cases} P^s_{\text{base}} = \mathcal{F}_{MAP}(y^s_{\text{base}}), \\ P^q_{\text{base}} = \mathcal{F}_{MAP}(y^q_{\text{base}}), \end{cases} \tag{19}$$

where $\hat{y}^s_{\text{novel}}$ is the ground-truth mask of support image, y^q_{novel} is the predicted mask of query image. Both y^s_{base} and y^q_{base} are predicted masks output by the base learner. The generation process of these prototypes can be seen in Fig. 3. Afterwards, the contrastive learning loss $\mathcal{L}_{CL}$ shown in Eq. (20) is introduced to optimize the above prototypes:

$$\mathcal{L}_{CL} = -\log \frac{e^{\cos(P^q,P^s)}}{\sum_{\mathcal{B}}\left(e^{\cos(P^q,P^q_{\text{base}})} + e^{\cos(P^q,P^s_{\text{base}})}\right)}, \tag{20}$$

where cos denotes cosine similarity and e is the natural constant. Finally, we use the new total loss $\mathcal{L}_{\text{total}}$ to supervise the training of our model, which can be summarised as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{final}} + \lambda \mathcal{L}_{\text{meta}} + \beta \mathcal{L}_{CL}, \tag{21}$$

where $\mathcal{L}_{\text{final}}$ is the BCE loss between the final prediction p_f and ground truth m^q. λ, β are adjustable loss weights and are set to 1.0 and 0.01, respectively.

4 Experiments

4.1 Setup

4.1.1 Datasets

We evaluate our model on PASCAL-5^i dataset which is widely used in the field of FSS. The PASCAL-5^i dataset is proposed by Shaban et al. [30], created from the PASCAL VOC 2012 [34] dataset with the SBD dataset (Semantic Boundaries Dataset and Bench-mark) [35] as augmentation, which includes images of ordinary objects in daily life in a total of 20 categories. PASCAL-5^i has randomly divided 20 classes by category into 15 base classes and five novel classes and evenly split the dataset into four folds, each containing five categories.

4.1.2 Evaluation metric

Following [36], we use mean intersection-over-union (mIoU) to quantitatively measure the experimental results. We use C_{novel} as the number of classes during the testing stage, the mIoU calculates the average of IoUs over all testing classes, which can be expressed by the following formula:

$$mIoU = \frac{1}{C_{\text{novel}}} \sum_{i=1}^{C_{\text{novel}}} IoU_i. \tag{22}$$

For an individual class, the IoU metric is defined as:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{23}$$

where the TP, FP, FN are the number of true positives, false positives and false negatives of the predicted masks. A higher IoU indicates a more accurate segmentation result.

4.1.3 Implementation details

All experiments are based on the ResNet50 [37] backbone and PASCAL-5^i. We adopt a two-stage training strategy similar to BAM [11], where the base learner is trained using fully-supervised learning protocol on each fold of the PASCAL-5^i dataset, and the meta learner is trained jointly with the ensemble module using a meta-training paradigm. In the first stage, we train PSPNet [15] as a base learner on 15 base classes and the background class for PASCAL-5^i. In each fold, we obtain a separate PSPNet after training for 100 epochs. The stochastic gradient descent (SGD) optimizer with a learning rate of 2.5e-3 is utilized to update the network parameters during training. In the meta-training phase, the two learners share the backbone of the base learner who trained in the first stage. We freeze the parameters of PSPNet during the meta learner training phase. To train the meta learner, we use the SGD optimizer on PASCAL-5^i for 200 epoches at a learning rate of 5e-2, and the training batch set is set to 8. We compute the average outcomes of 5 runs with various random seeds to reduce the performance effect of chosen support-query image pairings. To facilitate a comparison of results with BAM [11], we use the same data argumentation strategies as BAM. All experiments are implemented in the PyTorch 1.7.0 environment and conducted on the NVIDIA GeForce RTX 3090 GPUs. For a better comparison of performance, we output the results without performing any post-processing and fine-tuning.

4.2 Comparison with state-of-the-art methods

We compare the performance of our method with several state-of-the-art (SOTA) FSS methods [4, 6–11] using the PASCAL-5^i dataset. The experiments are conducted with the ResNet50 backbone under 1-shot and 5-shot settings. The performance of our method is illustrated in both quantitative and qualitative forms.

4.2.1 Quantitative comparison

Table 1 illustrates the performance comparison of our DBAM method equipped with the ResNet50 backbone with other FSS methods. Our method achieves the best performance. Specifically, under the 1-shot setting, the averaged mIoU of our method outperforms that of BAM* (our implementation) by 0.53. In the 5-shot setting, our method shows improvements of 0.87 over the reproduced BAM*.

4.2.2 Qualitative comparison

We visualize the segmentation results of some episodes in the meta-testing phase to better illustrate and understand the effect of our approach. In Fig. 5, the first and second columns are the support and query images with corresponding masks. The third and fourth columns show the segmentation result of the original BAM model and our method. In Fig. 5, our method reduces the activated base class regions better than the original BAM model. For example, the box behind the wine glass in the last one of the second row and the hand next to the pigeon in the last place of the third row are both well suppressed. This shows that the self-attention module can obtain more accurate adjustment factors and thus produce better segmentation results. Contrastive learning allows the two learners to interact with each other so that the meta learner can learn better image representations allowing the novel and base class samples to be further apart. However, the contribution of each module to the performance improvement cannot be seen from the figure. Hence we conduct extensive ablation experiments to observe the specific contribution of each module.

Table 1 Performance comparison of BAM with contrastive learning loss and self-attention on PASCAL-5^i. Results in bold indicate the top performance. * denotes the results that were obtained by our own implementation

Backbone	Method	1-shot					5-shot				
		Fold0	Fold1	Fold2	Fold3	Mean	Fold0	Fold1	Fold2	Fold3	Mean
ResNet50	CANet [6]	52.50	65.90	51.30	51.90	55.40	55.50	67.80	51.90	53.20	57.10
	PGNet [9]	56.00	66.90	50.60	50.40	56.00	57.70	68.70	52.90	54.60	58.50
	PPNet [10]	48.58	60.58	55.71	46.47	52.84	58.85	68.28	66.77	57.98	62.97
	PFENet [4]	61.70	69.50	55.40	56.30	60.80	63.10	70.70	55.80	57.90	61.90
	ASGNet [7]	58.84	67.86	56.79	53.66	59.29	63.66	70.55	64.17	57.38	63.94
	NTRENet [8]	65.40	72.30	59.40	59.80	64.20	66.20	72.80	61.70	62.20	65.70
	BAM [11]	68.97	73.59	67.55	61.13	67.81	70.59	75.05	70.79	67.20	70.91
	BAM* [11]	68.39	73.47	67.46	60.58	67.48	70.09	74.93	70.69	66.65	70.59
	DBAM (ours)	**69.34**	**73.71**	**67.66**	**61.36**	**68.01**	**71.71**	**75.42**	**71.19**	**67.50**	**71.46**

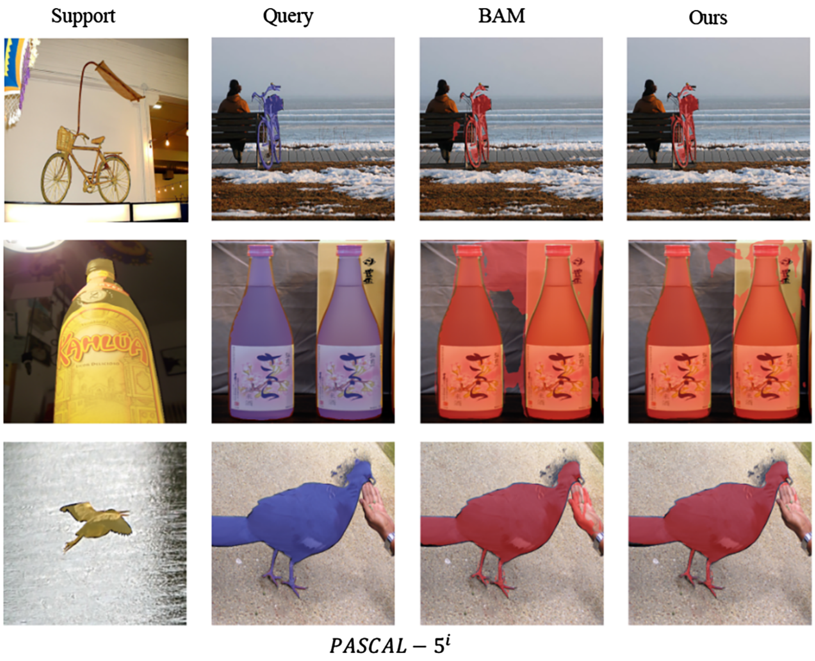

Figure 5 The examples of the segmentation results for the original BAM model and our method on PASCAL-5^i under the 1-shot FSS setting. The support images and corresponding masks are in yellow, the query images with ground-truth masks are in green, and the BAM and our method prediction results are in red

4.3 Ablation study

We conduct sufficient ablation studies using the ResNet50 backbone on PASCAL-5^i under a 5-shot setting to investigate the effect of each component on segmentation performance.

4.3.1 Ablation study on self-attention module

The adjustment factor ψ is an essential component of the ensemble module. It is derived from the scene differences between the feature maps of the support-query image pairs. Thus, selecting suitable feature maps from the backbone block is critical to the fusion results. We conduct extensive experiments on the impact of the feature maps extracted from each layer of the backbone network (i.e., ResNet50 [37]) on the segmentation performance. As can be seen from Fig. 6, the B2 feature map shows the optimal segmentation performance in B0-B3. We attribute this to the fact that B2 associates some low-level features, such as colour, texture, image style, etc., while B3 associates some more abstract high-level features that are not conducive to computing scene differences in query-supported image pairs. The Bl and B2 feature maps are insufficient to completely explore the features of the image pairs and perform relatively poorly.

Moreover, adding the self-attention module to the B2 feature map shows better segmentation results under the 1-shot and 5-shot settings. The quantitative result (see Ta-

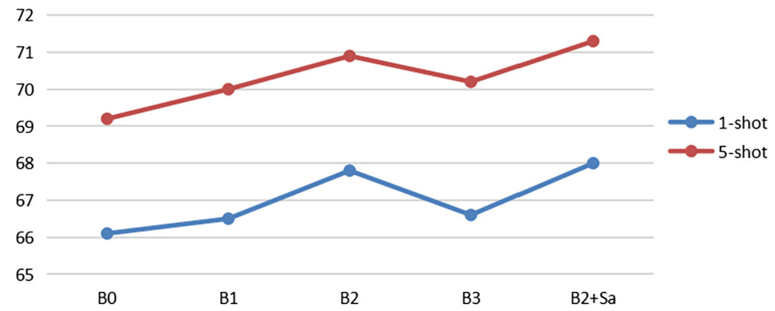

Figure 6 Ablation studies on the low-level features with ResNet50 backbone. "*Bi*" represents the feature maps that the backbone's i-th convolutional block produced, "B2+Sa" denotes the additional self-attention module after Block2

Table 2 Ablation study on self-attention module, 'S' denotes the self-attention module (Top performance in bold mIoU)

Backbone	Methods	1-shot					5-shot				
		Fold0	Fold1	Fold2	Fold3	Mean	Fold0	Fold1	Fold2	Fold3	Mean
ResNet50	DBAM(w/o S)	68.39	73.47	67.46	60.58	67.48	70.09	74.93	70.69	66.65	70.59
	DBAM(w/ S)	**69.10**	**73.71**	**67.66**	**61.10**	**67.89**	**71.48**	**75.62**	**71.07**	**66.89**	**71.27**

Table 3 Ablation study on contrastive learning loss, where 'CL' represents the contrastive learning loss (Top performance in bold mIoU)

Backbone	Methods	1-shot					5-shot				
		Fold0	Fold1	Fold2	Fold3	Mean	Fold0	Fold1	Fold2	Fold3	Mean
ResNet50	DBAM(w/o CL)	68.39	73.47	67.46	60.58	67.48	70.09	74.93	70.69	66.65	70.59
	DBAM(w/ CL)	**69.22**	**73.69**	**67.65**	**61.25**	**67.95**	**71.37**	**75.27**	**70.93**	**67.26**	**71.21**

ble 2) shows that our DBAM with the self-attention module improves by 0.41 and 0.68 under the 1-shot and 5-shot settings, respectively.

4.3.2 Ablation study on contrastive learning loss

The contrastive learning loss introduced by the PFOM helps the two branches to interact with each other and enables the meta learner to learn the knowledge of the base learner. Thus, to investigate the impact of contrastive learning loss on segmentation performance, we conduct experiments on it. As shown in Table 3, the mIoU of our DBAM with contrastive learning loss increases by 0.47 and 0.62 under the 1-shot and 5-shot settings, respectively.

In Table 4, we summarise the ablation experiments of the two modules. The table shows that the contribution of the self-attention module and contrastive learning to improving segmentation performance in these two settings is close. The most significant improvement is when the two modules are used together, which is about 0.53 and 0.87 in the 1-shot and 5-shot settings respectively.

5 Conclusion

In this project, we aims to address the potential problems of BAM [11]. We propose a new model based on the base-and-meta structure to more accurately exclude the distracting objects of base classes from the images. Particularly, the self-attention mechanism is introduced into the ensemble module for getting a more precise adjustment factor, so as to refine the coarse prediction from the meta learner. In addition, contrastive learning is leveraged to distinguish target objects from distracting objects of base classes by introducing base-learner knowledge into the meta learner. Extensive experiments and ablation studies validate the effectiveness of our method and demonstrate the superior performance of our method compared with other state-of-the-art approaches.

Appendix: Symbol list

D_{base}	base dataset	(A1)
D_{novel}	novel dataset	(A2)
C_{base}	base classes	(A3)
C_{novel}	novel classes	(A4)
$\mathcal{F}_{\text{conv}}$	convolutional block	(A5)
$\mathcal{F}$	Frobenius norm	(A6)
$\mathcal{F}_{1\times1}$	1×1convolutional block	(A7)

Table 4 Overall ablation study on two modules, where "Sa" denotes the self-attention module and "CL" represents contrastive learning loss (Top performance in bold mIoU)

Backbone	Methods		1-shot					5-shot				
	Sa	CL	Fold0	Fold1	Fold2	Fold3	Mean	Fold0	Fold1	Fold2	Fold3	Mean
ResNet50			68.39	73.47	67.46	60.58	67.48	70.09	74.93	70.69	66.65	70.59
	✓		69.10	73.71	**67.66**	61.10	67.89	71.48	**75.62**	71.07	66.89	71.27
		✓	69.22	73.69	67.65	61.25	67.95	71.37	75.27	70.93	67.26	71.21
	✓	✓	**69.34**	**73.71**	**67.66**	**61.36**	**68.01**	**71.71**	75.42	**71.19**	**67.50**	**71.46**

$\mathcal{R}$	reshape operation	(A8)
$\mathcal{F}_{\text{pool}}$	average pooling operation	(A9)
$\odot$	Hadamard product operation	(A10)
$\mathcal{P}$	dimensional expansion operation	(A11)
$\mathcal{C}$	concatenating operation	(A12)
f_b^q	feature map of query image	(A13)
$\mathcal{D}_b$	decoder network	(A14)
p_b	prediction of base learner	(A15)
N_b	number of base classes	(A16)
m_b^q	ground-truth of base learner	(A17)
v_s	class-related prototype from meta learner	(A18)
f_m^s	intermediate feature maps for support image	(A19)
f_m^q	intermediate feature maps for query image	(A20)
G_s	Gram metrics of the support image	(A21)
G_q	Gram metrics of the query image	(A22)
f_{low}^s	low-level feature from the support image	(A23)
$f_{\text{low}}^{s'}$	re-weighted low-level feature from the support image	(A24)
f_{low}^q	low-level feature from the query image	(A25)
$f_{\text{low}}^{q'}$	re-weighted low-level feature from the query image	(A26)
ψ	adjustment factor	(A27)
$\mathcal{M}_\psi$	adjustment map	(A28)
p_m^1	foreground obtain from the meta learner	(A29)
p_m^0	background obtain from the meta learner	(A30)
$\mathcal{L}_{\text{base}}$	loss from base learner	(A31)
$\mathcal{L}_{\text{meta}}$	loss from meta learner	(A32)
$\mathcal{L}_{CL}$	contrastive learning loss	(A33)
$\mathcal{L}_{\text{total}}$	total loss	(A34)
y_{novel}^q	the predicted mask of query image in novel classes	(A35)
$\hat{y}_{\text{novel}}^s$	the ground-truth mask of support image in novel classes	(A36)
y_{base}^q	the predicted mask of query image in base classes	(A37)
y_{base}^s	the predicted mask of support image in base classes	(A38)

Acknowledgements
We are very grateful to Dr. Shi for his guidance.

Funding
There are no sources of funding for the research.

Availability of data and materials
The data and material that support the findings of this study are available on request from the corresponding author.

Code availability
The code of this study is available on request from the corresponding author.

Declarations

Competing interests
Prof. Miaojing Shi is an editorial board member for Autonomous Intelligent Systems and was not involved in the editorial review, or the decision to publish, this article. All authors declare that there are no other competing interests.

Author contributions
XC: Conducting experiments, Writing and Revising the manuscript. YW: Methodology, Conducting experiments, Analysing the results. All authors read and approved the final manuscript.

Author details

[1]Informatics, King's College London, Strand, London, WC2R 2LS, United Kingdom. [2]College of Electronic and Information Engineering, Tongji University, Shanghai, 201804, China.

Received: 17 October 2022 Revised: 31 August 2023
Accepted: 12 November 2023 Published online: 27 November 2023

References

1. Z. Yang, M. Shi, C. Xu, V. Ferrari, Y. Avrithis, Training object detectors from few weakly-labeled and many unlabeled images. Pattern Recognit. **120**, 108164 (2021)
2. M. Zhang, M. Shi, L. Li, Mfnet: multiclass few-shot segmentation network with pixel-wise metric learning. IEEE Trans. Circuits Syst. Video Technol. **32**(12), 8586–8598 (2022)
3. Y. Du, M. Shi, F. Wei, G. Li, Boosting zero-shot learning via contrastive optimization of attribute representations (2022). arXiv preprint. arXiv:2207.03824
4. Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, J. Jia, Prior guided feature enrichment network for few-shot segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (2020). arXiv preprint. arXiv:2008.01449
5. K. Wang, J.H. Liew, Y. Zou, D. Zhou, J. Feng, Panet: few-shot image semantic segmentation with prototype alignment, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9197–9206
6. C. Zhang, G. Lin, F. Liu, R. Yao, C. Shen, Canet: class-agnostic segmentation networks with iterative refinement and attentive few-shot learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5217–5226
7. G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, J. Kim, Adaptive prototype learning and allocation for few-shot segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 8334–8343
8. Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, L. Shao, Learning non-target knowledge for few-shot semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 11573–11582
9. C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, R. Yao, Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9587–9595
10. Y. Liu, X. Zhang, S. Zhang, X. He, Part-aware prototype network for few-shot semantic segmentation, in *European Conference on Computer Vision* (Springer, Berlin, 2020), pp. 142–158
11. C. Lang, G. Cheng, B. Tu, J. Han, Learning what not to segment: a new perspective on few-shot segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 8057–8067
12. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440
13. O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, Berlin, 2015), pp. 234–241
14. F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions (2015). arXiv preprint. arXiv:1511.07122
15. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2881–2890
16. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
17. Y. Lifchitz, Y. Avrithis, S. Picard, A. Bursuc, Dense classification and implanting for few-shot learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 9258–9267
18. P. Tokmakov, Y.-X. Wang, M. Hebert, Learning compositional representations for few-shot recognition, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 6372–6381
19. W. Jiang, K. Huang, J. Geng, X. Deng, Multi-scale metric learning for few-shot learning. IEEE Trans. Circuits Syst. Video Technol. **31**(3), 1091–1102 (2020)
20. Y. Yang, F. Wei, M. Shi, G. Li, Restoring negative information in few-shot object detection. Adv. Neural Inf. Process. Syst. **33**, 3521–3532 (2020)
21. Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, G. Li, Learning to prompt for open-vocabulary object detection with vision-language model, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 14084–14093
22. O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra et al., Matching networks for one shot learning. Adv. Neural Inf. Process. Syst. **29** (2016). arXiv preprint. arXiv:1606.04080
23. A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, Meta-learning with memory-augmented neural networks, in *International Conference on Machine Learning* (2016), pp. 1842–1850. PMLR
24. J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning. Adv. Neural Inf. Process. Syst. **30** (2017). arXiv preprint. arXiv:1703.05175
25. F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: relation network for few-shot learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 1199–1208
26. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *International Conference on Machine Learning* (2017), pp. 1126–1135. PMLR
27. A. Nichol, J. Schulman, Reptile: a scalable metalearning algorithm, (2018). arXiv preprint. arXiv:1803.02999
28. A.A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, R. Hadsell, Meta-learning with latent embedding optimization (2018). arXiv preprint. arXiv:1807.05960
29. K. Lee, S. Maji, A. Ravichandran, S. Soatto, Meta-learning with differentiable convex optimization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10657–10665
30. A. Shaban, S. Bansal, Z. Liu, I. Essa, B. Boots, One-shot learning for semantic segmentation (2017). arXiv preprint. arXiv:1709.03410
31. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in *International Conference on Machine Learning* (2020), pp. 1597–1607. PMLR
32. W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, L. Van Gool, Exploring cross-image pixel contrast for semantic segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 7303–7313
33. M. Agarwal, M. Yurochkin, Y. Sun, On sensitivity of meta-learning to support data. Adv. Neural Inf. Process. Syst. **34**, 20447–20460 (2021)
34. M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
35. B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in *2011 International Conference on Computer Vision* (IEEE Press, New York, 2011), pp. 991–998
36. X. Zhang, Y. Wei, Y. Yang, T.S. Huang, Sg-one: similarity guidance network for one-shot semantic segmentation. IEEE Trans. Cybern. **50**(9), 3855–3865 (2020)
37. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.