Dual Contrastive Learning with Anatomical Auxiliary Supervision for Few-shot Medical Image Segmentation

Huisi Wu(✉), Fangyan Xiao, and Chongxin Liang

College of Computer Science and Software Engineering, Shenzhen University
hswu@szu.edu.cn, {2070276124, 2060271074}@email.szu.edu.cn

Abstract. Few-shot semantic segmentation is a promising solution for scarce data scenarios, especially for medical imaging challenges with limited training data. However, most of the existing few-shot segmentation methods tend to over rely on the images containing target classes, which may hinder its utilization of medical imaging data. In this paper, we present a few-shot segmentation model that employs anatomical auxiliary information from medical images without target classes for dual contrastive learning. The dual contrastive learning module performs comparison among vectors from the perspectives of prototypes and contexts, to enhance the discriminability of learned features and the data utilization. Besides, to distinguish foreground features from background features more friendly, a constrained iterative prediction module is designed to optimize the segmentation of the query image. Experiments on two medical image datasets show that the proposed method achieves performance comparable to state-of-the-art methods.[1]

Keywords: Few-shot Segmentation, Medical Image Segmentation, Contrastive Learning, Anatomical Information

1 Introduction

Automatic segmentation of medical images is widely used in clinical applications such as lesion localization, disease diagnosis, and prognosis. Recently, although segmentation networks based on fully supervised deep learning have achieved excellent performance [7, 34, 49, 48], they rely on large amounts of pixel-level annotations and are difficult to directly be applied to segment unseen classes. For medical images, especially CT and MRI, labeling images is often expensive and even requires years of clinical experience by experts. And it is generally impractical to retrain a model for each tissue or organ class. These problems lead to the challenges of medical image segmentation with few manual annotations and poor model generalization ability.

Few-shot learning is proposed as a potential developable scheme to alleviate those challenges, and the segmentation method based on few-shot learning is

[1] Code is available at: https://github.com/cvszusparkle/AAS-DCL_FSS

called few-shot semantic segmentation (FSS) [32]. The idea is to adopt the prior knowledge distilled on labeled samples (denoted as *support*) to segment unlabeled samples (denoted as *query*). Therefore, FSS often learns few-shot *tasks* composed of *base* classes in an episodic training manner, and segments *unseen* classes in the form of *tasks* in the inference stage.

Few-shot segmentation methods have made considerable progress in recent years, but most of them are applied to natural images [29, 40, 39, 44, 50, 21, 17, 46], while FSS approaches for medical imaging are still developing [24, 51, 31, 28, 9, 35, 36]. There are the following possible factors: due to the low contrast, the fundamental difference in modalities and the less amount of data of medical images (such as CT or MRI), methods based on natural images cannot be directly applied to medical images; due to most of the current FSS methods for medical imaging rely on images with target classes, while medical image datasets contain limited target classes images, and some images without target classes are discarded, resulting in low data availability. However, those images without target classes (denoted as *non-target* images in this paper) may be rich in some anatomical knowledge. Therefore, how to efficiently employ the existing medical image data (i.e. try not to discard any data) and design a segmentation network suitable for the characteristics of medical images are quite significant. Nevertheless, so far, it seems that only SSL-ALPNet [28] considered applying those medical images without target classes to FSS, other methods still only depend on images containing target classes [9, 31, 35, 36].

In order to enhance data availability and improve segmentation performance under the bottleneck of scarce training data, we consider the acquisition of meaningful knowledge from *non-target* images for contrastive learning. Intuitively, the information of irrelevant tissues contained in non-target slices are the anatomical background cues of the target classes. Taking those non-target slices into consideration to construct contrastive learning can greatly increase the capability of the network for learning discriminative features. At present, contrastive learning has been widely used in visual tasks recently, such as [4, 5, 12, 11]; it has also been gradually introduced into few-shot classification, such as [19, 27]; but it is developing slowly in few-shot segmentation, especially for medical imaging. The possible reason is that it is not as simple to construct positive and negative samples for medical images as compared to natural images, and there are challenges in both the number of data samples and the characteristics of the image modalities. How to construct effective contrastive learning to help few-shot medical image segmentation may be a developable thought.

In this paper, we propose a scheme for designing *dual contrastive learning with anatomical auxiliary supervision*, called *AAS-DCL*. The purpose is to adopt the tissue knowledge in the images without target classes as negative samples, and the support guidance information as positive samples, to construct prototypical contrastive learning and contextual contrastive learning, to strengthen the discriminability of the features learned by the network and the data utilization. Furthermore, in order to more effectively discriminate foreground and background features, we also build a *constrained iterative prediction* (CIP) mod-

ule to optimize the query segmentation. That is, by exploring the distribution consistency between the feature similarity map and mask similarity map to constrain the update of the prediction results of the classifier.

Our contributions are summarized as follows:

- We propose to make full use of the anatomical information in the medical images without target classes to construct *dual contrastive learning* (DCL), from both prototypical and contextual perspectives, to encourage the network to learn more discriminative features and improve the data utilization.
- We present a *constrained iterative prediction* module to optimize the prediction mask of the classifier, which can effectively discriminate the query features learned by the network equipped with the DCL module.
- Our proposed method set the start-of-the-art performance of few-shot segmentation in two famous medical image datasets.

2 Related Work

Few-shot Semantic Segmentation. The FSS methods for **natural images** are emerging in endlessly [6, 17, 21, 32, 37, 39, 40, 44, 46, 50]. OSLSM [32] proposed the pioneering two branches and generated weights from support images for few-shot segmentation; PL [6] proposed a prototypical framework tailored for few-shot natural image segmentation; PANet [40] designed an alignment loss to help the network learn the consistency between the support and query feature spaces. PMMs [44] applied the *Expectation-Maximization* algorithm to extract multiple partial prototypes to help the query image's segmentation; PPNet [21] proposed to build local prototypes based on superpixels; PFENet [37] designed to construct prior masks and effective feature enrichment module to reconcile the spatial inconsistency of query and support features. More recently, both RePRI [3] and CWT [22] considered some unique training schemes, such as transductive inference or multi-stage training, opening up some thought-provoking few-shot segmentation insights. And some of those innovative units for natural images may be conductive to few-shot medical image segmentation.

As for **medical imaging**, the FSS methods have also progressed in recent years, and most of them are implemented based on 2D frameworks [43, 45, 31, 35, 28, 36]. As a pioneering work, sSENet [31] designed a strong interactive segmentation network tailored for the traits of medical images; GCN-DE [35] improved part of the structure in sSENet by applying a global correlation module with discriminative embedding; PoissonSeg [33] applied *Poisson* learning to propagate supervised signals and spatial consistency calibration for representative learning; SSL-ALPNet [28] proposed to adopt the idea of self-supervision without annotations, designed adaptive local prototypical network and obtained state-of-the-art effect. RP-Net [36] designed a clever model combined class-level prototypical network with iterative refinement to segment the query image and got the state-of-the-art performance. Besides, AKFNet [41] proposed to subtly use the anatomical knowledge from support images to guide the query segmentation. There are also FSS methods based on 3D frameworks, such as BiGRU [14], which

proposed to capture the correlation between adjacent slices of medical volumes. It can be seen that most of the current few-shot segmentation methods rely on images with target classes. Differently, we propose to apply slices without target classes for auxiliary supervision to construct contrastive learning, which will help the network to learn more discriminative features.

Contrastive Learning. Contrastive learning aims to learn the similarity of various samples in embedding space, that is, to pull closer the more similar samples and push farther the dissimilar ones. Recent works have been increasingly used in computer vision tasks, including self-supervised ideas [5, 11, 12] and supervised methods [13, 42]. SimCLR [4] applied different data augmentations on same images for self-supervised contrastive learning; MoCo [12] proposed a momentum encoder to increase memory banks to store considerable views of different data transformations. Recently, contrastive learning has also been adopted to few-shot learning tasks [10, 19, 23, 27]. Liu et al. [19] proposed to use distinct data augmentations into a few-shot embedding network for contrastive learning; Ouali et al. [27] employed spatial features to build contrastive learning for few-shot classification. For few-shot segmentation, there is a slow development of contrastive learning. Liu et al. [20] presented to apply global and local contrastive losses to pre-train feature extractor for query prior features to help few-shot segmentation of natural images. DPCL [15] was designed as a dual prototypical contrastive learning network suitable for few-shot natural image segmentation. By exploring the traits and challenges of medical image datasets, we design contrastive learning for few-shot medical image segmentation.

Superpixel Segmentation. Superpixels are pixel groups generated by clustering pixels using statistical methods according to image features. Based on superpixels, one image can be segmented to different degrees, which can generate pseudo-labels to provide effective supervision for unlabeled scenarios. There are some popular methods for superpixel segmentation, such as SLIC [1], SEEDS [2], Graph-cuts [25] and Felzenszwalb's [8]. SSL-ALPNet [28] applied the way [8] to produce pseudo-labels of unlabeled images by superpixel segmentation for self-supervision; Li et al. [18] employed the same method [8] to generate superpixels as pseudo- classes and construct self-supervised tasks. In this work, different superpixel segmentation algorithms will be studied to generate pseudo-labels for *non-target* medical slices to construct contrastive learning.

3 Method

3.1 Problem Definition

The general formulation of few-shot semantic segmentation (FSS) is mainly followed by [32]. In the FSS setting, the idea is to train a model on an annotated training dataset $\mathcal{D}_{train}$ that can perform segmentation well on the testing data $\mathcal{D}_{test}$ with a few labeled samples, without re-training. The training classes $\mathcal{C}_{train}$

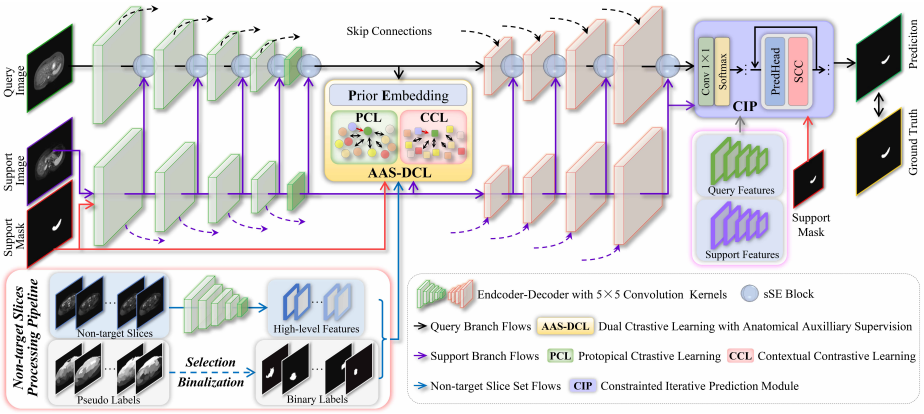

Fig. 1. Workflow of the proposed framework for few-shot medical image segmentation, which is based on the baseline sSENet [31] with skip connections, and instrumental schemes are designed including AAS-DCL and CIP modules

have no overlap with the testing classes $\mathcal{C}_{test}$, *i.e.*, $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$. The episodic training is adopted in FSS, where the model is trained with many epochs and one epoch contains many episodes. To be specific, in each episode, the few-shot learning consists of a support and query data pair, *i.e.*, $\mathcal{D}_{train}$ is composed of the support set $\mathcal{S}_{tr}$ and query set $\mathcal{Q}_{tr}$. Here, $\mathcal{S}_{tr} = \{(\mathbf{x}^s(c_i), \mathbf{y}^s(c_i))\}$ and $\mathcal{Q}_{tr} = \{(\mathbf{x}^q(c_i))\}$, where $x \in \mathcal{X}$ refers to the image and $\mathbf{y} \in \mathcal{Y}$ refers to corresponding binary mask, $i = 1, 2, ..., |c|$ is the number index of class $c \in \mathcal{C}_{train}$. In *inference*, $\mathcal{D}_{test}$ is defined in the same mode but for testing images and masks with the unseen class set $\mathcal{C}_{test}$. The background class (denoted as c_0) is not counted in $\mathcal{C}_{train}$ and $\mathcal{C}_{test}$.

3.2 Network Overview

Inspired by the strong interactive structure designed by sSENet [31] for medical images, we adopt it as a basic network and improve it subtly. As shown in Fig. 1, the framework includes the following parts: (1) Encoder-decoder structures: feature extraction and reconstruction referring to the sSENet with added skip connections; (2) *Non-target slices processing pipeline*: generating pseudo-labels for non-target slices and extracting features for randomly selected partial slices of *non-target slice set*; (3) *Dual contrastive learning with anatomical auxiliary supervision* (AAS-DCL): prototypes generation and feature processing for prototypical and contextual contrastive learning, respectively; (4) *Constrained iterative prediction* (CIP): a designed classifier for iterative query prediction.

Non-target Slices Processing Pipeline. For CT and MRI datasets, most raw scans have several slices without target organs, and their masks are all black. We define the slices as *non-target slice set* $\mathcal{D}_{nt}$. Since those non-target slices are rich in anatomical knowledge about irrelevant organs or tissues, they are inherently

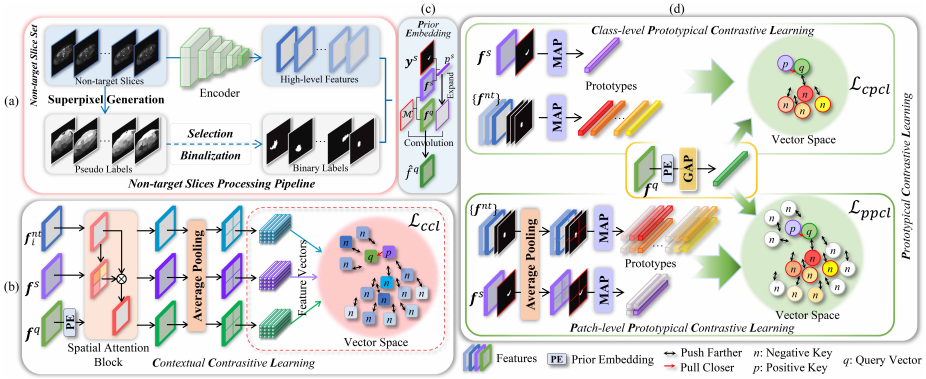

Fig. 2. The diagram of dual contrastive learning with anatomical auxiliary supervision (AAS-DCL). (a) Non-target slices processing pipeline: applying superpixel segmentation to generate pseudo-labels and then selecting randomly some slices to extract features by the encoder. (b) The workflow of contextual contrastive learning (CCL). (c) The structure of the prior embedding (PE) module. (d) The workflow of the prototypical contrastive learning, including two sub-modules (CPCL and PPCL)

discriminative from target classes. To effectively utilize these information, as shown in Fig. 2(a), we first use a superpixel segmentation algorithm ([1, 2] or [8]) to produce pseudo-labels offline for $\mathcal{D}_{nt}$. Besides, K non-target slices are randomly selected in each training episode, defined as $\mathcal{X}_{nt} = \{(\mathbf{x}^{nt})\}^K$. For their pseudo-labels, one superpixel (denoted as a pseudo-class) in each pseudo-label is randomly selected to binarize the label, and the K binary pseudo-labels are defined as $\mathcal{Y}_{nt} = \{(\mathbf{y}^{nt})\}^K$. Then the non-target slices $\mathcal{X}_{nt}$ are sent to the network encoder to extract non-target features $\mathcal{F}_{nt} = \{(\boldsymbol{f}^{nt})\}^K$; finally, the pseudo-labels $\mathcal{Y}_{nt}$ and the feature maps $\mathcal{F}_{nt}$ are input into the AAS-DCL scheme.

3.3 Dual Contrastive Learning with Anatomical Auxiliary Supervision

Due to the low contrast of medical images, the demarcation between the target class and the background tissues is not obvious, which makes it difficult to accurately segment target organs. Therefore, we explore the utilization of non-target slices with rich anatomical knowledge to provide more background guidance, which can constitute contrastive learning with query and support features. For one thing, that will help to enhance the discriminability of learned features of the model and improve the segmentation performance; for another, it will also take better advantage of medical image datasets in few-shot segmentation.

As shown in Fig. 2, in the AAS-DCL scheme, we design *prototypical contrastive learning* (PCL) and *contextual contrastive learning* (CCL) to form the DCL module in each training episode, from the perspectives of semantic class and spatial information, respectively, to make the features similar to the representation information of the target class closer, and the dissimilar ones are farther

away. The contrastive losses applied in the DCL module all refer to infoNCE [26]:

$$\mathcal{L}_{cl}(q, k^+, k^-) = -\log \frac{\exp\left(q \cdot k^+/\tau\right)}{\sum_{i=0}^{K} \exp\left(q \cdot k_i^-/\tau\right)} \tag{1}$$

where K denotes the number of negative keys and τ is a temperature factor.

Prior Embedding. To further guide the activation of foreground class-specific information in query features, inspired by PFENet [37], we design the *Prior Embedding* (PE) (Fig. 2(c)). The module needs to obtain the confidence map $\mathcal{M}$ by the normalized similarity map between the support and query feature [37], and the class-level prototype p^s generated from the support feature and its mask. Then the similarity map $\mathcal{M}$, the expanded prototype $\mathcal{E}(p^s)$ and the query feature $\boldsymbol{f}^q$ are concatenated and convolved to get an enhanced query feature $\hat{\boldsymbol{f}}^q$.

Prototypical Contrastive Learning. Prototypes are vectors rich in semantic information; and most few-shot methods obtain prototypes by global average pooling of features and corresponding masks, which are often denoted as class-level prototypes to guide the segmentation of the query image. However, since class-level prototypes are prone to lose intra-class information, many few-shot methods [44, 21, 28] proposed to produce partial prototypes to retain sufficient intra-class cues. Our PCL includes two sub-modules: *class-level prototypical contrastive learning* (CPCL) and *patch-level prototypical contrastive learning* (PPCL), which comprehensively discriminate semantic information from the global and local prototypes, respectively.

(i) Class-level Prototypical Contrastive Learning. Class-level prototype-based learning can distinguish foreground and background features from the perspective of the overall semantic class. To be specific, the support feature and support mask, non-target features and the corresponding binary pseudo-labels are used to generate their class-level prototypes by masked averaged pooling (MAP) operation [50], respectively. And the query feature $\hat{\boldsymbol{f}}^q$ is used to obtain a mean vector v^q by global average pooling (GAP). In addition, the vector is regarded as the *query vector*; the support prototype is regarded as a positive key; and those non-target prototypes are regarded as negative keys. As shown in the CPCL of Fig. 2(d), the contrastive loss is formulated as:

$$\mathcal{L}_{cpcl} = \mathcal{L}_{cl}(v^q, p^s, p^{nt}) = -\log \frac{\exp\left(v^q \cdot p^s/\tau\right)}{\sum_{i=0}^{K} \exp\left(v^q \cdot p_i^{nt}/\tau\right)} \tag{2}$$

(ii) Patch-level Prototypical Contrastive Learning. We propose the PPCL for two reasons: for one thing, since the class-level prototypes obtained by MAP will filter out other background information around the target class, which may be utilized as negative samples to enhance the discriminativeness of the semantics around the target class, so the PPCL is used to remedy this problem; for another, because contrastive learning largely requires amounts of effective

negative samples, it is not sufficient to only consider class-level prototypes. The generation of patch-level prototypes is inspired by the local prototypes in [28]. First, the entire feature map is evenly divided based on patches, and then MAP is applied to each feature patch and the mask patch of the corresponding position to obtain the local prototypes, we called them patch-level prototypes. Given the patch size (D_H, D_H) and a feature map $\boldsymbol{f} \in \mathbb{R}^{C\times H\times W}$, the process is formulated as follows:

$$p_{d_j} = \frac{\sum_{(h,w)\in d_j} \boldsymbol{f}_{d_j}(h,w)\cdot \mathbf{y}_{d_j}(h,w)}{\sum_{(h,w)\in d_j} \mathbf{y}_{d_j}(h,w)} \tag{3}$$

where the (h, w) are spatial coordinates in each patch, p_{d_j} refers to the prototype in a patch d_j and $j = 1, 2, ..., \frac{D_H}{H}$ denotes the patch index.

The support feature and its mask can obtain both some patch-level background prototypes $\{p_d^{s-}\}$ and the patch-level foreground prototype p_d^{s+} by setting a threshold [28]. Similarly, each non-target feature and corresponding pseudo-label are also used to produce mickle patch-level prototypes $\{p_d^{nt}\}$. The query feature $\hat{\boldsymbol{f}^q}$ is used to generate a mean vector v^q by GAP operation. When building the contrastive loss, the support foreground prototype p_d^{s+} is regarded as a positive key, and other support patch-level prototypes $\{p_d^{s-}\}$ and all non-target patch-level prototypes $\{p_d^{nt}\}$ are regarded as negative keys. This scheme increases the number of prototype samples, which is beneficial to learning the correlation between local features of similar classes and the discriminativeness among local features of dissimilar semantic classes. As shown in the PPCL of Fig. 2(d), the contrastive loss is written as:

$$\begin{aligned}\mathcal{L}_{ppcl} &= \mathcal{L}_{cl}\big(v^q, p_d^{s+}, (p_d^{s-}, p_d^{nt})\big) \\ &= -\log \frac{\exp\left(v^q \cdot p_d^{s+}/\tau\right)}{\sum_{i=0}^{K}\exp\left(v^q\cdot p_d^{nt}(i)/\tau\right) + \sum_{i=0}^{(\frac{D_H}{H}-1)}\exp\left(v^q\cdot p_d^{s-}(i)/\tau\right)}\end{aligned} \tag{4}$$

Contextual Contrastive Learning. In order to increase the discriminativeness of context features, we propose the contextual contrastive learning (CCL), as shown in Fig. 2(b). Specifically, the support feature, enhanced query feature $\hat{\boldsymbol{f}^q}$ and non-target features are first processed by a spatial attention block [30], to make all feature maps focus on richer contextual information. Then, these processed features are averagely pooled to obtain different feature vector sets, which will be used to construct contrastive learning. When calculating once a CCL loss, a query feature vector $\overline{\boldsymbol{f}}_j^q$ is regarded as the *query vector*, the support feature vector $\overline{\boldsymbol{f}}_j^s$ at the same position is regarded as a positive key, and all non-target feature vectors are regarded as negative keys. The number of loss calculations N is equal to the number of query feature vectors, and finally the mean value of the cumulative sum of all CCL losses obtained is adopted as the output contextual contrastive loss. The formula is as follows:

$$\mathcal{L}_{ccl} = \frac{1}{N}\sum_{j=0}^{N}\mathcal{L}_{cl}^{j}(\overline{\boldsymbol{f}}_j^q, \overline{\boldsymbol{f}}_j^s, \overline{\boldsymbol{f}}^{nt}) = \frac{1}{N}\sum_{j=0}^{N}\big[-\log\frac{\exp\left(\overline{\boldsymbol{f}}_j^q\cdot\overline{\boldsymbol{f}}_j^s/\tau\right)}{\sum_{i=0}^{K}\exp\left(\overline{\boldsymbol{f}}_j^q\cdot\overline{\boldsymbol{f}}_i^{nt}/\tau\right)}\big] \tag{5}$$

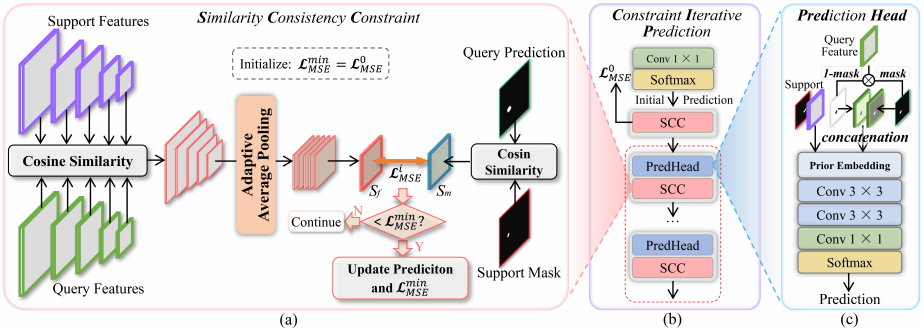

Fig. 3. The diagram of constrained iterative prediction (CIP). (a) The details of similarity consistency constraint (SCC). (b) The main iteration optimization process of CIP. (c) The structure of the designed classifier to generate the query prediction

3.4 Constrained Iterative Prediction

The design of the dual contrastive learning module is beneficial to enhance the features' discriminability of the encoder-decoder. To encourage the foreground and background of query feature to be segmented more effectively on the final classifier, so the CIP module is proposed to optimize the part, which can also as an assistant to the DCL module. As shown in Fig. 3(b), the basic components include *similarity consistency constraint* (SCC) and *prediction head*.

Prediction Head. As shown in Fig. 3(c), unlike only a generic classifier (1x1 convolution and softmax), in order to predict more meaningfully, we consider integrating the query prediction into the query feature by convolution, which is derived from the iterative optimization idea of RP-Net [36] and CANet [47]. RP-Net fuses the prediction and the high-level query feature by a correlation matrix, where a prototype-based classifier is further employed to obtain a new prediction. And CANet fuses the prediction and the middle-level query feature by residual concatenation, where an ASPP is further used to produce a new prediction. Differently, we first utilize the result of the generic classifier as the initial prediction mask, fuse it with the query feature of the decoder's tail and send into the *prior embedding*(PE) module, where the new segmentation result is obtained by a series of convolution operations. The application of PE is to make use of the guidance information of the support feature to promote the fusion of the predicted query mask. The new query prediction will be determined to be updated through the following constraints in SCC.

Similarity Consistency Constraint. As shown in Fig. 3(a), we consider the information distribution constraints and first compute two similarity maps, $\mathcal{S}_m$ and $\mathcal{S}_f$. The $\mathcal{S}_m$ is calculated between the predicted query mask and the support mask; $\mathcal{S}_f$ is calculated from the support and query features obtained by the encoder. Intuitively, the probability value distributions on $\mathcal{S}_m$ and $\mathcal{S}_f$ should be

as consistent as possible, because a more consistent distribution means that the predicted query mask is closer to ground truth. We minimize the MSE loss to quantify the better distribution consistency between $\mathcal{S}_m$ and $\mathcal{S}_f$, and employ an iterative strategy to use this loss to constrain whether to update the final prediction result and the iteration times I is experimentally set to 5.

3.5 Training Strategy

For each episode, one support and query pair is selected according to the strategy in [31] to form a 1-way 1-shot few-shot setting, and K non-target slices and corresponding pseudo-labels are randomly selected. All pairs are input to the network for end-to-end learning. For loss function, we adopt cross-entropy loss to supervise the segmentation of query image. And three contrastive losses are employed to learn the feature discriminability of the model, so the total loss is:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{cpcl} \cdot \mathcal{L}_{cpcl} + \lambda_{ppcl} \cdot \mathcal{L}_{ppcl} + \lambda_{ccl} \cdot \mathcal{L}_{ccl} \tag{6}$$

where the $\lambda_{cpcl}, \lambda_{ppcl}$ and λ_{ccl} are experimentaly set as 0.08, 0.12 and 0.06, respectively.

4 Experiments

4.1 Dataset and Evaluation Metric

In order to evaluate the effectiveness of the proposed method, we conducted experiments on two public medical image datasets: (1) **CHAOS-T2** is from the challenge (Task 5) of ISBI 2019 [38], and it contains 20 MRI scans of *T2SPIR* with 623 axial 2D slices in total, including 492 slices with target classes and 131 non-target slices. (2) **Synapse** is from the public challenge [16], and it contains 30 CT scans with 3779 axial abdominal slices, including 1755 slices with target classes involved in our experiments, the remaining are non-target slices.

Due to the 2D framework of our model, experiments are performed with images sliced from 3D scans along the 2D axis. Then five-fold cross-validation is applied in our experiments. To simulate the scarcity of annotated data in realistic situations, we performed 1-way 1-shot setting experiments for four organ classes, including the *liver*, *left kidney*(LK), *right kidney*(RK) and *spleen*. We employ one of the four classes as the unseen class and the other three classes as the base classes, so that four few-shot tasks are constructed. To evaluate the segmentation performance of 2D slices on 3D scans, we follow the protocol in the work [31] and apply the Dice coefficient as the major evaluation metric. In addition, we also conduct the statistical significance analysis, which is reported in the supplementary material.

4.2 Implementation Details

The network is implemented by Pytorch on an Nvidia RTX 2080Ti GPU. All 2D slices are resized to the resolution of 256 $\times$ 256. In PPCL, the number of

Table 1. Statistical comparison (in Dice score %) of different methods on the validation sets for CHAOS-T2 and Synapse datasets

Method	CHAOS-T2					Synapse				
	Liver	RK	LK	Spleen	Mean	Liver	RK	LK	Spleen	Mean
PANet [40]	41.87	44.55	50.64	49.94	46.75	36.18	25.09	32.41	22.13	28.95
sSENet [31]	27.25	59.22	54.89	48.45	47.45	41.87	34.55	40.60	39.94	39.24
PoissonSeg [33]	61.03	53.57	50.58	52.85	54.51	58.74	47.02	50.11	52.33	52.05
GCN-DE [35]	49.47	83.03	76.07	60.63	67.30	46.77	**75.50**	**68.13**	56.53	61.73
SSL-ALPNet [28]	68.71	79.53	71.14	61.11	70.12	73.26	57.73	64.34	64.89	65.05
RP-Net [36]	69.73	82.06	**77.55**	73.90	75.81	**74.11**	68.32	63.75	65.24	67.85
Ours	**69.94**	**83.75**	76.90	**74.86**	**76.36**	71.61	69.95	64.71	**66.36**	**68.16**

patches is experimentally set to 4. By using an initial learning rate of 0.0001, we use the loss function (Equ. 6) for episodic training with a batch size of 1. And the learning rate is based on a polynomial decay learning rate policy. Besides, the Adam algorithm is employed to optimize the learning process with power = 0.95 and weight decay = 1e-7. The model is trained by 70 epochs, each of which contains the number of episodes equal to the scale of the training set. For data augmentation, we focus more on random modifying the image sharpness, with the visibility factor range of [0.2, 0.5] and the lightness factor range of [0.5, 1.0].

4.3 Comparisons with State-of-the-art Methods

To demonstrate the superiority of the proposed method, we compare our method with the classical and current state-of-the-art FSS methods, as shown in Table 1 and Fig. 4. The PANet [40] proposed an alignment regularization loss, which is utilized widely by other few-shot methods; sSENet [31] is the pioneering work of FSS for medical imaging; GCN-DE [35] employs global correlation module to optimize the structure of sSENet; PoissonSeg [33] considers *Poisson* learning and spatial consistency calibration for few-shot medical image segmentation; SSL-ALPNet [28] utilizes self-supervision and adaptive local prototypes for FSS; RP-Net [36] applies global prototypes with recurrent refinement for prediction. Since GCN-DE and PoissonSeg are not open source, we directly compare the statistical results from their original works; besides, the results of other FSS methods are obtained by re-experiments in a unified experimental environment.

Table 1 shows that compared with other few-shot methods of medical imaging, for the **CHAOS-T2** dataset, our method achieves relatively best performance in most target classes, including right kidney, liver and spleen, as well as the mean Dice score; for the **Synapse** dataset, our method also get the best Dice scores on left kidney, spleen and mean Dice. It is precisely because our method makes full use of anatomical auxiliary knowledge of non-target slices to help the network learn more discriminative features through contrastive learning, and applies the CIP module for more effective prediction to make the segmentation performance better.

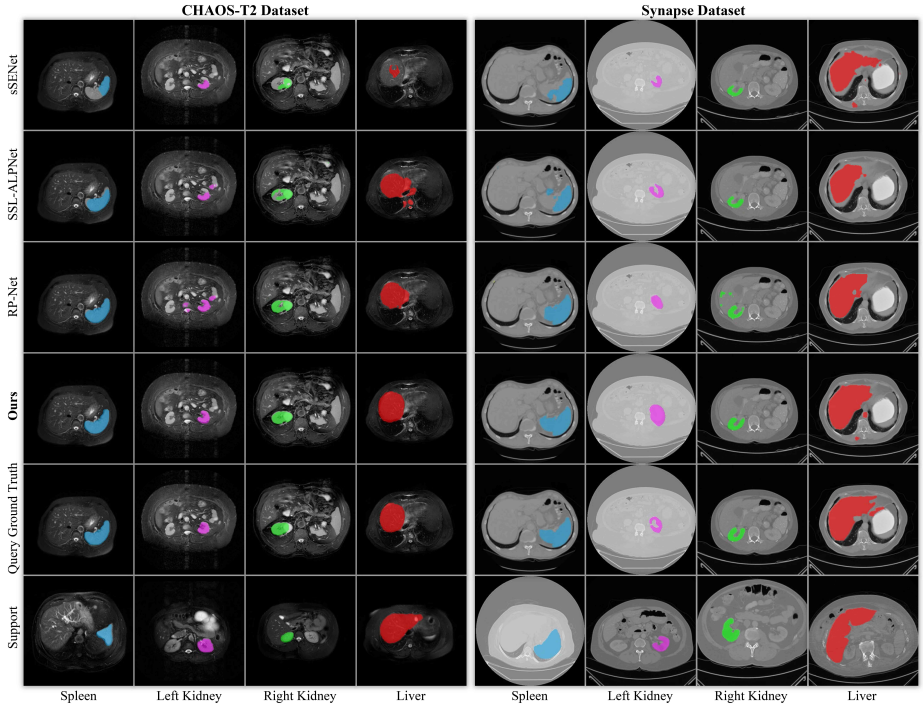

Fig. 4. Visual comparison of different methods on the CHAOS-T2 and Synapse datasets

4.4 Ablation Studies

Ablation experiments are performed on the CHAOS-T2 dataset. Based on the baseline sSENet [31], we focus on the roles of the proposed dual contrastive learning (DCL) and constrained iterative prediction (CIP), the impact of the sub-modules (CPCL, PPCL and CCL) in DCL, and other factors of the network.

Effect of the DCL module. To demonstrate the effect of the DCL module, we performed ablation experiments on it, the data results are shown in Table 2 and Table 3. And Fig. 5 shows the visualization results of the DCL module, which applies the visualized feature map after 1×1 convolution and before the *softmax* function. Table 2 shows that the DCL can effectively enhance the Dice performance due to its promotion to the network's ability of learning discriminative features. For the sub-modules CPCL and PPCL in PCL, Table 3 and Fig. 5 both show that when the two kinds of prototypical contrastive losses are combined concurrently, the performance gets better. Because the two can assist each other: CPCL is conducive to learning the gap among semantic classes, and PPCL effectively utilizes local prototypes to improve the identifiability among the local features within classes. Then the CCL can increase the discriminability of contextual features and also help improve the segmentation performance.

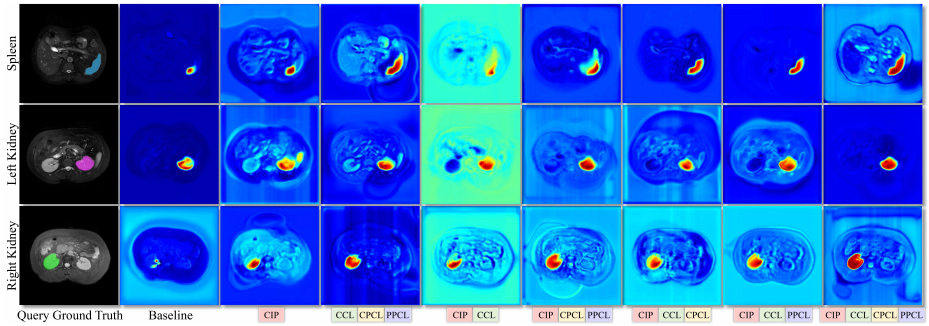

Fig. 5. Visual feature maps of different components based on the baseline‡(on the CHAOS-T2 dataset). The warmer colors represent the better discriminative features

Table 2. Ablation study results (in Dice score %) on the CHAOS-T2 dataset for the DCL and CIP modules. Baseline†: sSENet [31]; Baseline‡: sSENet with skip connections

Methods	Liver	RK	LK	Spleen	Mean
Baseline†[31]	27.25	59.22	54.89	48.45	47.45
Baseline‡	30.21	62.47	55.96	47.68	49.08
Baseline‡+CIP	51.02	69.84	63.77	63.65	62.07
Baseline‡+DCL	56.91	75.23	67.19	69.52	67.21
Baseline‡+DCL+CIP	**69.94**	**83.75**	**76.90**	**74.86**	**76.36**

Table 3. Ablation study results (in Dice score %) on the CHAOS-T2 dataset for sub-modules (CPCL, PPCL and CCL) of the DCL module

CCL	CPCL	PPCL	Liver	RK	LK	Spleen	Mean
✓			59.25	71.64	69.04	65.66	66.40
	✓	✓	67.63	79.14	71.09	70.79	72.16
✓	✓		68.00	80.26	72.22	71.85	73.08
✓		✓	68.53	80.80	73.17	72.54	73.76
✓	✓	✓	**69.94**	**83.75**	**76.90**	**74.86**	**76.36**

Importance of the CIP module. Table 2 and Fig. 5 also show the ablation experimental results of CIP. It can be seen that the design of CIP can help the network to further improve the segmentation performance. The PerdHead and SCC cooperate to perform iterative optimization, which is quite significant for discriminative features learned by DCL in the network. More Dice performances under different iteration numbers are shown in Fig. 6.

Influence of other factors. (1) For the **superpixel segmentation algorithms** generating pseudo-labels for non-target slices, we adopt the SLIC [1], SEEDS [2] and Felzenszwalb's [8] to produce pseudo-labels respectively, the results are shown in Table 4. It is shown that Felzenszwalb's method gets the relatively best Dice score; because it could generate more irregularly shaped superpixel pseudo-labels that better fit anatomical contours, enabling our PCL to efficiently utilize them. The superpixels generated by SLIC are too regular and not friendly to the learning of anatomical information. The SEEDs-based method has a moderate performance. Moreover, the superpixels produced by the latter two would contain more all-black backgrounds, which may also affect the contrastive learning. So we employ the Felzenszwalb's as main algorithm for our experiments. Examples of related superpixels and pseudo-labels are shown in Fig. 7. (2) For the **iteration numbers** of CIP module, we employ different numbers for ablation experiments, the results are shown in Table 4. To get a trade-off between the performance and memory usage, we utilize five iterations in our experiments. (3) For **the number of randomly selected non-target**

Table 4. Ablation study results (in Dice score) on the CHAOS-T2 dataset for other factors

Alation Factors	Settings	Liver(%)	RK(%)	LK(%)	Spleen(%)	Mean(%)
Superpixel Segmentation	SLIC [1]	45.23	70.45	68.72	61.00	61.35
	SEEDS [2]	49.05	80.83	71.19	67.22	67.07
	Felzenszwalb's [8]	**69.94**	**83.75**	**76.90**	**74.86**	**76.36**
Iteration Numbers	I = 3	67.53	78.47	72.39	71.93	72.58
	I = **5**	69.94	**83.75**	**76.90**	**74.86**	**76.36**
	I = 7	**70.11**	83.58	76.17	74.62	76.12
Number of Slices	K = 1	64.60	76.34	71.08	69.33	70.34
	K = **3**	**69.94**	**83.75**	**76.90**	**74.86**	**76.36**
	K = 5	69.35	82.83	75.81	71.22	74.80

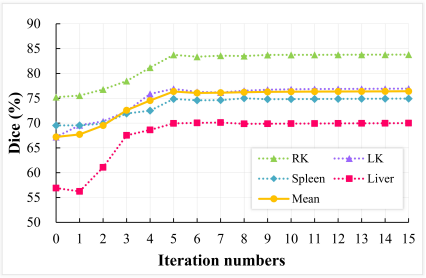

Fig. 6. Dice performance under different iteration numbers in the CIP module (on the CHAOS-T2 dataset)

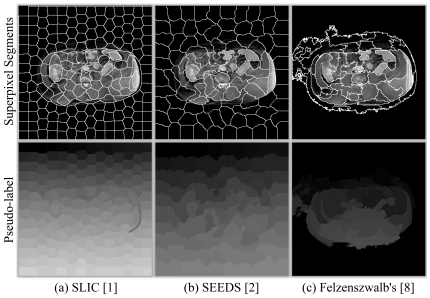

Fig. 7. Examples of superpixels and pseudo-labels obtained by different methods on a same non-target slice

slices, we consider the memory usage and performance and apply $K = 3$ for main experiments, the results are shown in Table 4.

5 Conclusion

We presented a network of dual contrastive learning with anatomical auxiliary information from medical images without target classes, including prototypical and contextual contrastive learning, and enhanced the discriminability of learned features and the data utilization. Besides, the designed constrained iterative prediction module optimized the query segmentation result. Experiments show the superiority of the proposed method for CT and MRI datasets. Moreover, The contrastive learning with the anatomical supervision from non-target images may provide a developable insight for few-shot medical image segmentation.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 61973221, the Natural Science Foundation of Guangdong Province of China under Grant 2018A030313381 and Grant 2019A1515011165.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. Tech. rep. (2010)
2. Bergh, M.V.d., Boix, X., Roig, G., Capitani, B.d., Gool, L.V.: Seeds: Superpixels extracted via energy-driven sampling. In: European Conference on Computer Vision. pp. 13–26. Springer (2012)
3. Boudiaf, M., Kervadec, H., Masud, Z.I., Piantanida, P., Ben Ayed, I., Dolz, J.: Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13979–13988 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
6. Dong, N., Xing, E.P.: Few-shot semantic segmentation with prototype learning. In: BMVC. vol. 3 (2018)
7. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Transactions on Medical Imaging **39**(11), 3619–3629 (2020)
8. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59**(2), 167–181 (2004)
9. Feng, R., Zheng, X., Gao, T., Chen, J., Wang, W., Chen, D.Z., Wu, J.: Interactive few-shot learning: Limited supervision, better medical image segmentation. IEEE Transactions on Medical Imaging (2021)
10. Gao, Y., Fei, N., Liu, G., Lu, Z., Xiang, T.: Contrastive prototype learning with augmented embeddings for few-shot learning. In: Uncertainty in Artificial Intelligence. pp. 140–150. PMLR (2021)
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
13. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020)
14. Kim, S., An, S., Chikontwe, P., Park, S.H.: Bidirectional rnn-based few shot learning for 3d medical image segmentation. arXiv preprint arXiv:2011.09608 (2020)
15. Kwon, H., Jeong, S., Kim, S., Sohn, K.: Dual prototypical contrastive learning for few-shot semantic segmentation. arXiv preprint arXiv:2111.04982 (2021)
16. Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI: Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge (2015)
17. Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., Kim, J.: Adaptive prototype learning and allocation for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8334–8343 (2021)

18. Li, Y., Data, G.W.P., Fu, Y., Hu, Y., Prisacariu, V.A.: Few-shot semantic segmentation with self-supervision from pseudo-classes. arXiv preprint arXiv:2110.11742 (2021)
19. Liu, C., Fu, Y., Xu, C., Yang, S., Li, J., Wang, C., Zhang, L.: Learning a few-shot embedding model with contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8635–8643 (2021)
20. Liu, W., Wu, Z., Ding, H., Liu, F., Lin, J., Lin, G.: Few-shot segmentation with global and local contrastive learning. arXiv preprint arXiv:2108.05293 (2021)
21. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: European Conference on Computer Vision. pp. 142–158. Springer (2020)
22. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8741–8750 (2021)
23. Majumder, O., Ravichandran, A., Maji, S., Achille, A., Polito, M., Soatto, S.: Supervised momentum contrastive learning for few-shot classification. arXiv preprint arXiv:2101.11058 (2021)
24. Mondal, A.K., Dolz, J., Desrosiers, C.: Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. arXiv preprint arXiv:1810.12241 (2018)
25. Moore, A.P., Prince, S.J., Warrell, J., Mohammed, U., Jones, G.: Superpixel lattices. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
26. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018)
27. Ouali, Y., Hudelot, C., Tami, M.: Spatial contrastive learning for few-shot classification. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 671–686. Springer (2021)
28. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: European Conference on Computer Vision. pp. 762–780. Springer (2020)
29. Rakelly, K., Shelhamer, E., Darrell, T., Efros, A.A., Levine, S.: Few-shot segmentation propagation with guided networks. arXiv preprint arXiv:1806.07373 (2018)
30. Roy, A.G., Navab, N., Wachinger, C.: Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. IEEE Transactions on Medical Imaging **38**(2), 540–549 (2018)
31. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: 'squeeze & excite'guided few-shot segmentation of volumetric images. Medical Image Analysis **59**, 101587 (2020)
32. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410 (2017)
33. Shen, X., Zhang, G., Lai, H., Luo, J., Lu, J., Luo, Y.: Poissonseg: Semi-supervised few-shot medical image segmentation via poisson learning. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1513–1518. IEEE (2021)
34. Shi, G., Xiao, L., Chen, Y., Zhou, S.K.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Medical Image Analysis **70**, 101979 (2021)
35. Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., Paisley, J.: Few-shot medical image segmentation using a global correlation network with discriminative embedding. Computers in Biology and Medicine **140**, 105067 (2022)

36. Tang, H., Liu, X., Sun, S., Yan, X., Xie, X.: Recurrent mask refinement for few-shot medical image segmentation. arXiv preprint arXiv:2108.00622 (2021)
37. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J.: Prior guided feature enrichment network for few-shot segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
38. Valindria, V.V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 547–556. IEEE (2018)
39. Wang, H., Zhang, X., Hu, Y., Yang, Y., Cao, X., Zhen, X.: Few-shot semantic segmentation with democratic attention networks. In: European Conference on Computer Vision. pp. 730–746. Springer (2020)
40. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9197–9206 (2019)
41. Wei, Y., Tian, J., Zhong, C., Shi, Z.: Akfnet: An anatomical knowledge embedded few-shot network for medical image segmentation. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 11–15. IEEE (2021)
42. Wu, Z., Efros, A.A., Yu, S.X.: Improving generalization via scalable neighborhood component analysis. In: European Conference on Computer Vision. pp. 685–701 (2018)
43. Xiao, J., Xu, H., Zhao, W., Cheng, C., Gao, H.: A prior-mask-guided few-shot learning for skin lesion segmentation. Computing pp. 1–23 (2021)
44. Yang, B., Liu, C., Li, B., Jiao, J., Ye, Q.: Prototype mixture models for few-shot semantic segmentation. In: European Conference on Computer Vision. pp. 763–778. Springer (2020)
45. Yu, Q., Dang, K., Tajbakhsh, N., Terzopoulos, D., Ding, X.: A location-sensitive local prototype network for few-shot medical image segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 262–266. IEEE (2021)
46. Zhang, B., Xiao, J., Qin, T.: Self-guided and cross-guided learning for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8312–8321 (2021)
47. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: 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. pp. 5217–5226 (2019)
48. Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., Yu, Y.: Cross-modality deep feature learning for brain tumor segmentation. Pattern Recognition **110**, 107562 (2021)
49. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Transactions on Medical Imaging **39**(7), 2531–2540 (2020)
50. Zhang, X., Wei, Y., Yang, Y., Huang, T.S.: Sg-one: Similarity guidance network for one-shot semantic segmentation. IEEE Transactions on Cybernetics **50**(9), 3855–3865 (2020)
51. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8543–8553 (2019)