Ye and Zhang
EURASIP Journal on Advances in Signal Processing (2023) 2023:49
https://doi.org/10.1186/s13634-023-00999-z

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

A dynamic few-shot learning framework for medical image stream mining based on self-training

Zhengqiang Ye[1] and Wei Zhang[2*]

*Correspondence:
weizhang@shmu.edu.cn

[1] Information Center, Eye and ENT Hospital, Fudan University, Shanghai, China
[2] Biomedical Informatics and Statistics Center, School of Public Health, Fudan University, Shanghai, China

Abstract

Few-shot semantic segmentation (FSS) has been widely used in the field of information medicine and intelligent diagnosis. Due to the high cost of medical data collection and the privacy protection of patients, labeled medical images are difficult to obtain. Compared with other semantic segmentation dataset which can be automatically generated in a large scale, the medical image data tend to be continually generated. Most of the existing FSS techniques require abundant annotated semantic classes for pre-training and cannot deal with its dynamic nature of medical data stream. To deal with this issue, we propose a dynamic few-shot learning framework for medical semantic segmentation, which can fully utilize the features of newly-collected/generated data stream. We introduce a new pseudo-label generation strategy for continuously generating pseudo-labels and avoiding model collapse during self-training. Furthermore, an efficient consistency regularization strategy is proposed to fully utilize the limited data. The proposed framework is iteratively trained on three tasks: abdominal organ segmentation for CT and MRI, and cardiac segmentation for MRI. Experiments results demonstrate significant performance gain on medical data stream mining compared with the baseline method.

Keywords: Information medicine, Few-shot learning, Semantic segmentation, Data steam mining

1 Introduction

Medical image segmentation is envisioned as a promising technique for future information medicine, as it has great potential in computer-assisted diagnosis, improving diagnostic accuracy and efficiency. One of the critical targets of medical image segmentation is to isolate the necessary areas, such as the brain or the lung, from raw image data, and then remove unuseful regions, such as the air. The images tend to be multi-modal, such as computerized tomography (CT), magnetic resonance imaging (MRI) and other scanned images. These identified/extracted regions of interest can be further used for a particular research or diagnosis, providing a more precise analysis of anatomical data as well as helping clinicians make an accurate diagnosis.

Early medical image segmentation researches mainly focus on model-driven approaches, such as template matching techniques, edge detection and statistical shape models. An edge detection algorithm based on mathematical morphology was proposed in [1], which achieved an initial segmentation on lung CT images. Machine learning techniques, such as support vector machines (SVMs) and Markov random fields (MRF), were also applied in this field, as the research works in [2] that focused on brain MRI and [3] that segmented body data.

In recent years, the development of deep learning and convolutional neural networks (CNN) has drawn researchers' attention due to their strong ability to extract and represent image features. With the help of deep learning, medical image segmentation has made great progress and also becomes a hot topic in the field of computer vision. These methods can be categorized into supervised learning, semi-supervised learning, and unsupervised learning on the basis of the proportion of labeled data. Among these methods, supervised learning methods gain the most popularity. The widely-implemented framework includes fully convolution network (FCN) [4], U-Net [5], ResNet [6] combined with some useful techniques such as skip connection and model cascade. Although supervised learning achieves high accuracy, it has a huge demand for a large amount of labeled data and high-quality labels. However, in the field of medical image segmentation, it is usually difficult and expensive to build a large-scale dataset.

Few-shot learning (FSL), compared with fully supervised learning, is able to carry on the segmentation of incomplete or imperfect datasets. In the field of medical image segmentation, FSL, or one-shot learning is viewed as an important technique to compensate for the lack of large-scale high-quality labels [7]. Among these types of methods, self-training strategies with pseudo-labels have received lots of interest in FSL areas. Pseudo-label method leverage models trained by labeled data to predict pseudo-labels for unlabeled data, and then retrain the models on pseudo-labeled and labeled datasets. The common strategies such as data augmentation [8, 9] have also achieved a significant performance improvement. Nevertheless, these methods still require large amount of data for pre-training and generate high-quality pseudo-labels. Most importantly, these models are designed for pre-generated static datasets. When the labels are dynamically generated in the form of the data stream, these methods failed to leverage the features of newly-generated data. However, in the field of information medicine, both the new image and labels tend to be continuously generated. Taking the small number of high-quality labels into consideration, it's important for the models to dynamically adjust and learn the features from the data stream.

To deal with this issue, we propose a dynamic FSL framework for medical semantic segmentation, which can fully utilize the features of the newly-generated data stream. The medical image data is continuously collected from the medical information system, based on which the labels are dynamically generated. Then, we propose a few-shot self-training framework to utilize unlabeled data, increasing the generalization performance of the semantic segmentation model. The common training objective of the self-training method is the cross-entropy loss, which is calculated on generated pseudo-labels and softmax probabilities of unlabeled data. By iteratively generating pseudo-labels for the unlabeled sequences and retraining the network with pseudo-labels, the model can gradually learn the feature expressions of unlabeled data without additional annotation.

However, both the pseudo-label generation and the predictions of unlabeled data rely on the accuracy of softmax probability scores. When the scores are not accurate enough, they may lead to wrong pseudo-labels and wrong predictions. The negative feedback of both even brings about training collapse.

To overcome the limitations of conventional few-shot self-training methods, we decouple the pseudo-label generation and unlabeled data predictions to alleviate mutual interference. First, we introduce a novel pseudo-label generation strategy that utilizes the proxy-based distance instead of the softmax probability. We project the features into an embedding space to generate more accurate pseudo-labels as well as avoid negative feedback to the model. Second, to capture the temporal correlations in unlabeled sequences, we augment the unlabeled data by several different operations and then recombine them before prediction output. The proposed augmentation-and-recombination strategy not only improves the stability of the framework but also alleviates over-segmentation errors.[1] The major novelties of this article can be summarized as follows.

- We propose a dynamic FSL framework for medical semantic segmentation. The proposed framework is specially designed for medical data stream mining where medical images are continuously generated for model training and intelligent diagnosis.
- We propose a novel self-learning algorithm in order to eliminate the requirement for large-scale annotations. We devise a proxy-based pseudo-label generation strategy that generates pseudo-labels in the embedding and presents an efficient consistency regularization strategy, augmentation-and-recombination, to exploit pixel-wise correlations on unlabeled images.
- The proposed framework is iteratively trained and evaluated on three tasks: abdominal organ segmentation for CT, abdominal organ segmentation for MRI and cardiac segmentation for MRI. Experiments results demonstrate significant performance gain on the medical data stream for semantic segmentation compared with the baseline methods.

This paper is organized as follows: Sect. briefly introduces the background and related works. In Sect. , the proposed dynamic medical image stream mining framework is introduced and analyzed. Section further proposes the few-shot medical image segmentation algorithm based on self-training. Experiments and discussion are shown in Sect. . Section presents conclusions.

2 Related work

2.1 Medical semantic segmentation

With the rapid development of deep learning, semantic segmentation has made a breakthrough. Segmentation is an important processing step in natural image scene understanding and medical image analysis, image-guided intervention, radiotherapy or improved radiology diagnosis. A large number of depth learning methods have been introduced in the

[1] Note that semantic segmentation focus on the pixel-level classification that assigns a corresponding category to each pixel in an image. Compared to semantic segmentation, instance segmentation not only needs to achieve pixel-level classification but also needs to distinguish instances on the basis of specific categories. In this paper, we only focus on semantic segmentation.

literature, including X-ray, visible light imaging (such as color dermatoscopy images), magnetic resonance imaging (MRI), positron emission tomography (PET), computer tomography (CT) and ultrasound (such as echocardiography). However, due to the low contrast and insufficient semantic information of medical images, many methods are still difficult to achieve good segmentation results in medical images. In the field of medical image segmentation, algorithms can generally be divided into two categories: fully supervised learning methods [10–12] and semi-supervised learning methods [13–16]. The former makes the model form the memory of features by using the labeled data to train the model [17]. For semi-supervised medical image segmentation, FSS is an important category as it only requires a limited number of labeled samples [14–16]. However, the existing FSS methods are incapable of exploiting the data stream or treating small-scale labels.

2.2 Few-shot learning

Few-shot learning (FSL) is of great significance and challenge in the field of machine learning and deep learning. As humans, we can build knowledge of a new object from just one or more images and identify it in other images. In contrast, most advanced machine learning or deep learning algorithms rely on a large amount of data training to achieve better performance. However, due to a range of factors such as privacy, security, or the high labeling cost of data, many real-world scenarios (such as medicine, military, and finance) do not have the conditions to obtain enough tagged training samples. Therefore, the implementation of FSL becomes an important task in the field of machine learning or deep learning [18]. With the booming development of in-depth learning, especially the great success of CNN [6, 19] in visual tasks, many FSL researchers began to shift their attention from non-depth models to depth models. In 2015, by proposing a twin convolution network, [20] took the lead in incorporating deep learning into the solution of FSL problems by learning category-independent similarity measures in paired samples, which also means a new era for FSL. Subsequent FSL methods take full advantage of the deep neural networks in feature representation and end-to-end model optimization to solve FSL problems from different perspectives, including data enhancement [21], metric learning [22] and meta-learning [23, 24].

2.3 Data stream mining

The requirement of mining stream data becomes essential to obtain valuable knowledge. Data stream mining refers to mining the data in real-time by storing, processing, and extracting feasible knowledge that can help in decision-making and understanding some phenomenal events [25]. Many data analysis techniques can help with studying stream data, such as data clustering [26] and classification [27]. Compared with conventional data processing, these methods applied on data stream mining processes and updates data as it comes one by one based on their updating features [28–30].

3 Proposed dynamic medical image stream mining framework

Figure 1 demonstrates our proposed framework, aiming to dynamically exploit and explore the collected medical data stream to achieve semantic segmentation. The framework consists of three layers: the data layer, the model layer and the application layer, respectively.

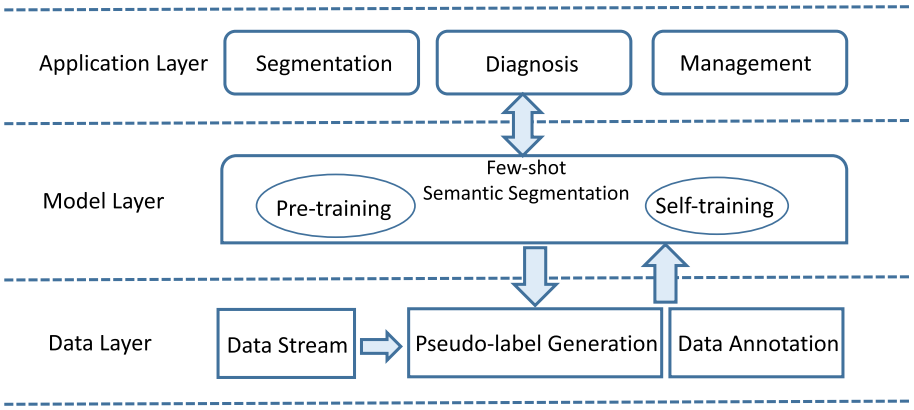

Fig. 1 Overview of the proposed dynamic medical image stream mining framework

The main function of the data layer is to collect raw data, store the data stream and integrate the corresponding labels. Note that the labels are from both model generation (pseudo labels) and manual annotation (real labels). Compared with other segmentation tasks where a big batch of labels can be automatically generated, the medical labels are gradually generated and keep dynamic in the framework. The data stream is continuous and multi-modal, thus requiring the data layer dynamically integrate the labels with the images.

In the model layer, semantic segmentation models are implemented based on FSL, where various techniques, such as pre-training and self-training, can be utilized. The model also outputs pseudo labels and transmits them to the data layer. The newly-annotated or newly-collected data can also be utilized by the model with a continuous training process. The segmentation results outputted by the model layer are then delivered to the application layer for various uses.

The application layer takes the segmentation results as the input, and then utilizes the results for corresponding applications designed for specific tasks, such as segmentation as a straightforward usage, intelligent diagnosis and medical case management. The different usage of the tasks raise different requests for the model, thus the model is also optimized and advanced according to the feedback of the application layer. In this paper, we focus on the task of semantic segmentation, which is often an initial task of various further tasks.

4 Proposed few-shot medical image segmentation algorithm based on self-training

4.1 Preliminaries

In the case of fully-supervised (FS) medical semantic segmentation, a batch of medical images $\mathbf{I} = \{I_t\}_{t=1}^{k}$ are fully annotated with annotated labels $\mathbf{Y} = \{Y_t\}_{t=1}^{k}$, where I_t and Y_t denotes the tth image and its mask, respectively. Training on the FS action segmentation models can be formulated as minimizing the following loss function:

$$\mathcal{L}_{FS} = \frac{1}{N}\sum_{t=1}^{k} H(c(f(I_t)), Y_t) \tag{1}$$

where N indicates the number of images. $f(\cdot)$ represents the feature extraction procedure. $H(\cdot)$ denotes the cross-entropy loss function, calculated by the ground truth y_t and

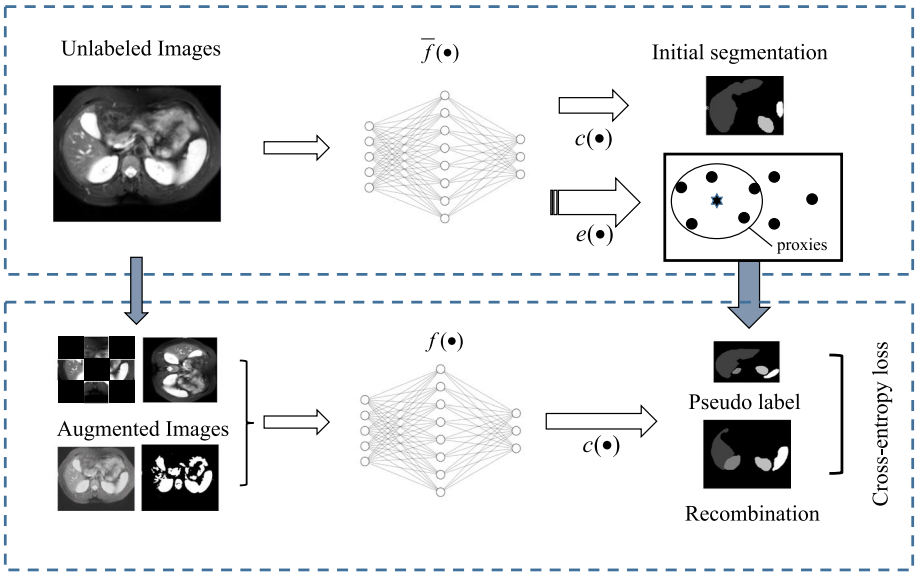

Fig. 2 Overview of the proposed self-training method for medical semantic segmentation. It mainly contains two parts. (1) Proxy-based pseudo-label generator (top half): we input the unlabeled frames into a momentum model, project the features into embedding space by $e(\cdot)$, and then generate pseudo-labels $\hat{Y}_U$ according to its distance to each proxy. (2) Augmentation-Recombination Module (bottom half): we split the unlabeled sequences into augmented images by different augmentation methods, such as cropped, rotate, and color jittering, and then input them to the model. Through the classifier $c(\cdot)$ and recombination, we obtain the prediction. Finally, we calculate the cross-entropy loss on $\hat{Y}_t$ and p_t. Note that the proxy

the predicted probabilities that are inferred by $c(\cdot)$. Note that the cross-entropy loss is calculated on pixel-wise on each image.

To exploit unlabeled data to improve generalization performance, we consider few-shot semantic segmentation with self-training. Accordingly, the loss function can be reformulated as follows:

$$\mathcal{L}_{SS} = \frac{1}{N_l} \sum_{I_t \in \mathbf{I}_l} H(c(f(I_l)), Y_l) + \frac{1}{N_u} \sum_{I_t \in \mathbf{I}_u} H(c(f(I_u)), \hat{Y}_u) \tag{2}$$

where $\mathbf{I}^l$ and $\mathbf{I}^u$ refer to the batch of the labeled and unlabeled data, respectively. $\hat{Y}_u$ indicates the pseudo-labels of the uth unlabeled image.

In this work, we develop a self-training framework to exploit the unlabeled data, as shown in Fig. 2. It includes two important parts: a proxy-based pseudo-label generator for generating pseudo-labels, and the other is an augmentation-and-recombination module (ARM) for learning unlabeled data. The former is used to generate higher-quality self-training pseudo-labels $\hat{Y}_u$ by projecting the output features of the segmentation network into the embedding space. The latter can be regarded as a novel data augmentation method for unlabeled data and a stable self-training framework. We introduce the two components in detail in the following sections.

4.2 Proxy-based pseudo-label generator

It is crucial to obtain the appropriately selected pseudo-labels to exploit the unlabeled data the self-training process. Typical self-training strategies suggest the unlabeled data with higher softmax probability scores according to a fixed confidence threshold:

$$\hat{Y}_t = \mathcal{G}(p_t), \tag{3}$$

where $p_t = c(f(I_t))$ is the softmax probability of I_t belonging to the mask, given by

$$p_t = \frac{\exp c(f(I_t))}{\exp \sum_{I_t \in \mathbf{I}} c(f(I_t))}, \tag{4}$$

where $\mathcal{G}(\cdot)$ denotes the operation of selecting the pixel with the highest score as the pseudo-label. However, the predicted p_t is not necessarily accurate, especially on the medical semantic segmentation task where the predicted accuracy is relatively low, which causes excessive noise in pseudo-labels. On the other hand, the wrong pseudo-labels will lead to a worse model since the model is optimized based on the cross-entropy loss $H(p_t, \hat{Y}_t)$, which in turn, leads to worse pseudo-labels. The high correlations between p_t and $\hat{Y}_t$ result in that small errors can be easily amplified. Such malignant negative feedback brings about instability in self-training, even making the training collapse.

Embedding space. To address the above issues, we try to decouple the inputs of cross-entropy loss, p_t and $\mathcal{G}(p_t)$, to alleviate the mutual interference. Specifically, we input the feature map $f(I_t)$ before the final prediction layer of the segmentation network into an additional projection branch $e(\cdot)$, and generate an embedding vector $e(f(I_t)))$ through two-layer convolution to project the feature into an embedding space, as illustrated in Fig. 2. This projection branch stops back-propagation during training, so as to avoid affecting the parameters of the original model. In the embedding space, each class will generate a corresponding proxy embedding, which is used to approximate the position of this class in the embedding space. Therefore, we propose to utilize a proxy-based distance score to replace the softmax probabilities score:

$$\hat{Y}_t = \mathcal{G}(\rho_t), \tag{5}$$

where ρ_t is the proxy-based distance score of the mask, calculated according to the distance between $e(f(I_t))$ and the proxy $\boldsymbol{\mu}_c$. This is based on the practical assumption that similar input data always have similar feature representations. In the embedding space of the mask, we calculate an anchor point as the proxy which can be used as a representative of all samples. If the distance between $e(f(I_t))$ and $\boldsymbol{\mu}_c$ is close, it's more likely that they belong to the same class. More concretely, we adopt cosine distance to measure the embedding distance:

$$d(\boldsymbol{e}_t, \boldsymbol{e}_k) = 1 - \frac{\boldsymbol{e}_t \cdot \boldsymbol{e}_k}{||\boldsymbol{e}_t|| \cdot ||\boldsymbol{e}_k||}. \tag{6}$$

To simplify the notation, we drop the full $e(f(I_t))$ notation and use $\boldsymbol{e}_t$ to denote the embedding of I_t.

Then, the proxy-based distance score can be calculated as:

$$\rho_t^{(c)} = \exp\left(-d(\boldsymbol{e}_t, \boldsymbol{\mu}_c)\right), \tag{7}$$

Proxy calculation. Instead of manually setting the proxies, we regard $\boldsymbol{\mu}_c$ as a learnable parameter. The network automatically updates it during the training process without our manual calculations. Motivated by [31], we try to make the proxy $\boldsymbol{\mu}_c$ more similar to the pixel belonging to the masks than other pixels, which is achieved by optimizing loss function $\mathcal{L}_{Pro}$:

$$\mathcal{L}_{Pro}(I_t, \boldsymbol{\mu}_c) = -\log\left(\frac{\exp\left(-d(\boldsymbol{e}_t, \boldsymbol{\mu}_c)/\tau\right)}{\sum_{\boldsymbol{\mu}_c'} \exp\left(-d(\boldsymbol{e}_t, \boldsymbol{\mu}_c')/\tau\right)}\right), \tag{8}$$

where $\boldsymbol{\mu}_c'$ denotes the proxies except $\boldsymbol{\mu}_c$, and τ denotes the scale factor to accelerate convergence. We set $\tau = 2$ in all experiments of this work. Our $\mathcal{L}_{Pro}$ is aimed to make $\boldsymbol{\mu}_c$ closer to $\boldsymbol{e}_t$ than any other proxies to $\boldsymbol{e}_t$. In the training process, we first train the mask proxies by the labeled data and then fine-tune them with a lower learning rate during the self-training process. Note that the proxies are dynamically updated with the newly-generated labeled data and pseudo-labels.

4.3 Augmentation-and-recombination module

In Sect. , we have discussed generating more stable pseudo-labels by using proxy-based distance score ρ to alleviate collapse caused by noise. In this section, we introduce consistency regularization into self-training. By adopting a novel data augmentation strategy on $c(f(I_u))$, we try to force the model to better exploit the temporal correlations in unlabeled data.

Consistency regularization is an efficient component of semi-supervised methods, which is based on the assumption that the model should output similar predictions even with perturbed versions of the same input. For semantic segmentation, the model should make similar predictions for the transformed images and the original images. In self-training, we can approximately regard the pseudo-labels as the prediction of original frames. Thus, the pseudo-labels should be similar with the predictions of transformed frames, which can be described by the cross-entropy function. Let I_t' be the transformed input. The cross-entropy loss for the unlabeled data can be formulated as:

$$\mathcal{L}'_u = \frac{1}{N_l} \sum_{I_t \in \mathbf{I}_u} H(c(f(I_t')), \hat{Y}_t), \tag{9}$$

where $\hat{Y}_t$ is the generated pseudo-labels for the original input. While Eq. (9) is similar to the pseudo-labeling loss in Eq. (2), it is crucially different in that the loss is computed on the model's output for a augmented image I_t'. By minimizing Eq. (9), we introduce a form of consistency regularization into the network.

However, the simple cropping operation on image-wise randomly drops an image clip, as shown in Fig. 3a. This kind of drop-based augmentation is regarded as strong data augmentation [32, 33], which forces the network to reduce the number of pixel correlations in input data and to restore the information loss according to context information, improving the ability to capture wide-range pixel correlations. Nevertheless, the basic assumption that this kind of information loss can be recovered is

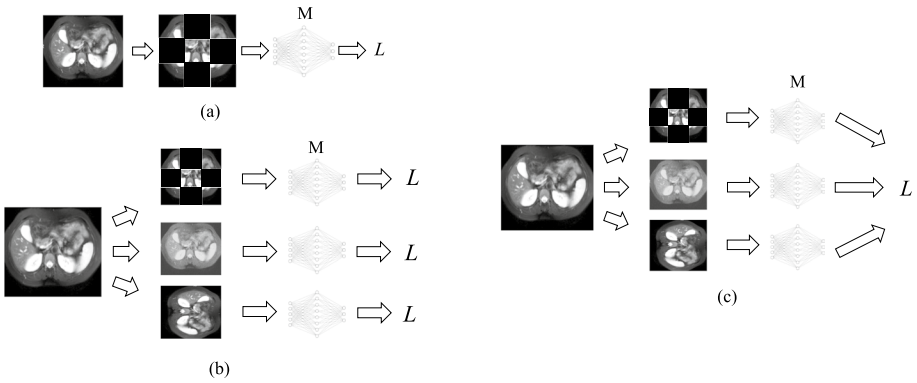

Fig. 3 Illustration of different implementations of data augmentation. **a** Single augmentation, taking cropping as an example; **b** multiple augmentation without recombination; **c** augmentation-and-recombination

that there is information correlation in its context. Otherwise, the information gap caused by strong enhancement will damage the performance. The simple cropping may cause overmuch losses on pixel correlations since the cropped part may contain several useful parts. To address this issue, a natural idea is that if we can provide the cropped parts for the input, then the problem of the information gap may be alleviated. Thus, we propose a data augmentation strategy as augmentation-and-recombination, which adopts a complementary structure where the input sequence is splitted into several symmetrical sub-images to ensure the integrity of the original information, as illustrated in Fig. 3c. We recombine the four symmetrical sub-images to make predictions of unlabeled videos. As shown in Fig 2, denoted as $x'(1)_t$, $x'(2)_t$, $x'(3)_t$ and $x'(4)_t$, respectively. Then, the output predictions is denoted by $p' = \frac{c(f(x'(1)_t))+c(f(x'(2)_t)+c(f(x'(3)_t)+c(f(x'(4)_t)}{4}$.

We discuss the implementation of data augmentation as illustrated in Fig. 3. Figure 3a shows only a single augmentation is implemented. In Fig. 3b, we adopt several different augmentation strategies while we calculate the cross-entropy loss on the pseudo-labels of the outputs separately. By contrast, strategies in Fig. 3c can not only improve the network's ability to recover the information gap but also make the augmented images complement each other to avoid information losses. Finally, the cross-entropy loss can be formulated as:

$$\mathcal{L}_{SPM} = \frac{1}{|\mathbf{I}_u|} \sum_{I_t \in \mathbf{I}_u} H\left(\frac{c(f(x'(1)_t)) + c(f(x'(2)_t) + c(f(x'(3)_t) + c(f(x'(4)_t)}{4}, \hat{Y}_t \right). \tag{10}$$

Algorithm 1 Our self-training framework

Input: unlabeled data $\mathbf{X}^U$, labeled data $\mathbf{X}^L$
Function: feature extraction $f(\cdot)$, classification $c(\cdot)$, embedding projection $e(\cdot)$, ZebraSplit $Z(\cdot)$

warm-up: using $\mathbf{X}^L$ to train the Model ($f(\cdot)$, $c(\cdot)$, $e(\cdot)$)
for x_u, x_l **in** zip($\mathbf{X}^U$, $\mathbf{X}^L$) **do**
  $\hat{y}_u = \mathcal{G}(\rho_u)$ # ρ_u *is calculated by Eq.7*
  $x'_u, x''_u = Z(x_u)$
  $\mathcal{L}_{unlabeled} = H(\frac{f(c(x'_u)), f(c(x''_u))}{2}, \hat{y}_u)$
  $\mathcal{L}_{labeled} = H(f(c(x_l)), y_l) + \mathcal{L}_{Pro}(x_l, \mu)$
  $loss = \mathcal{L}_{unlabeled} + \mathcal{L}_{labeled}$
  $loss.backward()$
  update params.
end for

5 Experiments and discussion

5.1 Datasets and evaluation indicators

We test the segmentation performance of our proposed method on three medical image segmentation datasets, which is the CT and MRI images for medical semantic segmentation task, respectively (Dataset-Abdomen-MRI, Dataset-Abdomen-CT and Card-MRI). The proposed method is used to segment medical images on three different datasets to verify the universal applicability of the proposed method in different situations.

Dataset-Abdomen-MRI is from ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge which includes 20 3D MRI scans for train task and 20 3D MRI scans for test task [34]. Dataset-Abdomen-CT is from MICCAI 2015 Multi-Atlas Abdomen Labeling challenge, with 20 clinical CT scans [35]. Card-MRI is from MICCAI 2019 Multi-sequence Cardiac MRI Segmentation Challenge (bSSFP fold) [36]. This dataset is divided into two parts, a training set containing 83 images of 3D abdominal scanning and the corresponding labels, and a test set containing 72 3D abdominal scanning images. It should be noted that this dataset stems from patients with clinical pathological characteristics.

5.2 Implementation details

In order to realize the universal use of the proposed method, it is necessary to convert the image into a unified format in the data preprocessing stage. All image data will be converted into 2D axial (Dataset-Abdomen-MRI and Dataset-Abdomen-CT) or 2D short-axis (Card-MRI) slices.[2] To improve the accuracy of the image segmentation algorithm, we use the gray transformation method in image enhancement to increase the dynamic range and contrast of the image, we process the image into 256×256 pixels. Additionally, it's also worth noting that gray transformation is widely used in medical image segmentation, as using grayscale can make it easier to identify fine details and subtle patterns in the image, which can be lost when using a full-color representation.

[2] The initial data format is dcm format. First, we convert it to nii format through shell command line to facilitate observation. All image data will be converted into 2D axial (Dataset-Abdomen- MRI and Dataset-Abdomen- CT) or 2D short-axis (Card-MRI). The axial refers to the cross-section perpendicular to the human body from foot to head.

For the datasets of two abdomens, in order to enable our model to distinguish between normal organs and organs with pathological characteristics, and to compare the differences between MRI images and CT images, we use these two datasets to jointly build a shared data set containing the left kidney, the right kidney, the spleen, and the liver. As for Card MRI, as it contains a scanning image of the heart, we can create a separate label set for this dataset, The label set includes three categories: left ventricular blood pool (LV-BP), left ventricular myocardium (LV-MYO), and right ventricle (RV).

To measure the overlapping between prediction and ground truth, we employ Dice score (0–100, 0: mismatch; 100: perfectly match), which is commonly used in medical image segmentation researches. To evaluate the quality of 3D volume image after segmentation into 2D, we follow the evaluation protocol established by [37].

In our experimental setup, the number of slices in each 3D scan is selected as 3, and for all slices to be segmented, the slices in the middle of adjacent slices are used as the reference for segmentation. It should be noted that the training set and set use data from different patients.

So as to test the generalization ability of the model to the test categories that have not been touched, in addition to the setting (setting 1) [37] of medical image segmentation with few samples for the experiment, we also use another setting(setting 2.) [15]. In setting 2, we implement the model segmentation test for the image of the training category by deleting the data of the same category in the test set and the training set.

In order to simulate the situation where the labeled clinical data are scarce in most cases in the real world, our experiments are conducted under the condition of one-way one-shot, that is, one category of data is used for training each time and one sample is used for each category.

The details of this experiment are described as follows. The network is implemented with PyTorch based on official PANet implementation [38]. In the feature map, we use fully-convolution to improve the resolution of features. It takes a $3 \times 256 \times 256$ image as input and produces a $256 \times 32 \times 32$ feature map. Pooling windows (L_H, L_W) for training and testing are set to 4×4 and 2×2 respectively.

5.3 Quantitative and qualitative results

Tables 4 and 6 show the comparison of image segmentation accuracy measured by Dice score between the algorithm we proposed and the existing algorithms, PANet [38] and SE-Net [37]. Both the two networks were efficient and widely used in the semantic segmentation tasks. PANet [38] enhances the entire feature hierarchy in lower layers by bottom-up path augmentation and adopts adaptive feature pooling. SE-Net [37] adaptively adjusts the feature responses of each channel by taking into account the interdependencies between channels. In the absence of manual annotation, the average Dice score of our proposed method always exceeds that of other methods. The results in Table 5 show that for unknown categories, our model can also achieve good segmentation and has strong generalization ability. This implies that the proposed superpixel-based self-supervised learning has successfully trained the network to learn more diverse and generalizable image representations from unlabeled images.

For qualitative comparison, we demonstrate the results as shown in Fig. 4. Our method has achieved good segmentation results on different types of data. We can observe that

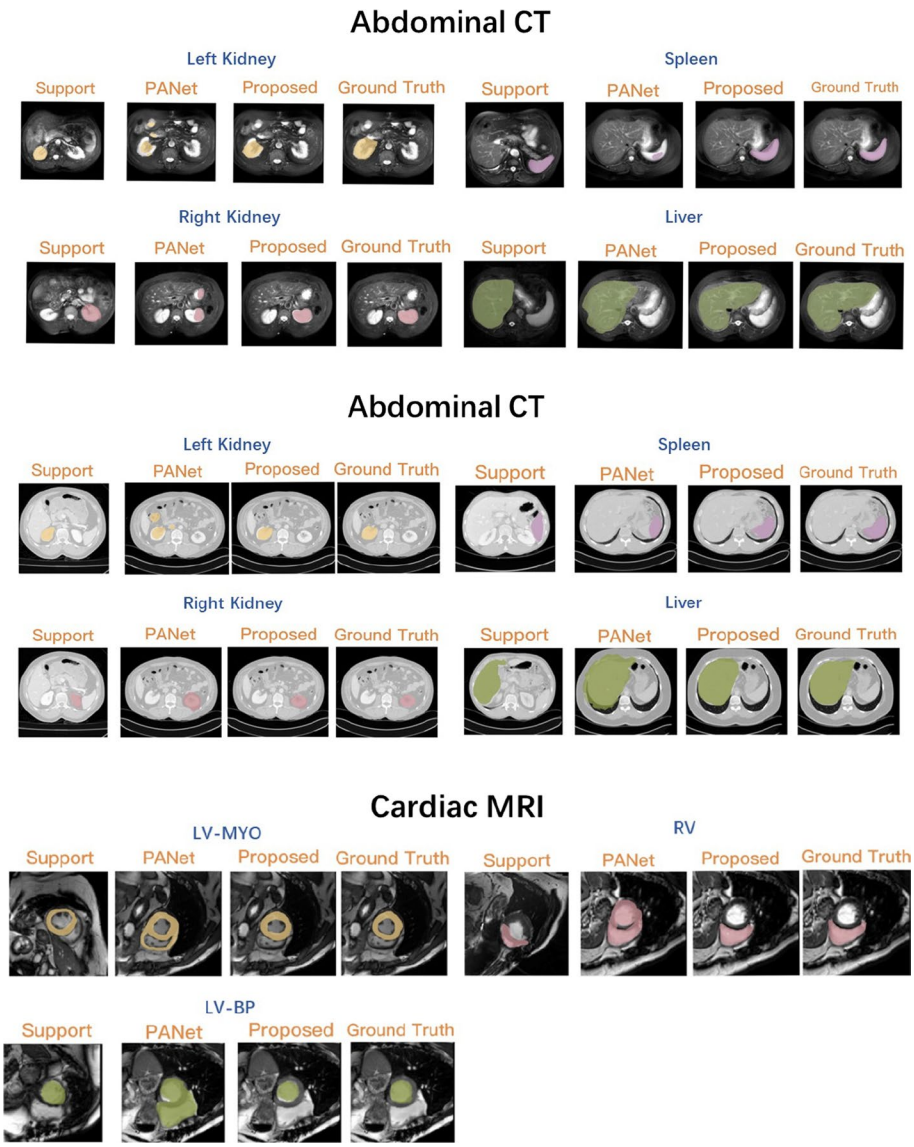

Fig. 4 The quantitative segmentation results of our proposed segmentation method on three different data sets. It can be seen that the segmentation effect achieved by our method is very close to the real situation. In order to demonstrate the good generalization ability of the proposed method, the segmentation of Abd CT and Abd MRI images in the figure is based on setting 2, that is, the training data does not contain the same category as the test set data

the segmentation effect of MRI images is better than that of the CT images because the part to be segmented in the MRI images has a more obvious contrast with the surrounding background, which is conducive to making the boundary clear so that the segmentation accuracy is higher.

5.4 Ablation study

To fully demonstrate the impact of each component of our model on the final performance of the model, we conducted ablation experiments on current MRI and CT images. As shown in Table 1, the self-training method can learn unlabeled data. To explore the

Table 1 Ablation study of the proposed method

Dataset	Method	Left kidney	Right kidney	Spleen	Liver	Mean
Abdomen-MRI	Self-training	71.33	76.52	66.02	71.63	71.38
	+PDS	72.02	76.78	66.54	72.25	71.90
	+ARM	72.98	76.53	68.15	72.46	72.53
	+PDS and ARM	74.57	78.74	68.01	73.84	73.79
Abdomen-CT	Self-training	59.68	51.24	58.23	70.23	59.85
	+PDS	61.22	51.96	60.37	71.88	61.36
	+ARM	61.68	55.49	60.13	70.94	62.06
	+PDS and ARM	64.59	55.57	61.11	74.56	63.95

PDS and ARM represent the proxy-based distance score and augmentation-and-recombination module, respectively

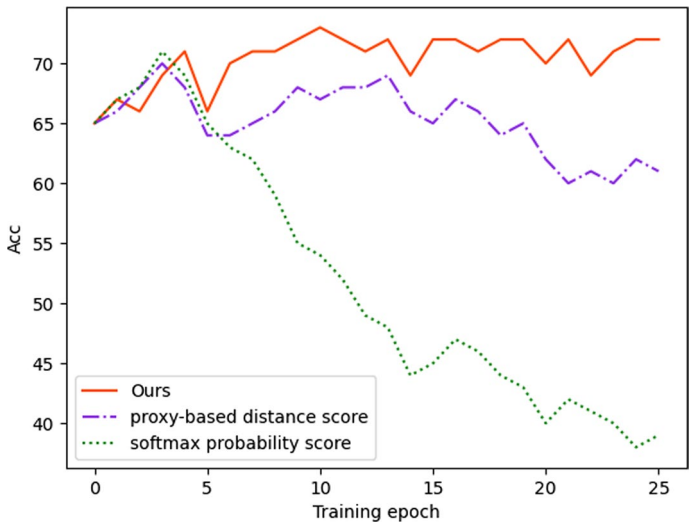

Fig. 5 The final results of various methods to deal with the negative feedback during self-training

impact of ARM, we applied this method to the basic self-training framework. In addition, our research shows that the method "PDS" and the method "ARM" have a common effect on improving the model performance. The results also show that the accuracy of using both methods achieves about 1% higher than that of the "ARM" method, and the effect of the basic self-training method is significantly improved.

Figure 5 shows the different results of different score indicators in dealing with negative feedback problems in the process of self-training. When using softmax as the confidence score, we can observe that the training accuracy rate crashes after a short increase. This is because the negative feedback of pseudo tags leads to the continuous deterioration of the model performance. In contrast, the proposed proxy-based distance score significantly alleviates the problem of the model crash and improves performance. However, there is still a downward trend in accuracy. Using our method, the segmentation accuracy of the model increases with the number of training epochs and converges to a stable value. Through the above experimental verification, our method solves the problem of model collapse caused by self-training negative feedback.

Table 2 compares the results of various data enhancement methods for the basic self-training model. The three types of data enhancement methods we compared have

Table 2 Influence of different data enhancement methods on model effect

Dataset	Method	Left kidney	Right kidney	Spleen	Liver	Mean
Abdomen-MRI	w/o augmentation	70.82	76.52	66.02	71.63	71.25
	Clip	70.99	77.15	67.22	71.96	71.83
	Augmentation Only	71.49	76.53	67.35	72.07	71.86
	ARM	72.98	76.53	68.15	72.46	72.53
Abdomen-CT	w/o augmentation	57.74	54.78	58.36	69.12	60.01
	Clip	58.27	54.03	58.69	70.05	60.26
	Augmentation only	61.64	55.81	58.95	70.66	61.77
	ARM	61.68	55.49	60.13	70.94	62.06

Table 3 Influence of different label scales on model effects

Dataset	Label proportion (%)	Mean
Abdomen-MRI	5	72.26
	10	72.53
	20	72.61
	50	74.28
Abdomen-CT	5	61.55
	10	62.06
	20	62.93
	50	63.42

improved compared with the original model. Interestingly, the accuracy of a single method of clipping the data image is not much better than that of the original model. We analyze that such methods lose some of the original data characteristics while enhancing data. This results in a small final performance improvement. Compared with the original model, the ARM algorithm has better performance in both MRI and CT datasets. The results in Table 2 show that our algorithm has a performance improvement of more than 1% compared with the original algorithm, which proves the superiority of our method in medical impact recognition tasks.

We further study the effect of the real label proportion on the model effect in the process of model self-training. Our basic experiment uses 10% of the true label ratio. For comparison, we also use 5%, 20%, and 50% of the true labels to observe the final results. The experiment shows that although the model effect will decrease with the decrease in the proportion of real tags, our method reduces these gaps to a certain extent. Table 3 shows that the accuracy rate obtained by the proportion of 10% and 20% is roughly the same, and there is only a decrease of about 1% compared with 50%. The results show that our method achieves good performance in the case of low proportion of real tags (Tables 4, 5, 6).

The effect on τ is given in Table 7 and the results are measured by Dice score. We can note that the Dice score achieves the highest value on average when τ equals to 2. It can also be noted that there is an improvement when increasing τ from 1 to 2 and that the Dice score decreases when increase τ to 2.

Table 4 Experiment results based on abdominal scans under setting 1

Method	Dataset-abdomen—MRI					Dataset-abdomen—CT				
	Left kidney	Right kidney	Spleen	Liver	Mean	Left kidney	Right kidney	Spleen	Liver	Mean
SE-Net [37]	46.48	47.23	47.79	28.82	42.58	23.51	14.40	42.59	35.34	28.96
Vanilla PANet [38]	31.11	32.78	41.74	50.40	39.01	20.56	20.91	36.30	49.55	31.83
ALPNet [39–41]	44.73	48.66	49.81	63.14	51.58	29.12	30.23	41.01	64.70	41.26
DARCNN [13]	57.54	60.41	61.32	72.15	62.85	56.37	51.24	55.72	61.68	56.25
MRNet [42]	61.28	63.83	63.99	72.46	65.39	65.59	68.33	69.14	70.01	68.26
Ours	82.61	85.68	72.26	77.48	79.50	73.96	72.18	70.94	78.68	73.94

Table 5 Experiment results on abdominal scans under setting 2

Method	Dataset-abdomen—MRI					Dataset-abdomen—CT				
	Lower		Upper		Mean	Lower		Upper		Mean
	Left kidney	Right kidney	Spleen	Liver		Left kidney	Right kidney	Spleen	Liver	
SE-Net [37]	60.19	59.98	50.12	26.43	49.18	33.05	15.43	0.24	0.30	12.26
Vanilla PANet [38]	54.33	37.61	49.81	44.15	46.48	32.86	17.23	29.98	39.17	29.81
ALPNet-init	18.25	13.31	23.77	36.63	22.99	13.77	10.85	16.49	41.52	82.63
ALPNet [39]	54.08	58.74	51.81	36.22	50.12	35.12	30.41	28.25	47.81	35.39
DARCNN [13]	44.72	44.83	58.27	64.75	53.14	37.64	35.89	44.78	61.10	44.85
MRNet [42]	52.21	56.55	57.45	69.27	58.87	38.67	35.71	50.03	46.22	42.65
Ours	74.57	78.74	68.01	73.84	73.79	64.59	55.57	61.11	74.56	63.95

Table 6 Experiment results based on cardiac scans under setting1

Method	LV-BP	LV-MYO	RV	Mean
SE-Net [37]	57.64	24.81	13.68	32.04
Vanilla PANet [38]	52.33	35.27	36.21	41.27
ALPNet [39]	72.95	48.35	56.05	59.12
DARCNN [13]	69.90	45.97	68.22	61.36
MRNet [42]	70.81	53.14	69.30	64.41
Ours	84.66	67.05	80.13	77.28

Table 7 Effect on τ (Dice score)

Dataset	τ	Left kidney	Right kidney	Spleen	Liver	Mean
Abdomen-CT	1.0	63.68	64.11	52.29	60.56	62.45
	1.5	63.96	53.45	60.47	74.21	62.47
	2.0	64.59	55.57	61.11	74.56	63.95
	2.5	64.98	52.98	59.21	73.98	61.34

6 Conclusion

In this paper, we proposed a dynamic FSL framework for medical semantic segmentation aiming at image stream mining. Compared with the conventional static frameworks, the proposed framework can learn from the continuously generated data streams. To compensate for the lack of large-scale high-quality labels, we proposed a proxy-based pseudo-label generation strategy, which proved to be effective in learning feature representation and avoiding model collapse. We further integrated an augmentation-and-recombination strategy to improve consistency regularization. The experiments on three widely used medical semantic segmentation datasets demonstrated significant performance gain compared with the baselines.

Abbreviations

FSS	Few-shot semantic segmentation
CT	Computerized tomography
MRI	Magnetic resonance imaging
SVM	Support vector machine
MRF	Markov random fields
CNN	Convolutional neural network
FCN	Fully convolution network
FSL	Few-shot learning
ARM	Augmentation-and-recombination module

Acknowledgements
Not applicable.

Author contributions
ZY conceived of the presented idea and wrote the manuscript. WZ developed the scheme and checked the manuscript. Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
We tested the performance of the proposed method based on three public medical image segmentation datasets accessible from websites https://www.synapse.org/#!Synapse:syn3193805/wiki/217789 and https://chaos.grand-challenge.org/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 22 November 2022 Accepted: 1 March 2023
Published online: 01 May 2023

References

1. Z. Yu-Qian, G. Wei-Hua, C. Zhen-Cheng, T. Jing-Tian, L. Ling-Yun, Medical images edge detection based on mathematical morphology, in 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 6492–6495 (2006). IEEE
2. K. Held, E.R. Kops, B.J. Krause, W.M. Wells, R. Kikinis, H.-W. Muller-Gartner, Markov random field segmentation of brain MR images. IEEE Trans. Med. Imaging **16**(6), 878–886 (1997)
3. S. Li, T. Fevens, A. Krzyżak, A SVM-based framework for autonomous volumetric medical image segmentation using hierarchical and coupled level sets, in *International Congress Series*, vol. 1268 (Elsevier, 2004), pp. 207–212.
4. 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
5. 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, 2015), pp. 234–241
6. 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
7. Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (csur) **53**(3), 1–34 (2020)
8. Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, Q.V. Le, Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)
9. D. Berthelot, N. Carlini, E.D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, C. Raffel, Remixmatch: semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 (2019)
10. W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, Y. Zheng, Learning calibrated medical image segmentation via multi-rater agreement modeling, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 12341–12351
11. S. Reiß, C. Seibold, A. Freytag, E. Rodner, R. Stiefelhagen, Every annotation counts: multi-label deep supervision for medical image segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 9532–9542
12. Q. Liu, C. Chen, J. Qin, Q. Dou, P.-A. Heng, Feddg: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1013–1023
13. J. Hsu, W. Chiu, S. Yeung, Darcnn: domain adaptive region-based convolutional neural network for unsupervised instance segmentation in biomedical images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1003–1012
14. Y. Shi, P. Sheng, J-net: asymmetric encoder–decoder for medical semantic segmentation. Secur. Commun. Netw. **2021**, 1–8 (2021)
15. C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, D. Rueckert, Self-supervision with superpixels: training few-shot medical image segmentation without annotation, in *European Conference on Computer Vision* (Springer, 2020), pp. 762–780
16. L. Sun, C. Li, X. Ding, Y. Huang, Z. Chen, G. Wang, Y. Yu, J. Paisley, Few-shot medical image segmentation using a global correlation network with discriminative embedding. Comput. Biol. Med. **140**, 105067 (2022)
17. M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges. J. Digit. Imaging **32**(4), 582–596 (2019)
18. J. Lu, P. Gong, J. Ye, C. Zhang, Learning from very few samples: a survey. arXiv preprint arXiv:2009.02653 (2020)
19. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
20. G. Koch, R. Zemel, R. Salakhutdinov, et al. Siamese neural networks for one-shot image recognition, in *ICML Deep Learning Workshop*, vol. 2 (Lille, 2015)
21. M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**(398), 528–540 (1987)
22. E. Xing, M. Jordan, S.J. Russell, A. Ng, Distance metric learning with application to clustering with side-information. Adv. Neural Inf. Process. Syst. **15**, 505–512 (2002)
23. R. Vilalta, Y. Drissi, A perspective view and survey of meta-learning. Artif. Intell. Rev. **18**(2), 77–95 (2002)
24. O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, in *Advances in Neural Information Processing Systems*, vol. 29, ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Curran Associates, Inc., 2016)
25. A.H.A. Rahnama, Distributed real-time sentiment analysis for big data social streams, in *2014 International Conference on Control, Decision and Information Technologies (CoDIT)* (IEEE, 2014), pp. 789–794

26. A.K. Jain, Data clustering: 50 years beyond k-means. Pattern Recognit. Lett. **31**(8), 651–666 (2010)
27. C.C. Aggarwal, Data classification, in *Data Mining* (Springer, 2015), pp. 285–344
28. C.C. Aggarwal, J. Han, J. Wang, P.S. Yu, A framework for projected clustering of high dimensional data streams, in *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, vol. 30 (2004), pp. 852–863
29. C.C. Aggarwal, S.Y. Philip, J. Han, J. Wang, A framework for clustering evolving data streams, in *Proceedings 2003 VLDB Conference* (Elsevier, 2003), pp. 81–92
30. C.C. Aggarwal, J. Han, J. Wang, P.S. Yu, On demand classification of data streams, in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004), pp. 503–508
31. Y. Movshovitz-Attias, A. Toshev, T.K. Leung, S. Ioffe, S. Singh, No fuss distance metric learning using proxies, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 360–368
32. Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34 (2020), pp. 13001–13008
33. T. DeVries, G.W. Taylor, Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
34. A.E. Kavur, N.S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D.D. Pham, S. Chatterjee, P. Ernst, S. Özkan et al., Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Med. Image Anal. **69**, 101950 (2021)
35. B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, A. Klein, Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge (2015). https://doi.org/10.7303/syn3193805
36. X. Zhuang, Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 2933–2946 (2018)
37. A.G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, C. Wachinger, 'squeeze & excite' guided few-shot segmentation of volumetric images. Med. Image Anal. **59**, 101587 (2020)
38. 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
39. X. Ren, J. Malik, Learning a classification model for segmentation, in *IEEE International Conference On Computer Vision*, vol. 2 (IEEE Computer Society, 2003), p. 10
40. D. Stutz, A. Hermans, B. Leibe, Superpixels: an evaluation of the state-of-the-art. Comput. Vis. Image Underst. **166**, 1–27 (2018)
41. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
42. P.C. Roth, D.C. Arnold, B.P. Miller, Mrnet: A software-based multicast/reduction network for scalable tools, in *SC'03: Proceedings of the 2003 ACM/IEEE Conference on Supercomputing* (IEEE, 2003), pp. 21–21

Publisher's Note