Recurrent Mask Refinement for Few-Shot Medical Image Segmentation

Hao Tang Xingwei Liu Shanlin Sun Xiangyi Yan Xiaohui Xie
Department of Computer Science
University of California, Irvine, California, 92697
{htang6, xingweil, shanlins, xiangyy4, xhx}@uci.edu

Abstract

Although having achieved great success in medical image segmentation, deep convolutional neural networks usually require a large dataset with manual annotations for training and are difficult to generalize to unseen classes. Few-shot learning has the potential to address these challenges by learning new classes from only a few labeled examples. In this work, we propose a new framework for few-shot medical image segmentation based on prototypical networks. Our innovation lies in the design of two key modules: 1) a context relation encoder (CRE) that uses correlation to capture local relation features between foreground and background regions; and 2) a recurrent mask refinement module that repeatedly uses the CRE and a prototypical network to recapture the change of context relationship and refine the segmentation mask iteratively. Experiments on two abdomen CT datasets and an abdomen MRI dataset show the proposed method obtains substantial improvement over the state-of-the-art methods by an average of 16.32%, 8.45% and 6.24% in terms of DSC, respectively. Code is publicly available [1].

1. Introduction

Medical image segmentation is a fundamental task in medical image analysis. It is used in many clinical applications, including disease diagnosis, treatment planning and treatment delivery. Segmentation of anatomical structures or lesions is usually done manually by experienced doctors, which is often tedious and labor-intensive. With the recent use of deep convolutional neural networks, automated segmentation tools using computer programs can achieve near human accuracy on multiple tasks with very short processing time. However, in order to achieve good performance, these systems are usually trained in a fully supervised fashion with large amounts of annotated data. Acquiring a dataset with abundant manual labels is often very expensive and time-consuming as it requires experts with many years' clinical experience. Moreover, the differences in image acquisition protocols among different medical equipment and institutes pose great challenges to the generalization ability of the learning based systems.

Few-shot learning has been proposed as one of the potential solutions to addressing these challenges in the low data regime [43, 46, 56, 8, 22]. The main few-shot image segmentation approach forms the problem as meta learning [9, 10, 16] and uses supervised learning to train few-shot learning models. A few-shot learning model is trained to extract class-specific features from the set of support images with annotations, and then perform segmentation on the query images by using distilled knowledge from the support images. During test time, by extracting features from a set of new support images (unseen classes), the model is able to segment novel classes. Many few-shot learning methods have been proposed and achieved great performance on natural image segmentation tasks [33, 39, 6, 41, 59, 67, 66, 62, 17]. However, applying few-shot learning models for medical image segmentation is still in early stages [31, 36].

Few-shot segmentation in medical images is different than that in natural images. Many approaches are based on prototypical networks [43], and often apply masked average pooling [6, 59, 67] to extract class prototypes from feature maps within the foreground mask. This step usually assumes the masked region contains sufficient features to distinguish different classes, especially foreground and background. However, this may not always be true in medical images. Distinct local appearances and context information are more critical in determining the boundary for foreground and background. A clear boundary to separate regions of interest from the background is of critical importance in medical image segmentation. Moreover, the background is usually large and spatially inhomogeneous while the foreground is small and homogeneous [30], and there exists the abundance of tissues that share very similar appearance to each other, all of which add ambiguity to define the foreground and background regions. To address this is-

[1] https://github.com/uci-cbcl/RP-Net

sue, we encourage the network to explicitly model the context relationship between foreground and background pixels, especially pixels around the boundary.

In this work, we introduce a new network framework for few shot medical image segmentation using prototypical network (RP-Net: **R**ecurrent **P**rototypical **Net**works). First, we propose a context relation encoder (CRE) on top of the extracted features, to explicitly model the relation between foreground and background feature maps. The relationships between foreground and background regions are more important in defining the boundary of the regions of interest in medical image segmentation. To force the model to distill and utilize the local context relation information, CRE uses correlation to capture the differences in the foreground and background regions. Pixel features are augmented with the context relation features. The explicit extraction of the context relationship poses a strong constraint to the features the model would learn and forces it to focus on the boundary of the region of interest. A prototypical network is followed to produce predicted masks using these augmented features.

Second, we propose a recurrent mask refinement module that iteratively refines the segmentation using CRE and prototypical networks. This design draws inspiration from recent works [53, 32, 18] that employ iterative refinement. More importantly, the prediction mask modifies the mask in the previous step, which results in updated local context relationship. The recurrent module serves the purpose to recapture the updated context relationship and recompute its context relationship based on new prediction. Starting from the segmentation mask from the previous step, the model uses the refined prediction mask in the previous step to compute new context features using CRE, and then feeds it to the same prototypical network. The weights of the module are shared among multiple iterations so it is fully recurrent. This recurrent module facilitates the learning and forces the model to learn to gradually refine the segmentation.

Our contributions are summarized as:

- A context relation encoder (CRE) that uses correlation between foreground and background to enhance context relationship features around the object boundary.

- A new framework for few-shot medical image segmentation that iteratively refines the prediction mask through a recurrent module that uses CRE and prototypical networks.

- We conducted experiments on two abdomen CT datasets and one abdomen MRI dataset. Experiments show that the proposed framework outperforms the SOTA few-shot framework for medical image segmentation by an average of 16.32% on ABD-110 dataset [49], 8.45% on MICCAI15 Multi-Atlas Abdomen Labeling challenge dataset [23] and 6.24% on ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge [21] in terms of DSC.

2. Related work

2.1. Medical image segmentation

In recent years, deep learning has brought significant progress to the field of medical image analysis [40], such as computer-aided diagnosis [38, 48, 50, 52], image registration [2, 1, 14], reconstruction [64, 7, 63], and etc. In terms of medical image segmentation, the development of the deep convolutional neural networks has lead to various successful applications, including segmentation of tissue [44, 58, 28], anatomical structures [47, 3, 55, 70, 11, 5, 45, 4, 25, 51] and lesions [12, 69, 57, 24, 37, 61]. One of the most famous and widely used network architecture is U-Net [34]. U-Net uses lateral connection to fuse features from encoders and decoders. Many its variants were proposed, with different focus on their designs. V-Net [26] extends the use of U-Net to 3D volume data. Attention U-Net [29] proposes to use gated mechanism to filter features. nnUNet [19] combines different U-Net like network architectures and automatically configure the optimal setting for different tasks, which is the best out of box U-Net. These SOTA methods require abundant manual annotations for their specific tasks to achieve good performance. They are designed to fully utilize the power of annotated dataset, and is limited when segmenting novel classes.

2.2. Few-shot learning

Few-shot learning can be categorized into three main focuses: data, model and algorithm [60]. One main stream of few-shot segmentation in natural image that focuses on the model is prototypical networks [43]. Prototypical network uses the idea of meta learning [9, 10, 16] and applies averaged mask pooling to pool class-specific features from the support set, which is called prototypes. Then, segmentation for the query image is done by computing the cosine distance with each class prototype . PANet [59] further improves upon this idea by proposing a prototype alignment network to better utilize the support set, by also predicting on support images using query images as support set.

In few-shot medical image segmentation, most works focus on generating new training data to enlarge the training set given only a few labels [68, 27, 31, 65]. However, this still requires retraining the model when a new class needs to be segmented. More recently, a few works focus on designing network architecture that does not require retraining the model. Squeeze and excite [36] first proposes a few-shot learning architecture specifically designed for medical image segmentation. They propose to use squeeze and excite modules to fuse information from support image on to query image to guide the segmentation arm. [30] proposes local prototypes to enrich the representation of class prototypes and a self-supervised training strategy using super pixels. Likewise, we focus on few-shot medical image segmenta-

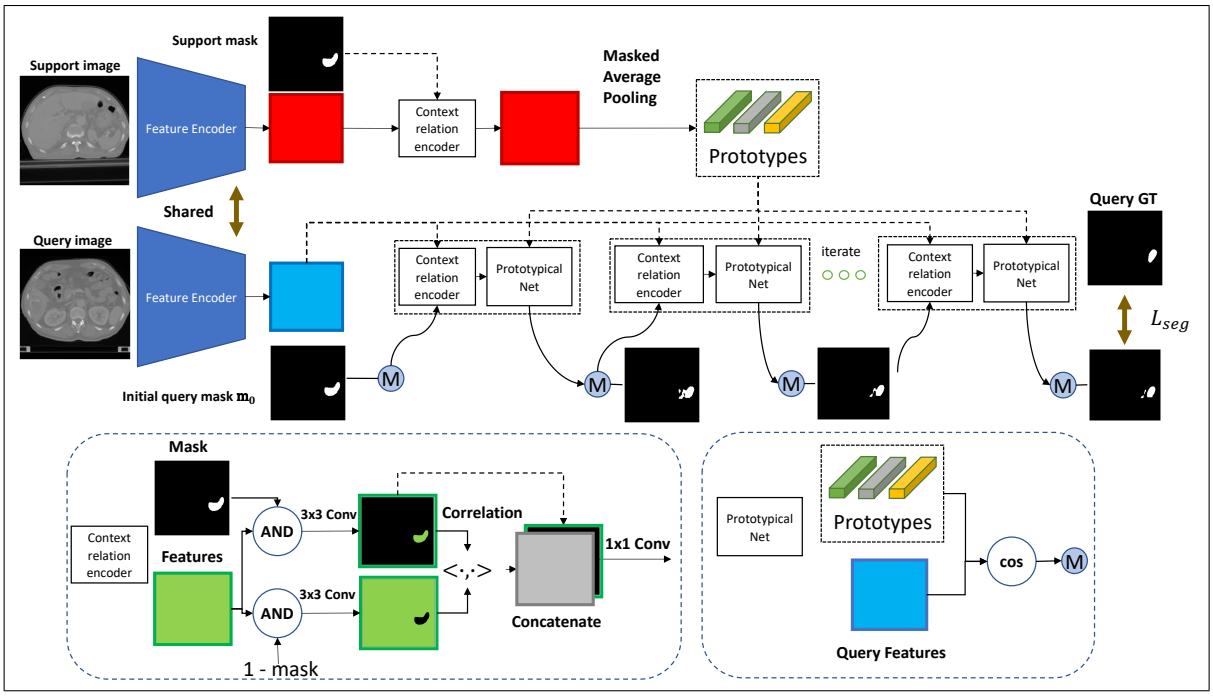

Figure 1. RP-Net consists of three main components: (1) A feature encoder that extracts features from both support and query images; (2) A context relation encoder (CRE) that use correlation to enhance the local context relationship features; (3) A recurrent mask refinement module that iteratively uses CRE and a prototypical network to recaptures the change of local context features and refines the mask.

tion without retraining the model, and we propose a new framework that uses CRE and recurrent mask refinement module to better capture local feature and shape differences around foreground object boundary.

3. Method

We first describe the formal definition of few-shot medical image segmentation. Next, we introduce the architecture of RP-Net, especially the context relation encoder (CRE) and recurrent mask refinement module.

3.1. Problem definition

In few-shot medical image segmentation task, the model is trained using images and a set of semantic labels C_{tr} drawn from a training dataset D_{tr}. During inference, the model segments a new set of semantic classes C_{te} from test images D_{te}, given a few labeled examples of C_{te}. Note that $C_{tr} \cap C_{te} = \emptyset$. For example, the model is trained using semantic labels $C_{tr} = \{\text{liver, left and right kidney}\}$ and during testing time the model needs to segment new semantic classes $C_{te} = \{\text{spleen}\}$. Let N be the number of semantic classes in C_{te}, and K be the number of examples for each semantic class in C_{te}. The few-shot learning problem is also referred to as N-way K-shot learning. In medical image segmentation, most works usually consider 1-way 1-shot learning [36, 30].

To achieve the goal of segmenting unseen classes in inference time, an episodic training strategy is used widely [59, 30, 36]. To simulate the situation in testing time where only K examples for each class are provided, the episodic training schema randomly draws each training example in the form of a support and query data pair $[(\mathbf{x}_s, \mathbf{y}_s), (\mathbf{x}_q, \mathbf{y}_q)]$ from D_{tr}. The model is trained to distill knowledge about a semantic class from the support set $(\mathbf{x}_s, \mathbf{y}_s)$ and then apply this knowledge to segment query set $\mathbf{x}_q$. In inference time, only the K support images $\mathbf{x}_s$ and their corresponding labels $\mathbf{y}_s$ are given, and the model performs segmentation on query images $\mathbf{x}_q$.

3.2. Proposed method

We now introduce RP-Net for few-shot learning in medical images. For the rest of this section, we consider a 1-way K-shot learning problem. The architecture of RP-Net is shown in Figure 1. Our approach consists of three steps: 1) extracting image features, 2) enhancing context relation features using CRE, 3) iteratively applying CRE and prototypical network to refine the segmentation mask. All stages are differentiable and can be trained end-to-end.

3.2.1 Feature extraction

The input to the network is a set of K support images $\mathbf{x}_s \in \mathbb{R}^{H \times W \times 1}$ and a query image $\mathbf{x}_q \in \mathbb{R}^{H \times W \times 1}$, padded to the

same height H and width W. The support and query images are first aligned globally using affine transformation, which is a common step in many medical image tasks.

The model first uses the same feature encoder f_θ to extract support features $\mathbf{F}_s \in \mathbb{R}^{H' \times W' \times Z}$ and query features $\mathbf{F}_q \in \mathbb{R}^{H' \times W' \times Z}$ respectively. H′ and W′ are the height and width of the feature map, and Z is the number of feature channels. An adapted version of the U-Net backbone was used as the feature encoder f_θ. Instead of upsampling the feature maps to the original resolution as implemented in the original U-Net, we remove the last two upsampling blocks in the U-Net to save GPU memory and computation. This results in the resolution of the support and query features being 1/4 of the image resolution ($H' = H/4, W' = W/4$).

3.2.2 Context relation encoder (CRE)

In medical image segmentation, the local context features are important to determine the boundary of foreground and background. To strengthen and emphasize these features, we propose the context relation encoder to enhance context features and force the model to focus on the shape and context of the region of interest rather than pixels themselves.

CRE takes the extracted features $\mathbf{F}$ (we drop subscript q and s for convenience) and foreground mask $\mathbf{m}$ as input and outputs augmented features $\mathbf{F}_{cre} = f_{cre}(\mathbf{F}, \mathbf{m}) \in \mathbb{R}^{H' \times W' \times Z}$. $\mathbf{m}$ is the mask of the foreground class from the support image ($\mathbf{y}_s$), or the proposed foreground mask of a query image. Features of foreground and background are first extracted by masking $\mathbf{F}$ using the mask $\mathbf{m}$: $\mathbf{F}_f = \phi_f(\mathbf{F} \odot \mathbf{m})$ and $\mathbf{F}_b = \phi_b(\mathbf{F} \odot (1 - \mathbf{m}))$. ϕ_f and ϕ_b denote 3×3 convolution. Next, a correlation computation is applied to acquire the context relation features between foreground and background feature vectors at each spatial location (x, y) of $\mathbf{F}_b$ and $(x - i, x - j)$ of $\mathbf{F}_f$ with offset i and j:

$$\mathbf{C}^{(x,y,i,j)} = \sum_z \mathbf{F}_f^{(x,y,z)} \mathbf{F}_b^{(x-i,x-j,z)} \tag{1}$$

Instead of computing correlation between every pair of pixels on $\mathbf{F}_f$ and $\mathbf{F}_b$, we limit the maximum displacement d for comparison at each location (x, y). Given a maximum displacement d, we only compute correlation $C^{(x,y,i,j)}$ in a neighborhood of size $2d + 1$ by limiting the range of (i, j). As a result, the context relation feature $\mathbf{C}$ is of size $H' \times W' \times (2d + 1)^2$. $\mathbf{C}^{(x,y)}$ effectively captures information of how a background pixel is related to foreground when it is close to the object boundary. Finally, we concatenate $\mathbf{C}$ and $\mathbf{F}_f$ along channel dimension and apply a 1×1 convolution to fuse foreground features and context relation features to obtain $\mathbf{F}_{cre}$. d is set to 5 based on empirical results (see Table 2 for details).

Compared to directly computing correlation between feature maps, separating feature map into foreground and background features is important. Correlation calculated this way is sparse and has only non-zero values around the boundary, which captures the shape of the object and clearly differentiate a pixel from the background. Correlation calculated between full feature maps is not able to achieve this because it does not have the sense of boundary of the region.

3.2.3 Prototypical networks

Following [30, 59], we use a relative simple method for calculating the prototypes, averaging feature vectors within the mask and across support images. Given the enhanced image features of support set $\mathbf{F}_{cre,s}$, we first compute the prototype of class c via masked average pooling:

$$\mathbf{p}_c = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{x,y} \mathbf{F}_{cre,s}^{(k,x,y)} \mathbf{y}_s^{(k,x,y,c)}}{\sum_{x,y} \mathbf{y}_s^{(k,x,y,c)}} \tag{2}$$

where (x, y) is the index of pixels on the feature map, (x, y, c) indexes the spatial locations of the binary mask of class c and K is the number of support images.

Segmentation is done using a non-parametric metric learning method. Prototypical network calculates the distance between the query feature vector and the computed prototypes $P = \{\mathbf{p}_c | c \in C\}$. A softmax over the distances is applied to produce a probabilistic output over all classes. Formally, for each pixel at location (x, y) of query feature map $\mathbf{F}_{cre,q}$, we have:

$$\mathbf{m}_{soft} = \cos(\mathbf{F}_{cre,q}, P), \text{and}$$
$$\cos(\mathbf{F}_{cre,q}, P)^{(x,y,c)} = \frac{\exp(-\alpha d(\mathbf{F}_{cre,q}^{(x,y)}, \mathbf{p}_c))}{\sum_{\mathbf{p}_j \in P} \exp(-\alpha d(\mathbf{F}_{cre,q}^{(x,y)}, \mathbf{p}_j))} \tag{3}$$

where the distance function d is a commonly used cosine distance and α is a scaling factor for this distance function to work best with the softmax function. α is set to 20 [59]. The class prediction can be obtained by:

$$\mathbf{m}^{(x,y)} = \arg\max_c \mathbf{m}_{soft}^{(x,y,c)} \tag{4}$$

3.2.4 Recurrent mask refinement

Since the mask $\mathbf{m}$ used to compute context relation features would change every time the network makes a prediction, we propose a recurrent mask refinement module to recapture this change and compute new context relation features based on the previous prediction.

The recurrent mask refinement module estimates a sequence of mask predictions $\{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_n\}$ from an initial mask which is the union of all support masks: $\mathbf{m}_0 =$

$\bigcup_{i=1}^{K} \mathbf{y}_s^i$. At each iteration t, it produces a new segmentation mask $\mathbf{m}_t$ based on $\mathbf{m}_{t-1}$. The design of this architecture mimics the steps of an optimization algorithm. For this purpose, all the weights in the recurrent module are shared across multiple iterations. The model is trained to learn to modify the mask gradually so that the final output mask $\mathbf{m}_n$ converges to an optimum solution. Note that, in this work the $\mathbf{m}_0$ is initialized using the average of support masks since images are affine aligned, but it is also possible to better initialize $\mathbf{m}_0$ using other methods.

This recurrent mask refinement module takes in support features $\mathbf{F}_s$, query features $\mathbf{F}_q$ and the mask $\mathbf{m}_{t-1}$ in previous step, uses CRE to enhance query features, and applies prototypical network to output a segmentation mask $\mathbf{m}_t$.

$$\mathbf{m}_{soft,t} = \cos(f_{cre}(\mathbf{F}_q, \mathbf{m}_{soft,t-1}), P) \tag{5}$$

We apply 4 iterations of the recurrent mask refinement module during training to save memory and computation cost. In inference time, we apply 10 iterations. We show in Figure 2 the performance at each iteration during inference time and 10 iterations are sufficient to obtain a stable result. The final prediction is obtained by upsampling $\mathbf{m}_n$ to the same resolution of the $\mathbf{x}_q$ using bilinear interpolation.

3.2.5 Loss function

We supervise our network using dice loss and cross entropy between the final predicted mask $\mathbf{m}_{soft,n}$ and ground truth segmentation mask $\mathbf{y}_q$:

$$\begin{aligned} L_{seg} &= \beta L_{dice} + L_{ce} \\ L_{dice} &= 1 - \frac{2\sum_{i,j,c} \mathbf{m}_{soft,n}^{(i,j,c)} \mathbf{y}_q^{(i,j,c)}}{\sum_{i,j,c} \mathbf{m}_{soft,n}^{(i,j,c)} + \sum_{i,j,c} \mathbf{y}_q^{(i,j,c)}} \\ L_{ce} &= -\frac{1}{HWC} \sum_{i,j,c} \mathbf{y}_q^{(i,j,c)} \log(\mathbf{m}_{soft,n}^{(i,j,c)}) \end{aligned} \tag{6}$$

where β is a constant controlling the strength of the two loss terms and is set to 1. Note that the use of the sum of dice loss and cross entropy is widely used in medical image segmentation tasks, such as [20].

4. Experiment

4.1. Setup

Dataset We conducted experiments using two abdomen CT datasets and one MRI dataset:

- ABD-110 is an abdomen dataset from [49] that contains 110 3D CT images from patients with various abdomen tumors and these CT scans were taken during the treatment planning stage.

- ABD-30 is an abdomen dataset from the MICCAI 2015 Multi-Atlas Abdomen Labeling challenge [23]. It contains 30 3D abdominal CT scans (ABD-30) from patients with various pathologies and has variations in intensity distributions between scans.

- ABD-MR is a MRI dataset from ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge [21]. It contains 20 3D T2-SPIR MRI scans.

We perform the same 5-fold cross validation and consider only 1-way 1-shot learning, following the same protocol as previous work setting 2 [30]. Liver, spleen and left and right kidney are used as semantic classes. Within each fold, one organ is considered as unseen semantic class for testing while the rest are used for training. Moreover, to reduce the variance by choosing only one support image during inference, following [59], for each query image in the test set we randomly sample one support image from the test set, repeat this process for 5 times and the final result is obtained by averaging the 5 runs.

Evaluation metric We use the same evaluation metric Sørensen–Dice coefficient (DSC) as in previous work [30, 36]. DSC measures the overlap of the prediction mask $\mathbf{m}$ and ground truth mask $\mathbf{g}$, and is defined as:

$$\text{DSC}(\mathbf{m}, \mathbf{g}) = \frac{2|\mathbf{m} \cap \mathbf{g}|}{|\mathbf{m}| + |\mathbf{g}|} \tag{7}$$

Implmentation details All images are resampled to have the same xy-plane spacing of 1.25mm $\times$ 1.25mm. For segmenting 3D volume data, we follow the same protocol used in [30, 36] by dividing the support and query images into 12 chunks and segmenting all slices in the query chunk by using the center slice in the corresponding chunk of the support image. During training, a pair of support and query images and their labels are both cropped to have a fixed size of 256 $\times$ 256 around the image center. Support and query images are aligned online using affine transformation before feeding into the network. RP-Net is trained from scratch using Adam as optimizer with initial learning rate 0.0001 for 50 epochs and the learning rate is reduced by a factor of 10 every 20 epochs. We also add the alignment loss to train RP-Net as in [59].

4.2. Comparison with the state-of-the-art methods

Table 1 shows the performance comparison of RP-Net with previous work on ABD-110, ABD-30, ABD-MR respectively. PANet [59] is an extended version of the widely used prototypical network [43] designed for natural image segmentation. PANet-init means directly using the pretrained VGG16 feature extraction backbone without any finetuning on the few-shot setting. SE-Net [36] is the first specifically designed architecture for few-shot medical image segmentation. SSL-ALPNet [30] is the state-of-the-art few-shot medical image segmentation framework that uses

Dataset	Method	Spleen	Kidney L	Kidney R	Liver	mean
ABD-110	PANet-init [59]	30.95±1.09	19.24±0.37	17.64±0.71	49.91±0.34	29.43
	PANet [59]	35.89±1.75	40.22±1.71	41.54±0.82	52.36±0.60	42.50
	SE-Net [36]	29.48±1.07	37.48±2.08	37.53±1.97	19.09±0.36	30.89
	SSL-ALPNet [30]	64.90±1.62	61.58±2.53	64.05±2.27	71.83±1.81	65.59
	Affine	50.42±0.91	53.04±1.57	52.025±2.17	66.99±1.20	55.62
	RP-Net (Ours)	**78.77±0.64**	**81.89±1.45**	**85.12±0.98**	**81.88±0.63**	**81.91**
	Fully supervised [49]	95.9	95.7	95.7	96.4	95.92
ABD-30	SE-Net [36]	0.23	32.83	14.34	0.27	11.91
	PANet [59]	25.59	32.34	17.37	38.42	29.42
	SSL-ALPNet [30]	60.25	63.34	54.82	73.65	63.02
	Affine	48.99	43.44	45.67	68.93	51.75
	RP-Net (Ours)	**69.85±2.34**	**70.48±2.55**	**70.00±0.89**	**79.62±0.91**	**72.48**
	Fully supervised [70]	96.8	95.3	92.0	97.4	95.4
ABD-MR	SE-Net [36]	51.80	62.11	61.32	27.43	50.66
	PANet [59]	50.90	53.45	38.64	42.26	46.33
	SSL-ALPNet [30]	67.02	73.63	78.39	73.05	73.02
	Affine	62.87	64.70	69.10	65	65.41
	RP-Net (Ours)	**76.35±0.66**	**81.40±2.10**	**85.78±1.12**	**73.51±1.55**	**79.26**
	Fully supervised [20]	-	-	-	-	94.6

Table 1. DSC comparison with other methods on ABD-110, ABD-30 and ABD-MR (unit: %).

Experiment	Method	Spleen	Kidney L	Kidney R	Liver	mean
Added components	Affine	50.42	53.04	52.025	66.99	55.62
	Affine + Grabcut	57.93	64.17	64.25	65.27	62.91
	Affine + Concat	56.41	52.39	54.99	70.87	58.66
	Affine + CRE	57.73	58.05	60.62	73.53	62.48
	Affine + Concat + Recurrent	59.99	60.65	62.31	83.03	66.50
	Affine + CRE + Recurrent	78.77	81.89	85.12	81.88	81.91
Backbone	VGG16	73.57	67.49	56.81	72.04	67.48
	Res18	72.39	79.13	81.61	80.89	78.50
	U-Net	78.77	81.89	85.12	81.88	81.91
Correlation radius	$d = 0$	78.40	81.90	82.12	83.89	81.58
	$d = 1$	80.03	81.87	82.09	82.1	81.52
	$d = 3$	79.12	81.79	83.41	81.32	81.41
	$d = 5$	78.77	81.89	85.12	81.88	81.91
	$d = 7$	77.56	80.25	81.77	80.22	79.95
Initialization	Affine	50.42	53.04	52.02	66.99	55.62
	Demons	63.60	63.89	61.89	73.59	65.74
	RP-Net (Affine)	78.77	81.89	85.12	81.88	81.91
	RP-Net (Demons)	80.31	83.55	85.01	82.86	82.93

Table 2. Ablation study on ABD-110 (unit: %). Underlined is the final configuration used in RP-Net.

self-supervised learning and prototypical networks. Affine is the result of the accuracy after globally aligning the support and query image using affine transformation, which we use as an initial mask. [30] reported performance for PANet-init, PANet, SE-Net and SSL-ALPNet on ABD-30 and ABD-MR, so these numbers are directly quoted. We ran these algorithms using public available code to report their performance on ABD-110.

First, compared to PANet, RP-Net outperforms PANet by 39.49%, 43.06% and 21.75% on the three datasets ABD-110, ABD-30 and ABD-MR respectively. Second, compared to SE-Net, RP-Net outperforms SE-Net by 51.02%, 60.57% and 27.42% on ABD-110, ABD-30 and ABD-MR respectively. Third, compared to the state-of-the-art method

SSL-ALPNet, RP-Net outperforms SSL-ALPNet by an average of 16.32%, 9.46% and 6.24% on ABD-110, ABD-30 and ABD-MR respectively.

These experiments demonstrate our approach can achieve the SOTA accuracy on medical image datasets with different image modalities (CT and MRI). Also, we focus on designing a new framework for few-shot medical image segmentation, which outperforms other approaches of the same motivation, e.g. SE-Net by a large margin. Additional gain may be obtained by combining our method with the self-supervised training schema proposed in SSL-ALPNet.

4.3. Ablation study

Ablation experiments are conducted using the ABD-110 dataset, because it has more data compared to the other two. Table 2 shows the results for the following experiments.

Effect of each component To verify the contribution of the two added components - context relation encoder and recurrent module, we conducted experiments by adding one component at a time: 1) model trained and tested without the CRE. To make use of the support mask which is used in CRE, we concatenate the mask to the feature map from backbone and apply a 3×3 convolution for a fair comparison (denoted as concat). 2) model trained without recurrent module. Note that if we remove both CRE and recurrent training, the model becomes the PANet [59]. Moreover, we compare with Grabcut [35] which is an unsupervised method that uses iterated Graphcut. Grabcut can be seen as an unsupervised version of our algorithm.

First, we verify the effect of using CRE. Affine + Concat is a naive way of integrating support masks by concatenating it directly to the feature maps, which outperforms the Affine by 3.04%. Affine + CRE implements the more sophisticated way of exploring local feature differences using CRE, which outperforms the Affine + Concat by 3.82%. This shows the CRE better captures the local difference via the use of correlation. However, the performance improvement is still not significant and the reason is that the mask prediction is changed each time and it lacks a mechanism to recapture this change and recompute the new local differences. The recurrent mask refinement module serves this purpose and we discuss its effect in the next paragraph.

Second, we compare the performance of using the recurrent mask refinement module. Affine + Concat + Recurrent means we apply the recurrent module to the concatenated feature map, which performs 7.84% better than not using the recurrent module (Affine + Concat). This shows that the recurrent training indeed helps the model to find the right mask prediction because the initial mask from support is a very rough estimation of the location of the region of interests. If we combine the two added components together (Affine + CRE + Recurrent), we can achieve a big improvement by 15.39% compared to Affine + Concat + Recurrent. This demonstrates that the integration of recurrent module to recapture local changes in the CRE is very important and can greatly boost the performance.

Third, we compare with Grabcut. Our method is in some sense similar to Grabcut - we both use an iterative update to refine the segmentation mask. Grabcut outperforms the baseline Affine by 7.29%, showing that iteratively refining a mask is indeed beneficial. RP-Net (Affine + CRE + Recurrent) outperforms Grabcut by 19%. There are mainly three reasons for this large improvement. First, Grabcut only uses one image, thus only image intensity is used to separate foreground and background region. On the contrary, RP-Net uses the support images to extract knowledge about the relationship between the foreground and background region, and utilize this knowledge to guide the segmentation of the new image. Second, Grabcut only refines the mask in the probable foreground region which is a human defined boundary and lacks the flexibility to attend other areas in the image, as well as the ability to correct error in the sure foreground region. RP-Net does not have these constraints and can potentially use information from the whole image. Third, RP-Net uses training data to train the feature extractor, while Grabcut is not a learning-based method and only uses information directly derived from pixel intensity.

Effect of feature extraction backbone We also experimented with three different feature extraction backbones - VGG16 [42], Res18 [13] and U-Net [34]. To make sure the output feature map is 1/4 of the original image resolution for a fair comparison, we only kept the first two downsampling operations in both VGG16 and Res18 backbones and the rest of the network architecture remained the same. As seen from Table 2, VGG16 backbone performs the worst among the three backbones, which is 8.03% lower than Res18. U-Net backbone outperforms Res18 backbone by an average of 2.32% which is mainly because of the lateral connection in U-Net that fuses both low-level and high-level features. This demonstrates that RP-Net is compatible with different backbones, and backbones that perform better on medical image segmentation task, such as U-Net, would result in similar gain when combined with RP-Net.

Effect of correlation radius We conducted experiments with different radius $d = 0, 1, 3, 5, 7$ in the correlation layer, which controls how many neighbouring pixels are included when computing correlation. $d = 0$ means the correlation computation is carried out only at a single point. Note that even with $d = 0$, the model is able to use features from the surrounding pixels because ϕ_f and ϕ_b are used to extract foreground and background specific features. Table 2 shows our approach is not very sensitive to the radius, and this is likely because RP-Net is designed to focus on a small region around the object boundary at a time, a larger context may not necessarily bring more benefits.

Effect of number of inference iterations We show in

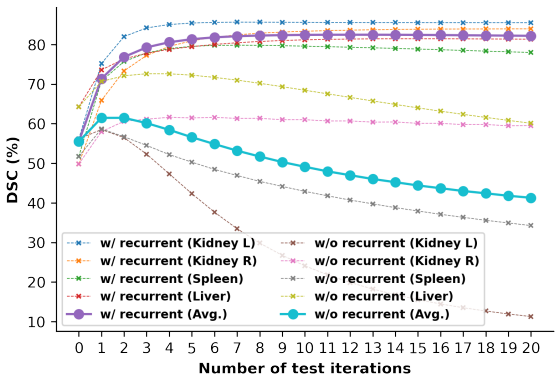

Figure 2. DSC at each refinement iteration. This figure shows the DSC performance of the proposed model per iteration. DSC of four organs and an average is shown for two models: one w/ recurrent training (purple) and one w/o recurrent training (cyan).

Figure 2 the performance at each inference iteration from one fold in ABD-110. Although the model is trained using 4 iterations of recurrent module, we can apply more iterations during inference. As seen from this figure, a model without recurrent training diverges after the 1st iteration, while a model with recurrent training quickly converges and does not diverge after 20 epochs. It demonstrates that with the recurrent training, the model learns to gradually refine its prediction and converges to a stable solution.

Effect of initialization Demons [54] is a medical image registration method that uses deformable registration, which performs 10.12% better than a simple affine transformation. As shown in Figure 2, using a better initialization (Demons), RP-Net achieves a 1.02% improvement. Although better initialization improves the result, the improvement is small compared to that of the initialization itself, and our network is less sensitive to the initial mask as long as it roughly locates the foreground region. For this reason, we only use initialization mask from Affine transformation for its simplicity. In many cases, a coarse map or a map derived through affine registration would suffice. Some recent registration methods (e.g., DEEDS [15] and its extensions) that can handle large anatomical variations, although missing details, can fit well to our method.

4.4. Qualitative result

We show in Figure 3 how the segmentation mask converges to the optimum solution in multiple iterations. In general, we can observe that RP-Net refines the initial mask gradually, finds a better segmentation mask at each iteration, and finally converges to an optimum solution. RP-Net is able to learn to distill knowledge about the relation between the foreground and background from the support image, and apply it to segment query images by comparing local differences and modifying its prediction to conform to

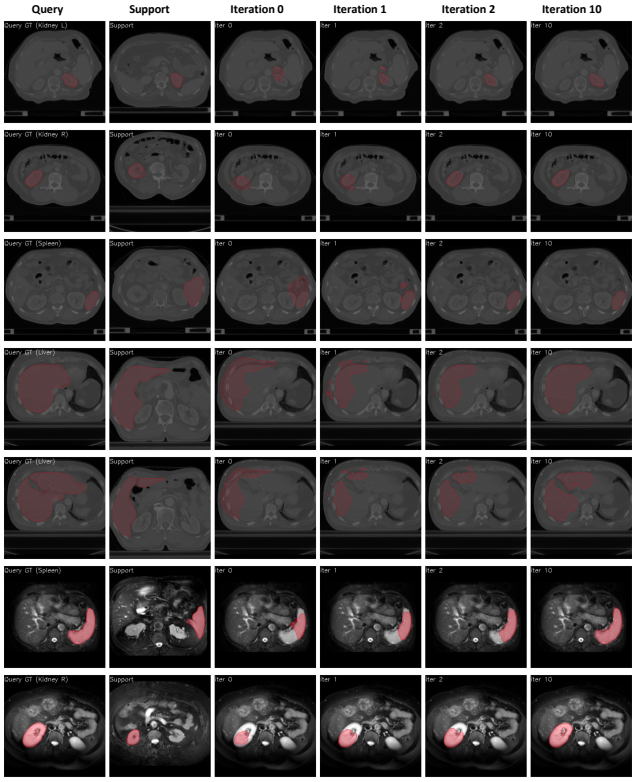

Figure 3. Examples of predication of RP-Net at different iterations. Each row represents one slice of the a test scan (row 1-5 are CT images, row 6-7 are MR images).

the shape and boundary. Moreover, RP-Net generates satisfying segmentation masks that have a clear boundary along the object boundary, demonstrating the successful design of the CRE and recurrent module.

5. Conclusion

In this work, we present a new few-shot medical image segmentation framework that refines the segmentation mask iteratively using a context relation encoder and a recurrent module. The proposed model learns to incrementally refine the segmentation mask to better align the object boundary. Experiments on three organ segmentation datasets demonstrate that RP-Net outperforms the previous state-of-the-art approach by as much as 16% in terms of DSC. Moreover, the proposed CRE and recurrent module are generic and can also be integrated into other types of network to enhance context relationship features.

Acknowledgement This work is partly supported by NSF grant IIS-1715017, a Simons Foundation grant (594598), and a hardware grant from NVIDIA.

References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

[2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[3] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170:446–455, 2018.

[4] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021.

[5] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE transactions on medical imaging*, 38(5):1116–1126, 2018.

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

[7] Tianming Du, Yanci Zhang, Xiaotong Shi, and Shuang Chen. Multiple slice k-space deep learning for magnetic resonance imaging reconstruction. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1564–1567, 2020.

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[10] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *arXiv preprint arXiv:1806.02817*, 2018.

[11] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.

[12] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Mattias P Heinrich. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 50–58. Springer, 2019.

[15] Mattias P Heinrich, Oskar Maier, and Heinz Handels. Multimodal multi-atlas segmentation using discrete optimisation and self-similarities. *VISCERAL Challenge@ ISBI*, 1390:27, 2015.

[16] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[17] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees GM Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8441–8448, 2019.

[18] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019.

[19] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[20] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[21] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[22] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[23] B Landman, Z Xu, J Igelsias, M Styner, T Langerak, and A Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[24] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[25] Haoyu Ma, Renhao Lei, Junxiao Sun, and Youyong Kong. Multi-session parcellation of the human brain using resting-state fmri. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*, pages 336–340. IEEE, 2018.

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[27] Arnab Kumar Mondal, Jose Dolz, and Christian Desrosiers. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241*, 2018.

[28] Jeffrey J Nirschl, Andrew Janowczyk, Eliot G Peyster, Renee Frank, Kenneth B Margulies, Michael D Feldman, and Anant Madabhushi. Deep learning tissue segmentation in cardiac histopathology images. In *Deep learning for medical image analysis*, pages 179–195. Elsevier, 2017.

[29] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[30] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European Conference on Computer Vision*, pages 762–780. Springer, 2020.

[31] Cheng Ouyang, Konstantinos Kamnitsas, Carlo Biffi, Jinming Duan, and Daniel Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 669–677. Springer, 2019.

[32] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2020.

[33] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. 2018.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[35] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[36] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020.

[37] Hyunseok Seo, Charles Huang, Maxime Bassenne, Ruoxiu Xiao, and Lei Xing. Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE transactions on medical imaging*, 39(5):1316–1325, 2019.

[38] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[39] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[40] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[41] Mennatullah Siam, Boris N. Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[43] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[44] Liyan Sun, Wenao Ma, Xinghao Ding, Yue Huang, Dong Liang, and John Paisley. A 3d spatially weighted network for segmentation of brain tissue from mri. *IEEE transactions on medical imaging*, 39(4):898–909, 2019.

[45] Shanlin Sun, Yang Liu, Narisu Bai, Hao Tang, Xuming Chen, Qian Huang, Yong Liu, and Xiaohui Xie. Attentionanatomy: A unified framework for whole-body organs at risk segmentation using multiple partially annotated datasets. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2020.

[46] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[47] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, pages 1–12, 2019.

[48] Hao Tang, Daniel R Kim, and Xiaohui Xie. Automated pulmonary nodule detection using 3d deep convolutional neural networks. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 523–526. IEEE, 2018.

[49] Hao Tang, Xingwei Liu, Kun Han, Xiaohui Xie, Xuming Chen, Huang Qian, Yong Liu, Shanlin Sun, and Narisu Bai. Spatial context-aware self-attention model for multi-organ segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 939–949, 2021.

[50] Hao Tang, Xingwei Liu, and Xiaohui Xie. An end-to-end framework for integrated pulmonary nodule detection and false positive reduction. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 859–862. IEEE, 2019.

[51] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Automatic pulmonary lobe segmentation using deep learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1225–1228. IEEE, 2019.

[52] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Nodulenet: Decoupled false positive reductionfor pulmonary nodule detection and segmentation. *arXiv preprint arXiv:1907.11320*, 2019.

[53] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[54] J-P Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[55] Nuo Tong, Shuiping Gou, Shuyuan Yang, Dan Ruan, and Ke Sheng. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical physics*, 45(10):4558–4567, 2018.

[56] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

[57] Eugene Vorontsov, An Tang, Chris Pal, and Samuel Kadoury. Liver lesion segmentation informed by joint liver segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1332–1335. IEEE, 2018.

[58] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7:53, 2019.

[59] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019.

[60] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[61] Fan Xu, Haoyu Ma, Junxiao Sun, Rui Wu, Xu Liu, and Youyong Kong. Lstm multi-modal unet for brain tumor segmentation. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pages 236–240. IEEE, 2019.

[62] Shipeng Yan, Songyang Zhang, Xuming He, et al. A dual attention network with semantic embedding for few-shot learning. In *AAAI*, pages 9079–9086, 2019.

[63] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE transactions on medical imaging*, 39(1):188–203, 2019.

[64] Chenyu You, Qingsong Yang, Hongming Shan, Lars Gjesteby, Guang Li, Shenghong Ju, Zhuiyang Zhang, Zhen Zhao, Yi Zhang, Wenxiang Cong, et al. Structurally-sensitive multi-scale deep neural network for low-dose ct denoising. *IEEE Access*, 6:41839–41855, 2018.

[65] Hanchao Yu, Shanhui Sun, Haichao Yu, Xiao Chen, Honghui Shi, Thomas S Huang, and Terrence Chen. Foal: Fast online adaptive learning for cardiac motion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4313–4323, 2020.

[66] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019.

[67] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020.

[68] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8553, 2019.

[69] Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Medical image analysis*, 43:98–111, 2018.

[70] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10672–10681, 2019.