Bidirectional RNN-based Few Shot Learning for 3D Medical Image Segmentation

Soopil Kim, Sion An, Philip Chikontwe, Sang Hyun Park
Department of Robotics Engineering, DGIST
soopilkim, sion_an, philipchicco, shpark13135@dgist.ac.kr

Abstract

Segmentation of organs of interest in 3D medical images is necessary for accurate diagnosis and longitudinal studies. Though recent advances using deep learning have shown success for many segmentation tasks, large datasets are required for high performance and the annotation process is both time consuming and labor intensive. In this paper, we propose a 3D few shot segmentation framework for accurate organ segmentation using limited training samples of the target organ annotation. To achieve this, a U-Net like network is designed to predict segmentation by learning the relationship between 2D slices of support data and a query image, including a bidirectional gated recurrent unit (GRU) that learns consistency of encoded features between adjacent slices. Also, we introduce a transfer learning method to adapt the characteristics of the target image and organ by updating the model before testing with arbitrary support and query data sampled from the support data. We evaluate our proposed model using three 3D CT datasets with annotations of different organs. Our model yielded significantly improved performance over state-of-the-art few shot segmentation models and was comparable to a fully supervised model trained with more target training data.

Introduction

Deep learning based segmentation models have achieved success in various applications (Gu et al. 2019; Zhou et al. 2018; Alom et al. 2019; Nie et al. 2019) for both natural and medical images. Despite their success, it is often difficult to make a robust model for segmentation especially for medical images since constructing a large-scale dataset incurs a high cost in scanning, and manually creating annotations for volumetric images is laborious and time-consuming. Above all, each hospital acquires images with different resolutions and modalities given that medical experts are interested in different tasks. Consequently, it would constitute designing a separate model for each task, which is not practical. Moreover, each task may have a low data regime with limited annotated samples, and training based on fine-tuning or transfer learning may fail and lead to overfitting.

Recently, few shot learning methods (Snell, Swersky, and Zemel 2017; Sung et al. 2018; Garcia and Bruna 2017; Liu et al. 2018) have been proposed in the field of machine learning to efficiently address these challenges. The key idea of few shot learning is to learn generalizable knowledge across several related tasks that can be used to predict the label of a query sample with support data and labels. Therefore, in the ideal setting, when this idea is applied to medical datasets, a model trained with several existing organ annotations should be able to accurately segment unseen target organs with only a few samples. However, most few shot methods that focus on natural 2D images are not directly applicable for 3D image based analysis since such models are prone to overfitting when trained with few observations.

In general, 3D operations for dense pixel-level high dimensional predictions incur increased high memory usage and often lead to constraining the batch size to be small. Recently, (Roy et al. 2020) proposed a framework for organ segmentation in 3D CT scans using few shot learning. Volumetric segmentation is performed in 2D slice by slice in the coronal view with careful selection of potential query and support slices. Though successful, their approach does not consider the relation between adjacent slices and 3D structural information, therefore the segmentation result may often be inaccurate and not smooth.

In this paper, we propose a novel few shot segmentation framework that models the relation between support and query data from other few shot tasks alongside 3D structural information between adjacent slices. We integrate a bidirectional gated recurrent unit (GRU) between the encoder and decoder of a 2D few shot segmentation model for efficient representation learning. In this way, encoded features of both the support set and adjacent slices capture key characteristics to predict the segmentation of a query image in the decoding layers. Furthermore, we propose a transfer learning strategy to adapt the characteristics of a target domain in multi-shot segmentation setting. For a given task, we re-train the model parameters using the given support data by arbitrarily dividing them into support and query data using data augmentation. We evaluated our method using 3 datasets (1 for internal test and the rest for external validation) to verify the generalization ability of the model.

In our experiments, the proposed model using less than 5 support samples could achieve performance comparable to a supervised learning method used a large number of data samples. The contribution points of this paper are summarized as follows:

- We propose a novel 3D few shot segmentation model able to capture key relationships between adjacent slices in volume via bidirectional GRU modules.

- We propose a transfer learning strategy to improve performance in multi-shot segmentation model.
- We empirically demonstrate the generalization ability of our method through several experiments on both internal and external datasets with various organs and different resolutions.

Related Works

Few Shot Learning for Segmentation

Few shot learning aims to learn transferable knowledge that can be used to generalize to unseen tasks with scarce labeled training data. It has been widely used for classification (Snell, Swersky, and Zemel 2017; Sung et al. 2018; Garcia and Bruna 2017; Liu et al. 2018), segmentation (Li et al. 2020; Wang et al. 2019; Zhang et al. 2019; Shaban et al. 2017), and regression tasks (Finn, Abbeel, and Levine 2017; Wang, Ramanan, and Hebert 2017; Zhou et al. 2019). For the segmentation task, (Shaban et al. 2017) first proposed a method with a fully convolutional network (FCN) that learns optimal parameters from a support image and its label to segment the relevant object from in a query image. (Dong and Xing 2018) defined a prototype, i.e., a feature vector with high-level discriminative information, through encoder and global average pooling from support data, then predicted segmentation based on similarity with the prototype of the query image. (Wang et al. 2019) additionally proposed an alignment loss that inversely guesses the support label by using the query image and the predicted query label as support data in order to select a more accurate prototype. (Zhang et al. 2019) proposed a model that refines the segmentation result by optimizing the model in an iterative manner, and (Liu et al. 2020) improved the performance by defining the prototype as a set of part-prototypes so that parts of an object can be recognized. However, prototype-based methods are limited in that the resolution of the predicted segmentation is often low since the relationship between the prototypes or parameters for prediction is learned in the down-sampled embedding space and then rapidly up-sampled using interpolation.

Recently, (Li et al. 2020) progressively performed upsampling of encoded features using a decoder with skip connections at different levels. In (Hu et al. 2019), segmentation was performed using features obtained at various stages of the encoder with an attention mechanism. Also, (Roy et al. 2020) proposed a few shot method for 3D CT organ segmentation using dense connections with squeeze and excitation blocks added between the modules for support and query data. However, most of the proposed methods have limitations in obtaining smooth segmentation for 3D organs because they estimate segmentation of the query image by only relying on support data without considering contextual information between adjacent slices. Unlike the existing methods, we propose a 3D few shot segmentation method that can consider the relationship between adjacent slices for more reliable predictions.

Recurrent Neural Networks for 3D Medical Image Segmentation

A number of deep learning based segmentation methods have been proposed for 3D medical image analysis. Although some architectures are fully implemented with 3D convolutions, many methods predict 3D segmentation based on 2D slices (Zhao, Chen, and Cao 2019; Oktay et al. 2018; Roth et al. 2015) or 3D patches (Korez et al. 2016; Zhu et al. 2017; Luna and Park 2018) and then perform aggregation due to the expensive computations of 3D operations and limited training data. Though such methods can learn complex tasks relatively well, the main disadvantage is the lack of global context information used for prediction. To alleviate this problem, several methods have been proposed to consider consistency between adjacent slices using recurrent neural networks (RNN). In this setting, final segmentation is achieved by obtaining predictions on adjacent slices via U-Net or FCN, then perform updates on the feature maps through RNNs. For example, (Xu et al. 2019) and (Cai et al. 2017) feed U-Net predicted segmentation maps to an LSTM module. On the other hand, (Chen et al. 2016), (Cai et al. 2018), and (Li et al. 2019) used a bidirectional LSTM. (Bai et al. 2018) integrated a convolution based bidirectional LSTM into U-Net in order to model the relationship of encoded features between adjacent slices. Though successful, previous methods require a large amount of labeled training data to learn a robust model. Instead, we propose to integrate an RNN in the few shot setting for segmentation. In this way, we alleviate the issue of low training data and learn robust features considering context between slices for several organ segmentation tasks.

Transfer Learning in Few Shot Learning

Transfer learning has been shown to improve model performance for different tasks by leveraging deep models already trained on larger datasets with related tasks or characteristics. Recently, several methods have been proposed to address fine tuning in the few shot setting with limited target support samples. For example, (Caelles et al. 2017) proposed a fine tuning approach for 1-shot video object segmentation using the first frame. This method was also used in metric-based few shot segmentation (Shaban et al. 2017) even though it was limited in application i.e., used to update only part of the model. In addition, optimization based few shot methods also employ fine-tuning in the intermediate stages of training. (Finn, Abbeel, and Levine 2017) proposed to temporarily update the model using support data and minimize the loss for each task, in turn good performance is obtained with only a few model updates. (Sun et al. 2019) proposed to only update the fine tuning module during the fine-tuning stage by disentangling learning of general and transferable knowledge in the modules. Inspired by prior methods, we introduce a method to learn optimal parameters for the target task by using a fine tuning process that performs additional updates by randomly dividing support data in the K-shot setting.

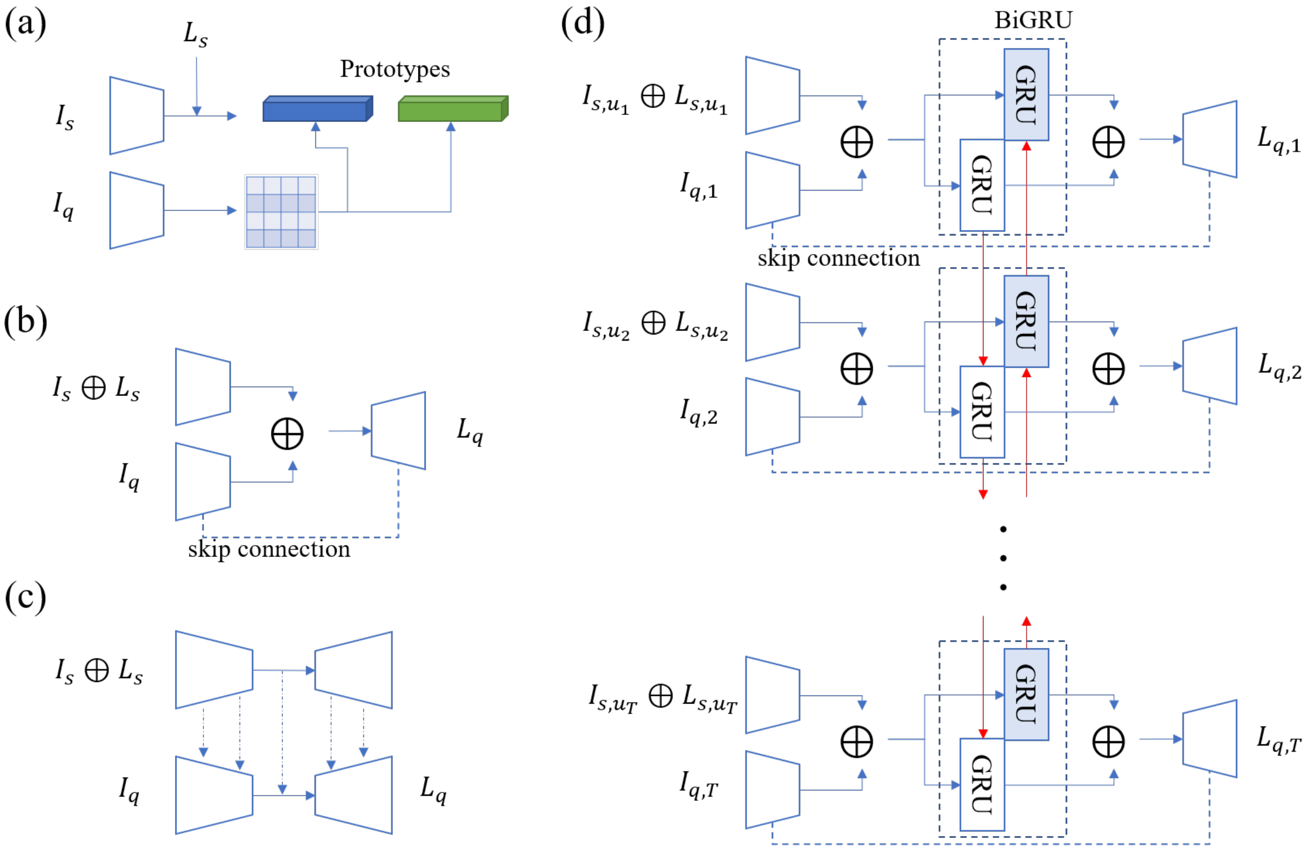

Figure 1: Segmentation models using few shot learning. (a), (b), and (c) are the existing few shot segmentation models (Wang et al. 2019; Li et al. 2020; Roy et al. 2020), while (d) is our proposed method.

Methods

Problem Setup

A few shot segmentation model FSS_θ learns parameters θ to segment a target object in a query image I_q using K pairs of support images and labels $\{I_s^1, L_s^1\}$, $\{I_s^2, L_s^2\}$, ..., $\{I_s^K, L_s^K\}$, where K defines the degree of supervision. Common architectures for 2D image based few shot segmentation are shown in Fig. 1 (a) to (c). (a) shows a prototypical network which defines prototypes for target object and background, and then performs segmentation using the distance to the defined prototypes, (b) is a relation network consisting of encoders and decoders for segmentation in a fully convolutional network, and (c) represents a network with dense connections between modules for support and query data. All of these methods learn the relations between support and query data of various segmentation tasks and utilize them to predict a target label in I_q using K support samples as:

$$L_q = FSS_\theta(\{I_s^k, L_s^k\}_{k=1}^K, I_q). \tag{1}$$

(Roy et al. 2020) extended this concept to 3D image few shot segmentation. They divided a query volume $\mathbf{I}_q$ into multiple 2D slices $I_{q,1}$, $I_{q,2}$, ..., $I_{q,T}$ where T is the number of slices in the axial view, and then segmented each query slice $I_{q,t}$ separately using the corresponding support slice I_{s,u_t} and label L_{s,u_t}, where u_t is the index of the support sample. To determine I_{s,u_t} and L_{s,u_t} in a 3D support volume $\mathbf{I}_s$ with label $\mathbf{L}_s$, it was assumed that the starting and ending slice locations of the organ of interest in $\mathbf{I}_q$ and $\mathbf{I}_s$ were known. The index of a support slice corresponding to $I_{q,t}$ is obtained via $u_t = round((t/T) \times \hat{T})$, where $\hat{T}$ is the number of interest slices in $\mathbf{I}_s$. This assumption is reasonable since the organ of interest is in a similar position from person to person, e.g., the liver is always located in the upper right part of the abdomen, even though it varies in size and shape. In this setting, few shot segmentation of a 3D image can be represented by the following formulation:

$$\mathbf{L}_q = \{L_{q,t}\}_{t=1}^T = \left\{FSS_\theta(\{I_{s,u_t}^k, L_{s,u_t}^k\}_{k=1}^K, I_{q,t})\right\}_{t=1}^T. \tag{2}$$

Most 2D based few shot segmentation models can follow this setting, but the relation between adjacent slices is not considered. In this study, we follow the mentioned problem setting but propose to incorporate adjacent slice information to accurately segment $I_{q,t}$. Formally,

$$L_{q,t} = FSS_\theta\Big(\Big\{\{I^k_{s,u_t}, L^k_{s,u_t}\}^K_{k=1}, I_{q,t}\Big\}^{t_0+n_a}_{t=t_0-n_a}\Big), \quad (3)$$

where $2n_a + 1$ is the number of adjacent slices and t_0 is the index of the center of multiple slices. To achieve this, we introduce a 3D few shot segmentation method using the bidirectional RNN.

Bidirectional RNN-based Few Shot Learning

Our model performs segmentation in three stages; (1) features of support and query images are extracted through two separate encoders E_s and E_q, respectively. (2) A bidirectional GRU models the relationship between features extracted from adjacent slices. (3) Finally, using updated feature maps and low level features in E_q, the decoder predicts the segmentation. An overview of the one-shot segmentation model is shown in Fig. 1 (d).

Feature Encoders We used two separate encoders E_s and E_q to extract features from support and query images since they receive inputs with different number of channels. E_s receives 2-channel input which is a concatenation of I_{s,u_t} and L_{s,u_t}, whereas E_q receives 1-channel $I_{q,t}$ as input. We used an ImageNet VGG16 (Simonyan and Zisserman 2014) model in the encoder module given its robust feature extraction ability. Since VGG16 takes a 3-channel image as input, we randomly initialized the parameters of the first layer and concatenate the features from the two encoders as:

$$x_t = E_s(\{I_{s,u_t}, L_{s,u_t}\}^K_{k=1}) \oplus E_q(I_{q,t}), \quad (4)$$

then feed x_t into the GRU model. Low-level features with different resolutions extracted from E_q are used again in the subsequent stage.

Bidirectional GRU After features $\{x_t\}^T_{t=1}$ are extracted by encoders from $\{I_{s,u_t}, L_{s,u_t}, I_{q,t}\}^T_{t=1}$, GRU models the change between adjacent slices. In particular, a bidirectional GRU has two modes i.e. both forward and backward directions for efficient feature representation. Features are sequentially feed into the forward GRU and later reversed for the backward model. Each GRU calculates two gate controllers z_t and r_t with x_t and a prior hidden state h_{t-1} for memory updates as:

$$z_t = \sigma(conv_z(h_{t-1} \oplus x_t) + b_z), \quad (5)$$

$$r_t = \sigma(conv_r(h_{t-1} \oplus x_t) + b_r). \quad (6)$$

z_t controls the input and output gates whereas r_t determines which part of the memory would be reflected on the hidden state h_t. Formally,

$$\hat{h}_t = tanh(conv_h(r_t \odot h_{t-1} \oplus x_t) + b_h), \quad (7)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z \odot \hat{h}_t. \quad (8)$$

In our GRU module, operations are replaced with 3×3 convolutions instead of weight multiplication in normal GRU cell. Sigmoid activation functions are used after the gate controller output, and a hyperbolic tangent function applied following the final hidden state output. Following, features extracted from the forward GRU h^f_t and backward GRU model h^b_t are concatenated as:

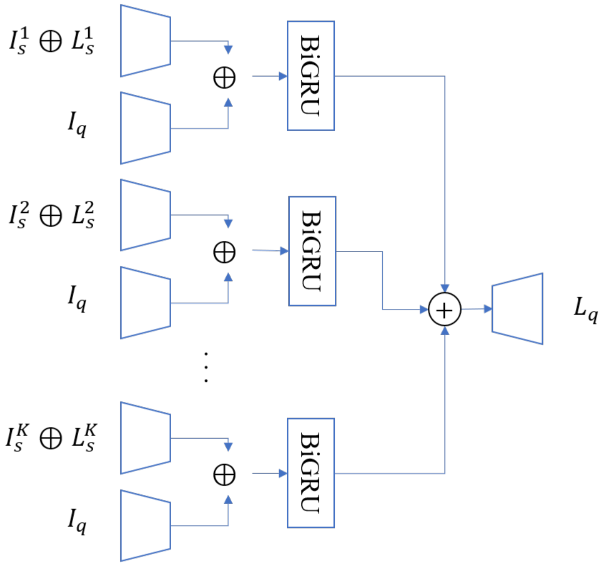

Figure 2: Model architecture for K shot segmentation. The features from bidirectional GRU are summed before the decoder.

$$h^{biGRU}_t = h^f_t \oplus h^b_t, \quad (9)$$

and then passed to the decoder. We used the GRU modules due to their low memory footprint, though any bidirectional RNN such as the Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber 1997) can alternatively be used for sharing features between adjacent slices.

When K support data $(I^k_s, L^k_s)^K_{k=1}$ are used (see Fig. 2), the GRU operation is performed for each support and query data pair, then the obtained features are summed as:

$$h^{biGRU}_t = \sum^K_{k=1} h^{biGRU,k}_t. \quad (10)$$

Finally, h^{biGRU}_t is passed to the decoder.

Decoder Our decoder has a similar architecture to UNet (Ronneberger, Fischer, and Brox 2015) for high resolution segmentation. The final segmentation is obtained by utilizing the features processed by the encoders and bidirectional GRU, as well as the low level features of the query slice image obtained from E_q. Low level features are connected to the decoder by skip connections, wherein the decoder predicts the segmentation using both low- and high-level information. Our cost function was defined as the summation of cross entropy loss and dice loss (Milletari, Navab, and Ahmadi 2016) between the prediction and label.

Transfer Learning-based Adaptation

Since the target organ is never seen in the training stage, it may be challenging for model to learn the optimal parameters for the target. Thus, in multi-shot scenarios, we perform transfer learning with few target-support samples prior to testing. Specifically, we temporarily sample support and query pairs from the support data and update our model. For example, in K shot testing stage, we collect existing pairs of support and query data by choosing $K-1$ samples from the K support set as temporary support samples and use the remainder as a temporary query image to retrain our model. Since we use $2n_a+1$ adjacent slices from a 3D volume, many different training pairs can be sampled enabling a robust fine tuning process. Moreover, to avoid overfitting and encourage training stability, we use random flipping and rotation based augmentation during training. Using this strategy, our model can effectively adapt to the novel characteristics of target data.

Experiments

Dataset

The proposed method was evaluated on the Multi-atlas labeling Beyond the Cranial Vault (BCV) dataset (Landman et al. 2015) which was used in a challenge at MICCAI 2015. The dataset includes 30 3D CT scans with segmentation labels for 15 organs. Among 15 organs, the labels of 9 organs (spleen, left kidney, esophagus, liver, stomach, aorta, inferior vana cava, bladder, and uturus) were used in our experiments since the other 6 organs were too small or hard to segment even with a supervised learning method due to large shape variations.

Moreover, we used two external datasets to see if the proposed model would be applicable to data with different characteristics. We employed CTORG (Blaine Rister and Rubin 2019) which contains 119 images with labels of 6 organs (lung, bones, liver, left and right kidneys, and bladder). It is worth noting that variations of in-plane resolution and thickness between images were significant since the dataset was collected from multiple sites. In our experiments, external tests were performed on the liver, kidney, and bladder, excluding the lung and brain since many CT scans did not include the whole part of the lung, and the brain samples were limited. Second, we also evaluated our method on the DECATHLON(Simpson et al. 2019) dataset. It consists of several images with 10 different organs (CT Liver, multimodal MRI brain tumors, mono-modal MRI hippocampus, CT lung tumors, Multimodal prostate, monomodal left atrium, CT pancreas, CT colon cancer primaries, CT hepatic vessels, and CT spleen). In a similar fashion with earlier experiments, spleen and liver data were used in the external tests, excluding organs that were too small or had severe shape changes.

Experimental Settings

The BCV dataset was divided into 15 volumes for training or selecting support data, 5 volumes for validation, and 10 volumes for testing for each organ. In training, pairs of support and query data were randomly sampled from the 15 volumes with 8 organs except a certain target organ to train the few shot models. For testing, support data was randomly sampled among the 15 volumes for the target organ, while the 10 volumes were used as query images. Since performing experiments for all organs was time-consuming, we tested our model on 4 clinically important organs (spleen, liver, kidney, and bladder) that are not too small. For example, the adrenal gland was excluded because it appears in limited slices of CT scans and was often hard to figure out the organ's 3D structure. For external validation, the models trained on the BCV dataset were applied to 65 liver, 63 kidney, and 53 bladder samples in CTORG, and 27 spleen and 87 liver data in the DECATHLON dataset. Voxel intensities of all images were normalized to range from 0 to 1 and slices were cropped into squares with the same size for each organ, then resized to 256×256.

To show the effectiveness of proposed model, we compared our method with a based supervised method and three few shot models (FSS_{base} (Li et al. 2020), $FSS_{prototype}$ (Wang et al. 2019), and FSS_{SE} (Roy et al. 2020)) recently proposed. A (Ronneberger, Fischer, and Brox 2015) trained with only one sample per organ was used as a lowerbound while that trained with all accessible data was used as a upperbound model. For fair comparison with our proposed model, was modified to use 5 adjacent axial slices as input and consisted of 2D convolutional encoder and decoder. He initialization (He et al. 2015) was used for all models with Adam (Kingma and Ba 2014) optimization and a learning rate of 10^{-4}. For every iteration in the training stage, support and query volumes were randomly selected from training data containing various organ segmentation labels except the target organ. For the bidirectional GRU models, a total 5 slices were feed into the model, i.e., n_a was set as 2. Similar settings were used for inference. In addition, the same parameter initialization and data augmentation (flipping and rotation) was applied across all evaluated models.

FSS_{base} is a baseline model with a similar architecture to the proposed model if the bidirectional GRU module is omitted. $FSS_{prototype}$ uses prototypes and an alignment process for predictions. It defines prototypes of foreground and background to implement distance-based pixel-wise classification on reduced feature maps extracted by an encoder. On the other hand, FSS_{SE} model uses squeeze and excitation blocks with skip connections trained from scratch, with a separate encoder and decoder for support and query data. Except for FSS_{SE}, we evaluated 1, 3, and 5 shot models on the internal and external testing datasets. Since FSS_{SE} was designed for the one-shot setting, 3 and 5 shot settings were not considered for evaluation. As for the proposed model, we note it as FSS_{BiGRU}.

Since the performance of segmentation may vary depending on how the support set is selected, the experiment was performed with different support sets randomly sampled 5 times for each query sample, and we report the average across trials. Segmentation performance was measured by the dice similarity score between the prediction and label.

Results and Discussion

Internal Test We show an overall comparison for methods trained and tested on the BCV dataset in Table 1. FSS_{BiGRU} with/without fine-tuning and its variants using different number of samples showed performance comparable to the fully supervised baseline. The margins was largely significant in the one-shot setting across all organs

Model	#	Spleen	Kidney	liver	bladder	mean
U-Net (lower)	1	0.695	0.793	0.645	0.574	0.677
U-Net (upper)	15	0.896	0.884	0.902	0.788	0.867
FSS_{base}	1	0.722±0.010	0.846±0.017	0.751±0.027	0.515±0.045	0.703
	3	0.781±0.033	0.829±0.021	0.788±0.008	0.496±0.038	0.711
	5	0.829±0.015	0.861±0.008	0.849±0.008	0.585±0.008	0.727
$FSS_{prototype}$	1	0.758±0.049	0.759±0.033	0.788±0.014	0.584±0.204	0.722
	3	0.834±0.018	0.791±0.006	0.836±0.011	0.696±0.016	0.790
	5	0.832±0.012	0.787±0.016	0.843±0.004	0.701±0.009	0.791
FSS_{SE}	1	0.767±0.022	0.824±0.010	0.750±0.062	0.607±0.033	0.737
FSS_{BiGRU}	1	0.816±0.028	0.829±0.061	0.817±0.039	0.648±0.017	0.778
	3	0.871±0.024	0.885±0.009	0.835±0.010	0.656±0.014	0.812
	5	0.888±0.013	0.889±0.005	0.864±0.006	0.698 ± 0.004	0.835
FSS_{BiGRU}+FT	5	**0.905±0.019**	**0.900±0.006**	**0.887±0.007**	**0.771±0.030**	**0.866**

Table 1: Performance comparison of the proposed model FSS_{BiGRU} against baseline models on the BCV dataset using the (dice score ± stdev) evaluation metric. The second column represents the number of training data (#) and FT denotes the fine-tuning. Boldface represents the best accuracy among the few shot comparison methods.

		DECATHLON		CTORG		
Model	#	Spleen	Liver	Kidney	liver	bladder
U-Net (BCV)	15	0.704	0.875	0.553	0.806	0.606
FSS_{base}	5	0.848±0.017	0.802±0.009	0.657±0.015	0.779±0.003	0.522±0.013
$FSS_{prototype}$	5	0.837±0.009	0.791±0.022	0.612±0.007	0.725±0.012	0.548±0.010
FSS_{SE}	1	0.788±0.002	0.644±0.052	0.529±0.005	0.610±0.052	0.663±0.004
FSS_{BiGRU}	1	0.841±0.005	0.792±0.029	0.623±0.031	0.751±0.015	0.703±0.009
	3	0.873±0.011	0.812±0.016	0.689±0.041	0.726±0.004	0.713±0.005
	5	0.878±0.029	0.852±0.013	**0.716±0.03**	**0.788±0.003**	0.729±0.005
FSS_{BiGRU}+FT	5	**0.900±0.022**	**0.888±0.009**	0.701±0.025	0.778±0.009	**0.731±0.001**
U-Net (lower)	1	0.571	0.625	0.621	0.727	0.243
U-Net (upper)	27,87,65,63,53	0.858	0.927	0.915	0.867	0.767

Table 2: Performance comparison on external datasets (dice score ± stdev). The (BCV) is the model trained by BCV dataset, while the (lower) and (upper) are trained using the same external datasets. Due to different numbers of volumes for each organ, we indicate 5 numbers used as the number of training data (#) for (upper), e.g. 27 training volumes for spleen(DECATHLON). Boldface represents the best accuracy among the few shot comparison methods.

with an approximate 20% mean score improvement. This clearly shows the proposed method is feasible for segmentation even under an extreme limited data regime.

Notably, our method became better (i.e., the accuracy increased) and robust (i.e., the standard deviation decreased) as the data samples were increased in most cases. Though unsurprising that the upperbound had high scores on most organs, it is rather significant that our model showed similar performance. In addition, we noted that transfer learning largely improved overall performance i.e. +3% mean score improvement over FSS_{BiGRU}. This implies that after additional updates our model was able to adapt the segmentation task of a target organ unseen in training. These results demonstrate that reliable segmentation can be achieved when considering multiple slices alongside 3D structural information to encode relationships between adjacent slices.

External Test In Table 2, we further evaluate our approach on external datasets to assess model performance under distribution shift. For simplicity, we considered the 5-shot setting for FSS_{base} and $FSS_{prototype}$. The upperbound trained on BCV as well as the upper- and lowerbound methods trained using all accessible data in the external set are also included for completeness.

In general, we observed that the performance of upperbound models trained with BCV dataset deteriorated on the external datasets. The model achieved significantly reduced scores across most organs except the liver in DECATHLON dataset compared to the results in Table 1. This is due to the different scanning protocols and machinery employed in the clinical setup. It may be challenging to achieve reliable segmentation on external datasets whose resolution is different since the model may overfit to appearance of specific resolution.

On the other hand, few shot learning based segmentation

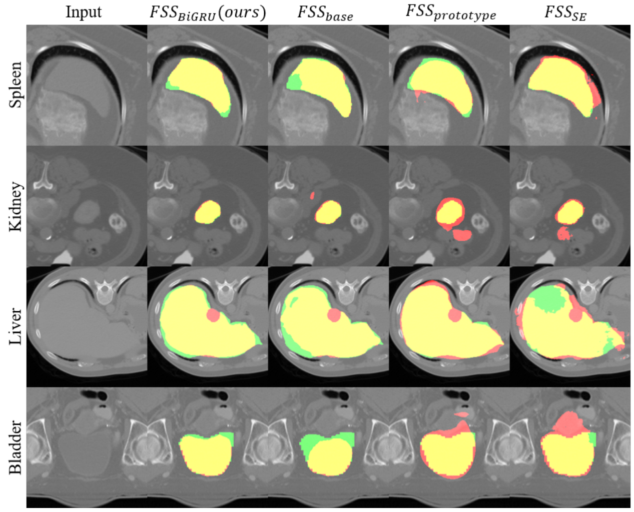

Figure 3: Qualitative results in the axial view. Yellow denotes overlapping region (true positive), while green and red denote false negative and false positive regions, respectively.

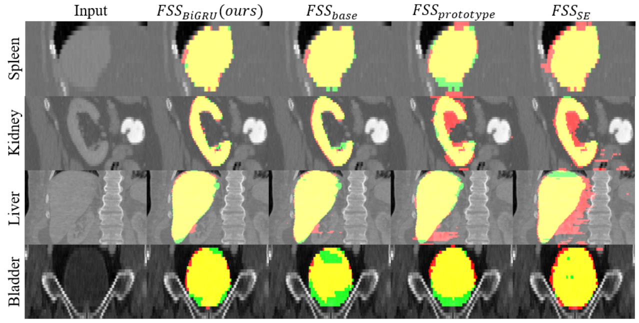

Figure 4: Qualitative results in the coronal and sagittal view. Spleen and kidney are shown in the sagittal view, while liver and bladder are shown in the coronal view.

methods can alleviate this effect by capturing similarities between query and support samples in both training and testing. Notably, our model obtained the comparable performance with the internal test on the two organs in DECATHLON dataset and the bladder in CTORG dataset. Especially, with the transfer learning updates, we obtained improved performance in the DECATHLON dataset i.e. +3% in both organs. This performance was comparable to the upperbound model in DECATHLON dataset.

The performance of our method on the kideny and liver in CTORG dataset was significantly lower that that on the internal test. The performance of supervised learning was good in the case of CTORG dataset since there were relatively many training data. On the other hand, when there is no image with a resolution similar to that of the query image among the few support data, the performance of the few shot learning methods deteriorated. In this sense, the transfer learning strategy was not significant as well if the resolutions of support and query images are inconsistent. However, for this challenging task, the proposed method achieved the best performance among all the few shot learning methods. We believe that better results can be obtained if data with multiple resolutions are contained in the support set.

Qualitative Results Fig.3 and Fig.4 show the qualitative results obtained in the axial view and other views, respectively. In most cases, the proposed method obtained similar segmentation to the ground truth label as opposed to the other few shot methods. Since the comparison methods do not consider information between adjacent slices, segmentation is often not smooth and false positives is obtained like noise in the outer part of the organ. This effect is more pronounced when the appearance of support and query images is different. It was confirmed that the difference in prediction between adjacent slices was relatively large in the sagittal or coronal views compared to the axial view in which the training was performed (see Fig. 4). On the other hand, as the proposed method considers the information between adjacent slices together, the boundary appears smooth even in the sagittal and coronal views.

Conclusion

In this paper, we propose a novel framework for CT organ segmentation under a limited data regime. Our model reliably incorporates multi-slice information to achieve precise segmentation of unseen organs in CT scans. In addition, we show that a bidirectional RNN module can effectively model 3D spatial information for improved feature learning. To learn the optimal parameters for unseen target task, we further introduce a transfer learning strategy. Extensive evaluation on public datasets show the effectiveness and generalization ability of our proposed model. Our method achieved the segmentation performance comparable to the U-Net based supervised learning model on internal and some of external datasets. For future work, we will develop few shot segmentation model that works on different body parts regardless of modality.

Acknowledgments

This work was supported by the grant of the medical device technology development program funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea)(20006006), and the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT)(No. 2019R1C1C1008727).

References

Alom, M. Z.; Yakopcic, C.; Hasan, M.; Taha, T. M.; and Asari, V. K. 2019. Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging* 6(1): 014006.

Bai, W.; Suzuki, H.; Qin, C.; Tarroni, G.; Oktay, O.; Matthews, P. M.; and Rueckert, D. 2018. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *International Conference on Medi-*

cal Image Computing and Computer-Assisted Intervention, 586–594. Springer.

Blaine Rister, Kaushik Shivakumar, T. N.; and Rubin, D. L. 2019. CT-ORG: CT volumes with multiple organ segmentations[Dataset], doi=10.7937/tcia.2019.tt7f4v7o, url=https://wiki.cancerimagingarchive.net/display/Public/CT-ORG\%3A+CT+volumes+with+multiple+organ+segmentations,.

Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; and Van Gool, L. 2017. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 221–230.

Cai, J.; Lu, L.; Xie, Y.; Xing, F.; and Yang, L. 2017. Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function. *arXiv preprint arXiv:1707.04912* .

Cai, J.; Lu, L.; Xing, F.; and Yang, L. 2018. Pancreas segmentation in CT and MRI images via domain specific network designing and recurrent neural contextual learning. *arXiv preprint arXiv:1803.11303* .

Chen, J.; Yang, L.; Zhang, Y.; Alber, M.; and Chen, D. Z. 2016. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In *Advances in neural information processing systems*, 3036–3044.

Dong, N.; and Xing, E. P. 2018. Few-Shot Semantic Segmentation with Prototype Learning. In *BMVC*, volume 3.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400* .

Garcia, V.; and Bruna, J. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043* .

Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging* 38(10): 2281–2292.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hu, T.; Yang, P.; Zhang, C.; Yu, G.; Mu, Y.; and Snoek, C. G. 2019. Attention-based multi-context guiding for few-shot semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8441–8448.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Korez, R.; Likar, B.; Pernuš, F.; and Vrtovec, T. 2016. Model-based segmentation of vertebral bodies from MR images with 3D CNNs. In *International conference on medical image computing and computer-assisted intervention*, 433–441. Springer.

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

Li, H.; Li, J.; Lin, X.; and Qian, X. 2019. MDS-Net: A Model-Driven Stack-Based Fully Convolutional Network for Pancreas Segmentation. *arXiv preprint arXiv:1903.00832* .

Li, X.; Wei, T.; Chen, Y. P.; Tai, Y.-W.; and Tang, C.-K. 2020. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2869–2878.

Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S. J.; and Yang, Y. 2018. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002* .

Liu, Y.; Zhang, X.; Zhang, S.; and He, X. 2020. Part-aware Prototype Network for Few-shot Semantic Segmentation. *arXiv preprint arXiv:2007.06309* .

Luna, M.; and Park, S. H. 2018. 3D patchwise U-Net with transition layers for MR brain segmentation. In *International MICCAI Brainlesion Workshop*, 394–403. Springer.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

Nie, D.; Wang, L.; Xiang, L.; Zhou, S.; Adeli, E.; and Shen, D. 2019. Difficulty-aware attention network with confidence learning for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1085–1092.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Roth, H. R.; Lu, L.; Farag, A.; Shin, H.-C.; Liu, J.; Turkbey, E. B.; and Summers, R. M. 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, 556–564. Springer.

Roy, A. G.; Siddiqui, S.; Pölsterl, S.; Navab, N.; and Wachinger, C. 2020. 'Squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis* 59: 101587.

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410* .

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Simpson, A. L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; Van Ginneken, B.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; et al. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* .

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, 4077–4087.

Sun, Q.; Liu, Y.; Chua, T.-S.; and Schiele, B. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 403–412.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, 9197–9206.

Wang, Y.-X.; Ramanan, D.; and Hebert, M. 2017. Learning to model the tail. In *Advances in Neural Information Processing Systems*, 7029–7039.

Xu, F.; Ma, H.; Sun, J.; Wu, R.; Liu, X.; and Kong, Y. 2019. LSTM Multi-modal UNet for Brain Tumor Segmentation. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, 236–240. IEEE.

Zhang, C.; Lin, G.; Liu, F.; Yao, R.; and Shen, C. 2019. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5217–5226.

Zhao, R.; Chen, W.; and Cao, G. 2019. Edge-Boosted U-Net for 2D Medical Image Segmentation. *IEEE Access* 7: 171214–171222.

Zhou, P.; Yuan, X.; Xu, H.; Yan, S.; and Feng, J. 2019. Efficient meta learning via minibatch proximal update. In *Advances in Neural Information Processing Systems*, 1534–1544.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 3–11. Springer.

Zhu, Z.; Xia, Y.; Shen, W.; Fishman, E. K.; and Yuille, A. L. 2017. A 3d coarse-to-fine framework for automatic pancreas segmentation. *arXiv preprint arXiv:1712.00201* 2.