Block Level Skip Connections Across Cascaded V-Net for Multi-Organ Segmentation

Liang Zhang, *Member, IEEE*, Jiaming Zhang, Peiyi Shen, Guangming Zhu, Ping Li, Xiaoyuan Lu, Huan Zhang, Syed Afaq Shah, *Member, IEEE*, and Mohammed Bennamoun, *Senior Member, IEEE*

***Abstract*— Multi-organ segmentation is a challenging task due to the label imbalance and structural differences between different organs. In this work, we propose an efficient cascaded V-Net model to improve the performance of multi-organ segmentation by establishing dense Block Level Skip Connections (*BLSC*) across cascaded V-Net. Our model can take full advantage of features from the first stage network and make the cascaded structure more efficient. We also combine stacked small and large kernels with an inception-like structure to help our model to learn more patterns, which produces superior results for multi-organ segmentation. In addition, some small organs are commonly occluded by large organs and have unclear boundaries with other surrounding tissues, which makes them hard to be segmented. We therefore first locate the small organs through a multi-class network and crop them randomly with the surrounding region, then segment them with a single-class network. We evaluated our model on SegTHOR 2019 challenge unseen testing set and Multi-Atlas Labeling Beyond the Cranial Vault challenge validation set. Our model has achieved an average dice score gain of 1.62 percents and 3.90 percents compared to traditional cascaded networks on these two datasets, respectively. For hard-to-segment small organs, such as the esophagus in SegTHOR 2019 challenge, our technique has achieved a gain of 5.63 percents on dice score, and four organs in Multi-Atlas Labeling Beyond the Cranial Vault challenge have achieved a gain of 5.27 percents on average dice score.**

***Index Terms*— Multi-organ segmentation, cascaded network, skip connections, inception-like structure, hard-to-segment.**

Manuscript received January 22, 2020; accepted February 13, 2020. Date of publication February 21, 2020; date of current version August 31, 2020. This work was supported in part by the Ningbo 2025 Key Project of Science and Technology Innovation under Grant 2018B10071, in part by the National Key Research and Development Plan under Grant 2019YFB1311600, and in part by the Shanghai Science and Technology Committee under Grant 18411952100 and Grant 17411953500. *(Jiaming Zhang and Huan Zhang are co-first authors.) (Corresponding author: Liang Zhang.)*

Liang Zhang, Jiaming Zhang, Peiyi Shen, and Guangming Zhu are with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: liangzhang@xidian.edu.cn; jmzhang.xidian@outlook.com; pyshen@xidian.edu.cn; gmzhu@xidian.edu.cn).

Ping Li and Xiaoyuan Lu are with the Shanghai BNC, Shanghai 200336, China (e-mail: pli@bnc.org.cn; xylu@bnc.org.cn).

Huan Zhang is with the Ruijin Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai 200070, China (e-mail: huanzhangy@126.com).

Syed Afaq Shah is with the College of Science, Health, Engineering and Education, Murdoch University, Murdoch, WA 6150, Australia (e-mail: afaq.shah@murdoch.edu.au).

Mohammed Bennamoun is with the School of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia (e-mail: mohammed.bennamoun@uwa.edu.au).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.2975347

I. Introduction

ORGAN segmentation in medical images is a crucial task for many applications, such as computer-aided diagnosis (CAD), diagnostic interventions, treatment planning and treatment delivery. The delineation of target organs, which is largely manual and tedious, is essential for radiotherapy planning. Deep learning has achieved significant attention in the recent years. Various deep learning based models, such as FCNs [1], U-Net [2], V-Net [3] and their variants [4]–[13] have achieved an excellent performance on image segmentation. However, accurate multi-organ segmentation is still a challenging task. The shape and location of different organs vary greatly, which requires a robust segmentation network. For some organs in the CT images, even the manual segmentation is challenging due to the low contrast between organs and the surrounding tissues, and the variable morphology. In addition, considering the safety and ethical issues, medical datasets are usually small, which makes it difficult to train a data-hungry deep neural network to segment these organs.

In recent years, few studies have only focused on single-organ segmentation, such as the liver [14]–[16], pancreas [17]–[20], blood vessels [21]–[23], or gliomas [24]–[26]. Single-class segmentation makes it easier for the network to handle specific organs and adopt special strategies [14]–[27]. In addition, single-class segmentation can alleviate the inter-class similarity between organs to a certain extent. It therefore usually performs better than a multi-class network. Considering the problem of extensibility, some studies focus more on the multi-organ segmentation [28]–[33]. In this work, we first validate our model on the multi-organ segmentation tasks, and then train a single-class network with cropped images as input to segment the esophagus, which is a small and hard-to-segment organ.

U-Net and V-Net are the most popular models for medical image segmentation. U-Net is a typical encoder-decoder structure, in which the encoder gradually reduces the spatial dimension of the pooling layer, and the decoder gradually repairs the details and spatial dimensions of the object.

There is usually a skip connection between the encoder and the decoder, which helps the decoder to accurately reconstruct the target. V-Net can be seen as a 3D U-Net combined with the residual network [34]. This architecture ensures convergence in a fraction of the time required by a similar network that does not learn residual functions [3]. In this work, we adopt a cascaded V-Net model as our backbone.

Based on the encoder-decoder structure of U-Net, some approaches apply it to the organ segmentation task in a cascaded fashion [5], [27], [33], [35]–[37]. Cascading networks is a classic and effective way to improve the performance. The most common method is to concatenate the final possibility map of the first stage network with the original image, and feeding the concatenated features to the second stage network. In cascaded networks, the first stage network can also play different roles. For instance, the first stage network in [31] is used to extract multi-scale features, the first stage in [33] is used to locate organs and the first stage in [37] is used to produce a weight map of organs. In general, a cascading network performs a coarse-to-fine segmentation.

However, there is a flaw in existing cascaded networks. They only benefit from the final output possibility map of the first stage network. Considering the structural consistency between the two cascaded networks, we argue that each intermediate layer of the first stage network should also provide useful information to the second stage network. Hence, we propose a novel cascaded network model with **Block Level Skip Connections** (**BLSC**) across cascaded models. This architecture allows the second stage network to capture the features learned by each block of the first stage network, and accelerate the convergence of the second stage network.

Trends in recent studies [34], [38]–[41] show that small convolution kernels with a stacking style are used to obtain the same receptive field as a large convolution kernel. Stacking small convolution kernels can reduce the model size and computation cost. Most importantly, stacking small convolution kernels can make the network deeper and have more nonlinear transformations, which makes the network perform better. In a recent study, Peng et al. proposed Global Convolution Networks (GCN) [42] and adopted large convolution kernels even as large as the size of the feature map. Their work proved that large convolution kernels can enlarge the valid receptive field, which enables the network to handle objects of different sizes. However, in multi-organ segmentation, large kernels tend to extract information from distant voxels, which may be unnecessary for voxels that come from the boundary of organs or small local organs. In order to capture information from distant voxels for organs and to retain the capability of dealing with local small organs, we combine stacked small and large convolution kernels with an inception-like structure.

In the multi-organ segmentation task, some small organs are easily occluded by the large organs. Fig. 1 shows four organs in SegTHOR 2019 and thirteen organs in Multi-Atlas Labeling Beyond the Cranial Vault. In Fig. 1, we can see that the different organs vary greatly in size, shape and location. In SegTHOR 2019, the heart and aorta have a large size and stable shape, which makes them easy to segment. However, from the first row (axial view), we can see that both the esophagus and the trachea have small sizes. From the second row (sagittal view), we can see that their locations vary greatly. Particularly the esophagus, which has a slender shape and varies between patients. In Multi-Atlas Labeling Beyond the Cranial Vault, there are several small organs with complex differences, which makes multi-organ segmentation more challenging. In this case, we crop a region of interest from the entire image for these hard-to-segment organs and train a single-class network for each organ separately. Detailed implementations are provided in section IV.

In summary, the major contributions of this work are as follows:

1) We propose a very efficient cascading network approach. In contrast to the existing cascaded networks, where the second stage network benefits only from the final possibility map of the first stage network, the second stage network in our cascaded network captures the features learned in each block of the first stage network. This significantly improves the performance of the cascaded network.
2) We combine stacked small convolution kernels and large convolution kernels on the block level with an inception-like structure, which enhances the capability of our model to handle different transformations of organs and better grasp information from distant voxels.
3) For some hard-to-segment organs, such as the esophagus, its shape and position vary greatly across patients. We therefore crop it with a proper bounding region and train a single-class network to segment it. This significantly improves the segmentation results and makes the segmentation of these hard-to-segment organs more practical.

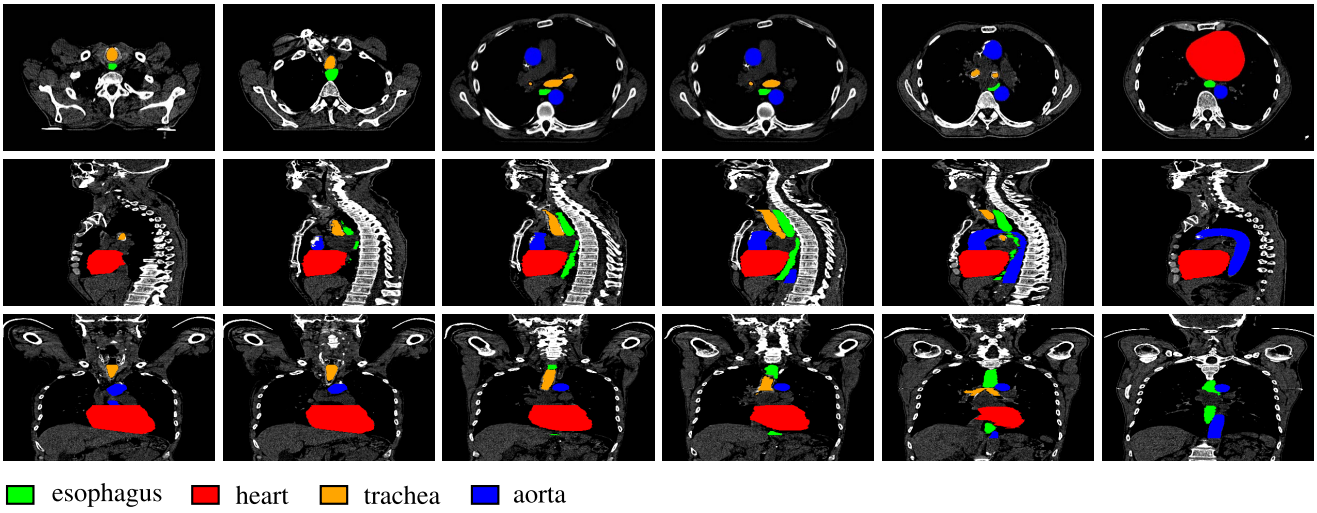
II. Materials

A. Dataset

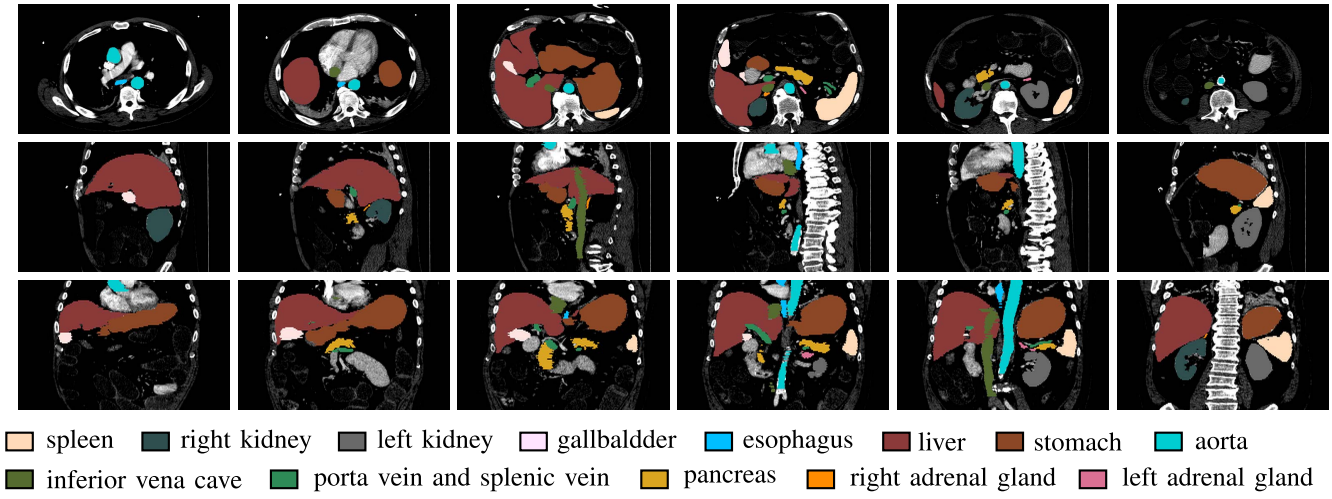
We evaluated our model on two publicly available datasets. The first dataset is from the SegTHOR 2019 challenge [43], and the second dataset is from the Multi-Atlas Labeling Beyond the Cranial Vault challenge. The SegTHOR 2019 challenge and Multi-Atlas Labeling Beyond the Cranial Vault challenge label 4 organs and 13 organs, respectively. The original resolution of the CT images for both datasets are 512×512. For the SegTHOR 2019 challenge, we trained our model on 40 CT scans from the published training set and evaluated our model on 20 CT scans from the unseen testing set online. For the Multi-Atlas Labeling Beyond the Cranial Vault challenge, there are 30 CT scans. The 30 CT scans are randomly divided into six subsets, we evaluated our model on these subsets with k-fold cross-validation and calculated the average results of dice score and Hausdorff distance on the six subsets.

B. Pre-Processing and Post-Processing

Intensity normalization is essential before feeding the image to the network, as this can accelerate the convergence of the network training and avoid pixel value explosion when the network is trained without Batch Normalization [44].



(a) Images of four different organs from SegTHOR 2019 challenge.



(b) Images of thirteen different organs from Multi-Atlas Labeling Beyond the Cranial Vault challenge.

Fig. 1. The axial view (the first row), sagittal view (the second row) and coronal view (the third row) of CT images. The CT images of a patient from SegTHOR 2019 training set and Multi-Atlas Labeling Beyond the Cranial Vault, respectively. The axial view shows the shape of the organ in 2D, the sagittal view shows the location distribution of the organ, the coronal view provides more information in 3D.

We truncated the Hounsfiled Unit (HU) within the range $[-k, k]$, and then the intensity values were linearly normalized into range $[-1, 1]$. The value of k is set to 1000 and 350 in SegTHOR dataset and Multi-Atlas Labeling Beyond the Cranial Vault dataset, which is large enough to cover most pixels of all the organs on each dataset, respectively. After intensity normalization, we scaled the thickness of the CT scans to the same thickness, the standard thickness of SegTHOR 2019 and Multi-Atlas Labeling Beyond the Cranial Vault are 2.5 mm and 3 mm, respectively. In order to reduce the computation cost, the original images were down-sampled from 512×512 to 256×256.

After segmentation, we only post-processed the segmentation results of SegTHOR 2019 challenge. For the esophagus and the heart, we only picked the largest 2D connected regions. For the aorta, we picked no more than two 2D connected regions with size over 150 voxels. For the trachea, we did not perform any processing of the segmentation results due to its unstable structure.

III. Method

In medical image segmentation, V-Net is a popular model that combines the residual networks with U-Net. V-Net encourages a much smoother gradient flow, which helps in the optimization and convergence. As the complexity of the segmentation task increases, in the case of multi-organ segmentation, the original V-Net or U-Net fails to handle various transformations of organs. Cascading two identical networks is a classic and effective approach to improve the capability of the network. In this work, we cascaded two V-Net networks as our backbone. Different from the original V-Net network, we replaced the convolution kernels of size $5 \times 5 \times 5$ with convolution kernels of size $3 \times 3 \times 3$ and added a Group Normalization [45] layer after each block, which helps the V-Net performs better in segmentation. In our backbone, the final output feature map of the first V-Net network is concatenated with the input image of the second V-Net network. Based on the backbone, we propose dense **Block Level Skip Connections (BLSC)** between the corresponding blocks,

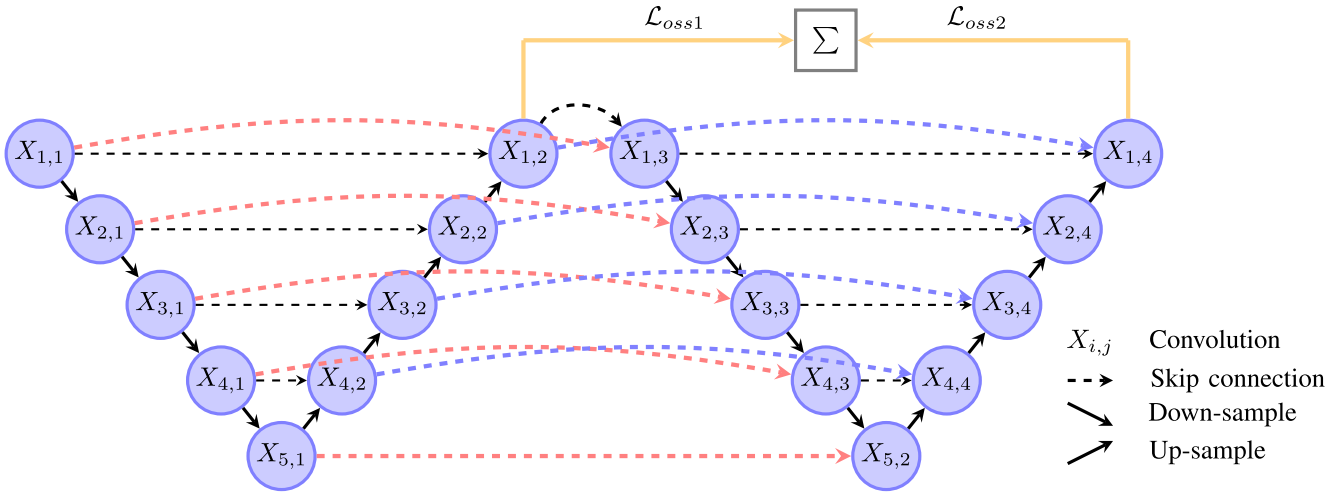


Fig. 2. The structure of our BLSC model. Our backbone consists of a cascaded network with two V-Net networks. The black dashed lines show the skip connections in our backbone, where only the output of first V-Net network is reused between the two V-Net networks. The red dashed lines and blue dashed lines show our block level skip connections between the corresponding encoder blocks and corresponding decoder blocks, respectively. $X_{i,j}$ represents the convolution block of the cascaded network, where i represents the convolution blocks are in the i-th horizontal line and j represents the index of the convolution block in the i-th horizontal line. $X_{i,1}$ and $X_{i,2}$ represent the encoder block and decoder block of the first stage network, respectively. $X_{i,3}$ and $X_{i,4}$ represent the encoder block and decoder block of the second stage network, respectively.

which effectively improves the performance of the cascaded network. Furthermore, we combine stacked small convolution kernels and large convolution kernels, which helps our network to better handle more transformation of different organs and learn more patterns. More details are provided below.

A. Block Level Skip Connections

Cascading networks is an effective way to improve the performance of an overall network. Traditional cascaded networks usually train the first stage network to produce a coarse possibility map and concatenate the possibility map with the original image. The concatenated features are then fed to the second stage network to produce a fine segmentation. The possibility map of the first stage network can be seen as a prior probability of the distribution of organs, which can help the second stage network to focus more on the regions of interest and achieve better results. However, a natural question is why only the last output possibility map of the first stage network is reused? Actually, the two cascaded networks have the same structure. In this work, these cascaded networks are two V-Nets. Therefore, **the features learned by each block of the first V-Net network can complement the second V-Net network.** This observation is the basis of our proposed model. In contrast to the traditional cascaded networks, our proposed model can take full advantage of the features from the first stage network by establishing dense Block Level Skip Connections (BLSC) between the two cascaded networks.

Our proposed model is illustrated in Fig. 2. It shows two V-Net networks cascaded by dense skip connections, where the black dashed lines show the skip connections of the traditional cascaded network, the red and blue dashed lines are our dense block level skip connections. The red dashed lines and blue dashed lines connect the corresponding encoder convolution blocks or decoder convolution blocks of the two cascaded V-Net networks, respectively. The skip connection between convolution block $X_{1,2}$ and convolution block $X_{1,3}$ is the only connection between the two cascaded V-Net networks in the traditional network. This means that the second V-Net network only benefits from the last possibility map of the first V-Net network and most of the information learned by the first V-Net network is lost. With our proposed Block Level Skip Connections, each block of the first V-Net network benefits the second V-Net network, and make the cascaded network more effective. The features from the blocks of the first V-Net network make the corresponding blocks of the second V-net network pay more attention to the regions of interest.

B. Block Level Mixed Convolution Kernels

Early works [3], [39] prefer to use large convolution kernel in neural networks. However, a large convolution kernel usually increases the model size and the computational cost, which is not conducive to building a deeper network. In the later models, such as VGG Net [38] and GoogLeNet [41], competitive results were achieved by stacking 3×3 convolution kernels. Stacking several 3×3 convolution kernels can obtain the same receptive field size as a large convolution kernel, but the model size and the computational cost are greatly reduced. It is generally accepted that stacking small convolution kernels makes the network handle more nonlinear transformations and achieve better results. But should the ideal of a large convolution kernel be completely abandoned? Peng et al proposed Global Convolutional Network (GCN) [42]. Their work proves that large kernels can enhance the capability of the network to handle different transformations and obtain a large visualization of valid receptive fields (VRF), introduced by [46].

In our work, we argue that a large convolution kernel can better grasp the information from distant voxels. Although a large convolution kernel can obtain the same receptive field as stacked small convolution kernels, the voxels at different

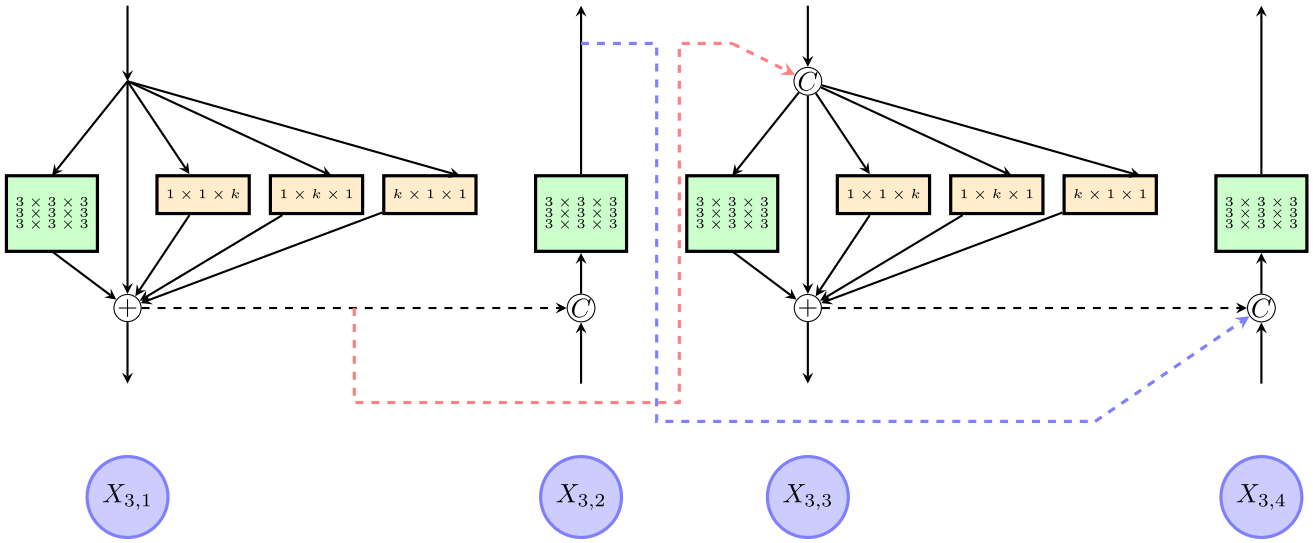


Fig. 3. The structure of the mixed convolution kernel based on R-BLSC model. This figure illustrates the structure of the four blocks in the third level of the horizontal line. The parameter k represents the size of the large convolution kernel. In the block $X_{3,1}$ and block $X_{3,2}$, there are three stacked small convolution kernels with size $3 \times 3 \times 3$, so the receptive field is $7 \times 7 \times 7$. In order to have the same receptive field with stacked small convolution kernel, the value of k is set to 7 in this case. We divided the large convolution kernel with size of $7 \times 7 \times 7$ into three convolution kernels, these three convolution kernels with size of $1 \times 1 \times 7$, $1 \times 7 \times 1$ and $7 \times 1 \times 1$ are in different branches. The branch of the stacked small convolution kernels and three sub-branches of the large convolution kernel are followed by a Group Normalization layer, respectively.

(a) two stacked small kernels (b) a large kernel

Fig. 4. (a) and (b) show how the two stacked small convolution kernels with size 3 and a large convolution kernel with size 5 works in one-dimensional convolution, respectively. In (a), the center voxels is calculated several times, which means that the center voxels will have a higher weight and the information from the boundary voxels or distant voxels becomes weaker.

locations contribute differently to the convolution. Fig. 4 demonstrates how the voxels contribute to the convolution. For the sake of simplicity, we only showed two different ways of convolution in one dimension. We adopted two stacked small convolution kernels with size 3 and a large convolution kernel with size 5 in Fig. 4 (a) and Fig. 4 (b), respectively. From Fig. 4, we can see that both types of convolution kernels obtain a receptive field of size 5. With stacked small convolution kernels, the voxels in the center of the original layer have a greater weight, which means that the distant voxels contribute less to the segmentation. As the number of network layers increases, the contribution of the distant voxels is reduced. However, in the case of a large convolution kernel, all voxels are counted only once. Therefore, we can regard that a large convolution kernel can focus more on the distant voxels compared to the stacked small convolution kernels.

In multi-organ segmentation, the shape and size of different organs or an organ from different patients varies greatly. For voxels from some small local organs or the boundary of the organs, the distant voxels are not key to the classification. Therefore, the large convolution kernels pay more attention to the distant voxels, which are useless and redundant. Given that stacked small convolution kernels can accommodate more linear variations and perform better in capturing the local information, we combined large convolution kernels and stacked small convolution kernels with an inception-like structure.

Fig. 3 shows the structure of our mixed convolution kernels. In each encoder block, we adopt several stacked convolution kernels of size $3 \times 3 \times 3$ and three large convolution kernels in different branches. The three large convolution kernels are separated from a large convolution kernel and only have a large size in one dimension. Given that the encoder blocks play the key role in feature learning, the mixed convolution kernels are not employed in the decoder blocks. We choose a proper large convolution kernel size that can obtain the same size of receptive field as the stacked small convolution kernels. After the convolution, Group Normalization (GN) is used at the end of each branch, then the feature maps of all branches are summed. The dashed lines show how the four blocks are connected, where the red and the blue dashed lines are our proposed block level skip connections.

In 3D V-Net, in order to reduce the model size and the computational cost, we adopt separable large filters. For a large kernel with size of $l \times l \times l$, we employ three $1 \times 1 \times l$, $1 \times l \times 1$ and $l \times 1 \times 1$ convolutions directly instead of the symmetrical separation of the large convolution kernels as in the case of the Global Convolution Network. This strategy makes large kernels more practical. For another, the three convolution kernels are not stacked in one branch, but in different branches, respectively. With this structure, the network can further learn more patterns from the convolution kernels of different shapes, which can help the network handle different transformations of multi-organs.

IV. Experiments

We validated our models on the two datasets by dice score and Hausdorff distance. We also performed repeated tests on each model and repeated measures Analysis of Variance (ANOVA) for the average results. If the value of PR is lower than 0.05 in ANOVA, we believe that the difference between different models is statistically significant.

In section A and B, we introduce the loss function of our experiments and the process of training and testing. In section C and D, we test our BLSC model and further adopt the mixed convolution kernels on it. In section E, we compared our BLSC model with mixed convolution kernels with other works. In section F, we provide a solution to produce better segmentation results on some hard-to-segment organs with single-class network, such as esophagus.

A. Loss Function

In multi-organ segmentation, in order to solve the problem of label imbalance, the commonly used and most popular loss functions are the weighted cross entropy loss or the dice loss. The weighted cross entropy loss function can avoid the loss imbalance between the small organs and the large organs by giving different weights to different organs. The dice loss subtly solves the problem of label imbalance by comparing the similarity between the prediction result and the label. In our work, we adopted the standard dice loss for each cascaded network as below:

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\frac{2\sum_{i}^{N} P_k(i)\, L_k(i)}{\sum_{i}^{N} P_k^2(i) + \sum_{i}^{N} L_k^2(i)} \tag{1}$$

where K is the number of classes, and k represents the k-th class. N is the total number of voxels in the entire image, and i is the i-th voxel. $P_k(i)$ and $L_k(i)$ represent the possibility of prediction and label for class k at voxel i, respectively. The total loss is the weighted average of the dice loss for all stages:

$$\mathcal{L}_{total} = \frac{1}{M}\sum_{i=1}^{M}\lambda_i \mathcal{L}_i \tag{2}$$

where $\mathcal{L}_i$ and λ_i represent the dice loss and the weight of the i-th network in the cascaded networks and M represents the number of cascaded networks. In our experiments, we set $M = 2$ and the weight $\lambda_{i=1,2} = 1$, then train our model with deep supervision.

B. Training and Testing

In the training phase of our experiments, for each CT scan, we choose all the slices which contain organs according to the labels and expand some slices which do not contain organs around both ends. Then, we divide these slices into several blocks of size $256 \times 256 \times 48$ in sequences. There are some overlapping slices between two adjacent blocks, and the number of the overlapping slices is controlled by a random number between 0 and 24. Finally, the blocks containing organs are fed into the network for training.

In the testing phase, we cannot reveal which slices contain organs without labels. We therefore test each CT scan twice for better results. **For the first test**, we divided all slices from each CT scan into some blocks of size $256 \times 256 \times 48$ in sequence and fed them into the network. The predictions of the first test can coarsely locate organs, which means we can coarsely choose the slices with organs based on this result. **In the second test**, we therefore only divided the slices that may contain organs into blocks of size $256 \times 256 \times 48$ and fed them into the network again for a fine segmentation. This method can discard most of the non-relevant information and help the network to focus on regions that are more likely to contain the target organs. Ultimately, it can effectively reduce the risk of false positive segmentation.

C. Block Level Skip Connections

In the traditional cascaded network, the first stage network usually plays the role of coarse segmentation and the second stage network only uses the possibility map of the first stage network. In contrast, the key idea of our cascaded network, with block level skip connections, is to help the second stage network to take full advantage of the features learned by the first stage network. We compared the performance of the traditional cascaded network (baseline) with our BLSC model. For both models, we fed the blocks of size $256 \times 256 \times 48$ to the first stage network, and then concatenated the output of the first stage network with the input blocks and fed that as the input to the second stage network. We trained the cascaded network with deep supervision. This means that the outputs from both stages were taken into account to calculate the dice loss. The final results on SegTHOR 2019 unseen testing set and Multi-Atlas Labeling Beyond the Cranial Vault validation set are reported in Table I.

From Table I, we note that our BLSC model (R-BLSC) improves the dice score by **1.17** (86.40 vs 87.57) percents and **2.80** (74.86 vs 77.66) percents compared to the traditional cascaded network (Baseline) on the two datasets, respectively. Our BLSC model can achieve a great improvement of dice score (76.04 vs 78.14) on the esophagus of SegTHOR 2019, which is the most challenging organ to segment and also weighted high by the challenge organizers. The decline in average Hausdorff distance also supports our results. According to Table II, the value of PR (Baseline vs BLSC (S)) on dice score is lower than 0.05, which proves that our results are statistically significant with a probability more than 0.95. However, the value of PR on Hausdorff distance show that our BLSC model does not gain obvious improvement on Hausdorff distance.

The experiments on the two datasets prove that our BLSC model makes the cascaded structure more effective. After establishing skip connections between the corresponding blocks of the two cascaded networks, the features of each block in the first V-Net network benefits the second V-Net network. The features of the first V-Net network can also be seen as the possibility map of each block, which helps the corresponding block in the second V-Net network to pay more attention to certain areas. With these dense skip connections,

TABLE I

COMPARISON OF THE TRADITIONAL CASCADED NETWORK WITH OUR MODELS. D-BLSC MEANS THERE ARE SKIP CONNECTIONS BETWEEN EACH BLOCK IN THE SAME HORIZONTAL LEVEL, AND R-BLSC MEANS ONLY THE SKIP CONNECTIONS BETWEEN THE CORRESPONDING BLOCKS ARE RETAINED. THE MIXED BLSC REPRESENTS THE BLSC MODEL WITH MIXED CONVOLUTION KERNELS. THE PERFORMANCE IS MEASURED WITH THE DICE SCORE (%) AND HAUSDORFF DISTANCE (mm)

SegTHOR 2019 unseen testing set

model / organ	Dice Score(%)				Hausdorff Distance(mm)			
	Baseline	R-BLSC	D-BLSC	Mixed BLSC	Baseline	R-BLSC	D-BLSC	Mixed BLSC
esophagus	76.04	78.14	78.11	**78.46**	0.6715	0.5569	0.5560	**0.5164**
heart	91.89	92.97	**93.05**	93.03	0.7891	**0.4208**	0.6159	0.6339
trachea	87.08	87.71	87.58	**88.99**	1.7137	2.0440	1.9340	**0.5719**
aorta	90.59	91.45	91.49	**91.61**	0.3364	0.2847	**0.2603**	0.3445
mean	86.40	87.57	87.56	**88.02**	0.8777	0.8266	0.8416	**0.5169**
Multi-Atlas Labeling Beyond the Cranial Vault validation set								
spleen	89.50	89.75	90.26	**91.47**	22.86	21.89	22.10	**21.41**
right kidney	91.34	91.77	**92.35**	91.93	20.25	21.10	19.67	**18.60**
left kidney	88.07	90.78	**92.01**	90.88	22.99	22.32	**20.23**	20.50
gallbladder	59.97	64.01	63.68	**68.18**	18.24	18.01	18.42	**17.86**
esophagus	63.69	68.18	68.63	**69.12**	21.98	22.37	22.85	**21.67**
liver	93.67	94.05	93.60	**94.50**	35.87	36.68	37.55	**35.85**
stomach	76.64	**78.68**	77.44	78.36	38.51	37.91	**37.07**	37.86
aorta	85.25	87.73	86.23	**87.74**	21.46	21.07	**19.98**	20.10
inferior vena cava	83.96	86.08	85.99	**86.54**	23.32	22.84	23.01	**20.98**
portal vein and splenic vein	62.84	68.59	67.49	**68.77**	29.18	27.31	27.52	**26.04**
pancreas	63.29	67.35	66.54	**69.43**	29.89	**29.43**	30.86	30.46
right adrenal gland	59.29	62.95	61.68	**65.14**	15.61	14.69	14.97	**14.49**
left adrenal gland	55.75	59.60	60.54	**61.90**	17.53	16.74	**15.85**	15.91
mean	74.86	77.66	77.42	**78.76**	24.44	24.03	23.85	**23.19**

TABLE II

RESULTS OF REPEATED MEASURES STATISTICAL ANALYSIS OF VARIANCE BETWEEN DIFFERENT MODELS. BLSC (S) IS THE BLSC MODEL WITH STACKED SMALL KERNELS, WHICH IS ALSO THE ORIGINAL BLSC MODEL. BLSC (L) IS THE BLSC MODEL WITH LARGE CONVOLUTION KERNELS, AND BLSC (M) IS THE BLSC MODEL WITH MIXED CONVOLUTION KERNELS. D-BLSC MEANS THERE ARE SKIP CONNECTIONS BETWEEN EACH BLOCK IN THE SAME HORIZONTAL LEVEL, AND R-BLSC MEANS ONLY THE SKIP CONNECTIONS BETWEEN THE CORRESPONDING BLOCKS ARE RETAINED. IF THE VALUE OF PR IS LOWER THAN 0.05, WE CAN BELIEVE THAT THE DIFFERENCE IS STATISTICALLY SIGNIFICANT

SegTHOR 2019 unseen testing set

models / metric	Baseline VS BLSC(S)		Baseline VS BLSC(M)		BLSC(S) VS BLSC(M)		BLSC(L) VS R-BLSC(M)		R-BLSC VS D-BLSC	
	F	PR(>F)	F	PR(>F)	F	PR(>F)	F	PR(>F)	F	PR(>F)
Dice Score	25.117373	0.007429	59.343135	0.001528	13.660266	0.020909	24.310004	0.007870	0.011388	0.920154
Hausdorff Distance	0.063778	0.813068	36.212740	0.003841	2.322864	0.202163	1.829869	0.247552	0.001775	0.968415
Multi-Atlas Labeling Beyond the Cranial Vault validation set										
Dice Score	40.514856	0.003123	77.121127	0.000927	60.662631	0.001466	380.918274	0.000041	3.225887	0.146908
Hausdorff Distance	1.275740	0.321838	14.358037	0.019282	19.474807	0.011570	125.232239	0.000363	0.521459	0.510179

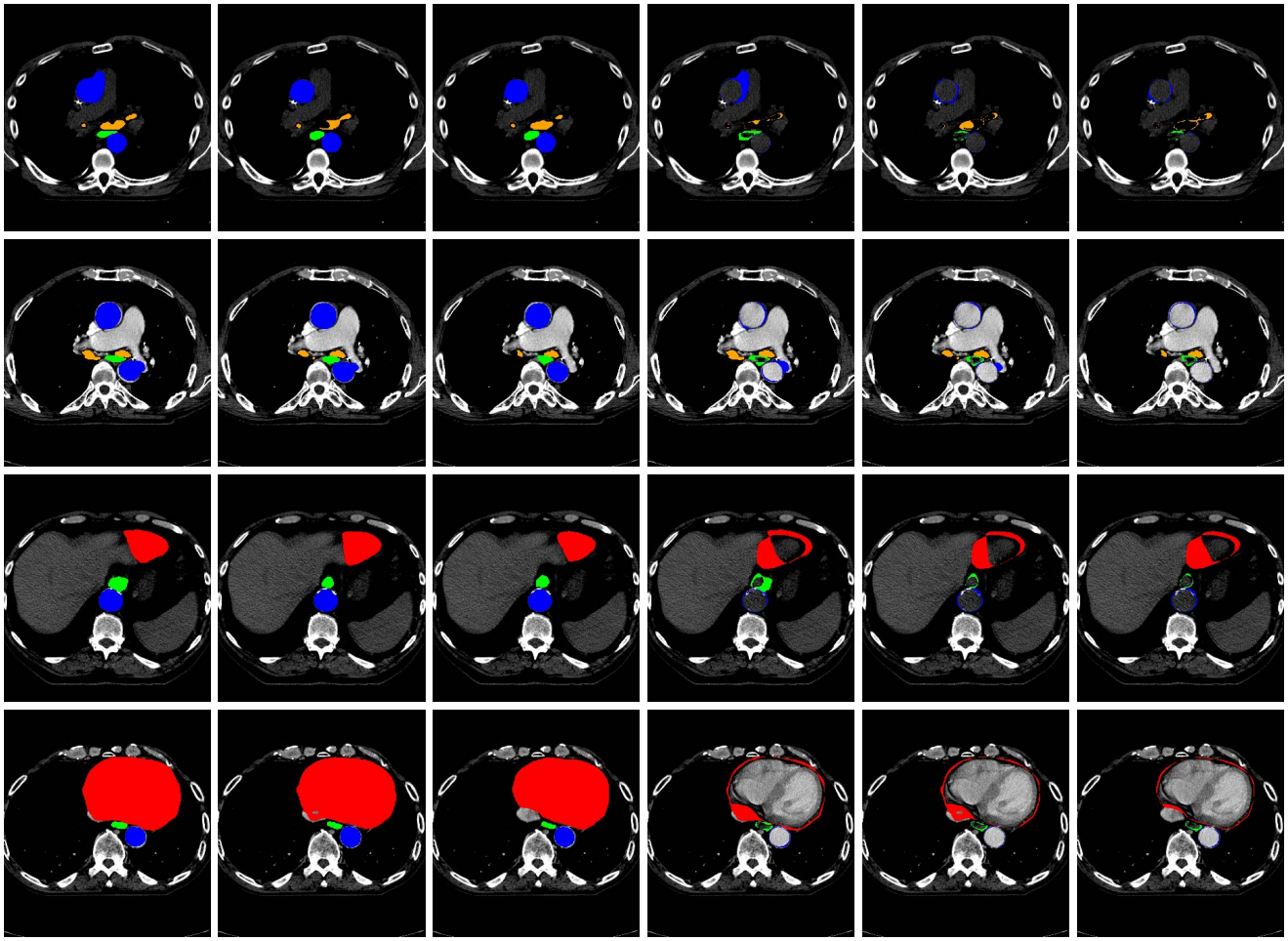
our model can avoid the dispersion of information caused by the depth of the network.

Discussion: The question here arises, why not connect the encoder of the first V-Net network to the decoder of the second V-Net network or the decoder of the first V-Net network to the encoder of the second V-Net network? DenseNet [47] takes full advantage of the features learned by the previous layers by concatenating all of these features with the subsequent layer. In this way, the performance of the network improves. However, this increases the computation cost of the network. In this work, we reduce the computation cost by deleting the redundant skip connections.

Fig. 5 shows the two different types of the proposed skip connections. In Dense BLSC (D-BLSC) model, each block in the same horizontal level is connected with all others through skip connections. In Reduced BLSC (R-BLSC) model, only the skip connections between the corresponding encoders or decoders are retained. The results on the two datasets are also reported in Table I. From Table I and Table II ($PR < 0.05$), we note that the R-BLSC model can achieve competitive results with D-BLSC model, but with a reduced computation cost and a smaller model size. According to these results, we can infer that the skip connections between the corresponding encoders and decoders is the key to improve our proposed cascaded network, and the features from the encoder blocks of the first stage network may bring more confusing information to the decoder of the second stage network.

(a) D-BLSC (b) R-BLSC

Fig. 5. Two ways of connecting different convolution blocks. (a) Illustrates the dense approach like DenseNet (D-BLSC) where each block in the i-th horizontal line is connected to all the others, (b) illustrates the reduced way (R-BLSC) of connecting blocks from the i-th horizontal line. $X_{i,j}$ represents the convolution block of the cascaded network, where i represents the convolution blocks which are in the i-th horizontal line and j represents the index of the convolution block which is in the i-th horizontal line. $X_{i,1}$ and $X_{i,2}$ represent the encoder block and the decoder block of the first stage network, respectively. $X_{i,3}$ and $X_{i,4}$ represent the encoder block and the decoder block of the second stage network, respectively.



(a) Baseline (b) BLSC (c) Mixed (d) Baseline difference (e) BLSC difference (f) Mixed difference

esophagus heart trachea aorta

Fig. 6. Example of the segmentation results of different models on SegTHOR 2019 challenge. The first three columns show the prediction results of different models, the following three columns show the difference between the prediction results and the ground truth. Red, green and blue delineations in the last three columns illustrate where the voxels were incorrectly segmented. The voxels from different organs are marked with different colors.

D. Block Level Mixed Convolution Kernels

Multi-organ segmentation is challenging as the shape or size of different organs varies greatly. We therefore tend to combine large convolution kernels and stacked small convolution kernels to better handle such differences. Compared with the stacked small convolution kernels, large convolution kernels perform better in capturing information from the distant voxels, but it does not definitely lead to an improvement of the segmentation results. For example, for some voxels from the small organs or boundary of the organs, the information from the local voxels has a more important influence than the distant voxels. A large convolution kernel can bring more confusing information from the distant voxels to these voxels, which may not improve or may even deteriorate the performance of the network.

In order to compare the role of the different convolution kernels, we adopted stacked small convolution kernels (original R-BLSC model) and a large convolution kernel based on our BLSC model, respectively. Table III reports the results of our experiment on the two dataset. As can be seen

TABLE III

EFFECT OF DIFFERENT SIZE OF CONVOLUTION KERNELS. LARGE REPRESENTS THE BLSC MODEL WITH LARGE CONVOLUTION KERNELS, AND MIXED REPRESENTS COMBINING THE STACKED CONVOLUTION KERNELS AND THE LARGE CONVOLUTION KERNELS BY SUMMATION. THE PERFORMANCE IS MEASURED WITH THE DICE SCORE (%) AND HAUSDORFF DISTANCE (mm)

SegTHOR 2019 unseen testing set

organ \ model	Dice Score(%)			Hausdorff Distance(mm)		
	small	large	Mixed	small	large	Mixed
esophagus	78.14	77.38	**78.46**	0.5569	0.5793	**0.5164**
heart	92.97	92.78	**93.03**	**0.4208**	0.6822	0.6339
trachea	87.71	87.21	**88.99**	2.0440	2.3367	**0.5719**
aorta	91.45	91.28	**91.61**	**0.2847**	0.2946	0.3445
mean	87.57	87.16	**88.02**	0.8266	0.9732	**0.5169**
Multi-Atlas Labeling Beyond the Cranial Vault validation set						
spleen	89.75	90.02	**91.47**	21.89	22.27	**21.41**
right kidney	91.77	85.88	**91.93**	21.10	24.45	**18.60**
left kidney	90.78	88.97	**90.88**	22.32	22.70	**20.50**
gallbladder	64.01	64.57	**68.18**	18.01	19.72	**17.86**
esophagus	68.18	60.27	**69.12**	22.37	**21.54**	21.67
liver	94.05	92.73	**94.50**	36.68	39.38	**35.85**
stomach	**78.68**	75.88	78.36	37.91	**37.50**	37.86
aorta	87.73	86.87	**87.74**	21.07	20.80	**20.10**
inferior vena cava	86.08	81.68	**86.54**	22.84	22.42	**20.98**
portal vein and splenic vein	68.59	64.39	**68.77**	27.31	27.00	**26.04**
pancreas	67.35	61.93	**69.43**	**29.43**	31.48	30.46
right adrenal gland	62.95	58.91	**65.14**	14.69	14.98	**14.49**
left adrenal gland	59.60	55.38	**61.90**	16.74	17.11	**15.91**
mean	77.66	74.57	**78.76**	24.03	24.72	**23.19**

TABLE IV

DIFFERENT MODELS ON THE TWO DATASETS, THE PERFORMANCE IS MEASURED WITH THE DICE SCORE (%) AND HAUSDORFF DISTANCE (mm)

SegTHOR 2019 unseen testing set

organ \ model	Dice Score(%)					Hausdorff Distance(mm)				
	U-Net[2]	V-Net[3]	Multi-scale[31]	UNet++[4]	ours	U-Net[2]	V-Net[3]	Multi-scale[31]	UNet++[4]	ours
esophagus	76.70	75.69	70.91	77.51	**78.46**	0.6955	0.7393	0.8846	0.5898	**0.5164**
heart	91.96	91.28	92.26	93.31	**93.03**	0.4986	0.9728	0.5004	**0.2540**	0.6339
trachea	87.18	87.31	87.62	**87.32**	88.99	1.1467	0.7440	1.6032	2.2808	**0.5719**
aorta	90.50	90.94	88.88	90.73	**91.61**	0.3185	**0.2718**	0.4273	0.3527	0.3455
mean	86.59	86.26	84.92	87.22	**88.02**	0.6648	0.6820	0.8539	0.8693	**0.5169**
Multi-Atlas Labeling Beyond the Cranial Vault validation set										
spleen	**91.61**	88.47	90.82	88.62	91.47	22.91	24.79	22.92	21.88	**21.41**
right kidney	90.79	90.28	87.02	90.50	**91.93**	19.60	20.14	23.36	21.72	**18.60**
left kidney	88.72	88.11	85.96	**90.94**	90.88	22.03	22.51	24.51	22.95	**20.50**
gallbladder	61.72	62.04	54.26	59.63	**68.18**	19.40	20.10	18.44	19.57	**17.86**
esophagus	62.56	62.42	58.56	64.94	**69.12**	22.90	22.96	22.58	21.92	**21.67**
liver	94.39	92.57	92.38	93.50	**94.50**	35.85	38.18	39.39	36.35	**35.85**
stomach	76.51	74.08	74.77	76.75	**78.36**	38.31	37.36	**37.06**	39.66	37.66
aorta	83.21	83.63	83.66	85.27	**87.74**	20.60	22.36	**18.39**	20.03	20.10
inferior vena cava	82.03	82.42	78.18	83.15	**86.54**	23.86	24.29	23.14	23.42	**20.98**
portal vein and splenic vein	64.02	63.22	62.60	66.00	**68.77**	26.86	29.45	29.42	29.02	**26.04**
pancreas	59.96	62.51	57.21	68.46	**69.43**	**28.65**	30.98	30.15	31.59	30.46
right adrenal gland	61.73	57.38	51.75	62.33	**65.14**	14.99	15.45	17.39	**13.99**	14.49
left adrenal gland	58.28	58.27	50.30	61.25	**61.90**	17.17	16.72	18.19	16.17	**15.91**
mean	75.04	74.26	71.34	76.26	**78.76**	24.09	25.02	25.00	24.48	**23.19**

in Table III, if only one type of convolution kernel is used, the stacked small convolution kernels perform better than the large convolution kernels. After combining stacked small convolution kernels and large kernels with an inception-like structure, our model can gain further improvement. This experiment proves that the mixed convolution kernels perform better on organs that need information from distant voxels. The value of PR (< 0.05) in Table II also proves that the difference is statistically significant.

For another, as can be seen in Table II and Table III, our BLSC model with mixed convolution kernels performs better on the second dataset, which has more organs and is therefore more challenging. This can be attributed to our mixed convolution kernels, which helps the network to learn more patterns.

TABLE V

RESULTS OF REPEATED MEASURES STATISTICAL ANALYSIS OF VARIANCE BETWEEN PREVIOUS MODELS AND OUR MODEL. IF THE VALUE OF PR IS LOWER THAN 0.05, WE CAN BELIEVE THAT THE RESULTS ARE STATISTICALLY SIGNIFICANT

metric \ models	U-Net VS ours		V-Net VS ours		Multi-scale VS ours		UNet++ VS ours	
	F	PR(>F)	F	PR(>F)	F	PR(>F)	F	PR(>F)
SegTHOR 2019 unseen testing set								
Dice Score	25.117373	0.007429	59.343135	0.001528	13.660266	0.020909	24.310004	0.007870
Hausdorff Distance	0.063778	0.813068	36.212740	0.003841	2.322864	0.202163	1.829869	0.247552
Multi-Atlas Labeling Beyond the Cranial Vault validation set								
Dice Score	40.514856	0.003123	77.121127	0.000927	60.662631	0.001466	380.918274	0.000041
Hausdorff Distance	1.275740	0.321838	14.358037	0.019282	19.474807	0.011570	125.232239	0.000363

TABLE VI

DICE SCORE (%) OF DIFFERENT BATCH SIZES IN THE CASE OF SINGLE-CLASS NETWORK ON THE ESOPHAGUS OF SEGTHOR 2019 UNSEEN TESTING SET. RESULTS OF REPEATED MEASURES STATISTICAL ANALYSIS OF VARIANCE BETWEEN MULTI-CLASS NETWORK WITH BATCH SIZE OF 1 AND SINGLE-CLASS NETWORK WITH DIFFERENT BATCH SIZE IS ALSO REPORTED IN THIS TABLE. IF THE VALUE OF PR IS LOWER THAN 0.05, WE CAN BELIEVE THAT THE RESULTS ARE STATISTICALLY SIGNIFICANT

metric \ batch size	Multi-class	Single-class			
	1	1	2	3	4
Dice Score(%)	78.46	**84.09**	80.57	79.94	79.32
F	-	336.612655	43.553581	44.320725	2.897793
PR(>F)	-	0.000052	0.002731	0.002644	0.163914
Hausdorff Distance(mm)	0.5164	**0.3954**	0.4554	0.4926	0.5050
F	-	506.836132	14.618783	32.425646	12.864496
PR(>F)	-	0.000023	0.018723	0.004699	0.023029

E. Comparison With Previous Works

We compared our model with previous works on the two datasets. Among these models, U-Net and V-Net are the classic models for medical image segmentation. Multi-scale pyramid [31] is a cascaded network with two FCN networks, but the image resolution in the two stages is different. For fair comparison, we replaced the two FCN networks with two V-Net networks. We down-sampled the image resolution to 128×128 and 256×256, and fed them to the first V-Net network and the second V-Net network, respectively. We also compared our model with UNet++ [4], which is a single-stage network and reused the features from different decoders in one U-Net network. UNet++ up-samples the feature map of each encoder block to the original resolution gradually and establishes dense skip connections between the blocks of the same horizontal line. The final results are reported in Table IV. We can see that our models can gain greatly improvement on dice score and Hausdorff distance. The value of PR (<0.05) in Table V can also prove that the results are statistically significant.

F. Single-Class Segmentation for Hard-to-Segment Organs

In the case of the multi-organ segmentation, the segmentation of the small organs is still challenging due to the small number of voxels in a single image. The esophagus from SegTHOR 2019 is a hard-to-segment organ due to its slender shape, small size and it varies greatly from patient to patient. Considering the difficulty of the esophagus segmentation, the esophagus is weighted high by the challenge organizers. In this experiments, we will discuss a series of strategies to achieve a better esophagus segmentation result. Finally, the average dice score of the esophagus improves from 0.7846 to 0.8409.

Our overall pipeline is as follows:

1) We roughly locate the esophagus through the previous multi-class segmentation results.
2) We randomly crop the area of the esophagus with a proper size.
3) We feed the cropped area to our model for a finer segmentation.

Cropping strategy: As shown in Fig. 1, while the esophagus only occupies a small portion of a slice, its position is not fixed across different slices. We crop out a block of size $112 \times 112 \times 112$ around the esophagus, which can completely cover the esophagus. We therefore can feed the entire esophagus into the 3D network at once. In order to increase the diversity of the training data, we randomly cropped a 112×112 slice area in a larger area of size 224×224. A cropping range twice the block size performs well in our experiments. A smaller cropped area may lead to over-fitting which requires more attention. In the test phase, we crop a block of size $112 \times 112 \times 112$ according to the coarse position of the esophagus that is obtained from the previous multi-class segmentation.

Batch Size: Due to the limitation of GPU memory, the batch size in the 3D multi-class network is set to 1. Therefore batch normalization is generally not used. In various models, a large and appropriate batch size is usually chosen empirically.

TABLE VII
PERFORMANCE OF THE MULTI-CLASS NETWORK AND THE SINGLE CLASS NETWORK WITH OUR PROPOSED STRATEGIES ON MULTI-ATLAS LABELING BEYOND THE CRANIAL VAULT DATASET. THE PERFORMANCE IS MEASURED WITH THE DICE SCORE (%) AND HAUSDORFF DISTANCE (mm). RESULTS OF REPEATED MEASURES STATISTICAL ANALYSIS OF VARIANCE BETWEEN TWO MODELS IS ALSO REPORTED IN THIS TABLE. IF THE VALUE OF PR IS LOWER THAN 0.05, WE CAN BELIEVE THAT THE RESULTS ARE STATISTICALLY SIGNIFICANT

organ	Dice Score(%)				Hausdorff Distance(mm)			
	Multi-class	Single-class	F	PR(>F)	Multi-class	Single-class	F	PR(>F)
portal vein and splenic vein	68.77	**72.53**	38.515602	0.003429	26.04	**22.24**	19.457966	0.011591
pancreas	69.43	**72.66**	11.411258	0.027835	30.46	**28.41**	8.182969	0.045902
right adrenal gland	65.14	**73.15**	56.365723	0.001684	14.49	**11.83**	26.254815	0.006867
left adrenal gland	61.90	**67.96**	14.589987	0.018783	15.91	**13.79**	17.670582	0.013656
mean	66.31	**71.58**	87.654494	0.000725	21.73	**19.07**	47.683323	0.002307

However, in medical image processing, there is a serious problem of data scarcity which results in small size medical datasets. For example, in these experiments, our training set has only 40 CT scans. Since the esophagus varies greatly across patients, the training set does not simulate the true distribution of the esophagus very well. Therefore, when the batch size is increased, the model is more easily over-fitted to the training set, which reduces the generalization of the model. We adopted different batch sizes to train our network. The results are shown in Table VI, which shows that when the batch size increases from 1 to 4, the dice score of the esophagus decreases. When the batch size is 1, the network achieves the best performance both on dice score and Hausdorff distance. We also performed repeated measures statistical analysis of variance between multi-class network with batch size of 1 and single-class network with different batch size, the results are also reported in Table VI. According to the value of the PR (<0.05) in Table VI, we believe that this results are statistically significant.

As a typical hard-to-segment organ, the segmentation strategies for the esophagus can also be used to segment other similar organs. We therefore also employed this method to four hard-to-segment organs in Multi-Atlas Labeling Beyond the Cranial Vault dataset. The results are reported in Table VII. From Table VII, we note that single-class networks with our strategies on dice score and Hausdorff distance performs much better than multi-class network. According to the value of PR (<0.05), we believe that the difference is statistically significant.

V. CONCLUSION

Organ segmentation is essential for the organ disease diagnosis and radiotherapy planning. A multi-organ system based on deep learning is expected to replace the tedious manual annotation and can be applied in computer-aided diagnosis. In this work, we proposed an efficient approach to boost the performance of cascaded network for multi-organ segmentation task. The key idea is to establish skip connections between the corresponding blocks of two cascaded networks. These skip connections can effectively pass the features learned in each block of the first network to the second cascaded network. Our experiments proved that our cascaded network, with block level skip connections, performs much better compared to the traditional cascaded network. Furthermore, we explored the combination of stacked small convolution kernels with several large kernels separated from a large convolution kernel to help the network to better capture the information from distant voxels and learn more patterns, which greatly improves the performance of multi-organ segmentation.

We also provided a solution for single-class network to segment hard-to-segment organs, which are small and vary greatly in shape across different patients. The proposed technique can also be used to segment other similar organs and make the segmentation of small organs more practical. However, training single-class networks for separate organs increases the model size and requires fine processing, which is not conducive to the extensibility of the multi-organ segmentation model. In addition, single-class network may not produce obvious improvement on large organs with good performance, which requires attention.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2015, pp. 234–241.

[3] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. New York, NY, USA: Springer, 2018, pp. 3–11.

[5] P. F. Christ *et al.*, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2016, pp. 415–423.

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, Oct. 2016, pp. 424–432.

[7] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019.

[8] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: Separable 3D U-Net for brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*. New York, NY, USA: Springer, 2018, pp. 358–368.

[9] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[10] O. Oktay *et al.*, "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: http://arxiv.org/abs/1804.03999

[11] B. Kayalibay, G. Jensen, and P. van der Smagt, "CNN-based segmentation of medical imaging data," 2017, *arXiv:1701.03056*. [Online]. Available: http://arxiv.org/abs/1701.03056

[12] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze and excitation' in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2018, pp. 421–429.

[13] Y. Qin *et al.*, "Autofocus layer for semantic segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2018, pp. 603–611.

[14] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2016, pp. 149–157.

[15] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 2, pp. 171–182, Sep. 2016.

[16] P. Hu, F. Wu, J. Peng, P. Liang, and D. Kong, "Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution," *Phys. Med. Biol.*, vol. 61, no. 24, pp. 8676–8698, Nov. 2016.

[17] H. R. Roth *et al.*, "DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2015, pp. 556–564.

[18] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal CT scans," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2017, pp. 693–701.

[19] J. Cai, L. Lu, Z. Zhang, F. Xing, L. Yang, and Q. Yin, "Pancreas segmentation in MRI using graph-based decision fusion on convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, Oct. 2016, pp. 442–450.

[20] M. Oda *et al.*, "Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, Oct. 2016, pp. 556–563.

[21] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Nov. 2016.

[22] J. I. Orlando, E. Prokofyeva, and M. B. Blaschko, "A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 16–27, Jan. 2017.

[23] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," *Neurocomputing*, vol. 149, pp. 708–717, Feb. 2015.

[24] D. Kwon, R. T. Shinohara, H. Akbari, and C. Davatzikos, "Combining generative models for multifocal glioma segmentation and registration," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2014, pp. 763–770.

[25] A. Gooya *et al.*, "GLISTR: Glioma image segmentation and registration," *IEEE Trans. Med. Imag.*, vol. 31, no. 10, pp. 1941–1954, Oct. 2012.

[26] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI," in *BrainLes*. New York, NY, USA: Springer, 2015, pp. 131–143.

[27] S. Li, Y. Chen, S. Yang, and W. Luo, "Cascade dense-unet for prostate segmentation in MR images," in *Proc. Int. Conf. Intell. Comput.* New York, NY, USA: Springer, 2019, pp. 481–490.

[28] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Trans. Med. Imag.*, vol. 32, no. 9, pp. 1723–1730, Sep. 2013.

[29] T. Okada, K. Yokota, M. Hori, M. Nakamoto, H. Nakamura, and Y. Sato, "Construction of hierarchical multi-organ statistical atlases and their application to multi-organ segmentation from CT images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2008, pp. 502–509.

[30] Y. Wang, Y. Zhou, P. Tang, W. Shen, E. K. Fishman, and A. L. Yuille, "Training multi-organ segmentation networks with sample selection by relaxed upper confident bound," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2018, pp. 434–442.

[31] H. R. Roth *et al.*, "A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2018, pp. 417–425.

[32] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation from CT images using conditional shape-location and unsupervised intensity priors," *Med. Image Anal.*, vol. 26, no. 1, pp. 1–18, Dec. 2015.

[33] H. R. Roth *et al.*, "An application of cascaded 3D fully convolutional networks for medical image segmentation," *Comput. Med. Imag. Graph.*, vol. 66, pp. 90–99, Jun. 2018.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] D. Nie *et al.*, "Segmentation of craniomaxillofacial bony structures from MRI with a 3D deep-learning based cascade framework," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* New York, NY, USA: Springer, 2017, pp. 266–273.

[36] D. Lachinov, E. Vasiliev, and V. Turlapov, "Glioma segmentation with cascaded UNet," in *Proc. Int. MICCAI Brainlesion Workshop*. New York, NY, USA: Springer, 2018, pp. 189–198.

[37] G. Aresta *et al.*, "iW-Net: An automatic and minimalistic interactive lung nodule segmentation deep network," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Aug. 2019.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[41] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[42] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4361–4535.

[43] R. Trullo, C. Petitjean, S. Ruan, B. Dubray, D. Nie, and D. Shen, "Segmentation of organs at risk in thoracic CT images using a SharpMask architecture and conditional random fields," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 1003–1006.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[45] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–9.

[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," Dec. 2014, *arXiv:1412.6856*. [Online]. Available: https://arxiv.org/abs/1412.6856

[47] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.