RESEARCH ARTICLE

Multi-scale U-like network with attention mechanism for automatic pancreas segmentation

Yingjing Yan^{*}, Defu Zhang

School of Informatics, Xiamen University, Xiamen, Fujian, China

^{*} 23020181154286@stu.xmu.edu.cn

Abstract

In recent years, the rapid development of deep neural networks has made great progress in automatic organ segmentation from abdominal CT scans. However, automatic segmentation for small organs (e.g., the pancreas) is still a challenging task. As an inconspicuous and small organ in the abdomen, the pancreas has a high degree of anatomical variability and is indistinguishable from the surrounding organs and tissues, which usually leads to a very vague boundary. Therefore, the accuracy of pancreatic segmentation is sometimes below satisfaction. In this paper, we propose a 2.5D U-net with an attention mechanism. The proposed network includes 2D convolutional layers and 3D convolutional layers, which means that it requires less computational resources than 3D segmentation models while it can capture more spatial information along the third dimension than 2D segmentation models. Then we use a cascaded framework to increase the accuracy of segmentation results. We evaluate our network on the NIH pancreas dataset and measure the segmentation accuracy by the Dice similarity coefficient (DSC). Experimental results demonstrate a better performance compared with state-of-the-art methods.

OPEN ACCESS

Citation: Yan Y, Zhang D (2021) Multi-scale U-like network with attention mechanism for automatic pancreas segmentation. PLoS ONE 16(5): e0252287. <https://doi.org/10.1371/journal.pone.0252287>

Editor: Jianjun Hu, University of South Carolina, UNITED STATES

Received: February 23, 2021

Accepted: May 13, 2021

Published: May 27, 2021

Copyright: © 2021 Yan, Zhang. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All dataset is available from <https://academictorrents.com/details/80ecfcbabede760cdbcdf63e38986501f7becd49>.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Segmenting organs such as the spleen, liver, and pancreas from abdominal CT scans is a critical prerequisite of computer-aided diagnosis (CAD) [1, 2], quantitative and qualitative analysis. In this paper, we are committed to automatic pancreas segmentation, which is an extensively useful CAD technique for the diagnosis and prognosis of pancreatic cancer. So far, automatic segmentation for other organs (e.g., lung, kidney and liver) has reached relatively a high accuracy. However, the pancreas has irregular shapes, great individual differences and vague boundary, while occupying only a very small part (e.g., less than 1%) of the whole CT volumes, which results in the difficulty of segmentation with high accuracy [3, 4].

In recent years, with the fast development of deep learning, natural image semantic models are widely used to complete medical image segmentation tasks, such as full convolution network (FCN) [5] and U-net [6]. Through skip connection, U-net can extract more features from the target tissue and obtain more accurate information than FCN while fusing features in the

Fig 1. Slices from three different viewpoints. Pancreas is highlighted in red and is a small organ with vague boundary and irregular shape.

<https://doi.org/10.1371/journal.pone.0252287.g001>

decoder part. Integrating deep supervision into the network can avoid the risk of gradient vanish in the training process and can effectively improve the performance of the model [7, 8]. Among these complex organ segmentation tasks, pancreas segmentation is one of the most challenging tasks in medical image segmentation [9].

In general, we can divide pancreas segmentation methods into two categories, namely methods based on 2D networks and methods based on 3D networks. 3D networks (e.g., 3D U-net [10] and V-net [11]) use the entire CT volume as the input of the network, which can capture 3D spatial information of the CT volume. However, the high requisition of computing resources, especially GPU memory, restricts the depth of the networks and the number of the feature maps, which are two significant factors for performance improvement [12]. In addition, 3D networks are also faced with the problem that there are only a few CT scans available for training. The main reason is that manual labeling requires a lot of time and labor costs. 2D networks use slices by cutting along three axes (the coronal, sagittal and axial views, see Fig 1) from the CT volume. Each slice from three viewpoints is processed separately in the network and prediction results of these viewpoints are fused via majority vote to construct a 3D pancreas volume [13, 14]. However, the pancreas merely occupies less than 2% of the entire CT volume, and its boundary with surrounding tissues and organs is blurred, which means the network may be confused by the background region. To alleviate this problem, a two-stage framework is taken into consideration. In order to reduce the interference of the background region, Zhou et al. [15] proposed a coarse-to-fine framework. Firstly, a 2D FCN is used to segment the slices roughly. Then, according to the results of the coarse segmentation, the pancreatic region is located and the CT volume is clipped to get a smaller block containing the pancreatic region, which is then inputted into another 2D FCN for fine segmentation. Cai et al. [16] presented an end-to-end stacked CNN-RNN segmentation model. The training process is to first train a 2D convolution neural network (CNN) to segment multi-layer adjacent pancreas regions and then the segmentation results are input into a recurrent neural network (RNN). RNN further obtains accurate results by integrating the information of its adjacent layers. Yu et al. [17] used two cascaded 2D FCN to segment the pancreas. First a FCN is used to

acquire a coarse segmentation probability map. Then the probability map is transformed to a spatial weight map by a saliency conversion module, and multiplied with the input slice to get a slice containing the coarse segmentation information. According to the coarse segmentation probability map, the entire CT volume is cropped into a smaller volume containing the pancreatic region. Finally, the loss function of the coarse and fine segmentation is optimized.

With the help of attention mechanism, the network can focus on the most relevant features without extra supervision. Therefore, many computer vision tasks have integrated attention mechanism into CNN, including emotion recognition [18], object detection [19], human pose estimation [20] and semantic segmentation [21]. Dual attention network (DANet) [22] applies attention mechanism to both spatial dimension and channel dimension. The output of these two attention modules is fused to get more robust feature expression, so as to obtain more accurate segmentation results. To generate dense and pixel-wise contextual information, these attention modules have to aggregate contextual information of each pixel from all pixels, which will generate a huge feature map, and result in high computational complexity and high occupation of GPU memory. To address these problems, Huang et al. [23] proposed a new cross attention module. The cross attention module merely aggregates contextual information for each position from pixel in horizontal and vertical directions. It can obtain the contextual information from all pixels through two continuous cross attention modules. Compared with the attention modules mentioned above, cross attention module has high computational efficiency and occupies less GPU memory, while achieving outstanding performance. Medical image segmentation tasks also benefit from attention mechanisms. For example, Oktay et al. [24] proposed attention U-net, which can easily integrate attention gates into U-net model with increasing minimal computational resources while improving the segmentation performance.

In this paper, inspired by the excellent performance of U-net and its variants in medical image segmentation, we propose a 2.5D U-like architecture, and apply a dense convolution block with skip connection to replace convolution operation in the encoder and the decoder. Then we integrate a hybrid attention module containing a spatial attention module and a channel attention module into the network. Finally, due to the high variability in shapes, size and location of the pancreas, we follow Zhou et al. [15] to use a coarse-to-fine framework for pancreas segmentation and apply the proposed network to both coarse and fine phase. In summary, our main contributions are two-fold:

- We propose a novel attention module, which contains a spatial attention module and a channel attention module. The output of these two modules is aggregated to obtain a more robust feature expression.
- We propose a 2.5D network, which combines 2D convolutional layers with 3D convolutional layers and uses 3 adjacent slices to form a 3-channel input image, so that our network can capture the inter-slice information compared with 2D models and needs less computational resources than 3D models. The proposed approach achieves state-of-the-art performance on the NIH pancreas segmentation dataset [9].

Materials and methods

Problem definition

In this section, we describe the pancreas segmentation problem as follows. Let $X \in \mathbb{R}^{n \times w \times h \times c}$ represent slices by cutting along the coronal, sagittal or axial plane, where n is the number of

slices containing pancreatic region, h and w represent the width and height of slices, and c is the number of channels. Let the binary segmentation mask $Y \in I^{n \times w \times h \times c}$ be the annotation of X , and it can be defined as follows:

$$Y_{n,w,h,c} = \begin{cases} 1, & X_{n,w,h,c} \in \text{pancreas} \\ 0, & X_{n,w,h,c} \notin \text{pancreas} \end{cases} \quad (1)$$

The voxel belongs to the pancreatic region while the value of its corresponding position in Y is equal to 1, otherwise the voxel belongs to the background. The pancreas segmentation task is to learn a pixel-wise segmentation model Seg to obtain the segmentation map $\text{Seg}(X)$ of each input image, and our goal is to minimize the difference between the ground truth Y and the prediction map $\hat{Y} = \text{Seg}(X; \theta)$, where θ indicates the parameters of the model. To evaluate the similarity between the ground truth Y and the segmentation map $\hat{Y}$, we follow [11] to use Dice Similarity Coefficient (DSC), which is widely applied to medical image segmentation tasks. DSC is defined as follows:

$$\text{DSC} = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (2)$$

where A and B are two voxel sets. We slightly modify the DSC into a loss function $L(Y, \hat{Y})$, and the segmentation task can be defined as follows:

$$L(Y, \hat{Y}) = 1 - \frac{2 \times \sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i} \quad (3)$$

where y_i and $\hat{y}_i$ are the i^{th} pixels of the segmentation map Y and the ground truth $\hat{Y}$, and the value of these pixels are 0 either 1.

Coarse-to-fine framework

The pancreas is an inconspicuous and extremely small organs of irregular shape, which usually occupies less than 2% of the entire CT volume. Furthermore, the pancreas has a high degree of anatomical variability and is indistinguishable from the surrounding organs and tissues, which results in a blurry boundary with other organs and tissues. Therefore, deep neural networks are easily confused by the surrounding organs and the model will bias towards the background. For example, FCN and U-net, which are widely used in semantic segmentation and have an outstanding performance in medical image segmentation, sometimes demonstrate unsatisfactory results for pancreas segmentation task. A smaller input image containing the pancreatic region can efficiently alleviate this problem, see Fig 2.

Inspired by this, we follow Zhou et al. [15] to apply a coarse-to-fine framework (see Fig 3), which crop the entire input image into a smaller input region for the fine stage, according to the prediction masks of the coarse stage. We cut a 3D volume into a set of slices along three axes, *i.e.*, the sagittal, axial and coronal views. Then We train two sets of models (each set includes three models, M_s , M_a and M_c for three views) for coarse and fine phase. We fuse these prediction masks of three different views into a 3D binary volume via majority vote. The fusion function $r(\hat{Y}_s, \hat{Y}_a, \hat{Y}_c)$ can be defined as follows:

$$\hat{Y} = r(\hat{Y}_s, \hat{Y}_a, \hat{Y}_c) = \frac{\hat{Y}_s + \hat{Y}_a + \hat{Y}_c}{3} \quad (4)$$

Fig 2. Illustration of the segmentation results of two different input regions. A smaller input region improves the segmentation performance.

<https://doi.org/10.1371/journal.pone.0252287.g002>

where $\hat{Y}_s$, $\hat{Y}_a$ and $\hat{Y}_c$ are the segmentation masks of the sagittal, axial and coronal viewpoints, and $\hat{Y}$ is the fusion result.

We add a frame around the coarse segmentation results, which is filled with the origin image, to make the network more robust and avoid over-fitting, then crop the volume into a smaller block.

Fig 3. Illustration of the coarse-to-fine framework. The framework segments the pancreas through coronal, sagittal and axial viewpoints and fuses the results into a new segmentation volume.

<https://doi.org/10.1371/journal.pone.0252287.g003>

Fig 4. Illustration of the proposed multi-scale attention net. Dense block is shown in Fig 5 and the proposed attention module is shown in Fig 6.

<https://doi.org/10.1371/journal.pone.0252287.g004>

Our approach

In this section, we introduce a novel 2.5D segmentation network in detail. Since U-net has been widely used for medical image segmentation and has achieved outstanding performance, we design a 2.5D U-net architecture (see Fig 4) and integrate a hybrid attention module into it. Currently, 3D segmentation networks need more computational resources and GPU memory, while 2D segmentation models are smaller and less accurate because 2D networks cannot capture spatial information along the third dimension. Therefore, we combine 2D convolutional layers with 3D convolutional layers and use 3 adjacent slices to form a 3-channel input image, so that our approach can extract inter-slice features from these adjacent slices and needs less computational resources than 3D networks. The input image of our approach can be simply described as follows:

$$I_i = \begin{cases} (S_i, S_i, S_i), & i = 0 \text{ or } i = n - 1 \\ (S_{i-1}, S_i, S_{i+1}), & \text{otherwise} \end{cases} \quad (5)$$

where S_i represents the i -th slice, and n is the number of slices.

Similar to U-net, the proposed architecture includes an encoder structure and a decoder structure, and both of them are composed of 4 blocks. The difference is that for the proposed network, the first two layers of the encoder structure are 3D convolutional layers and the remaining layers are 2D convolutional layers. We use denser convolution layers (see Fig 5) to obtain a larger receptive field, because a larger receptive field is helpful to extract effective features for large object segmentation, while after coarse segmentation the pancreas is relatively a large object compared with the background region. However, denser convolution layers bring about the risk of network degradation, gradient vanish and gradient explosion [25], so we introduce skip connections to alleviate this problem.

Due to the downsampling layers, the proposed network can extract semantic information at different image resolutions, and these feature maps are fused with the feature maps extracted

Fig 5. Illustration of the dense block. Dense block is composed of 6 convolution layers to obtain a large receptive field and skip connections are used to avoid the risk of network degradation, gradient vanish and gradient explosion.

<https://doi.org/10.1371/journal.pone.0252287.g005>

in the decoder structure. In addition to the U-like architecture, a hybrid attention module is integrated into the skip connections between the encoder structure and the decoder structure. This module is composed of a channel attention module and a spatial attention module.

As is shown in Fig 6, the attention module has two input feature maps $X \in R^{w \times h \times c}$ and $Y \in R^{w/2 \times h/2 \times 2c}$ from different dense blocks in the encoder path. For the spatial attention module, input Y is upsampled by a factor of 2 and is fused with input X , and the position contextual information is obtained by several 1×1 convolution layers, then the weight map is obtained through a sigmoid function. Finally, we multiply the weight map and input X to get the attention map, and spatial attention can be described as follows:

$$SA = X\sigma(\text{conv}(f(X, Y))) \quad (6)$$

where σ is the sigmoid function, and f represents the fusion layer. For the channel attention module, we concatenate input X and input Y in the channel dimension, then use the global average pooling layer and global max pooling layer to obtain two $1 \times 1 \times c$ feature maps, where c denotes the number of channels. Average pooling can smoothly extract features but it cannot extract salient features, so that we introduce max pooling to obtain remarkable features of each

Fig 6. Illustration of the proposed attention module. It is composed of a spatial attention module and a channel attention module, and make the network focus on the most relevant regions and channels.

<https://doi.org/10.1371/journal.pone.0252287.g006>

channel. These two pooling layers are defined as follows:

$$gap = \frac{1}{w \times h} \sum_i^w \sum_j^h I_c(i, j) \quad (7)$$

$$gmp = \max I_c \quad (8)$$

where I_c denotes c^{th} channel of input I , w and h are width and height of the input. After a shared convolution layer, a fusion layer and a sigmoid function, we can get the channel weight, then we multiply channel weight and input X to obtain the attention map. Finally, we fuse the spatial attention map with the channel attention map and integrate it into the skip connections between the encoder path and the decoder path. The proposed attention module can help the network focus on the most relevant regions and channels and alleviate the risk of disrupted by the background region of the input image.

Results

Dataset

We evaluate our approach on the public NIH pancreas dataset [9], which is the largest and most authoritative Open Source Dataset in pancreas segmentation. The dataset includes 82 contrast-enhanced abdominal 3D CT scans. The resolution of each CT scan is $512 \times 512 \times L$, where $L \in [181, 466]$ is the number of slices along the third axis. The slice thickness varies from 0.5 mm to 1.0 mm. The pixel values in a CT slice were clipped to $[-100, 240]$, then re-scaled to $[0, 255]$. We resize the slices to $[224, 224]$ and use 3 adjacent slices to form 3 channels. Following the standard cross-validation strategy, we split the dataset into 4 fixed folds, each of which contains approximately the same number of samples. 3 of 4 subsets are used for training and the remaining one is used for testing. We use DSC to evaluate the similarity between the ground truth and the segmentation mask and record the mean, max, min and standard deviation statistics over all testing cases.

Implementation details

We implement our approach in Keras framework and train our segmentation model on an NVIDIA GeForce GTX 1060 GPU. Each model of coarse and fine stage is trained for 30 epochs with an Adam optimizer [26]. The learning rate is set to 0.00001 and batch size is set to 1. We use 10% of the training slices as a validation set to check for overfitting. We cut all 3D CT scans along the coronal, sagittal and axial axis and feed these slices into segmentation models to obtain the predication masks. Subsequently, we construct these masks into a 3D volume and then compute the DSC of the whole 3D CT volume. We use slices with more than 100 pixels of pancreatic region for coarse segmentation training.

Evaluation and analysis

We use the proposed 2.5D U-net for both coarse and fine segmentation. Coarse segmentation results are shown in Table 1, the proposed approach achieves 75.33 ± 7.36 , 74.04 ± 9.75 and

Table 1. Evaluation for coarse segmentation of three axes and fusion segmentation is reported in the table.

Axes	Mean DSC±std	Max DSC	Min DSC
Coronal axis	75.33±7.36	86.22	57.79
Sagittal axis	74.04±9.75	86.17	52.09
Axial axis	76.89±7.44	87.74	59.60
Fusion	78.79±6.83	89.28	68.37

<https://doi.org/10.1371/journal.pone.0252287.t001>

Table 2. Comparison between our approach and the state-of-the-art approaches on NIH pancreas dataset.

Method	Mean DSC $\pm$ std	Max DSC	Min DSC
Zhou et al. [15]	82.37 $\pm$ 5.86	90.85	62.43
Roth et al. [28]	81.27 $\pm$ 6.27	88.96	50.69
Yu et al. [17]	84.50 $\pm$ 4.97	91.02	62.81
Asaturyan et al. [29]	79.3 $\pm$ 4.4	86.0	72.8
Zhang et al. [27]	84.61 $\pm$ 5.21	91.46	70.36
Li et al. [30]	84.19 $\pm$ 5.73	91.08	53.61
Our approach	86.61$\pm$3.47	92.03	73.89

<https://doi.org/10.1371/journal.pone.0252287.t002>

76.89 $\pm$ 7.44 average DSC for three views and the gap between maximum DSC and minimum DSC is more than 20%. This demonstrates that the network is confused by the background region because of the irregular shape and size of the pancreas. We improve the segmentation accuracy to 78.79 $\pm$ 6.83 by fusing these three models via majority voting, which means this fusion operation can make use of the 3D contextual information, to a certain extent.

Then we evaluate the coarse-to-fine framework on the NIH pancreas dataset and compare the segmentation results of our approach with the state-of-the-art approaches in recent years. The mean, maximum and minimum DSC are presented in Table 2 and an example of the segmentation results is shown in Fig 7. Our segmentation results outperform the best approach [27] by 2%. The minimum DSC has reached 73.89% and the standard deviation has reached 3.47%, which is a significant improvement especially compared with the approach proposed by Roth et al. [28]. This remarkable improvement on minimum DSC suggests that our approach is more stable and reliable on the NIH pancreas dataset. The segmentation accuracy has been significantly improved from 78.79 $\pm$ 6.83 to 86.61 $\pm$ 3.47, which indicates that a smaller input image can effectively alleviate the risk of being confused by the background region.

Segmentation results by ablation study of our approach are demonstrated in Table 3. The results show that our attention module outperforms the attention module proposed by Oktay et al. [24] and Fu et al. [22] and 2.5D U-net architecture achieves better performance. In addition, 2.5D U-net with our attention module costs less time for training phase (about 30 minutes per epoch) than 2.5D U-net with dual attention module [22] (about 70 minutes per epoch). Compared with 2D networks and 3D networks, 2.5D networks achieve higher segmentation accuracy and 2.5D models are much smaller than 3D models.

We also analyze the DSC distribution of all slices for 2 models. As is shown in Table 4, the number of slices with segmentation accuracy less than 0.5 has decreased from 9.42% to 3.09%,

Fig 7. An example of the segmentation results. Compared with the coarse segmentation, fine segmentation result has been improved from 70.45% to 84.31%. However, we can observe that we cannot predict the boundary of the pancreas very well.

<https://doi.org/10.1371/journal.pone.0252287.g007>

Table 3. Ablation study of our proposed 2.5D U-net and proposed attention module. AG: The attention gate proposed in [24]. DA: The dual attention module proposed in [22]. HA: Our proposed attention module. Params: The number of parameters.

Network	Mean DSC $\pm$ std	Min DSC	Max DSC	Params
2D U-net	79.50 $\pm$ 6.83	56.52	87.94	5.8M
3D U-net	81.37 $\pm$ 5.32	60.60	88.90	19.2M
2.5D U-net	81.79 $\pm$ 4.88	63.61	89.05	6.9M
2.5D U-net + AG	83.10 $\pm$ 5.32	67.36	91.02	7.2M
2.5D U-net + DA	84.07 $\pm$ 4.73	67.61	91.08	6.1M
2.5D U-net + HA	86.61 $\pm$ 3.47	73.89	92.03	7.3M

<https://doi.org/10.1371/journal.pone.0252287.t003>

Table 4. Analysis of the DSC distribution of all slices.

Method	U-net	Our approach
DSC [0, 0.5]	9.42%	3.09%
DSC (0.5, 0.6]	4.92%	3.70%
DSC (0.6, 0.7]	8.94%	10.14%
DSC (0.7, 0.8]	16.96%	18.97%
DSC (0.8, 0.9]	40.24%	44.84%
DSC (0.9, 1]	19.51%	19.26%

<https://doi.org/10.1371/journal.pone.0252287.t004>

and the number of slices with segmentation accuracy distributing between 0.5 and 0.6 has decreased from 4.92% to 3.70%. This denotes that our approach can effectively address the problem of the background region disruption and is stable on the NIH pancreas dataset.

Conclusion

In this paper, we propose a 2.5D U-net and a hybrid attention module for the pancreas segmentation task. The proposed attention module helps the network focus on the relevant region, which alleviate the risk of being confused by the background region. Our 2.5D U-net combines 2D convolutional layers with 3D convolutional layers, and uses 3 adjacent slices to form a 3-channel input image. Therefore, the proposed network achieves a higher accuracy because it can capture the inter-slice information and requires less computational resources than 3D networks. Furthermore, we follow Zhou et al. [15] to apply a coarse-to-fine framework to improve the segmentation accuracy.

Our approach is evaluated on the NIH pancreas dataset, and outperforms state-of-the-art approaches. The minimum DSC for testing cases has also been remarkably improved compared with state-of-the-art approaches, which proves that our approach is stable and can alleviate the risk of being disrupted by the background region.

Author Contributions

Methodology: Yingjing Yan.

Supervision: Defu Zhang.

Writing – original draft: Yingjing Yan.

Writing – review & editing: Defu Zhang.

References

1. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. *Medical image analysis*. 2017; 35:18–31. <https://doi.org/10.1016/j.media.2016.05.004> PMID: 27310171
2. Zhou Y, Xie L, Fishman EK, Yuille AL. Deep supervision for pancreatic cyst segmentation in abdominal CT scans. In: *International conference on medical image computing and computer-assisted intervention*. Springer; 2017. p. 222–230.
3. Karasawa K, Oda M, Kitasaka T, Misawa K, Fujiwara M, Chu C, et al. Multi-atlas pancreas segmentation: atlas selection based on vessel structure. *Medical image analysis*. 2017; 39:18–28. <https://doi.org/10.1016/j.media.2017.03.006> PMID: 28410505
4. Farag A, Lu L, Roth HR, Liu J, Turkbey E, Summers RM. A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. *IEEE Transactions on Image Processing*. 2016; 26(1):386–399. <https://doi.org/10.1109/TIP.2016.2624198> PMID: 27831881
5. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2015. p. 3431–3440.
6. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer; 2015. p. 234–241.
7. Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: *Artificial intelligence and statistics*. PMLR; 2015. p. 562–570.
8. Zeng G, Zheng G. Multi-stream 3D FCN with multi-scale deep supervision for multi-modality isointense infant brain MR image segmentation. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE; 2018. p. 136–140.
9. Roth HR, Lu L, Farag A, Shin HC, Liu J, Turkbey EB, et al. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: *International conference on medical image computing and computer-assisted intervention*. Springer; 2015. p. 556–564.
10. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *International conference on medical image computing and computer-assisted intervention*. Springer; 2016. p. 424–432.
11. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE; 2016. p. 565–571.
12. Song M, Hu Y, Chen H, Li T. Towards pervasive and user satisfactory cnn across gpu microarchitectures. In: *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE; 2017. p. 1–12.
13. Fu M, Wu W, Hong X, Liu Q, Jiang J, Ou Y, et al. Hierarchical combinatorial deep learning architecture for pancreas segmentation of medical computed tomography cancer images. *BMC systems biology*. 2018; 12(4):119–127. <https://doi.org/10.1186/s12918-018-0572-z> PMID: 29745840
14. Cai J, Lu L, Xing F, Yang L. Pancreas segmentation in CT and MRI images via domain specific network designing and recurrent neural contextual learning. *arXiv preprint arXiv:180311303*. 2018;.
15. Zhou Y, Xie L, Shen W, Wang Y, Fishman EK, Yuille AL. A fixed-point model for pancreas segmentation in abdominal CT scans. In: *International conference on medical image computing and computer-assisted intervention*. Springer; 2017. p. 693–701.
16. Cai J, Lu L, Xing F, Yang L. Pancreas segmentation in CT and MRI via task-specific network design and recurrent neural contextual learning. In: *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. Springer; 2019. p. 3–21.
17. Yu Q, Xie L, Wang Y, Zhou Y, Fishman EK, Yuille AL. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2018. p. 8280–8289.
18. Gupta A, Agrawal D, Chauhan H, Dolz J, Pedersoli M. An attention model for group-level emotion recognition. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*; 2018. p. 611–615.
19. Chen S, Tan X, Wang B, Hu X. Reverse attention for salient object detection. In: *Proceedings of the European Conference on Computer Vision (ECCV)*; 2018. p. 234–250.
20. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X. Multi-context attention for human pose estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017. p. 1831–1840.

21. Nie D, Gao Y, Wang L, Shen D. ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2018. p. 370–378.
22. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 3146–3154.
23. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 603–612.
24. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:180403999. 2018;.
25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
26. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014;.
27. Zhang Y, Wu J, Liu Y, Chen Y, Chen W, Wu EX, et al. A deep learning framework for pancreas segmentation with multi-atlas registration and 3D level-set. Medical Image Analysis. 2021; 68:101884. <https://doi.org/10.1016/j.media.2020.101884> PMID: 33246228
28. Roth HR, Lu L, Lay N, Harrison AP, Farag A, Sohn A, et al. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Medical image analysis. 2018; 45:94–107. <https://doi.org/10.1016/j.media.2018.01.006> PMID: 29427897
29. Asaturyan H, Gligorievski A, Villarini B. Morphological and multi-level geometrical descriptor analysis in CT and MRI volumes for automatic pancreas segmentation. Computerized Medical Imaging and Graphics. 2019; 75:1–13. <https://doi.org/10.1016/j.compmedimag.2019.04.004> PMID: 31103856
30. Li J, Lin X, Che H, Li H, Qian X. Probability map guided bi-directional recurrent UNet for pancreas segmentation. arXiv preprint arXiv:190300923. 2019;.