Xie *et al.*
Health Information Science and Systems (2023) 11:9
https://doi.org/10.1007/s13755-022-00200-z

Health Information Science and Systems

RESEARCH

Target area distillation and section attention segmentation network for accurate 3D medical image segmentation

Ruiwei Xie[1], Dan Pan[2*], An Zeng[1*], Xiaowei Xu[3*], Tianchen Wang[3], Najeeb Ullah[4] and Yuzhu Ji[1]

Abstract

3D medical image segmentation has an essential role in medical image analysis, while attention mechanism has improved the performance by a large margin. However, existing methods obtained the attention coefficient in a small receptive field, resulting in possible performance limitations. Radiologists usually scan all the slices first to have an overall idea of the target, and then analyze regions of interest in multiple 2D views in clinic practice. We simulate radiologists' recognition process and propose to exploit the 3D context information in a deeper manner for accurate 3D medical images segmentation. Due to the similarity of human body structure, medical images of different populations have highly similar shape and location information, so we use target region distillation to extract the common segmented region information. Particularly, we proposed two optimizations including Target Area Distillation and Section Attention. Target Area Distillation adds positions information to the original input to let the network has an initial attention of the target, while section attention performs attention extraction in three 2D sections thus with large range of receptive field. We compare our method against several popular networks in two public datasets including ImageCHD and COVID-19. Experimental results show that our proposed method improves the segmentation Dice score by 2–4% over the state-of-the-art methods. Our code has been released to the public (Anonymous link).

Keywords: 3D medical image segmentation, Section attention, Target area distillation, Transformer, U-Net

Introduction

3D medical image segmentation has always been one of the most important tasks in medical imaging research. Manual segmentation by pathologists is time-consuming and tedious especially on 3D medical images. Thus, automatic segmentation for 3D medical images has attracted tremendous attention in the community. Recently, with the advances in medical imaging technologies, the related 3D data has been increasing exponentially for decades [7]. Ponemon Institute survey found that 30% of the world's data storage resides in the healthcare industry by 2012 [10]. Recently, fully conventional networks [15] especially 2D/3D U-Net [6, 19] and nnU-net [32] have improved the segmentation performance by a large margin compared with traditional methods [13, 27] (Fig. 1).

Currently, attention mechanism has been extensively used to further boost the performance in various tasks. Attention gate mechanism [20] was proposed as a soft attention coefficient of the region in U-Net. However, the segmentation models obtained the attention coefficient in a small receptive field, resulting in possible performance limitations. As Transformer [12] works in an attention like manner and can effectively extract long-range or large-scale context information using position coding, TransUnet [5] combines Transformer and U-Net to form a powerful encoder to mitigate the disadvantage (focus on local details more than long-range context information) of U-Net.

On the other hand, radiologists in clinic practice usually follow some patterns, which can be regarded as

*Correspondence: pandan@gpnu.edu.cn; zengan@gdut.edu.cn; xiao.wei.xu@foxmail.com
[1] Guangdong University of Technology, Guangzhou, Guangdong, China
[2] Guangdong Polytechnic Normal University, Guangzhou, Guangdong, China
[3] Guangdong Provincial People's Hospital, Guangzhou, Guangdong, China
Full list of author information is available at the end of the article

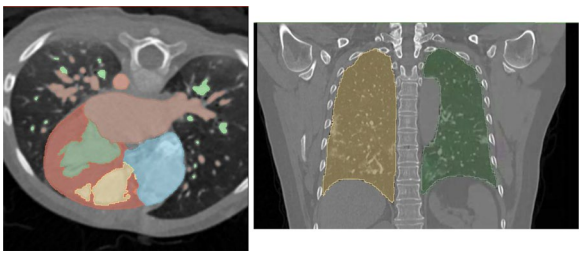

Fig. 1 Typical examples of: **a** a three-dimensional heart image, which are overlayed by segmentation masks of the heart, in seven colors, and **b** an three-dimensional lungs image, which are overlayed by segmentation masks of the lungs, in yellow, green. (Best viewed in color.)

specific forms of attention. First, radiologists usually scan all the slices first to have an overall idea of the target. In this way, a rough segmentation mask is obtained, with normal and abnormal decisions on regions are made during the initial rough scan. Second, they analyze abnormal regions (regions of interest) one-by-one to perform more detailed analysis. Particularly, multiple 2D views (including but not limit to axial, coronal, and sagittal views) are used to analyze the regions to know the large-field context thus with improved accuracy.

Inspired by Transformer and TransUnet which effectively make use of long-range context information, in this paper, we imitate the two patterns of radiologists, and take the step further to exploit the 3D context information in a deeper manner. Particularly, we proposed two novel modules, including target area distillation (TAD) and section attention (SA). TAD adds the positions information of the original input to let the network to learn an initial attention of the target, while SA performs attention extraction in three 2D sections thus with larger receptive field. We compared our method against several popular networks in two public datasets including ImageCHD and COVID-19. Experimental results show that our proposed method improves the segmentation Dice score by 2%-4% over the state-of-the-art methods. Our code has been released to the public [1]. Overall, our major contributions can be summarized as follows:

- We designed a 3D section attention module. Each section can obtain feature information from the sections of other dimensions, which is helpful to extract more context information, so as to improve the interest level of the region and strengthen the attention to the feature information of the divided region;
- We propose a voxel level target area distillation method. With the iteration of model training, the rectification of the target area is to update each element point in the random original input information with the corresponding position information of the label. Finally, all the data are integrated with the target location information of an element point in different data to form a unified target region location information, which is helpful to strengthen the model to distinguish the region of interest;
- We evaluated our proposed approach using a self-designed model of target area distillation and 3d sectional attention mechanism applied to multi-label 3D heart segmentation data and COVID-19 3D lung segmentation data. The results show that our 3d section attention mechanism and target area distillation can improve the prediction accuracy and obtain the best performance.

Related work

CNN-based segmentation

In the past years, substantial progress has been made on biomedical image segmentation with pixel based methods [8, 18, 22, 25] and structure based methods [3, 9, 11, 23]. Because these methods derive advanced results from a priori knowledge of hand-derived feature structures, their performance is greatly degraded when applied to severely deformed objects. Therefore, in the research of biomedical image segmentation, many researchers reduce the hand-made features or prior knowledge, focus on the fully convolutional network research, and get an effective performance improvement. Ronneberger et al. [19] proposed U-shaped deep convolutional network that contains a symmetric expanding path to enable precise localization. Similar to U-Net [19], a variety of works including 3D FCN with attention gate [21], Res-Unet [36], and Unet 3+ [14] have been proposed recently to further boost the segmentation performance. Wu et al. [33] designed a novel inception-residual block for a U-shape network and introduced four supervision paths with different convolution kernel sizes to utilize multi-scale features. In order to use the multi-scale features, Wu et al. [33] designed a novel inception-residual block, which uses four different convolution kernels for supervision. Song etal. [24]proposes AttaNet network, which is composed of two modules: Strip Attention module (SAM) and attention fusion module (AFM). SAM uses

striping operations to not only retain context information but also reduce the complexity of vertical global context encoding. AFM uses an attention strategy to weight features at different levels of pixels, to obtain efficient multi-layer representation. Wang et al. [31] proposed dual U-Nettthe model can combine an encoder extracting spatial information with an encoder extracting context. To improve the hierarchical representation capture ability of the model, Mou et al. [17] added self-atttention mechanism to the U-shape encoder-decoder.

Attention and transformer

Since the rise of the attention mechanism [29], researchers in the field of segmentation have paid extensive attention on this interesting topic. Attention gated network [20] integrates the skip connection with additive attention gate into 2D/3D U-Net [6, 19 for medical image segmentation. However, this method is still based on CNN, and the attention coefficient is obtained in a small range of the acceptance domain, instead of integrating larger context information. TransUnet [5] combined Transformer and CNN to form a powerful encoder, and the decoding is based on the expanding path of U-Net. This method uses a large amount of 2D image pre-training and can achieve the best performance, but when used for 3D image segmentation, the effect is very mediocre. Similar to TransUnet, the methods [28, 35] also take advantage of the complementarity of Transformer and CNN to improve the segmentation performance. Swin-Unet [4] is a U-type architecture based on Transformer and consists of encoder, bottleneck, decoder, and jump link. However, the input image is segmented into non-overlapping image blocks, resulting in limited context information.

Methodology

Overview

The overall architecture is illustrated in Fig. 2, which has a similar structure as TransUnet [5]. There are two encoders in our model: a Transformer encoder and a CNN encoder. Two optimization modules, TAD and SA, are added. TAD works at the input part, while SA is added to the skip connections. Particularly, the original data is downsampled to an uniform size. TAD is described in detail in Sect. 3.2, while SA is described in Sect. 3.3.

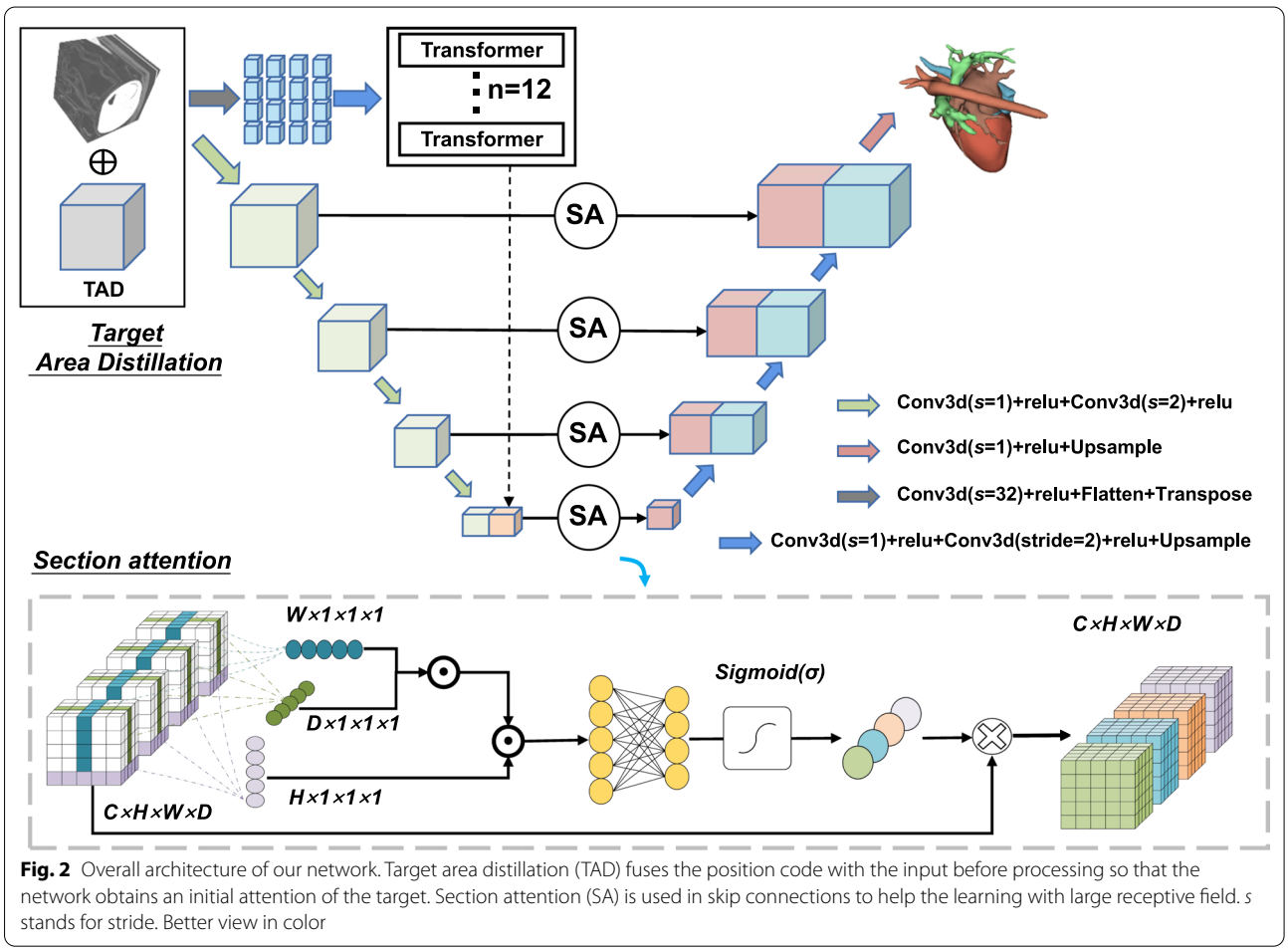

Fig. 2 Overall architecture of our network. Target area distillation (TAD) fuses the position code with the input before processing so that the network obtains an initial attention of the target. Section attention (SA) is used in skip connections to help the learning with large receptive field. *s* stands for stride. Better view in color

Target area distillation

Radiologists usually first perform a quick scan of the 2D slices (e.g., CT/MRI slices) to have an overall idea, then performs local analysis. Inspired by the aspect, we believe that the input should not only include the voxel information but also the location information of input pixels. Therefore, we add region location information to the voxel information in the original image, and the size of the region information is consistent with the size of the original image because each voxel should correspond to location information. In addition, the location region information is a trainable parameter, because the segmentation region of the input training original image dataset is unknown, so we need to use the parameter adjustment process of model training to distill the segmentation region of all training datasets. According to the research experience, we can adjust the parameters by using the loss function to fit the segmentation region we need. For the target area distillation, loss function is also used to adjust parameters, and the location information feature we randomly initialize is fitted and adjusted, and the size of the location information feature is consistent with the original image. In the training process, the location information features that belong to the segmented region will get higher scores, while those that do not belong to the segmented region will get less scores. Since our location information features are trained and updated in all data sets, it is equivalent to distilling the segmented region information of all data sets. Such a process is called target region distillation. In short, the approximate position of the target area is finally obtained by rectification of the target area, so that the network can focus on these areas and improve its efficiency.

Section attention

Similar to Attention Unet [20], SA works at the skip connections, and SA obtains attention in a large receptive field.This is inspired by the fact that radiologists usually check some uncertain target in the sagittal, coronal, and transverse planes. Because radiologists are looking for a more complete view of organ function,the medical images of organ segmentation region is usually in the center of the image, and image edge is usually other physiological structure. The more central the image, the more likely it is to be a segmented region. Particularly, SA first divides features into three widely used planes including the sagittal, the coronal, and the transverse planes in 3D medical images. Second, we evenly pool all sections of the same dimension of each channel to extract attention coefficients. Suppose C, H, W, and D are the number of channel, height, width, and depth of the feature maps, and $x_{c,h,w,d}$ represents the voxel with an index of c, h, w, and d on the corresponding axle, respectively. The calculation of H_{att}, W_{att} and D_{att} is as follows,

$$W_{att} = \frac{1}{CHD} \sum_{c=0}^{C} \sum_{h=0}^{H} \sum_{d=0}^{D} x_{c,h,w,d}, \tag{1}$$

$$H_{att} = \frac{1}{CWD} \sum_{c=0}^{C} \sum_{w=0}^{W} \sum_{d=0}^{D} x_{c,h,w,d}, \tag{2}$$

$$D_{att} = \frac{1}{CHW} \sum_{c=0}^{C} \sum_{h=0}^{H} \sum_{w=0}^{W} x_{c,h,w,d}. \tag{3}$$

Third, we could obtain H_{att}, W_{att} and D_{att} attention coefficients in the three plans, respectively. Based on the intuition that some point should be paid more attentions if it is important in all of the three plans, we propose the attention scores H_{att}, W_{att} and D_{att} to fuse the attention coefficients. The final attention coefficient S_{att} is expressed by Equation (4),

$$S_{att} = \sigma(\mathbf{A}^{\mathrm{T}}((W_{att} \cdot D_{att}) \cdot H_{att}) + b) \tag{4}$$

where σ is the Sigmoid activation function, and b is the bias term. We use the sigmoid function as the activation function here for two reasons. The first is that sigmoid function is continuous everywhere to facilitate derivation. The second is that it can control the output attention coefficient between 0 and 1, which is convenient for us to classify the feature layer into redundant information to be ignored or regional information to be focused on.

S_{att} is 3D section attention, We extract feature attention according to S_{att}, so that our feature maps pay more attention to the segmented region.

Post-processing

Generally speaking, when our segmentation model deduces and calculates segmented regions, some erroneous sporadic regions are predicted to be segmented. Therefore, post-processing procedures need to be added after our segmentation model to remove erroneous fragmentary segmentation regions, so that the segmentation region we get is closer to the region to be segmented, which can improve our segmentation accuracy. Since there usually exists a few incorrect islands in the segmentation results which are isolated from the target island. Thus, by analyzing voxel-wise morphological connectivity using scikit-image [30], only the largest island is extracted and the other small ones are disregarded. Algorithm 1 describes the detailed post-processing algorithm

Algorithm 1 Post-processing algorithm

Require: *predictionLabel*

function DELETESPORADICAREAS(*predictionLabel*)
  $recordLabel \leftarrow predictionLabel$
  $recordLabel[recordLabel > 0] \leftarrow 1$
  $maxConnectedRegion \leftarrow getMaxConnectedRegion(recordLabel)$
  $newPredictionLabel \leftarrow maxConnectedRegion * predictionLabel$
  return $newPredictionLabel$
end function

function GETMAXCONNECTEDREGION(*recordLabel*)
  from skimage.measure import label as osLabel
  from skimage.measure import regionprops
  $newPreLabel, connectedNum \leftarrow osLabel(record)$
  if $connectedNum < 1$ **then**
    return record
  end if
  $region \leftarrow regionprops(newPreLabel)$
  for $i = 1 \rightarrow connectedNum + 1$ **do**
    $numList[i-1] \leftarrow i$
  end for
  for $i = 0 \rightarrow connectedNum$ **do**
    $areaList[i] \leftarrow region[numList[i]]$
  end for
  $numListSorted \leftarrow sorted(numList, key = lambda\ x : areaList[x-1][::-1])$
  if $len(numListSort) > 1$ **then**
    for $i = 1 \rightarrow len(numListSort)$ **do**
      $newPreLabel[region[numListSort[i]-1].slice][region[numListSort[i]-1].image] \leftarrow 0$
    end for
  end if
  $newPreLabel[newPreLabel > 0] \leftarrow 1$
  return newPreLabel
end function

Experiments

Datasets

Two 3D medical datasets, ImageCHD and COVID-19 which are shown in Fig. 1, are adopted for evaluation. ImageCHD [34] consists of 110 patients, and the size of the images is $512 \times 512 \times (129 - 357)$, and the typical voxel size is $0.25 \times 0.25 \times 0.5$ mm^3. As is shown Fig. 3, the segmentation include seven substructures: left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (Myo), faorta (Ao) and pulmonary artery (PA). Figure 3

COVID-19 [16] contains 20 labeled COVID-19 CT scans. Left lung, right lung, and infections are labeled. There exist three segmentation benchmark tasks based on this dataset [2], and we focus on segmentation of left lung, right lung, and infections using pure but limited

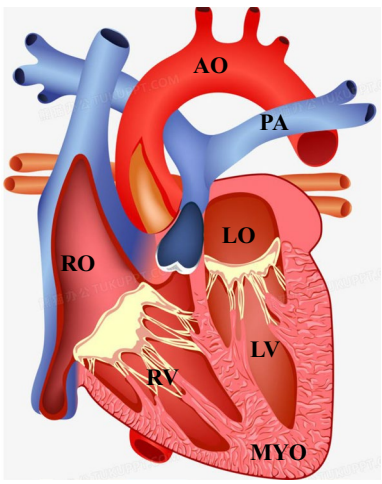

Fig. 3 The seven segmented organ regions of the heart

Table 1 Segmentation performance of our method and exiting methods on the ImageCHD dataset

Method	Dice	HD95	Iou	Rev
3DUnet	0.808	18.340	0.756	0.133
3D FCN	0.718	19.949	0.685	0.161
3DtransUnet	0.668	25.102	0.649	0.164
3DswinUnet	0.848	**11.973**	0.820	0.098
Ours	**0.868**	12.725	**0.852**	**0.061**

The experimental data are the average results of 5-fold Cross Validation tests

Bold values represent the current highest value for the data

Table 2 Segmentation performance of our method and exiting methods on the Covid-19 dataset

Method	Dice	HD95	Iou	Rev
3DUnet	0.913	15.769	0.847	0.080
3D FCN	0.896	9.828	0.788	0.035
3DtransUnet	0.868	12.680	0.756	0.076
3DswinUnet	0.921	11.000	0.859	0.092
Ours	**0.961**	**3.982**	**0.937**	0.023

The experimental data are the average results of 10-fold Cross Validation tests

Bold values represent the current highest value for the data

COVID-19 CT scans. In the splitting of the ImageCHD dataset, we used five-fold cross-validation (19 images for testing, and 76 images for training) In the splitting of the COVID-19 dataset, we used ten-fold cross-validation (2 images for testing, and 18 images for training). During the dataset split, all types of heart or lung diseases appear in subsets with equal probability.

Experimental setup

Compared method

In the experiments, we compare our method with four state-of-the-art 3D segmentation frameworks, 3D U-Net [6], 3D TransUnet [5], attention gated networks [20], and SwinNet [4].

Evaluation Metrics Four evaluation metrics including Dice coefficient (DSC), 95% Hausdorff distance (HD95), intersection-over-union (IoU), and Relative Volume Error (RVE) are adopted for a comprehensive evaluation.

IoU is the overlap region between predicted segmentation and label divided by the joint region between predicted segmentation and label (intersection of the two/union of the two).

Dice coefficient is defined as the intersection of two times divided by the sum of pixels, also known as the F1 score. The Dice coefficient is very similar to IoU, they are positively correlated.

HD95 is used for the segmentation index, which is mainly used to measure the segmentation accuracy of the boundary and a measure to describe the similarity between two sets of points. It is a definition form of the distance between two sets of points.

RVE builds microscale correlations between structure and performance.

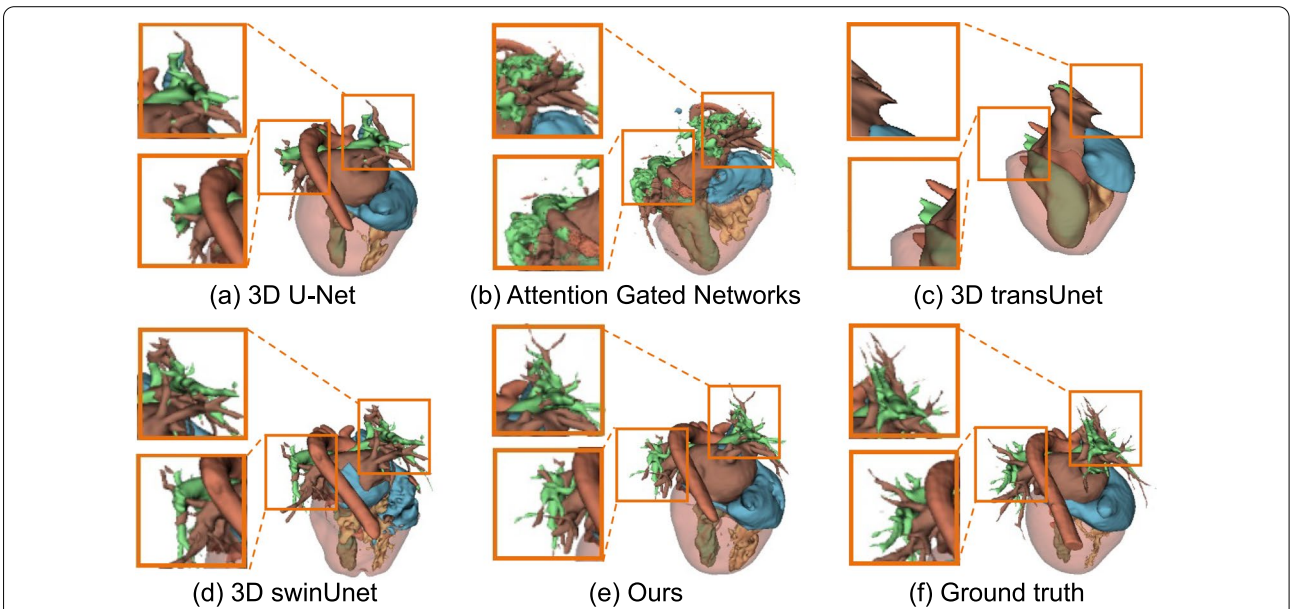

Fig. 4 Visualization comparison of our method and existing methods on the ImageCHD datase. Two boxes in the same positions Better view in color

Table 3 Ablation results on the ImageCHD dataset

Method	Dice	HD95	Iou	Rev
Baseline	0.792	21.345	0.739	0.073
Baseline+SA	0.834	17.129	0.772	0.069
Baseline+TAD	0.754	26.582	0.703	0.094
Baseline+TAD+SA	**0.868**	**12.725**	**0.852**	**0.061**

TAD for target area distillation and SA for section attention. The experimental data are the average results of 5-fold Cross Validation tests

Bold values represent the current highest value for the data

Table 4 Ablation results on the Covid-19 dataset

Method	Dice	HD95	Iou	Rev
Baseline	0.930	5.878	0.896	0.017
Baseline+SA	0.944	10.290	0.902	0.043
Baseline+TAD	0.924	7.143	0.870	0.045
Baseline+TAD+SA	**0.961**	**3.982**	**0.930**	**0.023**

TAD for target area distillation and SA for section attention. The experimental data are the average results of 10-fold Cross Validation tests

Bold values represent the current highest value for the data

Implementation details

We use 3D U-Net as the basic backbone, in which skip connection is replaced by our 3D section attention module. In addition, we added a Transformer encoder, which can extract high-resolution information in the CNN encoder. For the data input to the model, we downsample the original image and then get an image with a size of $256 \times 256 \times 256$. The position encoding is a matrix with the same size as that of the downsampled images, and the matrix is randomly initialized before training. We model each batch of two input image information, so each group of the input image and the target area coding together. The target area coding in the model is used to fuse the segmentation region feature information of all training data. With the training iteration, the parameter will eventually fuse the characteristic information of the cardiac segmentation region by itself. All the experiments run on a Nvidia GeForce RTX 3090 GPU with 24 GB memory. All models run in the PyTorch framework, and the training epochs are 400. We use the Adam optimizer with a learning rate of 0.0001,the betas of (B1,B2)=(0.9,0.999), an epsilon of 1e-8 and a weight decay of zero. Dice loss is adopted with a smooth parameter of 1e-5.

Comparison with existing methods

The result is shown in Tables 1 and 2. We can notice that the proposed method achieves the optimal performance on three of four metrics including Dice, HD95, IoU, RVE on the ImageCHD dataset, and on all the four metrics on the COVID-19 dataset. Compared with existing methods, the proposed method can improve the DSC by 2.2% and 4.0% on the ImageCHD dataset and the COVID-19 dataset, respectively. Our method has a lower performance only on HD95 on the ImageCHD dataset. Overall, our method achieves higher performance than state-of-the-art methods on the two public 3D medical image dataset. We can observe that 3D U-Net obtains much higher performance on all the four metrics in both the two adopted datasets which is expected asmU-Net has

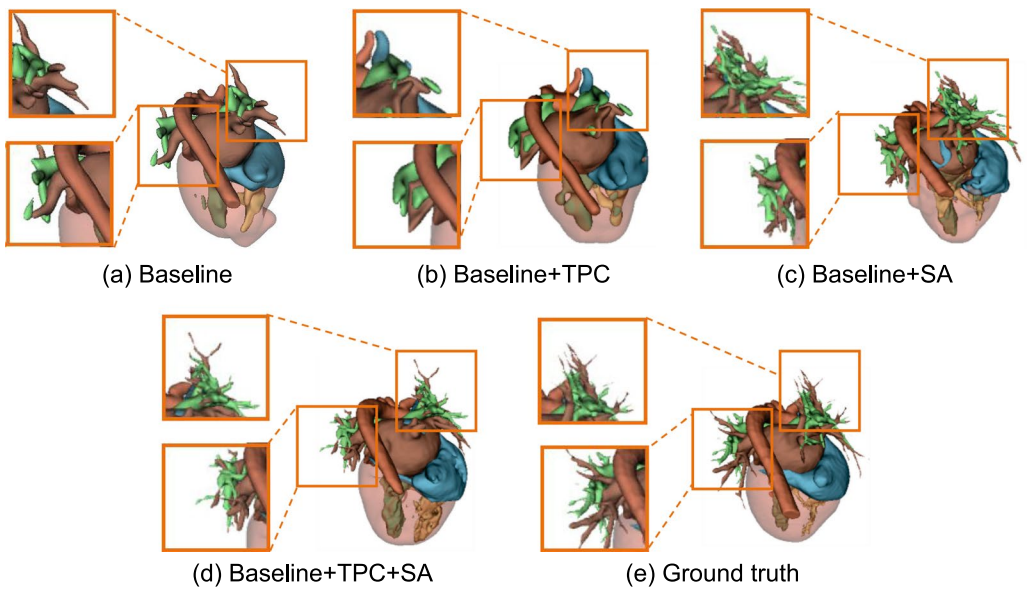

Fig. 5 3D visualisation of the ablation results on the ImageCHD database. Better view in color

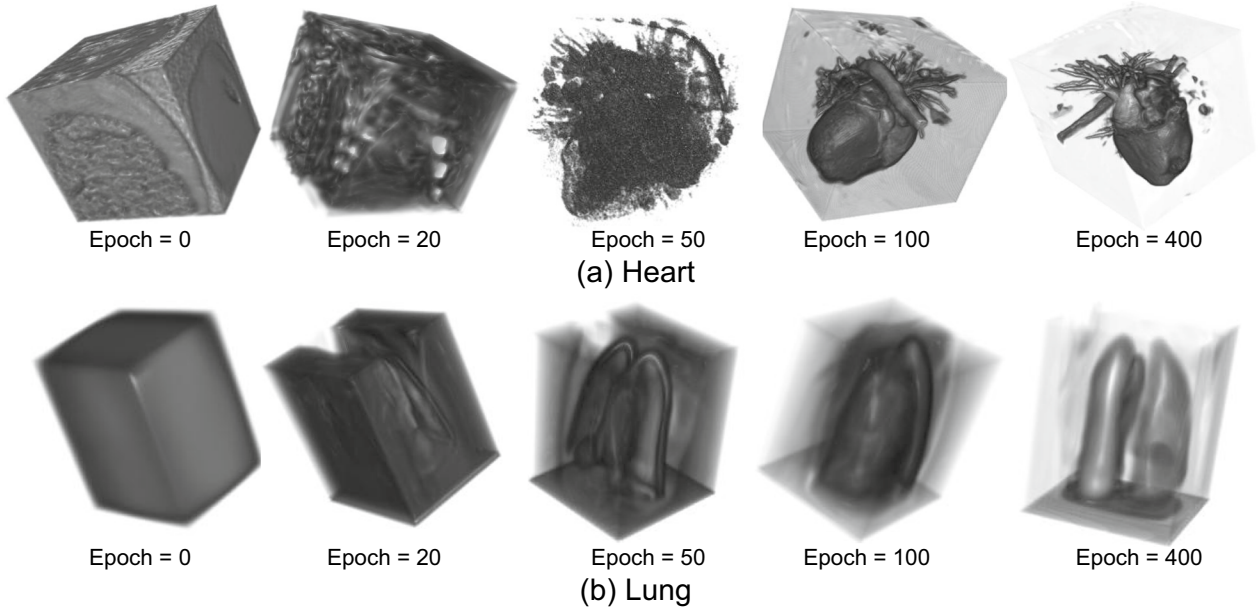

Fig. 6 3D illustration of target area updating during training in segmentation of **a** hearts in the ImageCHD dataset and **b** lungs in the COVID-19 dataset

skip connections to help location of the details. Surprisingly 3D transUnet obtains the worst performance. The reason is that pre-training of 2D transUnet using the large-scale ImageNet dataset is critical for performance enhancement. However, for 3D image segmentation, there doesn't exist such a large-scale dataset. Considering the high cost, large-scale 3D image dataset is not possible at least for a while.

We can also observe that a similar work, 3D swinUnet which also combines U-Net and Transformer, achieves a much better performance. The main reason is that as Transformer is too deep to be converged [26], two successive Swin Transformer blocks [4] are used to constructed the bottleneck to learn the deep feature representation. In this way, pre-training become much less critical for 3D swinUnet than transUnet.

Visualization of the comparison is shown in Fig. 4. 3D transUnet obtains the worst segmentation and all most all thin vessels are missing. 3D U-Net and attention gated networks can both detect part of these thin vessels, but 3D U-Net has much less isolated islands as the skip connection in 3D U-Net can help locate such details more precisely. 3D swinUnet can extract more details, while ours can detect almost all thin vessels compared with the ground truth. However, we can still notice obvious segmentation error in the thin vessels, which need further investigation on distinguishing local details.

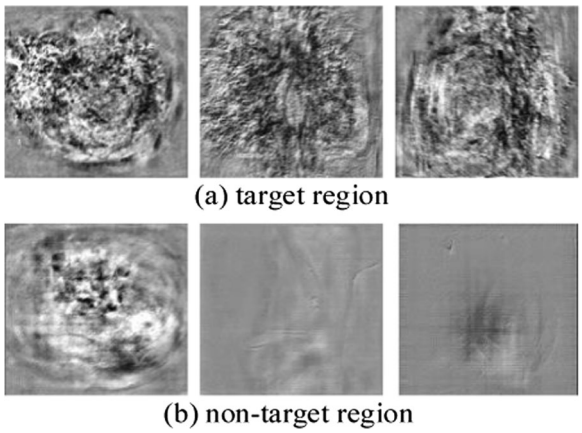

Fig. 7 Axial, Coronal , and Sagittal views of points in the **a** target region and the **b** non-target region on the ImageCHD dataset

Ablation study

Ablation study on the ImageCHD dataset and the COVID-19 dataset is shown in Tables 3 and 4, and 3D visualization of the ablation results on the ImageCHD dataset is shown in Fig. 5. Note that the baseline is the network shown in Fig. 2 with TAD and SA. We can notice that when SA and the combination of SA and TAD can effectively improve the performance.

We can also notice that TAD degrades the performance by 0.06–3.8% on DSC. However, more interestingly, when both TAD and SA are adopted, the performance is even

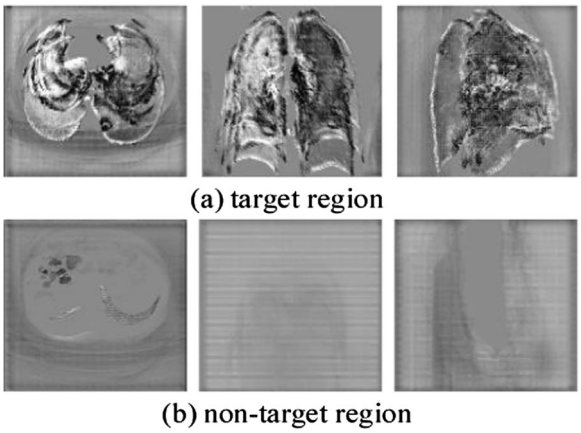

Fig. 8 Axial, Coronal , and Sagittal views of points in the **a** target region and the **b** non-target region on the Lung dataset

higher than that with only SA. A possible reason is that when only TAD is adopted, the network put too much effort to learn a good/precise target area which may discard some boundary regions. As a result, some thin vessels in the boundary regions may not be detected as shown in Fig. 5b compared with the baseline as shown in Fig. 5c. When TPC and SA are both adopted, there exists a tradeoff between a precise target area and the SA coefficient, and the network can learn a rough but a better estimation of the target regions. When SA is adopted, most of the thin vessels especially in the boundary regions can be detected as shown in Fig. 5c and d. In addition, the segmentation of the atrium, ventricle, and myocardium is also slightly improved by SA. When both TAD and SA are used, more thin vessels can be detected (Fig. 5d).

3D illustration of position code updating during training on the ImageCHD dataset and the COVID-19 dataset is shown in Fig. 6 We can notice that as training progresses, the position code focus more on the target regions, and finally, we even get an rough region of the segmentation. We can also notice that the the target regions are much larger than some segmentation.

For example, the vessels of the position code with an epoch of 400 shown in ImageCHD dataset is much thicker than common vessels,COVID-19 dataset contains multiple lung shapes, and the superposition of these lung shapes is not in the same position. The reason is that the position code can be rgarded as a voting of all the targets in the training dataset, and thus it contains all the target regions but with different weights.

Examples of axial, coronal, and sagittal views of TAD feature maps in the (a–b) target region and the (a–b) non-target region are shown in Figs. 7 and 8. We can easily notice that the TAD feature maps can roughly obtain the target regions. When the voxel is located at the edge of the target, the lower coefficient gets less attention,but When located in the center of the target, the voxels are with higher coefficient thus with higher attention. For the COVID-19 dataset, there exist a double image of the lung regions. A possible reason is that SA also affect the TAD feature maps. Thus, SA and the position code in TAD together determine the TAD feature maps during inference.Thus, SA and the target area in TAD together determine the TAD feature maps during inference.

Conclusion

In this paper, we have exploited the 3D context information in a deeper manner, and proposed two optimizations including target position coding and section attention. We compared our method against several popular networks in two public datasets including ImageCHD and COVID-19. Experimental results show that our proposed method improves the segmentation accuracy by 2–4% over the state-of-the-art methods. 3D visualization results show that our method can detect more thin vessels than existing works. The ablation study shows the effectiveness of our method, and also shows its potential in 3D medical image segmentation. Our code has been released to the public [1] to facilitate further research.

Acknowledgements

This study was supported by the Science and Technology Planning Project of Guangdong (Grant Nos. 2019A050510041, 2021A1515012300 and 2021B0101220006), National Natural Science Foundation of China (Grant Nos. 61976058 and 61772143) and Science and Technology Planning Project of Guangzhou (Grant Nos. 202103000034, 202206010007 and 202002020090).

Declarations

Conflict of interest

The author confirms that there is no conflict of interest.

Author details

[1]Guangdong University of Technology, Guangzhou, Guangdong, China. [2]Guangdong Polytechnic Normal University, Guangzhou, Guangdong, China. [3]Guangdong Provincial People's Hospital, Guangzhou, Guangdong, China. [4]Mardan University of Engineering and Technology, Mardan, Pakistan.

Received: 12 August 2022 Accepted: 21 October 2022
Published online: 30 January 2023

References

1. https://github.com/xieruiwei/TAD-and-SA.
2. https://gitee.com/junma11/COVID-19-CT-Seg-Benchmark.
3. Altunbay D, Cigir C, Sokmensuer C, Gunduz-Demir C. Color graphs for automated cancer diagnosis and grading. IEEE Trans Biomed Eng. 2010;57(3):665–74.
4. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. 2021. arXiv:2105.05537.
5. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y: Transunet: transformers make strong encoders for medical image segmentation. 2021. arXiv:2102.04306.
6. 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. New York: Springer; 2016. p. 424–32.
7. Dinov ID. Volume and value of big healthcare data. J Med Stat Inf. 2016;4.
8. Doyle S, Madabhushi A, Feldman M, Tomaszeweski J. A boosting cascade for automated detection of prostate cancer from digitized histology. MICCAI; 2006. p. 504–11.
9. Fu H, Qiu G, Shu J, Ilyas M. A novel polar space random field model for the detection of glandular structures. IEEE Trans Med Imaging. 2014;33(3):764–76.
10. Gerrity J. Health networks–delivering the future of healthcare. 2014.
11. Gunduz-Demir C, Kandemir M, Tosun AB, Sokmensuer C. Automatic segmentation of colon glands using object-graphs. Med Image Anal. 2010;14(1):1–12.
12. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in transformer. Adv. Neural Inf Process Syst. 2021;34.
13. Held K, Kops ER, Krause BJ, Wells WM, Kikinis R, Muller-Gartner HW. Markov random field segmentation of brain mr images. IEEE Trans Med Imaging. 1997;16(6):878–86.
14. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen YW, Wu J. Unet 3+: a full-scale connected unet for medical image segmentation. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2020. pp. 1055–1059.
15. 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. 3431–40.
16. Ma J, Ge C, Wang Y, An X, Gao J, Yu Z, Zhang M, Liu X, Deng X, Cao S, Wei H, Mei S, Yang X, Nie Z, Li C, Tian L, Zhu Y, Zhu Q, Dong G, He J. Covid-19 ct lung and infection segmentation dataset. 2020.
17. Mou L, Zhao Y, Fu H, Liu Y, Cheng J, Zheng Y, Su P, Yang J, Chen L, Frangi AF, et al. Cs2-net: deep learning segmentation of curvilinear structures in medical imaging. Med Image Anal. 2021;67: 101874.
18. Nguyen K, Sarkar A, Jain AK: Structure and context in prostatic gland segmentation and classification. In: MICCAI. New York: Springer; 2012. p. 115–23
19. 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. New York: Springer; 2015. p. 234–41.
20. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal. 2019;53:197–207.
21. Shen Y, Fang Z, Gao Y, Xiong N, Zhong C, Tang X. Coronary arteries segmentation based on 3d fcn with attention gate and level set function. IEEE Access. 2019;7:42826–35.
22. Sirinukunwattana K, Snead DR, Rajpoot NM: A novel texture descriptor for detection of glandular structures in colon histology images. In: SPIE medical imaging. International Society for Optics and Photonics; 2015. p. 94200S.
23. Sirinukunwattana K, Snead DR, Rajpoot NM. A stochastic polygons model for glandular structures in colon histology images. IEEE Trans Med Imaging. 2015;34(11):2366–78.
24. Song Q, Mei K, Huang R: Attanet: attention-augmented network for fast and accurate scene parsing. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35. 2021, p. 2567–75.
25. Tabesh A, Teverovskiy M, Pang HY, Kumar VP, Verbel D, Kotsianti A, Saidi O. Multifeature prostate cancer diagnosis and gleason grading of histological images. IEEE Trans Med Imaging. 2007;26(10):1366–78.
26. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 32–42.
27. Tsai A, Yezzi A, Wells W, Tempany C, Tucker D, Fan A, Grimson WE, Willsky A. A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans Med Imaging. 2003;22(2):137–54.
28. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. arXiv:2102.10662. 2021.
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems. 2017. pp. 5998–6008.
30. Van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T. scikit-image: image processing in python. PeerJ. 2014;2: e453.
31. Wang B, Qiu S, He H. Dual encoding u-net for retinal vessel segmentation. In: International conference on medical image computing and computer-assisted intervention. New York: Springer; 2019. pp. 84–92.
32. Wang X, Girshick R, Gupta A, He K: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. pp. 7794–803.
33. Wu Y, Xia Y, Song Y, Zhang D, Liu D, Zhang C, Cai W. Vessel-net: retinal vessel segmentation under multi-path supervision. In: International conference on medical image computing and computer-assisted intervention. New York: Springer; 2019. p. 264–72.
34. Xu X, Wang T, Zhuang J, Yuan H, Huang M, Cen J, Jia Q, Dong Y, Shi Y. Imagechd: a 3d computed tomography image dataset for classification of congenital heart disease. In: International conference on medical image computing and computer-assisted intervention. New York: Springer; 2020. p. 77–87.
35. Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and cnns for medical image segmentation. arXiv:2102.08005. 2021.
36. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. 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: Springer; 2018, pp. 3–11.