MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231101

Dual-Attention Model Fusing CNN and Transformer for Pancreas Segmentation

Yan ZHU[a,b], Peijun HU[a], Yu TIAN[b], Kaiqi DONG[b] and Jingsong LI[a,b,1]
[a] *Research Center for Healthcare Data Science, Zhejiang Laboratory, Hangzhou, China*
[b] *Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China*

Abstract. Pancreatic cancer is a highly malignant cancer of the digestive tract and is rapidly progressing and spreading clinically. Automatic and accurate pancreatic tissue segmentation in abdominal CT images is essential for the early diagnosis of pancreatic-related diseases. It is challenging that the pancreas is small in size and complex in morphology. To address this problem, we propose a dual-attention model fusing CNN and Transformer to effectively activate pancreas-related features expression. The CNN structure weights the importance of pancreas-related features at the channel level and weakens the background information. Transformer feature aggregation module constructs spatial correlations among long-distance pixels from a global perspective. This study is validated on the NIH-TCIA dataset and achieved a mean Dice Similarity Coefficient of 85.82%, which is outperforming than the state-of-the-art methods. The visualization of surface distance also demonstrates the effective segmentation of pancreas boundary details by the proposed model.

Keywords. Pancreas segmentation, Attention mechanism, Transformer, CT images

1. Introduction

In recent years, the incidence and mortality of pancreatic cancer was increasing significantly, making it the fourth leading cause of cancer death in the United States[1]. Accurate localization and segmentation of the pancreatic tissue is important for early diagnosis, prognostic assessment, and improving survival for patients with pancreatic disease[2]. The precise automatic segmentation of pancreatic tissue faces two critical problems. First, pancreas presents characteristics such as small size and irregular morphology. Second, the pancreas is adjacent to a variety of organs, thus the pancreas in CT images demonstrates blurred borders and mutual obscuration with other tissues.

To highlight features beneficial for pancreas segmentation and suppress irrelevant features, Oktay *et al.*[3] proposed a model named Attention-UNet by introducing the gate structure as attention modules in 3D U-Net[4], achieving a Dice Similarity Coefficient (DSC) of 83.10%. Li *et al.*[5] designed a bi-directional recurrent neural

[1] Corresponding Author: Jingsong Li, Tel: +86-571-87951564, Fax: +86-571-87951564, email: ljs@zju.edu.cn.

network fused with probabilistic maps to obtain a DSC of 83.02% on the pancreas segmentation task. The network constructs pixel-level probability maps to enhance feature extraction, and a bidirectional recurrent network ensures propagation of spatial information. Yu *et al.*[6] and Wang *et al.*[7] adopt multi-stage cascade framework, in which features extracted from multiple networks complement each other to optimize the segmentation results. Existing studies have paid insufficient attention to morphological features of pancreas and showed deficiencies in segmentation accuracy.

In this paper, by introducing a dual attention mechanism into the segmentation framework, we propose a novel model integrating CNN and Transformer[8] structure to enhance the representation of pancreas-related features. The model outperforms the state-of-the-art models with a mean DSC of 85.82%. In addition, the ablation experiments validate the effectiveness of the dual attention mechanism.

2. Methods

2.1. Datasets

The dataset adopted is collected by the National Institutes of Health Clinical Center and is known as NIH-TCIA dataset[9]. The dataset contains 55 male and 27 female subjects, who had neither major abdominal nor pancreatic cancer lesions. The dataset consists of 82 3D abdominal enhanced CT images. The horizontal plane resolution of each CT scan is 512 × 512, and the slice thickness ranges between 1.5 mm and 2.5 mm.

2.2. Dual-Attention Network

2.2.1 Channel attention module

In this study, channel attention modules are embedded in the shallow layers of the encoder to activate the morphological feature expression. The module extends the traditional Squeeze and Excitation (SE) paradigm in 3D format and performs cascading. The channel attention of 3D feature space filters irrelevant background information in the CT images and activates the pancreatic-related foreground features. What's more, this structure enriches the capacity of the encoder and optimizes the model parameters, while improving the network feature extraction capability.

$$Z_c = F_{squeeze}(M_c) = \frac{1}{XYZ}\sum_{i=1}^{X}\sum_{j=1}^{Y}\sum_{k=1}^{Z} M_c(i,j,k) \tag{1}$$

$$M_{CA}{=}F_{\text{excite}}(Z_c,W)\cdot M_c = \sigma(g(Z_c,W)) \tag{2}$$

where Z_c presents squeezed features. The attention-based feature in output channel M_{CA} is computed by dot product of output $F_{excite}(\cdot)$ and input feature map M_c . In addition, σ denotes softmax operation, g denotes activation function and W is weights.

2.2.2 Spatial attention module

However, CNN lacks attention to global features. In contrast, the vision Transformer structure acquires long-term contextual semantic dependencies and thus perceives global features. Therefore, we introduce spatial attention module by Transformer structure to capture information correlations and establish associations among feature maps at different scales. The spatial attention module unfolds in a multi-level form by dividing the window into multiple sub-windows of different sizes and computing attention within the local window. This design enables the model to reduce computational consumption by using self-attentive mechanism to ensure the feature extraction effectiveness.

2.2.3 Fusion model

Based on attention modules above, we propose a pancreas segmentation model fusing CNN and Transformer and display it in Figure1. The model adopts the architecture of encoding-feature aggregation-decoding. The encoding and decoding modules both use convolution for down-sampling and up-sampling operations, respectively. The feature aggregation part employs the Transformer structure for information interaction and fusion.

At scale i of the decoder, feature map for segmentation reconstruction M_{S_i} is computed by concatenation of spatial attention feature map M_{SA} and the feature map from former layer $M_{S_{i-1}}$, which is computed as:

$$M_{S_i} = F_{cat}(M_{SA}, M_{S_{i-1}}) \tag{3}$$

Considering the variable morphological features of pancreas, the features used for output inference are fusion features of four scales in the up-sampling layers as follows:

$$M_{out} = M_{S1} \cdot (1 + F_{up}(M_{S2} \cdot (1 + F_{up}(M_{S3} \cdot (1 + F_{up}(M_{S4} \cdot (1 + M_{in})))))) \tag{4}$$

where output feature map M_{out} consists of 4 scales feature aggregation of $M_{S1}, M_{S2}, M_{S3}, M_{S4}$ in the form of level-by-level residual connection.

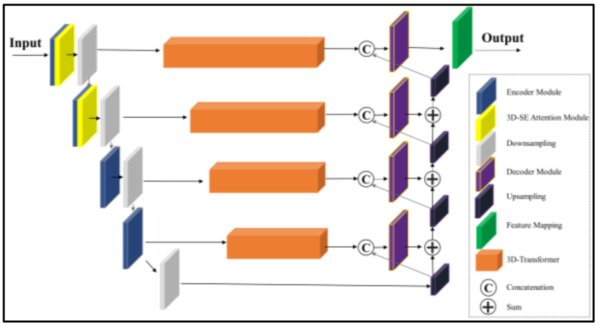

Figure 1. Overview of proposed dual-attention model.

3. Results

Experiments are conducted on NIH-TCIA dataset with 4-fold cross-validation. The results of our proposed model compared with state-of-the-art pancreas segmentation methods are listed in Table 1. The results indicate that the proposed model outperforms other advanced segmentation models with an average DSC of 85.82% ± 4.52%, and the best sample reached a segmentation performance of 92.12%. It was demonstrated that the proposed model is able to segment the pancreas tissue effectively and robustly.

We calculated the 3D surface distance of each point on predicted mask to ground truth mask, and then reconstructed the set of surface distances on real pancreatic tissue. The reconstruction is displayed in Figure 2. The closer the color to green in the figure, the smaller the deviation is. It can be observed that the model proposed in this paper segments the pancreas structure but also precisely predicts the boundary details.

Table 1. Comparison of state-of-the-art methods on NIH-TCIA dataset.

Method	Mean DSC(%)	Max DSC(%)	Min DSC(%)
Zhu et al.[10]	84.50±4.86	91.45	69.62
Xia et al.[11]	84.63±5.07	91.57	61.58
Ma et al.[12]	85.32±4.19	91.47	71.04
Chen et al.[13]	85.22±4.07	91.36	**71.40**
Ours	**85.82±4.32**	**92.12**	70.54

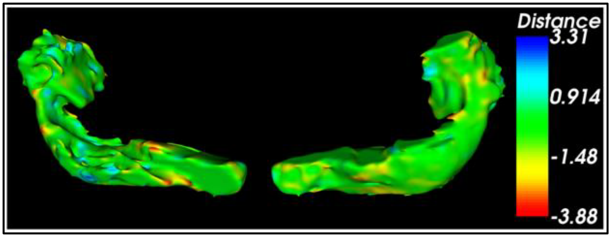

Figure 2. 3D surface distance reconstruction of predicted pancreas and ground truth.

4. Discussion

Table 2 shows the ablation study results of the proposed dual-attention structure. As can be seen, by introducing channel attention module, the model substantially improves in the mean DSC, mIoU, precision and recall metrics over baseline. Furthermore, the fusion model with spatial attention module is further improved, achieving a mIoU score of 71.13%. It is also evident by the precision metric that pancreatic tissue is more accurately identified. In summary, both of the attention modules introduced in this study provide an effective improvement in the pancreas segmentation task.

Table 2. Performance comparison of models with different attention modules.

Model	Mean DSC(%)	mIoU(%)	Precision(%)	Recall(%)
3D ResUNet	75.98±11.16	49.5	68.35	64.26
3D ResUNet+ Channel Attention	83.13±5.88	61.26	78.63	73.50
3D ResUNet+ Dual Attention	**85.82±4.32**	**71.13**	**84.30**	**82.00**

5. Conclusions

In this paper, we design a dual-attention model fusing CNN and Transformer for pancreas segmentation. The robust image comprehension capability of CNN is utilized to enhance feature representation by introducing channel attention mechanism. Moreover, the Transformer structure is employed to aggregate features at different scales, and establish long-term relationship among features in the spatial information level to ensure the effective activation of global features. The validation results demonstrate that this work achieves meaningful pancreas segmentation performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 12101571, No. 82172069) and the Key Research Project of Zhejiang Laboratory Grants (No. 2022ND0AC01).

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022 Jan;72(1):7-33. doi: 10.3322/caac.21708. Epub 2022 Jan 12. PMID: 35020204.

[2] Zhou Y, Xie L, Fishman EK, Yuille AL. Deep supervision for pancreatic cyst segmentation in abdominal CT scans. InMedical Image Computing and Computer Assisted Intervention− MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 2017 Sep (pp. 222-230). Cham: Springer International Publishing.

[3] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. 2018 Apr.

[4] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InMedical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 2015, October, Proceedings, Part III 18 2015 (pp. 234-241). Springer International Publishing.

[5] Li J, Lin X, Che H, Li H, Qian X. Pancreas segmentation with probabilistic map guided bi-directional recurrent UNet. Phys Med Biol. 2021 May 20;66(11). doi: 10.1088/1361-6560/abfce3. PMID: 33915526.

[6] 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. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 8280-8289).

[7] Wang W, Song Q, Feng R, Chen T, Chen J, Chen DZ, Wu J. A fully 3D cascaded framework for pancreas segmentation. In2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) 2020 Apr (pp. 207-211). IEEE.

[8] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017.

[9] Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B. and Summers, R.M., 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In Medical Image Computing and Computer-Assisted Intervention--MICCAI 2015: 18th International Conference, Munich, Germany, 2015, October, Proceedings, Part I 18 (pp. 556-564). Springer International Publishing..

[10] Zhu Z, Xia Y, Shen W, Fishman E, Yuille A. A 3D coarse-to-fine framework for volumetric medical image segmentation. In 2018 International conference on 3D vision (3DV) 2018 Sep (pp. 682-690). IEEE.

[11] Xia Y, Xie L, Liu F, Zhu Z, Fishman EK, Yuille AL. Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net. InMedical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, 2018,September, Proceedings, Part IV 11 2018 (pp. 445-453). Springer International Publishing.

[12] Ma J, Lin F, Wesarg S, Erdt M. A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation. InMedical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, 2018, September, Proceedings, Part IV 11 2018 (pp. 480-487). Springer International Publishing.

[13] Chen H, Wang X, Huang Y, Wu X, Yu Y, Wang L. Harnessing 2D networks and 3D features for automated pancreas segmentation from volumetric CT images. InMedical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, 2019, October, Proceedings, Part VI 22 2019 (pp. 339-347). Springer International Publishing.