Original Article

Subset selection strategy-based pancreas segmentation in CT

Yi Huang[1], Jing Wen[1], Yi Wang[1], Jun Hu[2], Yizhu Wang[3], Weibin Yang[4]

[1]School of Computer Science, Chongqing University, Chongqing, China; [2]Department of Neurology, the First Affiliate Hospital of Army Military Medical University, Chongqing, China; [3]Ziwei King Star Digital Technology Co., Ltd., Hefei, China; [4]Center for Intelligent Oncology, Chongqing University Cancer Hospital, Chongqing, China

Contributions: (I) Conception and design: Y Huang, J Wen, J Hu; (II) Administrative support: J Wen, J Hu; (III) Provision of study materials or patients: J Hu; (IV) Collection and assembly of data: Yizhu Wang, W Yang; (V) Data analysis and interpretation: Y Huang, J Wen, Yi Wang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Jing Wen. School of Computer Science, Chongqing University, Chongqing 400030, China. Email: wj@cqu.edu.cn; Jun Hu. Department of Neurology, the First Affiliate Hospital of Army Military Medical University, Chongqing, China. Email: hujuncq@163.com.

Background: Although convolutional neural network (CNN)-based methods have been widely used in medical image analysis and have achieved great success in many medical segmentation tasks, these methods suffer from various imbalance problems, which reduce the accuracy and validity of segmentation results.

Methods: We proposed two simple but effective sample balancing methods, positive-negative subset selection (PNSS) and hard-easy subset selection (HESS) for foreground-to-background imbalance and hard-to-easy imbalance problems in medical segmentation tasks. The PNSS method gradually reduces negative-easy slices to enhance the contribution of positive pixels, and the HESS method enhances the iteration of hard slices to assist the model in paying greater attention to the feature extraction of hard samples.

Results: The proposed methods greatly improved the segmentation accuracy of the worst case (samples with the worst segmentation results) on the public National Institutes of Health (NIH) clinical center pancreatic segmentation dataset, and the minimum dice similarity coefficient (DSC) was improved by nearly 5%. Furthermore, performance gains were also observed with the proposed methods in liver segmentation (the minimum DSC increased from 75.03% to 84.29%), liver tumor segmentation (the minimum DSC increased from 20.92% to 35.73%), and brain tumor segmentation (the minimum DSC increased from 21.97% to 30.38%) on different neural networks. These results indicate that the proposed methods are effective and robust.

Conclusions: Our proposed method can effectively alleviate foreground-to-background imbalance and hard-to-easy imbalance problems, and can improve segmentation accuracy, especially for the worst case, which guarantees the reliability of the proposed methods in clinical applications.

Keywords: Sample balancing; foreground-to-background imbalance; hard-to-easy imbalance

Submitted Aug 10, 2021. Accepted for publication Mar 02, 2022.
doi: 10.21037/qims-21-798
View this article at: https://dx.doi.org/10.21037/qims-21-798

Introduction

Accurate image segmentation plays an important role in both medical image analysis and computer-aided diagnosis (CAD). It is a prerequisite for many clinical applications, such as diabetes inspection and surgical planning. Recently, convolutional neural network (CNN)-based methods have achieved great success in many medical image segmentation tasks, such as liver segmentation (1-3), vessel segmentation (4-6), brain segmentation (7-9), and pancreas segmentation (10-13). However, these methods rely heavily on massive amounts of annotated data and inevitably encounter various imbalance problems in the training process. CNNs can be disrupted, and consequently, the final segmentation results become inaccurate, especially around the boundary regions.

Medical images include magnetic resonance imaging

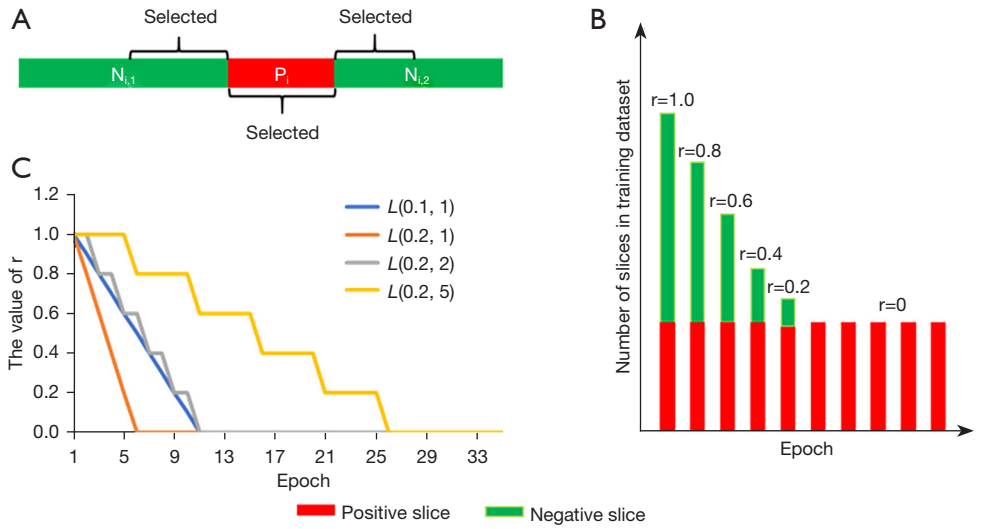

Figure 1 Example of the PNSS method. (A) The sample selection method on an individual 3D CT scan. (B) The sample selection method in the training process. (C) How $L(\alpha, \beta)$ regulates the r value in the training process. Red and green denote the positive slices and negative slices, respectively, while r represents the selected ratio on negative slices. PNSS, positive-negative subset selection; CT, computed tomography.

(MRI), computed tomography (CT), and X-ray scans, among others, which are visual representations of the internal organs or tissue functions of the human body. Due to the continuity of human tissues and the differences in various medical image acquisition equipment, the boundary imaging of some tissues and organs is not obvious. In addition, the size and morphology of various tissues and organs are varied. For patients with certain diseases, their organ morphology on imaging is distinct, and image analysis of these cases is more difficult.

In this paper, we focus on two common imbalance problems in medical segmentation tasks: foreground-to-background imbalance and hard-to-easy imbalance. Foreground-to-background imbalance manifests as an extreme inequality between the number of positive pixels and negative pixels, which is caused by an abundance of negative slices and small target regions in positive slices. To address these problems, some researchers (14) have used a coarse-to-fine strategy to remove abundant negative pixels (1,10,14). Regions of interest (ROIs) are cropped from coarse segmentation results to enhance the contribution rate of positive pixels in the fine stage (Figure S1). However, the fine segmentation results are independent of the coarse segmentation results. Omission or inaccurate segmentation of ROIs often leads to irreparable loss and degrades the final segmentation performance (15).

Different from foreground-to-background imbalance, distinguishing between hard samples and easy samples lacks suitable rules, and the segmentation difficulty of slices may vary according to different tasks (e.g., an easy slice in a liver segmentation task may be a hard one in a tumor segmentation task), so it is almost impossible to discriminate hard samples or easy samples before the training process. We observed that hard samples can give rise to self-defects (e.g., morphological variation, lesions, and tumors, among others). Therefore, accurate segmentation of these cases is of great significance in clinical application. To alleviate this problem, focal loss is used to enhance the contribution of hard pixels by adjusting their weight in loss function (16). However, these methods are unable to address the imbalance problem between hard slices and easy slices. Zhang *et al.* used the pretraining method to handle hard-to-easy imbalance in liver segmentation (17), but it required a two-stage training strategy and extra datasets.

To address the above problems, we propose two simple but effective sample balancing methods: positive-negative subset selection (PNSS) and hard-easy subset selection (HESS). Unlike traditional example mining-based methods such as online hard example mining (OHEM), they directly abandon most negative examples, which leads to an inefficient training effect (18). In PNSS, we gradually remove negative slices to enhance the contribution of positive pixels (*Figure 1*). In addition, it is observed that the majority of negative slices are easier to segment than positive ones (*Figure 2*); therefore, PNSS could also alleviate the hard-to-easy imbalance problem to a certain degree. Inspired by our previous study (17), we propose the HESS

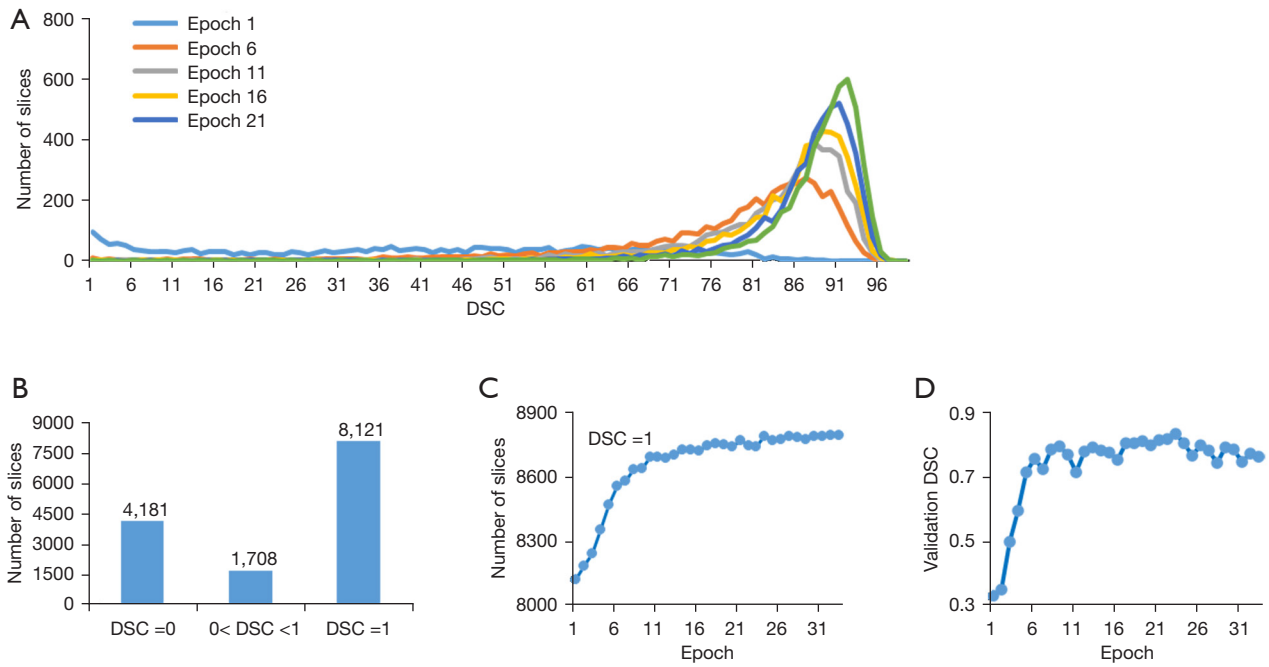

Figure 2 DSC information of training slices in training epochs. (A) Overall DSC distribution varies with training epochs. (B) DSC distribution during the initial training phase. (C) The number of slices (DSC =1) in each training epoch. (D) The validation DSC in each epoch. DSC, dice similarity coefficient.

method to alleviate the hard-to-easy imbalance problem, especially hard-to-easy imbalance between positive slices. HESS increases the iteration of hard slices to assist the model in paying greater attention to feature extraction. Unlike our research group's previous work (17), PNSS and HESS emphasize dynamic adjustment. Especially for HESS, the threshold for distinguishing hard and easy slices between positive slices is automatically changed according to the training state instead of setting a fixed value.

Among different abdominal organ segmentation tasks, accurate pancreas segmentation is especially challenging, as it may suffer from serious sample imbalance problems. For example, the pancreas often occupies a very small portion (e.g., <0.5%) of the entire CT volume, and the abundant negative pixels may overwhelm the contribution of target pixels in batch-manner training epochs, which can then lead to inaccurate results, especially for the pixels around boundary areas (10). Moreover, the pancreas has high interpatient variability in terms of its shape, position, and size, which could hinder the performance of the model.

In this work, we evaluated the effectiveness of our proposed method on the National Institutes of Health (NIH) dataset, a widely used pancreas segmentation CT dataset (13). We found that our method greatly enhanced the segmentation performance of the model on the worst case in the NIH dataset and achieved competitive performance compared to other state-of-the-art methods. In addition, performance gains in different 2D networks [i.e., fully convolutional network (FCN) (19), U-Net (20), high-resolution network (HRNet) (21), and transformers-based U-Net framework (TransUNet) (22)] and different segmentation tasks (i.e., liver and liver tumor segmentation from CT scans, brain tumor segmentation from MRI) validated its effectiveness and generalizability. We present the following article in accordance with the STARD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-21-798/rc).

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Datasets and preprocessing

(I) The NIH (https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT) dataset is a widely used pancreas segmentation CT dataset, which contains

82 contrast-enhanced abdominal CT scans with corresponding ground truth. The CT scans have a resolution of 512×512 pixels, the number of slices varies from 181 to 466, and the slice thickness is between 1.5 and 2.5 mm. The image intensity values for all CT scans are truncated to the range of –150 to 250 Hounsfield units (HU) to remove the irrelevant details. We performed a 4-fold cross-validation in a random split from 82 patients for training and testing folds, where each testing fold had 21, 21, 20, and 20 cases, respectively. In each round of 4-fold standard cross-validation, we employed 3 folds of data as training cases and the remaining fold for testing.

(II) The Liver Tumor Segmentation challenge (LiTS) dataset (https://competitions.codalab.org/competitions/17094) is available from the Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2017 liver tumor segmentation challenge, providing 131 CT scans with manual annotation labels. For image preprocessing, intensity values for all CT scans are truncated to the range of –200 to 250 HU. We randomly selected 105 CT scans for training and the remaining 26 volumes for testing.

(III) The Brain Tumors Task of Medical Segmentation Decathlon challenge (BT-MSD) dataset (http://medicaldecathlon.com/) provides 484 multiparametric MRI scans with manually annotated brain tumor subregions (i.e., edema, enhancing, and nonenhancing tumor) (23). In our study, each multiparametric sequence included native (T1), post-Gadolinium (Gd) contrast T1-weighted (T1-Gd), native T2-weighted (T2), and T2 fluid-attenuated inversion recovery (T2-FLAIR) volumes. All MRI scans were coregistered to a reference atlas space using the SRI24 brain structure template and resampled to an isotropic voxel resolution of 1 mm^3. We randomly split the dataset with annotations into 436 MRI volumes for training and the remaining 48 volumes for testing.

In this paper, we employed the dice similarity coefficient (DSC), precision, and recall to evaluate the segmentation accuracy,

$$DSC = \begin{cases} \frac{2\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i}, & \sum_i y_i \neq 0 \quad or \quad \sum_i \hat{y}_i \neq 0 \\ 1, & \sum_i y_i = 0 \quad and \quad \sum_i \hat{y}_i \neq 0 \end{cases} \quad [1]$$

where y_i and $\hat{y}_i$ denote the ground truth and prediction result, respectively.

$$precision = \frac{|V_s \cap V_g|}{|V_s|} \quad [2]$$

$$recall = \frac{|V_s \cap V_g|}{|V_g|} \quad [3]$$

Let V_g denote the voxel set of automatic segmentation volume, and V_g denote the voxel set of ground-truth volume.

Network architecture

The 2D network architecture [residual network (ResNet-34) variants] for pancreas segmentation was employed in this study (*Figure 3*). This included an encoder path with 5 resolution steps on the left and a decoder path with 3 resolution steps on the right. The left part employs 2D convolution layers and residual blocks to learn the low-level features and high-level features of medical images. We performed a convolution operation with a stride of 2 to reduce the spatial resolution of the feature map by half in the encoder path. Inspired by FCN-8s (19), we had the right part decompress the extracted high-level features into a finer resolution through deconvolution layers. In addition to encoder path and decoder path, we imposed a convolution block to bridge the skip connections between low-level features and high-level features. Moreover, the proposed network contained 3 auxiliary loss layers and 1 main loss layer. For auxiliary loss layers, deconvolution layers were applied to upsample feature maps to have the same spatial resolution as the inputs.

Loss function

The following binary cross entropy function was employed in this work:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log \hat{y}_i + (1-y_i)(1-\log \hat{y}_i)\right) \quad [4]$$

where y represents the ground truth, $\hat{y}$ denotes the predicted segmentation results, and y_i and $\hat{y}_i$ indicate the label and predicted probability for voxel i, respectively. The overall loss is formulated as follows:

$$L_{total} = L(y, \hat{y}_{main}) + \beta_1 L(y, \hat{y}_{aux1}) + \beta_2 L(y, \hat{y}_{aux2}) + \beta_3 L(y, \hat{y}_{aux3}) \quad [5]$$

where $L(y, \hat{y}_{main})$, $L(y, \hat{y}_{aux1})$, $L(y, \hat{y}_{aux2})$, and $L(y, \hat{y}_{aux3})$ denote the main loss and 3 auxiliary losses, respectively; and β_1, β_2, and β_3 are the balanced weights and are set as 0.2, 0.4,

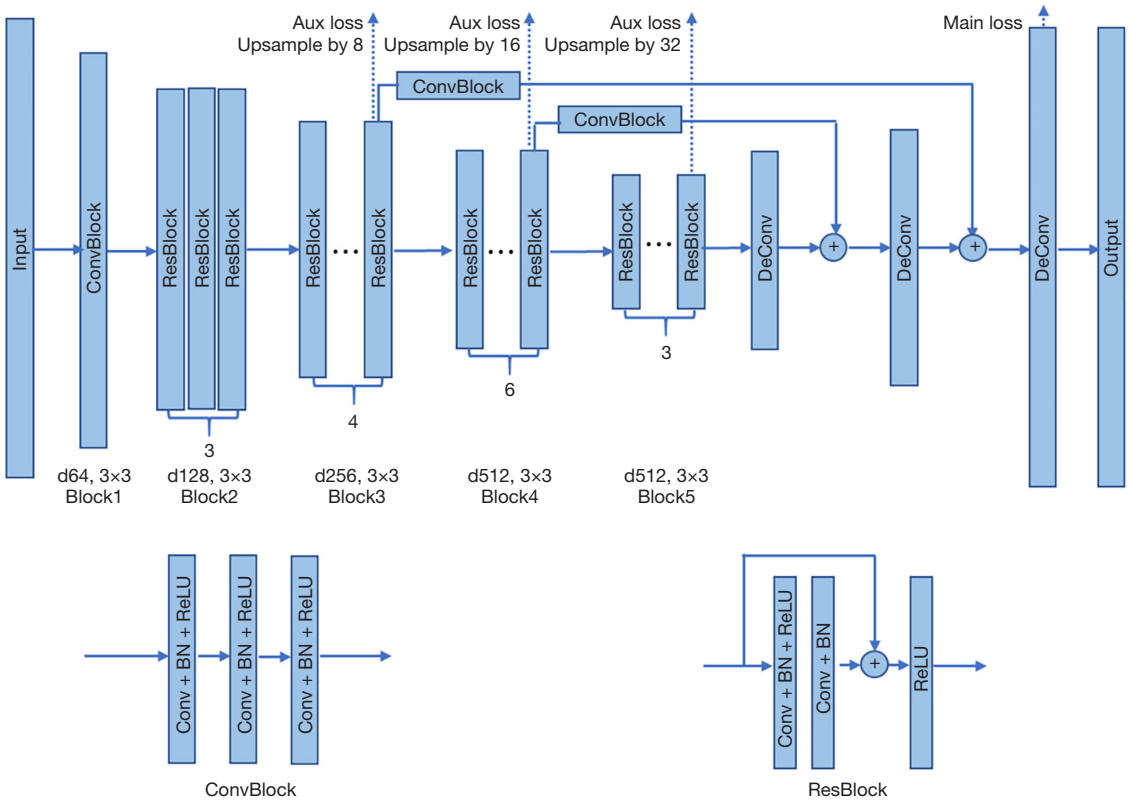

Figure 3 Illustration of the 2D convolutional network architecture of our method. DeConv, deconvolution; Conv, convolution layer; BN, batch normalization; ReLU, rectified linear units; ConvBlock, convolutional block; ResBlock, residual block.

and 0.8, respectively.

Proposed sample balancing methods

PNSS

With the assumption that there are m cases in the training dataset, X_i denotes the i-th case, $i=1,\cdots,m$, and n slices in X_i, $x_{i,j}$ denotes the *j-th* slices in X_i, $i=1,\cdots,n$. As shown in *Figure 1A*, there are two types of slices in X_i: negative slices and positive slices. Generally, the positive slices are at the middle, denoted as P_i, while the negative slices are at both ends, denoted as $N_{i,1}$ and $N_{i,2}$, respectively. We assume that there are k positive slices in X_i, and $x_{i,l}$ is the first positive slice, $N_{i,j}=\{X_{i,1},\cdots,x_n\}$, $P_{i,j}=\{X_{i,l},\cdots,x_{i,l+k-1}\}$, $N_{i,2}=\{x_{i,l+k},\cdots,x_n\}$. In medical volumetric images, there is a strong correlation between adjacent slices, which is useful in information complementation. In order to effectively extract the feature information of positive slices, all positive slices and their adjacent continuous negative slices were selected in each CT scan (*Figure 1A*). We set the selected ratio on negative slices in CT scans as r, $r\in[0,1]$. The selected negative slices are denoted as $SN_{i,1}$ and $SN_{i,2}$ along with slice number $n_{i,1}$ and $n_{i,2}$, respectively:

$$n_{i,1}=\lceil(l-1)\times r\rceil \quad [6]$$

$$n_{i,2}=\lfloor(n-l-k+1)\times r\rfloor \quad [7]$$

where $SN_{i,1}=\{x_{i,l-n_{i,1}},\cdots,x_{i,l-1}\}$ and $SN_{i,2}=\{x_{i,l+k},\cdots,x_{i,l+k+n_{i,2}-1}\}$. Selected samples in X_i are denoted as $SX_i^r=\{SN_{i,1},P_i,SN_{i,2}\}$, and all selected slices in the training dataset are denoted as $D^r=\{SX_1^r,\cdots,SX_n^r\}$. As the selection is between positive slices and negative slices, the proposed method is named "positive-negative subset selection". We introduced a function $r=L(\alpha,\beta)$ to regulate r in the training process, which represents the value of r reduced by α every β epochs until *r=0* (*Figure 1B*).

HESS

Following our previous work (17), we used the DSC to assess the difficulty of slices in the training process, with a

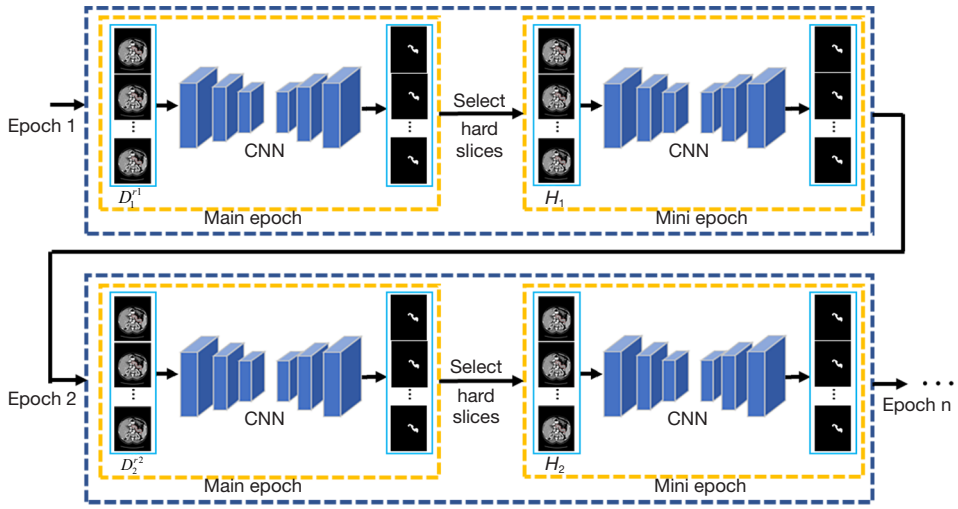

Figure 4 Illustration of the pipeline of the HESS method. CNN, convolutional neural network; HESS, hard-easy subset selection.

higher DSC indicating easier segmentation. According to Eq. [1], if 0< DSC <1, the corresponding slice is a positive sample. If DSC =1, the corresponding slice is a negative sample. However, we cannot determine the corresponding slice if DSC =0, as the DSC of both positive and negative slices could reach 0.

In this paper, we propose the HESS method, which alleviates hard-to-easy imbalance by increasing the iteration of hard slices. We denote the training dataset at *n-th* epoch as D_n^r with m slices, $D_n^r = \{d_1, \cdots, d_m\}$ and their corresponding DSC score as $S_n^r = \{s_1, \cdots, s_m\}$, with r representing the selection ratio on negative slices. As shown in *Figure 4*, after every epoch (called main epoch in this paper), we distinguished training slices by thresholding. Hard slices were selected as H_n and then randomly combined into a batch manner to participate in the training process (called mini epoch in this paper). For convenience, we defined samples as hard slices if 0.5< DSC < s_{med}, where s_{med} denotes the average DSC of training slices and is calculated as follows:

$$s_{med} = \frac{1}{m}\left(s_1 + \cdots + s_m\right) \quad [8]$$

We considered slices with DSC <0.5 to be outliers due to the fact that they exist stably even when the model is converged (*Figure 2A*).

Implementation details

Our proposed networks were implemented in the public TensorFlow platform. During the training process, we used stochastic gradient descent (SGD) with a batch size of 6 and a momentum of 0.9. Inspired by Li *et al.* (2), the initial learning rate was set at 10^{-4} and decayed by multiplying $\left(1-\frac{iterations}{total\ iterations}\right)^{0.9}$. We used the early-stop strategy in the training process, and the patience was set as 10 epochs. For data augmentation, we adopted random flip and mirror for all training slices to alleviate the overfitting problem.

The 3D network architecture for pancreas segmentation was employed in this work, which is the same as that used by Zhao *et al.* (14). We conducted experiments using a 32×96×96 patch size. Two deep networks of the same architecture were trained with downsampled and original 3D CT scans for the purpose of coarse ROI definition and refined segmentation.

Results

Analysis of hard-to-easy imbalance problems in the training process

To overcome the hard-to-easy imbalance problem, we recorded the DSC of training slices in each main training epoch. In the early training stage, approximately 92% of negative slices achieved perfect segmentation results (DSC =1), while most positive slices still had poor DSC results during this period (*Figure 2B*). The number of perfectly segmented slices continuously increased and tended to be stable in the later training stage (*Figure 2C*).

Table 1 Effectiveness of the PNSS method

Method	DSC (%)		
	Mean ± std	Max	Min
ResNet (baseline)	82.70±7.01	88.76	67.92
ResNet + *L*(0.1, 1)	82.37±6.79	87.62	71.95
ResNet + *L*(0.1, 3)	83.03±6.33	88.85	71.52
ResNet + *L*(0.1, 5)	82.67±6.85	88.29	69.24
ResNet + *L*(0.1, 7)	82.79±7.83	88.93	67.69
ResNet + *L*(0.1, 9)	82.16±7.33	88.67	68.72
ResNet + *L*(0.2, 1)	81.29±8.53	87.95	72.12
ResNet + *L*(0.2, 3)	81.89±7.83	88.45	70.69
ResNet + *L*(0.2, 5)	83.12±6.21	89.13	71.64
ResNet + *L*(0.2, 7)	82.75±7.12	88.87	70.36
ResNet + *L*(0.2, 9)	82.03±6.53	89.05	69.05

L(α, β) means the value of *r* reduced by α every β epochs until it reaches zero, *r* represents the selected ratio on negative slices. PNSS, positive-negative subset selection; DSC, dice similarity coefficient; std, standard deviation; max, maximum; min, minimum; ResNet, residual network.

The hard-to-easy imbalance problem also exists between positive slices. As shown in *Figure 2A*, the DSC score of positive slices has a left-skewed distribution: most of positive slices gather into the high DSC region, some in the low DSC region (0.5< DSC < s_{med}, hard samples), and just a few slices in the very low DSC region (DSC ≤0.5, outliers). With iteration increasing, a greater number of positive slices could obtain a high DSC. However, there were still some hard slices with under-segmentation even when the model was converged (*Figure 2D*).

The effectiveness of the PNSS method with different parameters

As shown in *Table 1*, rapidly taking away abundant negative slices could improve the minimum DSC, but the mean DSC and maximum DSC decreased. For example, at *L*(0.2, 1), the mean DCS and maximum DSC decreased by 2% and 1%, respectively, while the minimum DSC improved by 6%. When β increased, which means that the removal speed of negative slices decreased, the mean DSC, maximum DSC, and minimum DSC improved compared to baseline, but the improvement effect of the minimum DSC decreased. However, when β gradually increased, there was almost no difference between baseline and HESS. At *L*(0.2, 5) we obtained the best segmentation performance, and the mean DSC, maximum DSC, and minimum DSC improved by 0.5%, 0.4%, and 5%, respectively. Therefore, we used *L*(0.2, 5) for medical segmentation experiments in this paper.

The effectiveness of the HESS method

In this paper, we propose HESS, which easily handles hard-to-easy imbalance, especially for hard-to-easy imbalance between positive slices. HESS can enhance the contribution of hard slices by improving their iteration in the training stage. As shown in *Table 2*, we observed that using HESS alone could enhance the segmentation accuracy under the worst case, which outperformed the baseline by nearly 4% in terms of the minimum DSC. Further performance gains could be obtained by applying HESS and PNSS simultaneously, and the mean DSC, maximum DSC, and minimum DSC improved by nearly 1%, 1%, and 5%, respectively. The mean values of the other two indices, namely precision and recall, increased from 81.74% to 83.60% and from 79.00% to 82.58%, respectively, compared to baseline.

Moreover, it was observed that applying both HESS and PNSS could speed up the convergence of the training model (*Figure 5*), which means that HESS and PNSS can help the model pay greater attention to the extraction of ROIs.

We also conducted experiments on the NIH dataset using the method in Zhang *et al.* (17). We set the threshold to 0.6783, the minimum dice score of baselines, to distinguish hard and easy samples.

As can be seen in *Table 2*, the method in Zhang *et al.* (17) was slightly better than the baseline. Although Zhang *et al.*'s (17) overall segmentation performance of the method was at the same level as that of HESS and PNSS, HESS and PNSS had better segmentation performance for the worst case. Moreover, the training set partition strategy of for Zhang *et al.*'s method (17) may render a few samples unable to participate in the training process, which may lead to the insufficient learning of some details for the model.

Results of the experiments on 2D models

To further verify the effectiveness of the proposed methods, we conducted experiments on different classical 2D CNN structures, including HRNet, DeepLabv3+, and TransUNet. To evaluate the segmentation performance, 3 metrics (DSC, precision, and recall) were used. As shown

Table 2 Detailed performance of the proposed methods on the NIH dataset

Method	DSC (%)	Precision (%)	Recall (%)
ResNet	82.53±6.84 [67.83, 88.71]	83.74±6.47 [69.23, 91.47]	79.00±8.11 [66.48, 92.17]
ResNet + methods (17)	82.84±6.49 [70.23, 89.24]	82.83±7.02 [71.64, 91.27]	82.15±8.34 [69.47, 92.31]
ResNet + PNSS	83.24±6.26 [71.79, 89.64]	83.10±6.13 [71.47, 90.42]	79.86±7.64 [71.93, 90.39]
ResNet + HESS	82.91±6.74 [70.72, 89.03]	84.49±5.25 [71.81, 91.11]	82.30±7.95 [69.66, 92.31]
ResNet + PNSS + HESS	83.67±4.71 [72.51, 90.04]	85.60±4.89 [72.11, 92.15]	82.58±7.05 [72.81, 94.57]

The performance is described by mean ± std [min, max]. NIH dataset, the public National Institutes of Health clinical center pancreatic segmentation dataset; DSC, dice similarity coefficient; ResNet, residual network; PNSS, positive-negative subset selection; HESS, hard-easy subset selection; std, standard deviation; max, maximum; min, minimum.

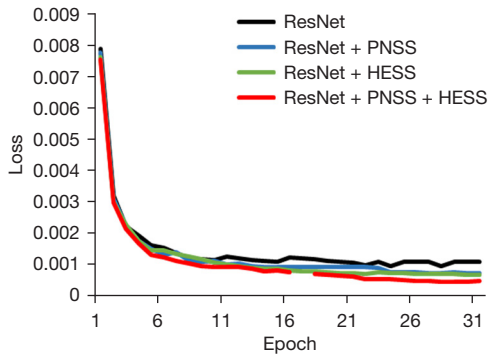

Figure 5 Training losses of 2D ResNet with and without PNSS, 2D ResNet with PNSS, and 2D ResNet with both methods. ResNet, residual network; PNSS, positive-negative subset selection; HESS, hard-easy subset selection.

in *Table 3*, HESS and PNSS could effectively improve the segmentation performance on various models, which was also observed on ResNet. For example, when using both PNSS and HESS methods simultaneously on TransUNet, the mean values of the DSC, precision, and recall increased from 81.72% to 83.52%, from 82.23% to 84.51%, and from 81.38% to 82.41%, respectively, especially for the worst case. The indices all increased by nearly 3%, which demonstrates the effectiveness of the proposed methods.

Results of the experiments on 3D models

The experiments were also conducted on 3D segmentation architecture (3D-ResUNet). As shown in *Table 4*, the values of the 3D-ResUNet using both PNSS and HESS were significantly higher than those of the 3D-ResUNet with PNSS/HESS and without our method, which demonstrates the effectiveness of our proposed method on 3D-ResUNet.

We observed that the effect of HESS and PNSS in 3D-ResUNet was different from that of ResNet (*Tables 2*,*4*). Compared with PNSS, HESS could improve the overall segmentation performance more significantly in 3D architecture, which may be due to the spatial continuity of the 3D patch. HESS focuses on the feature extraction of hard slices so that more independent connected domains can be connected. The 3D segmentation architecture has more opportunities to perform this operation on the z-axis, which makes the overall segmentation performance improvement more obvious.

Comparison with various imbalance strategies for pancreas segmentation

To further verify the effectiveness of our proposed methods, we performed experiments with various imbalance strategies, including focal loss, weighted cross entropy, and data oversampling for positive slices (the number of positive slices and negative slices were made equal). We set α=0.25, γ=2 for focal loss, and the best results were obtained. According to the proportion of positive and negative slices, we set the weight value as [0.75, 1.5] for weighted cross-entropy.

The above imbalance strategies were not better than our method in improving the overall segmentation performance (*Table 5*). There was not only foreground-to-background imbalance in medical images, but also hard-to-easy imbalance in positive slices. The above methods can alleviate foreground-to-background imbalance to some extent, but they ignore the hard-to-easy imbalance in positive slices, which makes the model focus more on the feature extraction of easy slices in positive slices and ignore more details.

Table 3 Segmentation results on the NIH dataset with HRNet, DeepLabv3+, and TransUNet

Model	PNSS	HESS	DSC (%)	Precision (%)	Recall (%)
HRNet	–	–	80.68±6.89 [67.28, 87.11]	80.62±6.68 [66.36, 87.96]	81.40±7.53 [68.22, 91.46]
	√	–	81.04±6.93 [70.23, 88.34]	83.51±7.21 [69.17, 92.62]	79.69±6.76 [71.32, 89.16]
	–	√	80.92±6.14 [69.47, 88.71]	83.92±6.82 [69.45, 91.92]	78.73±7.82 [69.49, 89.49]
	√	√	81.16±5.53 [71.47, 89.01]	83.74±6.21 [70.15, 92.47]	81.35±6.72 [72.84, 91.20]
DeepLabv3+	–	–	79.15±7.37 [67.12, 85.92]	82.99±7.23 [74.65, 90.52]	76.96±8.15 [60.97, 89.90]
	√	–	79.87±6.91 [70.02, 86.73]	81.69±5.38 [71.63, 93.04]	79.00±9.85 [68.48, 93.37]
	–	√	80.03±6.11 [70.13, 86.93]	83.30±5.60 [71.69, 93.85]	78.12±8.7 [68.61, 91.89]
	√	√	80.87±6.14 [71.42, 87.45]	84.25±5.68 [76.10, 93.43]	77.75±6.43 [67.28, 91.72]
TransUNet	–	–	81.72±5.93 [70.68, 88.23]	83.83±4.82 [71.61, 93.21]	81.38±6.97 [69.77, 92.92]
	√	–	82.57±5.27 [71.54, 89.97]	82.23±5.25 [71.52, 91.05]	82.30±7.95 [71.56, 94.91]
	–	√	83.22±5.21 [72.15, 89.10]	83.10±4.54 [71.47, 90.22]	81.95±8.47 [72.84, 92.73]
	√	√	83.52±5.03 [72.47, 89.56]	84.51±5.39 [73.11, 92.15]	82.41±8.12 [73.83, 93.68]

The performance is described by mean ± std [min, max]. NIH dataset, the public National Institutes of Health clinical center pancreatic segmentation dataset; HRNet, high-resolution network; TransUNet, transformers-based U-Net framework; PNSS, positive-negative subset selection; HESS, hard-easy subset selection; DSC, dice similarity coefficient; std, standard deviation; max, maximum; min, minimum.

Table 4 Segmentation results on the NIH dataset with 3D models

Model	PNSS	HESS	DSC (%)	Precision (%)	Recall (%)
3D-ResUNet	–	–	83.56±6.04 [69.54, 88.86]	83.64±6.21 [71.11, 90.15]	83.49±7.23 [68.04, 93.92]
	√	–	83.81±6.71 [70.23, 88.34]	83.92±6.47 [71.26, 90.34]	82.53±7.14 [69.23, 93.71]
	–	√	85.61±5.47 [71.34, 91.91]	85.92±7.11 [73.85, 90.39]	83.87±6.47 [70.83, 94.12]
	√	√	86.32±4.32 [72.92, 92.31]	85.52±6.03 [72.47, 92.41]	84.51±6.72 [73.37, 94.38]

The performance is described by mean ± std [min, max]. NIH dataset, the public National Institutes of Health clinical center pancreatic segmentation dataset; PNSS, positive-negative subset selection; HESS, hard-easy subset selection; DSC, dice similarity coefficient; 3D-ResUNet, 3D residual U-Net framework; std, standard deviation; max, maximum; min, minimum.

Table 5 Segmentation results on the NIH dataset with various imbalance strategies

Strategy	DSC (%)	Precision (%)	Recall (%)
ResNet + PNSS + HESS	83.67±4.71 [72.51, 90.04]	85.60±4.89 [72.11, 92.15]	82.58±7.05 [72.81, 94.57]
ResNet + focal loss	82.67±4.76 [70.98, 88.42]	86.23±6.82 [73.29, 94.97]	77.85±7.71 [66.28, 91.60]
ResNet + weighted cross entropy	82.57±4.82 [72.23, 87.93]	85.23±6.09 [71.63, 92.25]	80.37±7.77 [67.25, 89.74]
ResNet + data oversampling	82.51±5.87 [72.94, 88.31]	81.57±6.51 [71.85, 91.24]	80.64±6.23 [67.94, 91.45]

The performance is described by mean ± std [min, max]. NIH dataset, the public National Institutes of Health clinical center pancreatic segmentation dataset; DSC, dice similarity coefficient; PNSS, positive-negative subset selection; HESS, hard-easy subset selection; std, standard deviation; max, maximum; min, minimum; ResNet, residual network.

Table 6 Evaluation of different methods on the NIH dataset

Method	DSC (%)		
	Mean ± std	Max	Min
2D Roth *et al.* (13)	78.01±8.20	88.65	34.11
2D Zhou *et al.* (10)	82.65±5.47	90.85	63.02
2D Cai *et al.* (24)	82.40±6.70	90.10	60.00
3D Zhu *et al.* (25)	84.59±4.86	91.45	69.62
3D Zhao *et al.* (14)	85.99±4.51	91.20	57.20
3D Fang *et al.* (15)	85.46±4.80	92.24	67.03
2D ours (coarse)	83.67±4.71	90.04	72.51
2D ours (fine)	84.30±4.31	90.75	73.24
3D ours (coarse)	84.16±4.94	91.47	69.57
3D ours (fine)	86.32±4.32	92.31	72.92

NIH dataset, the public National Institutes of Health clinical center pancreatic segmentation dataset; DSC, dice similarity coefficient; std, standard deviation; max, maximum; min, minimum.

Comparison with state-of-the-art methods

After coarse-to-fine processing, the 2D network results showed that the segmentation accuracy of the worst case compared to the best case was from 69.62% to 73.24%, which was more than a 5% advancement (*Table 6*). Moreover, compared with other 2D network results, our proposed methods outperformed previous state-of-the-art methods by 2% in terms of the mean DSC. This may be attributable to the significantly improved segmentation accuracy of the worst cases.

Similarly, compared with other 3D methods, our methods also demonstrated a better segmentation performance on a 3D network. *Table 6* shows that the mean DSC and maximum DSC of the proposed methods were slightly higher than those of the 3D state-of-the-art method, which may be due to the improvement of the overall performance of the worst case and hard slices.

The segmentation results of 2D and 3D architecture are shown in *Figures 6*,*7*, respectively.

Generalizability of the proposed methods

Liver and liver tumor segmentation

Liver cancer is one of the most common malignancies worldwide and causes a massive number of deaths every year. Accurate segmentation of the liver and liver tumor from CT images is highly significant in clinical application, as it can assist doctors in making accurate disease condition assessments and surgical planning with precise contours. As shown in *Tables 7*,*8*, performance gains were observed on both FCN and UNet by adopting the proposed sample balancing methods. For liver segmentation, when both the PNSS and HESS methods were used simultaneously on FCN, the mean DSC and minimum DSC increased from 93.87% and 75.03% to 94.71% and 84.29%, respectively. For liver tumor segmentation, the mean DSC and minimum DSC increased by 4.82% and 62.32%, respectively. These results indicate that the proposed data balancing methods are beneficial, especially for hard CT scans. Segmentation results are shown in *Figure 8*.

Brain edema segmentation

For convenience, we conducted brain edema segmentation in our study by using T1-Gd on the BT-MSD dataset. As shown in *Table 9*, using the proposed sample balancing methods could effectively improve brain edema segmentation on both FCN and U-Net frameworks. For example, when PNSS and HESS were used simultaneously in the FCN framework, the mean DSC and minimum DSC increased from 70.69% and 22.77% to 71.58% and 32.03%, respectively. Segmentation results are shown in *Figure 9*.

Discussion

Findings and general discussion

Automatic medical image segmentation is crucial for clinical diagnosis, as it can provide the precise contour of organs and any lesion inside the anatomical segments of the organ, which assists doctors during the diagnosis process. In this work, we present two simple but effective sample balancing methods (PNSS and HESS) to address foreground-to-background imbalance and hard-to-easy imbalance problems in medical image segmentation. Through gradually removing negative slices and increasing the iteration of hard slices, segmentation performance for pancreas segmentation was improved, which demonstrates the effectiveness of the proposed methods. The performance gains in liver segmentation, liver tumor segmentation, and brain tumor segmentation with different CNN frameworks further validated the effectiveness and generalizability of the proposed methods.

Our proposed method makes the model focus more on hard slices and improves the overall performance of the

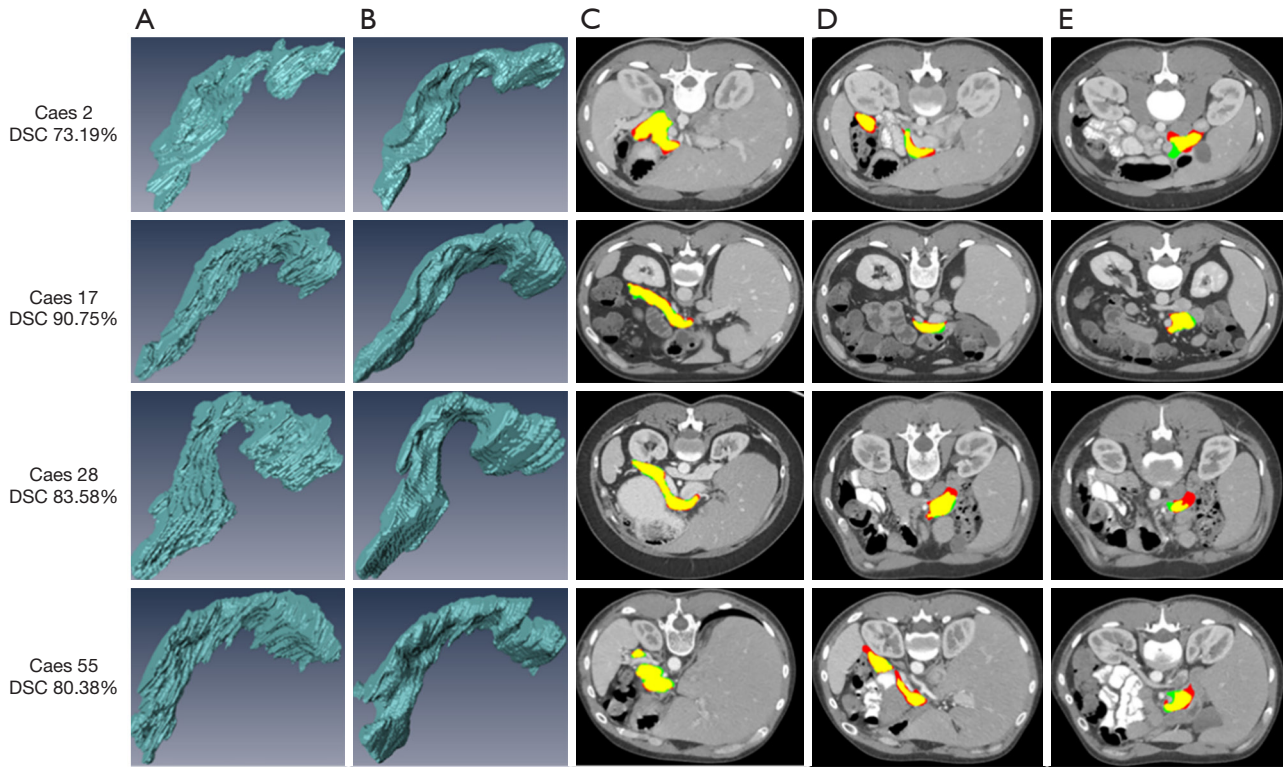

Figure 6 Examples of segmentation results of the proposed methods on the NIH dataset (2D segmentation architecture). (A,B) The 3D results of the ground truth and proposed methods, respectively. (C-E) The segmentation results on 2D slices, where red and yellow masks indicate the ground truth and prediction regions, respectively. DSC, dice similarity coefficient; NIH dataset, the public National Institutes of Health clinical center pancreatic segmentation dataset.

model, especially the segmentation performance of some worst cases. Compared with various modifications of the model by other methods, our method focuses on data enhancement, which is simple but more effective.

In clinical needs, special lesions (hard slices) are more difficult to distinguish than are ordinary ones. By improving the segmentation performance of special lesions, our method provides more help to doctors in diagnosing diseases.

Limitations

There are still some limitations in this work. The proposed methods were implemented only on the NIH, LiTS, and BT-MSD datasets and have not been verified on a large-scale dataset. Furthermore, the number of iterations of difficult samples is increased, which makes the training of the model relatively time-consuming. In addition, our method only focuses on the overall segmentation effect of difficult samples during training, ignoring the improvement of its boundary.

Future research

We observed that negative slices also have their contribution, especially in the early training stage. Furthermore, removing abundant negative slices could directly and rapidly lead to poor segmentation performance (*Table 1* and Table S1). Therefore, it is necessary to explore the roles of negative slices in future work.

Recently, transformer-based architectures have been widely used in image segmentation (26). They can use the attention mechanism to capture global contextual information to establish a long-distance dependence on the target, thereby extracting more powerful features. Many scholars have proposed methods that combine the advantage of the convolution and transformer block, making it possible to train on medical image datasets and obtain

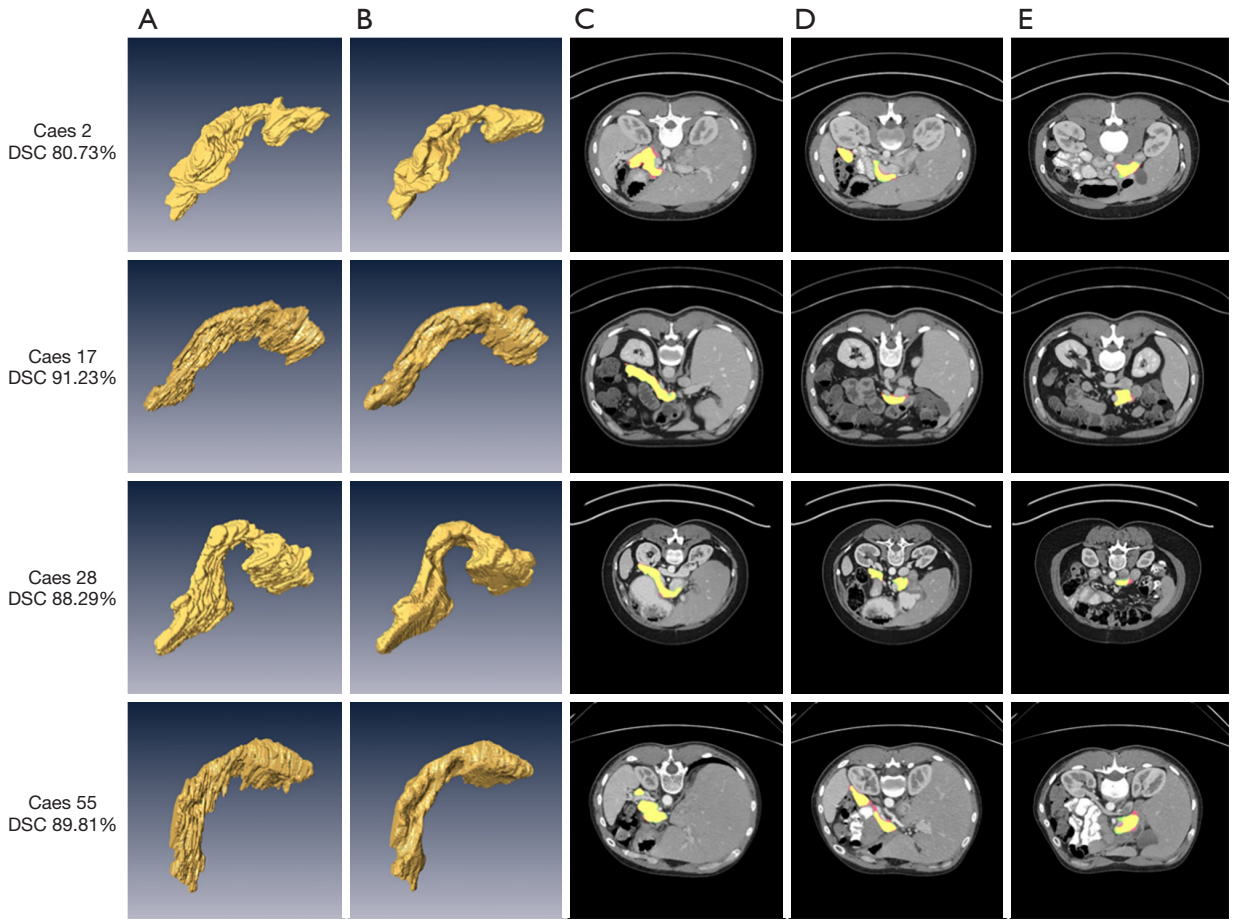

Figure 7 Examples of the segmentation results of the proposed methods on the NIH dataset (3D segmentation architecture). (A,B) The 3D results of the ground truth and proposed methods, respectively. (C-E) The segmentation results on 2D slices, where red and yellow masks indicate the ground truth and prediction regions, respectively. DSC, dice similarity coefficient; NIH dataset, the public National Institutes of Health clinical center pancreatic segmentation dataset.

Table 7 Segmentation results of the liver with different methods on the LiTS dataset

Model	PNSS	HESS	Liver, mean ± std [min, max]		
			DSC (%)	Precision (%)	Recall (%)
FCN	–	–	93.87±5.14 [75.03, 97.87]	92.54±6.23 [74.20, 98.12]	93.23±7.72 [75.88, 97.62]
	√	–	94.19±4.29 [81.11, 98.13]	94.32±5.47 [83.41, 98.32]	93.91±7.53 [78.93, 98.10]
	–	√	94.71±3.75 [83.73, 98.15]	93.74±5.14 [84.16, 98.62]	94.78±7.64 [83.30, 98.41]
	√	√	94.64±4.56 [84.29, 98.23]	94.49±5.05 [85.40, 98.71]	95.13±7.19 [83.20, 98.35]
U-Net	–	–	88.76±6.23 [68.73, 97.52]	89.42±7.13 [69.47, 96.71]	89.18±8.13 [68.00, 98.34]
	√	–	89.57±6.74 [72.34, 98.23]	88.89±6.93 [71.40, 97.59]	89.26±7.88 [73.33, 98.71]
	–	√	90.45±4.75 [80.15, 97.79]	90.13±6.83 [79.40, 98.13]	89.13±7.72 [80.91, 97.53]
	√	√	90.73±5.38 [80.06, 98.09]	90.52±6.12 [79.63, 98.15]	90.64±7.41 [80.49, 98.10]

LiTS dataset, MICCAI 2017 liver tumor segmentation challenge; PNSS, positive-negative subset selection; HESS, hard-easy subset selection; std, standard deviation; max, maximum; min, minimum; DSC, dice similarity coefficient; FCN, fully convolutional network.

Table 8 Segmentation results of liver tumor on the LiTS dataset

Model	PNSS	HESS	Liver tumor, mean ± std [min, max]		
			DSC (%)	Precision (%)	Recall (%)
FCN	–	–	62.35±20.15 [20.92, 90.83]	63.14±19.24 [23.54, 92.15]	61.14±20.67 [18.82, 89.71]
	√	–	63.43±18.57 [26.71, 89.97]	63.67±17.91 [30.15, 91.27]	63.14±18.63 [23.97, 90.15]
	–	√	63.98±14.88 [35.73, 90.42]	62.91±14.29 [34.62, 91.24]	65.01±15.17 [36.91, 89.83]
	√	√	65.36±16.58 [33.96, 90.91]	66.03±15.30 [34.19, 91.32]	64.64±16.24 [33.73, 90.50]
U-Net	–	–	56.44±25.15 [16.64, 88.47]	58.15±24.67 [18.31, 89.54]	54.82±23.71 [15.24, 88.23]
	√	–	57.63±26.68 [26.66, 89.05]	58.24±25.79 [28.47, 90.31]	58.13±24.30 [25.06, 88.92]
	–	√	59.64±19.75 [30.62, 89.44]	60.06±17.61 [31.41, 91.23]	60.22±19.30 [29.87, 90.11]
	√	√	61.89±18.60 [32.47, 90.27]	62.39±16.15 [34.59, 92.31]	61.57±19.13 [30.59, 90.31]

LiTS dataset, MICCAI 2017 liver tumor segmentation challenge; PNSS, positive-negative subset selection; HESS, hard-easy subset selection; std, standard deviation; max, maximum; min, minimum; DSC, dice similarity coefficient; FCN, fully convolutional network.

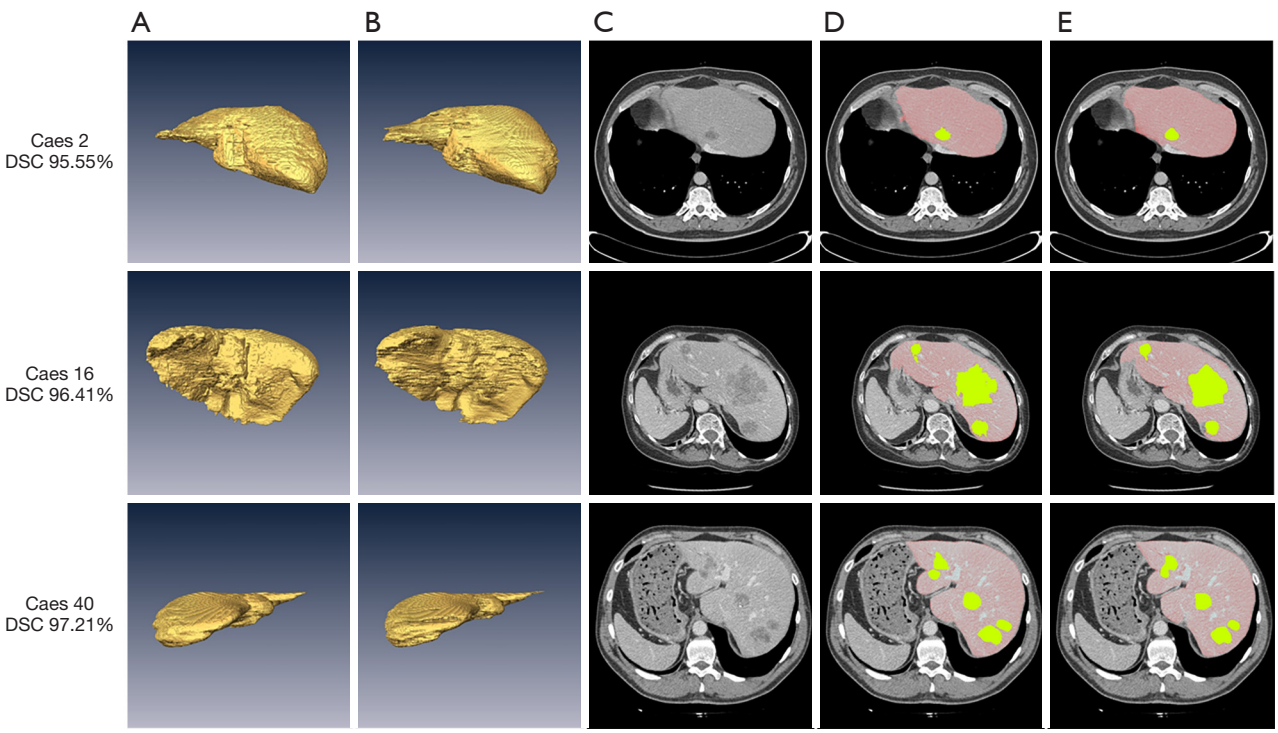

Figure 8 Examples of the segmentation results of the proposed methods on the LiTS dataset. (A,B) The 3D results of the ground truth and proposed methods, respectively. (C-E) The origin 2D slices, the ground truth, and the test result, respectively. The red and yellow masks indicate the liver and liver tumor, respectively. DSC, dice similarity coefficient; LiTS dataset, MICCAI 2017 liver tumor segmentation challenge.

better performance gains (22,27-29). In our experiments, we trained TransUNet and combined it with our methods, and promising results were achieved on the experimental datasets. How to better combine our methods with various models will be a potential direction we can take into consideration. Some emerging data augmentation techniques, such as mix-

Table 9 Brain edema segmentation on the BT-MSD dataset (DSC: %)

Model	PNSS	HESS	DSC (%)	Precision (%)	Recall (%)
FCN	–	–	70.69±15.45 [22.77, 91.16]	71.37±14.21 [26.31, 92.31]	69.47±15.34 [20.07, 91.13]
	√	–	70.46±16.22 [24.81, 92.18]	70.93±15.03 [25.67, 92.43]	69.59±15.82 [24.00, 92.30]
	–	√	71.04±15.35 [32.03, 91.77]	72.30±14.11 [35.49, 92.65]	70.37±16.11 [29.28, 91.91]
	√	√	71.58±15.91 [30.73, 93.08]	71.91±13.59 [32.37, 93.85]	71.45±14.10 [29.24, 92.45]
U-Net	–	–	68.87±18.94 [21.97, 91.58]	68.54±17.32 [22.54, 91.39]	68.87±17.51 [21.43, 92.17]
	√	–	69.14±16.49 [22.75, 92.53]	70.02±16.32 [23.24, 92.14]	68.64±16.92 [22.28, 92.92]
	–	√	70.58±15.37 [29.85, 92.36]	71.64±14.17 [30.45, 92.57]	69.64±14.89 [29.27, 92.15]
	√	√	70.86±14.45 [30.38, 92.47]	71.53±15.13 [31.34, 93.02]	70.41±14.32 [29.47, 92.64]

The performance is described by mean ± std [min, max]. BT-MSD dataset, Brain Tumors Task of Medical Segmentation Decathlon challenge dataset; PNSS, positive-negative subset selection; HESS, hard-easy subset selection; DSC, dice similarity coefficient; FCN, fully convolutional network; std, standard deviation; max, maximum; min, minimum.

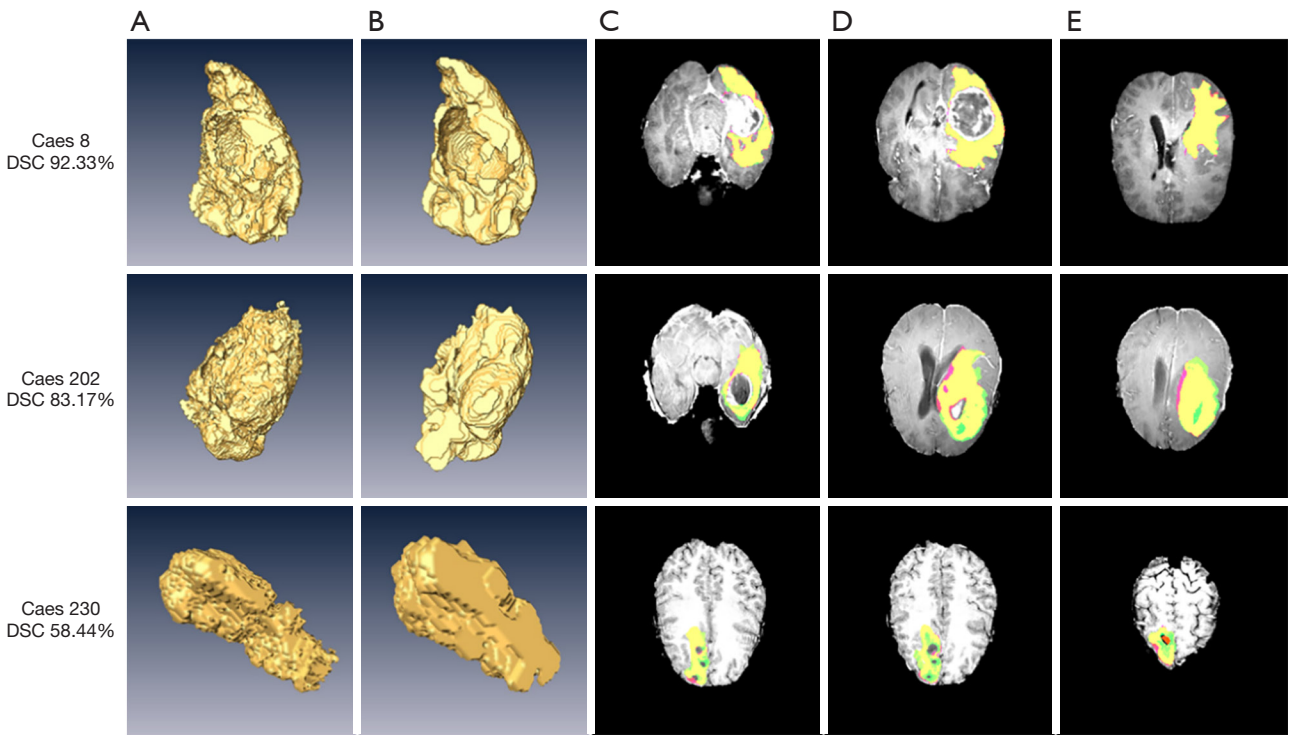

Figure 9 Examples of the segmentation results of the proposed methods on the BT-MSD dataset. (A,B) The 3D results of the ground truth and proposed methods, respectively. (C-E) The segmentation results on 2D slices. Purple and yellow masks indicate the ground truth and prediction regions, respectively. DSC, dice similarity coefficient; BT-MSD dataset, Brain Tumors Task of Medical Segmentation Decathlon challenge dataset.

up (30), adversarial data augmentation (31), and latent space data augmentation (32), can be used to expand hard samples. A recent study focused on the choice of loss functions in medical image segmentation, providing a new perspective for us to implement the loss function (33). Moreover, the imbalance problem in the 3D method is more serious, especially when segmenting small targets.

Recently, a trend of incorporating shape models into

CNNs to enhance boundary segmentation accuracy of the worst case has emerged (34,35). In future work, we intend to apply this mechanism to our sample balancing methods.

Conclusions

We present two simple but effective sample balancing methods (PNSS and HESS) to address foreground-to-background imbalance and hard-to-easy imbalance problems in medical segmentation tasks. Our methods greatly improved the segmentation accuracy of the worst case and achieved competitive performance compared to state-of-the-art methods on the NIH dataset. Moreover, the proposed method can be applied to other CNNs (2D models and TransUNet) and improve their performance. Experimental results showed that the proposed sample balancing methods could improve segmentation performance on different tasks including pancreas segmentation, liver segmentation, liver tumor segmentation, and brain tumor segmentation.

Acknowledgments

Funding: This work was supported by the Fundamental Research Funds for the Central Universities, China (No. 2019CDYGZD004 and No. 2020CDCGJSJ043), the Chongqing Key Research and Development Project, China (No. cstc2017shms-zdyfX0015), the National Natural Science Foundation of China (No. 61703062), the Special Project for Performance Incentive of Research Institutions in Chongqing (No. cstc2018jxjl130017), and the Humanities and Social Science Planning Fund from the Ministry of Education, China (No. 21YJAZH013).

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-21-798/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-21-798/coif). YZW worked for Ziwei King Star Digital Technology Co., Ltd. during the study. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Dou Q, Chen H, Jin Y, Yu L, Qin J, Heng PA. 3D deeply supervised network for automatic liver segmentation from CT volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2016:149-57.
2. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. IEEE Trans Med Imaging 2018;37:2663-74.
3. Liu Z, Song YQ, Sheng VS, Wang L, Jiang R, Zhang X, Yuan D. Liver CT sequence segmentation based with improved U-Net and graph cut. Expert Syst Appl 2019;126:54-63.
4. Zhou T, Tan T, Pan X, Tang H, Li J. Fully automatic deep learning trained on limited data for carotid artery segmentation from large image volumes. Quant Imaging Med Surg 2021;11:67-83.
5. Shin SY, Lee S, Yun ID, Lee KM. Deep vessel segmentation by learning graphical connectivity. Med Image Anal 2019;58:101556.
6. Jin Q, Meng Z, Pham TD, Chen Q, Wei L, Su R. DUNet: A deformable network for retinal vessel segmentation. Knowledge-Based Systems 2019;178:149-62.
7. Coupé P, Mansencal B, Clément M, Giraud R, de Senneville BD, Ta VT, Lepetit V, Manjon JV. AssemblyNet: A novel deep decision-making process for whole brain MRI segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2019:466-74.

8. Jog A, Hoopes A, Greve DN, Van Leemput K, Fischl B. PSACNN: Pulse sequence adaptive fast whole brain segmentation. Neuroimage 2019;199:553-69.
9. Chen H, Dou Q, Yu L, Qin J, Heng PA. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. Neuroimage 2018;170:446-55.
10. 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. Cham: Springer, 2017:693-701.
11. Liu S, Yuan X, Hu R, Liang S, Feng S, Ai Y, Zhang Y. Automatic pancreas segmentation via coarse location and ensemble learning. IEEE Access 2019;8:2906-14.
12. Xue J, He K, Nie D, Adeli E, Shi Z, Lee SW, Zheng Y, Liu X, Li D, Shen D. Cascaded MultiTask 3-D Fully Convolutional Networks for Pancreas Segmentation. IEEE Trans Cybern 2021;51:2153-65.
13. Roth HR, Lu L, Farag A, Shin HC, Liu J, Turkbey EB, Summers RM. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015:556-64.
14. Zhao N, Tong N, Ruan D, Sheng K. Fully automated pancreas segmentation with two-stage 3D convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2019:201-9.
15. Fang C, Li G, Pan C, Li Y, Yu Y. Globally guided progressive fusion network for 3D pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2019:210-8.
16. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell 2020;42:318-27.
17. Zhang Y, Wang Y, Wang Y, Fang B, Yu W, Long H, Lei H. Data Balancing Based on Pre-Training Strategy for Liver Segmentation from CT Scans. Applied Sciences 2019;9:1825.
18. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:761-9.
19. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:3431-40.
20. 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. Cham: Springer, 2015:234-41.
21. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:5693-703.
22. 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. arXiv preprint 2021. arXiv:2102.0430.
23. Simpson AL, Antonelli M, Bakas S, Bilello M, Farahani K, van Ginneken B, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint 2019. arXiv:1902.09063.
24. 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 2018. arXiv:1803.11303.
25. Zhu Z, Xia Y, Shen W, Fishman EK, Yuille AL. A 3D coarse-to-fine framework for automatic pancreas segmentation. arXiv preprint 2017. arXiv:1712.00201.
26. Jin Y, Han D, Ko H. Trseg: Transformer for semantic segmentation. Pattern Recognit Lett 2021;148:29-35.
27. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021:36-46.
28. Zhang Y, Higashita R, Fu H, Xu Y, Zhang Y, Liu H, Zhang J, Liu J. A Multi-branch Hybrid Transformer Network for Corneal Endothelial Cell Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021:99-108.
29. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022:574-84.
30. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. arXiv preprint 2017. arXiv:1710.09412.
31. Chen C, Qin C, Qiu H, Ouyang C, Wang S, Chen L,

Tarroni G, Bai W, Rueckert D. Realistic adversarial data augmentation for MR image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2020:667-77.
32. Chen C, Hammernik K, Ouyang C, Qin C, Bai W, Rueckert D. Cooperative Training and Latent Space Data Augmentation for Robust Medical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021:149-59.
33. Ma J, Chen J, Ng M, Huang R, Li Y, Li C, Yang X, Martel AL. Loss odyssey in medical image segmentation. Med Image Anal 2021;71:102035.
34. Ravishankar H, Venkataramani R, Thiruvenkadam S, Sudhakar P, Vaidya V. Learning and incorporating shape models for semantic segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2017:203-11.
35. Lee HJ, Kim JU, Lee S, Kim HG, Ro YM. Structure boundary preserving segmentation for medical image with ambiguous boundary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:4817-26.

Cite this article as: Huang Y, Wen J, Wang Y, Hu J, Wang Y, Yang W. Subset selection strategy-based pancreas segmentation in CT. Quant Imaging Med Surg 2022;12(6):3061-3077. doi: 10.21037/qims-21-798

Supplementary

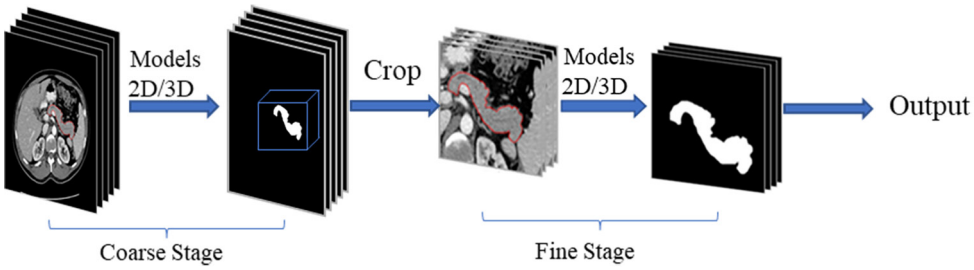

Figure S1 Illustration of the coarse-to-fine strategy.

Table S1 Pancreas segmentation results (DSC: %)

Methods	Mean ± std	Max	Min
AS	82.83±6.97	88.76	68.02
APS	81.73±4.42	87.43	73.53

DSC, dice similarity coefficient; AS, all samples, APS, all positive samples; std, standard deviation; max, maximum; min, minimum.

https://dx.doi.org/10.21037/qims-21-798