applied sciences

MDPI

Article

Voting-Based Contour-Aware Framework for Medical Image Segmentation

Qiao Deng [1], Rongli Zhang [1], Siyue Li [1], Jin Hong [2,3], Yu-Dong Zhang [4], Winnie Chiu Wing Chu [1,*] and Lin Shi [1]

[1] Department of Imaging and Interventional Radiology, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China
[2] School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130012, China
[3] Zhongshan Institute of Changchun University of Science and Technology, Zhongshan 528437, China
[4] School of Computing and Mathematic Sciences, University of Leicester, Leicester LE1 7RH, UK
* Correspondence: winniechu@cuhk.edu.hk; Tel.: +852-3505-2299

Abstract: Accurate and automatic segmentation of medical images is in increasing demand for assisting disease diagnosis and surgical planning. Although Convolutional Neural Networks (CNNs) have shown great promise in medical image segmentation, they prefer to learn texture features over shape information. Moreover, recent studies have shown the promise that learning the data in a meaningful order can make the network perform better. Inspired by these points, we aimed to propose a two-stage medical image segmentation framework based on contour-aware CNN and voting strategy, which could consider the contour information and a meaningful learning order. In the first stage, we introduced a plug-and-play contour enhancement module that could be integrated into the encoder–decoder architecture to assist the model in learning boundary representations. In the second stage, we employed a voting strategy to update the model using easy samples in order to further increase the performance of our model. We conducted studies of the two publicly available CHAOS (MR) and hippocampus MRI datasets. The experimental results show that, when compared to the recent and popular existing models, the proposed framework can boost overall segmentation accuracy and achieve compelling performance, with dice coefficients of 91.2 ± 2.6% for the CHAOS dataset and 88.2 ± 0.4% for the hippocampus dataset.

Keywords: MRI; two-stage CNN framework; contour-aware; voting strategy

check for updates

Citation: Deng, Q.; Zhang, R.; Li, S.; Hong, J.; Zhang, Y.-D.; Chu, W.C.W.; Shi, L. Voting-Based Contour-Aware Framework for Medical Image Segmentation. *Appl. Sci.* **2023**, *13*, 84. https://doi.org/10.3390/app13010084

Academic Editor: Marco Giannelli

Received: 28 October 2022
Revised: 30 November 2022
Accepted: 15 December 2022
Published: 21 December 2022

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Segmentation of anatomical structures in medical images is one of the most important topics in medical image analysis for assisting the diagnosis of anomalies, monitoring disease progression, and planning surgery. Manual segmentation is laborious, subjective, and time-consuming. With the advent of deep learning [1–3], automated segmentation algorithms have become increasingly reliable and potentially applicable for practical use, among which convolutional neural networks (CNNs) have shown enormous potential in open-data challenges in various segmentation tasks involving the liver [4–6], kidney [4,7], tumors [8], and bone segmentation [9].

Since 2012, many classification and segmentation CNNs have been proposed, including AlexNet [10], Google-Net [11], and the Deeplab series network [12–14]. Specialized deep learning algorithms have also been proposed for medical image segmentation. For example, a U-shaped model was proposed in [15] by combining a contracting path and expanding path, allowing the use of global location and context and enabling higher performance for segmentation tasks. Many encoder–decoder CNNs based on the same concepts have since been invented, including V-Net [16], Res-UNet [17], Dense-UNet [18], R2U-Net [19], Attention U-Net [20], and nnU-Net [21].

Considering human intelligence, humans prefer to classify new items, usually considering some object properties such as texture, color, and shape information [22–24]. CNNs are also capable of learning complicated representations of objects, including the features mentioned above [22–24]. Many researchers have proposed that CNNs can learn sophisticated information about the shape associated with each class [25], and the shape information of objects is more significant to CNNs than color [23,24]. It was reported in [22] that CNNs are more likely to recognize texture features than shape information and highlight the performance and robustness benefits of shape-based representations. Since CNNs have learning biases, it is necessary to consider the shape information when designing networks. Various research efforts have been dedicated to edge prediction and contour representation enhancement. A deep contour-aware network for gland segmentation was proposed to extract multi-level contextual features as explored with an auxiliary supervision [26]. A multi-label learning framework without deep supervision was introduced to improve the boundary identification [27]. In addition, the Thinned Edge Alignment approach was introduced, which included a basic contour thinning layer and a novel loss to generate thinner and more exact edges [28]. A two-stream CNN architecture consisting of a conventional and shaped stream was also established, with innovative gates connecting intermediate layers between these two streams [29]. An edge guidance module was integrated to learn edge-attention representations in the early encoding layers to guide the segmentation [30]. Several Edge-Gated blocks were cascaded and integrated with CNNs that accentuated and untangled texture and edge representations during the training [31]. A weighted self-information map was used to highlight the shape information of prediction maps in the shape-entropy-aware adversarial learning framework [6].

In human education, there is a curriculum plan to learn courses in a meaningful order, which means people always master basic knowledge before moving on to advanced courses [32]. Many researchers have pondered if this learning strategy could promote machine learning. Curriculum learning (CL) [32–34] imitated the human learning strategy to train the model with samples from easy to hard. More specifically, in contrast to the standard training process, which is usually a random data shuffle, CL initially trains the model with easier subsets of data and then gradually adds harder data into the subset, up to the entire dataset. CL has shown that it can avoid bad local minima and improve model performance and generalization in a variety of tasks, such as classification [35–37], detection [38], and medical image segmentation [39–41]. The core of most CL strategies is scheduling the training samples. Ranking functions were designed to improve the performance of the femur fracture classification using prior clinical knowledge and the uncertainty scores of model predictions [35]. An attention-guided curriculum learning framework [36] grouped the data using the radiologist-assessed disease severity and fed those subsets into the CNN in a severe to mild order. In this task, it was easier for a CNN to learn severe samples, which have obvious symptoms, while moderate and mild samples without obvious symptoms were more likely to be ambiguous for a CNN. Furthermore, high-confidence samples and heatmaps were utilized to guide model learning in the next iteration. Uncertainty could also be used to measure the difficulty of the samples. In [37], the small loss samples were deemed clean samples and first fed into the network for learning as simple samples. The authors proposed an online uncertainty sample mining strategy to solve the noisy labels in the skin lesion classification task, which selected the high uncertainty samples and stopped these data from updating the model. A cardiac MR motion artifact detection framework [38] employed a data augmentation based on k-space artifact creation and an easy-to-hard batching strategy based on increasing synthetic artifact severity. A CASED framework [40] proposed an adaptive sampling strategy that favors the predicted false samples for the current model. A sample selection stage [41] was proposed to filter the clean samples by measuring the uncertainty of model predictions.

Medical images mainly have texture and shape features, and a meaningful learning order can be beneficial to improve the performance of networks. Considering those contexts, we aimed to study whether the shape features and a meaningful learning order could

improve the performance of models for the automatic segmentation of medical images. During the inference, we only need to feed the images into three networks and then average their predictions, so it is convenient to implement this strategy. In this study, our contributions are summarized as follows:

1. We proposed a novel two-stage framework for medical image segmentation.
2. We developed the contour-aware CNN and employed the voting strategy in our joint optimization framework.
3. We validated our framework using two public datasets, abdomen MRI (CHAOS Challenge) and hippocampus MRI (the Medical Segmentation Decathlon), and could achieve comparable performance compared with other related methods.

2. Method

The proposed approach is described in this section. We presented a contour-aware network and employed a voting strategy during training inspired by [29,31,41]. Figure 1 depicts the pipeline of our two-stage framework. First, we built a contour-aware network by inserting contour enhancement modules into the encoder–decoder backbone CNN and supervising it with an auxiliary loss to learn shape information, as shown in Figure 1a. Secondly, we trained three contour-aware CNNs at the same time iteratively, and each target network could learn high-confidence samples in each mini-batch as chosen by a voting committee made up of its peer two networks, as seen in Figure 1c.

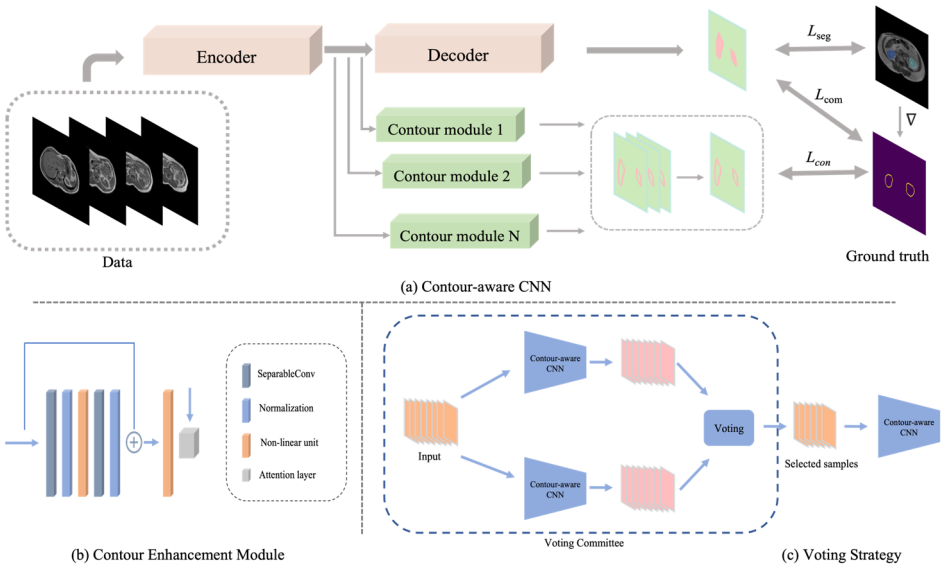

Figure 1. Illustration of two-stage segmentation framework. (**a**) Architecture of our contour-aware CNN in the first stage. (**b**) Detailed structure of the contour enhancement module. (**c**) Stage of voting strategy, where the voting committee filters the samples for the target network.

2.1. Backbone Network

Due to the outstanding performance of encoder–decoder network architecture in medical image segmentation, we adopted the nnU-Net and Segnet [42] as backbone networks in our framework. The detailed network configurations for the two datasets in our paper are shown in Tables 1 and 2. Because the nnU-Net can automatically configure its network architecture to adapt to different datasets, the nn-UNet has seven resolution stages for the CHAOS dataset and four resolution stages for the hippocampus dataset. There are two convolution blocks for each resolution stage in the encoder and decoder. Each block contains a convolution layer followed by an instance normalization and Leaky nonlinearity. Strided convolutions are utilized to downsample the features in the contracting path. The upsampling uses transposed convolutions to enlarge the feature maps in the expanding path. In our experiment, the Segnet has five resolution stages for the CHAOS dataset and

four resolution stages for the hippocampus dataset. In the first two resolution stages, each resolution stage contains two convolution blocks for both the encoder and decoder, with each block containing a convolution layer followed by a batch normalization and a ReLU layer. The encoder and decoder each have three convolution blocks for other resolution stages. Max-pooling is used for downsampling to achieve translation invariance. The decoder uses the max-pooling indices to upsample the feature maps from the corresponding resolution encoder. The difference between the nnU-Net and Segnet is that the nnU-Net use the entire feature maps to concatenate them with the related decoder feature maps. At the same time, Segnet passes the pooling indices to the corresponding decoder.

Table 1. The configurations of nnU-Net encoder and decoder layer in our experiments on CHAOS and hippocampus datasets.

Dataset	Encoder/Decoder	
	Resolution Stage	**Number of Feature Maps**
CHAOS	3×3 conv blocks 1	32
	3×3 conv blocks 2	64
	3×3 conv blocks 3	128
	3×3 conv blocks 4	256
	3×3 conv blocks 5	480
	3×3 conv blocks 6	480
	3×3 conv blocks 7	960
	Strided Convolution	
	Transposed Convolution	
Hippocampus	3×3 conv blocks 1	32
	3×3 conv blocks 2	64
	3×3 conv blocks 3	128
	3×3 conv blocks 4	256
	Strided Convolution	
	Transposed Convolution	

Table 2. The configurations of Segnet encoder and decoder layers in our experiments on CHAOS and hippocampus datasets.

Dataset	Encoder/Decoder	
	Resolution Stage	**Number of Feature Maps**
CHAOS	3×3 conv blocks 1	64
	3×3 conv blocks 2	128
	3×3 conv blocks 3	256
	3×3 conv blocks 4	512
	3×3 conv blocks 5	512
	Max pooling	
Hippocampus	3×3 conv blocks 1	64
	3×3 conv blocks 2	128
	3×3 conv blocks 3	256
	3×3 conv blocks 4	512
	Max pooling	

2.2. Contour-Aware CNN

We present the architecture of our contour-aware CNN in this stage as inspired by [29,31]. We employed the encoder and decoder CNNs as the backbone networks, as shown in Figure 1a, and we inserted contour enhancement modules to interconnect the encoder and decoder of intermediate layers in each resolution. We applied the bilinear up-sampling to the output of the contour enhancement module at each resolution and then concatenated all the feature maps. Furthermore, the contour-aware CNN is guided by a

semantic segmentation loss, a contour loss, and a compatibility loss, all of which ensure high quality of segmentation mask margins.

As shown in Figure 1b, the contour enhancement module can accentuate edge presentations and comprises a residual block and an attention layer. It requires two inputs in each resolution stage, one from the decoder and fed into the residual block, and one from the encoder and delivered into the attention layer. To minimize the computational cost of CNN, we employ separable convolution, followed by a normalization layer and a non-linear unit layer in the residual block. The attention layer has two inputs, and $I_{res,r}$ and $I_{e,r}$ indicate the inputs from the residual block and encoder, respectively, at resolution r. We concatenate $I_{res,r}$ and $I_{e,r}$, then perform a convolution operation using a normalized 1×1 convolution layer $Conv_{1\times1}$, followed by a sigmoid function σ. The intermediate result pixel-wise product with $I_{res,r}$ is used the output of the contour enhancement module at resolution r, which can be defined as follows:

$$Out_r = \sigma(Conv_{1\times1}(I_{res,r}||I_{e,r})) \cdot I_{res,r} \tag{1}$$

where operator $\cdot$ is a pixel-wise multiplication operator, and $\|$ signifies the concatenation of feature maps. The Out_r is then upsampled using bilinear interpolation for further fusing. The basic encoder–decoder CNN architecture typically has as many contour enhancement modules as intermediate resolutions. Backpropagation could well be undertaken end-to-end since the contour enhancement module is differentiable. In a nutshell, the contour enhancement module may be considered as superimposing an attention map on the intermediate representations, which can weigh the regions with critical edge semantics.

2.3. Loss Function

We propose training the contour-aware network using the following cascaded loss function:

$$L = L_{Seg} + \lambda L_{con} + \gamma L_{com} \tag{2}$$

where L_{Seg} denotes the standard loss function to learn the semantics representations of the contour-aware network, L_{con} is used to supervise the boundary information of the contour module, L_{com} is used to correctly align the predicted semantic segmentation with the ground-truth semantic boundaries, and λ and γ are hyper-parameters.

We defined our input training dataset as $S = \{(X_n, Y_n, Z_n), n = 1, \ldots, N\}$, where sample $X_n = \left\{x_j^n, j = 1, \ldots, |X_n|\right\}$ represents the raw input images, $Y_n = \left\{y_j^n, j = 1, \ldots, |Y_n|\right\}$, $y_j^n \in \{0, \ldots, C\}$ represents the corresponding ground truth for image X_n, and C is the segmentation class. $Z_n = \left\{Z_j^n, j = 1, \ldots, |Z_n|\right\}, Z_j^n \in \{0, 1\}$ denotes the binary contour map generated by ground truth mask Y_n. In the following formula, we omit n for ease of expression. To learn the semantic representations, we use cross entropy (CE) loss, which is calculated as follows:

$$L_{Seg}(Y, P_s) = -\sum_{i=1}^{C} y_{true,j}^{C} \log P_{s,j}^{C} \tag{3}$$

where $P_{s,j}^{C}$ is the semantic prediction of the contour-aware network at each pixel j. We supervise shape information holistically using a class-balanced cross-entropy loss [43] to counteract the imbalance between edge and non-edge, which may be characterized as follows:

$$L_{con} = -\beta \sum_{j \in Z_+} \log P_b(z_{pred,j} = 1|x; W) - (1 - \beta) \sum_{j \in Z_-} \log P_b(z_{pred,j} = 0|x; W) \tag{4}$$

where Z_- and Z_+ are the edge and non-edge ground truth label sets. Z is ground truth binary edge map for raw training image X. $\beta = |Z_-|/|Z|$ and $1 - \beta = |Z_+|/|Z|$ denote the ratio of edge pixels to all pixels and the ratio of non-edge pixels to all pixels, respectively.

W represents the parameters of the network, and $P_b\left(z_{pred,j}\right)$ represents the probability of the predicted binary contour map of outlines of targets at pixel j.

We use the compatibility loss function inspired by [29,31] to induce semantic segmentation to align with the ground-truth boundaries map, which could also be described as follows:

$$L_{com} = \sum_{j \in Z_-} \left(\|\nabla\left(\text{argmax}\left(P_{s,j}^c\right)\right)\| - \|\nabla y_{true,j}\| \right) \quad (5)$$

We take the spatial derivative ∇ to convert the semantic prediction into contour predictions. Because argmax is not differentiable, we substitute the Gumbel SoftMax trick [44].

2.4. Voting Strategy

The difficult level of samples can be measured by uncertainty, and minimizing the loss function is the goal of supervised segmentation, which means the predictions are close to the ground truth. As depicted in Figure 1c, we employed a voting strategy in our framework to pick up the high-confidence images as inspired by [41]. In this paper, we deemed high-confidence samples as easy samples in each mini-batch and used these samples to update the target network weights. Due to the random data shuffling in the training process, the model can select different samples in each iteration. We simultaneously initialized three of our contour-aware networks with identical architecture. Each network will be the target to learn the easy samples selected by their respective voting committees. For each target network, it has its voting committee composed of its peer two networks. In this case, the information is shared with the committee, which can result in convergence at a robust minimum. Note that, due to the three networks initialized with different conditions, they could learn different representations from the same samples; thus, the three networks will not produce the same predictions in each mini-batch of data.

In our work, we measured the agreement by computing the difference in loss values between the two networks in the committee, and for those samples with small loss differences, we deemed them to be high-confidence samples. We filtered out the high-confidence samples by choosing those showing agreed predictions of the two networks in the voting committee. The lower difference value means the predictions of the voting committee for this sample are more consistent, which also means the high confidence sample is more likely to obtain similar predictions from the voting committee. For the target network, its voting committee measures the agreement of samples according to the following equation:

$$\mu = \text{argmin}\left(L_{Seg}\left(Y_j, f_{v1}\left(P_j, W\right)\right) - L_{Seg}\left(Y_j, f_{v2}\left(P_j, W\right)\right)\right) \quad (6)$$

where μ denotes the agreement degree of the voting committee for each sample, L_{Seg} denotes cross entropy loss of semantic segmentation, and Y_j, $f_{v1,}$ f_{v2} denote the semantic ground truth and the semantic predictions of voting committee networks.

In this manner, we selected a certain proportion of the samples for the target network as usable samples and discarded the remaining samples in each mini-batch. The target network will only update its parameters by learning those useful samples. We describe the optimization details in Algorithm 1.

Algorithm 1: Voting strategy

Input: Training set X_n, Label set Y_n
Initialize: Initialize W_1, W_2, and W_3
t = 0
Repeat:
t = t + 1
Random samples from X_n
Compute the semantic predictions P_1, P_2, and P_3

1. Compute the loss function L_{Seg1}, L_{Seg2}, and L_{Seg3}
   Select the samples for each target network using Equation (6)
2. Compute the loss function L_{Seg1}, L_{Seg2} and L_{Seg3} using the selected sample cases
3. Compute the stochastic gradient and update W_1, W_2, and W_3

Until: convergence

3. Experiment

3.1. Dataset

CHAOS: CHAOS [4] provides MR data from healthy subjects for abdominal organ segmentation. There is a dataset of 20 patients with T1-DUAL in phase, opposite of phase, and T2-SPIR collected from the Department of Radiology, Dokuz Eylul University Hospital, Izmir, Turkey, and acquired on a 1.5 T Philips MRI. We included all 1594 slices with a resolution of 256 × 256. CHAOS data can be accessed with its DOI number via the zenodo.org webpage under CC-BY-SA 4.0 license. Further details and explanations are available on the CHAOS website (https://chaos.grand-challenge.org/Data/, accessed on 8 February 2021). Annotations include the liver, left kidney, right kidney, and spleen.

Hippocampus: The Medical Segmentation Decathlon [45,46] provided a 260 Hippocampus MR dataset gathered with a Philips Achieva scanner and acquired using a 3D T1-weighted MPRAGE sequence from Vanderbilt University Medical Center (Nashville, TN, USA). All data were made available online under Creative Commons license CC-BY-SA 4.0 and could be downloaded from the website (http://medicaldecathlon.com/, accessed on 8 February 2021). Annotations include the hippocampus anterior and posterior.

3.2. Evaluation Metrics

We use the dice similarity coefficient (DSC), mean intersection over union (mIoU), and 95 percentile Hausdorff distance (HD) to access segmentation accuracy. These metrics are used to measure the agreement between the predicted segmentation and the ground truth, which is described as follows:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \times 100\% \tag{7}$$

$$\mathrm{mIoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \times 100\% \tag{8}$$

$$\mathrm{HD}(X, Y) = \max(\mathrm{hd}(S, G), \mathrm{hd}(G, S)) \tag{9}$$

where TP, FP, and FN represent the number of true positive, false positive, and false negative pixels; S denotes the set of voxels in the segmentation result; G denotes the set of voxels in the ground truth; $\mathrm{hd}(S, G)$ and $\mathrm{hd}(G, S)$ are the one-sided HD from S to G and G to S, respectively.

3.3. Implementation Details

First, we crop the training data to their non-zero region. After the cropping, each patient scan is subjected to z-score normalization based on the mean and standard deviation of the intensities. We adopt data augmentation including random rotations, scaling, Gaussian blur, brightness, contrast, and mirroring. The five-fold cross-validation method is used to verify the robustness of the model while avoiding selection bias.

We implemented all models in Pytorch (Torch 1.7.1, torchvision 0.8.2, Meta Platforms, Menlo Park, CA, USA) and trained models on NVIDIA Tesla A100 40 GB GPUs (Nvidia, St. Clara, CA, USA) and NVIDIA Tesla V100 32 GB GPUs. We slightly modified our contour enhancement modules to adopt them on the different backbone networks. For Segnet, we used batchnorm and a rectified linear unit (ReLU) in each contour enhancement module. For nn-UNet, we adopted instance normalization and leakyReLU in each contour enhancement module. We used the grid search method to choose the suitable parameters λ = 20 and γ = 1 in Equation (2). We used a batch size of 8, 1000 epochs, the Stochastic gradient descent optimization algorithm with Nesterov momentum ($\mu = 0.99$), and the initial learning rate ($\alpha_0 = 0.01$), decreasing according to poly learning rate policy:

$$\alpha = \alpha_0 \times (1 - \text{epoch}/\text{epoch}_{\text{max}})^{0.9} \quad (10)$$

4. Results

In this part, we evaluate the segmentation performance of our framework on abdominal MR and hippocampal MR images. As described in Section 2, we established our framework by embedding a plug-and-play contour enhancement module into CNNs and training with a voting strategy.

CHAOS(MR): The CHAOS challenge aims to segment the abdominal organs accurately. Table 3 shows the ablation study results of our proposed framework for the metrics of DSC, mIoU, and HD for the dataset of CHAOS. We conducted two experiments in which we trained Segnet and nn-Unet (2d) as our baselines. We ebbed the contour enhancement modules in these two backbone networks to conduct our contour-aware networks (CN) in the first step and applied the stage of joint optimization using voting strategy (VS). It is distinct that our framework could achieve fairly good results for conducting different backbone networks and improve the accuracy of segmentation in each stage progressively. For the experiments conducted based on Segnet, the results of contour-aware networks can achieve 2.2% and 2.9% improvements and 3.1 mm reduction on DSC, mIoU, and HD, respectively, and the voting strategy can further improve slightly the average DSC and mIoU to 90.4 $\pm$ 3.0% and 83.7 $\pm$ 3.9% and can slightly reduce the average HD to 13.9 $\pm$ 8.4 mm. The results of our framework based on nnU-Net have a similar increasing tendency in each stage and can achieve 1.8% and 2.6% improvements and 5.8 mm reduction, compared to baseline, reaching 91.2 $\pm$ 2.6%, 84.8 $\pm$ 3.5%, and 12.6 $\pm$ 10.8 mm in DSC, mIoU, and HD, respectively. The qualitative results of our method on the CHAOS dataset are displayed in Figure 2.

Table 3. Ablation study results of our proposed framework for the CHAOS(MRI) dataset in terms of DSC (%), mIoU (%), and HD (mm).

Methods	Liver			Right Kidney			Left Kidney			Spleen			Average		
	DSC	mIoU	HD	DSC	mIoU	HD	DSC	mIoU	HD	DSC	mIoU	HD	DSC	mIoU	HD
Segnet	91.5 ± 4.5	85.8 ± 6.2	21.6 ± 28.7	87.5 ± 6.3	79.7 ± 7.6	17.7 ± 18.2	88.7 ± 4.8	80.7 ± 7.1	20.6 ± 20.0	84.0 ± 9.1	75.1 ± 11.3	20.1 ± 15.8	87.9 ± 5.4	80.3 ± 7.2	20.0 ± 20.4
CN	93.1 ± 3.1	88.3 ± 3.4	10.6 ± 7.9	90.0 ± 3.0	82.6 ± 4.3	18.9 ± 21.0	89.7 ± 5.0	82.4 ± 7.3	19.3 ± 13.2	87.5 ± 5.1	79.5 ± 6.3	18.8 ± 12.7	90.1 ± 2.8	83.2 ± 4.0	16.9 ± 10.7
CN+VS	92.9 ± 3.4	88.1 ± 3.9	10.9 ± 8.1	91.0 ± 2.2	84.2 ± 3.2	13.4 ± 10.1	90.9 ± 3.1	83.9 ± 4.8	12.6 ± 9.6	86.6 ± 5.8	78.7 ± 6.7	18.6 ± 13.2	90.4 ± 3.0	83.7 ± 3.9	13.9 ± 8.4
nnUNet [21]	91.3 ± 4.8	85.7 ± 6.7	20.9 ± 29.0	89.9 ± 5.2	82.6 ± 7.2	16.8 ± 18.6	89.3 ± 5.6	81.6 ± 8.1	17.2 ± 21.5	87.2 ± 5.6	78.9 ± 6.9	18.8 ± 16.0	89.4 ± 4.4	82.2 ± 6.3	18.4 ± 21.1
CN	92.0 ± 4.0	86.7 ± 5.2	20.1 ± 25.2	91.6 ± 2.4	84.8 ± 3.7	9.1 ± 5.7	91.3 ± 2.6	84.3 ± 4.0	13.9 ± 13.0	87.5 ± 5.1	79.3 ± 5.9	14.2 ± 7.1	90.6 ± 2.9	83.8 ± 4.0	14.3 ± 12.2
CN+VS	92.4 ± 3.8	87.3 ± 4.8	18.8 ± 22.8	92.3 ± 2.0	85.9 ± 3.3	7.7 ± 5.2	91.6 ± 2.8	85.0 ± 4.4	11.6 ± 12.3	88.4 ± 4.7	80.8 ± 5.4	12.3 ± 6.2	91.2 ± 2.6	84.8 ± 3.5	12.6 ± 10.8

Data are presented as mean $\pm$ standard deviation.

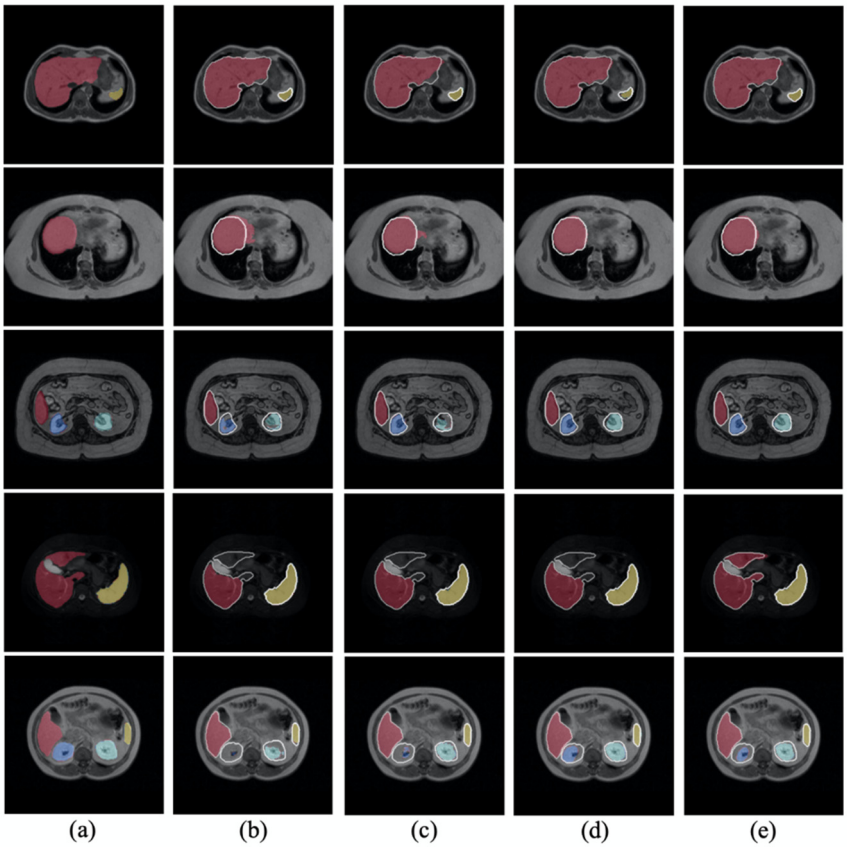

Figure 2. Qualitative comparison of segmentation results of our framework on the CHAOS dataset: (**a**) ground truth, (**b**,**c**) results of Segnet and nn-Unet, and (**d**,**e**) results of our method implemented on Segnet and nnUNet. The red region represents the liver, the blue region represents the right kidney, the cyan represents the left kidney, and the yellow represents the spleen. The white line denotes the boundary of the ground truth.

We presented a comparison of our proposed method with related works on the CHAOS(MRI) dataset. We compared with recurrent U-Net variant BI0-Net [47] and Neural Architecture Search (NAS) networks including Auto-DeepLab [48], NAS-UNet [49], MS-NAS [50], and BiX-NAS [51]. Table 4 summarizes the comparison, providing the metrics of DSC and mIoU for the abdominal organs. Our proposed framework could achieve the best performance among all methods in DSC and mIoU.

Table 4. Quantitative evaluation of different networks for the CHAOS(MRI) dataset in terms of DSC (%) and mIoU (%).

Methods	Liver		Right Kidney		Left Kidney		Spleen		Average	
	DSC	**mIoU**	**DSC**	**mIoU**	**DSC**	**mIoU**	**DSC**	**mIoU**	**DSC**	**mIoU**
Auto-DL [50]	93.5	87.9	85.1	75.6	84.2	73.1	87.3	78.2	87.5	78.7
NAS-UNet [50]	93.7	88.3	87.5	77.6	85.4	74.0	89.3	80.2	89.0	80.0
MS-NAS [50]	94.1	88.9	88.4	79.3	88.5	79.4	90.0	82.9	90.3	82.6
BIO-Net [51]	91.7	85.8	87.2	78.2	85.1	75.7	82.8	73.2	86.7	78.2
BiX-NAS [51]	89.8	82.6	82.1	71.0	82.7	71.9	76.5	66.0	82.8	72.9
Ours	92.4	87.3	92.3	85.9	91.6	85.0	88.4	80.8	91.2	84.8

Hippocampus: We further evaluated our framework by segmenting the hippocampus, which has two sub-regions, anterior and posterior. Similar to the ablation study for the CHAOS(MRI) dataset, we trained the same baseline networks and implemented our framework on them. Table 5 indicates the ablation study results of our framework on DSC, mIoU,

and HD for the hippocampus dataset. For the experiments based on Segnet, our framework could reach 88.0 ± 0.3%, 78.8 ± 0.5%, and 1.4 ± 0.1 mm on metrics of DSC, mIoU, and HD, respectively. Concerning the experiments based on nn-Unet, our method could achieve 1.3% improvements from 86.9 ± 0.5% to 88.2 ± 0.4% on DSC and achieve 1.9% improvements from 77.1 ± 0.7% to 79.1 ± 0.6 % on mIoU, reaching 1.4 ± 0.0 mm on HD. The qualitative results of our method on the hippocampus dataset are displayed in Figure 3.

Table 5. Ablation study results of our proposed framework for the hippocampus dataset in terms of DSC (%), mIoU (%), and HD (mm).

Methods	Anterior			Posterior			Average		
	DSC	mIoU	HD	DSC	mIoU	HD	DSC(%)	mIoU	HD
Segnet	88.4 ± 0.4	79.4 ± 0.6	1.3 ± 0.0	86.6 ± 0.5	76.7 ± 0.6	1.5 ± 0.1	87.5 ± 0.4	78.1 ± 0.5	1.4 ± 0.1
CN	88.6 ± 0.3	79.7 ± 0.5	1.3 ± 0.0	87.2 ± 0.2	77.4 ± 0.3	1.4 ± 0.0	87.9 ± 0.2	78.6 ± 0.4	1.4 ± 0.0
CN+VS	88.8 ± 0.4	80.1 ± 0.6	1.3 ± 0.0	87.2 ± 0.3	77.6 ± 0.5	1.4 ± 0.1	88.0 ± 0.3	78.8 ± 0.5	1.4 ± 0.1
nnUNet [21]	87.9 ± 0.4	78.6 ± 0.7	1.4 ± 0.1	86.0 ± 0.6	75.7 ± 0.8	1.5 ± 0.1	86.9 ± 0.5	77.1 ± 0.7	1.5 ± 0.1
CN	88.8 ± 0.3	80.0 ± 0.5	1.3 ± 0.0	87.2 ± 0.3	77.5 ± 0.4	1.4 ± 0.1	88.0 ± 0.2	78.8 ± 0.4	1.4 ± 0.0
CN+VS	89.0 ± 0.4	80.4 ± 0.6	1.3 ± 0.0	87.3 ± 0.5	77.7 ± 0.6	1.4 ± 0.1	88.2 ± 0.4	79.1 ± 0.6	1.4 ± 0.0

Data are presented as mean ± standard deviation.

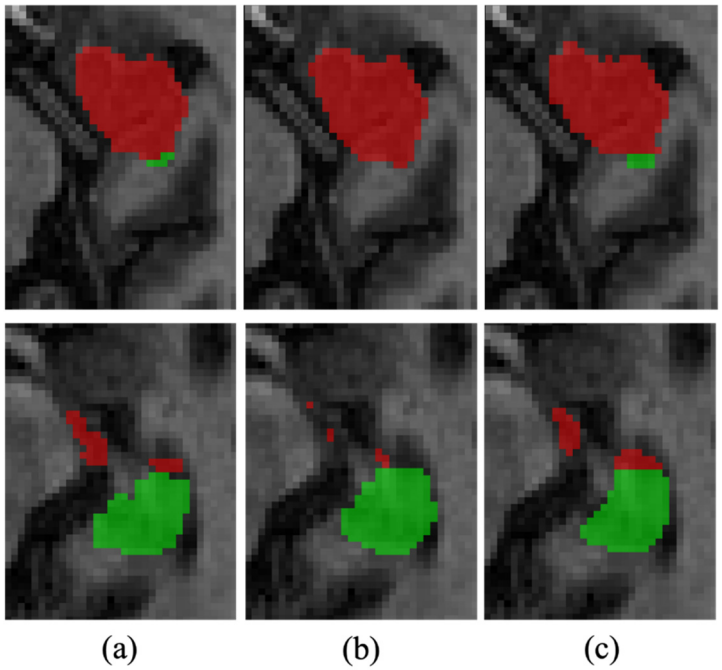

Figure 3. Qualitative comparison of segmentation results of our framework on hippocampus dataset. Column (**a**) is the ground truth, the first row of column (**b**) is the result of Segnet, and the second row of column (**b**) is the result of nn-Unet. The first row of column (**c**) and the second row of column (**c**) are the results of our method implemented on Segnet and nnUNet. The red and green regions represent the anterior and posterior of the hippocampus, respectively.

We compared our method with previous works on the hippocampus dataset. We compared our method with EMONAS-Net [52], C2FNAS [53], VFN [54], residual DSM [55], and distill DSM [56]. Table 6 summarizes the comparison of previous methods on the hippocampus dataset in terms of DSC. The performance of our proposed framework ranks third among these methods. The EMONAS-Net and distill DSM could achieve 88.3% and 88.8% in DSC, which is slightly higher than ours.

Table 6. Quantitative evaluation of different networks for the hippocampus dataset in terms of DSC (%).

Methods	DSC		
	Anterior	Posterior	Average
VFN [56]	89.2	86.6	87.9
Residual DSM [56]	89.0	86.5	87.8
C2FNAS [52]	83.1	83.0	83.1
EMONAS-Net [52]	89.1	87.5	88.3
Distill DSM [56]	89.6	87.9	88.8
Ours	89.0	87.3	88.2

5. Discussion

Our framework could achieve high DSC and mIoU on both CHAOS and the hippocampus dataset. There are two key components in our framework. We aim to encourage our model to focus more on the boundary representations and further improve the performance networks by learning high-confidence samples. To validate our framework, we chose two backbone networks to conduct our framework and evaluate our models using two different datasets in our experiments. As shown in Tables 3 and 5, our framework has a similar upward trend in performance, whether on different datasets or different backbones. For both datasets, compared with the baseline, our contour-aware CNNs can improve the performance overall, and the segmentation accuracy is further improved through the joint optimization stage with the voting strategy. It can be seen that the segmentation performance was improved by a relatively large amount after we embedded the contour enhancement modules into the backbone networks. Although the performance only increased slightly after we employed the voting strategy, the same upward trend exists in different evaluation matrices and backbone networks.

It is worth mentioning that the performance improvement of the framework on the CHAOS dataset is higher than that on the hippocampus dataset. The explanation for this might be that the two datasets have distinct properties. In Figure 2, the results of the baseline always segment incomplete organs or incorrectly segment non-organ parts. In addition, the results of our model can largely compensate for incomplete segmentation of the kidney and liver and decrease the false positive predictions of the liver. In Figure 3, the results show that our model is better able to detect the hippocampus than the baseline network. As abdominal organs in CHAOS have distinct margins, our method can learn the shape information well to fulfill the segmentation more completely and constrain the region of false positives. The Hippocampus is in a complex surrounding environment without distinct boundaries, and there is also no clear boundary between the anterior and posterior. In some cases the perception of the human eye is limited, but the machine may be able to detect it successfully. In our study, the network may detect the organs by learning the surroundings. In a word, the results show that our framework is useful not only for the dataset with distinct contours but also for the dataset with complex surroundings and no clear outlines.

Tables 4 and 6 show the comparison between our proposed method implemented on nn-UNet and other works related to the CHAOS and hippocampus dataset. The segmentation performance of our framework ranks first on the CHAOS dataset and ranks third on the hippocampus dataset. As a classical method, U-Net is consistently applied on medical images and can achieve robust performance. Many researchers have developed variants of the U-Net to achieve better performance on different medical images. We achieved results superior to the U-Net variant [47]. Recently, many works have aimed to design network architecture automatically because designing a specific neural network is time-consuming and difficult. Many neural architecture search (NAS) methods have been proposed for medical image segmentation. The performance of those models [48–51] was similar to that of our model, but slightly lower. Because the medical images are mostly 3D volumes in clinical diagnosis, numerous networks were designed for 3D medical image segmentation. EMONAS-Net [52] and C2FNAS [53] are NAS frameworks for 3D medical

images. There are also some works to bridge the gap between 2D and 3D. VFN [54] fused the 2D segmentation results from multi-views. Residual DSM [55] and Distill DSM [56] used 2D convolution to segment by extracting information and sharing it with neighboring slices. Although our model achieved lower scores on the hippocampus dataset than EMONAS-Net and Distill DSM, the EMONAS-Net trained 3D CNN, which was able to incorporate 3D features. The Distill DSM needs to input a stacked slice volume into the network. Our model was designed to train 2D CNN and segment with 2D slices, as 2D CNNs are known to be computationally lighter and have faster inference time than 3D CNNs.

We further measured the suitable value of λ and the appropriate parameter of the voting strategy. In Table 7, the results show that our contour-aware CNN based on nn-UNet has the best performance when $\lambda = 20$. When the value of λ is lower or higher, it decreases the performance of models. In Table 8, the results show that our framework based on nn-UNet prefers to learn half of the samples in each mini-batch. Learning 25% of cases per iteration may lead to underutilization of data and has the worst performance on both datasets. The 75% sampling rate results are slightly lower than the 50% sampling rate results on the CHAOS dataset, and they have the same performance on the Hippocampus dataset. It can be shown that our voting strategy performs well on both datasets when the experimental setting is greater than a 25% sampling rate.

Table 7. Segmentation results for different values of λ for our contour-aware CNN based on nn-UNet for the CHAOS dataset in DSC (%).

λ	DSC
$\lambda = 0$	90.4 ± 2.8
$\lambda = 10$	89.7 ± 3.9
$\lambda = 20$	90.6 ± 2.9
$\lambda = 30$	90.4 ± 3.5

Data are presented as mean ± standard deviation.

Table 8. Segmentation results for different numbers of subjects selected in our framework in DSC (%).

Dataset	25%	50%	75%
CHAOS	90.8 ± 3.3	91.2 ± 2.6	91.0 ± 2.8
Hippocampus	87.6 ± 0.4	88.2 ± 0.4	88.2 ± 0.3

Data are presented as mean ± standard deviation.

We first developed a novel joint optimization framework consisting of contour enhancement modules and a voting strategy in our study, which considered the contour features and a meaningful sample learning order. In actual clinical practice, we expect more accurate results, and we verified that the combination of these two techniques is effective. We demonstrated in ablation studies that each module is useful and improves the overall performance of state-of-the-art baseline networks at each stage on two public datasets. We assume that our framework can generalize to unseen datasets. Now that we have validated our framework on two public datasets, we will collect more datasets to validate our framework in the future. We expect that our work can be translated into high-precision segmentation tasks and can be used as a reference for future segmentation tasks that require high precision. The limitation of our framework is that the complexity is relatively high and the mechanism of action needs further study. The voting strategy does consume more computing resources and complicates training, but it should be adopted in deep learning in future research when high-precision results are required and the model size can be ignored.

6. Conclusions

In this study, we proposed a two-stage framework for medical image segmentation. We conducted a contour-aware network and trained with a voting strategy, which could

facilitate the shape representations and avoid bad local minima during training. We engineered the network to pay more attention to the shape information by plugging the contour enhancement modules and picking up easy samples for backpropagation to further improve the performance. We conducted our experiments on the CHAOS dataset and the hippocampus dataset to validate our framework. The results demonstrated that our framework is effective and can achieve competitive performance compared to the previous related work.

Author Contributions: Conceptualization, Q.D. and W.C.W.C.; data curation, Q.D., R.Z., S.L., J.H. and Y.-D.Z.; formal analysis, Q.D., R.Z., S.L., J.H. and Y.-D.Z.; funding acquisition, W.C.W.C. and L.S.; investigation, Q.D., R.Z., S.L., J.H. and Y.-D.Z.; Methodology, Q.D., R.Z., S.L., J.H. and Y.-D.Z.; Project administration, W.C.W.C. and L.S.; resources, W.C.W.C. and L.S.; software, Q.D. and R.Z.; supervision, W.C.W.C.; validation, Q.D. and R.Z.; visualization Q.D. and R.Z.; writing—original draft, Q.D.; writing—review & editing, Q.D., R.Z., S.L., J.H., Y.-D.Z., W.C.W.C. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the General Research Funding from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. CUHK 14200721].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data presented in this study may be provided upon request to the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hong, J.; Feng, Z.; Wang, S.-H.; Peet, A.; Zhang, Y.-D.; Sun, Y.; Yang, M. Brain age prediction of children using routine brain MR images via deep learning. *Front. Neurol.* **2020**, *11*, 584682. [CrossRef] [PubMed]
2. Hong, J.; Cheng, H.; Wang, S.-H.; Liu, J. Improvement of cerebral microbleeds detection based on discriminative feature learning. *Fundam. Inform.* **2019**, *168*, 231–248. [CrossRef]
3. Zuo, Q.; Lu, L.; Wang, L.; Zuo, J.; Ouyang, T. Constructing Brain Functional Network by Adversarial Temporal-Spatial Aligned Transformer for Early AD Analysis. *Front. Neurosci.* **2022**, *16*, 1087176. [CrossRef] [PubMed]
4. Kavur, A.E.; Gezer, N.S.; Barış, M.; Aslan, S.; Conze, P.-H.; Groza, V.; Pham, D.D.; Chatterjee, S.; Ernst, P.; Özkan, S.; et al. CHAOS Challenge-combined (CT-MR) healthy abdominal organ segmentation. *Med. Image Anal.* **2021**, *69*, 101950. [CrossRef]
5. Araújo, J.D.L.; da Cruz, L.B.; Diniz, J.O.B.; Ferreira, J.L.; Silva, A.C.; de Paiva, A.C.; Gattass, M. Liver segmentation from computed tomography images using cascade deep learning. *Comput. Biol. Med.* **2021**, *140*, 105095. [CrossRef]
6. Hong, J.; Yu, S.C.-H.; Chen, W. Unsupervised domain adaptation for cross-modality liver segmentation via joint adversarial learning and self-learning. *Appl. Soft Comput.* **2022**, *121*, 108729. [CrossRef]
7. Hong, J.; Zhang, Y.-D.; Chen, W. Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation. *Knowl. Based Syst.* **2022**, *250*, 109155. [CrossRef]
8. Wang, G.; Shapey, J.; Li, W.; Dorent, R.; Dimitriadis, A.; Bisdas, S.; Paddick, I.; Bradford, R.; Zhang, S.; Ourselin, S.; et al. Automatic Segmentation of Vestibular Schwannoma from T2-Weighted MRI by Deep Spatial Attention with Hardness-Weighted Loss. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Cham, Switzerland, 2019.
9. Noguchi, S.; Nishio, M.; Yakami, M.; Nakagomi, K.; Togashi, K. Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques. *Comput. Biol. Med.* **2020**, *121*, 103767. [CrossRef]
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
11. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
12. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
13. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
14. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.

15. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015.
16. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016.
17. Xiao, X.; Lian, S.; Luo, Z.; Li, S. Weighted Res-Unet for High-Quality Retina Vessel Segmentation. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018.
18. Guan, S.; Khan, A.A.; Sikdar, S.; Chitnis, P.V. Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 568–576. [CrossRef] [PubMed]
19. Alom, M.Z.; Yakopcic, C.; Hasan, M.; Taha, T.M.; Asari, V.K. Recurrent residual U-Net for medical image segmentation. *J. Med. Imaging* **2019**, *6*, 014006. [CrossRef] [PubMed]
20. Oktay, O.; Schlemper, J.; le Folgoc, L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
21. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
22. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* **2018**, arXiv:1811.12231.
23. Ritter, S.; Barrett, D.G.T.; Santoro, A.; Botvinick, M.M. Cognitive psychology for Deep Neural Networks: A Shape Bias Case Study. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
24. Hosseini, H.; Xiao, B.; Jaiswal, M.; Poovendran, R. Assessing Shape Bias Property of Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018.
25. Kriegeskorte, N. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **2015**, *1*, 417–446. [CrossRef]
26. Chen, H.; Qi, X.; Yu, L.; Heng, P.A. DCAN: Deep contour-aware networks for accurate gland segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
27. Yu, Z.; Feng, C.; Liu, M.Y.; Ramalingam, S. Casenet: Deep Category-Aware Semantic Edge Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
28. Acuna, D.; Kar, A.; Fidler, S. Devil is in the Edges: Learning Semantic Boundaries from Noisy Annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
29. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated Shape Cnns for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Replublic of Korea, 27 October–2 November 2019.
30. Zhang, Z.; Fu, H.; Dai, H.; Shen, J.; Pang, Y.; Shao, L. Et-net: A generic Edge-Attention Guidance Network for Medical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Cham, Switzerland, 2019.
31. Hatamizadeh, A.; Terzopoulos, D.; Myronenko, A. Edge-gated CNNs for volumetric semantic segmentation of medical images. *arXiv* **2020**, arXiv:2002.04207.
32. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009.
33. Peng, J.; Wang, Y. Medical image segmentation with limited supervision: A review of deep network models. *IEEE Access* **2021**, *9*, 36827–36851. [CrossRef]
34. Soviany, P.; Ionescu, R.T.; Rota, P.; Sebe, N. Curriculum learning: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 1526–1565. [CrossRef]
35. Jiménez-Sánchez, A.; Mateus, D.; Kirchhoff, S.; Kirchhoff, C.; Biberthaler, P.; Navab, N.; Ballester, M.A.G.; Piella, G. Curriculum learning for improved femur fracture classification: Scheduling data with prior knowledge and uncertainty. *Med. Image Anal.* **2022**, *75*, 102273. [CrossRef] [PubMed]
36. Tang, Y.; Wang, X.; Harrison, A.P.; Lu, L.; Xiao, J.; Summers, R.M. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Granada, Spain, 16 September 2018; Springer: Cham, Switzerland, 2018.
37. Xue, C.; Dou, Q.; Shi, X.; Chen, H.; Heng, P.A. Robust Learning At Noisy Labeled Medical Images: Applied to Skin Lesion Classification. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019.
38. Oksuz, I.; Ruijsink, B.; Puyol-Antón, E.; Clough, J.R.; Cruz, G.; Bustin, A.; Prieto, C.; Botnar, R.; Rueckert, D.; Schnabel, J.A.; et al. Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning. *Med. Image Anal.* **2019**, *55*, 136–147. [CrossRef]
39. Kervadec, H.; Dolz, J.; Granger, É.; Ben Ayed, I. Curriculum Semi-Supervised Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Cham, Switzerland, 2019.

40. Jesson, A.; Guizard, N.; Ghalehjegh, S.H.; Goblot, D.; Soudan, F.; Chapados, N. CASED: Curriculum Adaptive Sampling for Extreme Sata Imbalance. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 11–13 September 2017; Springer: Cham, Switzerland, 2017.
41. Xue, C.; Deng, Q.; Li, X.; Dou, Q.; Heng, P.A. Cascaded Robust Learning at Imperfect Labels for Chest X-ray Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Cham, Switzerland, 2020.
42. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
43. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
44. Jang, E.; Gu, S.; Poole, B. Categorical Reparameterization with Gumbel-Softmax. *arXiv* **2016**, arXiv:1611.01144.
45. Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B.A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R.M.; et al. The medical Segmentation Decathlon. *arXiv* **2021**, arXiv:2106.05735. [CrossRef]
46. Simpson, A.L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; van Ginneken, B.; Kopp-Schneider, A.; Landman, B.A.; Litjens, G.; Menze, B. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv* **2019**, arXiv:1902.09063.
47. Xiang, T.; Zhang, C.; Liu, D.; Song, Y.; Huang, H.; Cai, W. BiO-Net: Learning recurrent bi-directional connections for encoder-decoder architecture. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Cham, Switzerland, 2020.
48. Liu, C.; Chen, L.C.; Schroff, F.; Adam, H.; Hua, W.; Yuille, A.L.; Fei-Fei, L. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
49. Weng, Y.; Zhou, T.; Li, Y.; Qiu, X. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access* **2019**, *7*, 44247–44257. [CrossRef]
50. Yan, X.; Jiang, W.; Shi, Y.; Zhuo, C. Ms-Nas: Multi-Scale Neural Architecture Search for Medical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Cham, Switzerland, 2020.
51. Wang, X.; Xiang, T.; Zhang, C.; Song, Y.; Liu, D.; Huang, H.; Cai, W. Bix-nas: Searching efficient bi-directional architecture for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Cham, Switzerland, 2021.
52. Calisto, M.B.; Lai-Yuen, S.K. EMONAS-Net: Efficient multiobjective neural architecture search using surrogate-assisted evolutionary algorithm for 3D medical image segmentation. *Artif. Intell. Med.* **2021**, *119*, 102154. [CrossRef]
53. Yu, Q.; Yang, D.; Roth, H.; Bai, Y.; Zhang, Y.; Yuille, A.L.; Xu, D. C2fnas: Coarse-To-Fine Neural Architecture Search For 3d Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
54. Xia, Y.; Xie, L.; Liu, F.; Zhu, Z.; Fishman, E.K.; Yuille, A.L. Bridging the Gap between 2d and 3d Organ Segmentation with Volumetric Fusion Net. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Granada, Spain, 16 September 2018; Springer: Cham, Switzerland, 2018.
55. Lin, J.; Gan, C.; Han, S. Tsm: Temporal Shift Module for Efficient Video Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019.
56. Maheshwari, H.; Goel, V.; Sethuraman, R.; Sheet, D. Distill DSM: Computationally Efficient Method for Segmentation of Medical Imaging Volumes. In Proceedings of the Medical Imaging with Deep Learning, Lübeck, Germany, 7–9 July 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.