applied sciences

MDPI

Article

Attention V-Net: A Modified V-Net Architecture for Left Atrial Segmentation

Xiaoli Liu †, Ruoqi Yin † and Jianqin Yin *

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China; Liuxiaoli134@bupt.edu.cn (X.L.); ruoqiyin@bupt.edu.cn (R.Y.)
* Correspondence: jqyin@bupt.edu.cn
† These authors contributed equally to this work.

Abstract: We propose a fully convolutional neural network based on the attention mechanism for 3D medical image segmentation tasks. It can adaptively learn to highlight the salient features of images that are useful for image segmentation tasks. Some prior methods enhance accuracy using multi-scale feature fusion or dilated convolution, which is basically artificial and lacks the flexibility of the model itself. Therefore, some works proposed the 2D attention gate module, but these works process 2D medical slice images, ignoring the correlation between 3D image sequences. In contrast, the 3D attention gate can comprehensively use the information of three dimensions of medical images. In this paper, we propose the Attention V-Net architecture, which uses the 3D attention gate module, and applied it to the left atrium segmentation framework based on semi-supervised learning. The proposed method is evaluated on the dataset of the 2018 left atrial challenge. The experimental results show that the Attention V-Net obtains improved performance under evaluation indicators, such as Dice, Jaccard, ASD (Average surface distance), and 95HD (Hausdorff distance). The result indicates that the model in this paper can effectively improve the accuracy of left atrial segmentation, therefore laying the foundation for subsequent work such as in atrial reconstruction. Meanwhile, our model is of great significance for assisting doctors in treating cardiovascular diseases.

Keywords: 3D medical image; attention mechanism; semi-supervised learning; left atrial segmentation

Citation: Liu, X.; Yin, R.; Yin, J. Attention V-Net: A Modified V-Net Architecture for Left Atrial Segmentation. *Appl. Sci.* **2022**, *12*, 3764. https://doi.org/10.3390/app12083764

Academic Editors: Chunhua Su, Keping Yu, Celestine Iwendi and Thippa Reddy Gadekallu

Received: 20 March 2022
Accepted: 7 April 2022
Published: 8 April 2022

1. Introduction

Cardiovascular diseases have become an important factor affecting human life and health [1,2]. In recent years, cardiac interventional therapy, as an advanced diagnosis and treatment method between internal and surgical procedures, has been widely used to cure cardiovascular diseases [3,4]. Among them, atrial septal puncture location surgery is the key to the success of interventional treatment of cardiovascular disease. How to accurately locate the puncture point to quickly puncture the atrial septum is the key to successful surgery. However, due to the lack of accurate and reliable 3D imaging feedback, atrial septal puncture location is still a challenging process, which requires doctors to have a great deal of experience. Left atrial segmentation is of great significance for doctors to quickly and accurately locate the atrial septal puncture position, and to better complete interventional surgery.

In recent years, algorithms based on deep learning technology, especially convolutional neural networks (CNNs), have made great breakthroughs in left atrial image processing tasks. Ciresan et al. [5] first introduced CNN into medical image segmentation, using a sliding window to fetch the local area around the pixel to train the network. However, this strategy only uses high-level features, and does not make full use of features with more marginal information. Furthermore, it is very slow because of the great quantity of training data. Later, Shelhamer et al. [6] proposed FCN (Full convolutional network), for image semantic segmentation, extending the classification from image level to pixel

level. However, these methods did not fully consider the relationship between pixels, and ignored the spatial regularization steps used in the common segmentation methods, resulting in a lack of spatial consistency. Therefore, the results obtained are not detailed enough. Then, Ronneberger et al. [7] proposed U-Net, which has both a contraction path that captures context information and a symmetric expansion path that allows precise positioning. Meanwhile, it can be trained based on FCN with a small number of images end-to end. Despite the popularity of previous approaches, they usually can only process 2D images. Unfortunately, most clinically applied cardiology data consists of 3D volumes. Therefore, Milletari et al. [8] recently proposed V-Net, an FCN based on 3D images. The dataset used by the author is made up of 3D medical images, which is different from the common 2D data. Operations such as convolution in the V-Net network structure also use 3D processing mode, in which a residual function inspired by [9] is also learned, which ensures convergence in less training time and obtains good segmentation accuracy.

In addition, in view of the scarcity of labeled left atrial image data, many methods were proposed in recent years to develop high-performance left atrial segmentation models to reduce labeled data. Among them, the semi-supervised learning framework has achieved many successful results, which can directly learn from limited labeled data and a large amount of unlabeled data to obtain high quality segmentation results. These methods can be roughly divided into two categories: regularization based on data perturbation or model perturbation [10,11] and consistency constraints based on multi-task level [12,13]. Most of them take V-Net as the backbone network of the algorithm, and the skip connection structure of V-Net model improves the shortcomings of FCN, such as not considering global context information and insufficient segmentation. However, the hierarchical convolution structure in a V-Net encoder-decoder neglects the local region features of the segmentation target to some extent, which may lead to the misclassification of the target and other objects. The attention mechanism method can enable the network to focus on the local region of the feature map. This motivates us to seek a suitable framework to adaptively learn the regions of interest in the input object, highlighting the structural features that are meaningful to the task, and thus improving the accuracy of the model prediction.

Based on V-Net model and attention mechanism, this paper designs a segmentation algorithm for left atrial MR images which are mostly 3D data formats, different from classic algorithms such as FCN, U-Net, and other networks. The proposed method can use the interdependence between channel mappings, to emphasize the interdependent feature mapping, and improve the feature representation of specific semantics. Therefore, the model pays more attention to the salient features that are meaningful for specific tasks. The results demonstrate that our method achieves significant improvements in left atrial segmentation.

In summary, this paper mainly makes the following contributions to the problem of how to make the network adaptively focus on the region of interest in the feature map:

(1) We propose a 3D left atrium segmentation model based on the attention mechanism, Attention V-Net, to simulate the interdependence between channels. In contrast to the previous 2D segmentation models, it can fully use the information between the 3D sequences of medical images. It can adaptively learn to highlight the salient features that are useful for tasks in the image, thus effectively enhancing the ability of feature expression.

(2) The proposed algorithm is applied to the semi-supervised framework of left atrium segmentation. The experimental results show that compared with the baseline, the proposed method obtains improved performance in terms of Dice, Jaccard, ASD, and 95HD, and also outperforms other state-of-the-art semi-supervised methods.

The rest of the paper is organized as follows. A brief review of the related works in left atrial segmentation and attention mechanism is given in Section 2. The architectures of the proposed Attention V-Net model is presented in Section 3. We present the experimental settings and the corresponding results in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

2.1. Left Atrial Segmentation

2.1.1. Supervised Segmentation

In 2015, a full convolution semantic segmentation network [6] achieved excellent segmentation results, laying the foundation for the application of deep learning in image segmentation. In recent years, many end-to-end segmentation techniques were developed in the field of medical imaging, and some early atrial segmentation algorithms [14–16], based on supervised learning, have shown good results. For example, the champion of the 2018 MICCAI Left Atrium Segmentation Challenge proposed a segmented network with two V-Nets [14]. The first is used to roughly locate the atrial center, it crops out a fixed size area according to the prediction results, the second network finely divides the parts cropped in the previous stage. F Isensee et al. [15] proposed a robust adaptive framework, nnU-Met, based on 2D U-Net and 3D U-Net. It replaces the complex process of artificial optimization using a systematic approach based on explicit and interpretable heuristic rules. It can perform plug-and-play on a variety of datasets and achieve the same effect of the state-of-the-art methods. Ahmad et al. [16] proposed a method to segment the left atrium and left ventricle simultaneously on the 3D MRI data of the heart. This method uses the traditional neighborhood-based method to track and superimpose the upper and lower slices. Then, it reconstructs the 3D model of the segmented left atrium and left ventricle according to the 2D format. These methods can improve the segmentation accuracy of atrial structure to a certain extent, but it is still difficult to solve the actual situation of medical image data with few labels and small samples. Therefore, the recent development of semi-supervised learning has resulted in changes to atrial segmentation algorithms.

2.1.2. Semi-Supervised Segmentation

The training of deep neural network needs a large amount of annotated data, which can only be generated by experienced doctors, and the cost is high. To solve this problem, some methods based on semi-supervised learning framework [10–13] recently achieved successful results. The semi-supervised learning framework can directly learn from a limited number of labeled data and a large number of unlabeled data to obtain high quality segmentation results. These methods can be roughly divided into two categories: regularization based on data perturbation or model perturbation [10,11], and consistency constraints based on multi-task framework [12,13].

Regularization Based Methods

Similar to [17], Li and Yu [10] proposed a method to regularize the model by adding perturbation to the input data. An iterative model needs to propagate forward twice, the input is the unchanged image and the changed image, respectively. Then the results of the changed image are inverted transformed to build the consistency loss of the two predicted results. The idea is simple, but it works well. Yu and Wang [11] designed the uncertainty perception strategy on the basis of Mean Teacher [18], and they adopted the consistency loss function to improve the performance of the student model. The model perturbation regularization is realized by adding different perturbations to the teacher model and the student model, such as adding noise to the input or adding dropout to the network. This adds some extra computing overhead, but you obtain a performance boost.

Multi-Task Frameworks

Li, Zhang, and He [12] adopted a multi-task network structure to segment the image and perform the signed distance graph regression at the same time, and the network uses the discriminator as the regularization term. This design can make the prediction distribution of the whole unlabeled data set smooth. Meanwhile, it can introduce strong shape and position as prior information to ensure the stability and robustness of the segmentation results. The dual-task consistency algorithm [13] establishes the prediction disturbance between different tasks. The output of different task branches should be

transformed into the same predefined space, and the consistency regularization between the two prediction mappings is explicitly performed. It establishes a task-level regularization which is completely different from the previous data-level regularization. The model is simple, and the calculation cost is not large.

2.2. Attention Model

The Attention mechanism can retrieve the key features through the convolutional layer of the network to output relevant weights. Generally, Sigmoid or SoftMax are used to calculate weights to identify the important features. It can be applied to any sequence model [19,20]. There are two types of attention mechanisms: soft attention [21] and hard attention [22]. Soft attention pays more attention to regions [23] or channels [24]. For example, ref. [23] proposes a module called spatial transformer, which can carry out corresponding spatial transformation of spatial domain information in images, so as to extract key information. The most important thing is that soft attention is differentiable. It can optimize the parameters through backward propagation optimization in the model training process, learning to obtain the weight of attention. Hard attention differs from soft attention in that it pays more attention to points. Meanwhile, hard attention is a random prediction process, which does not use all hidden layer states, but extracts information from a certain area in the form of a one-hot. Monte Carlo sampling is needed to estimate the gradient because the backward propagation cannot be performed directly in this way. The key point is that hard attention is not differentiable, and the training process is usually completed through reinforcement learning. In recent years, the attention mechanism can be explained intuitively by using the human visual mechanism. For example, our visual system tends to pay attention to part of the information that assists judgment in the image and ignore irrelevant information [25]. Similarly, in problems involving language or vision, some parts of the input may be more helpful to the decision than others. Our goal is to be able to help the decoder have a reference of the weights of different inputs when generating feature maps. The attention module allows the model to dynamically focus on certain parts of the input that contribute to the current task, it is a good choice for semantic segmentation of image.

3. Methodology

3.1. The Proposed Framework

In this section, we show the structure of our proposed Attention V-Net. We use V-net, an encoder-decoder structure, as the backbone. The encoder part is used for feature extraction and the decoder part can restore the image resolution. The features are extracted from the early stages of the encoder part of the V-Net to the decoder part using horizontal connections. Furthermore, we apply the 3D attention gate we designed on the connection part to use the interdependence between channels to learn the spatial weight information combined with the feature map, and obtain some structural regions with strong correlation.

The main structure is shown in Figure 1. The network consists of four encoder blocks and four decoder blocks, and the encoder blocks and the decoder blocks are connected symmetrically by the skip-connections. The parameters of the convolutional neural layer are shown in Table 1. The entire network is divided into different stages according to different resolutions, and each stage includes one to three convolutional layers. The upper and lower sampling parts is also changed from pooling to transposed convolution. In addition, the structure of adding residual connections at each stage is designed. The last convolution layer is converted into probabilistic segmentation of foreground and background regions through SoftMax.

In addition, inspired by the previous work on Attention U-Net [26], we design a 3D attention module. We apply this module to the skip-connection part based on the standard V-Net network. Furthermore, the modified framework can simulate the interdependence between channels. In the image segmentation task of this paper, all hidden states are important, but not equally important. The V-Net deepens the network through convolution and pooling operations. Finally, the separated pixels in high-dimensional space will have

stronger semantic information. We need a module to combine the contextual information of adjacent layers, and then use this information to guide the network to learn the regions of interest in the feature map. In this case, self-attention is needed to dynamically adjust the importance of different hidden states. Compared with the rugged strategy of Squeeze-Excitation [24] in which each channel of the feature map is multiplied by a weight coefficient, our attention strategy is more detailed. The proposed method has a unique adjustment factor for each value of each channel in the feature map. Furthermore, the designed attention gate can learn the spatial weight information combined with feature maps, so that the output has stronger semantic information and less noise interference. As the core contribution of our paper, we will explain it in detail in the next section.

Table 1. The parameters of the convolutional neural layers.

Block Name	Layer Name	Layer Configuration	Remark
Encoder Block (1)	conv 1	$5 \times 5 \times 5$, 16	Down-sampling path
	de-conv 1	$2 \times 2 \times 2$, stride 2	
Encoder Block (2)	conv 2_1	$5 \times 5 \times 5$, 32	
	conv 2_2	$5 \times 5 \times 5$, 32	
	de-conv 2	$2 \times 2 \times 2$, stride 2	
Encoder Block (3)	conv 3_1	$5 \times 5 \times 5$, 64	Down-sampling path
	conv 3_2	$5 \times 5 \times 5$, 64	
	conv 3_3	$5 \times 5 \times 5$, 64	
	de-conv 3	$2 \times 2 \times 2$, stride 2	
Encoder Block (4)	conv 4_1	$5 \times 5 \times 5$, 128	
	conv 4_2	$5 \times 5 \times 5$, 128	
	conv 4_3	$5 \times 5 \times 5$, 128	
	de-conv 4	$2 \times 2 \times 2$, stride 2	
Decoder Block (5)	conv 5_1	$5 \times 5 \times 5$, 256	Up-sampling path
	conv 5_2	$5 \times 5 \times 5$, 256	
	conv 5_3	$5 \times 5 \times 5$, 256	
	up-conv 5	$2 \times 2 \times 2$, stride 2	
Decoder Block (6)	conv 6_1	$5 \times 5 \times 5$, 128	
	conv 6_2	$5 \times 5 \times 5$, 128	
	conv 6_3	$5 \times 5 \times 5$, 128	
	up-conv 6	$2 \times 2 \times 2$, stride 2	
Decoder Block (7)	conv 7_1	$5 \times 5 \times 5$, 64	
	conv 7_2	$5 \times 5 \times 5$, 64	
	conv 7_3	$5 \times 5 \times 5$, 64	
	up-conv 7	$2 \times 2 \times 2$, stride 2	
Decoder Block (8)	conv 8_1	$5 \times 5 \times 5$, 32	
	conv 8_2	$5 \times 5 \times 5$, 32	
	up-conv 8	$2 \times 2 \times 2$, stride 2	
	conv 9_1	$5 \times 5 \times 5$, 16	
	conv 9_2	$1 \times 1 \times 1$, 2	

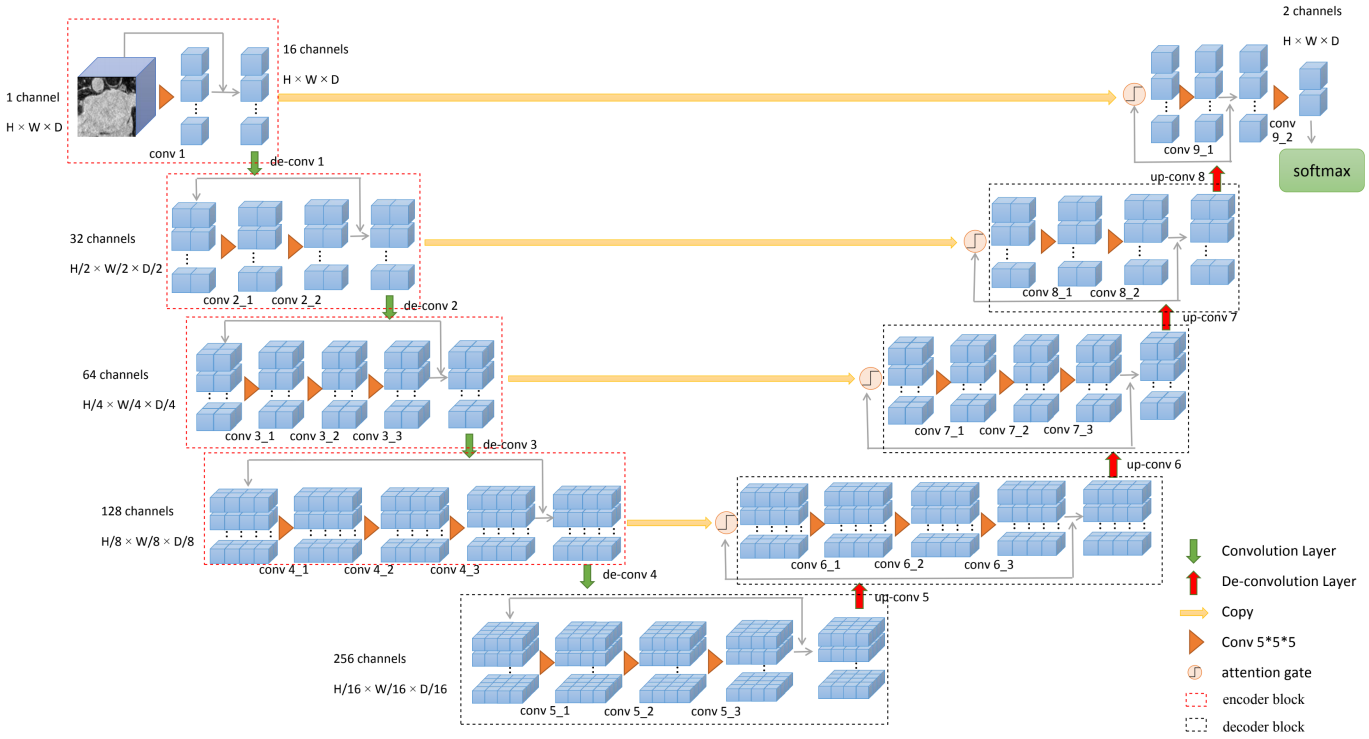


Figure 1. The proposed framework.

3.2. Attention Gates

Inspired by the work on Attention U-Net [26], we design a 3D attention gate for 3D data processing, combining it with the standard V-Net network. The 3D attention gate is applied before each level of skip-connection, which can make the network to put more weight on the characteristics associated with skip-connection. As shown in Figure 2, the 3D attention gate has two inputs: one is the feature map x transmitted from the extended pathway through a skip-connection, and the other is the feature map g output by the previous neural layer. Both x and g are sent to the $1 \times 1 \times 1$ convolution, turning them into the same number of channels without changing the size. After the upsampling operation to change the number of channels the same, they are accumulated along the direction of the channel and passed through the ReLU. Then, the output through another $1 \times 1 \times 1$ convolution and a sigmoid. Finally, we obtain an attention weight score, attention coefficients, $\alpha_i \in [0, 1]$, which can identify the salient features in the image. The output of the 3D attention gate is the element-wise multiplication of input feature-maps and attention coefficients: $\hat{x}^l_{i,c} = x^l_{i,c} \cdot \alpha^l_i$. In a default setting, a single scalar attention value is computed for each pixel vector $x^l_i = R^{F_l}$ where F_l corresponds to the number of feature-maps in layer l. The weight information can be added to the input feature map of this layer to eliminate the influence of irrelevant information in the skip-connection. As shown in Figure 2, the output of the 3D attention gate is connected to the next encoder through concatenate operation to integrate contextual information, where $C = C_x + C_g$, $H = H_x = H_g$, $W = W_x = W_g$, $D = D_x = D_g$. Therefore, the 3D attention gate module could help to achieve better segmentation performance.

We use additive attention [27] to obtain the attention weight coefficient, and the additive attention is formulated as follows:

$$\begin{aligned} q^l_{att} &= \psi^T \left(\sigma_1 \left(W_x^T x^l_i + W_g^T g_i + b_g \right) \right) + b_\psi \\ \alpha^l_i &= \sigma_2 \left(q^l_{att} \left(x^l_i, g_i; \Theta_{att} \right) \right) \end{aligned} \tag{1}$$

where $\sigma_2(x_{i,c}) = \frac{1}{1+\exp(-x_{i,c})}$ correspond to sigmoid activation function. The 3D attention gate is characterized by a set of parameters Θ_{att} containing linear transformations $W_x \in R^{F_l \times F_{\text{int}}}, W_g \in R^{F_g \times F_{\text{int}}}, \varphi \in R^{F_{\text{int}} \times 1}$ and bias terms $b_\psi \in R, b_g \in R^{F_{\text{int}}}$. The linear transformations are computed using channel-wise convolution for the input tensors. Furthermore, q_{att} defined the transformation operation of two inputs x and g under the parameters Θ_{att}.

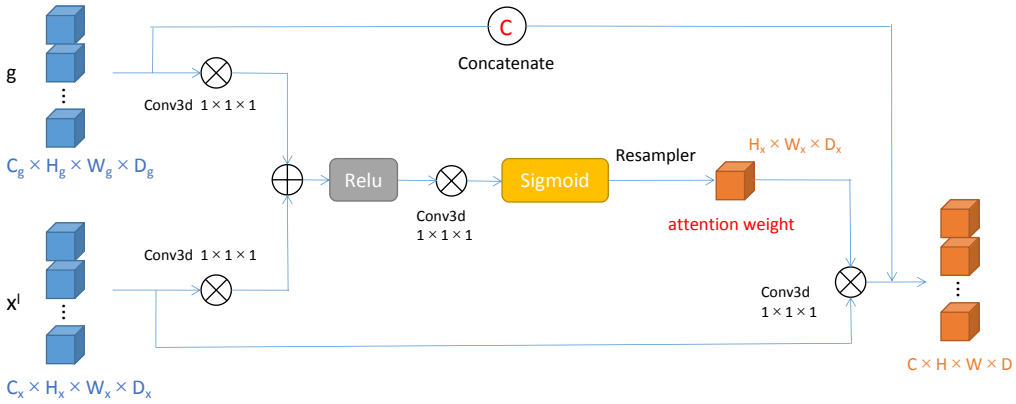


Figure 2. 3D attention gate.

4. Experiments and Results

4.1. Datasets and Pre-Processing

To evaluate the proposed method, we apply our algorithm on the left atrium dataset [28], which consists of 100 3D cardiac volume images. They are all obtained by GE-MRI (gadolinium-enhanced magnetic resonance imaging) from patients with atrial fibrillation. The original resolution of the data is $625 \times 625 \times 625$ mm^3. To fairly compare the advantages of the improved structure, we adopt the same data processing method as the semi-supervised learning algorithm DTC: 80 images are used for training, including 64 labeled images and 16 unlabeled images, and 20 images for testing. Meanwhile, we use the same pretreatment method.

4.2. Implementation Details and Evaluation Metrics

4.2.1. Implementation Details

In this part, we will make a brief introduction of the implementation of the Attention V-Net. All experiments are implements by Pytorch [29] library. Furthermore, Pytorch is an open source machine learning framework that accelerates the path from research prototyping to production deployment, which is provided by Facebook AI Research. More details can be found at https://pytorch.org/, (accessed on 20 March 2021). The experiments are carried out on a laboratory computer. The operating system is Ubuntu 16.04. The main required packages include python 3.6.13, CUDA9.0, cudnn7.6.5, Pytorch0.4.1.

In this work, we use the DTC algorithm as the baseline, where the V-Net network is the backbone. The dual-task V-Net is realized by adding a new regression layer at the end of the original V-Net network. The framework is trained by an SGD optimizer for 6000 iterations, which has an initial learning rate (lr) of 0.01 decayed by 0.1 every 2500 iterations. The batch size is four, consisting of two labeled images and two unlabeled images, the value of k is set to 1500 in this work. We randomly crop $112 \times 112 \times 80$ sub-volume as the network input. To avoid overfitting, we use the standard on-the-fly data augmentation methods during training stage. In the inference phase, we use a sliding window strategy to obtain the final results, which with a stride of $18 \times 18 \times 4$ for left atrium. At the inference time, we use the output of pixel-wise classification branch as the segmentation result.

4.2.2. Evaluation Metrics

We use overlap and surface distance measures to evaluate the segmentation, including Dice, Jaccard, the average surface distance (ASD), and the 95% Hausdorff Distance (95HD).

(1) Dice and Jaccard Coefficients : Given two binary segmentation masks, A and B, the Dice D and Jaccard coefficient J are defined as:

$$D = \frac{|A \cap B|}{|A| + |B|}, J = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where $|\cdot|$ gives the cardinality (i.e., the number of non-zero elements) of each set. Maximum and minimum values (1.0 and 0.0, repectively) for Dice and Jaccard coefficient occur when there is 100% and 0% overlap between the two binary segmentation masks, respectively.

(2) Average Surface Distance and 95% Hausdorff Distance: Let, S_A and S_B, be surfaces (with N_A and N_B points, respectively) corresponding to two binary segmentation masks, A and B, respectively. The average surface distance (ASD) S is defined as:

$$S = \frac{1}{2}\left(\frac{1}{N_A}\sum_{p \in S_A} d(p, S_B) + \frac{1}{N_B}\sum_{q \in S_B} d(q, S_A)\right) \tag{3}$$

Similarly, Hausdorff Distance (HD) H is defined as:

$$H = \max\left(\max_{p \in S_A} d(p, S_B), \max_{q \in S_B} d(q, S_A)\right) \tag{4}$$

where

$$d(p, S) = \min_{q \in S} d(p, q) \tag{5}$$

is the minimum Euclidean distance of point p from the points from the points $q \in S$. Hence, MSD computes the mean distance between the two surfaces, whereas, HD computes the largest distance between the two surfaces, and is sensitive to outliers.

Four complementary segmentation metrics are introduced to quantitatively evaluate the segmentation results. Dice and Jaccard, two region-based metrics, are used to measure the region mismatch. Average surface distance (ASD) and 95% Hausdorff Distance (95HD), two boundary-based metrics, are used to evaluate the boundary errors between the segmentation results and the ground truth.

4.3. Results and Analysis

4.3.1. Comparison with Other Semi-Supervised Methods

In this paper, we design an attention module and apply it to the V-Net network. It can use the interdependence between channels to learn the spatial weight information combined with the feature map, and to obtain some structural regions with strong correlation. The final feature of each channel is the weighted sum of the features filtered by the correlations between channels and the original features. The feature correlation of the channels simulates the remote semantic dependence between features. It helps to maintain the relationship between different channel feature maps, enlarge the inconsistency between categories, and make the feature maps transmitted by the skip-connection have stronger semantic information.

As shown in Table 2, on the 2018 left atrium segmentation dataset, we replace the backbone network V-Net of the DTC algorithm with Attention V-Net, comparing it with the recurring results of the basic framework DTC. The effect of Attention V-Net is 0.56% higher on Dice, 0.74% higher on Jaccard, 0.16 voxel greater on ASD, and 0.32 voxel greater on 95HD. Our method outperforms all the other semi-supervised networks in both Dice (89.08%) and Jaccard (80.48%), and achieves competitive results on other metrics. We compare our framework with four semi-supervised segmentation methods, including entropy minimization approch (Entropy Mini) [30], uncertainty-aware mean teacher model (UA-MT) [11], shape-aware adversarial network (SASSNet) [12], and dual-task consistency model (DTC). Please note that we use the official code and results of Entropy Mini and

UA-MT, and reimplement the SASSnet and DTC for left atrium segmentation. Table 2 shows the quantitative comparison of these methods. It can be found that our method achieved the better accuracy than other semi-supervised segmentation methods on all the evaluation metrics. It shows that our structure can improve the regional similarity of cardiac segmentation results, and also has a significant improvement in the accuracy of the boundary. Thus, our experiments can prove that our attention mechanism can enhance the accuracy of model segmentation, and help to improve the performance of left atrium segmentation.

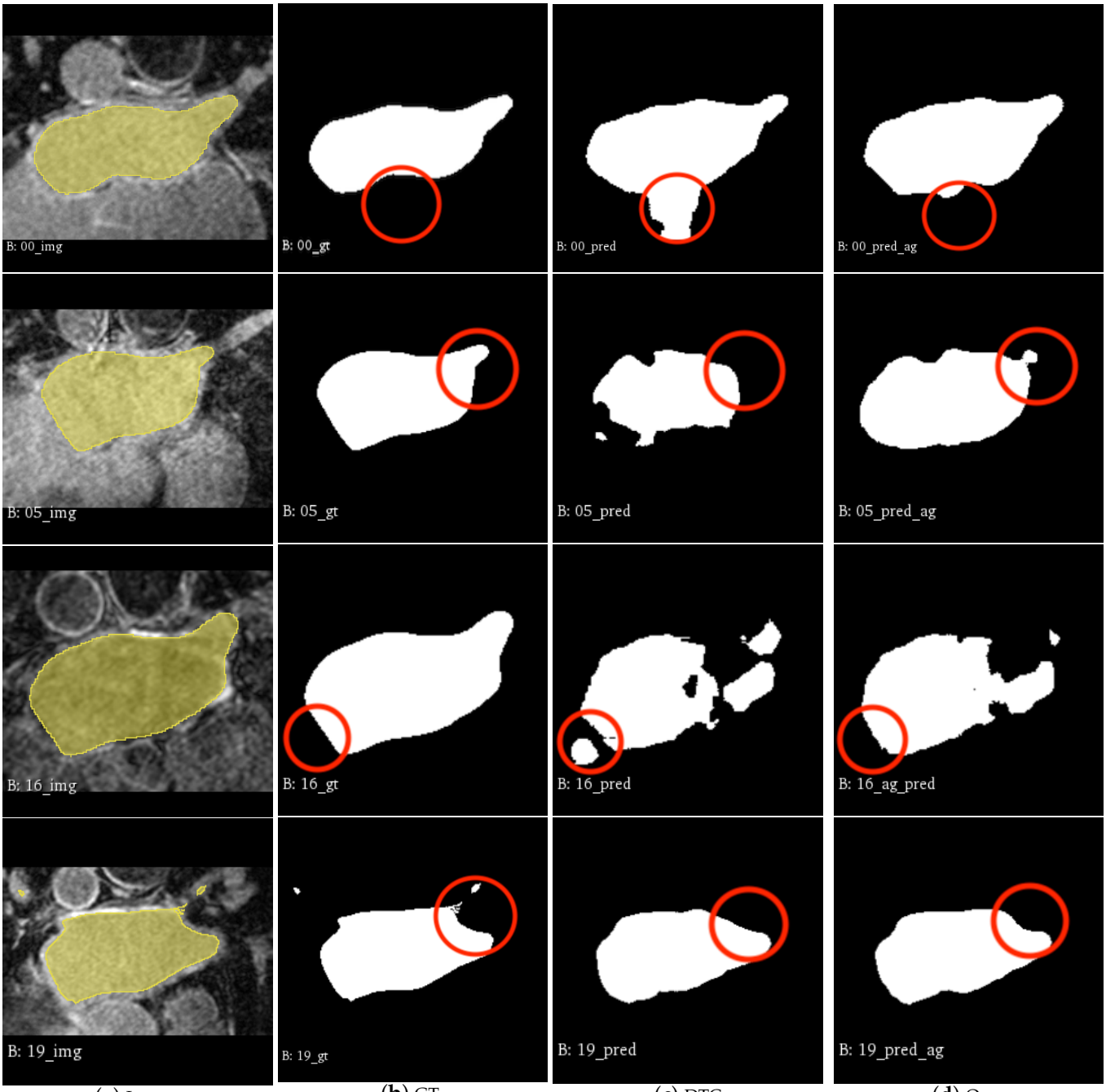
Table 2. Comparison with Other Semi-supervised Methods.

Method	Metrics			
	Dice (%)	**Jaccard (%)**	**ASD (Voxel)**	**95HD (Voxel)**
Entropy Mini (CVPR 2019)	88.45	79.51	3.72	14.14
UA-MT (MICCAI 2019)	88.88	80.21	2.26	7.32
SASSnet (MICCAI 2020)	88.78	80.04	2.96	10.30
DTC (AAAI 2021)	88.52	79.74	2.08	8.54
Attention V-Net	89.08	80.48	1.94	8.22

4.3.2. Visualization

The attention model can process a large amount of data information and generate weight probability information. The weight information can reflect the different degrees of importance of local regions, so as to allow the network to focus on those regions that are of great importance and more interesting to us. The Attention V-Net in this paper can use the interdependence between channel mappings to emphasize the interdependent feature mapping, and improve the feature representation of specific semantics. Therefore, the model pays more attention to the salient features that are meaningful for specific tasks, while suppressing any insignificant parts. Finally, the Attention V-Net plays a role in enhancing the ability of image feature representation.

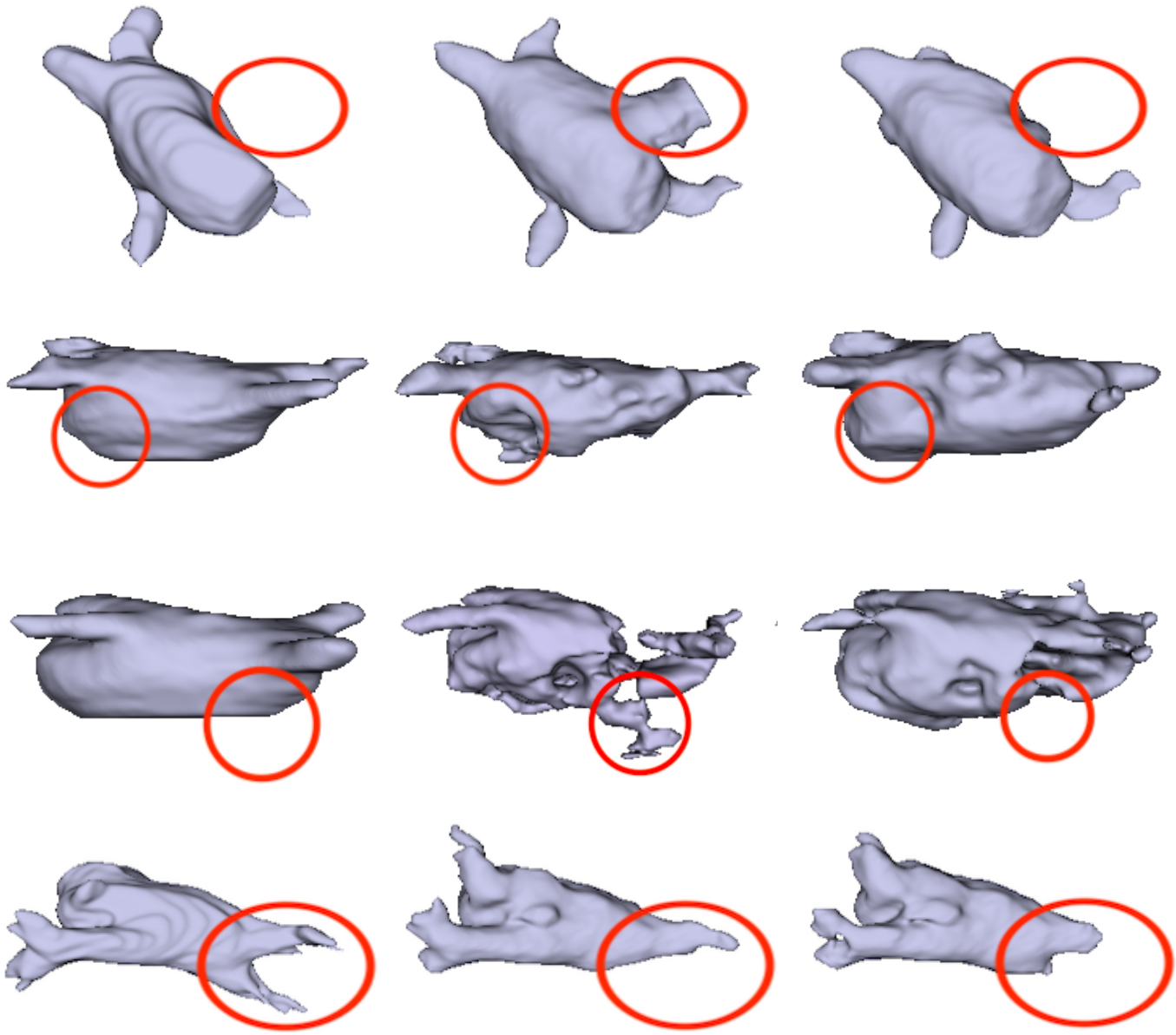
We use the Attention V-Net model based on DTC to randomly segment four 3D datas in the test dataset and reconstruct the left atrium image. Figure 3 is the reconstructed image of ground truth and prediction. Comparing the 2D visualization in Figure 3(1) and the 3D visualization in Figure 3(2), it can be seen that the overall prediction result of the Attention V-Net model is very close to the manual annotation. In contrast, DTC often misses the internal area of the target object, resulting in irregular shapes, while the model with the added attention mechanism can better simulate the segmentation results of the left atrium, improving the accuracy of the integrity of the internal area and boundary. Compared with other methods, our results have a higher overlap ratio with the ground truth, produce fewer false positives, and preserve more details, which further indicates the effectiveness, generalization, and robustness of our proposed method. The 3D representation of our structure is closer to the real left atrium model, but there are still deviations in the details, and cannot be completely consistent with the real shape.



(**a**) Image (**b**) GT (**c**) DTC (**d**) Ours

(1) Comparison of 2D segmentation visualization results

Figure 3. *Cont.*



(**e**) GT (**f**) DTC (**g**) Ours

(2) Comparison of 3D segmentation visualization results

Figure 3. 2D and 3D Visualization of the segmentation by DTC [9] and our method, where GT denotes ground truth segmentation.

5. Discussion and Conclusions

In this paper, we propose a fully convolutional neural network based on the attention mechanism, which can be used for 3D medical image segmentation tasks. In comparison with other end-to-end semantic segmentation networks, the proposed network can adaptively learn to highlight the salient features of the image that are useful for the task, by designing a new 3D attention module. The network also learns attention weights and concatenates them at each layer of the skip-connection part of the V-Net, which further improves accuracy. Meanwhile, it also can process the 3D image data, using the information between the 3D sequences of medical images synthetically.

We apply it to the left atrium segmentation framework based on semi-supervised learning, and we evaluate it on the dataset of the 2018 left atrial challenge. The experimental

results show that, compared with the original algorithm, the performance indexes such as Dice, Jaccard, ASD, and 95HD are improved. Moreover, compared with the current advanced semi-supervised segmentation algorithm, the experimental results show that our proposed Attention V-Net can improve the accuracy of medical image segmentation,w hich is of great significance to clinical diagnosis and treatment. The substantial increase in segmentation accuracy comes with a negligible increase in model complexity. Hence, our proposed 3D attention gate module can be extended to some other 3D medical organ segmentation tasks (e.g., brain structure or tumor segmentation) to boost performance. We believe that our model can be a crucial component for neural networks in many medical applications.

In the future related research of medical image analysis, we could pay more attention to adaptive feature learning and the multi-scale feature fusion. It may obtain better feature results and experimental performance, providing a reliable basis for clinical diagnosis and pathology research.

Author Contributions: Conceptualization, X.L. and J.Y.; methodology, R.Y.; software, X.L.; validation, X.L. and R.Y.; formal analysis, J.Y.; investigation, R.Y.; writing—original draft preparation, X.L.; writing—review and editing, J.Y.; visualization, R.Y.; project administration, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported partly by the National Natural Science Foundation of China (Grant No. 62173045, 61673192), and partly by the Fundamental Research Funds for the Central Universities(Grant No. 2020XD-A04-2), and partially supported by BUPT Excellent Ph.D. Students Foundation (CX2021314).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Our method is evaluated on the dataset of the 2018 left atrial challenge.

Conflicts of Interest: No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

References

1. Narayan, S.M.; Rodrigo, M.; Kowalewski, C.A.; Shenasa, F.; Meckler, G.L.; Vishwanathan, M.N.; Baykaner, T.; Zaman, J.A.B.; Paul, J.; Wang, P.J. Ablation of focal impulses and rotational sources: What can be learned from differing procedural outcomes. *Curr. Cardiovasc. Risk Rep.* **2017**, *11*, 27. [CrossRef]
2. Hansen, B.J.; Zhao, J.; Csepe, T.A.; Moore, B.T.; Li, N.; Jayne, L.A.; Kalyanasundaram, A.; Lim, P.; Bratasz, A.; Powell, K.A.; et al. Atrial fibrillation driven by micro-anatomic intramural re-entry revealed by simultaneous sub-epicardial and sub-endocardial optical mapping in explanted human hearts. *Eur. Heart J.* **2015**, *36*, 2390–2401. [CrossRef] [PubMed]
3. Njoku, A.; Kannabhiran, M.; Arora, R.; Reddy, P.; Gopinathannair, R.; Lakkireddy, D.; Dominic, P. Left atrial volume predicts atrial fibrillation recurrence after radiofrequency ablation: A meta-analysis. *EP Eur.* **2017**, *20*, 33–42. [CrossRef] [PubMed]
4. Higuchi, K.; Cates, J.; Gardner, G.; Morris, A.; Burgon, N.S.; Akoum, N.; Marrouche, N.F. The spatial distribution of late gadolinium enhancement of left atrial mri in patients with atrial fibrillation. *JACC Clin. Electrophysiol.* **2017**, *4*, 49–58. [CrossRef] [PubMed]
5. Ciresan, D.; Giusti, A.; Gambardella, L.M.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9.
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. 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; pp. 234–241.
8. 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; pp. 565–571.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
10. Li, X.; Yu, L.; Chen, H.; Fu, C.W.; Heng, P.A. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv* **2018**, arXiv:1808.03887.

11. Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; Heng, P.-A. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In Proceedings of the International Conference on Medical Imaging Computing for Computer Assisted Intervention, Shenzhen, China, 13–17 October 2019.
12. Li, S.; Zhang, C.; He, X. Shape-aware semi-supervised 3D semantic segmentation for medical images. In Proceedings of the International Conference on Medical Imaging Computing for Computer Assisted Intervention, Lima, Peru, 4–8 October 2020; pp. 552–561.
13. Luo, X.; Chen, J.; Song, T.; Wang, G. Semi-supervised medical image segmentation through dual-task consistency. *arXiv* **2020**, arXiv:2009.04448.
14. Xia, Q.; Yao, Y.; Hu, Z.; Hao, A. Automatic 3D atrial segmentation from GE-MRIs using volumetric fully convolutional networks. In Proceedings of the International Workshop on Statistical Atlases and Computational Models of the Heart, Granada, Spain, 16 September 2018; pp. 211–220.
15. Isensee, F.; Jäger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. Automated design of deep learning methods for biomedical image segmentation. *arXiv* **2019**, arXiv:1904.08128.
16. Ahmad, I.; Hussain, F.; Khan, S.A.; Akram, U.; Jeon, G. CPS-based fully automatic cardiac left ventricle and left atrium segmentation in 3D MRI. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4153–4164. [CrossRef]
17. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
18. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–10.
19. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
20. Banerjee, S.; Lyu, J.; Huang, Z.; Leung, H.F.F.; Lee, T.T.-Y.; Yang, D.; Su, S.; Zheng, Y.; Ling, S.-H. Light-Convolution Dense Selection U-Net (LDS U-Net) for Ultrasound Lateral Bony Feature Segmentation. *Appl. Sci.* **2021**, *11*, 180. [CrossRef]
21. Xiao, T.J.; Xu, Y.C.; Yang, K.Y.; Zhang, J.X.; Peng, Y.X.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 842–850.
22. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
23. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
25. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
26. Oktay, O.; Schlemper, J.; Folgoc, L.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.
27. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
28. Xiong, Z.; Xia, Q.; Hu, Z.; Huang, N.; Bian, C.; Zheng, Y.; Vesal, S.; Ravikumar, N.; Maier, A.; Yang, X.; et al. A global benchmark of algorithms for segmenting late gadolinium-enhanced cardiac magnetic resonance imaging. *Med. Image Anal.* **2020**, *67*, 101832. [CrossRef] [PubMed]
29. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
30. Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2517–2526.