Efficient Knowledge Distillation for Brain Tumor Segmentation

Yuan Qi 1, Wenyin Zhang 1,*, Xing Wang 1, Xinya You 2, Shunbo Hu 1 and Ji Chen 1

1 School of Information Science and Engineering, Linyi University, Linyi 276000, China
2 College of Engineering, University of Sydney, Sydney 2006, Australia
* Correspondence: zhangwenyin@lyu.edu.cn

Abstract: Deep learning has allowed great progress to be made in obtaining more accurate prediction results for brain tumor segmentation. The current mainstream research approaches obtain segmentation accuracy improvements by modifying deep-learning model architectures, while ignoring the computational and storage efficiency issues of segmentation. In this paper, we proposed an improved knowledge distillation method: coordinate distillation (CD), which integrates channel and space information and completes brain tumor segmentation by training the student network with the teacher network without changing the original network architecture. Experimental results showed that the method was effective and that it could enhance the segmentation accuracy of brain tumors without changing the segmentation efficiency.

Keywords: knowledge distillation; brain tumor segmentation; lightweight networks; medical image segmentation

1. Introduction

In the field of medical image processing and analysis [1], medical image segmentation is a critical and complex step. Deep learning is used to extract the important regions of medical images to assist doctors in clinical treatment and pathology research automatically. The task of segmenting brain tumors, as a type of medical image segmentation, is also a focus of researchers’ attentions. Tumors originating from the neuroepithelium are collectively called gliomas and are the most common intracranial primary tumors. Typically, gliomas are classified into four grades [2], where the higher the grade, the worse the prognosis is. Much research on brain tumor segmentation has been proposed; however, due to the uncertainty of a brain tumor’s shape, location and size, it is impossible to accurately locate it during segmentation, making brain tumor segmentation a severe challenge [3–5].

Meanwhile, a large majority of the approaches improve the original model accuracy by modifying the model architecture, inevitably adding various expensive computational components and expanding the required storage space. For a long time, people have discussed how to solve this dilemma and find a better balance between efficiency and accuracy. Hinton [6] introduced knowledge distillation (KD) to the deep-learning domain, which has gained much attention for its simplicity and efficiency. The core idea of KD is to use soft labels to learn class distribution. Specifically, first train a teacher network to obtain its output, and then, in the process of student network training, let the output of the student network learn the output of the teacher network. The final effect of the knowledge distillation method is that the output of the student network is similar to that of the teacher network. Many methods have been proposed to improve the learning accuracy of students’ networks following KD. For example, when using KD, Yang et al. [7] found that compared to the ground truth class, the secondary class could effectively learn inter-class similarity to obtain knowledge better. Xie et al. [8] focused the problem on a data issue; they first trained the teacher network to generate pseudodatasets, and then they combined it with the original datasets to form noise datasets to learn knowledge. Phuong et al. [9] stated that KD performed poorly when the difference between the student and the teacher capacity was
large, and they proposed a new ensemble distribution KD for training. Romero et al. [10] proposed learning knowledge through intermediate feature layers. Jiao et al. [11] proposed a new transformer distillation, which performed the transformer distillation in the pretraining and specific learning phases. Sanh et al. [12] just simply applied KD to Bert, making it 60% faster and guaranteeing 97% accuracy. Wang et al. [13] distilled the final transformer layer from the teachers and introduced teaching assistants to aid distillation. Unfortunately, the above approach did not take into account the validity of distillation in brain tumor segmentation scenarios. Only a few researchers have applied knowledge distillation methods to brain tumor segmentation; however, most of them focused on solving the problems of small datasets and multimodality [14–16], and they did not focus on the efficiency of brain tumor segmentation.

In this paper, considering brain tumor segmentation as a structured dense prediction task, we proposed a new knowledge distillation method. Different from channel attention, which just considers the importance of different channels, the proposed method of coordinate distillation (CD) considers spatial information. It encodes channels along two spatial directions, which is particularly important for spatial structure capture in segmentation tasks and helps the model to segment brain tumors accurately. Our extensive experiments on the public dataset BraTS2018 demonstrated the remarkable performance of the method.

In general, our contributions are summarized as follows.

1. By studying knowledge distillation for training networks, a simple and effective CD module is presented in the paper, which aims to model channel correlation and remote dependency in order to improve student network segmentation accuracy.

2. An effective knowledge distillation architecture is provided for brain tumor segmentation. Experiments on the public dataset BraTS2018 showed the feasibility of the method.

2. Related Work

2.1. Brain Tumor Segmentation

In recent years, research on brain tumor segmentation has been continuously developed. Ronneberger et al. [17] proposed the UNet, whose biggest feature was the U-shaped structure and the skip connection. The skip connection combined low-resolution information and high-resolution information; low-resolution information helped the category identification of objects, and high-resolution features provided accurate positions for segmentation. This network was perfectly suitable for medical image segmentation as the parameters reached 39.4 M, which also made it suitable for playing the teacher network. After UNet was proposed, a series of upgraded networks appeared, but most of them made some changes based on UNet. For example, Isensee et al. [18] reported an improved UNet for brain tumor segmentation that successfully avoided overfitting by using the Dice loss and data expansion. In [19], a comprehensive scheme of a long connection and a short connection was proposed. The parameter of the model was only 9.2 M, but the precision was low, so it was suitable to be a student network. Zhang et al. proposed DeepResUNet [20], pointing out that it was important to combine low-level semantic information with high-level semantic information using a skip connection; however, it was difficult to train such a deep neural network with limited medical training samples. The residual units were proposed to simplify the deep network, thus the model parameters were reduced to 32.6 M; however, the same problem of accuracy reduction still existed, so it was suitable to act as a student network in our experiments. Oktay et al. [21] indicated that there was a lot of redundant information in the map extracted from the encoder. Soft attention was proposed to add at the end of the skip connection to suppress the activation of irrelevant regions, in order to segment the medical images better; however, at the same time, it increased the computational complexity, causing the parameters to have values up to 39.7 M. This network was also suitable for playing a teacher network. Other methods, such as [22–24], also enhanced UNet to segment brain tumors better. In addition, some researchers have recognized the importance of capturing spatial continuity information, confirming that 3D brain tumor segmentation is more effective than 2D brain tumor segmentation [25–27].
2.2. Knowledge Distillation

Knowledge distillation is a method that improves the performance of lightweight models by transferring knowledge from powerful, but heavy, networks without losing efficiency. Hilton et al. formally defined distillation and proposed corresponding training methods in 2014. After that, many distillation strategies were developed, which were roughly classified into three categories [28]. The first was logits-based knowledge, which extracted knowledge from the output layer of the teacher network. Zhang et al. [29] broke this predefined “strong and weak relationship” and proposed deep mutual learning (DML), which allowed a group of student networks to learn from each other in the training process, rather than a static predefined one-way transition path between teachers and students. In [30], the author observed that when the teacher network segmentation performance was very great, the distilled student network segmentation performance might not be very great, presumably due to capacity mismatch, which caused the student not to imitate the teacher and instead brought out the main loss. The author proposed an early-stop teacher regularization for distillation, which should be stopped early, when near convergence. Next was feature-based knowledge, which extracted knowledge from the middle-hidden layer of the teacher network. Compared with the first category, it could better learn deeper knowledge. Liu et al. [31] included pairwise distillation techniques and GAN to extract knowledge. Qin et al. [32] proposed the idea of computing the interclass contrast between different tissue regions under the guidance of the marker segmentation template. Zagoruyko et al. [33] guided students’ attentional maps to learn those of the teachers so that the students’ attentional maps were similar to those of the teachers. Finally, relation-based knowledge, which learnt the relationship between layers to acquire knowledge, argued that learning the relationship between layers was similar to learning how to solve a problem. Yim et al. [34] did not fit teacher output but fitted the relationship between the layers of the teacher model. The relationship was defined by the inner product (dot product) between layers. Tung et al. [35] proposed a preserved similarity loss to motivate the student to mimic the teacher regarding the relational expressions within the data. Our proposed approach distilled knowledge from both the output and the intermediate hidden layer of the teacher network. Compared with the first category, it could better learn deeper knowledge. Liu et al. [31] included pairwise distillation techniques and GAN to extract knowledge. Qin et al. [32] proposed the idea of computing the interclass contrast between different tissue regions under the guidance of the marker segmentation template. Zagoruyko et al. [33] guided students’ attentional maps to learn those of the teachers so that the students’ attentional maps were similar to those of the teachers. Finally, relation-based knowledge, which learnt the relationship between layers to acquire knowledge, argued that learning the relationship between layers was similar to learning how to solve a problem. Yim et al. [34] did not fit teacher output but fitted the relationship between the layers of the teacher model. The relationship was defined by the inner product (dot product) between layers. Tung et al. [35] proposed a preserved similarity loss to motivate the student to mimic the teacher regarding the relational expressions within the data. Our proposed approach distilled knowledge from both the output and the intermediate hidden layers because it was easier to learn the answers directly than to learn the solution ideas.

3. Methods

The proposed integral distillation structure is shown in Figure 1. Let an MRI brain tumor image be denoted by $X \in \mathbb{R}^{h \times w}$, which was used as an input image, and a prediction image of the same size $Q \in \mathbb{R}^{h \times w}$ was finally output, where $h$ represents the height of the image and $w$ represents the width of the image.

The overall distillation structure was realized by the two green rectangular modules in the figure: CD and KD. KD drove the student’s final layer output to simulate that of the teacher. CD drove the student’s intermediate feature mapping $E_s^1, E_s^2$ to simulate the teacher’s intermediate feature mapping $E_t^1, E_t^2$. Finally, the student’s segmentation task loss $L_{seg}$ needed to be added to ensure the input domain’s basic performance. Through this distillation structure, the student network could learn the segmentation experience of the teacher network, and then better carry out its segmentation task.

3.1. Knowledge Distillation

Using the basic method of knowledge distillation KD [6], we treated the partitioned graph as a collection of pixelwise labeling problems, and by computing the difference between the last layer of the two networks, the student $s$ could learn from the teacher $t$. The training was calculated by the following loss function.

$$L_{KD} = \frac{1}{N} \sum_{i \in N} KL(P_s^i \parallel P_t^i)$$  \hspace{1cm} (1)
where $N = w \times h$ represents all the pixels in the segmented graph and $KL(\cdot)$ is the Kullback–Leibler divergence [36]. $P_i^s$ represents the probability of the $i$th pixel in the prediction map of the student network, and $P_i^t$ represents the probability of the $i$th pixel in the prediction map of the teacher network.

**Figure 1.** The proposed knowledge distillation architecture. The orange box represents the teacher network, and the blue box represents the student network. They obtained the same image as input. The green box between the two networks, from left to right, is the coordinate distillation (CD) module and the knowledge distillation (KD) module.

### 3.2. Coordinate Distillation

Inspired by [37], we recommended adding location information to the channel attention (AT) [33], prompting the attention graph to capture remote dependencies with accurate location information, which is especially important for spatial structure capture in segmentation tasks. Based on the works of channel attention transfer, we assumed that a neuron's absolute value indicated its importance. Specifically, given the input $E$, we summed the channel dimension along the horizontal $w$ and vertical $h$ coordinates, so the output could be described as follows.

$$
\phi_w(E) = \sum_{i=1}^{c} |E_i^w|^2
$$

$$
\phi_h(E) = \sum_{i=1}^{c} |E_i^h|^2
$$

where $c$ is the number of channels. Next, we multiplied the output $\phi_w$ and $\phi_h$ to obtain the final coordinate attention map $y$.

$$
y = \sigma(\phi_w(E)) \times \sigma(\phi_h(E))
$$

where $\sigma(\cdot)$ is the Sigmoid operation. This training was calculated by the following loss.

$$
L_{CD} = \sum_{(i,j) \in M} \left\| \frac{y_i^s}{\|y_i^s\|_1} - \frac{y_i^t}{\|y_i^t\|_1} \right\|_1
$$
where $M$ is the set of index pairs of all possible positions with the same embedding size, $(i, j)$ is the samples in $M$. The operation $\| \cdot \|_1$ is the $l_1$ normalization.

### 3.3. Network Optimization

The total loss function is shown below.

$$L_{total} = \alpha L_{seg} + \beta_1 L_{KD} + \beta_2 L_{CD}$$  \hspace{1cm} (6)

where $L_{seg}$ is the cross-entropy loss. The hyperparameter $\alpha$ was set to 0.9, $\beta_1$ was 0.1 and $\beta_2$ was 1.8. Our experiments demonstrated that such parameter settings were most conducive to the effectiveness of our method.

### 4. Experiments

#### 4.1. Datasets

The experiment was conducted on the publicly available brain tumor segmentation dataset BraTS2018 [38]. Multimodal brain tumor segmentation has the longest history of all MICCAI competitions and has the largest number of participants in almost all competitions each year. It is a good platform for understanding the cutting edge of segmentation methods. The dataset has a training set of 285 cases, using T1, T2, T1CE and flair modes, respectively, and needs to segment the whole tumor (WT), tumor core (TC) and enhanced tumor (ET). The size of each MR image is $240 \times 240 \times 155$.

#### 4.2. Implementation Details

To verify the effectiveness of our distillation method, we employed networks with good performances as teacher networks, such as UNet and AttUNet. We used networks with less effective performance as student networks, such as DeepResUNet and UNet++. We implemented all the segmentation networks in our experiments with PyTorch and ran them on a single NVIDIA GeForce RTX 2060 GPU (6 GB). The Adam algorithm was used to train the networks with a weight decay of 0.0001. The batch size was 8 and the initial learning rate was 0.0003. On the BraTs dataset, we first normalized each sequence of the brain MR images by z score, cropped the redundant background information around the brain images, then sliced them into 2D data to fit the 2D network. In addition, to mitigate the category imbalance problem, the parts of the slices that did not contain lesions were discarded, and, finally, because it was multimodal, the slices of each modality were combined into a training case with four channels. A total of 285 BraTS2018 cases were training sets. As BraTs only published a training set and did not have test set, the test set was part of the training set that BraTS2019 had more of than BraTS2018, including 49 HGGs and 1 LGG. There were 50 test sets in total. During the training, we used the $160 \times 160$ images as training samples, and we trained all the networks to reach the fusion of 20 epochs of training.

#### 4.3. Evaluation Metrics

The Dice score and Hausdorff distance are generally employed as measures of brain tumor segmentation. The Dice score measures the similarity between two samples, the range of values is 0–1, the best value is 1 and the worst value is 0. The definition of Dice is as follows.

$$\text{Dice} = \frac{2|Q \cap G|}{|Q| + |G|}$$  \hspace{1cm} (7)

where $Q$ represents the prediction mask and $G$ represents the ground truth.

The Hausdorff distance is a metric that evaluates the maximum segmentation boundary distance between $Q$ and $G$, the definition is as follows.

$$\text{HD} = \max \left\{ \min_{q \in Q} \max_{g \in G} d(q, g), \min_{g \in G} \max_{q \in Q} d(q, g) \right\}$$  \hspace{1cm} (8)
5. Ablation Study
5.1. Effectiveness of Distillation Methods

The proposed distillation methods, CD and KD, are collectively referred to as CK, and a large number of experiments were carried out on the BraTs2018 dataset to verify the effectiveness of CK. Figure 2 shows some visual examples, which show that our method corrected some of the students’ mistakes and made the predicted mask closer to the ground truth.

![Figure 2](image)

Figure 2. The segmentation results were obtained by applying our distillation method in four cases (a–d) of the BraTS validation dataset. ET—yellow, TC—yellow+red, WT—yellow+red+green. The teacher was UNet and the student was DeepResUNet.

We used UNet and AttUNet as our teachers and DeepResUNet and UNet++ as our students. Table 1 lists the performance of our distillation method, CK, for different combinations of teacher and student networks. The experimental results showed that the knowledge distillation architecture proposed in this paper played a positive role in improving the segmentation accuracy of the student network. After learning from UNet, DeepResUNet had maximum improvements of 1.25, 0.72 and 2.84 in ET, WT and TC Dice scores, respectively, and 0.08, 0.07 and 0.16 in Hausdorff distance, respectively. Meanwhile, after learning from the teacher AttUNet, the Dice score of the student network DeepResUNet in ET broke from 78.44 to 80.22, and the Dice score of the student network UNet++ in ET broke from 78.47 to 80.48. Thus, we honored DeepResUNet as the most outstanding student. To more visually demonstrate the effectiveness of our method, the obtained experimental results were compared with the currently popular networks in this experiment using AttUNet and UNet++ as the teacher network and the student network, respectively. Kong et al. [24] proposed HybridResNet, using resnet blocks for downsampling and normal double convolution blocks for upsampling. Compared to UNet, ResNet blocks helped to improve the performance of the network, so we also included [24] in our comparison network. In Table 2, it can be seen that the UNet++ using our method achieved the crown while keeping the lower parameters.
Table 1. The results of our experiments on different combinations of networks on BraTS2018. The bold values are the highest Dice scores and lowest Hausdorff distances in their columns. When the segmentation index of the teacher was lower than that of the student, theoretically, knowledge distillation was not applicable to this situation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice Score (%)</th>
<th>Hausdorff Distance</th>
<th>#Params (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ET</td>
<td>WT</td>
<td>TC</td>
</tr>
<tr>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1: UNet</td>
<td>78.83</td>
<td>85.69</td>
<td>85.81</td>
</tr>
<tr>
<td>T2: AttUNet</td>
<td>78.79</td>
<td>85.31</td>
<td>85.82</td>
</tr>
</tbody>
</table>

Students and their performances by our approach

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice Score (%)</th>
<th>Hausdorff Distance</th>
<th>#Params (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ET</td>
<td>WT</td>
<td>TC</td>
</tr>
<tr>
<td>S1: DeepResUNet</td>
<td>78.44</td>
<td>85.19</td>
<td>85.36</td>
</tr>
<tr>
<td>S1+T1 (ours)</td>
<td>79.69</td>
<td>85.91</td>
<td><strong>88.20</strong></td>
</tr>
<tr>
<td>S1+T2 (ours)</td>
<td>80.22</td>
<td><strong>86.49</strong></td>
<td>87.14</td>
</tr>
<tr>
<td>S2: UNet++</td>
<td>78.47</td>
<td>85.16</td>
<td>85.60</td>
</tr>
<tr>
<td>S2+T1 (ours)</td>
<td>79.13</td>
<td>85.62</td>
<td>87.64</td>
</tr>
<tr>
<td>S2+T2 (ours)</td>
<td><strong>80.48</strong></td>
<td>86.46</td>
<td>87.57</td>
</tr>
</tbody>
</table>

Table 2. Comparison with current popular methods.

<table>
<thead>
<tr>
<th>Models</th>
<th>GFLOPs</th>
<th>#Params</th>
<th>Dice Score (%)</th>
<th>Hausdorff Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ET</td>
<td>WT</td>
<td>TC</td>
<td>ET</td>
</tr>
<tr>
<td>UNet</td>
<td>18.03</td>
<td>39.4M</td>
<td>78.83</td>
<td>85.69</td>
</tr>
<tr>
<td>DeepResUNet</td>
<td>23.27</td>
<td>32.6M</td>
<td>78.44</td>
<td>85.19</td>
</tr>
<tr>
<td>AttUNet</td>
<td>18.47</td>
<td>39.7M</td>
<td>78.79</td>
<td>85.31</td>
</tr>
<tr>
<td>UNet++</td>
<td>13.52</td>
<td>9.2M</td>
<td>78.47</td>
<td>85.16</td>
</tr>
<tr>
<td>HybridResUNet</td>
<td>22.0</td>
<td>31.6M</td>
<td>76.81</td>
<td>84.08</td>
</tr>
<tr>
<td>UNet++ (ours)</td>
<td><strong>13.52</strong></td>
<td><strong>9.2M</strong></td>
<td><strong>80.48</strong></td>
<td><strong>86.46</strong></td>
</tr>
</tbody>
</table>

5.2. Comparison with AT

We compared our approach with AT. We performed this part of the experiment using the UNet and the DeepResUNet, and extracted feature maps at the same locations in the network to ensure the fair implementation of different distillation methods. We visualized their output, as shown in Figure 3. It can be seen intuitively that CD corrected some errors of AT when segmenting.

Table 3 shows the comparison results. Obviously, in contrast to AT, our distillation method CD improved results in every metric. It is worth mentioning that the Dice score of CD in TC improved from 85.66 to 87.34, and the Hausdorff distance of CD in TC improved from 1.64 to 1.55.

Table 3. Comparison with AT on the BraTS2018.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice Score (%)</th>
<th>Hausdorff Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ET</td>
<td>WT</td>
</tr>
<tr>
<td>T: UNet</td>
<td>78.83</td>
<td>85.69</td>
</tr>
<tr>
<td>S: DeepResUNet</td>
<td>78.44</td>
<td>85.19</td>
</tr>
<tr>
<td>S+AT</td>
<td>78.01</td>
<td>84.95</td>
</tr>
<tr>
<td>S+CD (ours)</td>
<td><strong>78.83</strong></td>
<td><strong>85.44</strong></td>
</tr>
</tbody>
</table>
5.3. Ablation Experiments

As the final part of the experiment, to better verify the effectiveness of CD, we performed ablation experiments on the BraTS2018 dataset. Table 4 lists the performance of the student with and without KD and CD when segmented. The results showed that each method brought positive effects to the student, especially when the two methods were combined; they enabled the student network to obtain better segmentation results. In addition, in Table 5, we also demonstrated the sensitivity of our method to hyperparameters, given random weights $a, \beta_1, \beta_2$ of Equation (6). Within a certain range, each experiment was trained by 20 epochs with different combinations of hyperparameters, and we finally chose 0.9, 0.1 and 1.8 to complete all our experiments.

Table 4. Ablation experiments of our method on the BraTS2018 dataset. The teacher was UNet and the student was DeepResUNet.

<table>
<thead>
<tr>
<th>Models</th>
<th>KD</th>
<th>CD</th>
<th>Dice Score (%)</th>
<th>Hausdorff Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ET</td>
<td>WT</td>
</tr>
<tr>
<td>T: UNet</td>
<td>×</td>
<td>×</td>
<td>78.83</td>
<td>85.69</td>
</tr>
<tr>
<td>S: DeepResUNet</td>
<td>×</td>
<td>×</td>
<td>78.44</td>
<td>85.19</td>
</tr>
<tr>
<td>S+ours</td>
<td>✓</td>
<td>×</td>
<td>79.03</td>
<td>85.73</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>✓</td>
<td>79.69</td>
<td>85.91</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 3. The segmentation results were obtained by applying our distillation method in four cases (a–d) of the BraTS validation dataset. ET—yellow, TC—yellow+red, WT—yellow+red+green. The teacher was UNet and the student was DeepResUNet.
6. Discussion

In this work, we proposed CD, which is a rare knowledge distillation method for brain tumor segmentation. One of the reasons why this method worked is that CD embeds location information into channel attention, and, hence, student networks can better solve the problem of uncertain tumor location. In [32], the author proposed a knowledge distillation method to solve the boundary ambiguity problem of medical image segmentation. This method used the ground truth to extract regional information by class from the feature map and calculated the regional contrast value by measuring the similarities between different categories of regions. However, besides the information on ground truth, the rest of the information was also important. Compared to this distillation method, our method obtained global information better and would be easier to use and understand.

Table 1 contains some interesting experimental results, the segmentation metrics of every student network were improved after using our method, and almost every student network outperformed the segmentation metrics of the teacher network after using our distillation method: this is also a rare phenomenon in [32]. Our knowledge distillation architecture is a combination of KD and CD. Similarly, more knowledge distillation modules could be added to this architecture. The current segmentation was based on 2D brain tumor images, and we believe that, with a slight extension, it could be applied to 3D brain tumor images as well. 3D images have rich spatial information, and, if using our method, the student network could obtain more spatial information.

7. Conclusions

In this article, the CD distillation method was proposed for brain tumor segmentation. A series of experiments were conducted on the BraTs2018 dataset. In future work, we will further explore the most appropriate combination of teacher and student networks. In addition, in some related studies, such as dataset distillation, we will apply the proposed knowledge distillation method to dataset distillation to address the problem of small datasets in medical images.

Author Contributions: Conceptualization, Y.Q., W.Z. and S.H.; methodology, Y.Q., W.Z. and X.W.; validation, Y.Q. and X.Y.; formal analysis, J.C. and S.H.; investigation, Y.Q. and X.Y.; writing—original draft preparation, Y.Q. and X.Y.; writing—review and editing, Y.Q., W.Z., J.C. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation of China (NSFC) 62006107 and by the Natural Science Foundation of Shandong Province (No. ZR2020MF058, ZR2020MF029).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.

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

References


