

Article

SL: Stable Learning in Source-Free Domain Adaptation for Medical Image Segmentation

Yan Wang ^{1,2,3}, Yixin Chen ⁴, Tingyang Yang ⁵ and Haogang Zhu ^{2,3,6,*}

¹ School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China; wangyan9509@buaa.edu.cn

² Key Laboratory of Data Science and Intelligent Computing and Zhongfa Aviation Institute, Beihang University, 166 Shuanghongqiao Street, Pingyao Town, Yuhang District, Hangzhou 311115, China

³ State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

⁴ Institute of Medical Technology, Peking University Health Science Center, Peking University, Beijing 100191, China; 2311110791@stu.pku.edu.cn

⁵ CNGC Institute of Computer and Electronics Application, Beijing 100089, China; tyang@buaa.edu.cn

⁶ Zhongguancun Laboratory, Beijing 100194, China

* Correspondence: haogangzhu@buaa.edu.cn

Abstract: Deep learning techniques for medical image analysis often encounter domain shifts between source and target data. Most existing approaches focus on unsupervised domain adaptation (UDA). However, in practical applications, many source domain data are often inaccessible due to issues such as privacy concerns. For instance, data from different hospitals exhibit domain shifts due to equipment discrepancies, and data from both domains cannot be accessed simultaneously because of privacy issues. This challenge, known as source-free UDA, limits the effectiveness of previous UDA medical methods. Despite the introduction of various medical source-free unsupervised domain adaptation (MSFUDA) methods, they tend to suffer from an over-fitting problem described as “longer training, worse performance”. To address this issue, we proposed the Stable Learning (SL) strategy. SL is a method that can be integrated with other approaches and consists of weight consolidation and entropy increase. Weight consolidation helps retain domain-invariant knowledge, while entropy increase prevents over-learning. We validated our strategy through experiments on three MSFUDA methods and two public datasets. For the abdominal dataset, the application of the SL strategy enables the MSFUDA method to effectively address the domain shift issue. This results in an improvement in the Dice coefficient from 0.5167 to 0.7006 for the adaptation from CT to MRI, and from 0.6474 to 0.7188 for the adaptation from MRI to CT. The same improvement is observed with the cardiac dataset. Additionally, we conducted ablation studies on the two involved modules, and the results demonstrated the effectiveness of the SL strategy.

Keywords: source-free unsupervised domain adaptation; training stability



Citation: Wang, Y.; Chen, Y.; Yang, T.; Zhu, H. SL: Stable Learning in Source-Free Domain Adaptation for Medical Image Segmentation. *Electronics* **2024**, *13*, 2878. <https://doi.org/10.3390/electronics13142878>

Academic Editor: Gemma Piella

Received: 5 June 2024

Revised: 13 July 2024

Accepted: 18 July 2024

Published: 22 July 2024



Copyright: © 2024 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

In recent years, deep convolutional neural networks (DCNNs) have emerged as highly effective tools [1–4], particularly in the field of medical image analysis [5–7]. However, the performance of deep networks tends to degrade when there are distribution differences between the training and test data. This is a common challenge in medical image processing, where data from different hospitals or cross-modalities may exhibit domain shifts. Additionally, the reliance of deep learning methods on costly manual annotations further complicates the widespread application of DCNNs.

To tackle these challenges, researchers have focused on designing unsupervised domain adaptation methods that transfer knowledge from labeled source domains to unlabeled target domains [8–11]. Consequently, unsupervised domain adaptation tasks have garnered significant attention, leading to the development of various methods [9,12–14].

However, most existing approaches still rely on source data, neglecting issues of data privacy and transmission in realistic application scenarios. In practical clinical scenarios, there are many other situations where source domain data cannot be obtained. For example, limitations in data transfer speed and resources between two hospitals often prevent data collected in Hospital A from being accessed by Hospital B. Additionally, data privacy is indeed a frequently considered factor in such scenarios. In addition to protecting patient privacy, it is also essential to preserve the value of imaging data, as many imaging datasets have significant analytical and educational value, especially those involving rare diseases. Even after de-identification, these images still constitute important information and are generally not allowed to be shared outside the institution. Taking the absence of source data into consideration, a more challenging variant has emerged: source-free unsupervised domain adaptation (SFUDA), where adaptation is performed without access to source data.

In SFUDA, our access is limited to a well-trained model from the source data. Previous works in this area can be broadly categorized into two perspectives: GAN-based methods [8,9,15] and non-GAN-based methods [16,17]. However, adversarial training has certain limitations. The training process is unstable and complex, leading to uncertainties, and clinicians often exhibit skepticism towards synthetic data. On the other hand, non-GAN-based methods employ a self-training approach to adapt the source model to the target domain [18–20]. While most previous works using self-training have successfully addressed the domain shift between the source and target domains without access to source data and target labels, there are still issues arising from the unstable self-training process. Experimental results demonstrate that existing SFUDA methods eventually face a dilemma: “longer training, worse performance”. For example, in [18], it is reported that they trained for only two epochs during the self-training stage. However, in the absence of labeled data from the target domain, it becomes challenging to determine the optimal number of self-training epochs. The performance of existing SFUDA methods [18–20] significantly deteriorates as the number of epochs increases.

In this paper, we conduct an analysis of the factors contributing to the dilemma of “longer training, worse performance”. Based on our analysis, we propose the Stable Learning (SL) framework as a solution to stabilize the self-training process. The SL framework consists of two main components: weight consolidation and entropy increase.

Weight consolidation. Datasets for domain adaptation tasks can be categorized into domain-specific and domain-invariant subsets. The domain-invariant knowledge acquired by the model can be applied to both the source and target domains. Conversely, domain-specific knowledge leads to a decline in the performance of the source model when applied to the target domain. Therefore, the self-training process in the target domain aims to eliminate the source domain-specific knowledge from the source model while preserving the domain-invariant knowledge to the greatest extent possible. Neural networks encode knowledge through trainable parameters. Consequently, preserving the domain-invariant knowledge is equivalent to selectively updating only certain parameters in the model. To address this, we propose weight consolidation, which encourages the model to update specific parameters selectively.

Entropy increase. The cause of “longer training, worse performance” can be described as over-learning of the samples. Therefore, we devised a strategy to reduce the impact of the samples that the model has learned and thus avoid over-learning. In Section 3.3, we prove that this idea is equivalent to entropy increase.

In summary, we propose the Stable Learning (SL) strategy for source-free domain adaptation in medical semantic segmentation, aiming to mitigate the over-fitting issue encountered by existing SFUDA methods due to the absence of labeled target domain data. The SL strategy consists of two components: weight consolidation and entropy increase. It can be integrated with various existing medical SFUDA methods. We have conducted experiments on four SFUDA tasks using two publicly available datasets to demonstrate the effectiveness of our approach.

2. Related Work

2.1. Source-Free Unsupervised Domain Adaptation

In recent years, there has been growing interest in source-free unsupervised domain adaptation (SFUDA) due to privacy and transmission concerns in realistic application scenarios [17,19,21]. In some of the earliest studies, source-free UDA (unsupervised domain adaptation) was also called source-relaxed UDA, such as the method proposed by Mathilde Bateson et al. [22]. The authors leverage entropy minimization and a class ratio prior to relaxing the need for a concurrent access to the source and target data. Unlike general unsupervised domain adaptation, SFUDA involves adapting a well-trained model without access to the source data. Solutions to source-free problems typically fall into two perspectives. First, they focus on aligning the source domain features with the target domain features. Second, they employ a self-training process to generate pseudo-labels and perform label training. Some methods solely focus on the former perspective [20], while others concentrate on the latter [16,18,19]. There are also approaches that incorporate both perspectives [17,19]. Additionally, in SFUDA, another approach involves generating and reconstructing source domain samples while aligning features using adversarial techniques [23,24]. However, since our proposed method addresses the instability of the self-training process, we will not delve too deeply into the generative models.

In their work, the authors of [19] focus on generating pseudo-labels for target data in the segmentation task. They propose a Label-Denoising (LD) framework that incorporates positive learning and negative learning. Positive learning addresses the issue of class imbalance by using an intra-class threshold. The pseudo-labels are then used to train the source model in a supervised manner. Similarly, ref. [18] introduces Denoised Pseudo-Labeling (DPL) to create low-noise pseudo-labels. They employ a prototype strategy to calibrate the noise in the pseudo-labels, similar to the approach used in [16]. On the other hand, the method named as Off-the-shelf Source (OS) [20] focuses on domain-wise alignment by adapting the batch normalization layer. They consider low-order batch statistics, such as mean and variance, as domain-specific, while high-order batch parameters like γ and β are considered domain-invariant. The method named as Domain Adaptive Semantic segmentation (DAS) [19] freezes the classifier parameters to minimize self-entropy and align the source domain features with the target domain features. They then use the aligned source model to generate pseudo-labels, employing the intra-class threshold technique introduced in [19]. We compare our method with LD [19], DPL [18], and OS [20] in our experiments. We believe that the DAS [19] method is basically the same as the LD [19] method and can be seen as an incremental experiment of LD, so DAS is not used as a separate comparative method.

2.2. Self-Training

Self-training is a widely used technique in semi-supervised learning, where the main idea is to train the current model using pseudo-labels generated by previous models [25,26]. This concept aligns well with the task settings of unsupervised domain adaptation (UDA) and source-free unsupervised domain adaptation (SFUDA), making self-training a common solution to these problems [13,18,19,27]. However, the introduction of unlabeled data poses challenges in improving the quality of predicted pseudo-labels as much as possible [25,28]. In domain adaptation tasks, generating reliable pseudo-labels becomes even more challenging due to domain shifts. Previous works have dedicated considerable effort to addressing this issue, as seen in the aforementioned studies [18,19]. However, only a few works have focused on the stability of the training process. For instance, ref. [29] discusses the phenomenon of “lazy mimicking”, where the student model learning from pseudo-labels reaches a plateau during self-training. To mitigate this issue, the authors propose an asynchronous teacher–student optimization algorithm, which demonstrates competitive performance. Similarly, our work also tackles the stability of the training process.

3. Methods

In this section, we primarily explain the two components of our proposed Stable Learning approach. The core component is weight consolidation (WC), which aims to address the issue of over-fitting. Additionally, we utilize entropy increase (EI) to smooth the over-fitting boundaries and further enhance the stability of the self-training process.

3.1. Unstable Self-Training

In the source-free unsupervised domain adaptation (UDA) setting, we have a model $f^s : X^s \rightarrow Y^s$ trained on an inaccessible source domain $D^s = (X^s, Y^s)$, as well as an unlabeled target domain $D^t = (X^t)$. The objective of source-free UDA is to obtain an adapted model $f^{x \rightarrow y}$ that performs well on the target domain distribution. In this case, we specifically consider the task of image segmentation, which involves multi-label segmentation with $X \subset \mathbb{R}^{H \times W \times 3}$.

We observed a common dilemma when using state-of-the-art source-free UDA methods, such as LD [19] and DPL [18]. During the self-training process using target domain data to refine the source model, we encountered instability issues, leading to the model easily over-fitting. It has been noted in [18] that they trained the target model for only two epochs. Other studies have also shown that self-training reaches saturation after two epochs, with accuracy starting to decline in the third epoch [29]. In our experiments, we found that the performance of DPL and LD initially improved rapidly within the first or second epoch but then deteriorated with subsequent epochs. However, this training process is problematic, as it requires a labeled target domain validation dataset for supervision. Without such supervision, these source-free UDA approaches become ineffective.

These methods have made contributions by improving the performance of the model. However, they suffer from a critical flaw: the performance improvement is short-lived and unstable. In this paper, we propose Stable Learning (SL) to enhance the stability of self-supervised training without compromising performance.

3.2. Weight Consolidation

Weight consolidation can be defined by adding a weight penalty to the loss function:

$$L'(\theta) = L(\theta) + \sum_i |\theta_i - \theta_i^*|_1. \quad (1)$$

This formula explains the motivation for weight consolidation, in which θ represents the parameter set of the model $f^{s \rightarrow t}$, and θ^* represents the parameter set of the source model. In detail, $L'(\theta)$ is the new loss function after adding the weight penalty, $L(\theta)$ is the original adaptation loss function, θ represents the parameter set of the current model, while θ^* represents the parameter set of the source model. By introducing the L1 norm penalty term, only a subset of the model parameters are updated during training, which helps retain similarity to the source model and reduces the risk of overfitting. Additionally, $\sum_i |\theta_i - \theta_i^*|_1$ is the L1 norm penalty term, which measures the difference between the current model parameters θ_i and the source model parameters θ_i^* .

The deep convolutional network contains a large number of convolution layers and a nonlinear activation function. The training process involves adjusting the set of trainable parameters (weights, biases) of the convolution layers in order to optimize performance. Many configurations of trainable parameters will result in the same performance [30,31]. Therefore, this over-parameterization makes it possible to generate a solution that causes the model to achieve the same performance under the constraint of weight consolidation. In addition, the experimental results also prove that weight consolidation can stabilize the MSFUDA methods without affecting the accuracy. The relevant experimental results can be found in Table 1.

Table 1. Comparison of model training stability with SOTA unsupervised domain adaptation methods for abdominal dataset in Dice. The Dice value is the average of the four categories of Dice, and the four categories are liver, right kidney, left kidney, and spleen.

Abdominal CT → MRI	Epoch 1	Epoch 5	Epoch 10	Epoch 20	Epoch 50	Best Performance
Source Model	0.5167	-	-	-	-	-
LD [19]	0.6843	0.6803	0.6589	0.6146	0.5690	0.6904 (epoch 1)
fairLD	0.6861	0.6843	0.6546	0.6114	0.5578	0.6861 (epoch 1)
DPL [18]	0.6673	0.6506	0.6128	0.5918	0.5375	0.6692 (epoch 1)
OS [20]	0.6771	0.6762	0.6680	0.6796	0.6765	0.6846 (epoch 25)
fairLD with SL (ours)	0.6802	0.6918	0.6844	0.6893	0.6821	0.7006 (epoch 16)
DPL with SL (ours)	0.6768	0.6812	0.6762	0.6896	0.6866	0.6961 (epoch 36)
OS with SL (ours)	0.6667	0.6734	0.6574	0.6649	0.6673	0.6867 (epoch 17)
Abdominal MRI → CT	Epoch 1	Epoch 5	Epoch 10	Epoch 20	Epoch 50	Best Performance
Source Model	0.6474	-	-	-	-	-
LD [19]	0.5374	0.2141	0.1845	0.1717	0.1624	0.5374 (epoch 1)
fairLD	0.6936	0.7034	0.6846	0.6777	0.6532	0.7061 (epoch 5)
DPL [18]	0.6534	0.6364	0.6156	0.6076	0.6097	0.6534 (epoch 1)
OS [20]	0.6460	0.6580	0.6220	0.6251	0.5835	0.6580 (epoch 4)
fairLD with SL (ours)	0.6437	0.6766	0.7025	0.7068	0.7128	0.7188 (epoch 47)
DPL with SL (ours)	0.6484	0.6361	0.6484	0.6426	0.6403	0.6535 (epoch 17)
OS with SL (ours)	0.6397	0.6217	0.6429	0.6317	0.6304	0.6606 (epoch 23)

3.3. Entropy Increase

The primary objective of weight consolidation is to mitigate the forgetting of domain-invariant knowledge by the model and weight consolidation also helps prevent overfitting. However, it is important to note that weight consolidation alone does not entirely eliminate all instances of overfitting.

As mentioned earlier, it has been noted that “Many configurations of trainable parameters will result in the same performance” [30,31]. This statement also implies that many configurations of trainable parameters can lead to the same overfitting dilemma. Therefore, the over-parameterization of the model allows for the possibility of encountering the same overfitting dilemma despite the constraint of weight consolidation.

To mitigate this possibility, we introduce entropy increase (EI) as a means to prevent overfitting during the self-training process. The overfitting phenomenon can be attributed to excessive learning from the available samples. Hence, we have devised a strategy to reduce the influence of learned samples on the model, thereby avoiding over-learning. Following the approach in [18,19], we employ cross-entropy loss as the adaptation loss.

The adaptation loss, denoted as L_a , is defined as follows:

$$L_a = -\hat{y} \log p, \quad (2)$$

where $\hat{y}$ represents the pseudo-label created using various methods, and p represents the softmax prediction value of the model.

The concept of “learned samples” refers to the samples for which the model exhibits high confidence. To address this, the entropy increase (EI) strategy aims to reduce the loss weight assigned to high-confidence samples. The adaptation loss with the EI strategy can be decomposed into two terms: the cross-entropy adaptation loss and the self-entropy term. It is formulated as follows:

$$L_a = -(\hat{y} - p) \log p = -\hat{y} \log p + p \log p. \quad (3)$$

Thus, the EI strategy aims to maximize the self-entropy of the model’s predictions.

This conclusion contradicts the findings of previous studies. Entropy minimization has been demonstrated to be effective in semi-supervised learning [32,33], and some multi-source domain adaptation (MSFDA) methods also incorporate entropy minimization loss

as part of their approach [19,20]. We believe that both self-entropy and cross-entropy are avenues for entropy minimization, which aids in the model's convergence. However, this can also lead to the model falling into the overfitting dilemma, resulting in an unstable training process. In semi-supervised tasks, having a labeled validation set allows the model to select the optimal performance moment during training. In the case of source-free unsupervised domain adaptation (SFUDA), where labeled target domain data are unavailable, a stable training process becomes a necessary prerequisite. In the experimental section, we conduct ablation experiments to compare the effects of entropy minimization and entropy maximization on both model stability and performance.

4. Experiments

Dataset and Metrics

We evaluated our proposed method on two public datasets that are used in other domain adaptation research [8,9,34,35]. The abdominal dataset is made up of two groups of data: 20 MRI scans from the CHAOS challenge [36] and 30 CT scans from Multi-Atlas Labeling Beyond the Cranial Vault Workshop and Challenge [37]. Each MRI scan is a 3D volume of $256 \times 256 \times L$ voxels, where L is the length of the long axis. Each CT scan is a $512 \times 512 \times L$ 3D volume. For the cardiac dataset, there are unpaired 20 MRI and 20 CT 3D images with golden standard segmentation labels from the Multi-Modality Whole Heart Segmentation Challenge 2017 dataset [38]. All CT data cover the whole heart from the upper abdominal to the aortic arch, and the slices were acquired in the axial view. The in-plane resolution is about 0.78×0.78 mm, and the average slice thickness is 1.60 mm. The MRI data were acquired using 3D balanced steady-state free precession (b-SSFP) sequences with about 2 mm acquisition resolution at each direction and were resampled into about 1 mm. The ground truth labels consist of the ascending aorta (AA), the left atrium blood cavity (LAC), the left ventricle blood cavity (LVC), and the myocardium of the left ventricle (MYO). According to the setting of [8], we divide the training set and the validation set according to the ratio of 4:1. The ground truth masks are annotated as liver, right kidney (R-Kid), left kidney (L-Kid), and spleen. We proceeded with two domain adaptation tasks, from CT to MRI and from MRI to CT. The Dice coefficient is the basic metric used in segmentation tasks, and we also employed it to evaluate the performance. Dice is the measurement of volume overlap between the predictions and the ground truth annotations in 3D.

5. Implementation Details

5.1. Training Configuration

The configuration of the source model follows the approach described in papers [18,19], and we utilize the deeplabv3-resnet50 segmentation model. To enhance the model's generalization ability, we employ the following image augmentation strategies: Blur, ShiftScaleRotate, RandomBrightnessContrast, and RandomGridShuffle. During training, we set the batch size to 4 and utilize the Adam optimizer with a learning rate of 3×10^{-5} and a weight decay of 3×10^{-5} .

5.2. Fairness Optimization

We found that the pseudo-labels generated by the LD method [19] depend on the dataset. In our experiments, it was found that the pseudo-labels generated by LD [19] in the abdominal data contain a lot of noise, which affects the performance of self-training. We believe that this impact is due to the poor adaptability between the dataset and the method, not a problem with the method itself. Therefore, the comparison under this premise is unfair and ineffective. We extend and improve the algorithm of generating pseudo-labels using the LD [19] method, which is called Double Threshold Pseudo-Label (DTPL).

The LD [19] method uses intra-class confidence to select the pixels as pseudo-labels, which can avoid the imbalanced selection and "winner-takes-all" dilemma (the model

would be biased towards the majority classes and ignore the minority classes). They select the pixels with high intra-class confidence. The intra-class threshold is defined as follows:

$$\delta^{(c)} = \tau_{\alpha}(p_t^{(c)}), \quad (4)$$

where $p_t^{(c)}$ represents the prediction softmax values with respect to category c , and τ_{α} means the top α (%) value. Therefore, for each category, we select the labels whose softmax value is larger than threshold $\delta^{(c)}$. However, the quality of pseudo-labels depends very much on the setting of α . LD [19] uses the same α between different categories, which is set as 0.3, but this does not work for any datasets. As shown in Figure 1, we found that the LD method creates a lot of noise. Different categories use the same α as their intra-class threshold, which causes each category in the pseudo-label to have a similar area. We can avoid this problem by setting α separately for each category, but too many hyperparameters can bring a lot of complexity.

Experimentally, we found that the generated noise can be filtered out by a global threshold, so we use double thresholds (global threshold and intra-class threshold) to generate pseudo-labels. The Double Threshold Pseudo-Label can be defined as follows:

$$\hat{y}^{h,w,c} = \begin{cases} 1 & \text{if } c = (\arg\max_c p^{h,w,c}) \cap (p^{h,w,c} > \delta^{(c)}) \\ & \cap (p^{h,w,c} > \lambda) \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where $\hat{y}^{h,w,c}$ is the pixel-wise pseudo label, $(p^{h,w,c} > \delta^{(c)})$ represents the intra-class threshold, and $(p^{h,w,c} > \lambda)$ means the global threshold. In our research, the α is set to 0.3, and λ is set to 0.2.

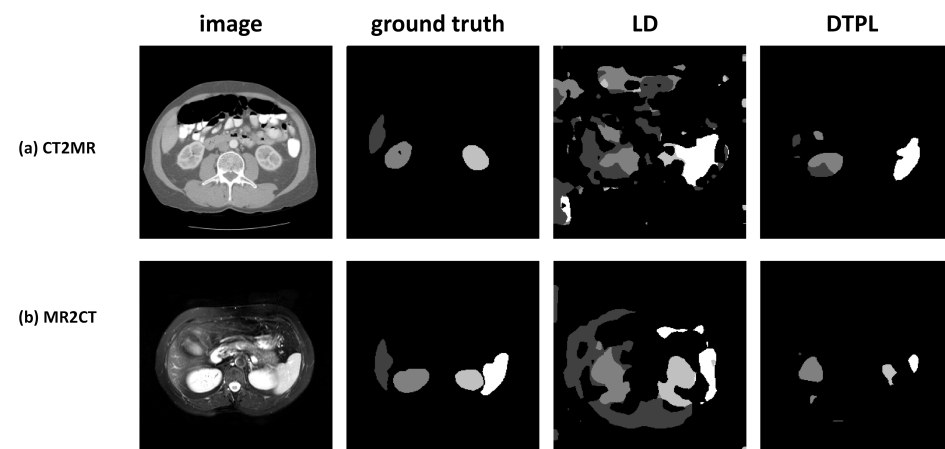


Figure 1. Comparison of pseudo-labels in LD [19] and DTPL in the abdominal dataset. The four columns from left to right represent images, annotations, pseudo-labels generated by LD, and pseudo-labels generated by DTPL.

5.3. Abdominal Comparison Experiments

In Table 1, we compare state-of-the-art approaches with or without the Stable Learning strategy and show that approaches with SL are more stable. The table above shows the results of the adaptation task from the CT to the MRI modality, and the table below shows the results of the adaptation task from MRI to CT. We conducted an in-depth analysis of the experimental results from two perspectives. Firstly, since this is a domain adaptation task, we assessed the effectiveness of addressing the domain shift problem by examining the best performance results. As shown in Table 1, for the task from CT to MRI, directly inferring the source model results in a low Dice performance of 0.5167, and the MSFUDA methods improve the Dice score to at least 0.6861, which shows the effectiveness of domain adaptation methods. Among these improvements, the best is fairLD, with our Stable

Learning strategy. Additionally, all methods show improved final adaptation performance when combined with our strategies. Secondly, we focus on the epoch at which each method achieves its best results during training. We compared snapshots of different methods at the 1st, 5th, 10th, 20th, and 50th epochs and found that the MSFUDA methods [18–20] without an SL policy would suffer a “longer training, worse performance” dilemma. For example, in task $MRI \rightarrow CT$, DPL performance at the 50th epoch is 0.6097, but the best performance of DPL is 0.6580 at the 1st epoch. Similarly, the performance of OS [20] is 0.5835 at the 50th epoch, but its best performance is 0.6580 at the 4th epoch. The methods with SL outperform those without SL. The experimental data in Table 1 can fully prove this idea.

In addition, we also compare the training time for the different MSFUDA methods on the adaptation from CT to MRI on the abdominal dataset. Because we aim to address the issue of instability during training and need validation set labels to determine the best performance, if we need to compare the time, it would be best to compare the time at which the best performance occurs, i.e., the complete training time. The results are shown in Table 2. Some methods become slower after adding the SL strategy because the number of epochs required for training increases, such as with DPL. However, this is actually more reasonable in terms of experimental results, as our focus is on stable training. In contrast, methods without the SL strategy tend to show a sudden deterioration in performance with continued training. As shown in Figure 2, we also show the segmentation visualization results of two CT samples, showing the best segmentation performance of fairLD with Stable Learning for the MRI to CT adaptation on the abdominal dataset.

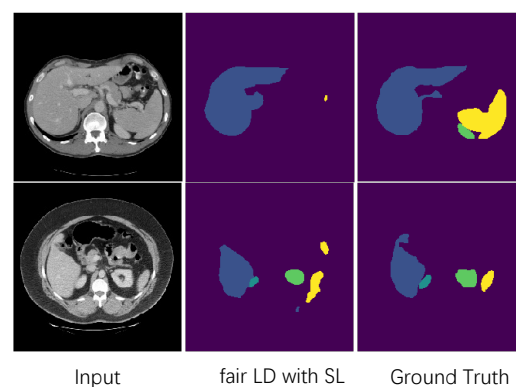


Figure 2. Visualization results for the segmentation of CT images on the abdominal dataset. The structure of the liver, right kidney, left kidney, and spleen are shown in blue, dark green, light green, and yellow, respectively.

Table 2. Training time for each method.

Method	Training Time (Hours)
LD	1.2
fairLD	1.5
DPL	3.2
OS	28.4
fairLD with SL	18.6
DPL with SL	90.5
OS with SL	20.2

5.4. Cardiac Comparison Experiments

In Table 3, we compare state-of-the-art methods with or without our Stable Learning strategy. Similarly to the analysis of the abdominal results, we first focus on the adaptation’s Dice performance. For the adaptation from CT to MRI, the source model yields a performance of 0.3753. All the methods improve this performance to at least 0.4089, which represents a 9.0% improvement over the baseline. For the adaptation from MRI to

CT in particular, there is a specific case worth mentioning: the OS method demonstrates significant instability during training. It struggles to improve upon the source model's performance of 0.49, possibly due to differences between the datasets used by the original method. However, with the application of SL strategies, there is still an improvement based on the OS method from 0.4609 to 0.5083. Secondly, we show that most methods exhibit an unstable phenomenon. In the 50th epoch, the performance of fairLD and DPL are 0.3661 and 0.3961, respectively, which has a huge drop off from the best performance (0.4202 and 0.4238). However, with the help of the Stable Learning strategy, the performance of fairLD and DPL are 0.4136 and 0.4213 in the 50th epoch, which is consistent with the best performance (0.4264 and 0.4238, respectively).

Table 3. Comparison of model training stability for the cardiac dataset in Dice. The Dice value is the average of the four categories of Dice, which are: ascending aorta, left atrium blood cavity, left ventricle blood cavity, and myocardium of the left ventricle.

Cardiac CT → MRI	Epoch 1	Epoch 5	Epoch 10	Epoch 20	Epoch 50	Best Performance
Source Model	0.3753	-	-	-	-	-
LD [19]	0.4089	0.3874	0.3761	0.3678	0.3535	0.4089 (epoch 0)
fairLD	0.4155	0.4124	0.4022	0.3768	0.3661	0.4202 (epoch 2)
DPL [18]	0.4036	0.4216	0.4125	0.4128	0.3961	0.4238 (epoch 8)
OS [20]	0.3974	0.4141	0.4107	0.4154	0.4186	0.4261 (epoch 36)
fairLD with SL (ours)	0.4073	0.4171	0.4165	0.4101	0.4136	0.4264 (epoch 13)
DPL with SL (ours)	0.4109	0.4174	0.4151	0.4177	0.4213	0.4238 (epoch 44)
OS with SL (ours)	0.4125	0.4154	0.4195	0.4195	0.4141	0.4202 (epoch 24)
Cardiac MRI → CT	Epoch 1	Epoch 5	Epoch 10	Epoch 20	Epoch 50	Best Performance
Source Model	0.4951	-	-	-	-	-
LD [19]	0.5548	0.5280	0.5569	0.5557	0.5563	0.5990 (epoch 8)
fairLD	0.5840	0.6033	0.5985	0.5712	0.5829	0.6234 (epoch 3)
DPL [18]	0.5447	0.4795	0.4641	0.5071	0.5049	0.5447 (epoch 0)
OS [20]	0.4334	0.4479	0.4464	0.4609	0.4375	0.4609 (epoch 19)
fairLD with SL (ours)	0.5772	0.6213	0.6211	0.6119	0.6223	0.6289 (epoch 28)
DPL with SL (ours)	0.5211	0.5418	0.5591	0.5509	0.5456	0.5608 (epoch 14)
OS with SL (ours)	0.4834	0.4991	0.4867	0.4950	0.4912	0.5083 (epoch 22)

5.5. Ablation Study

The results of ablation experiments are shown in Table 4. The ablation experiment consists of four source-free domain adaptation tasks: two tasks on the abdominal dataset and the other two on the cardiac dataset. In this section, we analyze three aspects: first, whether self-entropy minimization plays a positive role in the SFUDA task; second, whether Stable Learning stabilizes the self-training; and third, the impact of entropy increase in Stable Learning.

Table 4. The results of the ablation experiments. The table above shows the experimental results on the abdominal dataset, and the table below is about the cardiac dataset. “Ab.” and “Ca.” are abbreviations for abdominal and cardiac, respectively. “Best” refers to the best performance of the model over 200 epochs. “Entropy” contains three states: “-” not used; “Min” entropy minimization; and “Max” entropy maximization.

Group	Ab. CT → MRI	Entropy	Epoch 50	Epoch 200	Best
A	fairLD	-	0.5823	0.5312	0.6832
	OS	-	0.6772	0.6751	0.6878
	DPL	-	0.5375	0.4912	0.6692
B	fairLD	Min	0.5578	0.5053	0.6861
	OS	Min	0.6765	0.6698	0.6846
	DPL	Min	0.5198	0.4789	0.6718

Table 4. Cont.

Group	Ab. $CT \rightarrow MRI$	Entropy	Epoch 50	Epoch 200	Best
C	fairLD + WC	-	0.6811	0.6679	0.7011
	OS + WC	-	0.6771	0.6698	0.6834
	DPL + WC	-	0.6858	0.6699	0.6912
D	fairLD + WC	Max	0.6821	0.6897	0.7006
	OS + WC	Max	0.6673	0.6733	0.6867
	DPL + WC	Max	0.6866	0.6881	0.6961
Group	Ab. $MRI \rightarrow CT$	Entropy	Epoch 50	Epoch 200	Best
A	fairLD	-	0.6772	0.6219	0.7101
	OS	-	0.6009	0.5722	0.6611
	DPL	-	0.6097	0.5792	0.6534
B	fairLD	Min	0.6532	0.5979	0.7061
	OS	Min	0.5835	0.5301	0.6580
	DPL	Min	0.5977	0.5521	0.6598
C	fairLD + WC	-	0.7098	0.6881	0.7159
	OS + WC	-	0.6375	0.6198	0.6639
	DPL + WC	-	0.6462	0.6219	0.6559
D	fairLD+WC	Max	0.7128	0.7072	0.7188
	OS + WC	Max	0.6304	0.6408	0.6606
	DPL + WC	Max	0.6426	0.6482	0.6535
Group	Ca. $CT \rightarrow MRI$	Entropy	Epoch 50	Epoch 200	Best
A	fairLD	-	0.3823	0.3629	0.4233
	OS	-	0.4177	0.4152	0.4234
	DPL	-	0.3961	0.3638	0.4238
B	fairLD	Min	0.3661	0.3329	0.4202
	OS	Min	0.4186	0.4159	0.4261
	DPL	Min	0.3872	0.3511	0.4244
C	fairLD + WC	-	0.4144	0.4003	0.4283
	OS + WC	-	0.4102	0.3998	0.4232
	DPL + WC	-	0.4254	0.3978	0.4225
D	fairLD + WC	Max	0.4136	0.4182	0.4264
	OS + WC	Max	0.4141	0.4164	0.4202
	DPL + WC	Max	0.4213	0.4129	0.4238
Group	Ca. $MRI \rightarrow CT$	Entropy	Epoch 50	Epoch 200	Best
A	fairLD	-	0.6021	0.5641	0.6254
	OS	-	0.4477	0.4202	0.4596
	DPL	-	0.5049	0.4593	0.5447
B	fairLD	Min	0.5829	0.5545	0.6234
	OS	Min	0.4375	0.4149	0.4609
	DPL	Min	0.4982	0.4521	0.5455
C	fairLD + WC	-	0.6218	0.6022	0.6247
	OS + WC	-	0.4899	0.4728	0.5052
	DPL + WC	-	0.5421	0.5372	0.5583
D	fairLD + WC	Max	0.6223	0.6198	0.6289
	OS + WC	Max	0.4912	0.4955	0.5083
	DPL + WC	Max	0.5456	0.5508	0.5608

The experiments of Group A and Group B in Table 4 are related to the role of self-entropy minimization in SFUDA tasks. The experimental results for Group A did not use an entropy minimization strategy, and the experimental results in Group B used it. From the best performance of the two groups, it can be found that an entropy minimization

strategy cannot improve the best performance of SFUDA methods. For example, in the abdominal $CT \rightarrow MRI$ task, the best performances of fairLD, OS, and DPL without entropy minimization are 0.6832, 0.6878, and 0.6692, respectively. They did not improve significantly after using entropy minimization (0.6861, 0.6846, and 0.6718). However, entropy minimization might aggravate the over-fitting phenomenon. The experimental results show that the performance of the method using entropy minimization at the 50th epoch and the 200th epoch are worse than those without entropy minimization. In general, the effect of self-entropy minimization in SFUDA tasks does not provide a significant positive impact.

The experiments of Group D in Table 4 are related to the role of Stable Learning. For the effectiveness and stability of SL in Table 1, we have considered such a possibility: by slowing down the convergence speed of the model, the process of self-training is stabilized. If this assumption holds, it means that if the training epoch is prolonged, the final model will still fall into the dilemma of over-fitting. We extended the training epoch from 50 to 200 and found that SL can indeed stabilize the training process, not delay the convergence speed. In Group D, the performance of fairLD with SL changed from 0.6821 (50th epoch) to 0.6897 (200th epoch) in the abdominal $CT \rightarrow MRI$ task. Similarly, the performance of fairLD with SL changed from 0.6223 (50th epoch) to 0.6198 (200th epoch) in the cardiac $MRI \rightarrow CT$ task. In general, the effect of Stable Learning can indeed stabilize the model and prevent the model from falling into the over-fitting dilemma.

The experiments of Group C and Group D in the Table 4 are related to the role of entropy increase in Stable Learning. The experimental results in Group C did not use an entropy increase strategy, and the experimental results in Group D used it. Weight consolidation was used in both groups. From the 200th epoch of the two groups, it can be found that the method with the WC strategy only had a decline in performance after a long training time. For example, in the abdominal $CT \rightarrow MRI$ task, the performance of fairLD (Group C) suffered a decline from 0.6811 (50th epoch) to 0.6679 (200th epoch); that of DPL also had a drop from 0.6858 (50th epoch) to 0.6699 (200th epoch). However, the performance of fairLD and DPL (Group D) is more stable, which is from 0.6821, 0.6866 (50th epoch) to 0.6897, 0.6881 (200th epoch), respectively. The reason for this phenomenon is that over-parameterization makes it possible to have a solution that makes the model suffer from the same over-fitting dilemma under the constraint of weight consolidation (more details can be found in Section 3.3).

6. Conclusions

We introduce the novel Stable Learning (SL) strategy to address the challenges of source-free domain adaptation in medical semantic segmentation. Our approach specifically tackles the overfitting dilemma that current source-free unsupervised domain adaptation (SFUDA) methods encounter due to the lack of a labeled target domain validation dataset. The SL strategy integrates two essential components, weight consolidation and entropy increase, which work together to enhance model stability and performance. Weight consolidation adds a weight penalty to the loss function, allowing the model to retain knowledge of source domains. Entropy increase aims to maximize the self-entropy of the model's predictions, which is proposed to avoid over-learning. It is designed to be compatible with various existing medical SFUDA methods. We combined this training strategy with other methods and validated its effectiveness on two public datasets.

Author Contributions: Y.W.: conceptualization, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; Y.C.: conceptualization, investigation, methodology, validation, writing—original draft; T.Y.: investigation, software, validation, visualization, writing—original draft, writing—review and editing; H.Z.: conceptualization, data curation, project administration, software, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China under Grant No. 2021ZD0140407, the National Natural Science Foundation of China under

Grants No. U21A20523, No. L222152, and No. 61971017, and the Beijing Natural Science Foundation (72443251, L222152).

Data Availability Statement: The abdominal dataset is publicly available, and the link addresses are <https://chaos.grand-challenge.org/> and <https://www.synapse.org/Synapse:syn3193805/wiki/89480> accessed on 11 August 2021. The public cardiac dataset is also available, and the link is <https://zmiclab.github.io/zxh/0/mmwhs/>, accessed on 12 September 2021.

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

References

- Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
- 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: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
- 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, 27–30 June 2016; pp. 770–778.
- Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [[CrossRef](#)] [[PubMed](#)]
- Reiss, S.; Seibold, C.; Freytag, A.; Rodner, E.; Stiefelhagen, R. Every Annotation Counts: Multi-Label Deep Supervision for Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 9532–9542.
- He, Y.; Yang, D.; Roth, H.; Zhao, C.; Xu, D. DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 5841–5850.
- Chang, Q.; Qu, H.; Zhang, Y.; Sabuncu, M.; Chen, C.; Zhang, T.; Metaxas, D.N. Synthetic Learning: Learn From Distributed Asynchronous Discriminator GAN Without Sharing Medical Image Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
- Chen, C.; Dou, Q.; Chen, H.; Qin, J.; Heng, P.A. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 865–872.
- Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 1989–1998.
- Dou, Q.; Ouyang, C.; Chen, C.; Chen, H.; Heng, P.A. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv* **2018**, arXiv:1804.10916.
- Chen, C.; Dou, Q.; Chen, H.; Qin, J.; Heng, P.A. Unsupervised Bidirectional Cross-Modality Adaptation via Deeply Synergistic Image and Feature Alignment for Medical Image Segmentation. *arXiv* **2020**, arXiv:2002.02255.
- Zhang, Q.; Zhang, J.; Liu, W.; Tao, D. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *arXiv* **2019**, arXiv:1910.13049.
- Zou, Y.; Yu, Z.; Kumar, B.; Wang, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 289–305.
- Mei, K.; Zhu, C.; Zou, J.; Zhang, S. Instance adaptive self-training for unsupervised domain adaptation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 415–430.
- Yang, S.; Wang, Y.; van de Weijer, J.; Herranz, L.; Jui, S. Unsupervised domain adaptation without source data by casting a bait. *arXiv* **2020**, arXiv:2010.12427.
- Kim, Y.; Cho, D.; Han, K.; Panda, P.; Hong, S. Domain adaptation without source data. *IEEE Trans. Artif. Intell.* **2021**, *2*, 508–518. [[CrossRef](#)]
- Liang, J.; Hu, D.; Feng, J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In Proceedings of the International Conference on Machine Learning, PMLR, Vienna, Austria, 13–18 July 2020; pp. 6028–6039.
- Chen, C.; Liu, Q.; Jin, Y.; Dou, Q.; Heng, P.A. Source-Free Domain Adaptive Fundus Image Segmentation with Denoised Pseudo-Labeling. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2021*; Springer: Cham, Switzerland, 2021; pp. 225–235.
- You, F.; Li, J.; Zhu, L.; Chen, Z.; Huang, Z. Domain Adaptive Semantic Segmentation without Source Data. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 3293–3302.
- Liu, X.; Xing, F.; Yang, C.; El Fakhri, G.; Woo, J. Adapting off-the-shelf source segmenter for target medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2021*; Springer: Cham, Switzerland, 2021; pp. 549–559.

21. Kundu, J.N.; Venkat, N.; Babu, R.V. Universal source-free domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4544–4553.
22. Bateson, M.; Kervadec, H.; Dolz, J.; Lombaert, H.; Ben Ayed, I. Source-relaxed domain adaptation for image segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I* 23; Springer: Berlin/Heidelberg, Germany, 2020; pp. 490–499.
23. Kurmi, V.K.; Subramanian, V.K.; Namboodiri, V.P. Domain impression: A source data free domain adaptation method. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 615–625.
24. Xia, H.; Zhao, H.; Ding, Z. Adaptive Adversarial Network for Source-Free Domain Adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9010–9019.
25. 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**, 1195–1204.
26. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Proceedings of the Workshop on Challenges in Representation Learning, ICML, Atlanta, GA, USA, 16–21 June 2013; Volume 3, p. 896.
27. Pan, F.; Shin, I.; Rameau, F.; Lee, S.; Kweon, I.S. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3764–3773.
28. Yang, C.; Xie, L.; Qiao, S.; Yuille, A.L. Training deep neural networks in generations: A more tolerant teacher educates better students. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5628–5635.
29. Huo, X.; Xie, L.; He, J.; Yang, Z.; Zhou, W.; Li, H.; Tian, Q. ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1235–1244.
30. Hecht-Nielsen, R. Theory of the backpropagation neural network. In *Neural Networks for Perception*; Elsevier: Amsterdam, The Netherlands, 1992; pp. 65–93.
31. Sussmann, H.J. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Netw.* **1992**, 5, 589–593. [\[CrossRef\]](#)
32. Grandvalet, Y.; Bengio, Y. Semi-supervised Learning by Entropy Minimization. In *Advances in Neural Information Processing Systems*; Saul, L., Weiss, Y., Bottou, L., Eds.; MIT Press: Cambridge, MA, USA, 2005; Volume 17.
33. Springenberg, J.T. Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. *arXiv* **2016**, arXiv:1511.06390.
34. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
35. Tomar, D.; Lortkipanidze, M.; Vray, G.; Bozorgtabar, B.; Thiran, J.P. Self-attentive spatial adaptive normalization for cross-modality domain adaptation. *IEEE Trans. Med. Imaging* **2021**, 40, 2926–2938. [\[CrossRef\]](#) [\[PubMed\]](#)
36. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; Klein, A. Multi-Atlas Labeling beyond the Cranial Vault. 2015. Available online: <https://www.synapse.org/Synapse:syn3193805/wiki/89480> (accessed on 11 August 2021).
38. Zhuang, X.; Shen, J. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Med. Image Anal.* **2016**, 31, 77–87. [\[CrossRef\]](#) [\[PubMed\]](#)

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.