ADAPTIVE PSEUDO LABELING FOR SOURCE-FREE DOMAIN ADAPTATION IN MEDICAL IMAGE SEGMENTATION

Chen Li[†], *Wei Chen*[†✉], *Xin Luo, Yulin He, Yusong Tan*

College of Computer, National University of Defense Technology, Changsha, China

ABSTRACT

Domain adaptation is common but challenging in signal processing tasks due to the intrinsic discrepancy, especially in difficult-to-label medical image segmentation application scenarios. Pseudo labeling methods are widely utilized to compensate for the scarcity of annotation. However, most existing methods set the fixed thresholds to select highly-confident predictions as pseudo labels, inevitably generating false labels with noise. In this paper, we combine the dual-classifiers consistency and predictive category-aware confidence to form a novel regularization for pseudo-label denoising. The dual-classifiers consistency helps promote the robustness of pseudo labels. Meanwhile, category-aware confidence is utilized as adaptive pixel-wise weights, avoiding the need for handcrafted thresholds. The adapted model is refined by the rectified pseudo labels without source domain samples. The proposed method is model-independent and thus can be plug-and-play to improve existing UDA methods. We validated it on the cross-modality medical image segmentation and obtained more competitive results.

***Index Terms*—** Unsupervised domain adaptation, Source-free, Adaptive pseudo labeling, Medical image segmentation.

1. INTRODUCTION

Deep learning-based methods have brought remarkable advances in the signal processing field, such as image analysis [1, 2, 3, 4]. Supervised learning is one of the most accomplished research, relying on massive annotated data to provide direct supervision for feature extraction and reconstruction. This mechanism is beneficial to promote performance but limits the generalization ability of the networks at the same time. Intuitively, the well-trained networks are prone to suffer from performance degradation when transferred from source domain to target domain. This phenomenon is caused by the feature distribution gap between different domains, which is called domain shift. Domain shift exists in cross-modality image analysis tasks, such as learning the domain-shared structural features from Magnetic Resonance Imaging dataset (MRI) and then segmenting multi-organs from Computed Tomography (CT). When applying the supervised training method to domain generalization, it is challenging to collect sufficient pixel-wise annotations due to privacy protection or unaffordable expense. Meanwhile, it is hard to utilize conventional methods to get promising performance without any target domain labels [5]. In this case, unsupervised domain adaptation (UDA) [6] becomes a preferring choice.

Related work: Most UDA methods try to align domains by minimizing the discrepancy between source domain and target domain in the feature space, where annotations of target samples are absent for training. In collaboration with the above process, pseudo labeling is the self-training method to generate pseudo labels based on the learned domain-invariant representation. Then the pseudo labels are utilized as annotations to fine-tune the adapted model in return. Lee et al. [7] drew on the proposed above idea and leveraged the pseudo labels of unlabeled samples to semi-supervised train the deep neural networks. Zou et al. [8] proposed the CBST and first introduced the idea of pseudo labeling into domain adaptation. Through iterative optimization, the knowledge extracted from source domain is distilled to target domain and achieve domain adaptation. Zou et al. [9] adapted the pseudo labeling method to the semantic segmentation and obtained soft confident results by regularization. Zheng et al. [10] proposed the uncertainty estimation method to calculate the variance between predictions and extract domain-shared features.

Most threshold-based methods generated confident pseudo labels through handcrafted thresholds. However, they ignored that there was still noise inside the prediction. Both noise and semantic information in the prediction were treated equally, which compromised the subsequent training. Meanwhile, the optimal handcrafted thresholds are different for different tasks. Other methods still need to utilize source domain samples for regularization. So it was prone to overfit the source domain-specific features during pseudo labeling. Besides, previous works failed to take advantage of the complementary relationship between confidence and consistency.

This paper focuses on the unsupervised source-free domain adaptation and proposes an adaptive pseudo labeling method for cross-modality medical images segmentation. Specifically, we combine the dual-classifiers consistency and predictive category-aware confidence to form a novel reg-

[†]These authors contributed equally to this work. This research was funded by the National Key Research and Development Program of China (No. 2018YFB0204301).

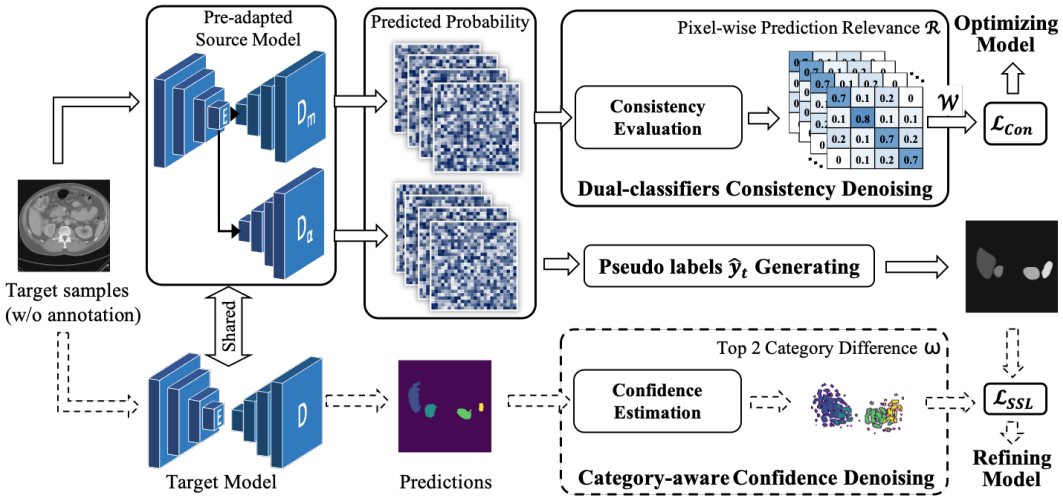

Fig. 1. **An overview of our proposed method.** Generated from the pre-adapted model, pseudo labels are then denoised with the dual-classifiers consistency. Category-aware confidence is utilized to refine the target model with rectified pseudo labels.

ularization for pseudo-label denoising. The dual-classifiers consistency helps promote the robustness of pseudo labels, which is calculated through the weighted relevance matrix. Furthermore, the category-aware confidence is utilized as adaptive pixel-wise weight to avoid the need for handcrafted thresholds, which is obtained by calculating the probability difference of the top two classes. After that, the pseudo labels are optimized and then used to carry out semi-supervised training, aiming to refine the adapted model and get better performance. We name the proposed **A**daptive **P**seudo **L**abeling method as **APL**. We validate it on two public clinical medical images datasets, i.e., MRI $\Rightarrow$ CT. The above two images are entirely different in appearance but essentially have shared implicit domain-invariant representation. The main contributions of this paper are listed as follows: (1) We are the first to quantify the consistency and confidence in noisy prediction sequentially and then integrate them for adaptive pseudo labels rectification; (2) The proposed APL method is model-independent and thus can be plug-and-play to promote existing UDA methods without any extra parameters or source domain samples; (3) The proposed APL method is validated on the cross-modality (MRI $\Rightarrow$ CT) medical image segmentation and achieved SOTA performance.

2. METHODOLOGY

2.1. Definition and Motivation

In the unsupervised domain adaptation, we have access to source domain X_S with annotation Y_S, where image $x_s \in X_S$ and label $y_s \in Y_S$. Besides, we also have access to target domain X_T without annotation, where image $x_t \in X_T$. It is vital that both X_S and Y_S share the same implicit semantic feature space, where two domains describe the same object with different data distributions. The goal of UDA is to construct projection function to learn the domain-invariant representation with the assistance of $\{X_S, Y_S, X_T\}$. Based on UDA, source-free domain adaptation is to distill the learned representation to target domain without source samples.

After finishing domain adaptation, the adapted model is utilized to predict the pixel-wise classification results of target samples as $\hat{Y}_T$. Existing works mostly calculated the prediction probability and then filtered the pixels with the handcrafted max probability threshold or uncertainty regularization.However, the handcrafted threshold-based methods still require manually design thresholds to separate high-confident objects, which were not an automatic learning process and neglected confident pixels. Meanwhile, the uncertainty regularization-based method ignored the inconsistency of different noisy predictions and failed to take advantage of the complementary relationship between consistency and confidence. These defects greatly limited the robustness of the methods.

For the above reasons, we investigate two critical properties for the generated pseudo labels, i.e., consistency and confidence. **Consistency** is measured by calculating the variance of predictions from different branches and evaluated as the internal reliability of the pseudo label. **Confidence** is measured by calculating the discrepancy of pseudo label and evaluated by the external reliability of the pseudo label. Higher consistency and confidence denote better prediction. For example, there are two branches that generate results (A and B) for the same input. The probability in A is [0.34,0.33,0.33] while B is [0.35,0.32,0.33]. The variance between A and B is small, and the maximum values of A and B are also small, which means these predictions are consistent in distribution but not confident. Meanwhile, there are another two branches that generate results (C and D) for the same input. The probability in C is [0.80,0.15,0.05] while D is [0.70,0.15,0.15]. The vari-

ance between C and D is large, and the maximum values of C and D are also large, which means these pseudo labels are confident but inconsistent. The above two comparisons denote that there is massive noise in the pseudo labels. Directly using these noisy labels as accurate annotations to optimize the model would compromise the subsequent training.

To alleviate the impact of noise, we propose a full-automatic adaptive pseudo-labeling method. Based on the rectified labels, the dynamic pixel-wise weights replace manually designed thresholds. Following most research in UDA, the modified DeepLabv2 with dual-branch classifiers are selected as the baseline model. After obtaining the adapted model from the existing UDA methods, we generated confident pseudo labels with low uncertainty to refine the trained model. The proposed method is orthogonal to the existing UDA methods and yields competitive performance of semantic segmentation.

2.2. Dual-Classifiers Consistency Denoising

Dual-classifiers in the baseline model are composed of decoders with two branch paths, including the main classifier $D_m(\cdot)$ and the auxiliary classifier $D_a(\cdot)$. The process of feature extraction through the encoder is denoted as $E(\cdot)$. Referring to the BCDM proposed by [11], we redesign the loss function to measure the pixel-wise consistency between the prediction results of the $D_m(E(\cdot))$ and $D_a(E(\cdot))$. Both of dual-predictions are matrixs with size $C \times 1$. Firstly, we define the weighted Dual-Classifiers Prediction Relevance $\mathcal{R}$ based on the pixel-wise softmax probabilities:

$$\mathcal{R}_{x_t} = \mathcal{W}_{x_t} \times D_m(E(x_t)) \times D_a(E(x_t))^T. \quad (1)$$

$\mathcal{W}_{x_t}$ is a weight matrix with size $C \times C$, where the value in the i-th row and j-th column is the number of pixels.

Obviously, $\mathcal{R}$ is a $C \times C$ matrix, which is used to evaluate the relevance across different classes between the dual-classifiers predictions. $\mathcal{R}^{i,j}(i, j \in \{1, 2, 3, ..., C\})$ is the element in the i-th row and j-th column of $\mathcal{R}$. The values of the diagonal element $\mathcal{R}^{i,i}(i \in \{1, 2, 3, ..., C\})$ determine the consistency of the pseudo labels. The larger the sum, the more concentrated the class distribution in the predictions and the higher the consistency. With the help of Relevance $\mathcal{R}$, the pseudo labels are denoised and consistent by maximizing the sum of the diagonal elements. Finally, we define the Dual-classifiers Consistency loss $\mathcal{L}_{Con}$ as the follows:

$$\min_{\theta_{D_m, D_a}} \mathcal{L}_{Con} = \sum_{x_t}^{X_T} [\sum_{i,j=1}^{C} \mathcal{R}_{x_t}^{i,j} - \sum_{i=1}^{C} \mathcal{R}_{x_t}^{i,i}], \quad (2)$$

2.3. Category-aware Confidence Denoising

The whole training and optimizing process of our framework is described in Algorithm 1. The confidence of each pseudo label is estimated by the difference between the probability values of the top two categories. Firstly, we merge the dual-classifier predictions and obtain the pseudo label as

Algorithm 1 Optimizing process with the proposed APL.

Input: The pre-adapted feature extractor E and classifiers D_m, D_a; The un-annotated target domain samples, X_T.
Output: Adaptive annotated pseudo labels of target domain $\hat{Y_T}$ and refined models;

for $i = 1$ to N **do** (do Semi-supervised Learning)
  Predict the target domain samples ($x_t \in X_T$);
  Obtain the predictions $D_m(E(x_t)), D_a(E(x_t))]$ from the dual-classifiers;
  Evaluate the the weighted Dual-Classifiers Prediction Relevance $\mathcal{R}$ of predictions with Eq.(1);
  Merge the predictions and obtain the pseudo label $\hat{y_t}$;
  Evaluate the Category-aware Confidence with Eq.(3);
  Optimize the models θ_{D_m, D_a} by maximizing the Dual-classifiers Consistency with Eq.(2);
  Update the ensemble of target pseudo label $\hat{y_t} \in \hat{Y_T}$;
  Refine the models θ_E by maximizing the the Category-aware Confidence with Eq.(4);
end for
return $\hat{Y_T}, E, D_m, D_a$;

$\hat{y_t} = arg\ max[D_m(E(x_t)) + D_a(E(x_t))]$. And then, the Category-aware Confidence is calculated and formulated as follows.

$$\omega = |\delta_1(\hat{y_t}) - \delta_2(\hat{y_t})|, \quad (3)$$

where ω is the pixel-wise measured confidence, $\delta_1(\hat{y_t})$ denotes the largest category probability among the pseudo label $\hat{y_t}$. In the same way, $\delta_2(\hat{y_t})$ is the second largest category probability. ℓ_1 distance is used to measure the discrepancy between above two probability distributions.

Through the proposed adaptive pseudo labeling method, we automatically obtain the confident labels with high consistency. Based on the denoised labels, we directly utilize them as the original target domain dataset's annotation and conduct Semi-Supervised Learning to refine the adapted model.

$$\min_{\theta_E} \mathcal{L}_{SSL} = \sum_{x_t \in X_T, \hat{y_t} \in \hat{Y_T}} \omega \cdot [-\hat{y_t} \log D(E(x_t))] \quad (4)$$

3. EXPERIMENTS AND RESULTS

To evaluate the effectiveness and superior of the proposed adaptive pseudo labeling method (APL), extensive experiments are performed on four different medical image segmentation tasks of diverse modalities (MRI $\Rightarrow$ CT).

3.1. Experimental settings

As for the experimental dataset, the Combined Healthy Abdominal Organ Segmentation (CHAOS)[1] and Multi-Atlas Labeling Beyond the Cranial Vault (MALBCV)[2] are chosen for cross-modality medical image segmentation. These datasets

[1] https://chaos.grand-challenge.org/
[2] https://www.synapse.org/Synapse:syn3193805/wiki/217789

Table 1. Experimental results of the promoted different domain adaptation methods with the proposed pseudo-labeling method for MRI ⇒ CT abdominal organs segmentation.

UDA Methods (Mean)	Dice (%)				
	LV	RK	LK	SP	Avg
W/o Adaptation	73.10	47.30	57.30	55.10	58.20
Supervised Train	92.80	86.40	87.40	88.20	88.70
SynSegNet [12]	85.00	82.10	72.70	81.00	80.20
SynSegNet +APL	85.19	81.26	79.24	84.41	82.53
AdaOutput [13]	85.40	79.70	79.70	81.70	81.60
AdaOutput +APL	87.65	83.85	83.88	80.77	84.04
CycleGAN [14]	83.40	79.30	79.40	77.30	79.90
CycleGAN +APL	86.79	78.54	79.24	84.62	82.30
CyCADA [15]	84.50	78.60	80.30	76.90	80.10
CyCADA +APL	89.06	81.43	83.30	81.33	83.78
ADVENT [16]	89.28	77.05	81.37	83.45	82.79
ADVENT +APL	90.37	85.10	81.48	84.82	85.44
SIFA-v1 [17]	87.90	83.70	80.10	80.50	83.10
SIFA-v1 +APL	89.56	89.74	83.66	82.31	86.32
SIFA-v2 [18]	88.00	83.30	80.90	82.60	83.70
SIFA-v2 +APL	89.56	87.88	89.19	82.80	87.11

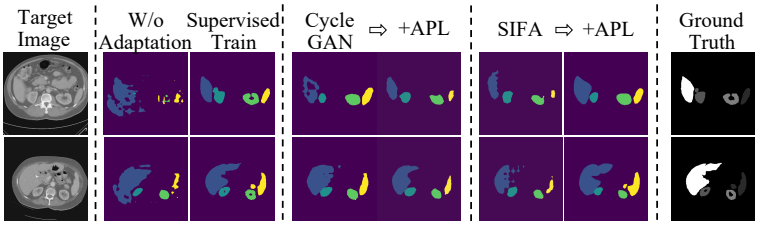

Fig. 2. **Comparison of segmentation results produced by different methods for abdominal organs segmentation.**

are derived from the public MICCAI and ISBI challenges, containing multi-abdominal organs described in different modalities, including 30 CT volumes and 20 MRI volumes.

As for the preprocessing method, the volume data is sliced as 2D images through transverse view and normalized with the size of 256x256. The segmentation objects consist of four abdominal organs, including liver (LV), right kidney (RK), left kidney (LK), and spleen (SP). Among the preprocessed data, the labels of four organs are obtained through pixel-level annotation by experienced specialists. As for the evaluation metric, the Dice similarity coefficient [19, 20] is utilized to compare the performance discrepancy of methods by calculating the overlap between predicted result and ground truth. Higher Dice values represent better performance.

3.2. Promotion based on adaptive pseudo labeling method

Firstly, we quantitatively measured the domain shift by calculating performance dependency in the abdominal organs segmentation. The bottom bound was to directly transfer the well-trained model from source domain to target domain without any domain adaptation technique. The top bound was to use the labels of target samples to train the target model based on the supervised train. Both of their performance were reported in the Table 1. We can find the enormous performance gap between top bound and bottom bound, which also indicated the severe domain shift between cross-modality images (MRI ⇒ CT) and the challenge of unsupervised source-free domain-adaptive medical image segmentation.

Table 2. Comparison of quantitative results between our APL method and other SOTA pseudo-label methods on cross-modality abdominal organ segmentation. Dice (%) of four organs and their average are compared here.

Pseudo Methods	LV	RK	LK	SP	Avg
Threshold [7]	81.44	75.18	73.18	79.98	77.59
CBST [8]	89.49	80.60	82.14	**84.19**	84.11
MRNet [10]	87.77	86.67	81.97	83.10	84.88
APL ($\mathcal{L}_{Con}$)	88.12	89.70	81.87	81.77	85.37
APL (+$\mathcal{L}_{SSL}$)	**89.56**	**89.74**	**83.66**	82.31	**86.32**

Second, in order to evaluate the promotion of adaptive pseudo labeling method based on the adapted models, we applied the proposed **APL** method to other seven UDA methods to segment multi-abdominal organs from cross-modality medical images. The quantitative results were reported in the Table 1, where the results of the above SOTA methods all referred from the paper [18]. Compared with the original adapted models based on the UDA methods, the introduced APL method made remarkable improvements by refining these models with rectified pseudo labels. For example, SIFA [17] increased the average Dice from 83.10% to 86.32% after introduced the APL method.

Third, we also compared the different promotions with different pseudo labeling methods. Adapted SIFA-v1 [17] was used as the basic model. The results were reported in Table 2. Our APL consistently outperformed the state-of-the-art methods as well as existing handcrafted-threshold strategies.

Finally, the qualitative segmentation results in Fig.2 also showed the challenge of UDA in medical image analysis. Our method can successfully locate the four organs and generate semantically meaningful masks. All the above qualitative and quantitative results validated the effectiveness of our method.

4. CONCLUSION

In this paper, we proposed a novel regularization for adaptive pseudo-label denoising, combining the dual-classifiers consistency and predictive category-aware confidence. Besides, the proposed method was orthogonal and thus can be plug-and-play to improve existing UDA methods. The pseudo labels were denoised and then used to refine the adapted model without source domain samples. The experimental results demostrated that the proposed method was superior to other methods in the cross-modality medical image segmentation.

5. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[3] Chen Li, Yusong Tan, Wei Chen, Xin Luo, Yulin He, Yuanming Gao, and Fei Li, "Anu-net: Attention-based nested u-net to exploit full resolution features for medical image segmentation," *Computers & Graphics*, 2020.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2014, pp. 580–587.

[5] Chen Li, Wei Chen, Mingfei Wu, Xin Luo, Yulin He, and Yusong Tan, "Tri-directional tasks complementary learning for unsupervised domain adaptation of cross-modality medical image semantic segmentation," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 1406–1411.

[6] Chen Li, Xin Luo, Wei Chen, Yulin He, Mingfei Wu, and Yusong Tan, "Attent: Domain-adaptive medical image segmentation via attention-aware translation and adversarial entropy minimization," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 952–959.

[7] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013, vol. 3, p. 896.

[8] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *European Conference on Computer Vision (ECCV)*, September 2018.

[9] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang, "Confidence regularized self-training," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5981–5990.

[10] Zhedong Zheng and Yi Yang, "Unsupervised scene adaptation with memory regularization in vivo," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 1076–1082.

[11] Shuang Li, Fangrui Lv, Binhui Xie, Chi Harold Liu, Jian Liang, and Chen Qin, "Bi-classifier determinacy maximization for unsupervised domain adaptation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[12] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K. Moyo, Michael R. Savona, Richard G. Abramson, and Bennett A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging (TMI)*, vol. 38, no. 4, pp. 1016–1025, 2019.

[13] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng, "Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning (ICML)*, 2018, vol. 80, pp. 1989–1998.

[16] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2512–2521.

[17] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," *in AAAI Conference on Artificial Intelligence (AAAI)*, pp. 865–872, 2019.

[18] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Transactions on Medical Imaging (TMI)*, vol. 39, no. 7, pp. 2494–2505, 2020.

[19] Chen Li, Wei Chen, and Yusong Tan, "Render u-net: A unique perspective on render to explore accurate medical image segmentation," *Applied Sciences*, vol. 10, no. 18, 2020.

[20] Chen Li, Wei Chen, and Yusong Tan, "Point-sampling method based on 3d u-net architecture to reduce the influence of false positive and solve boundary blur problem in 3d ct image segmentation," *Applied Sciences*, vol. 10, no. 19, 2020.