SEMI-SUPERVISED MEDICAL IMAGE SEGMENTATION VIA DUAL NETWORKS

Yunyao Lu[1], Yihang Wu[1], Reem Kateb[2], Ahmad Chaddad[1,3,]*

[1]AIPM, School of Artificial Intelligence, Guilin University of Electronic Technology, China
[2]College of Computer Science and Engineering, Jeddah University, Jeddah, Saudi Arabia.
[3]Laboratory for Imagery, Vision and Artificial Intelligence, École de Technologie Supérieure, Canada
Correspondence:ahmad8chaddad@gmail.com

ABSTRACT

Traditional supervised medical image segmentation models require large amounts of labeled data for training; however, obtaining such large-scale labeled datasets in the real world is extremely challenging. Recent semi-supervised segmentation models also suffer from noisy pseudo-label issue and limited supervision in feature space. To solve these challenges, we propose an innovative semi-supervised 3D medical image segmentation method to reduce the dependency on large, expert-labeled datasets. Furthermore, we introduce a dual-network architecture to address the limitations of existing methods in using contextual information and generating reliable pseudo-labels. In addition, a self-supervised contrastive learning strategy is used to enhance the representation of the network and reduce prediction uncertainty by distinguishing between reliable and unreliable predictions. Experiments on clinical magnetic resonance imaging demonstrate that our approach outperforms state-of-the-art techniques. Our code is available at `https://github.com/AIPMLab/Semi-supervised-Segmentation`.

Index Terms— Deep learning, medical image segmentation, semi-supervised learning, contrastive learning

1. INTRODUCTION

Medical image segmentation is a fundamental technique designed to partition images into distinct regions, playing an important role in the medical field. The identification and quantification of abnormal regions (e.g., abnormal areas) provide clinicians with critical diagnostic and therapeutic insights, enabling more informed medical decision making [1]. Many deep learning architectures [2, 3, 4] have shown promising results, achieving state-of-the-art (SOTA) performance in various medical image segmentation tasks [5].

However, traditional supervised learning methods are based on large-scale labeled datasets, which are not practical in the medical domain. Furthermore, pixel-level annotations in medical images are particularly challenging and labor intensive [6]. To address the impact of limited labeled data, in [5], they proposed semi-supervised learning (SSL), which can leverage unlabeled data with labeled data for model training to reduce dependence on large labeled datasets. Due to the unavailability of ground truth for unlabeled samples, pseudo-supervision is proposed for SSL. The key idea is to use the pseudo-labels predicted by the segmentation model for model training [7].

Despite advances in pseudo-labeling methods, two major challenges remain in the task of semi-supervised medical image segmentation: 1) *Noisy pseudo-label*, and 2) *insufficient supervision in the feature space*. Pseudo-supervision relies on segmentation models to generate pseudo-labels for unlabeled images. However, these models are highly susceptible to label noise, which weakens the effectiveness of the pseudo labeling process. Similarly, insufficient supervision in the feature space also limits the performance of SSL [8]. Many methods only provide supervision in the label space and lack explicit guidance in the feature space, resulting in insufficient class separability [9, 10, 11, 12]. Although recent studies have attempted to mitigate noise by filtering out predictions with classification scores below a certain threshold [12], this approach does not fully eliminate false predictions (e.g., some incorrect predictions may still yield high classification scores, leading to overconfidence or miscalibration [13]). Furthermore, using a high threshold to exclude unreliable predictions considerably decreases the number of pseudo labels, which in turn impacts the effectiveness of SSL. This reduction in pseudo-labels can lead to class imbalance in the training data, reducing overall segmentation performance [14].

Motivated by the previous challenges, this study proposes a dual-stream network architecture inspired by the idea of Knowledge Distillation (KD) with S-T frameworks, where each sub-network has a 3D encoder-decoder module. Furthermore, we introduce a supervised loss function to guide the network in learning the features of each class. Additionally, to stabilize the training process, we involve consistency regularization by minimizing the cosine distance between the predictions of the student and teacher networks. The teacher network generates pseudo labels for unlabeled data, while the student network uses those data as supervisory features for learning. In addition, to alleviate the negative impact of noisy pseudo-labels on model performance, we dynamically adjust
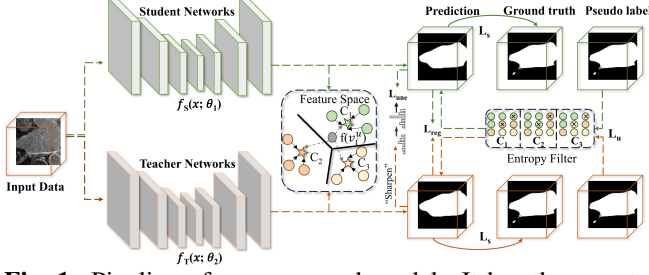
Fig. 1: Pipeline of our proposed model. It has three parts: consistency regularization, uncertainty estimation, and a contrastive learning method to improve student and teacher performance.

the contribution of pseudo-labels to model training based on uncertainty estimation. Specifically, uncertainty estimation introduces adaptive voxel weighting for pseudo-supervised loss, where voxels with higher confidence are assigned higher weights, and vice versa. Finally, we use contrastive learning to align the uncertain pixel features with the prototypes of reliably predicted classes, while increasing the distance between different class prototypes. Our contributions can be summarized as follows.

1. We introduce a consistency regularization term to reduce discrepancies between student and teacher network predictions. It leads to stabilize the training process.

2. We propose an uncertainty estimation mechanism to dynamically adjust pseudo-label contributions, mitigating noise and avoiding over depended on uncertain regions.

3. We use contrastive loss to align unreliable pixel features with reliable class prototypes, while increasing separation between different class prototypes.

2. PROPOSED METHOD

Figure 1 shows the pipeline of our proposed method. It consists of two subnetworks: the Student network $f_S(x; \theta_1)$ and the Teacher network $f_T(x; \theta_2)$ both employing a 3D encoder-decoder design with comparable performance. It uses a combination of labeled data $D_l = \left\{ \left(x_i^l, y_i^l \right) \right\}_{i=1}^{N_l}$ and a larger set of unlabeled data $D_u = \{x_i^u\}_{i=1}^{N_u}$ to enhance the segmentation performance. Furthermore, we introduce consistency regularization, uncertainty estimation, and contrastive loss to improve the performance of the student and teacher model.

In general, the total loss is divided into supervision loss $\mathcal{L}_s$, self-supervised contrastive loss $\mathcal{L}_c$, semi-supervised pseudo-labeling loss $\mathcal{L}_u$, and regularization loss $\mathcal{L}_{reg}$.

$$\mathcal{L} = \underbrace{\mathcal{L}_s^S + \mathcal{L}_s^T}_{\text{labelled}} + \lambda_c \underbrace{\mathcal{L}_c + \mathcal{L}_{reg} + \mathcal{L}_{une}}_{\text{self}} + \underbrace{\mathcal{L}_u^S + \mathcal{L}_u^T}_{\text{semi}} \quad (1)$$

For the labeled data, our aim is to minimize the Cross-Entropy loss and Dice loss, as defined by the following Equation 2.

$$\mathcal{L}_s = \mathcal{L}_{ce}(p_i^l, y_i^l) + \text{Dice}(\sigma(p_i^l), y_i^l) \quad (2)$$

Here, y_i^l represents the manually annotated mask label for the i-th labeled image, while p denotes the model prediction.

For the unlabeled data, we apply pixel-level entropy filtering to remove unreliable pseudo-labels prior to computing the unsupervised loss. We introduce an unsupervised loss function based on entropy minimization, aimed at reducing the uncertainty in model predictions and preventing the model from being misled by identifying and excluding areas with high uncertainty. Specifically, we calculate the entropy of each pixel to assess the confidence of the model in its predictions. Regions with the greatest uncertainty are filtered out according to a predefined ratio and assigned a special ignore label. Subsequently, we employ a cross-entropy loss function with an ignored index to compute the loss, thereby effectively minimizing the influence of noise on the model.

$$Ent_S = -\sum_{c=1}^{C} p_i^u(c) \cdot \log_2 \left(p_i^u(c) + \epsilon \right), \quad (3)$$

where p_c^u represents the predicted probability for class c. The Ent_T count in the same way.

Define a mask for valid points based on a threshold τ, while ignoring high-entropy pixels:

$$\hat{y}_i^u = \begin{cases} \arg\max_c \ p_i^u(c), & \text{if } Ent < \tau, \\ \text{ignore}, & \text{otherwise.} \end{cases}$$

We calculate the unsupervised loss defined in Equation 4:

$$\mathcal{L}_u = \frac{\sum_{i=1}^{B \times H \times W \times D} \mathcal{L}_{ce}(f(x; \theta), \mathbb{I}\,[\hat{y}_i^u \neq \text{ignore}])}{\sum_{i=1}^{B \times H \times W \times D} \mathbb{I}\,[\hat{y}_i^u \neq \text{ignore}]}, \quad (4)$$

2.1. Consistency Regularization

Relying on the smoothness assumption, where small perturbations to an input should not result in significantly different predictions [15], model consistency becomes a crucial objective. Specifically, model consistency demands that when the same image is passed through different models, the predictions should be similar. However, using pseudo-labels can lead to increased risk of training instability. To address this, we introduce a consistency regularization term that enforces prior constraints based on these assumptions, ensuring more stable training.

To further encourage consistency between the two subnetworks S and T, we employ a cosine similarity-based consistency regularization. This regularization aims to minimize the cosine distance between the predictions of the subnets S

and T, reducing their disagreement, and thus stabilizing the training. The consistency regularization loss is defined as:

$$\mathcal{L}_{reg} = 1 - \cos(p_i^S, p_i^T) \tag{5}$$

Where p_i^S refers to the i-th output of the student model, and p_i^T represents the i-th output of the teacher model. The cosine similarity between the outputs of the two subnetworks is calculated as:

$$\cos(p_i^S, p_i^T) = \frac{p_i^S \cdot p_i^T}{\|p_i^S\|\|p_i^T\|} \tag{6}$$

By minimizing this loss, we ensure that the predictions from different models of the same input remain consistent, thus reducing variance and improving model generalization.

2.2. Uncertainty Estimation

Pseudo supervision can sometimes be unreliable due to label noise. A common method to reduce this noise is confidence threshold, but this approach can fail in segmentation tasks, as it may bias the model towards easier classes, leaving harder classes underrepresented. To address this challenge, we propose using prediction uncertainty to refine pseudo supervision. By estimating uncertainty through KL-divergence, we adjust the contribution of the pseudo labels based on prediction confidence.

The uncertainty loss is defined as:

$$\mathcal{L}_p(p_i^S, p_i^T) = \mathcal{L}_{ce}(p_i^S, \sigma(p_i^T / Tp)), \tag{7}$$

$$\mathcal{L}_{une}(p_i^S, p_i^T) = e^{-\mathcal{D}_{KL}(p_i^S, p_i^T)} \mathcal{L}_p(p_i^S, p_i^T) + \mathcal{D}_{KL}(p_i^S, p_i^T) \tag{8}$$

$$D_{KL}(p_i^S, p_i^T) = p_i^T \log\left(\frac{p_i^T}{p_i^S}\right). \tag{9}$$

where, $\mathcal{D}_{KL}(p_i^S, p_i^T)$ represents the KL-divergence between the model prediction and the pseudo label. We sharpen the softmax probabilities by dividing the logits by a temperature hyperparameter Tp.

2.3. Contrastive Loss for Uncertain Voxels

To address the issue of insufficient supervision in the feature space, we incorporate a contrastive loss function. The goal is to align uncertain voxels with their respective class prototypes, reducing the risk of misclassification and minimizing prediction uncertainty. This approach estimates the confidence level of each voxel prediction and classifies the predictions into two categories: reliable and unreliable. Based on the reliable predictions, class prototypes are constructed by averaging the representations of these reliable voxels.

Initially, we calculate the foreground and background prototypes, which are the mean vectors of the reliable voxel representations. These prototypes represent the class identity

within the feature space. The foreground class prototype is calculated as:

$$c_f = \frac{1}{|S_f|} \sum_{(\mathbf{v}_i^r, y_i) \in S_f} f(\mathbf{v}_i^r) \tag{10}$$

The calculation for the background class prototype, c_b, follows the same equation.

Where S_f denotes the sets of reliable voxels for the foreground and background classes, respectively. $f(\mathbf{v}_i^r)$ represents the feature embedding of a reliable voxel $\mathbf{v}_i^r$.

The contrastive loss function is composed of three terms, designed to align uncertain voxels with their corresponding class prototypes while maximizing the separation between different prototypes:

$$\mathcal{L}_c = d(f(\mathbf{v}_i^u), c_f) + d(f(\mathbf{v}_i^u), c_b) + d(c_f, c_b) \tag{11}$$

The function $d(\cdot, \cdot)$ calculates the distance between the voxel's feature embedding and the class prototypes, encouraging the alignment of uncertain voxels to the most appropriate class prototype.

Table 1: Summary of test metrics in LA dataset (MRI images). The best two results are marked in bold.

Method	Dice (%)	Jaccard (%)	95HD (voxel)	ASD (voxel)
MT [11]	85.89 ± 0.024	76.58 ± 0.027	12.63 ± 5.741	3.44 ± 1.382
UA-MT [16]	85.98 ± 0.014	76.65 ± 0.017	9.86 ± 2.707	2.68 ± 0.776
SASSNet [17]	86.21 ± 0.023	77.15 ± 0.024	9.80 ± 1.842	2.68 ± 0.416
DTC [18]	86.36 ± 0.023	77.25 ± 0.020	9.92 ± 1.015	2.40 ± 0.223
MC-Net [19]	87.65 ± 0.011	78.63 ± 0.013	9.70 ± 2.361	3.01 ± 0.700
MCF [20]	88.71 ± 0.018	80.41 ± 0.022	6.32 ± 0.800	1.90 ± 0.187
LUSEG [14]	89.10 ± 0.012	81.62 ± 0.024	6.30 ± 0.850	1.80 ± 0.020
Ours	**90.19 ± 0.012**	**82.42 ± 0.019**	**5.74 ± 0.757**	**1.75 ± 0.227**

Table 2: Summary of test metrics in Pancreas dataset (CT images).

Method	Dice (%)	Jaccard (%)	95HD (voxel)	ASD (voxel)
MT [11]	74.43 ± 0.024	60.53 ± 0.030	14.93 ± 2.000	4.61 ± 0.929
UA-MT [16]	74.01 ± 0.029	60.00 ± 3.031	17.00 ± 3.031	5.19 ± 1.267
SASSNet [17]	73.57 ± 0.017	59.71 ± 0.020	13.87 ± 1.079	3.53 ± 1.416
DTC [18]	73.23 ± 0.024	59.18 ± 0.027	13.20 ± 2.241	3.83 ± 0.925
MC-Net [19]	73.73 ± 0.019	59.19 ± 0.021	13.65 ± 3.902	3.92 ± 1.055
MCF [20]	75.00 ± 0.026	61.27 ± 0.030	11.59 ± 1.611	3.27 ± 0.919
LUSEG [14]	**76.40 ± 0.018**	**62.96 ± 0.027**	**10.69 ± 1.603**	2.79 ± 0.0954
Ours	75.25 ± 0.014	61.65 ± 0.017	11.15 ± 1.604	**2.49 ± 0.293**

3. EXPERIMENTS

3.1. Datasets

The Left Atrial Dataset (LA). The LA [21] consists of 100 3D gadolinium-enhanced MR volumes with manual left atrial annotations at an isotropic resolution of $0.625 \times 0.625 \times 0.625$ mm³. The dataset is divided into 5 folds of 20 volumes each. Preprocessing includes normalization to zero mean and unit variance according to [20, 14]. During training, random cropping generates input volumes of size $112 \times 112 \times 80$. For inference, a sliding window approach of the same dimensions and a stride of $18 \times 18 \times 4$ is used. *The Pancreas Dataset.*

Table 3: Ablation study with loss functions.

$\mathcal{L}_{une}$	$\mathcal{L}_{reg}$	Left Atrium				Pancreas-CT			
		Dice (%)	Jaccard (%)	HD95 (voxel)	ASD (voxel)	Dice (%)	Jaccard (%)	HD95 (voxel)	ASD (voxel)
✓		88.84 ± 0.011	80.38 ± 0.015	6.76 ± 0.802	2.16 ± 0.292	**75.44 ± 0.018**	**61.80 ± 0.021**	11.37 ± 1.70	3.17 ± 0.128
	✓	89.28 ± 0.012	81.02 ± 0.017	6.58 ± 0.628	2.06 ± 0.269	73.92 ± 0.021	60.10 ± 0.025	**10.13 ± 1.47**	**2.17 ± 0.248**
✓	✓	**90.19 ± 0.012**	**82.42 ± 0.019**	**5.74 ± 0.757**	**1.75 ± 0.227**	75.25 ± 0.014	61.65 ± 0.017	11.15 ± 1.60	2.49 ± 0.293
		89.03 ± 0.012	80.61 ± 0.016	6.68 ± 0.759	2.24 ± 0.294	74.62 ± 0.023	60.83 ± 0.027	11.17 ± 0.95	2.81 ± 0.302

These datasets include 82 contrast-enhanced 3D CT volumes with manual pancreas annotations [22]. Each volume size is $512\times512\times D$, where D ranges from 181 to 466 slices. The dataset is split into four folds. Preprocessing involves applying a soft tissue window (-120 to 240 HU) and cropping around the pancreas with a 25-voxel margin according to [20, 14]. Training uses random cropping to resize volumes to $96\times96\times6$, and inference is performed with a sliding window of $16\times16\times16$.

Implementation details Following [20, 14], we adopted ResNet and V-Net as the two-stream network backbone, and used K-fold cross-validation (i.e., 5 folds for LA and 4 folds for Pancreas) to provide a comprehensive evaluation. The SGD optimizer (weight decay: 0.0001, momentum: 0.9) is used to optimize the network parameters, starting with a learning rate of 0.01, which decreased by a factor of 10 every 2500 iterations for a total of 6000 iterations. Each training iteration included both labeled and unlabeled samples, with a batch size of two. In Eq. 1, we set $\lambda_c = 0.1 \times e^{4(1-\frac{t}{t_{\max}})^2}$, where t and $t_{\max}$ represent the current and maximum iterations, respectively. The experiment environment is based on the Windows 11 operating system and features an Intel 13900KF CPU with 128 GB of RAM and an L40 GPU. We considered the Dice score, Jaccard, Hausdorff distance (HD95), and Average Surface Distance (ASD) to compare the performance of the models [23].

Results Table 1 and Table 2 report the performance metrics using the LA and the Pancreas dataset. In Table 1, compared to LUSEG, our model shows an improvement in Dice, with Dice increasing from 89.10 to 90.19. In addition, our method maintains a low performance variance, which improves the stability and reliability. Table 2 demonstrates that our model achieves feasible performance in ASD. Figure 2 illustrates the visual results of the left atrial and pancreas datasets on different segmentation methods. Within the pancreas dataset, our model shows lower error rates than the MCF and LUSEG models.

Ablation study Table 3 presents the effects of the regularization and uncertainty estimation elements within our model. In particular, these two losses ($\mathcal{L}_{une}$ and $\mathcal{L}_{reg}$) have different effects on the two datasets. For LA dataset, removing the regularization module resulted in a drop in the Dice from 90.19 to 88.84. Similarly, removing the uncertainty loss resulted in a drop in the Dice score from 90.19 to 89.28. Dice score decreases from 90.19 to 89.03 when both losses are absent.
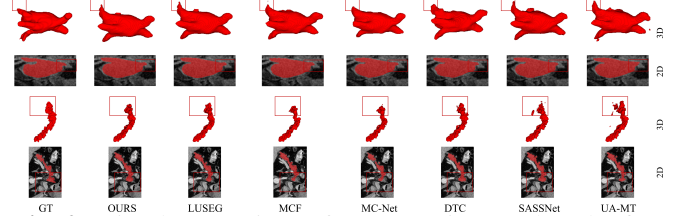

Fig. 2: Visual comparison of segmentation results: the first and second rows show the left atrium (LA), while the third and fourth rows show the pancreas, respectively.

However, in pancreas datasets, we observed that the combination of both losses ($\mathcal{L}_{reg}$ and $\mathcal{L}_{une}$) resulted in lower Dice and Jaccard scores compared to using only uncertainty loss ($\mathcal{L}_{une}$). Yet, the combination led to better 95HD and ASD scores than using only the regularization loss ($\mathcal{L}_{reg}$). This result may be assigned to the varying features of the two datasets. For example, the Pancreas-CT dataset contains more complex structures and varied image features, which may cause the combination of losses to overfit or struggle to generalize well. The additional complexity of combining both losses might lead to better boundary (HD and ASD) performance, but at the cost of a reduction in overlap metrics (Dice and Jaccard). This suggests that for some complex datasets, focusing on individual loss components may yield better results, while the combined effect could introduce noise that affects segmentation performance.

4. CONCLUSION

In this paper, we presented a 3D network, an innovative contrastive consistency segmentation model designed to effectively leverage limited annotations for medical image segmentation. By incorporating an enhanced contrastive learning strategy alongside consistency and uncertainty estimation schemes, our approach achieves promising results on CT and MRI images. In future work, we will involve domain adaptation [24] to solve domain shifts that exist among different modalities to further boost the performance of our model in multi-source situations.

5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

7. REFERENCES

[1] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi, "Medical image segmentation using deep learning: A survey," *IET image processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[4] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[5] Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin, "Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation," *Computers in Biology and Medicine*, p. 107840, 2023.

[6] Ziyang Wang and Congying Ma, "Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 870–879.

[7] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8934–8954, 2022.

[8] Xiangyu Zhao, Zengxin Qi, Sheng Wang, Qian Wang, Xuehai Wu, Ying Mao, and Lichi Zhang, "Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[9] Yassine Ouali, Céline Hudelot, and Myriam Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12674–12684.

[10] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2613–2622.

[11] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[12] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18408–18419, 2021.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[14] Sanaz Karimijafarbigloo, Reza Azad, Yury Velichko, Ulas Bagci, and Dorit Merhof, "Leveraging unlabeled data for 3d medical image segmentation through self-supervised contrastive learning," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[15] Kai Han, Victor S Sheng, Yuqing Song, Yi Liu, Chengjian Qiu, Siqi Ma, and Zhe Liu, "Deep semi-supervised learning for medical image segmentation: A review," *Expert Systems with Applications*, p. 123052, 2024.

[16] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*. Springer, 2019, pp. 605–613.

[17] Shuailin Li, Chuyu Zhang, and Xuming He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 552–561.

[18] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 8801–8809.

[19] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part II 24*. Springer, 2021, pp. 297–306.

[20] Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao, "Mcf: Mutual correction framework for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15651–15660.

[21] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al., "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical image analysis*, vol. 67, pp. 101832, 2021.

[22] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer, 2015, pp. 556–564.

[23] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, pp. 210, 2022.

[24] Ahmad Chaddad, Yihang Wu, Yuchen Jiang, Ahmed Bouridane, and Christian Desrosiers, "Simulations of common unsupervised domain adaptation algorithms for image classification," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–1, 2025.