FRCNet: Frequency and Region Consistency for Semi-supervised Medical Image Segmentation

Along He[1], Tao Li[1,2](✉), Yanlin Wu[1], Ke Zou[3], and Huazhu Fu[4]

[1] Tianjin Key Laboratory of Network and Data Security Technology, College of Computer Science, Nankai University, Tianjin, China
litao@nankai.edu.cn
[2] Haihe Lab of ITAI, Tianjin, China
[3] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, Sichuan, China
[4] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore, Singapore

Abstract. Limited labeled data hinder the application of deep learning in medical domain. In clinical practice, there are sufficient unlabeled data that are not effectively used, and semi-supervised learning (SSL) is a promising way for leveraging these unlabeled data. However, existing SSL methods ignore frequency domain and region-level information and it is important for lesion regions located at low frequencies and with significant scale changes. In this paper, we introduce two consistency regularization strategies for semi-supervised medical image segmentation, including frequency domain consistency (FDC) to assist the feature learning in frequency domain and multi-granularity region similarity consistency (MRSC) to perform multi-scale region-level local context information feature learning. With the help of the proposed FDC and MRSC, we can leverage the powerful feature representation capability of them in an effective and efficient way. We perform comprehensive experiments on two datasets, and the results show that our method achieves large performance gains and exceeds other state-of-the-art methods.

Keywords: Semi-supervised learning · Medical image segmentation · Consistency regularization

1 Introduction

Medical image segmentation aims to segment the anatomical structure, organs or lesions for clinical analysis and diagnosis such as skin disease diagnosis [22] and fundus screening [6]. It has been significantly developed with the rise of convolutional neural networks (CNNs) [17, 28, 11, 21]. While existing approaches rely on high-quality labeled data, which is labor-intensive and time-consuming in the medical domain. It is not always feasible to collect sufficient labeled data in real world applications, especially for pixel-wise segmentation tasks.

Training with limited data may easily lead to over-fitting and catastrophic forgetting [9], causing the performance drop. In clinical practice, there are a large number of unlabeled data available, and the dependence on labeled data can be alleviated if these data can be utilized effectively. Therefore, it is promising to develop semi-supervised learning (SSL) methods to reduce the burden of annotation in medical domain. Therefore, many attempts have been made at SSL [19, 18]. They introduce data-level [12, 23] and model-level consistency [19, 16, 15], or generate pseudo labels using the model's predictions [4].

Despite achieving good results, previous methods still face the following limitations. **Firstly**, current SSL methods mainly focus on RGB domain, ignoring the frequency domain. The intensity of most lesion regions in medical images is flat and they are low-frequency signals [27]. Segmenting lesions with only the RGB domain information is an challenging task without explicitly considering frequency information. Therefore, we aim to explore the rich information of frequency and RGB domains. **Secondly**, most of previous methods focus on pixel-level consistency, and lack semantic consistency at region level. Considering multiple region-level consistency can effectively deal with lesions of different sizes and enhance the multi-scale features modeling ability.

In this paper, we come up with two consistency regularization strategies for SSL, one is frequency domain consistency (FDC), and the other is multi-granularity region similarity consistency (MRSC). For FDC, we apply the Discrete Cosine Transform (DCT) [1] to convert the RGB domain features into the frequency domain. To discover the cues of lesion information in frequency space, we design a frequency enhancement module (FEM) with Transformer [20] to encode the features in frequency domain, which can boost the capabilities of the model in processing frequency signals. For MRSC, we aim to provide multi-granularity regional context information, rather than only pixel-wise consistency. We adopt the Mean Teacher (MT) [19] framework as base framework, and the proposed consistency regularization is incorporated to form the Frequency and Region Consistency (FRCNet) for SSL medical image segmentation.

The main contributions of our work: (1) We propose FDC and MRSC consistency regularization strategies, which can learn from sufficient unlabeled data to deal with insufficient annotation. (2) Our method is plug-and-play and it can be easily combined with existing SSL methods. In the inference time, FDC and MRSC can be safely removed, reducing the inference complexity. (3) Finally, our FRCNet outperforms the state-of-the-art (SOTA) methods on two datasets by a large margin.

2 Method

2.1 Notations and Overview.

For SSL, training set contains M labeled and N unlabeled data, where $N \gg M$. The labeled pairs and unlabeled images are denoted as $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^{M}$ and $\mathcal{D}^u = \{x_j\}_{j=1}^{N}$, respectively, where i and j are indexes of labeled and unlabeled data, respectively, and the training dataset is $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$.

The framework of our method is illustrated in Fig 1, mean teacher (MT) [19] is employed as the baseline, which consists of a teacher F_t and a student

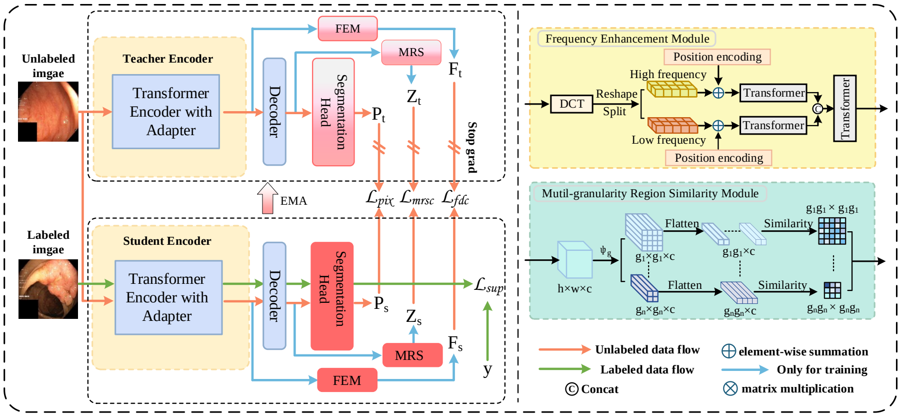

Fig. 1: Overall framework of the proposed FRCNet, which is based on MT structure. For labeled data, they are used to train the student network directly with the labels, and unlabeled data is used to train the student network through the proposed consistency regularization.

model F_s. For F_s, it contains a Transformer encoder with adapter [3], a decoder, a segmentation head, MRS and FEM modules. MRS is to obtain multi-scale region-level features, and the FEM branch is used to transform the RGB features to frequency domain. Formally, the outputs of teacher and student models are denoted as: segmentation prediction $\mathcal{P}$, frequency domain features $\mathcal{F}$ and regional features Z. For labeled data, student model is trained with cross-entropy and Dice loss, they are formulated as:

$$\mathcal{L}_{sup} = \frac{1}{2|\mathcal{D}_l|} \sum_{(x_i,y_i)\in\mathcal{D}_l} \mathcal{L}_{ce}(y_i, \mathcal{P}_i^s) + \mathcal{L}_{dice}(y_i, \mathcal{P}_i^s), \tag{1}$$

where s denotes student model.

For unlabeled data, student model can be optimized through the pixel-level loss, and the proposed frequency domain and region-level consistency, which is described in Fig 1. During training, we only fine-tune the parameters of segmentation head, adapters, FEM and MRS for parameter efficiency. Exponential moving average (EMA) is used to update teacher model [19].

2.2 Frequency Domain Consistency.

For RGB features from the last stage of Transformer encoder $X \in \mathbb{R}^{h\times w\times d}$, where h, w and d denote the height, width and channel number, respectively. First, we transform the RGB features X using DCT [1] and obtain frequency features. Then, we divide it into N patches of size $p \times p$, and $N = \frac{h\cdot w}{p^2}$, and the frequency domain features are denoted as $\mathcal{F}_{raw} \in \mathbb{R}^{\frac{h}{p}\times\frac{w}{p}\times c}$, $c = p \cdot p \cdot d$, and thus the RGB features are transformed to the frequency domain, and p is set to 2.

Frequency Enhancement Module. After the transformation of DCT, we adopt Transformer [20] to enhance the raw frequency features, as shown on the right of Fig 1. To model lesion features that belong to low-frequency, we decompose the raw frequency domain $\mathcal{F}_{raw} = \{\mathcal{F}_l, \mathcal{F}_h\}$, where $\mathcal{F}_l \in \mathbb{R}^{\frac{h}{p} \times \frac{w}{p} \times \frac{c}{2}}$ and $\mathcal{F}_h \in \mathbb{R}^{\frac{h}{p} \times \frac{w}{p} \times \frac{c}{2}}$ denote the low-frequency component and the high-frequency component, respectively. To boost the signals in the frequency domain, the low-frequency component $\mathcal{F}_l$ and high-frequency component $\mathcal{F}_h$ are fed into two independent Transformer blocks to encode frequency features, which can be denoted as

$$\mathcal{F}_l' = \Phi_1(\mathcal{F}_l + E_l^{pos}), \mathcal{F}_h' = \Phi_2(\mathcal{F}_h + E_h^{pos}), \tag{2}$$

where Φ_1 and Φ_2 are the two Transformer blocks, each of them consists of a multi-head self-attention layer and multi-layer perceptron. $E_l^{pos} \in \mathbb{R}^{\frac{h}{p} \times \frac{w}{p} \times \frac{c}{2}}$ and $E_h^{pos} \in \mathbb{R}^{\frac{h}{p} \times \frac{w}{p} \times \frac{c}{2}}$ are the learnable position embeddings to preserve the frequency location information. Then, another Transformer block Φ_3 is adopted to encode all the low-frequency and high-frequency components to better fuse the features in frequency space: $\mathcal{F} = \Phi_3([\mathcal{F}_l', \mathcal{F}_h'])$, $[.,.]$ denotes concatenation along the channel dimension, and $\mathcal{F}$ is the final fused high and low frequency features.

Frequency Domain Consistency Loss. For the same input images, after the feature encoding in frequency domain, the outputs of student and teacher models should be consistent. We can obtain the unsupervised consistency loss on unlabeled data. For simplicity, we use Mean Square Error (MSE) loss as the consistency regularization:

$$\mathcal{L}_{fdc} = \frac{1}{|\mathcal{D}_u|} \sum_{x_j \in \mathcal{D}_u} \left\| \mathcal{F}_j^t - \mathcal{F}_j^s \right\|_2^2, \tag{3}$$

where $\mathcal{L}_{fdc}$ is the frequency domain consistency loss, $\mathcal{F}_j^t$ and $\mathcal{F}_j^s$ are the frequency space features of teacher and student models, respectively.

2.3 Multi-granularity Region Similarity Consistency.

Although the SSL methods [19, 15] based on pixel-level consistency have achieved decent results, they ignore the local information in 2D images, and thus they cannot model the relationship between local regions well. Therefore, we propose multi-granularity region similarity consistency, which builds region-level contextual relationships with different scales, as shown on the right of Fig. 1. Note that MRSC does not contain any trainable parameters, it only includes the feature transformation.

Specifically, in order to obtain the features at region level, we adopt the output from the last stage of decoder as the region features. The features from the teacher and student models are denoted as Z^t and $Z^s \in \mathbb{R}^{h' \times w' \times c'}$. First, they are projected and flatten into small regional features R,

$$R = \Gamma(\Psi_g(Z)) \in \mathbb{R}^{gg \times c'}, \tag{4}$$

where Γ is the flatten operation and Ψ_g is the region projection with the granularity of g. The region R is with the size of $g \times g$, and each element in R represents a local region in original image. It can be implemented with arbitrary downsampling methods, such as linear interpolation and pooling based methods, and we use linear interpolation in our method (See Table. 3). Then, we compute region similarity matrix $A = RR^T \in \mathbb{R}^{gg \times gg}$, where the elements indicates the similarity between regions, they can be used to model the similarity of region features. The region similarity consistency loss is to minimize the difference between the teacher and student model:

$$\mathcal{L}_{mrsc} = \frac{1}{|\mathcal{D}_u|} \sum_{x_j \in \mathcal{D}_u} \sum_{g \in G} \left\| A^t_{j,g} - A^s_{j,g} \right\|_2^2 , \tag{5}$$

G denotes the total granularities, $A^t_{j,g}$ and $A^s_{j,g}$ are region similarity scores from teacher and student with granularity g of sample x_j. In this paper, we adopt three granularities $G = \{16, 24, 32\}$ as a default setting (See Table. 3). To further model the pixel-level information, we use MSE to maximize the output similarity predicted by the student and teacher models, and thus the loss is formulated as:

$$\mathcal{L}_{pix} = \frac{1}{|\mathcal{D}_u|} \sum_{x_j \in \mathcal{D}_u} \left\| \mathcal{P}^t_j - \mathcal{P}^s_j \right\|_2^2 \tag{6}$$

The overall training loss on labeled and unlabeled data is defined as:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda(\mathcal{L}_{fdc} + \mathcal{L}_{mrsc} + \mathcal{L}_{pix}), \tag{7}$$

where λ is a time-dependent Gaussian warming up function [26].

3 Experiments

3.1 Dataset and Evaluation Metrics.

Datasets. Kvasir-SEG dataset [8] is designed for gastrointestinal polyp segmentation and includes 1000 polyp images along with their corresponding labels. We randomly choose 80% of them as training set, and the remaining as testing set. Skin lesion dataset (ISIC 2016) [7] contains dermoscopy images, which are used for automated diagnosis of skin cancer. It includes 900 training and 379 testing images.
Evaluation Metrics: We evaluate performance using the following metrics: pixel-wise Accuracy (Acc), Mean Absolute Error (MAE), Dice Similarity Coefficient (Dice), and Intersection-over-Union (IoU).

3.2 Implementation Details.

The pre-trained backbone is SegFormer-B4 [24], and the adapter in backbone is AdaptFormer [3]. The AdamW optimizer with weight decay 0.0001 and momentum 0.9 is adopted for training 90 epochs, and the learning rates is set to 0.0005. The images and labels are resized to 512×512, with a batch size of 10. All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU using the Pytorch. Code will be available at https://github.com/NKUhealong/FRCNet.

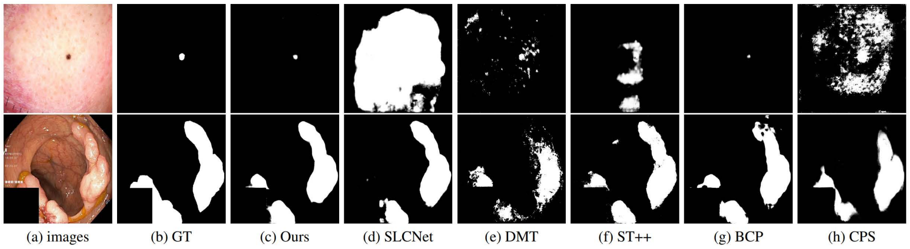

Fig. 2: The segmentation results of skin (first row) and polyp segmentation (second row) tasks using 10% labeled data.

3.3 Comparisons with SOTA Methods.

We compare our method against recent SOTA methods on two datasets. To make a fair comparison, we reimplemented these SOTA methods and obtained the results with the same backbone. Considering the reproducibility, we report the average results of 3 runs. The quantitative results are shown in Table 1, when we only adopt the labeled data to train the supervised network, we achieve inferior performance than the SSL method, which indicates that the problem of insufficient annotation can be alleviated by utilizing the unlabeled data. Our FRCNet achieves 90.77% Dice and 84.70% IoU on skin dataset with 10% labeled data, surpassing the second best method BCP [2] by 1.75% and 1.77% in Dice and IoU scores. Our method still outperforms the second best method ST++ [25] by 2.54% and 2.20% in Dice and IoU scores on polyp lesion segmentation task. Table 1 shows all the previous SSL methods exhibit sub-optimal performance. In contrast, our method consider both region-level and frequency domain information, and outperforms all these methods.

With 20% labeled data, our method also achieved the best results, showing the effectiveness of the proposed consistency regularization strategies. Our approach can make use of large amounts of unlabeled data with less labeled data. When using only 20% of the labeled data, the results are already very close to the supervised results. For qualitative analysis, we show the visual comparisons of challenging cases in Fig. 2. It can be seen that the predictions of other methods fail to capture the details of small lesions (the first row) or can not be able to completely segment the edges of lesions (the second row), and our method has better scalability to lesions with different shapes and scales.

3.4 Ablation Study.

Analysis of FDC and MRSC. To study the importance of high frequency and low frequency for medical image segmentation, we integrated the two components with baseline, respectively. It can be seen from Table 2 that low frequency component plays a more important role, because many lesions are similar to the color and texture features of surrounding normal regions, and they belong to

Table 1: Comparison of the proposed FRCNet and other SOTA SSL methods. Boldface and underline represent the best and second best results of each setting, respectively.

Ratio	Method	Kvasir-SEG				ISIC 2016			
		MAE↓	Acc↑	Dice↑	IoU↑	MAE↓	Acc↑	Dice↑	IoU↑
100%	SegFormer-B4	**3.01**	**97.48**	**91.79**	**86.86**	**3.37**	**95.93**	**92.88**	**87.31**
10%	SegFormer-B4	5.39	94.52	78.83	71.07	6.27	93.51	87.15	79.54
	MT [19]	4.82	94.97	83.45	76.32	5.96	94.03	88.01	81.42
	SASSNet [10]	4.91	94.78	82.78	75.21	5.84	93.78	88.42	81.46
	CCT [16]	4.39	94.88	83.74	76.45	5.79	93.78	88.22	81.34
	URPC [15]	4.89	94.69	82.43	75.49	5.74	93.74	88.12	81.21
	DTC [14]	4.98	94.60	80.75	73.84	5.82	93.62	87.94	81.23
	CPS [4]	4.67	94.91	82.81	75.34	5.23	94.28	88.89	82.12
	SLCNet [13]	5.01	94.79	83.43	75.64	5.21	94.82	88.23	82.09
	DMT [5]	4.34	94.89	84.98	77.23	4.94	94.93	88.03	82.34
	ST++ [25]	4.04	95.31	86.23	80.41	5.03	94.98	88.17	82.46
	BCP[2]	4.21	95.09	85.79	78.43	4.65	95.09	89.02	82.93
	FRCNet (Ours)	**3.51**	**96.49**	**88.77**	**82.61**	**4.59**	**95.41**	**90.77**	**84.70**
20%	SegFormer-B4	4.86	95.27	82.16	74.69	5.76	94.32	88.65	81.58
	MT [19]	4.67	95.47	84.45	77.43	5.05	95.43	89.83	83.15
	SASSNet [10]	4.45	95.56	83.97	77.21	5.03	95.19	89.94	83.67
	CCT [16]	4.01	95.78	84.89	77.94	5.09	95.42	89.59	82.99
	URPC [15]	4.35	95.52	85.12	78.21	5.03	95.21	89.91	82.81
	DTC [14]	4.31	95.65	84.22	78.01	5.11	95.39	89.79	83.32
	CPS [4]	4.23	95.88	85.83	78.86	4.90	95.11	89.03	84.67
	SLCNet [13]	4.81	95.74	85.42	78.35	4.86	95.34	89.34	84.54
	DMT [5]	4.03	95.83	86.47	79.32	4.45	95.56	89.21	84.32
	ST++ [25]	3.47	96.36	87.94	82.17	4.12	95.59	90.21	85.23
	BCP[2]	3.54	96.23	87.43	80.45	4.19	95.32	89.98	85.01
	FRCNet (Ours)	**2.78**	**97.22**	**90.62**	**85.01**	**3.70**	**96.30**	**92.44**	**86.73**

the low frequency signals. When high and low frequencies are integrated, the results can be further improved, indicating that the two frequency components are complementary. We visualized the features after considering frequency domain in Fig. 3, and the results show that FDC can capture lesion information well by learning in frequency domain with FEM.

For MRSC, the performance is improved by more than 1.60% in Dice and IoU scores and no trainable parameters are introduced. The best performance is achieved when FDC and MRSC are aggregated, outperforming the baseline by more than 5% in Dice and IoU scores. These above results fully demonstrate the effectiveness of our proposed method for semi-supervised medical image segmentation. Finally, please note that our proposed FDC and MRSC are only used for training. During the inference time, they can be removed and the computational complexity of inference is further reduced.

Analysis of granularities in MRSC. We study four granularity combinations in Table 3, and they are $\{16\}, \{16, 24\}, \{16, 24, 32\}$ and $\{16, 24, 32, 48\}$, respectively. The results show that MRSC with three granularities {16,24,32} can obtain the best performance. The results demonstrate that considering multiple region-level consistency can make the structural information and relationships among different image regions be effectively used, improving the feature representation ability of the model.

Analysis of region projection in MRSC. We can perform region projection in different ways, including linear interpolation, max pooling, average pooling,

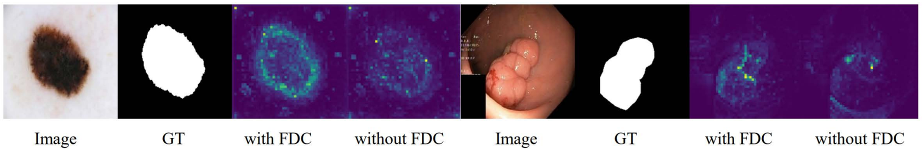

Fig. 3: The feature response with and without frequency domain consistency regularization. We can see that after frequency domain feature learning, the feature response of lesion region can be highlighted, and thus it can provide a more accurate decision boundary for the final segmentation.

and convolution with different strides. We report the influence of region projection Ψ_g in Table 3. The results show that linear interpolation can achieve the best results, and max pooling is failed, the reason behind it is that max pooling only focuses on the maximum value, ignoring regions where the feature response is small but still important for lesion segmentation. Convolution can also achieve good results, but it will bring additional parameters. Based on the results, we choose linear interpolation as our default setting.

Table 2: Ablation studies of the proposed consistency regularizationon strategies on Kvasir-SEG dataset with 10% labeled data.

Method	MAE↓	Acc↑	Dice↑	IoU↑
MT baseline	4.03	94.97	83.45	76.32
+ $\mathcal{L}_{fdc}^{low}$	3.98	95.45	86.01	79.62
+ $\mathcal{L}_{fdc}^{high}$	4.10	95.21	85.89	78.91
+ $\mathcal{L}_{fdc}$	3.98	95.98	86.84	80.93
+ $\mathcal{L}_{mrsc}$	3.95	95.84	86.17	79.74
+ $\mathcal{L}_{fdc}$ + $\mathcal{L}_{mrsc}$	**3.51**	**96.49**	**88.77**	**82.61**

Table 3: Analysis of different region granularities and projection operations in MRSC on Kvasir-SEG dataset with 10% labeled data.

Granularities	MAE↓	Acc↑	Dice↑	IoU↑
$\{16\}$	3.89	95.72	87.69	81.56
$\{16, 24\}$	3.67	96.13	88.04	82.02
$\{16, 24, 32\}$	3.51	**96.49**	**88.77**	**82.61**
$\{16, 24, 32, 48\}$	**3.34**	96.33	88.23	82.21
Projection Ψ_g	MAE↓	Acc↑	Dice↑	IoU↑
Linear	**3.51**	**96.49**	**88.77**	**82.61**
AvgPool	3.69	96.11	88.01	82.01
MaxPool	13.50	86.50	37.81	25.45
Conv	4.01	96.02	87.43	81.63

4 Conclusion

In this work, we propose a semi-supervised medical image segmentation framework FRCNet and its core components are the two consistency regularization strategies FDC and MRSC, which are plug-and-play. FRCNet can improve the model's capability for modeling consistency in both region-level and frequency domain. The results on two datasets reveal that it is critical to consider the consistency of the two domains, which allows the model to be trained efficiently with low labeled data regime and reduces the annotation cost by almost 80%. These comprehensive experimental results verify the effectiveness of our proposed model.

Acknowledgements This work is partially supported by the National Natural Science Foundation (62272248), the Huazhu Fu's Agency for Science, Technology and Research (A*STAR) Career Development Fund (C222812010) and Central Research Fund (CRF) and the Natural Science Foundation of Tianjin (23JC-QNJC00010).

Disclosure of Interests The authors have no competing interests in this paper.

References

1. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE transactions on Computers **100**(1), 90–93 (1974)
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11514–11524 (2023)
3. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. arXiv preprint arXiv:2205.13535 (2022)
4. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021)
5. Feng, Z., Zhou, Q., Gu, Q., Tan, X., Cheng, G., Lu, X., Shi, J., Ma, L.: Dmt: Dynamic mutual training for semi-supervised learning. Pattern Recognition **130**, 108777 (2022)
6. Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D.W.K., Liu, J., Cao, X.: Disc-aware ensemble network for glaucoma screening from fundus image. IEEE Transactions on Medical Imaging **37**, 2493–2501 (2018)
7. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1605.01397 (2016)
8. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., Lange, T.d., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. pp. 451–462. Springer (2020)
9. Kirkpatrick, J., Pascanu, R., Rabinowitz, N.C., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**, 3521 – 3526 (2016)
10. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 552–561. Springer (2020)
11. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
12. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems **32**(2), 523–534 (2020)

13. Liu, J., Desrosiers, C., Zhou, Y.: Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 140–150. Springer (2022)
14. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8801–8809 (2021)
15. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 318–329. Springer (2021)
16. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684 (2020)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
18. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)
19. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
21. Wei, X.S., Xu, Y.Y., Zhang, C.L., Xia, G.S., Peng, Y.X.: Cat: a coarse-to-fine attention tree for semantic change detection. Visual Intelligence **1**(1), 3 (2023)
22. Wu, H., Pan, J., Li, Z., Wen, Z., Qin, J.: Automated skin lesion segmentation via an adaptive dual attention module. IEEE Transactions on Medical Imaging **40**, 357–370 (2021)
23. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 297–306. Springer (2021)
24. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
25. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: St++: Make self-training work better for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4268–4277 (2022)
26. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019)
27. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4504–4513 (2022)

28. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)