Spatial and Planar Consistency for Semi-Supervised Volumetric Medical Image Segmentation

Yanfeng Zhou[1,2]
zhouyanfeng2020@ia.ac.cn
Yiming Huang[1,2]
huangyiming2023@ia.ac.cn
Ge Yang[1,2]
ge.yang@ia.ac.cn

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract

Semi-supervised volumetric medical image segmentation has achieved remarkable success with the development of deep neural networks (DNNs). Consistency regularization is a common strategy for semi-supervised volumetric segmentation. These models use perturbations (such as noise, distance mapping, dropout, etc.) to construct consistency losses. So far, however, few studies exploit the differences of spatial and planar information between 2D and 3D models for consistency learning. In this study, we propose a spatial and planar consistency (SPC) strategy, which outperforms previous state-of-the-art models in semi-supervised volumetric medical image segmentation. SPC consists of a 3D spatial branch and a 2D planar branch. The 3D spatial branch focuses on the complete spatial structure of segmentation objects, while the 2D planar branch focuses on the planar details. The outputs of two branches can be used for semi-supervised consistency learning. Extensive experiments on two public 3D datasets demonstrate the effectiveness of our model. Code is available at https://github.com/Yanfeng-Zhou/SPC.

1 Introduction

Semantic segmentation is a fundamental task in medical image analysis, where the goal is to assign a class label to each pixel (voxel). Medical image semantic segmentation has been advanced with the development of DNNs [13, 20, 43, 44]. Some studies extend methods to 3D and achieve promising results on volumetric segmentation [7, 17, 28, 42, 43].

Fully-supervised models require large-scale labeled images, which are costly and time-consuming to produce. To alleviate this problem, researchers propose semi-supervised models that learn with with a small number of labeled images and a substantial number of unlabeled images [23, 25, 34]. The common solutions include adversarial training [18, 22], pseudo-labeling [4, 37], consistency regularization [1, 19] and contrastive learning [27, 29].

Current semi-supervised volumetric medical image segmentation models rely primarily on consistency regularization [10, 12]. These models construct consistency losses through

various strategies, such as adding noise to input patches [35], predicting additional distance maps [14], perturbing network structures and parameters [8], etc.

However, few studies consider exploiting potential differences between 3D and 2D models to construct consistency losses. To be specific, 3D models can effectively extract the spatial structure information of segmentation objects, but also ignore some details of each plane. While 2D models lack complete 3D view, but can focus on planar details (The quantitative comparison in Figure 2 intuitively shows the differences between 2D and 3D models). Therefore, 3D and 2D models can exploit the complementary consistency of learning information for semi-supervised training.

In this study, we propose a spatial and planar consistency (SPC) strategy for semi-supervised volumetric medical image segmentation. SPC consists of two branches, one is 3D spatial branch and the other is 2D planar branch. The 3D spatial branch uses volume patches as input to directly produce corresponding volume segmentation predictions. The 2D planar branch first splits the volume patches into a series of 2D slices, then feeds these slices into 2D model to generate segmentation predictions for each slice and restores these slices to the corresponding volume segmentation predictions. The 3D spatial branch focuses on the spatial structure, while the 2D planar branch focuses on the local details. The complementary consistency of two branches can be used for semi-supervised learning. Our model achieves state-of-the-art in semi-supervised volumetric medical image segmentation. Extensive benchmarking on two public 3D datasets demonstrate the effectiveness of our model.

Motivation. For volumetric semantic segmentation, there is a debate: which is better, 3D or 2D models? The 3D model can capture the complete spatial structure information of the segmentation objects, but it has a large number of parameters and requires patch-based training, which leads to convergence difficulty and overfitting. In contrast, 2D models are easier to train, but the receptive field is limited to single plane. Therefore, as in some related studies [16, 36], a clever combination of them can achieve better results. But different from previous studies, we focus on semi-supervised semantic segmentation. We exploit the complementary consistency of spatial and planar information between 3D and 2D to improve the model performance. Despite its simplicity, it has demonstrated competitive results compared to the state-of-the-art models, as discovered in our experimental study.

2 Related Work

Volumetric Medical Image Segmentation. For medical image segmentation, efficient encoder-decoder architecture achieves superior performance, such as UNet [20], UNet++ [44], UNet 3+ [9], etc. Some studies extend these architectures to 3D to meet the needs for volumetric segmentation. [17] proposes a 3D fully convolutional neural network (CNN) VNet. [2] extends UNet to 3D. ConResNet [38] proposes inter slice context residual learning. Recently, several volume segmentation methods incorporated transformers and CNNs to achieve superior results [7, 28, 32]. Transformers can capture long range dependencies [6, 24] to compensate for the limited receptive fields of CNNs. UNETR [7] uses a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information. [32] proposes CoTr to efficiently bridge a CNN and a Transformer. nnFormer [42] exploits the combination of interleaved convolution and self-attention operations, and introduces local and global volume-based self-attention mechanism to learn volume representations.

Semi-Supervised Semantic Segmentation. To alleviate the lack of labeled images, semi-

supervised semantic segmentation has become popular. The current dominant strategies include adversarial training [18, 22], pseudo-labeling [6, 34], consistency regularization [1, 19] and contrastive learning [26, 39, 40, 41]. Adversarial training use generative adversarial networks [6] to continuously improve the performance of both generator that generates segmentation predictions and discriminator that judges the authenticity of predictions. Pseudo-labeling utilizes high confidence predictions to improve model performance, such as DMT [6], ST++ [34], etc. Consistency regularization [1, 19, 23] enhances the learning from unlabeled images by enforcing consistency between different predictions. [23] proposes Mean-Teacher (MT) architecture that uses teacher model to generate pseudo-labels to guide student model and updates teacher model with exponential moving average (EMA). [1] proposes a novel consistency regularization approach, called cross pseudo supervision (CPS). CCT [19] uses multiple decoders and adds different perturbations to the decoders, then enforces prediction invariance on the decoder outputs. Contrastive learning [39, 40, 41] uses input images or intermediate features to generate positive and negative pairs, and then trains the model to pull positive pairs closer while push negative pairs away, which enables the model to extract latent features from sample pairs, such as PC^{2}Seg [41], Semi-CML [39], RCPS [40], etc.

Semi-Supervised Volumetric Medical Image Segmentation. This is a research branch of semi-supervised semantic segmentation that focuses on 3D volumetric medical images. The common strategy is perturbation-based consistency regularization [30, 31], which perturbs input images, intermediate features, network architecture or output predictions, allowing models to learn consistency from the perturbation. [15] proposes an uncertainty rectified pyramid consistency (URPC) strategy to perform consistent regularization on output predictions at different scales. CC-Net [8] uses a main model and two auxiliary models, and performs complementary consistency training with the disturbance between the main model and the auxiliary models. HRC-MT [11] proposes multi-scale deep supervision and hierarchical consistency regularization. The current strategy for generating consistent predictions is overly rigid and forced. In contrast, we exploit differences in the features of 3D and 2D networks to generate consistency predictions. Our approach is simpler and more natural, but also achieves competitive results.

Model	Strategy	Strength ↑ and Weakness ↓
MT	Noise perturbation of input images	↑ Simple, easy to implement ↓ Noise makes learning difficult
DTC SASSNet	Predict additional distance maps	↑ Only one network can output both segmentation predictions and distance maps. ↑ High computational efficiency. ↓ The generation of distance maps for multi-class segmentation is complicated
URPC HRC-MT	Consistency of multiple upsampled low-size segmentation predictions	↑ Efficient generation of low-size segmentation predictions ↓ Excessive upsampling ratio may cause distortion ↓ The loss function is more complex
MC-Net+ CC-Net	Three network output consistency	↑ Pseudo-labels are more reliable and accurate ↓ Computationally expensive
SPC (Ours)	Attention differences of 2D and 3D networks	↑ Take advantage of 2D and 3D networks simultaneously ↓ Only applicable to 3D volumetric segmentation

Table 1: Comparison of different consistency regularization strategies.

Consistency Regularization Strategies. We summarize the common consistency regularization strategies for semi-supervised models, including MT [23], DTC [14], SASSNet [10], URPC [15], HRC-MT [11], MC-Net+ [31] and CC-Net [8]. We also compare the strength and weakness of these strategies. The comparison results are shown in Table 1.

3 Method

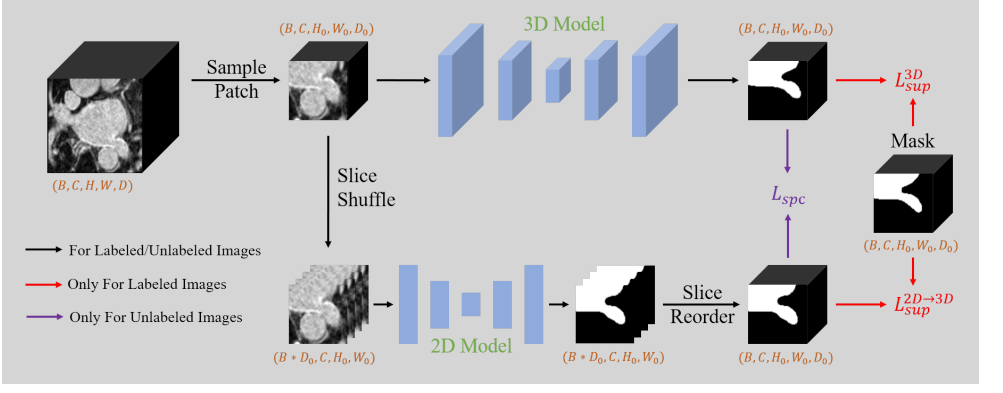

Figure 1: The overview of our proposed model SPC. SPC consists of a 3D spatial branch and a 2D planar branch. SPC is optimized by minimizing supervised loss L_{sup}^{3D} and $L_{sup}^{2D\to3D}$ on labeled images, and spatial and planar consistency loss L_{spc} on unlabeled images.

Figure 1 shows an overview of our model SPC. SPC consists of two branches: 3D spatial branch and 2D planar branch. 3D spatial branch focuses on the complete spatial structure of segmentation objects and uses volume patches as input to directly produce corresponding volume segmentation predictions by 3D model. While 2D planar branch focuses on planar details. It first splits the volume patches into a series of 2D slices and shuffles the input order of these slices. Then these slices are fed into 2D model as a batch to generate segmentation predictions. Finally, these segmentation predictions of 2D slices are reordered and restored to the corresponding volumetric segmentation predictions. The segmentation predictions of two branches are used for semi-supervised training. In this study, we use the common UNet [20] and 3D UNet [2] as 2D and 3D models, respectively.

We take a 3D volume with the number of channels, height, width and depth as C, H, W and D as an example to show the training process. We set the training batch size as B. We first sample (C,H_0,W_0,D_0) patches from the raw 3D volume, where H_0, W_0 and D_0 represent the height, width and depth of patches, respectively. We feed a batch of patches into 3D model to generate the corresponding 3D segmentation predictions p_{3D}. Simultaneously, we split and shuffle input patches into $(B\times D_0,H_0,W_0)$ 2D slices, where $B\times D_0$ represents batch size of 2D slices. We feed these slices into 2D model to generate corresponding 2D segmentation predictions $\{p_{2D}^i\}_{i=1}^{B\times D_0}$. Then We reorder and restore $\{p_{2D}^i\}_{i=1}^{B\times D_0}$ into 3D segmentation predictions $p_{2D\to3D}$. Our model learns from labeled images by minimizing supervised loss L_{sup} and learns from unlabeled images by minimizing spatial and planar consistency loss L_{spc}.

The total loss L_{total} is defined as:

$$L_{total} = L_{sup} + \lambda L_{spc}, \tag{1}$$

where λ is a weight to control the balance between L_{sup} and L_{spc}. λ increases linearly with training epochs:

$$\lambda = \lambda_{max} * \frac{epoch}{max\_epoch}, \tag{2}$$

where λ_{max} is the maximum weight, $epoch$ represents training epoch, max_epoch represents the maximum training epoch. We compare the performance of different λ_{max} in ablation studies of Section 4.5.

The supervised loss L_{sup} is defined as:

$$L_{sup} = L_{sup}^{3D}(p_{3D}, y_{3D}) + L_{sup}^{2D\to 3D}(p_{2D\to 3D}, y_{3D}), \tag{3}$$

where y_{3D} represents the ground truth of patches.

The spatial and planar consistency loss L_{spc} is achieved by CPS loss [1]: Use one prediction as pseudo-label to supervise the other, and vice versa. L_{spc} is defined as:

$$L_{spc} = L_{spc}^{3D}(p_{3D}, \hat{p}_{2D\to 3D}) + L_{spc}^{2D\to 3D}(p_{2D\to 3D}, \hat{p}_{3D}), \tag{4}$$

where $\hat{p}_{2D\to 3D}$ and $\hat{p}_{3D}$ represent pseudo-labels generated by $p_{2D\to 3D}$ and p_{3D}, respectively. We compare the performance of L_{sup} and L_{spc} with different loss functions in ablation studies of Section 4.5.

The specific training process is shown in Algorithm 1.

4 Experiments

4.1 Datasets

We evaluate our model on two public 3D datasets (LA [33] and P-CT [21]).

LA. This is a left atrial dataset from 2018 Atrial Segmentation Challenge. It consists of 100 3D MRI images, with a resolution of 0.625×0.625×0.625mm. Following [8, 14], we use 80 images for training and 20 images for testing and all images are cropped centering at the heart region.

P-CT. This is pancreas dataset, which includes 82 abdominal contrast-enhanced CT images. These images are collected on Philips and Siemens MDCT scanners and have a fixed resolution of 512×512 with varying thicknesses from 1.5 to 2.5 mm. Following [14, 31], we use the soft tissue CT window range of [-125, 275] HU, resample all images to an isotropic resolution of 1.0×1.0×1.0mm and crop the images centering at the pancreas region. We use 62 images for training and 20 images for testing.

4.2 Evaluation Metrics

We use Dice coefficient (Dice), Jaccard index (Jaccard), 95th percentile Hausdorff distance (95HD), and average surface distance (ASD) as evaluation metrics. Dice and Jaccard emphasize pixel-wise accuracy, while 95HD and ASD emphasize boundary accuracy. These metrics are widely used for benchmarking performance of biomedical image segmentation.

Algorithm 1: Training process

Input:

Label set $D^l = \{(X^i_{3D}, Y^i_{3D})\}_{i=1}^{M}$, where M represents the number of labeled images

Unlabel set $D^u = \{X^j_{3D}\}_{j=1}^{N}$, where N represents the number of unlabeled images

Parameter:

Input Channel C

Batch size B

Patch size (H_0, W_0, D_0)

Maximum weight λ_{max}

Maximum training epoch max_epoch

2D model F_{2D}, 3D model F_{3D}

Output:

Fully trained F_{2D} and F_{3D}

---

for $epoch = 1$ **to** max_epoch **do**

  $L_{sup}, L_{spc} = 0$

  for (X^i_{3D}, Y^i_{3D}) **in** D^l **do**

    $(X_{3D}, Y_{3D}) \leftarrow$ Group B (X^i_{3D}, Y^i_{3D}) into a batch // $X_{3D}, Y_{3D} \in R^{(B,C,H,W,D)}$

    $(x_{3D}, y_{3D}) \leftarrow$ Sample patches from (X_{3D}, Y_{3D}) // $x_{3D}, y_{3D} \in R^{(B,C,H_0,W_0,D_0)}$

    $x_{2D} \leftarrow$ Split and shuffle x_{3D} // $x_{2D} \in R^{(B\times D_0,C,H_0,W_0)}$

    $p_{3D} \leftarrow F_{3D}(x_{3D})$ // $p_{3D} \in R^{(B,C,H_0,W_0,D_0)}$

    $p_{2D} \leftarrow F_{2D}(x_{2D})$ // $p_{2D} \in R^{(B\times D_0,C,H_0,W_0)}$

    $p_{2D\to 3D} \leftarrow$ Reorder and restore p_{2D} // $p_{2D\to 3D} \in R^{(B,C,H_0,W_0,D_0)}$

    $L_{sup} += L^{3D}_{sup}(p_{3D}, y_{3D}) + L^{2D\to 3D}_{sup}(p_{2D\to 3D}, y_{3D})$

  end

  for X^j_{3D} **in** D^u **do**

    $X_{3D} \leftarrow$ Group B X^j_{3D} into a batch // $X_{3D} \in R^{(B,C,H,W,D)}$

    $x_{3D} \leftarrow$ Sample patches from X_{3D} // $x_{3D} \in R^{(B,C,H_0,W_0,D_0)}$

    $x_{2D} \leftarrow$ Split and shuffle x_{3D} // $x_{2D} \in R^{(B\times D_0,C,H_0,W_0)}$

    $p_{3D} \leftarrow F_{3D}(x_{3D})$ // $p_{3D} \in R^{(B,C,H_0,W_0,D_0)}$

    $p_{2D} \leftarrow F_{2D}(x_{2D})$ // $p_{2D} \in R^{(B\times D_0,C,H_0,W_0)}$

    $p_{2D\to 3D} \leftarrow$ Reorder and restore p_{2D} // $p_{2D\to 3D} \in R^{(B,C,H_0,W_0,D_0)}$

    $\hat{p}_{3D}, \hat{p}_{2D\to 3D} \leftarrow p_{3D}, p_{2D\to 3D}$ // $\hat{p}_{3D}, \hat{p}_{2D\to 3D} \in R^{(B,C,H_0,W_0,D_0)}$

    $L_{spc} += L^{3D}_{spc}(p_{3D}, \hat{p}_{2D\to 3D}) + L^{2D\to 3D}_{spc}(p_{2D\to 3D}, \hat{p}_{3D})$

  end

  $\lambda \leftarrow \lambda_{max} * \frac{epoch}{max\_epoch}$

  $L_{total} \leftarrow L_{sup} + \lambda L_{spc}$

  Update F_{3D} and F_{2D} to minimize L_{total}

end

return F_{3D}, F_{2D}

Dataset	Model	# Labeled	# Unlabeled	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
LA	MT [23]	16	64	88.23	79.29	2.73	10.64
	SASSNet [10]	16	64	89.17	80.69	2.86	8.57
	DTC [14]	16	64	89.43	81.00	2.12	7.39
	MC-Net [30]	16	64	90.12	82.12	1.99	8.07
	MC-Net+ [31]	16	64	91.05	83.64	1.69	5.81
	TraCoCo [12]	16	64	90.94	83.47	1.79	**5.49**
	CC-Net [8]	16	64	**91.27**	**84.02**	**1.54**	5.75
	SPC (Ours)	16	64	**92.52**	**86.08**	**1.40**	**4.59**
P-CT	MT [23]	12	50	76.79	62.33	2.94	10.97
	EM [25]	12	50	75.98	61.26	3.77	12.80
	UAMT [35]	12	50	77.14	62.79	3.85	14.91
	SASSNet [10]	12	50	77.81	63.67	3.06	9.15
	DTC [14]	12	50	78.25	64.26	2.14	**7.17**
	MC-Net [30]	12	50	77.71	63.54	2.74	9.02
	MC-Net+ [31]	12	50	**78.87**	**65.11**	**1.89**	8.15
	SPC (Ours)	12	50	**79.82**	**66.42**	**1.83**	**6.68**

Table 2: Comparison with state-of-the-art models on LA and P-CT test set. All models are trained with 20% labeled images and 80% unlabeled images, which is the common semi-supervised experimental partition. **Red** and **bold** indicate the best and second best performance.

4.3 Implementation Details

We implement our model using PyTorch. Training and inference of all models are performed on four NVIDIA GeForce RTX3090. We use SGD with momentum to train models, the momentum is set at 0.9, and the weight decay is set at 0.00005. Batch size is set at 4. The number of epochs is set at 200. The initial learning rate is set at 0.5 and the learning rate decays by 0.5 every 50 epochs. All models are trained with 20% labeled images and 80% unlabeled images, which is the common semi-supervised experimental partition. For training, we use flip, biasfield, noise and blur for data augmentation. Following [8, 14, 31], patch size is set at 112×112×80 and 96×96×96 for LA and P-CT, respectively. For inference, following previous studies, we use a sliding window strategy to obtain complete segmentation results, with strides of 18×18×4 and 16×16×16 for LA and P-CT, respectively.

4.4 Comparison with State-of-the-art Models

We compare our model with previous state-of-the-arts, including MT [23], UAMT [35], EM [25], SASSNet [10], DTC [14], MC-Net [30], MC-Net+ [31], TraCoCo [12] and CC-Net [8]. From Table 2, we can see that our model outperforms previous state-of-the-art models by a large margin. Our model can simultaneously focus on spatial structure and planar details, 2D and 3D networks optimize each other with the pseudo-labels produced by each other and finally achieve better pixel-wise accuracy (Dice and Jaccard) and boundary contours (ASD and 95HD). A more precise organ structure is critical for organ localization and surgical planning.

4.5 Ablation Studies

To verify effectiveness of each component, we perform the following ablation studies on LA and P-CT.

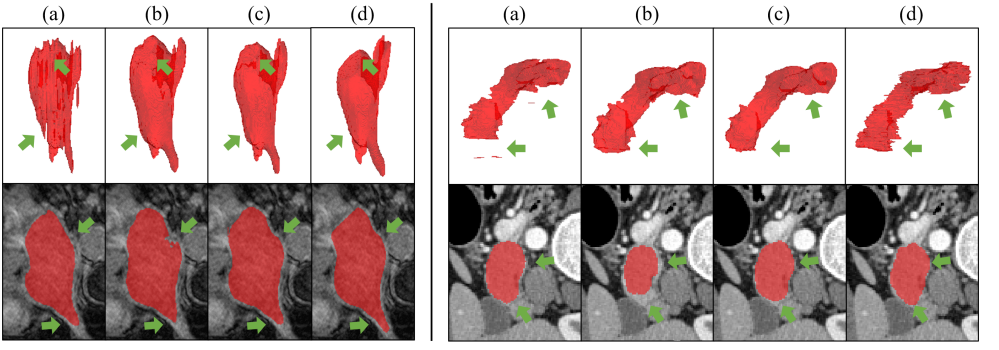

Figure 2: Qualitative comparison of 2D and 3D models on LA (left) and P-CT (right). (a) two 2D models. (b) two 3D models. (c) SPC. (d) Ground truth. The green arrows highlight the difference among of the results.

Characteristics of 2D and 3D models. We compare the qualitative results of two 2D models, two 3D models, one 2D model and one 3D model (SPC). The results are shown in Figure 2. We find that 2D model cannot predict the overall contours of segmentation objects very well and it also misses some parts of the segmentation objects or predicts redundant small regions, but it is more accurate for some slices. This is because the 2D models lack full 3D view, but can better focus on each plane slice. Segmentation predictions of 3D models have smoother overall contours but less accuracy in slice details. This is because 3D models can effectively extract the spatial structure information, but ignore planar details. While our model can simultaneously focus on spatial structure and planar details to achieve better performance.

Dataset	2D	3D	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
LA	✓		85.31	74.39	2.36	8.17
		✓	89.79	81.48	2.03	6.86
	✓	✓	**92.52**	**86.08**	**1.40**	**4.59**
P-CT	✓		75.30	60.39	3.45	17.33
		✓	78.70	64.88	2.03	8.76
	✓	✓	**79.82**	**66.42**	**1.83**	**6.68**

Table 3: Ablation on effectiveness of various components, including using two 2D models, two 3D models, one 2D model and one 3D model.

Effectiveness of spatial and planar consistency. We compare the quantitative performance of two 2D models, two 3D models and SPC. The results are shown in Table 3. compared to using only 2D models or 3D models, using both 2D and 3D models for complementary consensus learning achieves the best performance. To be specific, for LA, SPC improves the performance to 92.52% in Dice, 86.08% in Jaccard, 1.40 voxels in ASD and

4.59 voxels in 95HD. For P-CT, SPCi mproves the performance to 79.82% in Dice, 66.42% in Jaccard, 1.83 voxels in ASD and 6.68 voxels in 95HD.

Dataset	λ_{max}	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
LA	1	91.49	84.32	1.63	5.35
	3	91.89	85.00	1.54	5.16
	5	**92.52**	**86.08**	**1.40**	**4.59**
	7	92.00	85.18	1.51	5.13
P-CT	0.5	76.53	61.99	2.79	11.63
	1	**79.82**	**66.42**	**1.83**	**6.68**
	3	79.46	65.92	1.85	7.41

Table 4: Comparison of different λ_{max} for LA and P-CT.

Comparison of λ_{max}. The comparison results are shown in Table 4. For LA, segmentation is relatively easy, λ should increase faster (i.e., λ_{max} is large) to highlight the role of many unlabeled images to prevent overfitting. For P-CT, segmentation is more difficult, λ should change smoothly (i.e., λ_{max} is small), so that the model can better use the labeled images in the early training stage and further improve from unlabeled images in the later training stage. We finally set λ_{max} are 5 and 1 for LA and P-CT, respectively, and apply it to related experiments in Table 2.

Comparison of Loss Functions. Dice and cross entropy (CE) are common segmentation loss functions, and we compare their performance in Table 5. We find that for our model, CE loss achieves better performance. Based on the above experiments, we apply CE loss to related experiments in Table 2.

Dataset	L_{sup}	L_{spc}	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
LA	Dice	Dice	92.10	85.36	1.45	**4.52**
	Dice	CE	91.79	84.82	1.52	4.89
	CE	Dice	91.41	84.18	1.66	5.71
	CE	CE	**92.52**	**86.08**	**1.40**	4.59
P-CT	Dice	Dice	79.61	66.13	**1.81**	7.13
	Dice	CE	79.09	65.41	1.95	6.71
	CE	Dice	78.18	64.18	1.97	7.58
	Dice	Dice	**79.82**	**66.42**	1.83	**6.68**

Table 5: Comparison of different loss functions for L_{sup} and L_{spc}, including Dice and CE loss.

4.6 Qualitative Results

Figure 3 shows some qualitative results of different models, including MT [23], DTC [14], MC-Net [30], MC-Net+ [31] and SPC. Compared with other models, SPC combines the advantages of 2D and 3D models to achieve better spatial structure and planar details.

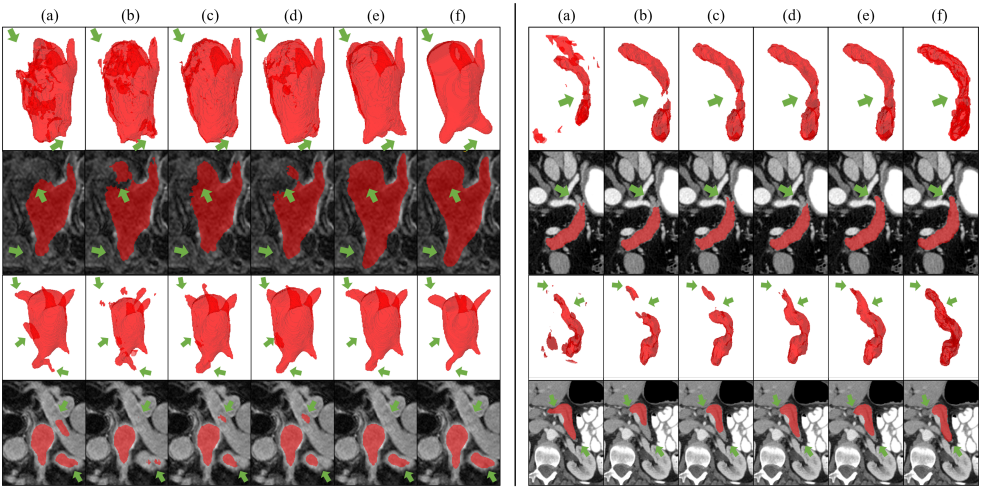

Figure 3: Qualitative results on LA (left) and P-CT (right). (a) MT. (b) DTC. (c) MC-Net. (d) MC-Net+. (e) SPC. (f) Ground truth. The green arrows highlight the difference among of the results.

5 Conclusion

We propose a spatial and planar consistency model, which achieves state-of-the-art in semi-supervised volumetric medical image segmentation. Our model consists of a 3D spatial branch and a 2D planar branch. The 3D spatial branch focuses on the complete spatial structure of segmentation objects, while the 2D planar branch focuses on the planar details. Two-branch complementary consistency can improve semi-supervised learning performance. Extensive experiments on two public 3D datasets demonstrate the effectiveness of our model.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (grants 31971289, 91954201) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant XDB37040402).

References

[1] 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*, pages 2613–2622, 2021.

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. *arXiv preprint arXiv:2004.08514*, 1(2):5, 2020.

[5] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, page 108777, 2022.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing systems*, 27, 2014.

[7] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[8] Hejun Huang, Zuguo Chen, Chaoyang Chen, Ming Lu, and Ying Zou. Complementary consistency semi-supervised learning for 3d left atrial image segmentation. *arXiv preprint arXiv:2210.01438*, 2022.

[9] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[10] 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*, pages 552–561. Springer, 2020.

[11] Shumeng Li, Ziyuan Zhao, Kaixin Xu, Zeng Zeng, and Cuntai Guan. Hierarchical consistency regularized mean teacher for semi-supervised 3d left atrium segmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3395–3398. IEEE, 2021.

[12] Yuyuan Liu, Yu Tian, Chong Wang, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Translation consistent semi-supervised segmentation for 3d medical images. *arXiv preprint arXiv:2203.14523*, 2022.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[14] 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*, volume 35, pages 8801–8809, 2021.

[15] Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517, 2022.

[16] Peiqing Lv, Jinke Wang, and Haiying Wang. 2.5 d lightweight riu-net for automatic liver and tumor segmentation from ct. *Biomedical Signal Processing and Control*, 75: 103567, 2022.

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[18] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[19] 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*, pages 12674–12684, 2020.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[21] 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*, pages 556–564. Springer, 2015.

[22] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

[23] 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*, 30, 2017.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[26] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79:102447, 2022.

[27] Tao Wang, Jianglin Lu, Zhihui Lai, Jiajun Wen, and Heng Kong. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 1444–1450, 2022.

[28] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[29] Huisi Wu, Zhaoze Wang, Youyi Song, Lin Yang, and Jing Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11666–11675, 2022.

[30] 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*, pages 297–306. Springer, 2021.

[31] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022.

[32] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

[33] 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*, 67:101832, 2021.

[34] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022.

[35] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

[36] Qihang Yu, Yingda Xia, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Thickened 2d networks for 3d medical image segmentation. *arXiv preprint arXiv:1904.01150*, 2019.

[37] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8229–8238, 2021.

[38] Jianpeng Zhang, Yutong Xie, Yan Wang, and Yong Xia. Inter-slice context residual learning for 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):661–672, 2020.

[39] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, and Zhenghua Xu. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 83:102656, 2023.

[40] 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. *arXiv preprint arXiv:2301.05500*, 2023.

[41] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021.

[42] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

[43] Yanfeng Zhou, Jiaxing Huang, Chenlong Wang, Le Song, and Ge Yang. Xnet: Wavelet-based low and high frequency fusion networks for fully- and semi-supervised semantic segmentation of biomedical images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21085–21096, October 2023.

[44] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019.