Efficient Semi-Supervised Multi-Organ Segmentation Using Uncertainty Rectified Pyramid Consistency

No Author Given

No Institute Given

Abstract. Although many automatic multi-organ segmentation methods based on deep learning have been proposed, the problem of their great dependence on fully annotated data and spatio-temporal inefficiency still remain unsolved. To meet these problems, an efficient semi-supervised framework with uncertainty rectified pyramid consistency regularization is proposed. Specifically, inspired by the fact that the predictions of the same input should be similar under different disturbance, we extend a backbone to produce pyramid predictions for unlabeled images and encourage them to be consistent. Since the multi-scale predictions have different resolution, directly encouraging them to be consistent may bring problems including lost of fine detail or model collapse. So a rectified scale-level uncertainty-aware module is introduced to enable the framework to gradually learn from reliable prediction regions. To deal with the domain gaps among multi-center datasets, a number of prepocessing methods are utilized, such as resampling the multi-center CT volumes to the same spacing and adjusting the window level and width. The proposed method has been evaluated on the FLARE2022 validation cased, and get a relatively good segmentation performance.

Keywords: Semi-supervised learning · Multi-organ segmentation · Uncertainty rectifying · Pyramid consistency.

1 Introduction

Whole abdominal organ segmentation plays an important role in the diagnosis of abdominal lesions, radiotherapy and follow-up. Manual organ delineation is time-consuming and error-prone[7]. Although many automatic segmentation methods based on deep learning have achieved good performances in abdominal organ segmentation, most of them rely heavily on large-scale fully annotated data, which is often difficult to obtain due to cost and privacy issues. Additionally, these methods are often difficult to be implemented into clinical practice due to the large model size and the extensive computational resources.

To meet the needs of fast inference and low computational cost while only use a small number of labeled cases, we develop an efficient semi-supervised framework with uncertainty rectified pyramid consistency to fully make use of the unlabeled data. Concretely, we extend a U-Net[9] backbone to produce pyramid

predictions at different scales and encourages these multi-scale predictions to be consistent for a given input. A standard supervised loss is used for learning from labeled data. For unlabeled cases, we encourages the pyramid predictions to be consistent, which served as a regularization. Since the predictions of the unlabeled cases are not necessarily reliable because of its insufficient supervision information, which may cause the model to collapse and lose details[6], we propose to estimate the uncertainty via the prediction discrepancy among multi-scale predictions. Different from those that estimate the uncertainty of each target prediction with Monte Carlo sampling[5], which needs massive computational costs as it requires multiple forward passes to obtain the uncertainty in each iteration, our proposal just needs a single forward pass. Under the guidance of the estimated uncertainty, the model strengthens the consistent learning of reliable regions and weakens that of unreliable ones. Meanwhile, to further improve the utilization of unlabeled data information and increase the segmentation efficiency, we introduce the uncertainty minimization[11] to reduce the prediction variance during training.

To address the domain gaps among FLARE2022 training, validation, and test sets, we employ a series of data preprocessing methods. For instance, resampling the images' spacing to $2 \times 2 \times 2.5$mm, adjusting the window level and window width of the CT images, data augmentation with CLAHE, gamma correction, random noise, random rotate and random flip.

2 Method

The overview of the proposed semi-supervised segmentation network is illustrated in Fig.1, which consists of a 3D U-Net[9] backbone with a pyramid prediction structure at the decoder and an uncertainty rectifying module.

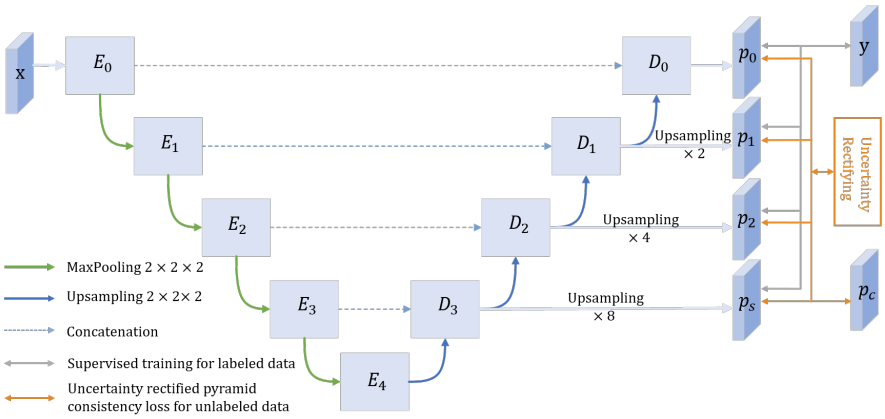

Fig. 1. Network architecture.

2.1 Multi-Scale Prediction Network with Pyramid Consistency

Firstly, we introduce pyramid prediction network (PPNet) for the multi-organ segmentation task, which can produce prediction at different scales. In this work, a 3D U-Net is employed as backbone and is modified to produce pyramid predictions by adding a prediction layer after each upsampling block in the decoder, where the prediction layer is implemented by $1 \times 1 \times 1$ convolution followed by a softmax layer. A dropout layer and a feature-level noise addition layer are inserted before the prediction layer to introduce more perturbations to the network.

For an input image x_i, the PPNet produces a set of multi-scale predictions $[p'_0, p'_1, \cdots, p'_s, \cdots, p'_{S-1}]$,where S means the number of the scales, p'_s means the prediction at scale s. A smaller s presents a higher resolution in the decoder. In our work, we set $S = 4$ for the multi-organ segmentation. Then, the multi-scale predictions are rescaled to the input size, and the the corresponding results are denoted as $[p'_0, p_1, \cdots, p_s, \cdots, p_{S-1}]$. For the labeled data, we use a combination of Dice and cross-entropy loss at multiple scales as the supervised loss:

$$\mathcal{L}_{sup} = \frac{1}{S} \sum_{s=0}^{S-1} \frac{\mathcal{L}_{dice}(p_s, y_i) + \mathcal{L}_{ce}(p_s, y_i)}{2} \tag{1}$$

where $y_i, \mathcal{L}_{dice}, \mathcal{L}_{ce}$ denote the ground truth of input x_i, the Dice loss and the cross entropy loss, respectively.

To efficiently leverage unlabeled data, we design a pyramid consistency loss to encouraging the multi-scale predictions to be consistent by minimizing their discrepancy(i.e., variance) with L_2 distance:

$$\mathcal{L}_{pyc} = \frac{1}{S} \sum_{s=0}^{S-1} \|p_s - p_c\|_2 \tag{2}$$

where p_c is the average prediction across scales, which is denoted as:

$$p_c = \frac{1}{S} \sum_{s=0}^{S-1} p_s \tag{3}$$

2.2 Uncertainty Rectified Pyramid Consistency Loss

As the pyramid predictions have different spatial resolutions, imposing voxel-level consistency directly between these predictions may lead to problems due the different spatial frequencies, such as lost of fine detail or model collapse. Inspired by existing works [1,11], a scale-level uncertainty-aware method is introduced to address these problems, which only requires a single forward pass and thus need less computational cost and running time than exiting methods[5].

We use the KL-divergence between the prediction at scale s and the mean prediction as the uncertainty measurement:

$$\mathcal{D}_s \approx \sum_{j=0}^{C} p_s^j \cdot \log \frac{p_s^j}{p_c^j} \tag{4}$$

where $C = 14$ is the class(i.e., channel) number in this work, and p_s^j is the jth channel of p_s. We can get a set of uncertainty maps $\mathcal{D}_0, \mathcal{D}_1, \cdots, \mathcal{D}_s, \cdots, \mathcal{D}_{S-1}$, where $\mathcal{D}_s$ represents the uncertainty of p_s. A larger value of $\mathcal{D}_s$ indicates the prediction for that pixel at scale s has high uncertainty, which means the prediction is unreliable and may be ignored for stable unsupervised learning.

Based on the uncertainty maps, we automatically select reliable voxels(with low uncertainty) to rectify the pyramid consistency loss for better using the information of unlabeled data:

$$\mathcal{L}_{unsup} = \underbrace{\frac{1}{S}\frac{\sum_{s=0}^{S-1}\sum_v (p_s^v - p_c^v)^2 \cdot w_s^v}{\sum_{s=0}^{S-1}\sum_v w_s^v}}_{uncertaintyrectification} + \underbrace{\frac{1}{S}\sum_{s=0}^{S-1}\|\mathcal{D}_s\|_2}_{uncertaintyminimization} \tag{5}$$

where p_s^v and $\mathcal{D}_s^v$ are the corresponding prediction and uncertainty values for voxel v. Rather than use the threshold-based cut off approaches which is hard to determine[1], we follow the policy in [11] to use a voxel- and scale-wise weight w_s^v to automatically rectify the MSE loss, which is defined as: $w_s^v = e^{-\mathcal{D}_s^v}$, where a higher uncertainty leads to lower weight. Meanwhile, in order to encourage the PPNet to produce more consistent predictions at different scales, we use the uncertainty minimization term as a constraint.

2.3 The Overall Loss Function

The proposed semi-supervised segmentation network learns form both labeled data and unlabeled data by minimizing the following combined objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda \cdot \mathcal{L}_{unsup} \tag{6}$$

where λ is a widely-used time-dependent Gaussian warming up function [5] to control the balance between the supervised loss and the unsupervised loss. The formula of $\lambda(t)$ is: $\lambda(t) = w_{max} \cdot e^{(} - 5(1 - \frac{1}{t_{max}})^2$, where w_{max} means the final regularization weight, t denotes the current training step and t_{max} is the maximal training step.

2.4 Preprocessing

To reduce the domain gaps between multi-center data, the following preprocessing techniques are employed:

- Resampling the anisotropic data:
  Due to differences in scanners or acquisition protocols, data from different centers usually have different spacing. The convolution operation of CNN requires the image to be isotropic for better feature extraction. So all data, including training, validation and test data, need to be resampled to keep the image isotropic. Observing the resolution information of the training data, combined with the consideration of the trade-off between the amount of

contextual information in the networks patch size and the details retained in the image data, the sampling layer spacing is set to 2.5mm. To save memory overhead and speed up training, the intra-layer resolution is resampled to 2mm × 2mm.

- Adjusting window level and window width:
  The method of adjusting the window width and window level was leveraged to improve the contrast of abdominal CT images. And the window level and width were adjusted 50 and 400, respectively.
- Intensity normalization method:
  The data were normalized with z-score normalization based on the mean and standard deviation of the intensity to avoid the problem of data being compressed after normalization.
- Data augmentation method:
  Due to the lack of label data and to avoid the problem of overfitting, channel-wise gamma correction, random noise, random flip, random rotate were used to augment the training data.

2.5 Proposed Method

- Network architecture details: The network architecture is shown in Fig.1.
- Loss function: the loss function is shown as Formula 6
- Number of the parameters of PPNet: 1330104.

2.6 Post-processing

A connected component analysis of segmentation mask is applied on the outputs to remove small connected areas. And then the results are resampled back to original spacing for the convenience of the following evaluation.

3 Experiments

3.1 Dataset and evaluation measures

The FLARE2022 dataset is curated from more than 20 medical groups under the license permission, including MSD [10], KiTS [3,4], AbdomenCT-1K [8], and TCIA [2]. The training set includes 50 labelled CT scans with pancreas disease and 2000 unlabelled CT scans with liver, kidney, spleen, or pancreas diseases. The validation set includes 50 CT scans with liver, kidney, spleen, or pancreas diseases. The testing set includes 200 CT scans where 100 cases has liver, kidney, spleen, or pancreas diseases and the other 100 cases has uterine corpus endometrial, urothelial bladder, stomach, sarcomas, or ovarian diseases. All the CT scans only have image information and the center information is not available.

The evaluation measures consist of two accuracy measures: Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), and three running efficiency measures: running time, area under GPU memory-time curve, and area under CPU utilization-time curve. Moreover, the GPU memory consumption has a 2 GB tolerance.

3.2 Implementation details

Environment settings The development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements.

Windows/Ubuntu version	Ubuntu 20.04.4 LTS
CPU	Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz
RAM	16×4GB; 2.67MT/s
GPU (number and type)	Four NVIDIA Corporation TU102(2080Ti) 10G
CUDA version	11.4.48
Programming language	Python 3.7.13
Deep learning framework	Pytorch (Torch 1.11.0, torchvision 0.12.0)
Specific dependencies	None

Training protocols The training protocols of the baseline method is shown in Table 2.

Table 2. Training protocols.

Network initialization	“he" normal initialization
Batch size	4
Patch size	80×96×96
Total epochs	100
Optimizer	Adam with momentum ($\mu = 0.9$)
Initial learning rate (lr)	0.001
Lr decay schedule	halved by 20 epochs
Training time	24 hours
Number of model parameters	15.3M[1]
Number of flops	59.32G[2]
CO_2eq	1 Kg[3]

4 Results and discussion

4.1 Quantitative results on validation set

Table 3 illustrates the results on the 20 validation cases whose ground truth are publicly provided by FLARE2022. Large-volume organs like liver, Aorta, spleen and kidney have relatively good performance. However, the poor performance

on small organs and diseased organs reduces the overall average performance. Indeed, the segmentation results have the problem of organ disappearance, that is, the segmentation results of some organs, especially small organs, are not predicted. Another problem is that due to the deformation of some diseased organs as well as the lack training of model on such data, the segmentation prediction of diseased organs is completely wrong. The above two situations are the main reasons that lead to the corresponding DSC and NSD of some cases' organs segmentation valuing 0, which is also a part to be further explored and sovled. It is noted that the values of NSD are generally better than that of DSC, indicating that the method proposed in this paper has a relatively better performance on organ boundary segmentation.

Table 3. Quantitative results on validation set (best).

Organs	DSC(%)	NSD(%)
Liver	94.95	93.66
RK	76.71	75.60
Spleen	85.07	84.87
Pancreas	70.24	80.09
Aorta	91.59	96.13
IVC	81.03	81.99
RAG	67.02	83.44
LAG	70.20	83.91
Gallbladder	59.43	61.40
Esophagus	72.12	81.77
Stomach	74.68	78.28
Duodenum	61.62	75.40
LK	80.38	82.76
mean	77.98	81.48

Compared with the segmentation results using only label data under the same experimental conditions, as shown in Table 4, the semi-supervised method proposed in this paper is improved on DSC, which directly demonstrate that effective utilization of unlabeled data can improve segmentation performance. The poor results using only labeled data also demonstrate the importance of the distribution of training data to model performance. Especially in the data of FLARE2022, there is the problem of different organ lesions between the training set and validation set, making their distributions vary and leading to extremely poor segmentation of some organs.

4.2 Qualitative results on validation set

Figure 2 presents the qualitative results on the validation set, where row 1 and 2 show some cases that are relatively easy to segment while row 3 and 4 show

Table 4. Quantitative results on validation set b (best).

Organs	DSC(our propose, %)	DSC(only labeled,%)
Liver	94.06	94.71
RK	85.02	83.17
Spleen	89.41	80.08
Pancreas	73.44	65.11
Aorta	90.09	90.97
IVC	79.90	76.97
RAG	63.17	63.56
LAG	66.31	53.13
Gallbladder	72.71	54.09
Esophagus	74.07	66.04
Stomach	78.54	75.81
Duodenum	65.44	55.69
LK	81.56	79.14
mean	77.98	72.19

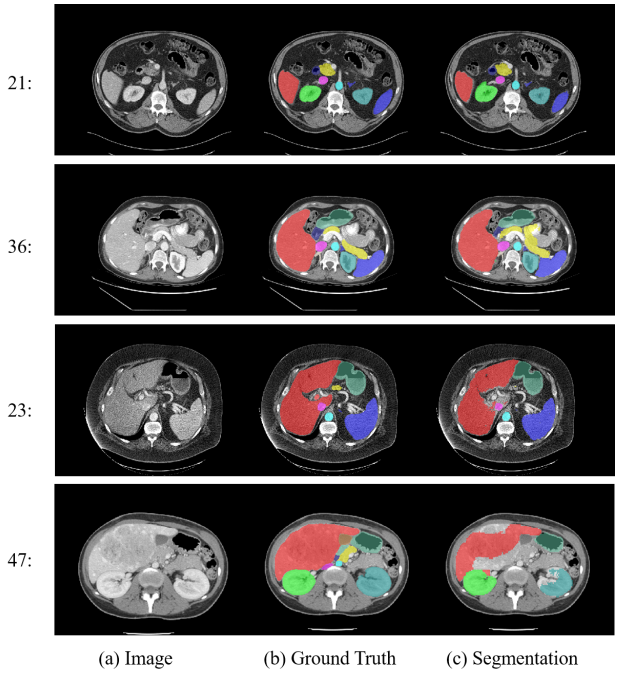

Fig. 2. Qualitative evaluation of model performance on validation set. Row 1 and 2:Well-segmented examples. Row 3 and 4: challenging examples.

cases that are challenging to segment. It can be find that the cases of row 1 and 2 have clear boundaries and good contrast for the organs, and there exist no severe artifacts or lesions in the organs. Compared with the well-segmented cases, the challenge cases usually have noise(row 3) or organs lesion(row 4), which create difficulties in correctly segmenting the organ.

5 Conclusion

Due to the rush of time in this work, in fact, only half a month was spent on research, the segmentation performance did not achieve our goal. In this work, it can be seen that the segmentation results of small organs is not very good. The deformation of the diseased organ is also not considered in the segmentation process. In the future, further research can be carried out in terms of boundary constraints and attention mechanisms.

Acknowledgements We declare that the segmentation method that implemented for participation in the FLARE 2022 challenge has not used any pretrained models nor additional datasets other than those provided by the organizers. The proposed solution is fully automatic without any manual intervention.

References

1. Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L.: Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. IEEE transactions on medical imaging **40**(1), 431–443 (2020) 3, 4
2. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) 5
3. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 5
4. Heller, N., McSweeney, S., Peterson, M.T., Peterson, S., Rickman, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Rosenberg, J., et al.: An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging. American Society of Clinical Oncology **38**(6), 626–626 (2020) 5
5. Liu, X., Yuan, Q., Gao, Y., He, K., Wang, S., Tang, X., Tang, J., Shen, D.: Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. Pattern recognition **122**, 108341 (2022) 2, 3, 4
6. 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) 2

7. Luo, X., Liao, W., Xiao, J., Song, T., Zhang, X., Li, K., Wang, G., Zhang, S.: Word: Revisiting organs segmentation in the whole abdominal region. arXiv preprint arXiv:2111.02403 (2021) 1
8. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2021.3100536 5
9. 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 (2015) 1, 2
10. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 5
11. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. International Journal of Computer Vision **129**(4), 1106–1120 (2021) 2, 3, 4