

# Dual-Consistency Semi-Supervised Learning with Uncertainty Quantification for COVID-19 Lesion Segmentation from CT Images

Yanwen Li<sup>\*1</sup>, Luyang Luo<sup>\*2</sup>,  
Huangjing Lin<sup>1,2</sup>, Hao Chen<sup>3</sup>, Pheng-Ann Heng<sup>2,4</sup>

<sup>1</sup>Insight AI Research Lab, Shenzhen, China  
liyanwen@insightmed.com

<sup>2</sup>Department of Computer Science and Engineering,  
The Chinese University of Hong Kong, Hong Kong, China  
lyluo@cse.cuhk.edu.hk

<sup>3</sup>Department of Computer Science and Engineering,  
The Hong Kong University of Science and Technology, Hong Kong, China.  
jhc@cse.ust.hk

<sup>4</sup>Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine  
Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese  
Academy of Sciences, China

**Abstract.** The novel coronavirus disease 2019 (COVID-19) characterized by atypical pneumonia has caused millions of deaths worldwide. Automatically segmenting lesions from chest Computed Tomography (CT) is a promising way to assist doctors in COVID-19 screening, treatment planning, and follow-up monitoring. However, voxel-wise annotations are extremely expert-demanding and scarce, especially when it comes to novel diseases, while an abundance of unlabeled data could be available. To tackle the challenge of limited annotations, in this paper, we propose an uncertainty-guided dual-consistency learning network (UDC-Net) for semi-supervised COVID-19 lesion segmentation from CT images. Specifically, we present a dual-consistency learning scheme that simultaneously imposes image transformation equivalence and feature perturbation invariance to effectively harness the knowledge from unlabeled data. We then quantify the segmentation uncertainty in two forms and employ them together to guide the consistency regularization for more reliable unsupervised learning. Extensive experiments showed that our proposed UDC-Net improves the fully supervised method by 6.3% in Dice and outperforms other competitive semi-supervised approaches by significant margins, demonstrating high potential in real-world clinical practice. <sup>2</sup>

**Keywords:** COVID-19 · Semi-supervised learning · Uncertainty · Segmentation.

<sup>1</sup> The first two authors contributed equally.

<sup>2</sup> Code is available at <https://github.com/poiuohke/UDC-Net>.

## 1 Introduction

By the end of 2020, the coronavirus disease 2019 (COVID-19) [36] characterized by atypical pneumonia has spread over 220 countries and areas, infected more than 81 million people, and caused near 1.8 million losses of lives<sup>1</sup>. For early screening of the COVID-19, chest computed tomography (CT) plays a vital role as a noninvasive and fast technique, which is reported to have high sensitivity for detecting COVID-19-related abnormal findings [6,1,13,7]. To improve the screening efficiency and alleviate radiologists' reading burden, various automatic COVID-19 chest CT analysis methods have been proposed from whole-volume classification and triaging [20,27,8,17,4], weakly-supervised lesion localization [16,31], to accurate segmentation of lesion regions [5,26]. Among previous studies, segmentation of COVID-19 often provides more accurate descriptions of the lesions, which has significant potential in assisting doctors with the diagnosis, treatment planning, and follow-up monitoring.

Currently, advanced segmentation methods are often fully supervised and heavily rely on pixel-wise or voxel-wise annotations. For novel diseases like COVID-19, acquiring such annotations is extremely expertise-demanded and time-consuming, while unlabeled data are often abundant due to increasing positive cases. Therefore, semi-supervised learning (SSL) that utilizes both labeled and unlabeled data is of great value to develop robust and accurate COVID-19 lesion segmentation algorithms. Thus far, many SSL approaches have been developed and successfully applied to various tasks [25]. Many works [23,19,2,9,14] adopts the smoothness assumption that two data samples that are close in the input space share the same label. This assumption is further expanded to the deep feature space, where similarities of feature maps are used for cluster assignment [28,21,29]. Despite the achievement, these approaches do not ensure robust learning from samples with low uncertainty. To reduce the influence of uncertain samples, uncertainty guidance has been introduced into the literature of SSL [34,33,30,15]. Nevertheless, semi-supervised segmentation of COVID-19 lesions remains a challenging task, of which the annotations are extremely scarce, and the lesions often have irregular and ambiguous contours.

To tackle the above challenges, we propose a novel deep neural network with a uncertainty-guided dual-consistency learning scheme for COVID-19 lesion segmentation from chest CT scan volumes. Specifically, we impose *image-level transformation equivalence* out of the observation that the prediction of a sample should obtain the same transformation of the input. Meanwhile, we adopt *feature-level perturbation invariance* to a multi-decoder V-Net, where auxiliary decoder paths take perturbed features as inputs and form output consistency with a main decoder. Dual-consistency comprehensively enforces smoothness assumption into the SSL model from both input space and feature space, and hence the network could learn more invariant representations to diverse input or feature variants. Moreover, deep neural networks could memorize and easily overfit to noisy and uncertain contour points of COVID-19 lesions [35], which leads to

<sup>1</sup> <https://covid19.who.int>

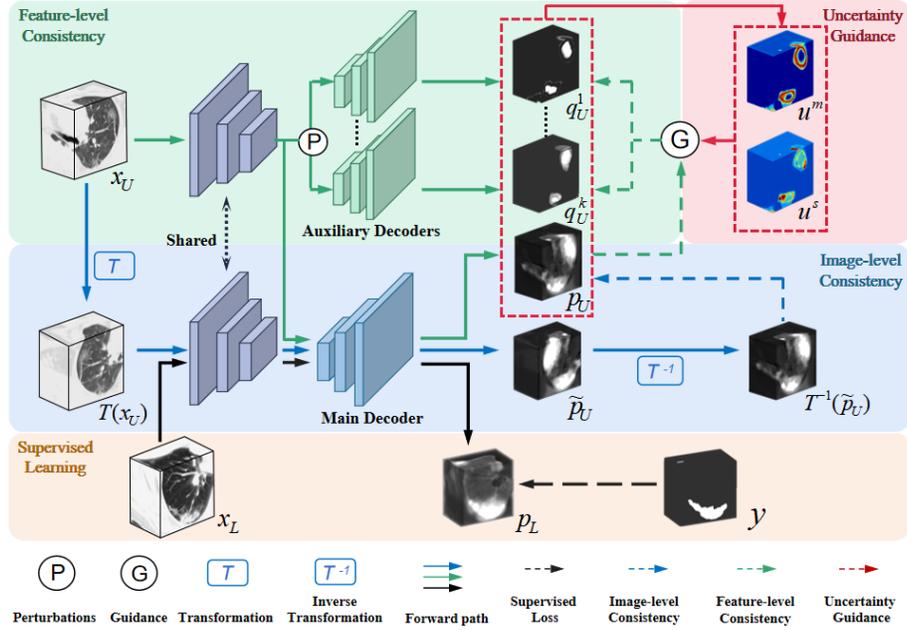


Fig. 1: Overview of UDC-Net. Feature-level consistency (in green) is formed by the main decoder’s prediction  $p_U$  and auxiliary decoders’ predictions  $\{q_U^1, \dots, q_U^k\}$ . Image-level consistency (in blue) is formed by  $p_U$  and the prediction  $\tilde{p}_U$  of transformed image. The confidence uncertainty  $u^m$  and the consensus uncertainty  $u^s$  are quantified by mean and standard deviation of the multi-decoders’ predictions, which are then used to guide the consistency learning (in red). A supervised loss is also used on the labeled data (in orange).

poor generalization in real-world clinical practice. Hence, we further introduce a novel uncertainty guidance to the consistency learning process. Particularly, we quantify both the confidence uncertainty and the consensus uncertainty based on the multi-decoder structure. The estimated uncertainties are then used together in an indicator function to filter out uncertain samples during training. The proposed uncertainty-guided dual-consistency network (UDC-Net) is evaluated on a large-scale COVID-19 dataset with 852 whole-volume chest CT scans. Extensive experiments show that our approach outperforms other competitive SSL-based segmentation approaches, yielding state-of-the-art performance on semi-supervised COVID-19 lesion segmentation.

## 2 Method

As shown in Fig. 1, our UDC-Net consists of a modified 3D multi-decoder V-Net [18] as its backbone. Apart from the supervised loss, our method makes full use of

the unlabeled data by both feature-level and image-level consistency modules. Moreover, both the confidence uncertainty and the consensus uncertainty are estimated to guide more robust consistency learning.

## 2.1 Dual-consistency Learning for Semi-supervised Segmentation

**Image-level Consistency Learning** via transformation equivalence of deep segmentation models  $f_{\text{seg}}$  indicates that while a transformation  $T(\cdot)$  is applied to an input image  $x$ , there should be  $f_{\text{seg}}(T(x)) = T(f_{\text{seg}}(x))$  [32]. We conduct random transformation on the images to get the perturbed version  $T(x)$  as the input to our network. Subsequently, we have the corresponding prediction  $f(T(x))$  given by the V-Net and the inverse transformation to the output  $T^{-1}(f(T(x)))$ , which should be consistent to the output of input data without transformation  $f(x)$ . Following the notations set before, let  $p = f(x)$  and  $\tilde{p} = f(T(x))$ , we introduce an image-level consistency regularization by minimizing the L2 loss between the two versions of output:

$$\mathcal{L}_{\text{IC}} = \frac{1}{N} \sum_{i=1}^N \|p_i - [T^{-1}(\tilde{p})]_i\|_2^2 \quad (1)$$

where  $i$  and  $N$  are the index and the total number of voxels, respectively.

**Feature-level Consistency Learning** via perturbation invariance can also enrich the learned representation of the model [19]. Particularly, different perturbed versions of the same feature maps should maintain the same predictions. Following [21], we append several auxiliary decoders to the V-Net and inject shared encoder’s outputs with various types of perturbations. Each auxiliary decoder receives a different version of the perturbed feature map, while the main decoder receives the un-perturbed feature map. Denoting the prediction from the main decoder as  $p$ , the prediction from the  $k$ -th auxiliary decoder as  $q^k$ , the feature-level consistency is achieved by regularizing  $p$  and each  $q^k$  as follows:

$$\mathcal{L}_{\text{FC}} = \frac{1}{N \cdot K} \sum_{i=1}^N \sum_{k=1}^K \|p_i - q_i^k\|_2^2 \quad (2)$$

where  $K$  is the total number of extra decoders. Following [21], seven types of feature perturbations, i.e., Feature noise, Feature dropout, Object masking, Context masking, Guided cutout, Intermediate VAT, and Random dropout, were introduced to seven auxiliary decoders, respectively. Detailed descriptions of each perturbation strategy can be found in the supplementary. All extra decoders were required to generate consistent prediction with the main decoder.

## 2.2 Dual Uncertainty Quantification for Robust Learning

The perturbation of the hidden representations during the consistency learning process could amplify the feature noises and uncertainty caused by the difficulty

of accurately delineating the lesion contours of COVID-19. To this end, we propose to quantify both the confidence uncertainty and the consensus uncertainty of the multi-decoders, to guide more robust unsupervised learning.

**Confidence Uncertainty** indicates whether the model generates confident predictions. Previous works [34,15] used the entropy of the mean prediction of multiple perturbed inputs from self-ensembling models to estimate the prediction uncertainty. In our case, this form of uncertainty can be easily quantified using the main decoder and the  $K$  auxiliary decoders as below:

$$\mu_i = \frac{1}{K+1} \left[ \left( \sum_{k=1}^K q_i^k \right) + p_i \right] \quad \text{and} \quad u_i^m = -\mu_i \log \mu_i \quad (3)$$

where  $i$  indicates the voxel index,  $K$  is the total number of auxiliary decoders,  $\mu$  is the mean prediction, and  $u^m$  is the estimated uncertainty. The higher  $u_i^m$  is, the less confidence the model is on its prediction.

**Consensus Uncertainty** indicates whether the model generates consistent predictions over multiple runs with perturbed data [11,9]. Supposing the average prediction of a suspicious infection area is high but the outputs from different branches vary severely, this means the area is sensitive to perturbation. By the smoothness assumption [3], the predictions for the target should be robust to perturbation, and the sensitive prediction hence highly suggests a noisy sample. Hence, we quantify the consensus uncertainty  $u^s$  as the standard deviation over the multi-decoders' predictions:

$$u_i^s = \frac{1}{K+1} \sqrt{\left[ \sum_{k=1}^K (q_i^k - \mu_i)^2 \right] + (p_i - \mu_i)^2} \quad (4)$$

Here,  $u^s$  essentially indicates the consensus among different decoders, which is complementary with  $u^m$  which measures the confidence of the model.

### 2.3 Uncertainty-guided Dual-consistency Learning for Segmentation

The quantified uncertainties are used to filter out uncertain voxels and consequently guide the model to learn from more reliable unlabeled data. Denoting  $i$  as the voxel index for the prediction volume, the reliable voxels are selected from a set  $\Omega = \{i | u_i^s < \tau^s \ \& \ u_i^m < \tau^m\}$ , where  $\tau^s$  and  $\tau^m$  are two thresholds. The cross consistency loss among decoders is then guided by:

$$\mathcal{L}_{\text{UFC}} = \sum_{k=1}^K \sum_{i \in \Omega} \|p_i - q_i^k\|_2^2 \quad (5)$$

Here, the uncertainty guidance is applied onto feature-level consistency learning as the uncertainties are generated with feature perturbations. Thus, the total loss for our uncertainty-guided dual-consistency learning UDC-Net for semi-supervised lesion segmentation is as follows:

$$\mathcal{L} = \mathcal{L}_S + \alpha\mathcal{L}_{IC} + \beta\mathcal{L}_{UFC} \quad (6)$$

where  $\mathcal{L}_S$  is the supervised loss consists of a Dice loss and a cross entropy loss,  $\alpha$  and  $\beta$  are two hyper-parameters weighing the contributions from different losses.

During training, we first trained a supervised V-Net and then added the extra decoders for finetuning with uncertainty-guided consistency learning. The training process was terminated if the Dice coefficient on the validation dataset stagnated. Adam [10] was used as the optimizer with an initial learning rate of 0.001 and a learning decay rate of 0.95 per epoch. As widely adopted by SSL works [24,21],  $\alpha$  and  $\beta$  were set to be two sigmoid-shape monotonically functions of the training steps with maximum of 1. The threshold  $\tau^m$  and  $\tau^s$  were set to 0.34 and 0.12 after tuning on the validation set. For testing, we carried out sliding window inference and took only the main decoder’s prediction. All implementation was done with Pytorch [22] on an NVIDIA TITAN X GPU.

### 3 Experiments

#### 3.1 Datasets and Evaluation Metrics

**Datasets.** In total, 852 chest CT volumes acquired from December 2019 to April 2020 were collected and enrolled in this study, among which 144 were voxel-annotated by four experienced radiologists. The labeled data were divided into: (1) 65 cases as labeled training dataset; (2) 9 cases as the validation set; and (3) 70 cases as the testing set. The remained 708 chest CT scans were used as the unlabeled training data.

**Evaluation Metrics.** We adopted Dice Score (DSC), Jaccard similarity coefficient (Jaccard), and Average Symmetric Surface Distance (ASD) to evaluate the segmentation performance.

#### 3.2 Ablation Study on Different Components

We conduct ablation studies to analyze the contributions of our proposed methods, and the quantitative results can be seen in Table 1. Regarding the testing set performance, image-level consistency (IC) shows increases of 2.4% in DSC, 2.5% in Jaccard, and 3.7 in ASD comparing to 3D V-Net. Meanwhile, feature-level consistency (FC) regularization leads to a large improvement of 4.5% in DSC, 5.5% in Jaccard, and 6.0 in ASD comparing to 3D V-Net. Unifying dual consistencies further improves DSC and Jaccard with about 1%, which demonstrates the effectiveness of learning from the unlabeled data. Further, introducing either the confidence uncertainty or the consensus uncertainty guidance consistently benefit the learning of the unlabeled data. Moreover, our method with dual uncertainty achieves better DSC and Jaccard with a comparable ASD to those of the single-uncertainty models, further demonstrating that dual uncertainties are complementary for guiding more robust learning.

Table 1: Ablation study of different components. All results are reported as validation/testing results. (FC: feature-level consistency; IC: image-level consistency; UM: confidence uncertainty computed by the mean of the multi-decoders’ predictions; US: consensus uncertainty computed by the standard deviation of the multi-decoders’ predictions)

Components				Evaluation Metrics		
IC	FC	UM	US	DSC[%] $\uparrow$	Jaccard[%] $\uparrow$	ASD[mm] $\downarrow$
				70.0 / 71.1	56.5 / 56.8	12.1 / 12.1
✓				70.3 / 73.5	56.7 / 59.3	12.2 / 8.4
	✓			71.4 / 75.6	58.4 / 62.3	12.1 / 6.1
✓	✓			71.9 / 76.7	58.9 / 63.7	11.7 / 5.8
✓	✓	✓		72.2 / 77.0	59.0 / 64.0	11.3 / <b>3.2</b>
✓	✓		✓	72.4 / 77.2	59.5 / 64.3	11.4 / 4.1
✓	✓	✓	✓	<b>72.7 / 77.4</b>	<b>59.9 / 64.5</b>	<b>10.9 / 3.9</b>

### 3.3 Comparison with State-of-the-art Methods

We compare our method against other state-of-the-art semi-supervised segmentation approaches. Several recent models were implemented, including Mean-Teacher (MT) [24], Uncertainty-aware mean teacher [34], Transformation-consistent Self-ensembling Model (TCSM) [12], and Cross Consistency Training (CCT) [21]. We run each methods four times with different random seeds.

**Quantitative comparison** results are reported in Table 2. For a fair comparison, we implemented all methods with the 3D V-Net as backbone. As observed, UDC-Net outperforms all other methods with at least 1.8% in Dice, 2.2% in Jaccard, and 2.2 in ASD, showing outstanding unsupervised learning efficacy.

**Qualitative comparison** is illustrated by visualizing the segmentation results in Figure 2. As demonstrated, Our UDC-Net delineates more accurate lesion contours than other methods regarding diverse shapes and sizes of lesion. Visualization of the two uncertainties can be found in the supplementary.

Table 2: Quantitative comparison with other semi-supervised methods.

Methods	Evaluation Metrics		
	DSC[%] $\uparrow$	Jaccard[%] $\uparrow$	ASD[mm] $\downarrow$
V-Net [18]	71.1 $\pm$ 0.40	56.8 $\pm$ 0.45	12.1 $\pm$ 2.1
Mean Teacher [24]	72.5 $\pm$ 0.25	58.2 $\pm$ 0.36	11.3 $\pm$ 1.8
UA-MT [34]	74.0 $\pm$ 0.11	60.1 $\pm$ 0.15	9.2 $\pm$ 0.9
TCSM [12]	72.9 $\pm$ 0.46	58.9 $\pm$ 0.58	9.1 $\pm$ 1.4
CCT [21]	75.6 $\pm$ 0.11	62.3 $\pm$ 0.19	6.1 $\pm$ 0.7
UDC-Net(ours)	<b>77.4 <math>\pm</math> 0.14</b>	<b>64.5 <math>\pm</math> 0.15</b>	<b>3.9 <math>\pm</math> 0.5</b>

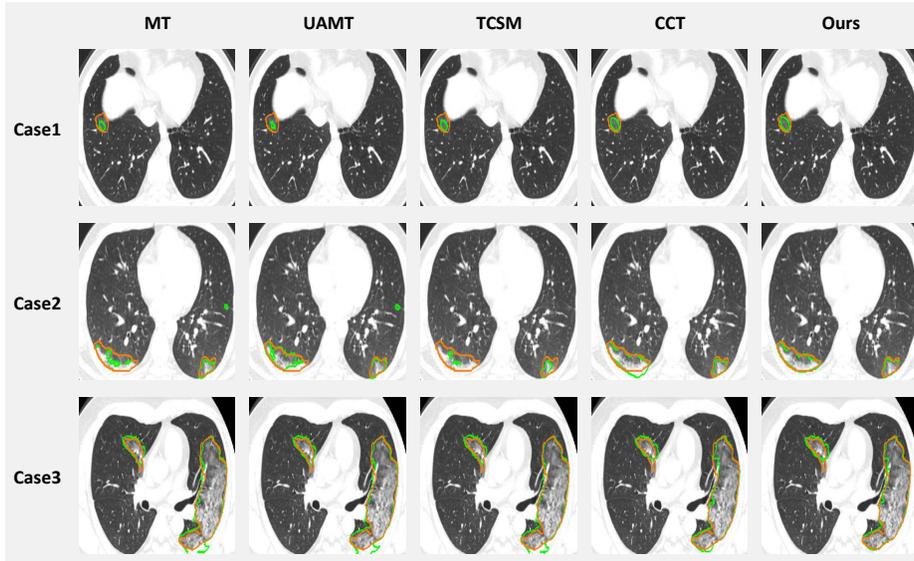


Fig. 2: Qualitative comparison. Green and orange curves delineate the model prediction and ground truth, respectively. Best viewed in color.

### 3.4 Analysis on Efficacy of Leveraging Unlabeled Data

We further evaluate our UDC-Net’s effectiveness by varying the ratios of labeled and unlabeled training data. Table 3 shows that UDC-Net consistently improves the baseline V-Net with significant margins in both DSC, Jaccard, and ASD, whenever 32 or 65 labeled scans are provided. Moreover, the proposed approach

Table 3: Quantitative performance comparison under different numbers of training labeled/unlabeled data.

Method	# scans used		Evaluation Metrics		
	Labeled	Unlabeled	DSC[%] ↑	Jaccard[%] ↑	ASD[mm] ↓
V-Net [18]	32	0	70.4	56.0	4.3
CCT [21]	32	140	74.4	60.7	5.8
<b>UDC-Net (ours)</b>	32	140	<b>75.0</b>	<b>61.5</b>	<b>4.8</b>
CCT [21]	32	708	75.2	61.6	7.0
<b>UDC-Net (ours)</b>	32	708	<b>76.9</b>	<b>64.0</b>	<b>4.4</b>
V-Net	65	0	71.1	56.8	12.1
CCT [21]	65	140	75.1	61.8	5.8
<b>UDC-Net (ours)</b>	65	140	<b>76.6</b>	<b>63.5</b>	<b>5.4</b>
CCT [21]	65	708	75.6	62.3	6.0
<b>UDC-Net (ours)</b>	65	708	<b>77.4</b>	<b>64.5</b>	<b>3.9</b>

consistently outperforms CCT [21] (the best model among those compared with ours) under all different scenarios. Notably, when less data are given, UDC-Net shows comparable or even better results than CCT. For instance, UDC-net achieves 75.0% DSC, 61.5% Jaccard, and 4.8 ASD with 32 labeled scans and 140 unlabeled scans (3rd row), which is comparable to the performance of CCT with double labeled scans (7th row). With 65 labeled scans and 140 unlabeled scans, UDC-Net (8th row) shows superior performance than CCT with 5 times unlabeled data (9th row). These findings demonstrate that our method enables more efficient unsupervised learning, suggesting

## 4 Conclusions

In this paper, we present an uncertainty-guided dual-consistency learning method for semi-supervised COVID-19 lesion segmentation from chest CT scans. Image-level transformation equivalence and feature-level perturbation invariance are both introduced to form dual consistency learning from unlabeled data. Meanwhile, the dual uncertainty mechanism further improves the learning process with more reliable and robust guidance. Extensive experiments on a large COVID-19 dataset demonstrate the efficiency of our method in real-world scenarios. Future work will include improving the method with more robust knowledge distillation and generalizing to other semi-supervised learning tasks.

**Acknowledgement.** This work was supported by Key-Area Research and Development Program of Guangdong Province, China (2020B010165004), Hong Kong Innovation and Technology Fund (Project No. ITS/311/18FP and Project No. ITS/426/17FP.), and National Natural Science Foundation of China with Project No. U1813204.

## References

1. Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., Xia, L.: Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology* p. 200642 (2020)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: *NeurIPS*. pp. 5049–5059 (2019)
3. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning. *IEEE TNNLS* **20**(3), 542–542 (2009)
4. Di, D., Shi, F., Yan, F., Xia, L., Mo, Z., Ding, Z., et al.: Hypergraph learning for identification of covid-19 with ct imaging. *MedIA* p. 101910 (2020)
5. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., et al.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI* (2020)
6. Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., et al.: Sensitivity of chest ct for covid-19: comparison to rt-pcr. *Radiology* p. 200432 (2020)

7. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al.: Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet* **395**(10223), 497–506 (2020)
8. Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., et al.: Development and evaluation of an artificial intelligence system for covid-19 diagnosis. *Nat. Commun* **11**(1), 1–14 (2020)
9. Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W.: Dual student: Breaking the limits of the teacher in semi-supervised learning. In: *ICCV*. pp. 6728–6736 (2019)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: *ICLR* (2015)
11. Lee, J., Chung, S.Y.: Robust training with ensemble consensus. In: *ICLR* (2020), <https://openreview.net/forum?id=ryxOUTVYDH>
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 TNNLS* (2020)
13. Liang, W., Yao, J., Chen, A., Lv, Q., Zanin, M., Liu, J., Wong, S., et al.: Early triage of critically ill covid-19 patients using deep learning. *Nat. Commun* **11**(1), 1–7 (2020)
14. Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A.: Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE TMI* (2020)
15. Luo, L., Yu, L., Chen, H., Liu, Q., Wang, X., Xu, J., et al.: Deep mining external imperfect data for chest x-ray disease screening. *IEEE TMI* **39**(11), 3583–3594 (2020)
16. Ma, J., Nie, Z., Wang, C., Dong, G., Zhu, Q., He, J., Gui, L., Yang, X.: Active contour regularized semi-supervised learning for covid-19 ct infection segmentation with limited annotations. *Physics in Medicine & Biology* (2020)
17. Mei, X., Lee, H.C., Diao, K.y., Huang, M., Lin, B., Liu, C., et al.: Artificial intelligence-enabled rapid diagnosis of patients with covid-19. *Nat. Med* pp. 1–5 (2020)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *3DV*. pp. 565–571. *IEEE* (2016)
19. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE TPAMI* **41**(8), 1979–1993 (2018)
20. Oh, Y., Park, S., Ye, J.C.: Deep learning covid-19 features on cxr using limited training data sets. *IEEE TMI* (2020)
21. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: *CVPR*. pp. 12674–12684 (2020)
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., et al.: Pytorch: An imperative style, high-performance deep learning library. In: *NeurIPS*. pp. 8026–8037 (2019)
23. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: *ECCV*. pp. 135–152 (2018)
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS* **30**, 1195–1204 (2017)
25. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. *Mach Learn* **109**(2), 373–440 (2020)
26. Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., et al.: A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. *IEEE TMI* **39**(8), 2653–2663 (2020)

27. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci. Rep* **10**(1), 1–12 (2020)
28. Wang, X., Chen, H., Ran, A.R., Luo, L., Chan, P.P., Tham, C.C., et al.: Towards multi-center glaucoma oct image screening with semi-supervised joint structure and function multi-task learning. *MedIA* **63**, 101695 (2020)
29. Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., Heng, P.A.: Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Medical image analysis* **70**, 102010 (2021)
30. Wang, X., Tang, F., Chen, H., Luo, L., Tang, Z., Ran, A.R., et al.: Ud-mil: Uncertainty-driven deep multiple instance learning for oct image classification. *IEEE JBHI* (2020)
31. Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., et al.: A weakly-supervised framework for covid-19 classification and lesion localization from chest ct. *IEEE TMI* (2020)
32. Worrall, D.E., Garbin, S.J., Turmukhambetov, D., Brostow, G.J.: Harmonic networks: Deep translation and rotation equivariance. In: *CVPR*. pp. 5028–5037 (2017)
33. Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., et al.: Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *MedIA* **65**, 101766 (2020)
34. 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: *MICCAI*. pp. 605–613. Springer (2019)
35. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: *ICLR* (2017)
36. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al.: A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine* (2020)

## 5 Supplementary Materials

Table 4: List of perturbations used in feature-level consistency learning.

<b>Perturbations</b> <sup>[21]</sup>	<b>Description</b>
Feature Noise	A noise tensor $N$ is applied to the output of encoder $z$ to get $\tilde{z} = z * N + z$ .
Feature Dropout	Generating a randomly dropout mask $M_{drop}$ to obtain perturbed $\tilde{z} = z * M_{drop}$ .
Object Masking	Generating a object mask $M_{obj}$ using the output of main decoder to get $\tilde{z} = z * M_{obj}$ .
Context Masking	Generating a context mask $M_{con} = 1 - M_{obj}$ to obtain $\tilde{z} = z * M_{con}$ .
Guided cutout	Zero-out a random crop within each object’s bounding box from the corresponding feature map $z$ .
Intermediate VAT	Using VAT to push the distribution to be isotropically smooth. Finding the adversarial perturbation $r_{adv}$ alter its prediction the most and injected into $z$ to obtain $\tilde{z} = r_{adv} + z$ . <sup>[19]</sup>
Random dropout	Spacial dropout applied to $z$ as a random perturbation

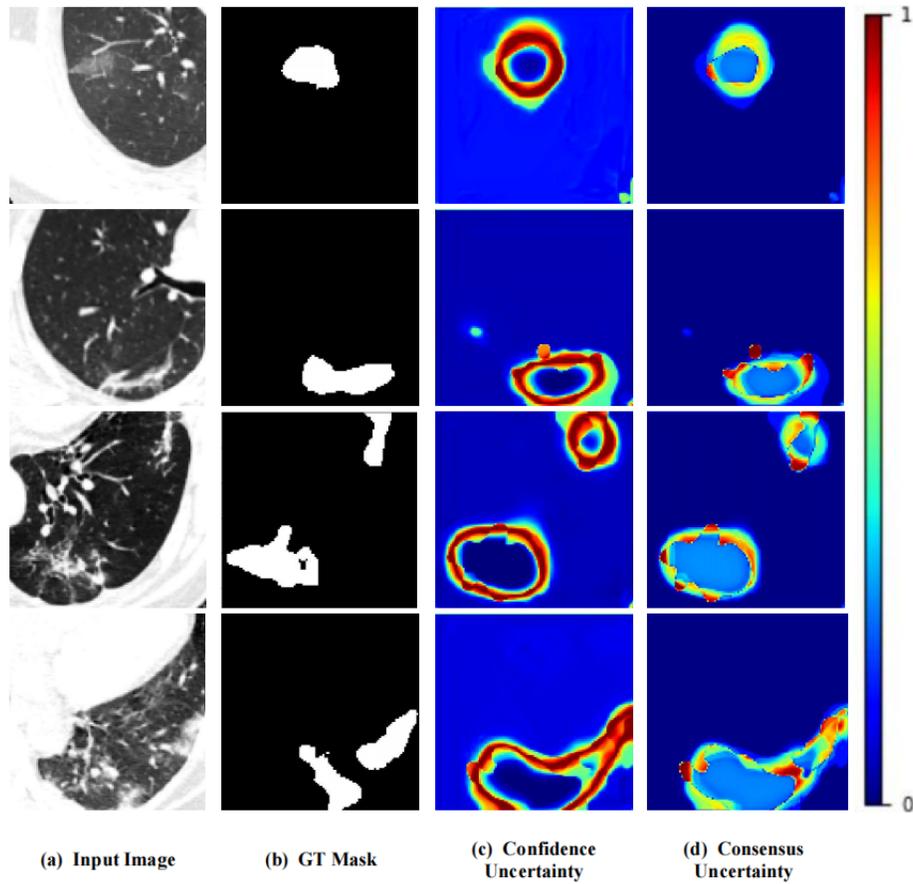


Fig. 3: Visualization of input images, ground truth masks, confidence uncertainty map, and consensus uncertainty map. The visualization results demonstrates that our proposed confidence and consensus uncertainties are complementary.

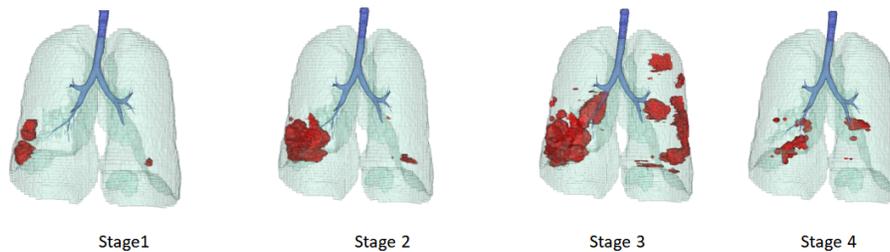


Fig. 4: An example of monitoring the lesion development of a patient from mild infection (stage 1), to common infection (stage 2), severe infection (stage 3), and finally to recovery stage (stage 4).