Uncertainty-Guided Cross Attention Ensemble Mean Teacher for Semi-supervised Medical Image Segmentation

Meghana Karri[1], Amit Soni Arya[2], Koushik Biswas[1], Nicolò Gennaro[1], Vedat Cicek[1],
Gorkem Durak[1], Yuri S. Velichko[1], Ulas Bagci[1]
[1]Department of Radiology, Northwestern University, Chicago, IL, USA
[2]School of Computer Science Engineering and Technology,
Bennett University, Greater Noida, UP, India
meghana.karri@northwestern.edu

Supplementary Material

1. Qualitative analysis

Figure 1 provides a detailed visualization of the uncertainty maps during the training phase, highlighting the impact of our consistency regularization approach within the UG-CEMT framework. As demonstrated in our **qualitative analysis (main paper)**, initially, the maps exhibit higher uncertainty, reflecting the model's early learning stage. As training progresses, the maps show increasingly higher confidence, demonstrating the effectiveness of our method. Consistency regularization helps the model produce stable predictions under various perturbations, enhancing the confidence and accuracy of predictions over time. It indicates that UG-CEMT effectively leverages consistency regularization to improve learning from both labeled and unlabeled data, resulting in more reliable and accurate segmentation.

2. Effect of Cross-Attention Ensemble Mean Teacher Framework

Figure 2 illustrates the impact of our Cross-Attention Ensemble Mean Teacher (CEMT) framework on model performance. The CEMT integrates cross-attention mechanisms (CA) and exponential weighted averaging (EWA) between the student and teacher models to enhance overall segmentation performance. As demonstrated in our ablation study (**Effect of different components in main paper**), the inclusion of EWA in the baseline student-teacher (ST) setup significantly improved the Dice score and reduced 95HD, indicating the positive effect of averaging model weights over time. Incorporating CA further enhanced these metrics, highlighting the importance of effective feature alignment and information exchange between the student and teacher models. The visual results in Figure 2 showcase how the EWA and CA components in the CEMT framework progressively reduce uncertainty and increase confidence in the predictions during the training process. The CA facilitates robust feature interactions, while EWA ensures stable and generalized learning. This combination leads to more accurate and consistent segmentation results over time.

3. Comparison with state-of-the-art methods

In this supplementary section, we provide additional quantitative comparisons of our proposed UG-CEMT method with recent state-of-the-art semi-supervised learning (SSL) methods such as PLGC, CauSSL, and MCF. Due to space constraints, these detailed results are presented here to complement the qualitative analysis provided in the main manuscript. Table 1 compares the performance of UG-CEMT and other SSL methods on the LA dataset across different labeled data ratios (5%, 10%, and 20%). UG-CEMT consistently outperforms other methods, particularly in terms of Dice and 95HD metrics. For example, with 5% labeled data, UG-CEMT achieves a Dice score of 85.89%, compared to 85.02% from MCF and 84.17% from CauSSL. As the labeled data increases, UG-CEMT continues to show its superiority, achieving 89.73% Dice with 20% labeled data, while other methods, such as CauSSL and MCF, reach only 88.87% and 89.05%, respectively.

This shows that UG-CEMT's innovative integration of cross-attention, uncertainty-guided regularization, and SAM significantly enhances segmentation performance, even with limited labeled data. Table 2 shows a similar trend on the multi-site prostate dataset. UG-CEMT consistently achieves the best performance across different labeled data percentages. With 5% labeled data, UG-CEMT yields a Dice score of 65.68%, outperforming MCF (64.18%) and PLGC (62.59%). As the labeled data increases, UG-CEMT continues to dominate, achieving 72.02% Dice with

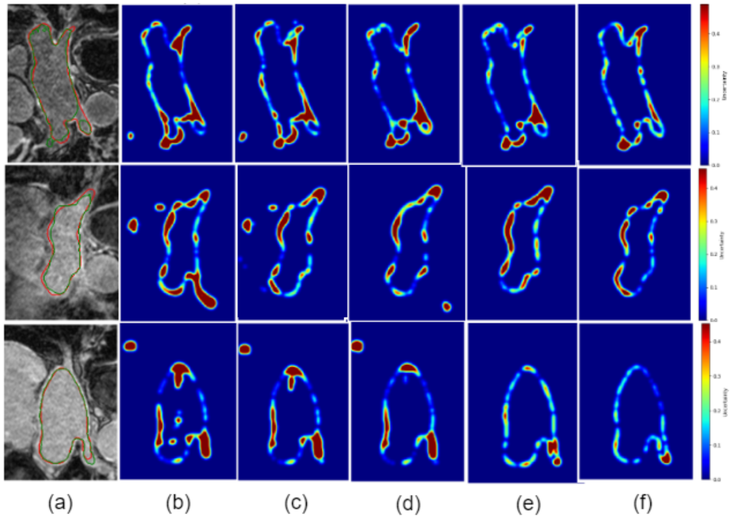

Figure 1. Visualization of uncertainty maps. (a) The overlap predicted (green line) and GT (red line) regions. (b), (c), (d), (e), and (f) are the uncertainty maps at 2000, 3000, 4000, 5000, and 6000 iterations.

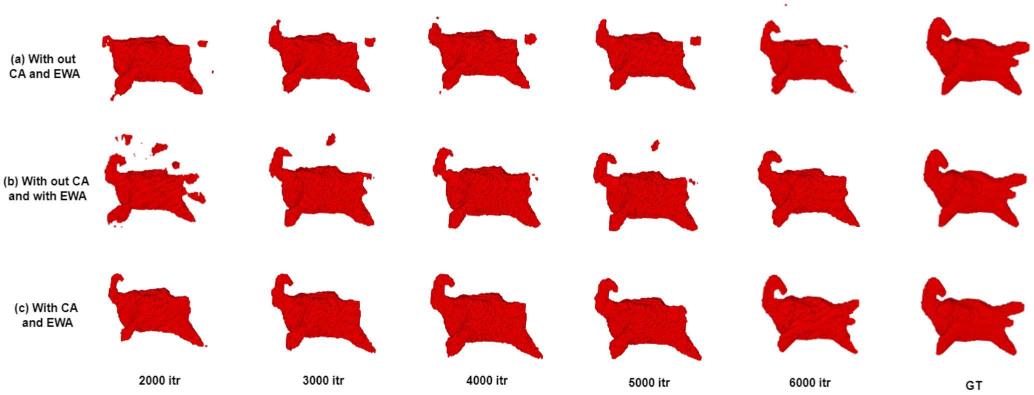

Figure 2. Effect of cross attention ensemble mean teacher framework visualized outcomes. (a) Without CA and EWA. (b) Without CA and with EWA, (c) With CA and EWA.

Table 1. Comparison of the proposed UG-CEMT with other state-of-the-art SSL methods on LA dataset for 6000 iterations.

Method	(%) of images used		Metrics			
	labeled	unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
PLGC	4(5%)	76	83.98	72.16	11.35	4.34
CauSSL	4(5%)	76	84.17	72.89	8.68	5.23
MCF	4(5%)	76	85.02	73.78	9.78	4.08
CEMT(Ours)	4(5%)	76	85.23	75.16	5.12	1.32
UG-CEMT(Ours)	4(5%)	76	**85.89**	**76.23**	**3.39**	**0.69**
PLGC	8(10%)	72	85.23	77.32	7.59	3.49
CauSSL	8(10%)	72	85.89	76.58	7.25	3.28
MCF	8(10%)	72	87.12	78.03	6.82	3.62
CEMT(Ours)	8(10%)	72	87.03	78.26	3.39	0.67
UG-CEMT(Ours)	8(10%)	72	**88.16**	**79.83**	**3.08**	**0.51**
PLGC	16(20%)	64	88.23	80.09	6.25	2.89
CauSSL	16(20%)	64	88.87	79.68	6.02	3.16
MCF	16(20%)	64	89.05	81.12	5.16	2.75
CEMT(Ours)	16(20%)	64	89.12	80.94	3.78	0.66
UG-CEMT(Ours)	16(20%)	64	**89.73**	**81.63**	**2.20**	**0.50**

20% labeled data. The superior performance of UG-CEMT demonstrates its robustness across multiple imaging modalities and highlights its generalization capability on multi-site datasets. These quantitative results confirm that our proposed UG-CEMT method achieves better segmentation accuracy than existing SSL methods. By incorporating cross-attention and uncertainty-guided consistency regularization, UG-CEMT delivers more reliable results, especially when the available labeled data is scarce.

4. computational cost analysis

This supplementary material provides a detailed analysis of the computational costs for our UG-CEMT framework. We focus on the key metrics: FLOPs, parameters, memory usage, and training time, providing insight into the practicality of our method for real-world applications.

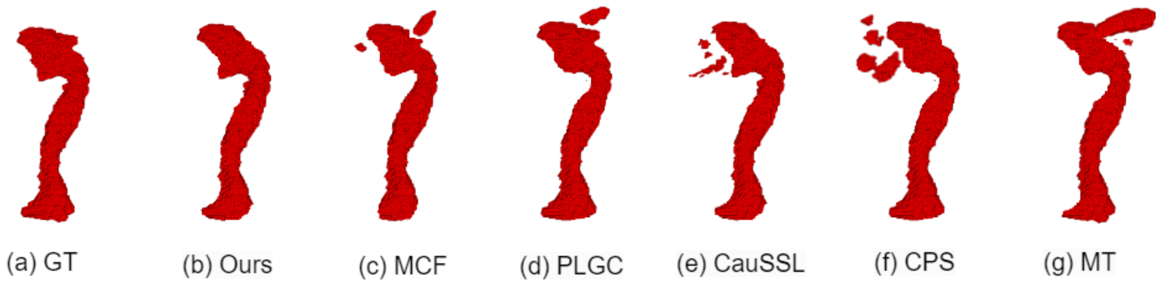

Figure 3. Visualization of 3D segmentation outcomes of various SSL methods for 20% labeled data on pancreas CT dataset.

Table 2. Comparison of the proposed UG-CEMT with other state-of-the-art methods on multi-site prostate dataset for 6000 iterations.

Method	(%) of images used		Metrics	
	labeled	unlabeled	Dice ↑	Jaccard ↑
PLGC	5(5%)	87	62.59	53.86
CauSSL	5(5%)	87	62.78	52.62
MCF	5(5%)	87	64.18	54.69
CEMT(Ours)	5(5%)	87	63.68	55.82
UG-CEMT(Ours)	5(5%)	87	**65.68**	**56.87**
PLGC	9(10%)	83	68.13	57.19
CauSSL	9(10%)	83	67.05	57.52
MCF	9(10%)	83	69.54	58.78
CEMT(Ours)	9(10%)	83	69.23	59.26
UG-CEMT(Ours)	9(10%)	83	**70.36**	**60.73**
PLGC	18(20%)	74	69.23	59.02
CauSSL	18(20%)	74	68.09	58.65
MCF	18(20%)	74	70.89	59.89
CEMT(Ours)	18(20%)	74	70.13	60.16
UG-CEMT(Ours)	18(20%)	74	**72.02**	**61.29**

UG-CEMT includes cross-attention mechanisms after each of the four decoder blocks in both student and teacher networks, enhancing segmentation performance. Table 3 shows a summary of the FLOPs and parameters. While the cross-attention mechanism adds computational complexity, the overall cost remains manageable for deployment on modern GPUs. UG-CEMT requires approximately 1843 MB of GPU memory, including a small overhead from the cross-attention mechanism (∼10MB). This makes the model suitable for single-GPU setups, standard in research and clinical environments. It was trained on an NVIDIA A6000 GPU (48GB). The average training time for 6000 iterations is approximately 1 hour 50 minutes. The cross-attention mechanism adds little overhead in terms of training time, ensuring efficient model training. Moreover, we compared the computational cost of UG-CEMT with other state-of-the-art SSL methods in Table 4. Although UG-CEMT slightly increases computational complexity due to cross-attention, it achieves significant improvements in segmentation performance while remaining computationally efficient.

Table 3. Layer wise summary of the FLOPs and parameters.

Layer/Component	**Params (Million)**	**FLOPs (GFLOPs)**
Encoder Blocks (total)	2.17	15.23
Decoder Blocks (total)	0.92	15.62
Cross-Attention Mechanism	≈ 0.2% overhead	≈ 0.001% overhead
Total	9.66M	47.1G

Table 4. Comparison of computational costs and training time for different semi-supervised learning methods.

Model	**FLOPs (GFLOPs)**	**Params(Million)**	**Training time (hrs)**
MCF	36.8	8.7	≈ 2.5
CauSSL	39.6	9.0	≈ 4.0
PLGC	40.2	8.8	≈ 3.2
MT	24.8	7.5	≈ 1.4
UA-MT	27.5	7.2	≈ 1.3
UG-CEMT (ours)	47.1	9.66	≈ 1.5

5. Experiments on Additional Modality: Pancreas CT Dataset

5.1. Dataset details:

The NIH Pancreas dataset is a publicly available dataset comprising 82 contrast-enhanced abdominal 3D CT volumes with manual annotations. Each CT volume has a size of $512 \times 512 \times D$, where $D \in [181, 466]$, representing the number of slices. For evaluation purposes, we divide the NIH pancreas dataset into 62 for training and 20 for testing. In the preprocessing stage, we apply the soft tissue CT window $[-120, 240]$ HU and crop the CT scans centered at the pancreas region with an additional margin of 25 voxels. The training volumes are randomly cropped to $96 \times 96 \times 96$ for training, and a stride of $16 \times 16 \times 16$ is applied during inference. This setup is consistent with the preprocessing strategy used in prior works, ensuring a fair comparison.

5.2. Implementation details:

The same network architecture and hyperparameter settings described in the main manuscript for LA and prostate segmentation are used for the pancreas CT dataset. For detailed parameter settings, network architecture, and training procedures, please refer to the implementation section of the main manuscript. The hardware setup, including the single NVIDIA A6000 GPU, is used to train the model with a total training time of approximately 1.5 hours for 6000 iterations.

5.3. Results and analysis

To further assess the generalizability of UG-CEMT across different medical imaging modalities, we conducted experiments on the NIH Pancreas CT dataset. The qualitative and quantitative results demonstrate the robustness of our model compared to other state-of-the-art SSL methods such as MT, CPS, PLGC, CauSSL, and MCF.

Quantitative Analysis: Table 5 presents the quantitative comparison of the proposed UG-CEMT framework with other SSL methods on the pancreas CT dataset. As shown, UG-CEMT outperforms the baseline methods across different metrics, including Dice and 95HD scores, mainly when using limited labeled data (10% and 20 %). With 10% labeled data, UG-CEMT achieved a Dice score of 71.23% and a 95HD of 17.01 mm, and outperforms methods such as MCF and PLGC. For the 20% labeled data setting, UG-CEMT attained the highest Dice score of 73.49% and a 95HD of 10.26 mm, indicating the model's superior performance in segmenting complex CT structures with limited labeled data. Compared to the fully supervised baseline V-VNet and B-VNet using 100% labeled data, UG-CEMT achieves competitive performance even with a significantly smaller labeled dataset, showcasing its effectiveness in semi-supervised learning setups. This further validates the ability of UG-CEMT to leverage unlabeled data effectively for improving segmentation outcomes.

Table 5. Comparison of the proposed UG-CEMT with other state-of-the-art methods on pancreas CT dataset for 6000 iterations.

Method	(%) of images used		Metrics	
	labeled	unlabeled	Dice	95HD
V-VNet	62(100%)	0	80.65	8.56
B-VNet	62(100%)	0	80.02	8.89
V-VNet	12(20%)	0	64.53	19.84
B-VNet	12(20%)	0	63.87	20.76
MT	6(10%)	56	68.07	18.68
CPS	6(10%)	56	67.28	22.59
PLGC	6(10%)	56	69.83	18.73
CauSSL	6(10%)	56	69.76	19.32
MCF	6(10%)	56	70.12	17.67
UG-CEMT(Ours)	6(10%)	56	**71.23**	**17.01**
MT	12(20%)	50	71.28	14.93
CPS	12(20%)	50	72.16	18.02
PLGC	12(20%)	50	71.58	13.69
CauSSL	12(20%)	50	72.02	12.87
MCF	12(20%)	50	72.18	11.59
UG-CEMT(Ours)	12(20%)	50	**73.49**	**10.26**

Qualitative Analysis: Figure 3 provides a qualitative comparison of the 3D segmentation results generated by different SSL methods on the pancreas CT dataset for 20% labeled data. The visualization clearly shows that UG-CEMT produces segmentation outputs that closely resemble the ground truth (GT) and are superior to other methods in terms of capturing the fine details of the pancreas region. In contrast, methods like MCF, CPS, and CauSSL exhibit over-segmentation or under-segmentation artifacts, leading to less accurate delineations of the pancreatic structure. UG-CEMT's ability to produce high-confidence predictions from uncertainty-guided maps (UGMs) further enhances its segmentation accuracy, especially in challenging areas. These results confirm that UG-CEMT can generalize effectively to different imaging modalities beyond MRI, such as CT, and provide reliable segmentation results across diverse datasets.