Supplementary Material: Annotation Ambiguity Aware Semi-Supervised Medical Image Segmentation

Suruchi Kumari Pravendra Singh*
Indian Institute of Technology Roorkee
{suruchi_k@cs.iitr.ac.in, pravendra.singh@cs.iitr.ac.in}

1. Architecture Design Analysis

To demonstrate the impact of the framework, an analysis is presented using different architectures. In architecture a, only the original encoder E^b_θ and decoder D^b_θ are utilized to process both labeled and unlabeled data. Initially, the pseudo-label set is generated from the decoder D^b_θ and is used to facilitate learning from unlabeled data. The same pseudo-label set is also employed in the posterior network to build a shared latent space. Similarly, for architecture b, E^b_θ and D^b_θ are utilized; additionally, the decoder D^b_ϕ is included. Here, D^b_θ is trained solely on labeled data. The pseudo-label set generated by D^b_θ is combined with the pseudo-label set from D^b_ϕ and provided to the posterior network. However, to train D^b_ϕ on unlabeled data, it only utilizes the pseudo-labels generated by D^b_θ. For architecture c, E^b_θ, D^b_ϕ, and D^b_ξ are used. Two pseudo-label sets are generated from both decoders and provided to the posterior network. Furthermore, cross-supervision is enabled by training D^b_ϕ and D^b_ξ on each other's pseudo-labels for the unlabeled data, ensuring mutual refinement and consistency between the decoders. A comparison of all three architectures with our framework is shown in Table 1.

Architecture	GED ↓	$Dice_{soft}$ ↑
a	0.3089	81.56
b	0.2967	82.33
c	0.2541	84.13
Ours	**0.2444**	**85.87**

Table 1. Comparison of different architectures on the ISIC dataset with 20 % labeled and 80 % unlabeled data.

2. Additional Implementation Details

For weak augmentation, we followed the approach described in [3] for data pre-processing and applied standard flip and rotation operations to augment the samples for both datasets. For strong augmentation, we applied techniques such as Gaussian noise, brightness adjustment, contrast enhancement, and salt-and-pepper noise to simulate diverse conditions and introduce random pixel-level perturbations.

Our method consists of a total of 31.60 million parameters. While the training process involves all three decoders, D^b_θ, D^b_ϕ, and D^b_ξ, once the training phase is complete. During the testing phase, only the decoder D^b_ϕ, the encoder E^b_θ, and the prior network E^{prior}_θ are utilized to generate final predictions. A testing time of 0.0110 seconds per image ensures practical applicability in real-world scenarios.

3. Impact of Strong Augmentations and Decoder Pruning on Model Performance

In this section, we analyze the impact of using strong augmentations alongside randomly pruned decoders on the final performance. Strong augmentations introduce data perturbations, while the random pruning of decoders adds diversity for unlabeled data by enabling varied feature representations. To evaluate the effect of these components (Table 2), we first present results where cross-decoder supervision occurs between D^b_ϕ and D^b_ξ without pruning or strong augmentations, referred to as **CDS-I**. Next, we apply strong augmentations to the unlabeled data, pass it through D^b_ξ, and perform cross-supervision between the decoders, naming this module **CDS-II**. Finally, our complete architecture, which incorporates both strong augmentations and random pruning to enhance feature diversity, is referred to as **CDS**.

Architecture	GED ↓	$Dice_{soft}$ ↑
CDS-I	0.2832	83.55
CDS-II	0.2742	84.08
CDS	0.2444	85.87

Table 2. Ablation study on showing the contribution of strong augmentation and pruned decoders in the final performance on the ISIC dataset with 20 % labeled and 80 % unlabeled data.

*Corresponding Author.

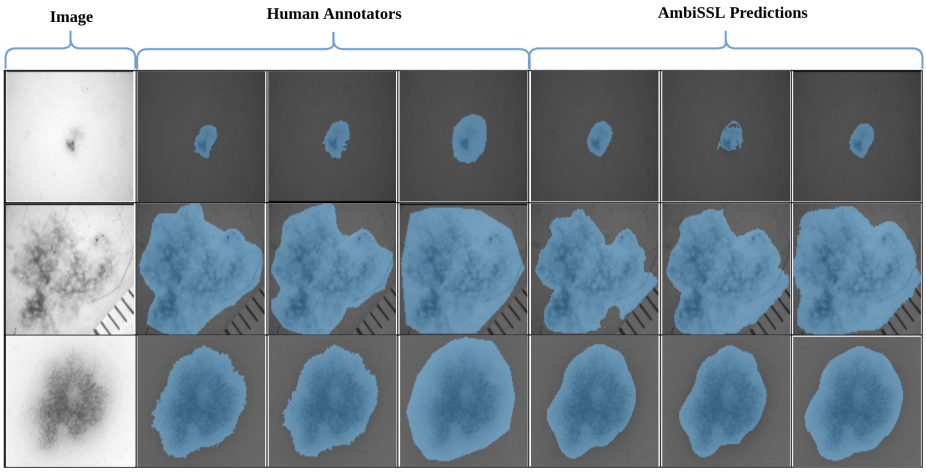

Figure 1. Comparison of segmentation results of our proposed AmbiSSL framework on the ISIC dataset with human annotators.

4. Visualization Results

The visualization results of the AmbiSSL framework for the ISIC dataset are presented in Figure 1. These results demonstrate its ability to produce diverse predictions that align with those of human annotators.

5. Additional Experiments

To further showcase the efficacy of our method, we evaluate it on the Gleason19 dataset [1]. The dataset consists of 333 Tissue Microarrays (TMAs) of prostate cancer, annotated by six different pathologists, and contains four classes. Among these, 244 images are publicly available with labels, as the test annotations from the challenge are not provided. Following the original dataset protocol [1], we resize all images to 1024×1024 pixels and create four cross-validation splits. With only 10% labeled data, our method achieves the lowest GED (0.341) and the highest Dice scores compared to other baselines (Table 3), demonstrating its effectiveness in utilizing unlabeled data for improved segmentation.

Method	Ratio		Diversity Performance		Personalized Performance (%)	
	Labeled	Unlabeled	$GED \downarrow$	$Dice_{soft} \uparrow$	$Dice_{max} \uparrow$	$Dice_{match} \uparrow$
Upper Bound	244(100%)	0	0.325	82.67	84.78	84.21
Pionono [2]	25(10%)	0	0.451	73.26	74.54	73.86
Baseline I	25 (10%)	219 (90%)	0.402	75.22	76.13	75.89
Baseline II			0.381	77.78	78.65	78.21
Baseline III			0.372	76.64	78.23	77.71
Ours			**0.341**	**80.22**	**81.78**	**81.34**

Table 3. Performance of our proposed framework on the Gleason19 dataset with 10% labeled data, rest is utilized as unlabeled data.

References

[1] Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180, 2018. 2

[2] Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina. Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In *ICCV*, pages 21097–21106, 2023. 2

[3] Lin Wang, Xiufen Ye, Lie Ju, Wanji He, Donghao Zhang, Xin Wang, Yelin Huang, Wei Feng, Kaimin Song, and Zongyuan Ge. Medical matting: Medical image segmentation with uncertainty from the matting perspective. *Computers in Biology and Medicine*, 158:106714, 2023. 1