A Appendix

Table 6: Details of ablation study on mixing ratio for mix-supervision based on PROMISE12.

GT	Noisy	Unlabeled	Dice (%)↑	JI (%)↑	HD(voxel)↓	ASD (voxel)↓
0%	0%	100%	76.68	63.14	7.85	2.64
0%	25%	75%	77.70	63.92	7.31	2.79
25%	0%	75%	79.02	65.56	6.93	2.37
0%	50%	50%	79.34	66.09	7.63	2.42
0%	75%	25%	80.15	67.00	7.02	2.52
25%	75%	0%	80.58	67.68	7.10	2.25
0%	100%	0%	80.83	68.10	6.68	2.10

Table 7: Details of ablation study on mixing ratio for mix-supervision based on LA dataset.

GT	Noisy	Unlabeled	Dice (%)↑	JI (%)↑	HD(voxel)↓	ASD (voxel)↓
0%	0%	100%	88.69	79.86	8.99	2.61
0%	25%	75%	88.79	80.10	9.88	2.87
25%	0%	75%	88.36	79.44	8.71	2.61
0%	50%	50%	88.60	79.67	8.25	2.32
0%	75%	25%	89.28	80.76	8.61	2.57
25%	75%	0%	89.51	81.18	8.15	2.51
0%	100%	0%	89.17	80.61	7.41	2.35

In this section, we give more details about the experiments setting and results. And code is publicly available at `https://anonymous.4open.science/r/MLB-Seg-C80E`.

Ablation study on different noise levels. We have conducted experiments on different noise levels where L$_1$, L$_2$, and L$_3$ represents that the corrupted ratios are around 60%, 40% , and 20% respectively. As shown in Table 8, we report the averaged dice coefficient over 5 repetitions for each series of experiments. The standard deviation for all experiments is within 0.5%. We could notice that while the noise level increases, performances of baseline drop from 80.03% to 59.77%, but performances of MLB-Seg only drop from 82.01% to 77.70% which indicates that our MLB-Seg is robust to different noisy levels and shows larger improvements under a much severer noisy situation.

Table 8: Ablation study on different noise levels

Method	Dice (%)↑
baseline - L$_1$	59.77
MLB-Seg - L$_1$	77.70
baseline - L$_2$	73.74
MLB-Seg - L$_2$	80.83
baseline - L$_3$	80.03
MLB-Seg - L$_3$	82.01

Experiments on different number of augmentations in PLE w/ and w/o mean teacher. Table 9 shows the averaged results over 5 repetitions for each series of experiments of the number of augmentations in PLE w/ and w/o mean teacher and standard deviations are all within 0.5%. We could notice that, while increasing the augmentation number in PLE, combining with mean teacher could help stabilize model performances and even improve the results when $Q = 4, 6$.

Table 9: Results of different number of augmentations in PLE w/ and w/o mean teacher based on PROMISE12

Method	Dice (%)↑
1 × Zoom in	74.34
1 × Zoom in + mean teacher	73.94
2 × Zoom in	74.99
2 × Zoom in + mean teacher	74.77
4 × Zoom in	72.07
4 × Zoom in + mean teacher	76.63
6 × Zoom in	70.91
6 × Zoom in + mean teacher	75.84

Visualization of weight map. As shown in Fig. 4, we also display the visualization of the weight maps. The blue/purple represents for imperfect annotation/prediction in $y^{\tilde{n}}$/ y^{p}. The red indicates pixels in w^{p*} have higher values. This shows that during training, the meta-learned weight maps could reassign higher values to pixels that are more accurate in pseudo labels and thus, could effectively alleviate the negative effects of imperfect pixels.

Data augmentation details in PLE. While applying zoom in/out in PLE at each training step, we would randomly pick a cropping or padding size from 4 to 30, denoted as c or p. Specifically, for

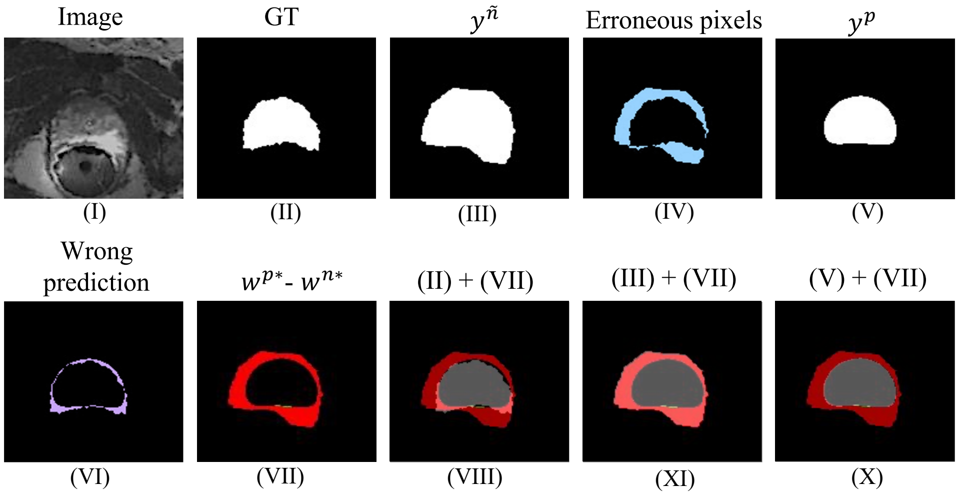

Figure 4: Weight map visualization.

zoom in, the input will first be resized to $[144 + 2 * c, 144 + 2 * c]$ and then center cropped to the size of $[144, 144]$. As for zoom out, the input will first be resized to $[144 - 2 * p, 144 - 2 * p]$, and then be zero padding to $[144, 144]$. As for flip, we use horizontally, vertically, or horizontally and vertically flip with a possibility of $\frac{1}{3}$ for each flip type.

More training details on meta-learning. For meta-learning, we apply SGD to optimize network parameters with a learning rate at 0.005 and the decay in the learning rate is at $\frac{20}{20+epoch}$. Under the setting of 4 × PLE w/ or w/o mean teacher, we set 1 as the batch size for the imperfect training data and 2 for the clean data used in the meta-update process. For 2 × PLE w/ or w/o mean teacher, we use 2 and 4 as the batch size for the imperfect and clean data respectively. For MLB without PLE strategy, we use 4 as the batch size for both imperfect and clean data. And during the meta-update process, we also apply the same PLE strategy used in the imperfect training data to the clean data. and in each experiment, we train the network for 100 epochs.

More details on synthesizing noisy annotations. For each ground-truth label, we discard labels of small size and set them all zero. Then we apply random rotation to the target. Rotation degree is randomly selected from $-20°$ to $20°$. Then we randomly apply erosion or dilation with a possibility of 0.5 for each operation.

More examples of the averaged meta-learned weights w/ and w/o mean teacher. To further show the instability using MLB-Seg w/o mean teacher and the benefit of MLB-Seg w/ mean teacher while increasing augmentation numbers for PLE, we show more examples in Fig.5. And Results are acquired using fixed networks and applying augmentations multiple times of meta-update respectively.

Examples of the corrupted noisy labels. In Fig.6, we give some examples of the corrupted noisy labels (the second row) and its corresponding ground-truth labels (the first row). They are generated using a combination of random rotation, erosion or dilation, following **??**.

Limitations and social impact. Our proposed MLB-Seg could be a very effective method for medical image segmentation under imperfect supervision (*e.g.,* semi- and noisy-supervision) which is quite common in the real world. Future work needs to address different types of imperfect supervision including weak supervision (e.g., bounding boxes), etc. Besides, given that there are currently no real-world noisy medical image segmentation benchmarks publicly accessible, our experiments are only conducted on the datasets with synthesized noisy annotations. And more experiments should be done on the different real-world noisy datasets to further evaluate the robustness of MLB-Seg.

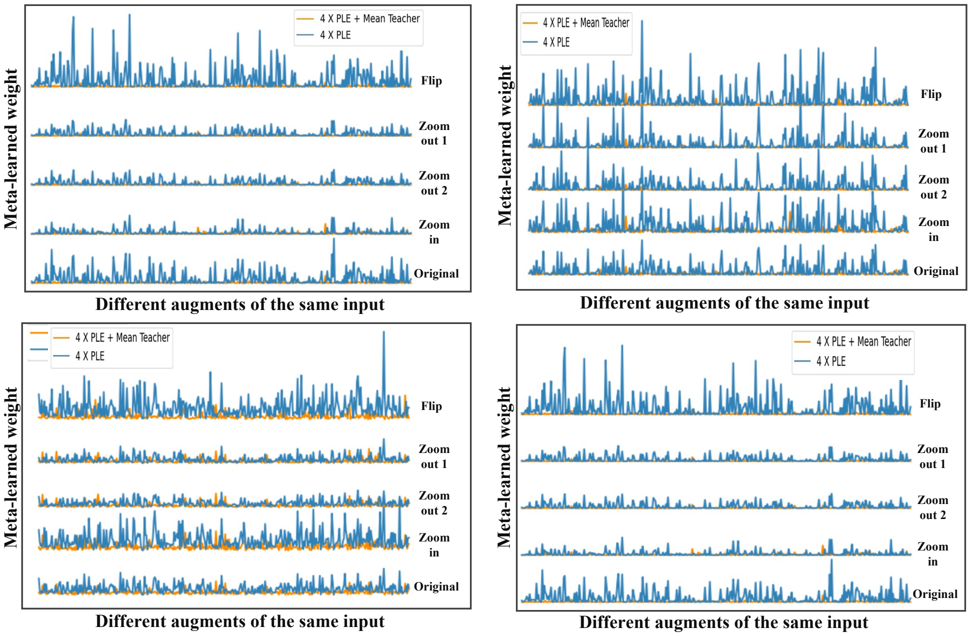

Figure 5: Average meta-learned weights of augmented variants w/ and w/o mean teacher. Blue line represents the average meta-learned weights of different augmented samples from one sample while using 4 × PLE and orange line represents using 4 × PLE w/ mean teacher.

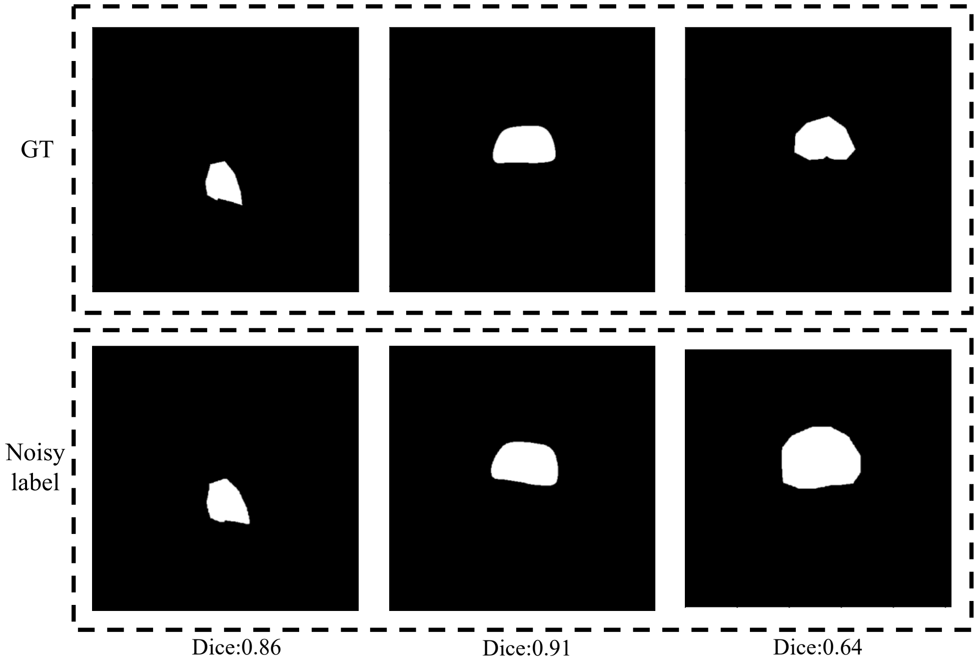

Figure 6: Examples of the corrupted noisy labels.