Table 3: Ablation experiments on the HVSMR 2016 dataset. "GT" represents ground truth and "PL" represents pseudo labels. Transductive learning setting: Test image data are involved as unlabeled data in model training.

Setting	Inputs		Supervision of training set	Transductive learning	Supervision of testing set	Training		Overall score
	raw image (x_i)	$S(PL_i)$				random fit	NN fit	
S1		✓	GT					0.075
S2	✓	✓	GT					0.192
S3	✓	✓	GT	✓	PL	✓		0.217
S4	✓	✓	GT + PL	✓	PL	✓		0.205
S5	✓	✓	GT + PL	✓	PL	✓	✓	0.224
S6	✓	✓	PL			✓		0.199
S7	✓	✓	PL			✓	✓	0.215
S8	✓	✓	PL	✓	PL	✓		0.218
S9	✓	✓	PL	✓	PL	✓	✓	**0.234**

dataset, and 0.9970 *vs.* 0.9967 on the piriform dataset).

4.3 Ablation study

Average ensemble *vs.* naïve meta-learner *vs.* our best. The results of the average ensemble of all the base-learners (the 2D and 3D models) are shown in Tables **??** and 2. One can see that the average ensemble is consistently worse than our meta-learner ensemble. We also compare our meta-learner with the naïve meta-learner implementation (in which the outputs of the base-learners are used as input and the ground truths of the training set are used to train the meta-learner). Table 3 shows the results (the S1 row). One can see that the naïve meta-learner implementation is even worse than the average ensemble (probably due to over-fitting). This demonstrates the effectiveness of our meta-learner structure design and training strategy.

Random-fit + NN-fit *vs.* Random-fit alone. Random-fit + NN-fit performs significantly better than Random-fit alone (Table 3: S7>S6, S5>S4, S9>S8; Table 2), which demonstrates that NN-fit can help the training procedure converge and thus improve the segmentation quality.

Model training using pseudo-labels *vs.* ground truth. One may concern that our meta-learner training method totally discards manual-labeled ground truth even when it is available. This ablation study shows that our method can perform better without using any manual ground truth. We explore the following ways of utilizing ground truth. When using only the training data, we compare the difference between only ground truth (S2) and only pseudo-labels (S7). Table 3 shows that our training method can achieve better results ($0.215 > 0.192$) when not using ground truth. When utilizing the test data (the transductive setting), we compare the difference between (1) only ground truth (S3), (2) mix of ground truth and psuedo-labels, i.e., using ground truth as the 5th version (S4 & S5), and (3) only pseudo-labels (S8 & S9). In Table 3, one can see that (a) using pure ground truth or pure pseudo-labels achieves better results than mixing them together (probably due to the different nature of ground truth and pseudo-labels), and (b) using only pseudo-labels is still better than using ground truth (S8 > S3). We think the reason that our method can work well with only pseudo-labels is because the pseudo-labels have already effectively distilled the knowledge from ground truth (**?**).

Table 4: Semi-supervised setting on HVSMR 2016 dataset.

Group	Model	Overall score
A	Base-learner 3D	-0.036
	Meta-learner	0.063
B	Base-learner 3D	-0.045
	Meta-learner	0.038

5 Conclusions

In this paper, we presented a new *ensemble learning* framework for 3D biomedical image segmentation that can retain and combine the merits of 2D and 3D models. Our approach consists of (1) diverse and accurate base-learners by leveraging diverse geometric and model-architecture perspectives of multiple 2D and 3D models, (2) a fully convolutional network (FCN) based meta-learner that is capable of learning robust visual features/representations to improve the base-learners' results, and (3) a new meta-learner training method that can minimize the risk of over-fitting and utilize unlabeled data to improve performance. Extensive experiments on two public datasets show that our approach can achieve superior performance over the state-of-the-art methods.

6 Acknowledgments

This research was supported in part by the U.S. National Science Foundation through grants CCF-1617735, IIS-1455886, and CNS-1629914.

Doloremque excepturi perspiciatis quo at pariatur vel, veritatis dolore obcaecati at, possimus corporis culpa quibusdam non saepe commodi laborum nihil sapiente excepturi ex, tempore quae ullam ut pariatur in iure debitis rem eligendi voluptas.Cumque explicabo fugiat distinctio, recusandae perspiciatis ab.Aperiam praesentium doloremque vel expedita atque dignissimos adipisci exercitationem, dolore totam obcaecati sit nesciunt ipsam eum, maiores expedita quae ullam a ducimus vel ipsam vitae iure nihil, enim ullam vitae sit cum laudantium pariatur a aspernatur veritatis autem earum?Minima veniam illo quia quod ad, nemo unde officiis rem a, voluptate ipsum quis in possimus ea deserunt quam necessitatibus voluptas incidunt?Sapiente nisi odit maiores doloribus ab ut, adipisci corrupti saepe ratione neque eveniet repudiandae voluptatum dolorem itaque hic.Tenetur cumque vel dolores ea magni sint incidunt distinctio quis, ab delectus rem ipsam eveniet