Supplementary Material: Self-similarity Student for Partial Label Histopathology Image Segmentation

Hsien-Tzu Cheng*[1], Chun-Fu Yeh*[1], Po-Chen Kuo[1,3], Andy Wei[1], Keng-Chi Liu[1], Mong-Chi Ko[1], Kuan-Hua Chao[1], Yu-Ching Peng[2], and Tyng-Luh Liu[1,4]

[1] Taiwan AI Labs
[2] Taipei Veterans General Hospital
[3] National Taiwan University College of Medicine
[4] Institute of Information Science, Academia Sinica, Taiwan

1 Augmentations for Model Training

We apply augmentations η_t and η_s for training teacher-studnet models [1, 2, 6] on partial label WSIs. Several augmentations are chosen as shown in Table. 1. For network dropout, we set all teacher models' dropout to 0.0 whereas student models' following Table. 2. We use Noisy augmentation hyperparameters to train the student model in Noisy Student, and Normal augmentation hyperparameters to train the other teacher-student models.

Table 1. Augmentations.

Augmentations		Hyperparameters	
		Normal	Noisy
Stain/Color	Contrast (delta)	(0.75, 1.25)	(0.5, 1.875)
	Brightness (delta)	(-0.2, 0.2)	(-0.3, 0.3)
	Hue (delta)	(-0.05, 0.05)	(-0.075, 0.075)
	Saturation (delta)	(0.8, 1.2)	(0.533, 1.8)
Deformation	Flip (probability)	0.5	
	Resize (scale)	(0.9, 1.1)	(0.6, 1.35)
	Crop (resolution)	(224, 224)	
	Rotation (degree)	(-180, 180)	
	Translation (delta)	(-0.05, 0.05)	(-0.075, 0.075)
Network	Dropout (ratio)	0.2	0.5

2 Implementation Details of Baseline Methods

In this section, we describe more about the implementation of our baseline previous art methods. We implement Mean Teacher [5], Noisy Student [6], and

* Both authors contributed equally to this work.

Algorithm 1 Teacher-student Model

Require: $\{P_{train}, P_{val}\} = P$ ▷ Noisy set of patches sampled from D
Require: $\alpha^b_{mt}, \alpha^e_{mt}, \alpha_{pred}, \lambda \in (0, 1) \subset \mathbb{R}$ ▷ EMA momentum and $\mathcal{L}_{cs}$ weight
Require: $ep_{max} \in \mathbb{N}$ ▷ Distance threshold and max epoch
Require: $\mathcal{O}$ = model weights gradient optimizer, e.g. Adam
  Initialize $f_t(\theta_t, \eta_t)$ ▷ Initialize teacher model
  Initialize $f_s(\theta_s, \eta_s)$ ▷ Initialize student model
  $\hat{P} \leftarrow P_{train}$ ▷ Initialize all pseudo label $(p, \hat{y})$
  for $ep \leftarrow 0$, ep_{max} **do** ▷ Main training loop
    for all $(p, \hat{y}) \in \hat{P}$ **do**
      $y_s, z_s \leftarrow f_s(p)$ ▷ Student forward
      $y_t, z_t \leftarrow f_t(p)$ ▷ Teacher forward
      $loss_{ce} \leftarrow \mathcal{L}_{CE}(\hat{y}, y_s)$ ▷ Cross entropy loss
      $loss_{cs} \leftarrow \mathcal{L}_{CS}(z_s, z_t)$ ▷ Consistency loss (Eq. 1)
      $\theta_s \leftarrow \mathcal{O}(\theta_s, loss_{ce} + \lambda loss_{cs})$ ▷ Update student's weights
      $\theta_t \leftarrow \alpha^b_{mt}\theta_t + (1 - \alpha^b_{mt})\theta_s$ ▷ Update teacher's weights per batch
    end for
    $\theta_t \leftarrow \alpha^e_{mt}\theta_t + (1 - \alpha^e_{mt})\theta_s$ ▷ Update teacher's weights per epoch
    for all $(p, \hat{y}) \in \hat{P}$ **do** ▷ Predictions ensemble
      $\hat{y} \leftarrow \alpha_{pred}\hat{y} + (1 - \alpha_{pred})y_t$ ▷ Update pseudo label
    end for
  end for

Table 2. Hyper parameters of baseline previous arts.

	α^b_{mt}	α^e_{mt}	α_{pred}	λ
Mean Teacher	0.999	1	0	1
Noisy Student	1	0	0	0
Pred-ensemble	0.999	1	0.9	0

Pred-ensemble [4] for comparison with our Self-similarity Student method. The teacher-student model training pipeline is illustrated in Algorithm 1. In each epoch, both teacher model and student model inference the noisy label patch p. The supervised loss between the (pseudo) label $\hat{y}$ and student model prediction y_s is the calculated by cross entropy loss L_{CE}. Following [5], the consistency loss L_{CS} (Eq. 1) is used to constrain the feature map outputs of teacher model f_t and student model f_s to have similar predictions:

$$\mathcal{L}_{CS}(z_t, z_s) = \|z_t - z_s\|^2 \tag{1}$$

where z_t and z_s are feature maps from fully-connected layers of teacher-student models. After the optimizer $\mathcal{O}$ backpropagated the loss, model weights ensemble and predictions ensemble are applied according to different settings respectively. Table. 2 shows all the hyperparameters settings of different previous art baselines. For per-batch EMA momentum α^b_{mt} and per-epoch EMA momentum α^e_{mt}, 1 denotes no weights update and 0 denotes entirely weights replacement of teacher model from student model. For predictions ensemble momentum α_{pred}, 1 denotes no label update and 0 denotes the entirely pseudo label update by

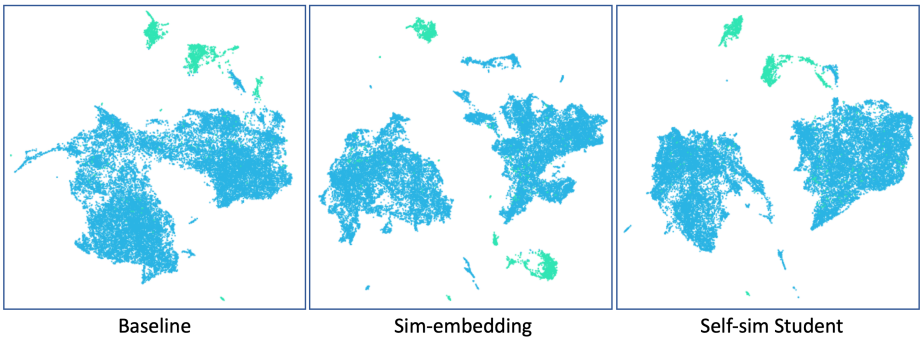

Fig. 1. Qualitative result of the distribution of feature embeddings with UMAP. The cyan colored points denote benign patches and light-green colored points denote cancerous patches. Our method is able to learn a more compact feature representation for both benign and cancerous patches.

the teacher model prediction. As indicated in Table. 2, following their original settings, we update Mean Teacher and Pred-ensemble every batch while update Noisy Student every epoch. We set $\alpha_{pred} = 0.9$ for Pred-ensemble and only apply $\mathcal{L}_{CS}$ on Mean Teacher. Note that we use pseudo label mechanism, instead of eliminating patches by noisy label filtering, to stabilize the training process without overly filtering.

3 Feature Embedding of Self-similarity Student

To illustrate the effectiveness of similarity learning, we use UMAP [3] dimension reduction algorithm to visualize the feature embeddings derived from our method and the baselines. Specifically, we sample 30000 patches from our testing set $\mathring{P}$ and apply UMAP on the feature embeddings (n dimension=1024). As shown in Fig. 1, with similarity learning, Self-similarity Student can learn a more compact distribution of feature embeddings, which support its advantage in identifying cancerous patches over the baseline.

4 Additional Qualitative Result

More qualitative results on TVGH TURP dataset are illustrated in Fig. 2. Moreover, the $k_{top} = 1$ results on CAMELYON16 dataset are shown in Fig. 4, and Fig. 3. Our Self-sim Student consistently outperforms other previous arts in multiple morphology patterns on TVGH TURP cancer dataset and CAMELYON16 dataset.

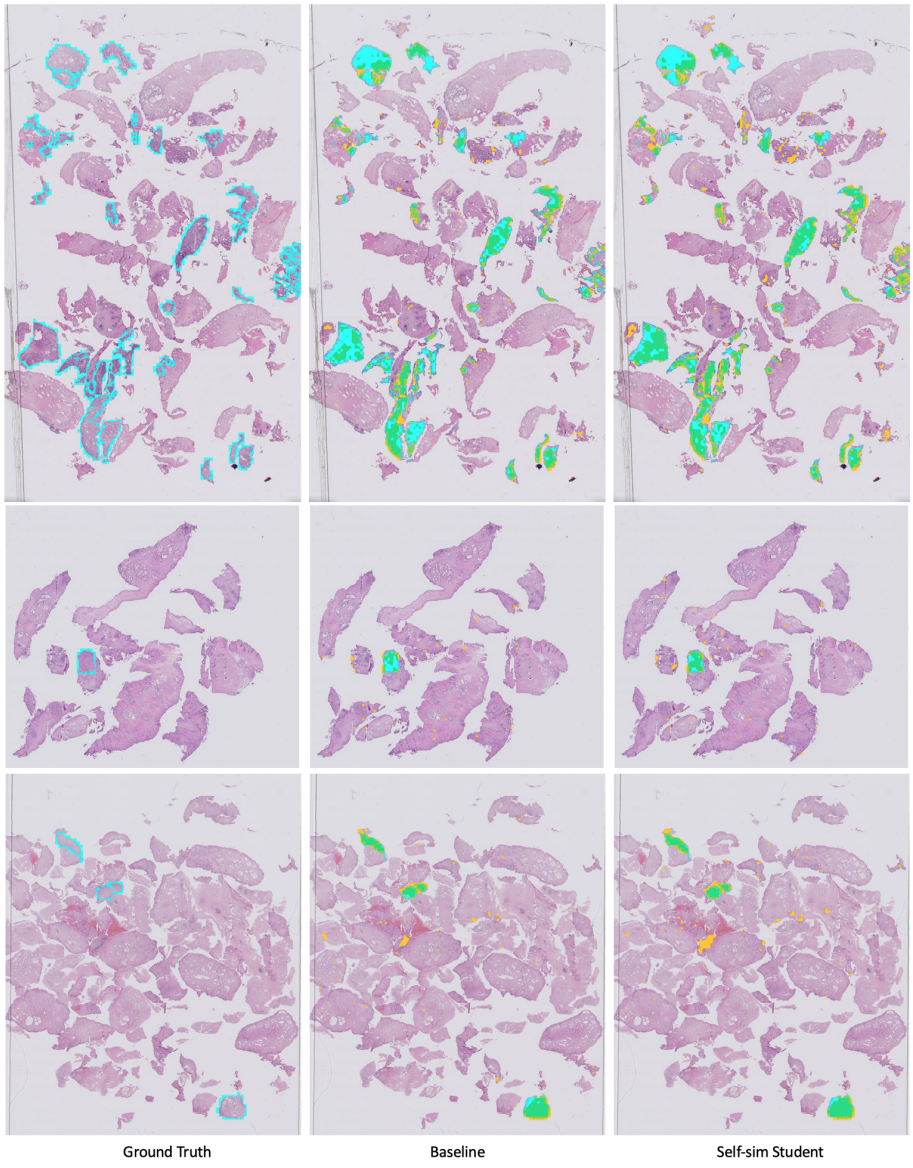

Fig. 2. Additional qualitative result on TVGH TURP dataset. The regions in green color indicate true positives and yellow indicate false positives. The cyan color denotes ground truth (false negatives if no prediction overlapped).

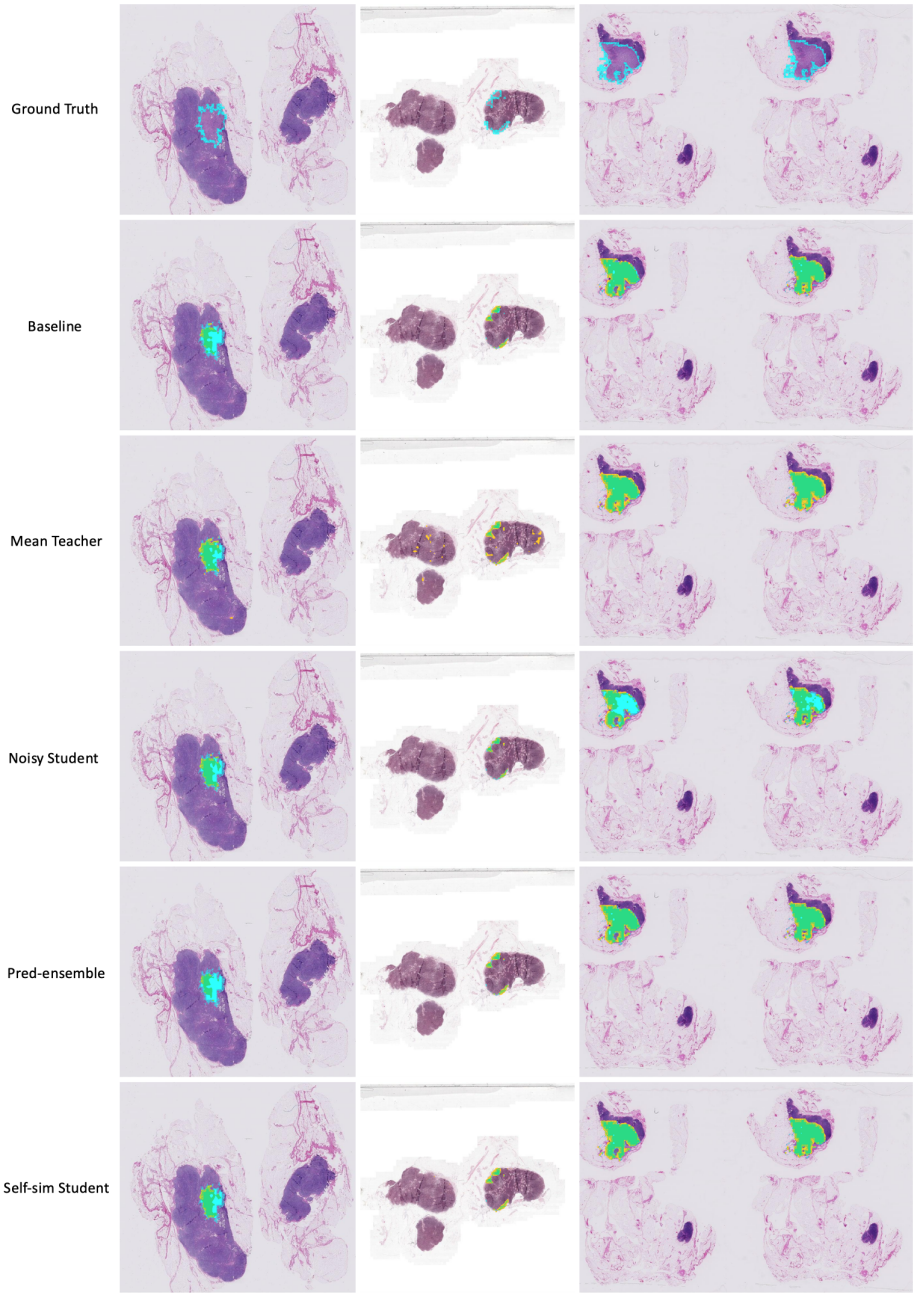

Fig. 3. Additional qualitative result on CAMELYON16 dataset. The regions in green color indicate true positives and yellow indicate false positives. The cyan color denotes ground truth (false negatives if no prediction overlapped).

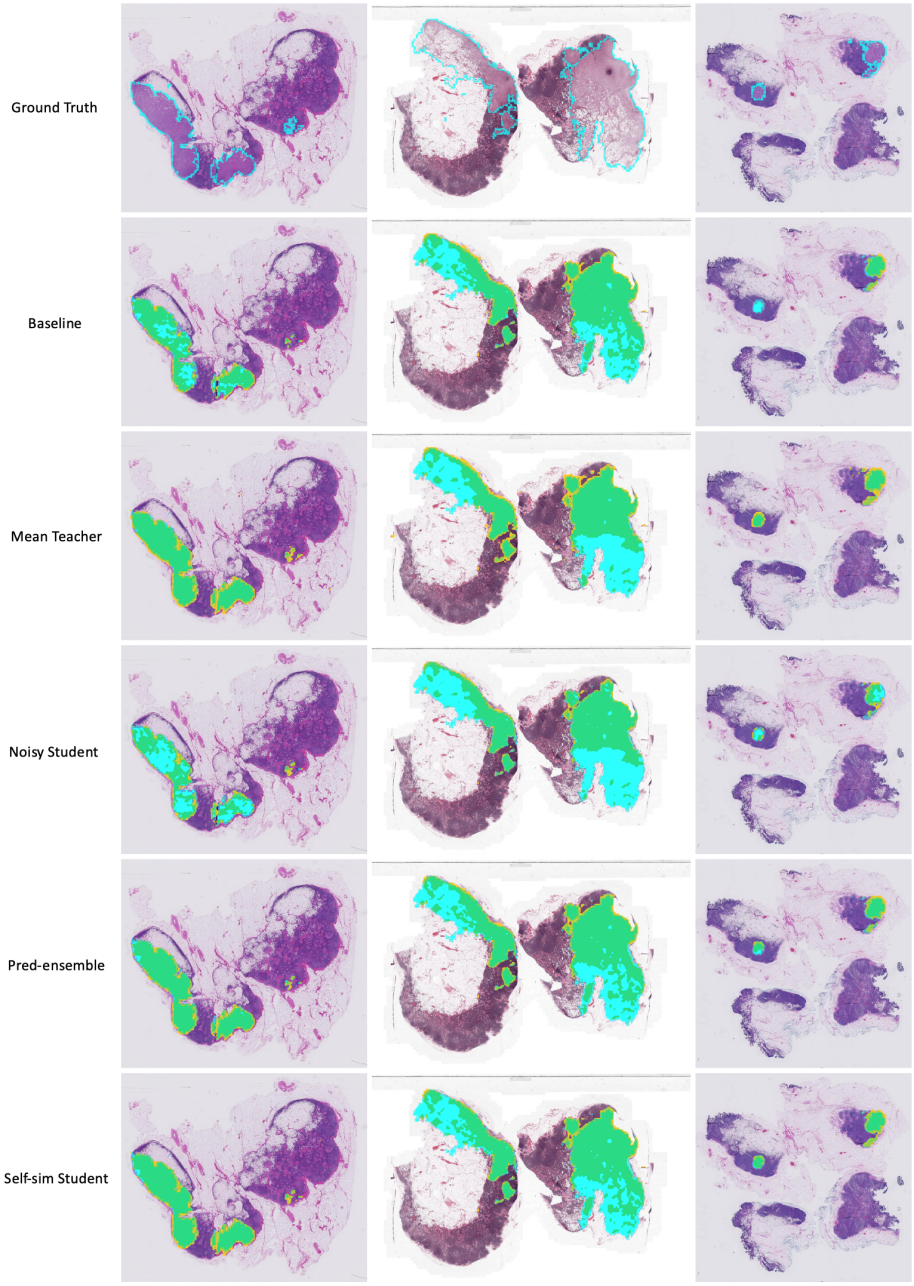

Fig. 4. Additional qualitative result on CAMELYON16 dataset. The regions in green color indicate true positives and yellow indicate false positives. The cyan color denotes ground truth (false negatives if no prediction overlapped).

References

1. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. arXiv preprint arXiv:1909.13719 (2019)
2. Lee, B., Paeng, K.: A robust and effective approach towards accurate metastasis detection and pn-stage classification in breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 841–850. Springer (2018)
3. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints (Feb 2018)
4. Nguyen, D.T., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Beggel, L., Brox, T.: Self: Learning to filter noisy labels with self-ensembling. arXiv preprint arXiv:1910.01842 (2019)
5. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems. pp. 1195–1204 (2017)
6. Xie, Q., Hovy, E., Luong, M.T., Le, Q.V.: Self-training with noisy student improves imagenet classification. arXiv preprint arXiv:1911.04252 (2019)