Semi-MoE: Mixture-of-Experts meets Semi Supervised Histopathology Segmentation

Nguyen Lan Vi Vu*[1,2], Thanh-Huy Nguyen*[1], Thien Nguyen*[3], Daisuke Kihara[4], Tianyang Wang[5], Xingjian Li[1], Min Xu[1]

[1] Carnegie Mellon University [2] Ho Chi Minh City University of Technology [3] Posts and Telecommunications Institute of Technology [4] Purdue University [5] University of Alabama at Birmingham

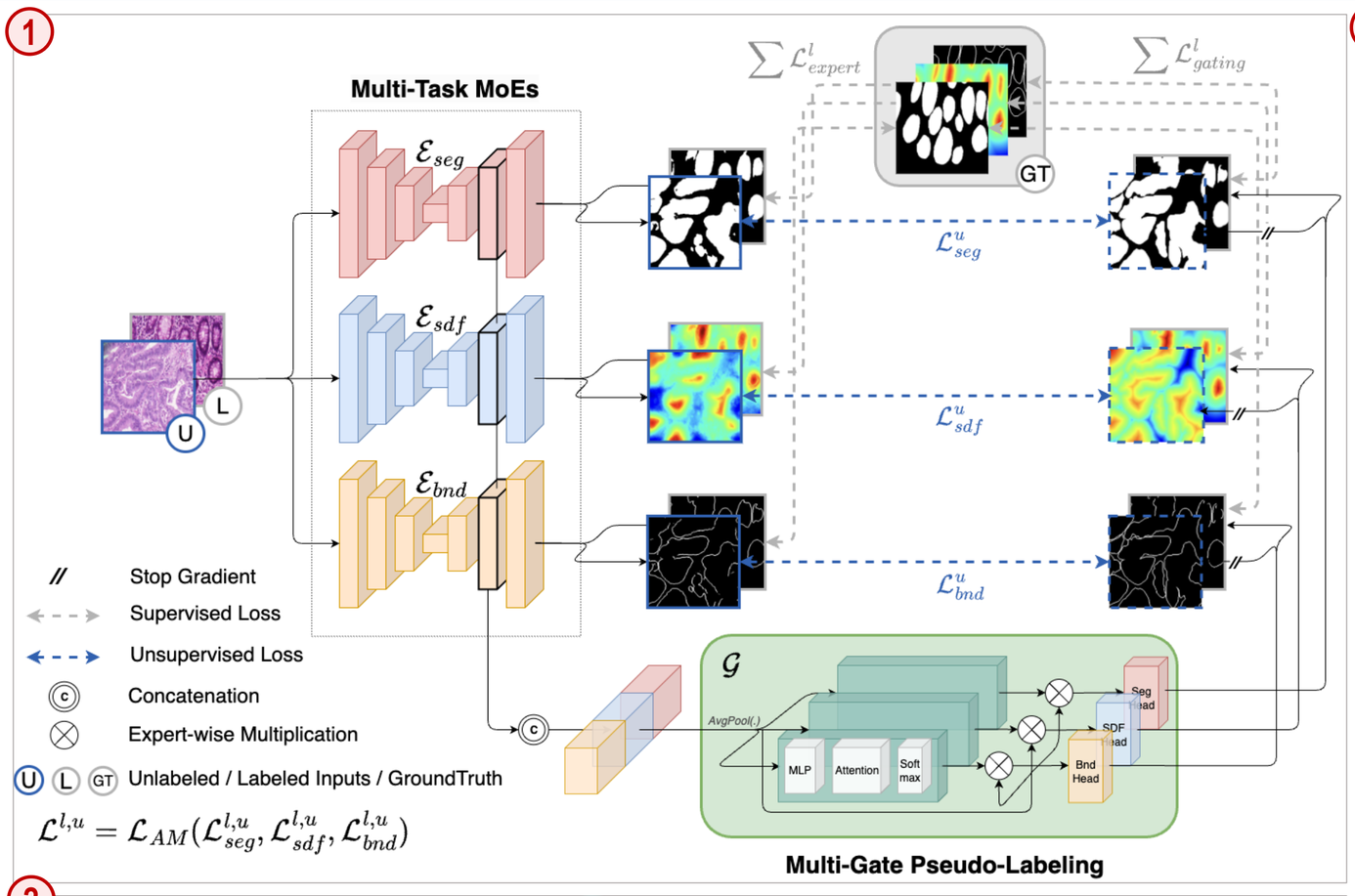
Abstract

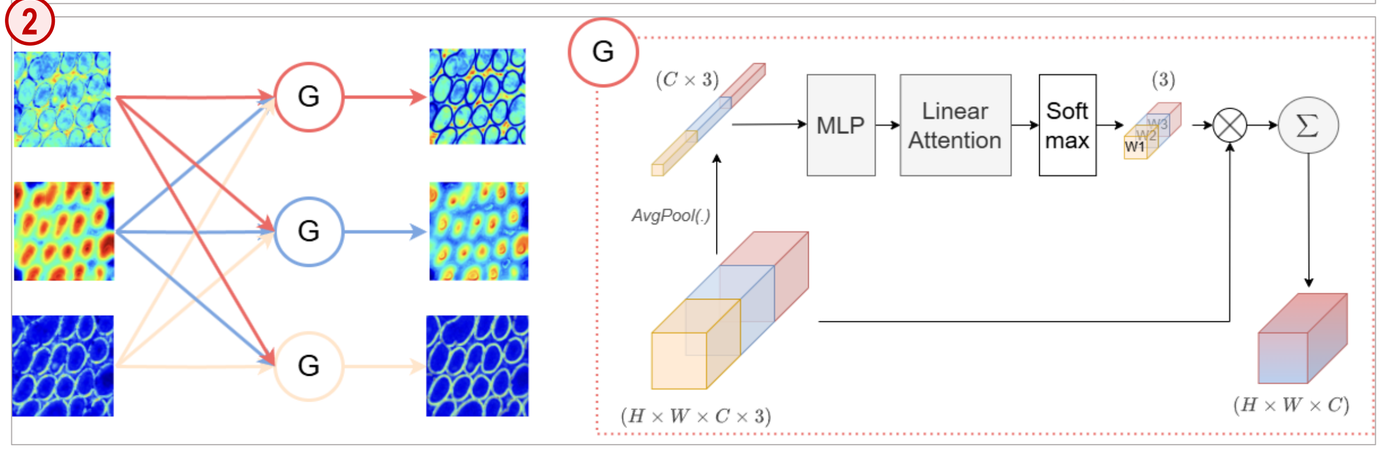
Semi-supervised learning has been employed to alleviate the need for extensive labeled data for histopathology image segmentation, but existing methods struggle with noisy pseudo-labels due to ambiguous gland boundaries and morphological misclassification. This paper introduces **Semi-MOE**, to the best of our knowledge, the **first multitask Mixture-of-Experts framework for semi-supervised histopathology image segmentation**. Our approach leverages **(1) three specialized expert networks**: A main segmentation expert, a signed distance field regression expert, and a boundary prediction expert, each dedicated to capturing distinct morphological features. Subsequently, the **(2) Multi-Gating Pseudo-labeling** module dynamically aggregates expert features, enabling a robust fuse-and-refine pseudo-labeling mechanism. Furthermore, to eliminate manual tuning while dynamically balancing multiple learning objectives, we propose an **(3) Adaptive Multi-Objective Loss**. Extensive experiments on GlaS and CRAG benchmarks show that our method outperforms state-of-the-art approaches in low-label settings, highlighting the potential of MoE-based architectures in advancing semi-supervised segmentation.

Contributions

1. **Semi-MoE:** A multi-task Mixture-of-Experts framework combining segmentation, boundary, and signed distance experts for robust tissue representation.
2. **Multi-Gate Pseudo-Labeling (MGP):** Dynamically aggregates expert predictions to refine pseudo labels and enhance semi-supervised learning reliability.
3. **Adaptive Multi-Objective Loss:** Balances task-specific objectives via uncertainty weighting, reducing manual tuning.
4. **Results:** Achieves state-of-the-art Dice and Jaccard scores on CRAG and GlaS histopathology datasets.

Proposed Method





3 **Goal**: Dynamically balance segmentation, boundary, and SDF objectives without manual weight tuning.

Idea: Each task learns its own uncertainty parameter (σ), which automatically adjusts its contribution to the total loss.

Formulation:

$$\mathcal{L}_{AM} = e^{-\sigma_{\text{seg}}}\mathcal{L}_{\text{seg}} + e^{-\sigma_{\text{sdf}}}\mathcal{L}_{\text{sdf}} + e^{-\sigma_{\text{bnd}}}\mathcal{L}_{\text{bnd}} + \gamma\sum_j e^{\sigma_j}$$

Benefit: Provides **stable training** and **adaptive weighting** across heterogeneous experts, improving convergence and performance consistency.

Algorithm 1 Training Semi-MoE

Require: Labeled set $\mathbf{D}_l = \{(X^l_i, Y^l_i)\}_{i=1}^N$, unlabeled set $\mathbf{D}_u = \{X^u_k\}_{k=1}^K$; experts $\{\mathcal{E}_m, \mathcal{C}_m\}_{m\in\{\text{seg,sdf,bnd}\}}$ with task uncertainty parameters $\{\sigma_m\}$; multi-gating module $\mathcal{G}$.
Ensure: Trained Semi-MoE model.

for each training iteration **do**
  Sample mini-batches $X^l \sim \mathbf{D}_l$, $X^u \sim \mathbf{D}_u$.
  Obtain expert predictions $P^{u,l}_m = \mathcal{C}_m(\mathcal{E}_m(X^{u,l}))$.
  Fuse expert representations via multi-gating: $G^{u,l}_m = \mathcal{G}(\oplus_m \mathcal{E}_m(X^{u,l}))$.
  if $X \in \mathbf{D}_l$ **then** ▷ Supervised phase
    Compute $\mathcal{L}_{\text{sup}}$ between P^l_m and Y^l_m (Eq. 4), task loss weighted by $\{\sigma_m\}$ (Eq. 11).
    Update parameters of $\{\mathcal{E}_m, \mathcal{C}_m, \mathcal{G}, \sigma_m\}$ via backpropagation.
  end if
  if $X \in \mathbf{D}_u$ **then** ▷ Unsupervised phase
    Derive pseudo-labels $\hat{Y}^u_m$ from G^u.
    Compute $\mathcal{L}_{\text{unsup}}$ between P^u_m and $\hat{Y}^u_m$ (Eq. 6), task loss weighted by $\{\sigma_m\}$ (Eq. 11).
    Update parameters of $\{\mathcal{E}_m, \mathcal{C}_m, \sigma_m\}$ via backpropagation.
  end if
end for

Experimental Results & Conclusion

Dataset & Evaluation

CRAG: 213 images (173 train / 40 test, 1512×1516 px), challenging gland morphology variability.
GlaS: 85 train (37 benign, 48 malignant) / 80 test images, 775×522 – 589×453 px, diverse cancer stages.
Training setup: 10 % and 20 % labeled data, 3-fold cross-validation.

Implementation

Backbone: U-Net, trained 200 epochs on RTX 3090 Ti.
Pre-processing: Resize 256×256, random crop 224×224, augment with flips & rotations.

Table 1: Quantitative experiments on GlaS and CRAG datasets. **Bold** and underline denote the best and second-best performance.

Data	Labeled ratio	20%		10%	
	Method	DSC (%) ↑	Jaccard (%) ↑	DSC (%) ↑	Jaccard (%) ↑
GlaS	MT [29]	89.46 ± 0.83	80.94 ± 1.35	87.65 ± 0.78	78.02 ± 1.23
	EM [30]	88.53 ± 0.34	79.42 ± 0.56	83.64 ± 0.56	71.89 ± 0.79
	UA-MT [37]	89.27 ± 0.12	80.62 ± 0.19	84.41 ± 0.56	73.03 ± 0.83
	CCT [23]	89.49 ± 0.09	80.98 ± 0.15	85.46 ± 0.42	74.51 ± 0.64
	CPS [8]	90.25 ± 0.53	82.23 ± 0.92	89.15 ± 0.44	80.43 ± 0.31
	URPC [19]	88.92 ± 0.45	80.05 ± 0.74	81.59 ± 0.62	68.9 ± 0.88
	CT [28]	88.45 ± 0.3	79.29 ± 0.49	81.65 ± 0.61	68.99 ± 0.87
	XNet [40]	88.68 ± 0.42	79.67 ± 0.67	84.44 ± 0.5	73.07 ± 0.75
	TopoSemiSeg [33]	89.5 ± 0.0	81.8 ± 0.0	87.8 ± 0.0	79.7 ± 0.0
	Semi-MoE (Ours)	**90.4 ± 0.32**	**82.48 ± 0.51**	**89.23 ± 0.41**	**80.55 ± 0.45**
CRAG	MT [29]	89.77 ± 0.1	81.43 ± 0.15	86.82 ± 0.09	76.71 ± 0.14
	EM [30]	87.55 ± 0.5	77.93 ± 0.8	84.13 ± 0.78	72.61 ± 1.17
	UA-MT [37]	88.84 ± 0.11	80.01 ± 0.06	83.97 ± 0.46	72.75 ± 0.6
	CCT [23]	89.5 ± 0.12	80.99 ± 0.19	85.28 ± 0.12	74.34 ± 0.18
	CPS [8]	89.42 ± 0.15	80.87 ± 0.25	85.52 ± 0.31	74.7 ± 0.47
	URPC [19]	88.75 ± 0.73	79.25 ± 0.25	82.5 ± 0.31	72.1 ± 0.61
	CT [28]	86.96 ± 1.19	75.9 ± 0.33	80.33 ± 0.68	67.13 ± 0.95
	XNet [40]	88.1 ± 0.15	79.1 ± 0.29	86.67 ± 0.05	76.48 ± 0.09
	TopoSemiSeg [33]	89.8 ± 0.0	82.0 ± 0.0	88.4 ± 0.0	79.8 ± 0.0
	Semi-MoE (Ours)	**91.06 ± 0.08**	**83.59 ± 0.11**	**89.27 ± 0.12**	**80.61 ± 0.18**

Table 2: Ablation study on the impact of task-specific experts (seg is the main task).

Expert	Seg	Sdf	Bnd	20%		10%	
Data				DSC	JC	DSC	JC
GlaS	✓	✓		87.93	78.46	85.34	74.43
	✓		✓	89.04	80.25	88.81	79.87
	✓	✓	✓	**90.4**	**82.48**	**89.23**	**80.55**
CRAG	✓	✓		90.73	83.03	85.85	75.21
	✓		✓	90.71	83.00	88.37	79.16
	✓	✓	✓	**91.06**	**83.59**	**89.27**	**80.61**

Table 3: Computation cost of methods.

Method	Params	Memory (GB)
MT	155M	4.58G
URPC	116M	3.62G
CCT	133M	4.11G
CPS	272M	15.71G
X-Net	326M	15.27G
Semi-MoE	181M	12.36G

Table 4: Ablation study for MGP and $\mathcal{L}_{AM}$ with 10% labeled data.

Data	Metric	w/o MGP	w/o $\mathcal{L}_{AM}$
Glas	DSC	-3.32	-1.24
	JC	-5.11	-2.0
CRAG	DSC	-0.22	-1.91
	JC	-0.34	-3.06

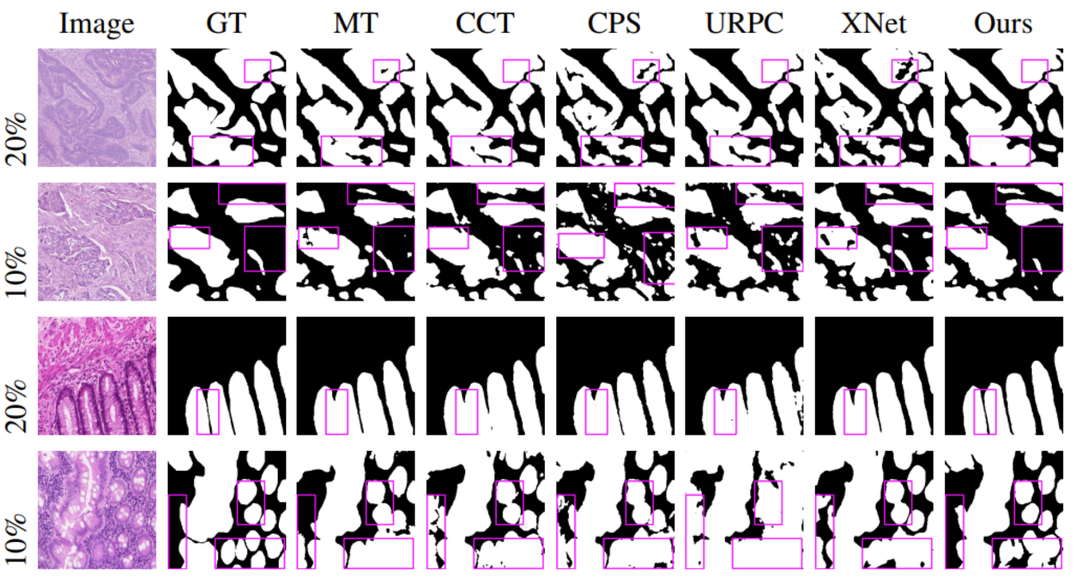


Figure 3: Qualitative results on CRAG (top two rows) and GlaS (bottom two rows).