MATCH: Multi-faceted Adaptive Topo-Consistency for Semi-Supervised Histopathology Segmentation

Meilong Xu[1,†] **Xiaoling Hu**[2,†] **Shahira Abousamra**[3] **Chen Li**[1] **Chao Chen**[1]

[1]Stony Brook University, NY, USA
[2] Massachusetts General Hospital and Harvard Medical School, MA, USA
[3]Department of Biomedical Data Science, Stanford University, CA, USA

Abstract

In semi-supervised segmentation, capturing meaningful semantic structures from unlabeled data is essential. This is particularly challenging in histopathology image analysis, where objects are densely distributed. To address this issue, we propose a semi-supervised segmentation framework designed to robustly identify and preserve relevant topological features. Our method leverages multiple perturbed predictions obtained through stochastic dropouts and temporal training snapshots, enforcing topological consistency across these varied outputs. This consistency mechanism helps distinguish biologically meaningful structures from transient and noisy artifacts. A key challenge in this process is to accurately match the corresponding topological features across the predictions in the absence of ground truth. To overcome this, we introduce a novel matching strategy that integrates spatial overlap with global structural alignment, minimizing discrepancies among predictions. Extensive experiments demonstrate that our approach effectively reduces topological errors, resulting in more robust and accurate segmentations essential for reliable downstream analysis. Code is available at https://github.com/Melon-Xu/MATCH.

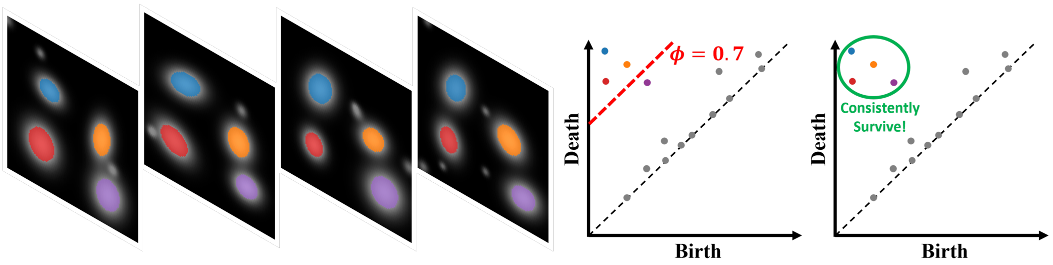
1 Introduction

Accurate segmentation of glands and nuclei in histopathology images is critical for digital pathology, significantly influencing diagnosis, prognosis, and treatment planning by enabling precise quantification of morphological and structural tissue features [8, 43, 27]. Numerous fully-supervised segmentation methods [49, 85, 11, 12, 25, 17, 41, 18] have demonstrated substantial success. However, densely distributed cellular structures in histopathology images often induce topological errors, including false merges or splits, severely impacting clinical reliability. Additionally, fully supervised methods demand extensive annotated datasets, which are costly, time-consuming, and not scalable [53, 32]. This limitation motivates exploring semi-supervised learning (SSL) strategies capable of leveraging abundant unlabeled data alongside limited annotations.

Recent SSL approaches have significantly enhanced segmentation accuracy in contexts of limited supervision [79, 55, 37–39, 64, 78, 81, 77, 82, 1, 83, 76, 30, 84, 73, 74, 45]. Nevertheless, these methods typically do not explicitly target topological errors, resulting in seemingly small segmentation errors with consequential significant topological inaccuracies that affect segmentation robustness. To explicitly address topological errors, persistent homology [7] offers a rigorous mathematical framework that captures and characterizes topological features, such as connected components and loops in data across multiple scales. The output, persistence diagram, summarizes these structures as dots in a 2D diagram. For each dot, the coordinate difference $(y - x)$ captures the *persistence* of the

[†] Equal contribution. Email: Meilong Xu (meixu@cs.stonybrook.edu).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) lh_1 to lh_4 (b) TopoSemiSeg (c) Ours

Figure 1: Intuition of the proposed framework. (a) Colored likelihood maps are coming from the MC dropout. Connected components consistently matched in at least three predictions retain identical colors across instances, indicating topological stability; components shown in grey fail to reach this consensus and are therefore treated as topologically transient. (b) Limitation of TopoSemiSeg [69], which relies on a fixed persistence threshold ($\phi = 0.7$, red dashed line) and therefore overlooks less-persistent yet meaningful structures (e.g. the violet point). (c) Our method adaptively identifies relevant topological structures without the need for human-selected thresholds.

topological structure across scales. Building upon this mathematical foundation, TopoSemiSeg [69] introduces topology-aware constraints into SSL frameworks, utilizing persistent homology to enforce topological consistency between teacher and student model predictions. Despite its effectiveness, TopoSemiSeg mainly relies on a predefined, hand-picked persistence threshold to identify meaningful topological structures. Such fixed thresholds are not data-driven, potentially biased, and can exclude relevant structures or retain irrelevant ones, as shown in Figure 1.

To address this issue, we investigate how to identify reliable topological structures from model predictions in a robust and adaptive manner, and enforce model consistency over these structures. We first revisit the fundamental principles of semi-supervised learning – robustness against perturbations. For an image without a training label, to identify reliable information, semi-supervised approaches typically add perturbations at the input level (i.e., augmentation) and at the model level (i.e., Monte Carlo Dropout). Pixel-level predictions that persist across these perturbations are considered reliable and used to self-supervise the model.

Our main idea is to tightly couple this SSL robustness-against-perturbation principle with topological reasoning. Moving beyond pixel-level, we identify topological structures that persist across different perturbations. These structures are considered reliable and used to self-supervise the model. This idea avoids a hand-picked threshold to determine reliable topological structures, and adaptively identifies truly relevant structures to enhance the model's topological reasoning power in an SSL setting.

Building on this idea, we propose a novel SSL segmentation framework employing **dual-level topological consistency**. Our method identifies significant topological features by examining predictions generated with different model perturbations. We formulate the structure correspondence task as a contrastive learning problem, distinguishing stable features, i.e., those consistently detected across multiple predictions, from transient or noisy structures. To identify the stable topological structures, we introduce an advanced matching algorithm that integrates spatial overlap, topological persistence, and spatial proximity criteria to associate topological structures across diverse predictions reliably.

As for perturbations, we propose to employ Monte Carlo (MC) dropout perturbations [9]. Meanwhile, we stress the importance of a temporal view of SSL. Previous works, such as [33, 36, 35, 51], demonstrate that evaluating the predictions in different training snapshots can reveal informative signals for robust prediction. Inspired by this, we propose to also compare topological structures across model snapshots at different training epochs. By explicitly optimizing for dual-level topological consistency, our framework enhances structural coherence within the student model without relying on extensive pixel-wise annotations. Our key contributions can be summarized as follows:

- We propose a novel integration of topological reasoning into the semi-supervised segmentation framework to robustly identify and preserve meaningful topological structures.
- We introduce dual-level topological consistency, measuring structural stability from intra-perturbed predictions (MC dropout) and temporal training snapshots, to effectively utilize unlabeled data.

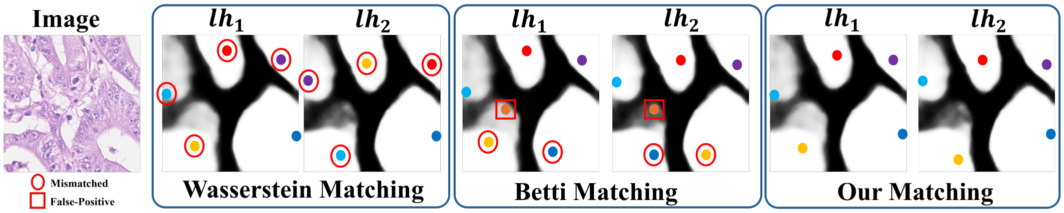


Figure 2: Comparison of our matching with Betti Matching [56] and Wasserstein Matching [21]. We match two likelihood maps obtained from the same input histopathology patch. The birth critical points of the matched pairs are highlighted in the same color. Note that Wasserstein Matching gets most matches wrong, and Betti Matching also gets two matches wrong while pairing biologically unrelated features when lacking the guidance of the ground truth.

- We develop a novel matching algorithm that integrates spatial overlap, topological persistence, and spatial proximity to accurately match topological structures across predictions.

Through extensive experiments on three widely used histopathology image datasets, our method significantly improves the topological accuracy while achieving comparable pixel-wise performance with limited annotations.

2 Related Works

Segmentation with Limited Supervision. Semi-supervised learning enhances medical image segmentation by effectively utilizing limited labeled data together with abundant unlabeled data. Consistency regularization approaches, such as the Mean Teacher model [57], ensure stable segmentation despite input variations [57, 37, 62, 46, 69]. Pseudo-labeling progressively improves accuracy by leveraging confident model predictions on unlabeled data [75, 50, 82]. Adversarial training aligns segmentation outputs from labeled and unlabeled datasets using discriminator networks [24, 34]. Additionally, uncertainty estimation methods such as MC dropout and Bayesian neural networks enhance reliability by effectively handling uncertainty during pseudo-label generation [9, 79, 44, 39, 70], while entropy minimization is used to reduce prediction uncertainty [13, 3, 66]. Contrastive learning strengthens segmentation robustness by training models to differentiate similarities and distinctions among data pairs, thereby boosting overall segmentation quality [78, 1, 77, 76].

Topology-Aware Image Segmentation. Topology-aware methods have been proposed to enforce correct topology, like connectivity or correct counts in segmentation tasks [21, 23, 52, 5, 71, 72, 20, 15, 61, 56, 60, 68, 80, 40]. These methods typically use differentiable loss functions derived from topological data analysis tools, including persistent homology [21, 5, 56], discrete Morse theory [23, 22, 16], topological interactions [15, 2], homotopy warping [20], centerline-based comparisons [52, 61]. These methods generally rely heavily on precisely annotated labels. Xu *et al.* [69] propose TopoSemiSeg to combine SSL with topological constraints. Classical persistent homology-based segmentation methods rely on Wasserstein matching [21, 69], which compares persistence diagrams based solely on feature lifespans. However, this approach may produce ambiguous or incorrect correspondences, as illustrated in Figure 2. To alleviate spatial inconsistencies, several methods were proposed [56, 63]. Betti Matching [56] embeds predictions and ground truth into a shared super-level filtration, ensuring alignment only among overlapping topological features. However, as shown Figure 2, it cannot ensure fully correct matching when the ground truth is missing and is too sensitive to preserve some transient structures. Our proposed MATCH-Pair could achieve almost completely accurate matching without the ground truth.

3 Methodology

The motivation of our proposed framework is to identify meaningful topological structures directly from perturbed predictions without the ground truth. The main challenge is to accurately match corresponding topological structures across multi-facet predictions that often contain substantial noise and variability. To overcome this challenge, we introduce MATCH-Pair, a pairwise matching algorithm, and MATCH-Global, an extended global matching algorithm, to robustly identify stable structures across multiple predictions. Building upon these matching algorithms, we propose dual-level topological consistency constraints: intra-topological consistency, enforcing consistency across multiple stochastic predictions, and temporal-topological consistency, ensuring stability across consecutive training snapshots. These consistency constraints directly optimize the student model, enabling it to learn robust segmentation representations from limited labeled data.

Our method overview is shown in Figure 3. The proposed MATCH framework leverages labeled data via supervised loss and unlabeled data through pixel-wise and dual-level topological consistency.

In this section, we will start by introducing the preliminaries of classic SSL. Next, we will use 3 subsections to introduce MATCH-Pair, MATCH-Global, and the dual-level topological consistency.

Preliminaries: SSL training. We address the semi-supervised image segmentation problem by leveraging a teacher-student framework, a widely adopted paradigm in semi-supervised learning [57]. Let $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$ denote the labeled dataset, where x_i^L represents the input image and $y_i^L \in \{0,1\}^{H \times W}$ is the corresponding pixel-wise annotation. Let $\mathcal{D}_U = \{x_j^U\}_{j=1}^{N_U}$ denote the unlabeled dataset. In our setting, the number of labeled samples is significantly smaller than the number of unlabeled samples, i.e., $N_L \ll N_U$. Our objective is to train a segmentation model f_θ, parameterized by θ, that accurately predicts segmentation masks using labeled and unlabeled data.

In this framework, the student model f_{θ_s} is trained using both supervised and unsupervised losses, while the teacher model f_{θ_t} provides stable targets for the student by being updated as an exponential moving average (EMA) of the student's parameters: $\theta_t^{(\tau+1)} = \alpha\theta_t^{(\tau)} + (1-\alpha)\theta_s^{(\tau+1)}$, where α controls the update rate. For the supervised loss on labeled data, we employ a combination of Dice loss and cross-entropy loss to capture both overlap and pixel-wise discrepancies, $\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{Dice}}(f_{\theta_s}(x^L), y^L) + \mathcal{L}_{\text{CE}}(f_{\theta_s}(x^L), y^L)$.

To leverage the unlabeled data, we enforce consistency between the student and teacher predictions. Specifically, the student receives a strongly augmented version of an unlabeled image x^U, while the teacher processes a weakly augmented version. The pixel-wise consistency loss is defined as the cross-entropy between the student and teacher outputs, $\mathcal{L}_{\text{cons}} = \mathcal{L}_{\text{CE}}(f_{\theta_s}(\mathcal{A}_s(x^U)), f_{\theta_t}(\mathcal{A}_w(x^U)))$, where $\mathcal{A}_s$ and $\mathcal{A}_w$ denote strong and weak augmentations, respectively.

3.1 MATCH-Pair: Spatial-Aware Pairwise Matching

Accurate identification of corresponding topological structures between the likelihood maps is crucial for robust histopathology image segmentation. We employ persistent homology with a **super-level set filtration** to extract 0-D topological features from likelihood maps, producing persistence diagrams that characterize each component by its persistence and critical points. To find correspondence between different persistence diagrams, traditional methods based on Wasserstein distance emphasize topological persistence without considering spatial relationships, often leading to incorrect associations between spatially distant yet similarly persistent features. In contrast, approaches based solely on spatial overlap tend to match transient structures of minimal significance incorrectly. To address these limitations, we propose MATCH-Pair, a Hungarian overlap-matching algorithm that integrates spatial overlap, topological persistence, and spatial proximity. The overall pipeline is depicted in Figure 4.

Given two likelihood maps $lh_1, lh_2 \in [0,1]^{H \times W}$, which are the softmax-activated outputs of the final UNet layer, we compute the persistence diagrams: $\mathrm{Dgm}(lh_k) = \{(b_i, d_i), \quad k \in \{1,2\}$ with the persistence $\mathrm{pers}_i = |d_i - b_i|$. Each persistence pair (b_i, d_i) yields a connected spatial region M_i, defined by flood-fill algorithm [54]. This algorithm generates a binary mask M_i starting from the birth pixel b_i. The region is expanded iteratively to neighboring pixels, provided that their likelihood exceeds the threshold $1 - d_i$.

To compute the relative significance of each structure, the persistence values are normalized to derive a weighting factor: $w_{k,i} = \frac{\mathrm{pers}_{k,i}}{\max\limits_j \mathrm{pers}_{k,j}}, \quad k \in \{1,2\}$. Here, i and j index the topological features from the 1_{st} and 2_{nd} persistence diagrams respectively, where $i \in 1, ..., n_1$ and $j \in 1, ..., n_2$ with n_1 and n_2 being the number of features in each diagram. k distinguishes between the two likelihood maps being compared. The notation $w_{k,i}$ refers to the normalized persistence weight of the i-th topological feature in the k-th likelihood map.

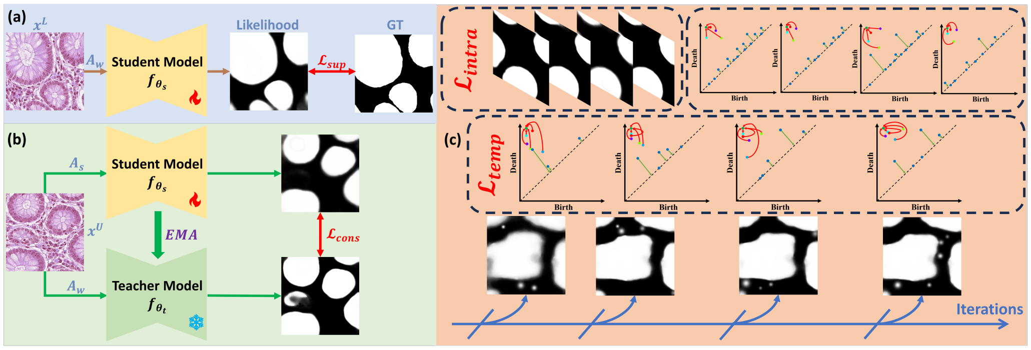


Figure 3: Overview of the proposed MATCH framework with dual-level topological consistency. Note that the $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{temp}}$ are used to directly optimize the parameters of the student model.

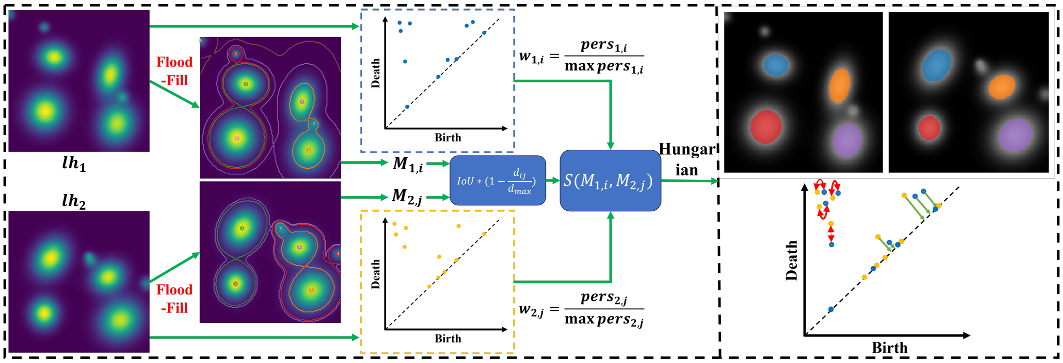

lh_1
lh_2
Flood -Fill
Flood -Fill
Death
Birth
$w_{1,i} = \frac{pers_{1,i}}{\max pers_{1,i}}$
$w_{2,j} = \frac{pers_{2,j}}{\max pers_{2,j}}$
$M_{1,i}$
$M_{2,j}$
$IoU * (1 - \frac{d_{ij}}{d_{max}})$
$S(M_{1,i}, M_{2,j})$
Hungarian
Death
Birth

Figure 4: Pipeline of the MATCH-Pair algorithm between two persistence diagrams.

To evaluate the similarity between spatial masks $M_{1,i}$ and $M_{2,j}$, a combined metric that integrates spatial overlap, normalized persistence weights, and spatial proximity is defined as:

$$S_{ij} = w_{1,i}\, w_{2,j} \frac{|M_{1,i} \cap M_{2,j}|}{|M_{1,i} \cup M_{2,j}|} \left(1 - \frac{d_{ij}}{d_{\max}}\right)$$

where d_{ij} is the Euclidean distance between birth critical points of the corresponding masks, and $d_{\max}$ denotes the maximum distance among all mask pairs. This similarity metric ensures the prioritization of spatially close, persistent, and well-overlapping structures.

A global one-to-one assignment between features from the two maps is obtained via the Hungarian algorithm [31], minimizing the cost (defined as the complement of similarity):

$$\min_{\pi_{ij}} \sum_{i,j} (1 - S_{ij})\, \pi_{ij}, \quad \pi_{ij} \in \{0, 1\}$$

Pairs achieving scores above a predefined threshold τ_{primary} constitute valid matches.

3.2 MATCH-Global: Multi-faceted Global Matching

While MATCH-Pair addresses an optimal correspondence between two persistence diagrams, many practical scenarios often involve multiple stochastic predictions (facets). Finding the corresponding topological structures among multiple facets is a challenge. Thus, we extend MATCH-Pair to MATCH-Global, a global multi-facet matching approach to link homologous 0-dimensional components consistently across all facets, assigning global indices to anatomical or topological structures.

Given a series of likelihood maps $\mathcal{L} = \{lh_t\}_{t=1}^{T}$, $lh_t \in [0,1]^{H \times W}$, each generates a persistence diagram: $\text{Dgm}_t = \{(b_{t,i}, d_{t,i})\}_{i=1}^{n_t}$. Each pair $(b_{t,i}, d_{t,i})$ corresponds to a spatial mask $M_{t,i}$, the normalized persistence weight $w_{t,i} = |d_{t,i} - b_{t,i}| / \max_{t',j} |d_{t',j} - b_{t',j}|$, and birth-critical point $c_{t,i}$.

Matching is performed sequentially across facets. For each adjacent pair of facets $(t, t+1)$ we form the weighted overlap matrix:

$$S_{ij}^{(t)} = \mathrm{w_{t,i}}\, \mathrm{w_{t+1,j}} \mathrm{IoU}\big(M_{t,i}, M_{t+1,j}\big) \Big(1 - \tfrac{\|c_{t,i} - c_{t+1,j}\|_2}{d_{\max}^{(t)}}\Big)$$

with $d_{\max}^{(t)} = \max_{i,j} \|c_{t,i} - c_{t+1,j}\|_2$ introduces a soft spatial penalty. Optimal assignments are solved via the proposed MATCH-Pair algorithm.

These matches form an undirected graph $G = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{(t,i) \mid 1 \le t \le T,\ 1 \le i \le n_t\}$, representing the structures and edges $\mathcal{E} = \bigcup_{t=1}^{T-1} \mathcal{E}^{(t)}$ indicating matches. Connected components $\{\mathcal{C}_k\}_{k=1}^{K}$ of G are identified by breadth-first search, providing globally consistent identities: $\mathcal{C}_k = \{(t,i) \mid \text{mask } M_{t,i} \text{ belongs to identity } k\}$.

Thus, the global multi-facet matching framework integrates pairwise correspondences into globally coherent tracks, robustly accommodating missing detections, splits, and merges, thereby ensuring topological consistency across multiple facets.

3.3 Dual-Level Topological Consistency

After identifying consistent topological structures across multiple facets, we propose dual-level topological consistency losses to enhance segmentation reliability and coherence. Specifically, we introduce two complementary loss terms: *intra-topological consistency*, which ensures consistency among stochastic predictions from MC dropout realizations [9], and *temporal-topological consistency*, which maintains consistency across consecutive training iterations.

In both scenarios, topological features are extracted from multiple prediction facets. We then apply our proposed MATCH-Global algorithm (see Section 3.2) to classify these topological structures into two distinct categories: **matched** ($\mathcal{C}_{\text{intra}}^{\text{match}}$, $\mathcal{C}_{\text{temp}}^{\text{match}}$), representing features consistently identified across multiple predictions, and **unmatched** ($\mathcal{C}_{\text{intra}}^{\text{unmatch}}$, $\mathcal{C}_{\text{temp}}^{\text{unmatch}}$), denoting features that are inconsistent or unstable across predictions. Specifically, matched structures are encouraged toward optimal probability distributions at their birth and death critical points, whereas unmatched structures, indicative of prediction uncertainty or instability, are driven toward shorter topological lifespans. Formally, we define the associated losses as:

$$\mathcal{L}_{\text{match}}(t,i) = \big(P_{b_{t,i}}^{(t)}\big)^2 + \big(1 - P_{d_{t,i}}^{(t)}\big)^2, \quad \mathcal{L}_{\text{diag}}(t,i) = \big(P_{b_{t,i}}^{(t)} - P_{d_{t,i}}^{(t)}\big)^2.$$

where $P_{b_{t,i}}^{(t)}$ and $P_{d_{t,i}}^{(t)}$ represent the predicted probability values at the birth ($b_{t,i}$) and death ($d_{t,i}$) critical points, respectively, of the i-th topological feature extracted from the t-th prediction.

The intra-topological consistency loss aggregates these penalties over multiple stochastic predictions through MC dropout within each iteration:

$$\mathcal{L}_{\text{intra}} = \frac{1}{B_{\text{intra}}} \sum_{b=1}^{B_{\text{intra}}} \left[\frac{1}{|\mathcal{C}_{\text{intra}}^{\text{match},(b)}|} \sum_{(t,i) \in \mathcal{C}_{\text{intra}}^{\text{match},(b)}} \mathcal{L}_{\text{match}}(t,i) + \frac{1}{|\mathcal{C}_{\text{intra}}^{\text{unmatch},(b)}|} \sum_{(t,i) \in \mathcal{C}_{\text{intra}}^{\text{unmatch},(b)}} \mathcal{L}_{\text{diag}}(t,i) \right]$$

where B_{intra} indicates the number of MC dropout predictions within each iteration. Similarly, the temporal-topological consistency enforces the constraints across predictions from consecutive training snapshots:

$$\mathcal{L}_{\text{temp}} = \frac{1}{B_{\text{temp}}} \sum_{b=1}^{B_{\text{temp}}} \left[\frac{1}{|\mathcal{C}_{\text{temp}}^{\text{match},(b)}|} \sum_{(t,i) \in \mathcal{C}_{\text{temp}}^{\text{match},(b)}} \mathcal{L}_{\text{match}}(t,i) + \frac{1}{|\mathcal{C}_{\text{temp}}^{\text{unmatch},(b)}|} \sum_{(t,i) \in \mathcal{C}_{\text{temp}}^{\text{unmatch},(b)}} \mathcal{L}_{\text{diag}}(t,i) \right]$$

where B_{temp} presents the number of temporal training snapshots. Finally, our dual-level topological consistency losses are integrated into the overall training objective alongside the supervised and pixel-wise consistency terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{intra}} \mathcal{L}_{\text{intra}} + \lambda_{\text{temp}} \mathcal{L}_{\text{temp}}$$

where hyperparameters λ_{cons}, λ_{intra}, and λ_{temp} balance their respective contributions, ensuring the model jointly meets pixel-level accuracy and robust topological coherence.

Table 1: Quantitative results on three histopathology image datasets. We compare our method with several state-of-the-art semi-supervised medical image segmentation methods on two settings of 10% and 20% labeled data. The statistically significant best results are highlighted in **bold**, while the second-best are marked with underline.

Dataset	Label Ratio (%)	Method	Pixel-wise	Topology-wise		
			Dice_Obj ↑	BE ↓	BME ↓	DIU ↓
CRAG	10	MT [57]	0.821 ± 0.006	2.238 ± 0.153	62.250 ± 3.127	74.630 ± 2.967
		EM [59]	0.789 ± 0.007	2.178 ± 0.147	80.100 ± 3.809	78.210 ± 3.298
		UA-MT [79]	0.837 ± 0.005	1.703 ± 0.112	66.450 ± 3.218	65.420 ± 2.847
		URPC [79]	0.829 ± 0.005	1.732 ± 0.118	74.600 ± 3.407	68.300 ± 3.004
		XNet [83]	0.872 ± 0.004	0.578 ± 0.053	15.050 ± 1.118	55.880 ± 2.516
		PMT [10]	0.876 ± 0.004	0.520 ± 0.051	14.200 ± 1.013	57.100 ± 2.638
		TopoSemiSeg [69]	0.884 ± 0.002	0.227 ± 0.014	10.475 ± 0.458	49.690 ± 1.947
		Ours	**0.888** ± **0.002**	**0.197** ± **0.012**	**9.175** ± **0.580**	**45.950** ± **1.880**
	20	MT [57]	0.858 ± 0.008	2.603 ± 0.161	99.025 ± 3.912	95.215 ± 3.487
		EM [59]	0.869 ± 0.006	1.933 ± 0.136	75.225 ± 3.772	63.823 ± 3.139
		UA-MT [79]	0.859 ± 0.006	1.822 ± 0.129	70.850 ± 3.586	61.138 ± 2.918
		URPC [39]	0.849 ± 0.007	2.489 ± 0.152	99.500 ± 4.085	87.681 ± 3.276
		XNet [83]	0.883 ± 0.005	0.422 ± 0.055	10.900 ± 1.127	50.537 ± 2.547
		PMT [10]	0.889 ± 0.004	0.460 ± 0.062	11.800 ± 1.203	48.300 ± 2.321
		TopoSemiSeg [69]	0.898 ± 0.004	0.226 ± 0.019	8.575 ± 0.736	43.712 ± 1.842
		Ours	**0.909** ± **0.005**	**0.188** ± **0.018**	**7.425** ± **0.570**	**40.250** ± **1.720**
	100 (Full)	Fully-Supervised	0.928 ± 0.002	0.149 ± 0.015	5.650 ± 0.223	29.425 ± 1.782
GlaS	10	MT [57]	0.790 ± 0.005	2.392 ± 0.162	31.125 ± 3.274	76.130 ± 2.965
		EM [59]	0.819 ± 0.006	1.431 ± 0.143	19.188 ± 3.846	61.245 ± 3.302
		UA-MT [79]	0.845 ± 0.004	2.086 ± 0.117	26.650 ± 3.245	68.025 ± 2.873
		URPC [79]	0.849 ± 0.004	1.155 ± 0.123	19.588 ± 3.408	54.832 ± 3.017
		XNet [83]	0.874 ± 0.003	0.843 ± 0.051	14.238 ± 1.154	40.912 ± 2.422
		PMT [10]	0.872 ± 0.004	0.798 ± 0.052	13.920 ± 1.097	39.850 ± 2.487
		TopoSemiSeg [69]	0.878 ± 0.003	0.551 ± 0.014	8.300 ± 0.478	35.845 ± 1.965
		Ours	**0.884** ± **0.003**	**0.501** ± **0.023**	**7.850** ± **0.391**	**30.525** ± **1.641**
	20	MT [57]	0.863 ± 0.005	2.126 ± 0.171	29.963 ± 3.987	64.275 ± 3.496
		EM [59]	0.865 ± 0.006	1.255 ± 0.138	17.275 ± 3.783	58.673 ± 3.255
		UA-MT [79]	0.866 ± 0.005	1.123 ± 0.132	18.038 ± 3.599	53.014 ± 3.069
		URPC [39]	0.878 ± 0.004	0.759 ± 0.067	14.350 ± 1.212	42.587 ± 2.601
		XNet [83]	0.884 ± 0.004	0.735 ± 0.065	10.188 ± 1.154	35.298 ± 2.328
		PMT [10]	0.887 ± 0.003	0.698 ± 0.062	9.980 ± 1.118	34.805 ± 2.271
		TopoSemiSeg [69]	**0.895** ± **0.003**	0.510 ± 0.053	9.825 ± 0.813	30.462 ± 1.978
		Ours	**0.894** ± **0.004**	**0.392** ± **0.056**	**7.925** ± **0.725**	**26.175** ± **1.633**
	100 (Full)	Fully-Supervised	0.917 ± 0.006	0.273 ± 0.026	6.875 ± 0.276	19.620 ± 0.712
MoNuSeg	10	MT [57]	0.748 ± 0.006	10.210 ± 0.486	292.857 ± 6.542	1526.079 ± 35.842
		EM [59]	0.757 ± 0.006	10.339 ± 0.503	257.071 ± 5.445	1319.815 ± 31.784
		UA-MT [79]	0.741 ± 0.007	10.227 ± 0.497	255.428 ± 5.983	1316.272 ± 30.216
		URPC [79]	0.774 ± 0.004	6.829 ± 0.319	214.428 ± 5.327	1098.372 ± 24.392
		XNet [83]	0.762 ± 0.005	7.152 ± 0.338	220.405 ± 4.611	1122.799 ± 25.116
		PMT [10]	0.764 ± 0.004	7.515 ± 0.352	227.650 ± 4.805	1210.400 ± 26.954
		TopoSemiSeg [69]	**0.783** ± **0.003**	6.661 ± 0.376	196.357 ± 3.067	1068.401 ± 17.500
		Ours	**0.785** ± **0.003**	**5.594** ± **0.361**	**192.863** ± **1.137**	**1011.857** ± **12.648**
	20	MT [57]	0.767 ± 0.005	12.522 ± 0.547	246.786 ± 8.018	1350.751 ± 32.407
		EM [59]	0.777 ± 0.006	7.160 ± 0.335	198.571 ± 6.731	1142.661 ± 27.581
		UA-MT [79]	0.772 ± 0.007	9.406 ± 0.444	246.857 ± 7.944	1336.684 ± 31.268
		URPC [39]	0.779 ± 0.004	5.325 ± 0.254	193.429 ± 6.105	1025.431 ± 23.799
		XNet [83]	0.776 ± 0.003	6.750 ± 0.316	198.525 ± 5.421	1117.406 ± 26.014
		PMT [10]	0.778 ± 0.006	6.500 ± 0.308	195.125 ± 6.289	1080.476 ± 25.145
		TopoSemiSeg [69]	**0.793** ± **0.004**	5.150 ± 0.145	188.642 ± 3.215	1105.946 ± 18.486
		Ours	**0.790** ± **0.006**	**4.930** ± **0.156**	**179.225** ± **2.383**	**982.286** ± **14.953**
	100 (Full)	Fully-Supervised	0.817 ± 0.010	2.491 ± 0.460	142.429 ± 4.674	729.017 ± 17.662

4 Experiments

We conduct comprehensive evaluations on three publicly available histopathology image datasets on both pixel-wise and topology-wise metrics. We benchmark our method against classic and recent state-of-the-art semi-supervised segmentation methods, including MT [57], EM [59], UA-MT [79], URPC [39], XNet [83], PMT [10], and TopoSemiSeg [69].

Implementation Details. The implementation details will be provided in the Supplementary.

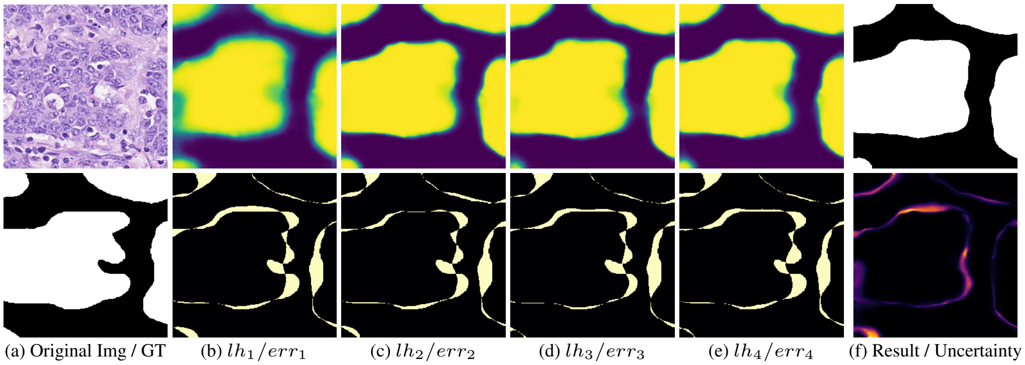


Figure 5: Qualitative illustration of MC dropout predictions (after the model convergence). **Top row:** original patch, the four likelihood maps, and the final segmentation. **Bottom row:** ground-truth mask, corresponding error maps, and the pixel-wise variance (uncertainty) map.

Datasets. We evaluate our proposed method on **Colorectal Adenocarcinoma Gland (CRAG)** [11], **Gland Segmentation in Colon Histology Images Challenge (GlaS)** [53], and **Multi-Organ Nuclei Segmentation (MoNuSeg)** [32]. More details are provided in the Supplementary.

Evaluation Metrics. To better evaluate our proposed method, we use pixel-wise metrics including **Object-level Dice Score (Dice_obj)** [67]; topology-wise metrics including **Betti Error** [21], **Betti Matching Error** [56], and **Discrepancy between Intersection and Union (DIU)** [40]. More details are provided in the Supplementary.

4.1 Results

Uncertainty Throughout the Topological Consistency. As illustrated in Figure 5, our proposed MATCH not only produces a robust segmentation result (top, (f)) but also furnishes an informative pixel-wise uncertainty map without any uncertainty-specific training objective or doing post hoc calibration. Visually, the variance map (bottom, (f)) concentrates along the gland boundaries where the four likelihood maps disagree, and these regions coincide almost perfectly with the binary error maps (bottom, (b) - (e)). Quantitatively, the Pearson correlation coefficients (PCC) [48] between the uncertainty and the error maps are 0.768, 0.728, 0.757, and 0.753 for the four facets, respectively. This confirms that the uncertainty is tightly coupled with prediction errors. Hence, reliable uncertainty estimation and the attendant suppression of spurious structures emerge naturally as a by-product of the proposed consistency mechanism, with no additional supervision or model modification required.

Quantitative Results. As shown in Table 1, across the three histopathology image datasets, our proposed method consistently achieves superior performance compared to state-of-the-art semi-supervised segmentation methods, under both 10% and 20% labeled data settings. Specifically, our method yields higher topology-wise accuracy with comparable pixel-wise performance. These results collectively illustrate that our framework effectively leverages limited annotations to achieve robust segmentation accuracy and enhanced topological fidelity.

Qualitative Results. We provide the qualitative results in Figure 6. The qualitative comparison highlights that our proposed method consistently outperforms other semi-supervised methods in preserving accurate glandular structures and topology across various histopathology samples. The comparative methods exhibit notable topological errors, including fragmentation, merging, and boundary leakage, as indicated by the red boxes. In contrast, our method effectively mitigates these errors, demonstrating superior robustness in maintaining topological integrity and accurate boundary delineation, thereby underscoring its effectiveness for precise medical image analysis tasks.

4.2 Ablation Study

To comprehensively explore the robustness and efficacy of our proposed strategy, hyperparameter-selection, and experimental settings, we conduct the ablation experiments on the CRAG dataset using 20% labeled data.

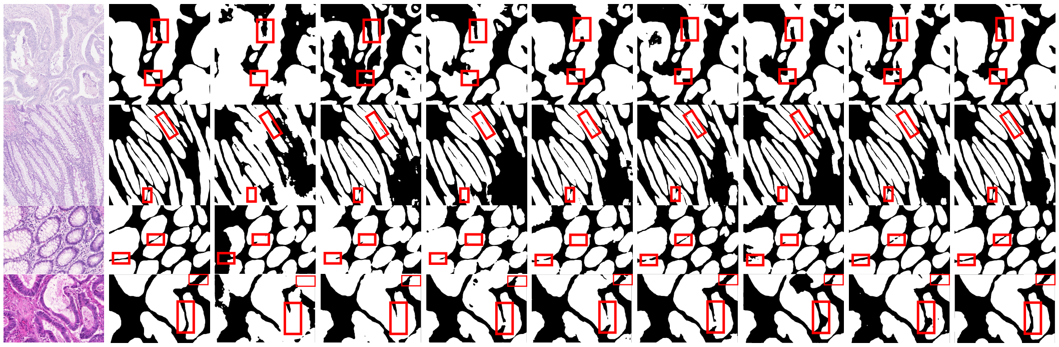


Figure 6: Qualitative results for semi-supervised methods on 10% and 20% labeled data. Rows 1-2 correspond to CRAG dataset, rows 3-4 correspond to the GlaS dataset. From left to right: (a) raw image, (b) ground-truth mask, (c) to (i) present the 7 baselines. (j) indicates the results of our method. The regions prone to topological errors are highlighted in red boxes.

Table 2: Ablation study on matching algorithms.

Matching	Pixel-wise	Topology-wise		
	Dice_obj ↑	BE ↓	BME ↓	DIU ↓
Wasser. [21]	0.864 ± 0.007	0.423 ± 0.026	9.647 ± 0.846	58.592 ± 2.574
Betti [56]	0.889 ± 0.005	0.237 ± 0.021	8.216 ± 0.717	44.157 ± 2.146
Ours	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**

Table 3: Effect of IoU & spatial-proximity (SP).

IoU	SP	Pixel-wise	Topology-wise		
		Dice_obj ↑	BE ↓	BME ↓	DIU ↓
✓	✗	0.890 ± 0.005	0.233 ± 0.020	8.300 ± 0.650	43.750 ± 2.100
✗	✓	0.882 ± 0.006	0.247 ± 0.022	9.600 ± 0.680	46.200 ± 2.250
✓	✓	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**

Ablation Study on Matching Algorithm. To validate the effectiveness of our proposed matching algorithm, we compare it against established alternatives, including Wasserstein Matching [21] and Betti-Matching [56]. As shown in Table 2, our algorithm consistently achieves superior performance in both pixel- and topology-wise metrics. Specifically, Wasserstein Matching, relying exclusively on persistence values without spatial information, exhibits the worst results. Although Betti-Matching incorporates spatial context, it still performs suboptimally compared to our method.

Ablation Study on IoU and Spatial Proximity (SP). To validate the effectiveness of the individual items of our matching cost, we conduct an ablation study on IoU and spatial proximity. The results in Table 3 quantitatively substantiate the complementary roles of the IoU and the spatial proximity factor in our Hungarian assignment cost. Removing either the proximity or the overlap item could degrade the performance. The overlap itself cannot fully distinguish spatially adjacent structures. These results demonstrate that both items are necessary to achieve topologically accurate matching.

Sensitivity Analysis on B_{intra} and B_{temp}. We further analyzed the sensitivity of our method to the number of MC dropout samples B_{intra} and temporal training snapshots B_{temp}. Table 4 shows that employing too few facets yields unreliable estimation of topological consistency, resulting in suboptimal segmentation performance. Conversely, increasing the number of facets beyond an optimal point introduces redundant information and additional variability, degrading model performance. Therefore, 4 is the optimal number that strikes a practical balance, ensuring the best performance while remaining computationally efficient.

Table 4: Influence of B_{intra} and B_{temp}.

B_{intra}	B_{temp}	Pixel-wise	Topology-wise		
		Dice_obj ↑	BE ↓	BME ↓	DIU ↓
2	2	0.878 ± 0.010	0.255 ± 0.025	9.350 ± 0.620	48.600 ± 2.300
3	3	0.892 ± 0.007	0.214 ± 0.020	8.105 ± 0.600	44.105 ± 2.050
5	5	0.872 ± 0.011	0.275 ± 0.023	10.050 ± 0.630	54.250 ± 2.253
4	4	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**

Table 5: Efficacy of $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{temp}}$.

$\mathcal{L}_{\text{intra}}$	$\mathcal{L}_{\text{temp}}$	Pixel-wise	Topology-wise		
		Dice_obj ↑	BE ↓	BME ↓	DIU ↓
✗	✗	0.862 ± 0.011	0.460 ± 0.022	11.680 ± 0.610	59.930 ± 2.150
✓	✗	0.898 ± 0.006	0.215 ± 0.020	7.920 ± 0.590	44.750 ± 1.970
✗	✓	0.882 ± 0.008	0.238 ± 0.031	8.540 ± 0.450	45.310 ± 2.040
✓	✓	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**

Ablation Study on Loss Components. To evaluate the contributions of individual loss terms in our dual-level topological consistency framework, we conduct experiments selectively enabling or disabling the $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{temp}}$. As presented in Table 5, each loss individually improves the pixel- and

Table 6: The results on the Roads dataset.

Labeled Ratio	Method	BE ↓	BME ↓	DIU ↓
10%	TopoSemiSeg	8.324 ± 0.729	9.681 ± 0.647	10.952 ± 0.671
	Ours	**7.892 ± 0.634**	**8.147 ± 0.521**	**9.376 ± 0.583**
20%	TopoSemiSeg	7.467 ± 0.582	8.213 ± 0.514	9.387 ± 0.538
	Ours	**6.983 ± 0.507**	**7.024 ± 0.436**	**8.149 ± 0.492**

Table 7: Density-aware quantitative results.

Setting	Dice_obj ↑	BE ↓	BME ↓
Sparse (Ours, $\leq$ 30 cells)	0.804 ± 0.004	4.620 ± 0.140	163.132 ± 2.136
Crowded ([69], $\geq$ 100 cells)	0.756 ± 0.009	6.890 ± 0.240	198.525 ± 3.125
Crowded (Ours, $\geq$ 100 cells)	0.774 ± 0.007	5.610 ± 0.198	186.313 ± 2.715
Ours (whole test image)	0.790 ± 0.006	4.930 ± 0.156	179.225 ± 2.383

topology-wise performance compared to the baseline without these constraints. Combining both losses achieves the strongest overall performance, confirming that $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{temp}}$ complement each other by addressing different sources of topological inaccuracies—stochastic noise within single facets and structural inconsistencies across training iterations.

Ablation Study on 1-D Topological Features. We mainly focus on 0-D topological features due to the following factors: For the primary application in our study (gland and nuclei segmentation), the most critical topological errors involve incorrect splitting or merging of individual structures, which are well-captured by 0-D persistent homology. For the validation on 1-dimensional structures, we conducted additional experiments on the Roads dataset [42]. The results are shown in Table 6. It verifies that our method could learn good topological representations from unlabeled data on 1-dimensional topological features.

Crowding-Aware Ablation Study. To quantify the influence of nuclei density on model performance, we randomly cut the test images into patches of size 256×256. For every patch, we count nuclei in the ground-truth instance map. Patches with <= 30 nuclei are labeled Sparse; those with >=100 nuclei are labeled Crowded. We sampled 14 samples for a fair comparison and show the results below in Table 7. The experiments above verify that our approach is density-aware. It achieves state-of-the-art performance on typical tissue, excels in sparse fields, and maintains a clear advantage over the strongest baseline under extreme nuclear crowding.

Ablation Study on the Alternatives of MC-dropout. We choose two alternative perturbation methods: Variational Inference (VI) [26], which generates multiple predictions by sampling from the learned variational posterior distribution, and Temperature Scaling [14], which produces diverse predictions through multiple sampling from temperature-modulated probability distributions. The experiments are conducted on CRAG 20% labeled data, and the results are shown in Table 8.

Table 8: Ablation of perturbation methods.

Method	Dice_obj ↑	BE ↓	BME ↓
Variational Inference	0.895 ± 0.006	0.242 ± 0.022	9.125 ± 0.685
Temperature Scaling	0.891 ± 0.007	0.258 ± 0.025	9.850 ± 0.795
MC-Dropout	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**

Table 9: Comparison with foundation models.

Method	Dice_obj ↑	BE ↓	BME ↓
LoRA–SAM	0.882 ± 0.006	0.440 ± 0.042	27.300 ± 2.937
LoRA–MedSAM	0.898 ± 0.005	0.268 ± 0.025	11.275 ± 1.899
Ours	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**

Comparison with Self-Supervised Methods Finetuned on Limited Labeled Data. To comprehensively evaluate the effectiveness of our method, we compare our method against some foundation models, like SAM [29] and MedSAM [41]. We use LoRA [19] to finetune these two models using 20% labeled data on the CRAG dataset and report the performance in Table 9. The results show that even with powerful foundation models, like SAM or MedSAM, topological errors can still exist without explicit topological modeling.

5 Conclusion

We present a semi-supervised segmentation framework that preserves significant topological structures in histopathology with limited annotations. Dual-level topological consistency across Monte Carlo dropout predictions and temporal training snapshots separates stable biological patterns from noise. For alignment, MATCH-Pair achieves spatially accurate matching between noisy persistence diagrams by combining spatial overlap, persistence, and proximity, and MATCH-Global scales to multiple facets. Experiments show consistent gains in robustness and substantial reductions in topological errors, enabling more reliable downstream analyses in digital pathology.

Acknowledgement. This work was partially supported by grants NSF CCF-2144901, NIH R01NS143143, and NIH R01CA297843.

References

[1] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *CVPR*, 2023.

[2] Alexander H Berger, Laurin Lux, Nico Stucki, Vincent Bürgin, Suprosanna Shit, Anna Banaszak, Daniel Rueckert, Ulrich Bauer, and Johannes C Paetzold. Topologically faithful multi-class segmentation in medical images. In *MICCAI*, 2024.

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[5] James R Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew P King. A topological loss function for deep-learning based image segmentation using persistent homology. *TPAMI*, 2020.

[6] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have l p-stable persistence. *Foundations of Computational Mathematics*, 2010.

[7] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[8] Matthew Fleming, Sreelakshmi Ravula, Sergei F Tatishchev, and Hanlin L Wang. Colorectal carcinoma: Pathologic aspects. *Journal of gastrointestinal oncology*, 2012.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[10] Ning Gao, Sanping Zhou, Le Wang, and Nanning Zheng. Pmt: Progressive mean teacher via exploring temporal consistency for semi-supervised medical image segmentation. In *ECCV*, 2024.

[11] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *MedIA*, 2019.

[12] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *MedIA*, 2019.

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2004.

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[15] Saumya Gupta, Xiaoling Hu, James Kaan, Michael Jin, Mutshipay Mpoy, Katherine Chung, Gagandeep Singh, Mary Saltz, Tahsin Kurc, Joel Saltz, et al. Learning topological interactions for multi-class medical image segmentation. In *ECCV*, 2022.

[16] Saumya Gupta, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, and Chao Chen. Topology-aware uncertainty for image segmentation. In *NeurIPS*, 2023.

[17] Hongliang He, Jun Wang, Pengxu Wei, Fan Xu, Xiangyang Ji, Chang Liu, and Jie Chen. Toposeg: Topology-aware nuclear instance segmentation. In *ICCV*, 2023.

[18] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *MedIA*, 2024.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[20] Xiaoling Hu. Structure-aware image segmentation with homotopy warping. In *NeurIPS*, 2022.

[21] Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In *NeurIPS*, 2019.

[22] Xiaoling Hu, Dimitris Samaras, and Chao Chen. Learning probabilistic topological representations using discrete morse theory. In *ICLR*, 2023.

[23] Xiaoling Hu, Yusu Wang, Li Fuxin, Dimitris Samaras, and Chao Chen. Topology-aware segmentation using discrete morse theory. In *ICLR*, 2021.

[24] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

[25] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2021.

[26] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 1999.

[27] Takahiro Karasaki, David A Moore, Selvaraju Veeriah, Cristina Naceur-Lombardelli, Antonia Toncheva, Neil Magno, Sophia Ward, Maise Al Bakir, Thomas BK Watkins, Kristiana Grigoriadis, et al. Evolutionary characterization of lung adenocarcinoma morphology in tracerx. *Nature medicine*, 2023.

[28] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

[30] Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In *CVPR*, 2025.

[31] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955.

[32] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *TMI*, 2019.

[33] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[34] Tao Lei, Dong Zhang, Xiaogang Du, Xuan Wang, Yong Wan, and Asoke K Nandi. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. *TMI*, 2022.

[35] Chen Li, Xiaoling Hu, Shahira Abousamra, and Chao Chen. Calibrating uncertainty for semi-supervised crowd counting. In *ICCV*, 2023.

[36] Chen Li, Xiaoling Hu, and Chao Chen. Confidence estimation using unlabeled data. In *ICLR*, 2023.

[37] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semi-supervised medical image segmentation. *TNNLS*, 2020.

[38] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *AAAI*, 2021.

[39] Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *MedIA*, 2022.

[40] Laurin Lux, Alexander H Berger, Alexander Weers, Nico Stucki, Daniel Rueckert, Ulrich Bauer, and Johannes C Paetzold. Topograph: An efficient graph-based framework for strictly topology preserving image segmentation. In *ICLR*, 2025.

[41] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 2024.

[42] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[43] Rodolfo Montironi, Roberta Mazzuccheli, Marina Scarpelli, Antonio Lopez-Beltran, Giovanni Fellegara, and Ferran Algaba. Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies. *BJU international*, 2005.

[44] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *MedIA*, 2020.

[45] Thanh-Huy Nguyen, Nguyen Lan Vi Vu, Hoang-Thien Nguyen, Quang-Vinh Dinh, Xingjian Li, and Min Xu. Semi-supervised histopathology image segmentation with feature diversified collaborative learning. In *AAAI Bridge Program on AI for Medicine and Healthcare*, 2025.

[46] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020.

[47] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[48] Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 1895.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[50] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *AAAI*, 2022.

[51] Wooseok Shin, Hyun Joon Park, Jin Sob Kim, and Sung Won Han. Revisiting and maximizing temporal knowledge in semi-supervised semantic segmentation. *arXiv preprint arXiv:2405.20610*, 2024.

[52] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *CVPR*, 2021.

[53] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *MedIA*, 2017.

[54] Alvy Ray Smith. Tint fill. In *Proceedings of the 6th annual conference on Computer graphics and interactive techniques*, 1979.

[55] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

[56] Nico Stucki, Johannes C Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In *ICML*, 2023.

[57] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[58] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, et al. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *TMI*, 2021.

[59] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.

[60] Fan Wang, Huidong Liu, Dimitris Samaras, and Chao Chen. Topogan: A topology-aware generative adversarial network. In *ECCV*, 2020.

[61] Haotian Wang, Min Xian, and Aleksandar Vakanski. Ta-net: Topology-aware network for gland segmentation. In *WACV*, 2022.

[62] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *MedIA*, 2022.

[63] Bo Wen, Haochen Zhang, Dirk-Uwe G Bartsch, William R Freeman, Truong Q Nguyen, and Cheolhong An. Topology-preserving image segmentation with spatial-aware persistent feature matching. *arXiv preprint arXiv:2412.02076*, 2024.

[64] Huisi Wu, Zhaoze Wang, Youyi Song, Lin Yang, and Jing Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *CVPR*, 2022.

[65] Kesheng Wu, Ekow J. Otoo, and Kenji Suzuki. Optimizing two-pass connected-component labeling algorithms. *Pattern Analysis and Applications*, 2009.

[66] Junsong Xie, Qian Wu, and Renju Zhu. Entropy-guided contrastive learning for semi-supervised medical image segmentation. *IET Image Processing*, 2024.

[67] Yutong Xie, Hao Lu, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Deep segmentation-emendation model for gland instance segmentation. In *MICCAI*, 2019.

[68] Meilong Xu, Saumya Gupta, Xiaoling Hu, Chen Li, Shahira Abousamra, Dimitris Samaras, Prateek Prasanna, and Chao Chen. Topocellgen: Generating histopathology cell topology with a diffusion model. In *CVPR*, 2025.

[69] Meilong Xu, Xiaoling Hu, Saumya Gupta, Shahira Abousamra, and Chao Chen. Semi-supervised segmentation of histopathology images with noise-aware topological consistency. In *ECCV*, 2024.

[70] Zhe Xu, Yixin Wang, Donghuan Lu, Xiangde Luo, Jiangpeng Yan, Yefeng Zheng, and Raymond Kai-yu Tong. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *MedIA*, 2023.

[71] Jiaqi Yang, Xiaoling Hu, Chao Chen, and Chialing Tsai. 3d topology-preserving segmentation with compound multi-slice representation. In *ISBI*, 2021.

[72] Jiaqi Yang, Xiaoling Hu, Chao Chen, and Chialing Tsai. A topological-attention convlstm network and its application to em images. In *MICCAI*, 2021.

[73] Jiaqi Yang, Nitish Mehta, Gozde Demirci, Xiaoling Hu, Meera S Ramakrishnan, Mina Naguib, Chao Chen, and Chia-Ling Tsai. Anomaly-guided weakly supervised lesion segmentation on retinal oct images. *MedIA*, 2024.

[74] Jiaqi Yang, Nitish Mehta, Xiaoling Hu, Chao Chen, and Chia-Ling Tsai. A multimodal approach combining structural and cross-domain textual guidance for weakly supervised oct segmentation. *arXiv preprint arXiv:2411.12615*, 2024.

[75] Huifeng Yao, Xiaowei Hu, and Xiaomeng Li. Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In *AAAI*, 2022.

[76] Chenyu You, Weicheng Dai, Fenglin Liu, Yifei Min, Nicha C Dvornek, Xiaoxiao Li, David A Clifton, Lawrence Staib, and James S Duncan. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *TPAMI*, 2024.

[77] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David Clifton, S Kevin Zhou, Lawrence Staib, and James Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. In *NeurIPS*, 2023.

[78] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *TMI*, 2022.

[79] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, 2019.

[80] Guoqing Zhang, Caixia Dong, and Yang Li. Topology-preserving hard pixel mining for tubular structure segmentation. In *BMVC*, 2023.

[81] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *CVPR*, 2022.

[82] Zhenxi Zhang, Chunna Tian, Harrison X Bai, Zhicheng Jiao, and Xilan Tian. Discriminative error prediction network for semi-supervised colon gland segmentation. *MedIA*, 2022.

[83] Yanfeng Zhou, Jiaxing Huang, Chenlong Wang, Le Song, and Ge Yang. Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In *ICCV*, 2023.

[84] Yanfeng Zhou, Lingrui Li, Zichen Wang, Guole Liu, Ziwen Liu, and Ge Yang. Xnet v2: Fewer limitations, better results and greater universality. In *BIBM*, 2024.

[85] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018*, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims of our paper are well reflected in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation of this work is discussed in the Supplementary.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This work does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details are provided in the Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided the code in the abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Some ablation studies are provided in the main paper and the rest of training and test details is provided in the Supplementary.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: This paper reports error bars suitably and correctly defined.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the GPU usage, learning rate, batch size, etc., in the Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The discussion of both potential positive societal impacts and negative societal impacts of the work performed is provided in the Supplementary.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work doesn't pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper are properly credited and the license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This work does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

MATCH: Multi-faceted Adaptive Topo-Consistency for Semi-Supervised Histopathology Segmentation

—Supplementary Material—

6 Overview

In the supplementary, we begin with a brief introduction to the persistent homology in Section 7, followed by detailed introductions of the datasets in Section 8 and the evaluation metrics in Section 9. Then, we provide the implementation details in Section 10, followed by the references of our baselines in Section 11. We also provide additional ablation studies in Section 12 to further illustrate the efficacy and robustness of our method and hyper-parameter selections. The limitations are provided in Section 13, followed by an analysis on the broader impact in Section 14.

7 Brief Introduction to Persistent Homology

Persistent homology [6, 7], a fundamental concept in topological data analysis (TDA), offers a robust framework for capturing and quantifying the topological features of data across multiple scales. In the context of image segmentation, particularly when dealing with likelihood maps that represent the probability of each pixel belonging to a specific class, persistent homology provides a means to analyze the underlying topological structures inherent in these probabilistic representations.

Given a likelihood map $f : \Omega \to [0, 1]$, where $\Omega \subset \mathbb{R}^2$ represents the image domain, we construct a filtration of super-level sets:

$$\mathcal{F}_\alpha = \{x \in \Omega \mid f(x) \geq \alpha\}, \quad \alpha \in [0, 1].$$

As α decreases from 1 to 0, the super-level set $\mathcal{F}_\alpha$ transitions from empty regions to encompass the entire domain Ω, revealing the sequential emergence, merging, and disappearance of connected components and loops. Persistent homology tracks these topological changes across the filtration, recording the corresponding birth and death thresholds of each feature in a persistence diagram.

A persistence diagram is a multiset of points $\{(b_i, d_i)\}$ in the extended plane $\mathbb{R}^2$, where each point corresponds to a topological feature that appears (birth b_i) and disappears (death d_i) during the filtration process. Features that persist across a wide range of α values (i.e., with large $|d_i - b_i|$) are considered topologically significant, while those with short lifespans are often attributed to noise.

8 Dataset Details

Colorectal Adenocarcinoma Gland (CRAG) [11] consists of 213 hematoxylin and eosin (H&E)-stained colorectal adenocarcinoma image tiles acquired at 20× magnification, each with detailed annotations at the instance level. Most images are in approximately 1512×1516 pixels. Officially, the dataset is partitioned into 173 training samples and 40 testing samples. For our experiments, the training subset is further divided into 153 images for model training and 20 images for validation. For semi-supervised scenarios with 10% and 20% labeled data, we randomly select 16 and 31 labeled images, respectively, for training.

Gland Segmentation in Colon Histology Images Challenge (GlaS) [53] comprises 165 images sourced from 16 H&E-stained histological slides of colorectal adenocarcinoma at stages T3 or T4. The official split includes 85 training images and 80 testing images. In our experimental setup, the training set is divided into 68 images for model training and 17 for validation. We randomly select 7 and 14 labeled images to represent 10% and 20% of labeled training data scenarios, respectively.

Multi-Organ Nuclei Segmentation (MoNuSeg) [32] dataset contains 44 H&E-stained histology images of dimensions 1000×1000 pixels, encompassing nuclei annotations from seven distinct organs. Officially, it consists of 30 training images with a total of $21,623$ annotated nuclei and 14 images designated for testing. For our experiments, we reserve 20% (6 images) of the training set for validation. In experiments involving 10% and 20% labeled data splits, we randomly select 3 and 5 labeled images, respectively, for training.

9 Evaluation Metrics

We evaluate the segmentation quality from both pixel- and topology-wise. **Object-level Dice coefficient (Dice_Obj)** [67] is selected to measure pixel-wise performance, which measures instance-wise overlap between predicted and ground-truth masks and is thus well suited to the precise delineation of individual structures required in digital pathology.

To evaluate the topological accuracy, we select three topological evaluation metrics, **Betti Error** [21], **Betti Matching Error** [56], and **Discrepancy between Intersection and Union (DIU)** [40]. Betti Error (BE) mainly computes the mean absolute difference in 0-dimensional Betti numbers over 256×256 sliding-window patches. Betti Matching Error (BME), which extends BE by enforcing spatial correspondence when pairing topological features, thereby penalizing misplaced components even when counts are preserved. Introduced in [40], DIU quantifies how faithfully the topology of the common and combined foreground regions agrees.

10 Implementation Details

Our model is trained in two distinct stages. In the initial stage, we perform pretraining using only supervised loss and pixel-wise consistency loss. For all three datasets, the pretraining stage proceeds for $12,000$ iterations. The second stage involves fine-tuning the model by integrating our proposed dual-level topological consistency constraints, which last for an additional $1,000$ iterations. We use UNet [49] as the backbone for both the student and teacher models.

All training is implemented using PyTorch [47] and optimized using the Adam optimizer [28]. Training hyperparameters are set as follows: the batch size is 16 and the learning rate is 5×10^{-4}. Both labeled and unlabeled data undergo pre-processing through random cropping (with cropping size of 256×256), followed by data augmentation procedures including random rotation and flipping as weak augmentations, and color adjustments and morphological shifts for stronger augmentations.

In particular, we adopt a random cropping strategy for enforcing intra-topological consistency, while a fixed patch cropping strategy is used for temporal-topological consistency. **The inputs to the student model to estimate the intra- and temporal-topological consistency are all original patches, without any transformations.** The EMA decay rate α is set to 0.999. Within the supervised loss, the weights assigned to the cross-entropy loss and Dice loss are equally set to 0.5. The weight of the pixel-wise consistency loss is calculated by the Gaussian ramp-up function $\lambda_{\text{cons}} = k * e^{-5*(1-\frac{\tau}{T})^2}$, where $k = 0.1$ and T is the total number of iterations.

Additionally, λ_{intra} and λ_{temp} are both set to 0.001. This balanced configuration ensures effective integration of topological constraints while maintaining stable training dynamics. **Note that dual-level topological consistency is used to optimize the student model directly, and we use the student model to do the inference.** The experiments are conducted on an NVIDIA RTX A6000 GPU (48 GB), using a 24-core Intel® Xeon® Gold 6248R CPU @ 3.00 GHz and 192 GB RAM. The training time of one iteration is 1020.04 ms, and GPU memory consumption is 25.726 GB using UNet with batch size 16. The training time of TopoSemiSeg for one iteration is 610.80 ms, and the GPU memory consumption is 15.235 GB. For the non-PH baseline, like PMT [10], the training time per iteration is 582.34 ms,

11 Baseline Reference

We select 7 classical and recent state-of-the-art methods as comparatives. The implementations of some of them are based on publicly available repositories. Here, we provide the source of our baselines for reference and greatly appreciate their efforts in building the open-source community:

MT [57], EM [59], UA-MT [79] and URPC [39] are based on the implementations from: https://github.com/HiLab-git/SSL4MIS.

XNet [83] is based on the implementations from:

https://github.com/guspan-tanadi/XNetfromYanfeng-Zhou.

PMT [10] is based on the implementations from: http://github.com/Axi404/PMT.

TopoSemiSeg [69] is based on the implementations from:

https://github.com/Melon-Xu/TopoSemiSeg.

12 Additional Ablation Study

Here, we provide additional ablation studies to illustrate the efficacy and robustness of our selected backbone and hyperparameters.

Ablation Study on λ_{intra} and λ_{temp}. The results shown in Table 10 demonstrate the impact of varing weights of intra- and temporal-topological consistency (λ_{intra} and λ_{temp}). When two weights are both 0.001, the performance is the best across both pixel-wise and topology-wise metrics. As these weights increase from 0.001 to 0.01, there's a clear degradation in performance, indicating that excessively large consistency constraints may introduce unnecessary regularization, thus impairing the segmentation quality. When both weights are reduced to 0.0005, the dual-level consistency regularization becomes too weak to meaningfully optimize the student model, leading to diminished topological guidance and a corresponding drop in both pixel-wise and topology-wise performance.

Ablation Study on EMA Decay α. Table 11 investigates the influence of the EMA decay parameter α. $\alpha = 0.999$ yields the best performance. When decreasing α from 0.999 to 0.996, the results remain competitive but slightly deteriorate, highlighting that a higher EMA decay value effectively leverages historical model parameters for improved topological and segmentation robustness. In contrast, very high values (e.g. $\alpha = 0.9999$) excessively rely on historical information, marginally weakening the adaptability and performance of the model.

Table 10: Influence of λ_{intra} and λ_{temp}.

λ_{intra}	λ_{temp}	Pixel-wise	Topology-wise		
		Dice_obj ↑	BE ↓	BME ↓	DIU ↓
0.01	0.01	0.865 ± 0.012	0.275 ± 0.023	10.650 ± 0.630	53.500 ± 2.300
0.005	0.005	0.892 ± 0.007	0.214 ± 0.020	8.105 ± 0.600	44.105 ± 2.050
0.001	0.001	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**
0.0005	0.0005	0.895 ± 0.007	0.235 ± 0.020	8.950 ± 0.580	44.225 ± 1.850

Table 11: Impact of the EMA decay α.

α	Pixel-wise	Topology-wise		
	Dice_obj ↑	BE ↓	BME ↓	DIU ↓
0.9999	0.890 ± 0.007	0.230 ± 0.022	9.500 ± 0.630	46.000 ± 2.100
0.999	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**
0.996	0.902 ± 0.006	0.205 ± 0.020	8.250 ± 0.610	42.500 ± 1.900
0.99	0.882 ± 0.008	0.260 ± 0.025	11.000 ± 0.700	50.000 ± 2.300

Ablation Study on Different Backbones. To further verify the robustness of our proposed method, we conduct ablation experiments on different backbones. The results are shown in Table 12. Specifically, DeepLabV3+ [4] and UNet++ [85] show modest but clear improvements in both pixel-wise and topology-wise metrics. The UNet [49] backbone achieves the most substantial gains, particularly in topology-wise metrics. These results demonstrate that integrating our MATCH framework consistently improves performance across multiple backbones.

Table 12: Performance comparison of different backbones w or w/o our MATCH.

Backbone	Pixel-Wise	Topology-Wise		
	Dice_Obj ↑	BE ↓	BME ↓	VOI ↓
DeepLabV3+ [4]	0.889 ± 0.010	0.272 ± 0.023	11.782 ± 0.690	50.867 ± 2.221
DeepLabV3+ [4]+Ours	0.892 ± 0.008	0.245 ± 0.022	10.129 ± 0.638	47.412 ± 2.047
UNet++ [85]	0.886 ± 0.008	0.245 ± 0.023	9.210 ± 0.603	45.517 ± 2.041
UNet++ [85]+Ours	0.890 ± 0.006	0.238 ± 0.020	9.021 ± 0.580	45.073 ± 1.995
UNet [49]	0.894 ± 0.006	0.232 ± 0.019	8.872 ± 0.579	44.281 ± 1.881
UNet [49]+Ours	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**

Ablation Study on Applying Dual-Level Topo-Consistency between Teacher and Student Models. We conduct an ablation study to assess the impact of enforcing dual-level topological consistency in a teacher-student framework. Specifically, the dual-level topological consistency is estimated from the teacher model's multiple predictions, and consistency constraints are applied between the student output and the most recent prediction from the teacher. We compare this teacher-student configuration with a student-only model, both trained under identical consistency constraints. The results in Table 13 reveal that the student-only model consistently achieves superior performance. The relatively poorer performance of the teacher-student configuration suggests that leveraging the teacher's predictions, potentially noisy or outdated, introduces additional uncertainty and adversely affects the student's ability to effectively capture stable topological structures.

Table 13: Ablation study on applying dual-level topo-consistency between teacher and student models.

Mode	Pixel-wise	Topology-wise		
	Dice_Obj ↑	BE ↓	BME ↓	DIU ↓
Teacher–Student	0.885 ± 0.007	0.217 ± 0.021	8.102 ± 0.620	42.520 ± 1.880
Student Only	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**

Extension from Binary to Multi-Class Segmentation. To extend our method to the multi-class setting, we choose a multi-class nuclei segmentation dataset, MoNuSAC [58], to conduct experiments. This dataset contains four cell types: Epithelial, Lymphocyte, Macrophage, and Neutrophil. We conducted experiments using 20% labeled data and report the class-wise performance of TopoSemiSeg and our method in Table 14. As demonstrated in our class-specific results on MoNuSAC (Epithelial, Lymphocyte, Macrophage, and Neutrophil), our approach consistently outperforms TopoSemiSeg across all cell types, with particularly notable improvements in topological metrics (BE and BME) that are crucial for distinguishing overlapping structures.

Table 14: The multi-class segmentation results on the MoNuSAC dataset.

Class	Method	Dice_obj ↑	BE ↓	BME ↓
Epithelial	TopoSemiSeg	0.778 ± 0.009	5.342 ± 0.187	195.158 ± 4.627
	Ours	**0.781 ± 0.008**	**5.128 ± 0.189**	**186.847 ± 3.958**
Lymphocyte	TopoSemiSeg	0.751 ± 0.013	6.089 ± 0.223	218.394 ± 5.841
	Ours	**0.756 ± 0.012**	**5.794 ± 0.235**	**207.693 ± 4.672**
Macrophage	TopoSemiSeg	0.765 ± 0.011	5.687 ± 0.201	206.732 ± 4.985
	Ours	**0.769 ± 0.010**	**5.423 ± 0.208**	**195.381 ± 4.127**
Neutrophil	TopoSemiSeg	0.738 ± 0.016	6.521 ± 0.267	234.576 ± 6.123
	Ours	**0.742 ± 0.015**	**6.187 ± 0.281**	**221.459 ± 5.894**

Ablation Study on the Sensitivity of $\tau_{primary}$. We provide the ablation study on the sensitivity of $\tau_{primary}$ in Table 15. The results have shown that our method is robust to selecting $\tau_{primary}$. Moreover, the low threshold of 0.1 was chosen to be inclusive rather than restrictive: it allows more potential matches to be considered valid while letting the Hungarian algorithm determine optimal assignments based on our comprehensive similarity metric (combining spatial overlap, persistence weights, and proximity).

Table 15: Effect of τ_{primary}.

τ_{primary}	**Dice_obj ↑**	**BE ↓**	**BME ↓**	**DIU ↓**
0.05	0.906 ± 0.006	0.195 ± 0.019	7.850 ± 0.620	41.750 ± 1.850
0.1 (current)	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**	**40.250 ± 1.720**
0.2	0.908 ± 0.005	0.191 ± 0.020	7.680 ± 0.590	41.100 ± 1.780
0.3	0.905 ± 0.006	0.201 ± 0.021	8.150 ± 0.650	42.850 ± 1.920

Table 16: Effect of dropout rates.

Dropout Rate	**Dice_obj ↑**	**BE ↓**	**BME ↓**
10%	0.898 ± 0.006	0.210 ± 0.020	8.200 ± 0.650
20% (current)	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**
30%	0.910 ± 0.005	0.185 ± 0.017	7.350 ± 0.560
50%	0.890 ± 0.007	0.220 ± 0.022	8.800 ± 0.720

The Impact of Different Dropout Rates. We also add complementary ablation studies on the dropout rate of the MC-dropout. Other settings are kept unchanged. We conduct the ablation experiments on CRAG 20% labeled data and report the performance in Table 16. The results reveal an optimal dropout rate range of 20%-30% for our framework, where performance plateaus with minimal differences between these rates. Lower dropout rates provide insufficient perturbation diversity for reliable topological matching. In contrast, excessive dropout introduces detrimental noise that degrades both pixel- and topology-wise performance, confirming that moderate stochasticity is essential for effective topological consistency estimation.

Ablation Study on $\mathcal{L}_{\text{cons}}$. We conducted the ablation study on the $\mathcal{L}_{\text{cons}}$ and the results are shown in Table 17. Based on the results of the ablation study and the principles of semi-supervised learning, removing the pixel-wise consistency term in the training stages would result in significant performance degradation across all metrics.

Table 17: Ablation Study on $\mathcal{L}_{\text{cons}}$.

Method	Dice_obj ↑	BE ↓	BME ↓
w/o $\mathcal{L}_{\text{cons}}$	0.875 ± 0.008	0.285 ± 0.025	9.850 ± 0.680
Ours	**0.909 ± 0.005**	**0.188 ± 0.018**	**7.425 ± 0.570**

Downstream Analysis: Cell counting To further analyze the impact of our method on downstream analysis, we conducted a cell counting study on the same MoNuSeg test cohort. We used the connected component analysis [65] to identify the cells and calculate the total cell count, the predicted total cell count, and the absolute counting error (mean ± std). The results are shown in Table 18. Note that the Total GT cell count and the predicted cell count are reported for the entire test cohort, while the absolute count error is reported per image (with a total of 14 test images).

We observed that our method yields noticeably smaller counting errors than both baseline approaches (one topo method and one non-topo method). This confirms that although the pixel-wise segmentation performances are comparable, fixing the topological errors on a few pixels leads to more accurate biological readouts.

Table 18: Downstream Analysis on Cell Counting.

Method	Total GT Cell Count	Predicted Cell Count	Absolute Counting Error (Mean ± Std)	Dice_obj
PMT [8]	6024	8106	148.71 ± 99.41	0.778 ± 0.006
TopoSemiSeg [57]	6024	7877	132.36 ± 56.09	0.793 ± 0.004
MATCH	6024	**7511**	**106.21 ± 49.30**	0.790 ± 0.006

13 Limitations

A potential limitation of our MATCH framework arises from its reliance on stable feature extraction from persistence diagrams, which can be challenged when predictions exhibit extreme noise or minimal structural differences. In addition, the framework introduces nontrivial computational overhead: computing persistence diagrams and performing MATCH-Global/MATCH-Pair alignments across Monte Carlo dropout samples and temporal snapshots require multiple forward passes and matching steps, resulting in longer training times and increased memory usage.

14 Broader Impact

Our method significantly contributes to enhancing segmentation robustness by effectively leveraging unlabeled data, reducing reliance on extensive annotations, and ensuring topological accuracy crucial for clinical and biomedical analysis. This approach not only facilitates efficient utilization of limited labeled data but also provides insightful uncertainty estimates beneficial for downstream diagnostic applications.

A negative broader impact could include inadvertent propagation of segmentation inaccuracies if poorly matched topological structures influence model learning, potentially affecting reliability in critical medical decisions.