Beyond Stabilization: Dual-EMA Teachers for Global–Local Semantic Learning in Semi-Supervised Medical Image Segmentation

Kaibo Liao[1] **Yazhou Ren**[1,2,*] **Song Wu**[1]
Kecheng Chen[3] **Xiaorong Pu**[1,2] **Lixin Duan**[1,2]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, China; [2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, China
[3]City University of Hong Kong, Hong Kong SAR, China

Abstract

Semi-supervised medical image segmentation has gained increasing attention for its outstanding performance with limited annotations. Most methods follow the Mean Teacher (MT) framework, where Exponential Moving Average (EMA) is primarily used to stabilize predictions, while its potential roles beyond stabilization are largely overlooked. In this work, we interpret EMA as a first-order infinite impulse response (IIR) low-pass filter with inherent frequency selectivity, which enables the preservation of semantic information at different levels. Motivated by this insight, we propose a dual-EMA teachers framework, including a long-term teacher focusing on global structures and a short-term teacher emphasizing local details. Furthermore, to help the student better exploit the guidance from both teachers, a First-In-First-Out (FIFO) replay queue is designed to reuse historical high-confidence pseudo-labeled pairs, leveraging pseudo labels provided at different training stages and reducing the sensitivity to sampling bias. Meanwhile, we impose feature-level global semantic and local relation consistency constraints to alleviate the impact of noisy pseudo labels, facilitating robust transfer of complementary representations from dual teachers to the student model. Experiments on three public medical image datasets demonstrate that our method outperforms current SoTA methods, verifying its effectiveness. The code is available at `https://github.com/kaiboliao/DualEMA-MedSeg`.

1 Introduction

Medical image segmentation [1] plays a fundamental role in clinical diagnosis and treatment planning. Fully supervised deep learning methods [29] have achieved promising results. However, such methods heavily rely on large-scale pixel-level annotations, which are costly and time-consuming to obtain. Semi-supervised medical image segmentation (SSMIS) [2, 9, 20, 36, 38] has therefore emerged as a more practical solution, aiming to improve segmentation performance using only a small set of labeled data together with abundant unlabeled data.

Among existing SSMIS approaches, the Mean Teacher (MT) paradigm [27] and its variants have emerged as a dominant framework (Fig. 1(a)), where an EMA teacher provides pseudo labels to train the student by minimizing a consistency loss. Despite its widespread adoption, EMA is often treated as an empirical mechanism, using default coefficients without further analysis. This raises a

*Corresponding author (yazhou.ren@uestc.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

key question: ***beyond prediction stabilization, does EMA play additional roles during training?*** Simply put, can different EMA coefficients provide different types of information, and how do such differences shape the semantic representations learned by the teacher? This issue is particularly relevant in medical image segmentation, as medical images possess strong structural characteristics. Accurate segmentation requires capturing both stable global anatomical layouts and fine-grained local details simultaneously.

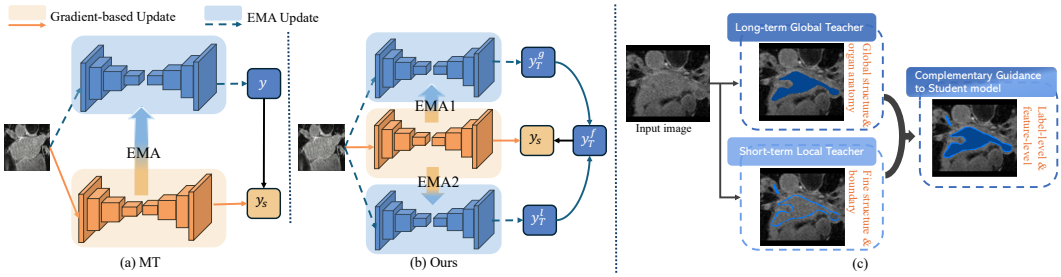

Figure 1: (a) Mean Teacher (MT) framework, (b) the proposed dual-EMA teachers framework, (c) illustration of semantic complementarity induced by dual-EMA teachers.

In this work, we show that EMA can be interpreted as a first-order Infinite Impulse Response (IIR) low-pass filter [23, 24], which is inherently frequency-selective, as illustrated in Fig. 2(a). Furthermore, we conduct experiments and find that this frequency selectivity suggests a natural connection to semantic preferences: larger EMA coefficients emphasize low-frequency, long-term trends corresponding to stable global structures, whereas smaller coefficients preserve short-term, high-frequency signals related to local details and boundaries.

Motivated by this observation, we propose a dual-EMA teachers framework that leverages complementary EMA settings for global–local semantic learning. As illustrated in Fig. 1(b)(c), our framework maintains a long-term teacher model dedicated to stable global anatomy and a short-term teacher sensitive to fine-grained local details. Their predictions are fused to provide complementary and more reliable supervision to the student model.

Although the dual-EMA mechanism is feasible and enables us to maintain two teacher models that capture different semantic cues, the key is how to enable the student to efficiently and robustly absorb their complementary guidance for improved segmentation capability. To this end, two challenges must be addressed. ***First, mini-batch pseudo supervision is susceptible to bias, leading to limited learning of long-term semantic knowledge.*** Existing consistency learning methods usually generate pseudo labels only for the current mini-batch, and immediately use them to supervise the student update. As a result, the student is exposed only to a small set of transient pseudo-supervised samples at each iteration. This makes the training more susceptible to batch sampling imbalances and random fluctuations, hindering stable and sufficient distillation of the long-term teacher's global semantic trends. ***Second, the impact of pseudo-label noise.*** Even with more comprehensive dual-EMA teachers for pseudo-label generation, erroneous pseudo labels are inevitable; if consistency is enforced only at the label level, noisy pseudo labels can be amplified and gradually propagated to the student model, thereby harming training stability and final performance.

To tackle the first challenge, we introduce a replay-queue-based pseudo-supervision reuse mechanism in addition to the base supervision. High-confidence unlabeled samples and their pseudo labels are stored in a First-In-First-Out (FIFO) replay queue and reused in subsequent iterations, allowing the student to efficiently learn from pseudo-supervision signals generated by the teachers at different training stages. Moreover, as pseudo labels become progressively more reliable during training, replay queue can more fully exploit these supervision signals, further improving learning efficiency and stability. ***To address the second challenge, we further impose feature-level global–local consistency constraints,*** including global semantic consistency and local relation consistency, which better leverage the complementary advantages offered by the dual-EMA teachers framework. These constraints then complement label-level supervision, mitigating the adverse impact of noisy pseudo labels and enhancing model robustness. In summary, the contributions of this work are as follows:

- We interpret EMA from signal processing perspective, revealing its frequency-selective behavior and motivating a dual-EMA teachers framework that simultaneously captures complementary global and local semantics.
- We design a FIFO replay queue to reuse high-confidence pseudo-labeled pairs across iterations, providing more diverse and stable pseudo supervision, and further impose feature-level global-local consistency constraints in addition to label-level supervision for more robust semantic distillation.
- Extensive experiments are conducted on three public datasets to verify the advantages of the proposed method in segmentation accuracy and generalization.

2 Related Work

Semi-supervised Medical Image Segmentation: Various SSMIS methods have been proposed, with the most commonly used being pseudo-labeling [3, 19, 26] and consistency regularization [13, 18, 39]. Pseudo-labeling methods supervise unlabeled data using generated labels, where ensuring label reliability is critical. For instance, Gao et al. [7] leveraged collaborative training and difference-driven alignment regularization for more reliable pseudo labels. Su et al. [26] filtered high-quality pseudo labels based on prediction confidence and intra-class similarity. Consistency regularization enforces the model to generate consistent predictions under different perturbations. Specifically, Li et al. [13] introduced a transformation-consistent self-ensembling framework and generalized data perturbations to enhance consistency regularization. Zhang et al. [39] combined intra-task and cross-task consistency to selectively leverage reliable unlabeled data.

EMA-Teacher Paradigms in SSMIS: Following the MT paradigm [27], many methods adopt EMA-updated frameworks due to stability and simplicity. UAMT [37] incorporated uncertainty-aware scheme into EMA-based consistency learning, and FUSSNet [33] combined EU-guided unsupervised learning with AU-guided supervised learning. Other works further explore multi-branch or dual-network designs. For example, Wang et al. [28] proposed a dual-decoder student for decoupling training guided by an EMA teacher. Li et al. [14] built on the MT framework by introducing an auxiliary student network and a deep-guided cross-modal mutual learning strategy. However, most approaches adopt a single EMA teacher with a fixed update rate, even in multi-branch architectures.

3 Method

In the semi-supervised learning setting, the training set consists of a small number of labeled data $D_L = \{(x_i^l, y_i^l)\}_{i=1}^N$, and a large number of unlabeled data $D_U = \{x_i^u\}_{i=N+1}^{N+M}$, where $N \ll M$. $x_i \in \mathbb{R}^{H\times W\times D}$ represents the image, and $y_i \in \{0, 1, 2, \ldots, C-1\}^{H\times W\times D}$ is ground truth with C categories. Our goal is to train a high-quality medical image segmentation model by effectively leveraging D_L and D_U. To achieve this, we propose an SSMIS framework based on dual-EMA teacher models, as shown in Fig. 3.

3.1 Dual-EMA Teachers as Frequency-Selective Filters for Global–Local Semantics

The standard form of EMA [27] can be defined as:

$$\theta_T^{(t)} = \alpha\theta_T^{(t-1)} + (1-\alpha)\theta_S^{(t)}, \tag{1}$$

where θ_S and θ_T denote the student's parameter and the teacher's parameter, respectively. α represents the smoothing coefficient, and t means the training iteration.

In digital signal processing, the standard difference equation of first-order IIR filter [23] is defined as:

$$y[n] = ay[n-1] + bx[n]. \tag{2}$$

By aligning Eq. (1) with Eq. (2), we can establish the correspondence: $y[n] \leftrightarrow \theta_T^{(t)}, y[n-1] \leftrightarrow \theta_T^{(t-1)}, x[n] \leftrightarrow \theta_S^{(t)}, a = \alpha, b = 1-\alpha$. This shows that the EMA update rule exactly conforms to the standard first-order IIR filter formulation. Therefore, EMA is mathematically equivalent to a

first-order IIR low-pass filter, enabling its behavior to be analyzed in the frequency domain. The transfer function of a first-order IIR filter in the Z-domain is:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b}{1 - az^{-1}}. \tag{3}$$

Then, since $b = 1 - \alpha$, the transfer function of the EMA can be written as: $H(z) = \frac{1-\alpha}{1-\alpha z^{-1}}$. Following standard discrete-time signal processing practice [23], evaluating the transfer function on the unit circle ($z = e^{j\omega}, \omega \in [0, \pi]$) yields the frequency response of the EMA:

$$H(e^{j\omega}) = \frac{1-\alpha}{1-\alpha e^{-j\omega}}, \left|H(e^{j\omega})\right| = \frac{1-\alpha}{\sqrt{1+\alpha^2-2\alpha\cos\omega}}. \tag{4}$$

Fig. 2(a) visualizes the magnitude response $\left|H(e^{j\omega})\right|$ of EMA as the first-order IIR low-pass filter: low-frequency variations ($\omega \to 0$) are largely preserved, while high-frequency oscillations ($\omega \to \pi$) are strongly attenuated. From this view, the smoothing coefficient α controls the temporal scale of the EMA, determining how quickly the teacher responds to parameter changes: a larger α enhances smoothing effects, capturing low-frequency long-term trends, while a smaller α enables faster adaptation to high-frequency short-term changes.

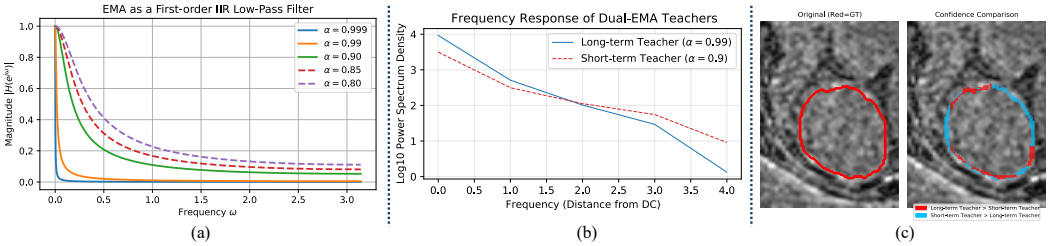

Figure 2: (a) The frequency response of EMA under different smoothing coefficients, with its frequency selectivity controlled by the weighting coefficients. (b) FFT analysis: the short-term teacher (red) retains more high-frequency energy, while the long-term teacher (blue) dominates in low-frequency; (c) Prediction confidence comparison. Blue indicates higher short-term teacher confidence, and red the opposite.

Moreover, this frequency selectivity of EMA is related to parameter update dynamics in SSMIS. High-frequency components, such as edges and fine boundaries, may contribute larger or more fluctuating gradients [4, 25], leading to faster-changing parameter updates, whereas low-frequency components, such as smooth regions and overall shapes, usually provide more stable supervision signals and induce smoother parameter evolution. Therefore, slowly evolving (long-term) parameters are expected to emphasize stable global anatomical structures, whereas rapidly varying (short-term) parameters can better preserve rapidly changing local details. This observation inspires our dual-EMA teachers design, including a long-term teacher $f_\theta^{T_g}$ that captures global semantics and a short-term teacher $f_\theta^{T_l}$ that focuses on local structural details, providing complementary guidance to the student. In practice, these two teachers are instantiated using EMA coefficients $\alpha_g = 0.99$ and $\alpha_l = 0.9$, respectively.

To examine whether the two teachers indeed exhibit the expected complementarity, we further conduct feature map FFT and boundary-confidence analyses. By calculating the radially averaged power spectral density of the feature maps from the final encoder layer within the same training iteration (as shown in Fig. 2(b)), we clearly observe their distinct frequency responses. It shows that the long-term teacher favors low frequencies, while the short-term teacher retains significantly higher energy in high-frequency regions ($>$1.8), which directly verifies the difference in frequency domain selection between the two teachers. Moreover, Fig. 2(c) shows that the short-term teacher tends to exhibit higher confidence along the boundary band, especially near the outer contour, suggesting greater sensitivity to local details. This provides empirical support that frequency-selection properties of the two teachers are reflected in complementary semantic tendencies.

Having established dual-EMA teachers that respectively encode global and local semantics, the remaining challenge is how to distill such complementary guidance into the student effectively during training. To address this, we first introduce a FIFO replay queue to reuse reliable pseudo-labeled pairs across iterations to mitigate mini-batch sampling bias (Sec. 3.2), and then apply global–local feature consistency to make the distillation robust to pseudo-label noise (Sec. 3.3).

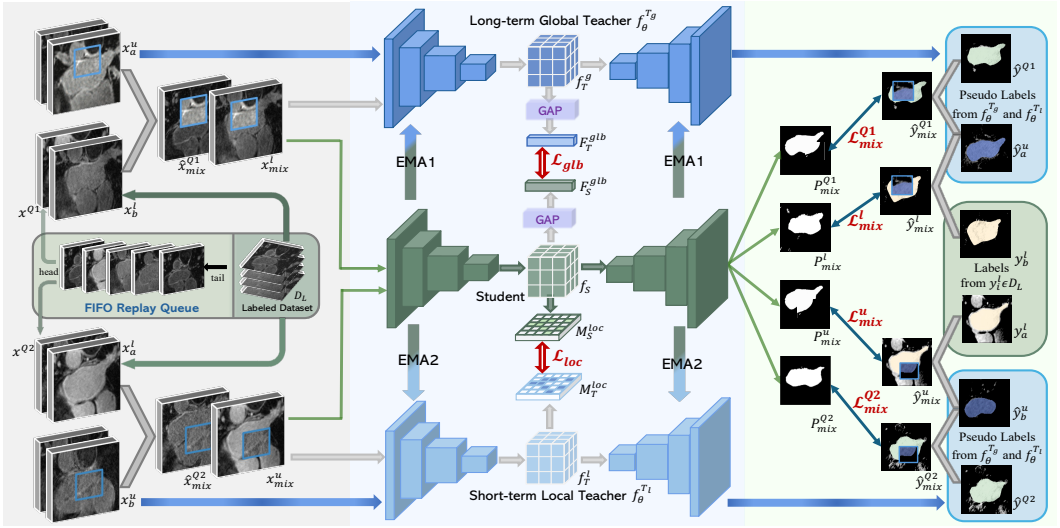

Figure 3: Overview of the proposed framework. The central blue region presents the dual-EMA teachers framework (EMA1$\leftrightarrow \alpha_g$, EMA2$\leftrightarrow \alpha_l$) and the feature-level consistency constraints ($\mathcal{L}_{glb}$ and $\mathcal{L}_{loc}$). The left gray part shows the FIFO replay queue for reusing high-confidence pseudo labels. The right green part illustrates the replay-enhanced label-level supervision, including labeled-guided and replay-guided.

3.2 Replay Queue for Pseudo Supervision Reuse

We maintain a FIFO replay queue $\mathcal{Q}$ to store high-confidence unlabeled samples and their pseudo labels generated by the dual-EMA teachers. Each iteration enqueues one pseudo-labeled pair, and the queued pairs are later replayed to provide additional training signals for the student.

Pseudo-Label Generation and Confidence Evaluation: At each iteration, we sample unlabeled volumes x_a^u and x_b^u from the unlabeled set D_U and generate predictions:

$$P_a^g = f_\theta^{T_g}(x_a^u), \quad P_a^l = f_\theta^{T_l}(x_a^u), \quad P_b^g = f_\theta^{T_g}(x_b^u), \quad P_b^l = f_\theta^{T_l}(x_b^u). \tag{5}$$

Here, $P_{a/b}^g$ and $P_{a/b}^l$ denote the predictions from the long-term global teacher and the short-term local teacher, respectively. Since the two teachers encode complementary global and local cues, we fuse their predictions to obtain more robust pseudo labels: $P_{a/b}^f = 0.5 \times P_{a/b}^g + 0.5 \times P_{a/b}^l$. This fusion can be viewed as a lightweight two-teacher ensemble, which reduces prediction variance and mitigates the bias of relying on a single teacher. As the relative reliability of the two teachers may vary during training and no prior favors either one, we adopt equal weights for simplicity and robustness. The pseudo labels are then obtained by $\hat{y}_{a/b}^u = argmax(\sigma(P_{a/b}^f))$, where σ denotes the softmax function.

To assess the reliability of pseudo labels for replay selection, we compute the average maximum softmax probability as a confidence metric:

$$\begin{aligned} C(x_{a/b}^u)^{g/l} &= \frac{1}{|\Omega|} \sum_{p \in \Omega} \max_{k \in \{0,\dots,C-1\}} \sigma\big(f_\theta^{T_g/T_l}(x_{a/b}^u)\big)_k, \\ C(x_{a/b}^u) &= 0.5 \times C(x_{a/b}^u)^g + 0.5 \times C(x_{a/b}^u)^l, \end{aligned} \tag{6}$$

where Ω denotes the set of voxels. The higher-confidence pair among $\{(x_a^u, \hat{y}_a^u), (x_b^u, \hat{y}_b^u)\}$ is enqueued as $(x^Q, \hat{y}^Q)$:

$$(x^Q, \hat{y}^Q) = \begin{cases} (x_a^u, \hat{y}_a^u), & C(x_a^u) \geq C(x_b^u), \\ (x_b^u, \hat{y}_b^u), & \text{otherwise.} \end{cases} \tag{7}$$

FIFO Replay Queue Update: At each iteration, the selected pair $(x^Q, \hat{y}^Q)$ is enqueued, and the oldest entry is removed once the queue is full:

$$\mathcal{Q}_t = \text{FIFO}\big(\mathcal{Q}_{t-1}, (x_t^Q, \hat{y}_t^Q); K\big). \tag{8}$$

In this way, $\mathcal{Q}$ acts as a lightweight *teacher-memory*, enabling the reuse of supervision signals from previous iterations. Furthermore, we enqueue $(x_t^Q, \hat{y}_t^Q)$ without comparing its confidence against items $\mathcal{Q}_{t-1}$ already in the queue. This always-enqueue strategy maintains the diversity of replay samples and prevents the queue from being dominated by a small set of "easy-to-segment/high-confidence" samples.

Replay-Enhanced Label-Level Supervision: In this section, we first present the base label-level supervision derived from labeled data, and then describe how the replay queue is used to provide replay-guided label-level supervision. Both are implemented using the bidirectional copy-paste introduced in BCP [2] to improve performance and generalization.

1. Labeled-guided supervision: Given labeled samples (x_a^l, y_a^l) and (x_b^l, y_b^l), along with unlabeled samples and their pseudo labels, we generate a random binary mask $M \in \{0,1\}^{H\times W\times D}$ with a zero-valued region of size $\beta H \times \beta W \times \beta D, \beta \in (0,1)$, and construct mixed samples as follows:

$$\begin{aligned} x_{mix}^l &= M \odot x_a^u + (1-M) \odot x_b^l, \quad \hat{y}_{mix}^l = M \odot \hat{y}_a^u + (1-M) \odot y_b^l, \\ x_{mix}^u &= M \odot x_a^l + (1-M) \odot x_b^u, \quad \hat{y}_{mix}^u = M \odot y_a^l + (1-M) \odot \hat{y}_b^u, \end{aligned} \tag{9}$$

where $\odot$ denotes element-wise multiplication. Then, mixed samples are fed into the student model to obtain predictions $P_{mix}^{l/u} = f_\theta^S(x_{mix}^{l/u})$, which are optimized by minimizing:

$$\mathcal{L}_{mix}^{l/u} = \mathcal{L}_{DiceCE}(P_{mix}^{l/u}, \hat{y}_{mix}^{l/u}; M) + \frac{1}{2}\mathcal{L}_{DiceCE}(P_{mix}^{l/u}, \hat{y}_{mix}^{l/u}; (1-M)), \tag{10}$$

where $\mathcal{L}_{DiceCE} = \frac{1}{2}\mathcal{L}_{Dice} + \frac{1}{2}\mathcal{L}_{CE}$.

2. Replay-guided supervision: Once sufficient samples have accumulated in the queue, we randomly sample two historical pairs $(x^{Q_1}, \hat{y}^{Q_1})$ and $(x^{Q_2}, \hat{y}^{Q_2})$ from the queue $\mathcal{Q}_t$ and mix them with current unlabeled samples:

$$\begin{aligned} \hat{x}_{mix}^{Q1} &= M \odot x_a^u + (1-M) \odot x^{Q_1}, \hat{y}_{mix}^{Q1} = M \odot \hat{y}_a^u + (1-M) \odot \hat{y}^{Q_1}, \\ \hat{x}_{mix}^{Q2} &= M \odot x^{Q_2} + (1-M) \odot x_b^u, \hat{y}_{mix}^{Q2} = M \odot \hat{y}^{Q_2} + (1-M) \odot \hat{y}_b^u. \end{aligned} \tag{11}$$

Similarly, for these mixed samples, we obtain the student model's prediction $P_{mix}^{Q1/Q2} = f_\theta^S(\hat{x}_{mix}^{Q1/Q2})$ and then optimize using the same loss:

$$\mathcal{L}_{mix}^{Q1/Q2} = \mathcal{L}_{DiceCE}(P_{mix}^{Q1/Q2}, \hat{y}_{mix}^{Q1/Q2}; M) + \frac{1}{2}\mathcal{L}_{DiceCE}(P_{mix}^{Q1/Q2}, \hat{y}_{mix}^{Q1/Q2}; (1-M)). \tag{12}$$

The replay queue preserves historical pseudo-labeled pairs, which encode semantic information from dual teachers across different iterations. Replay-guided supervision helps the student leverage this information more effectively.

3.3 Feature-level Global–Local Consistency

To reduce the impact of noisy pseudo labels and better exploit complementary semantics introduced by dual-EMA teachers, we further introduce feature-level consistency losses.

Global Semantic Consistency Loss: Owing to the smoothly updated parameters of the long-term global teacher $f_\theta^{T_g}$, it provides stable and reliable global representations to guide the student. To transfer such global semantic knowledge, we extract a compact global feature from the deepest encoder feature map f_T^g using global average pooling (GAP):

$$F_T^{glb} = GAP(f_T^g). \tag{13}$$

Similarly, the student generates the corresponding global feature $F_S^{glb} = GAP(f_S)$ from the same input. We then impose a global semantic consistency loss based on $L1$ distance:

$$\mathcal{L}_{glb} = \left\| F_S^{glb} - F_T^{glb} \right\|_1. \tag{14}$$

This loss encourages the student to align with the long-term global teacher in terms of global semantic representations, such as organ-level structure and overall anatomical layout.

Local Relation Consistency Loss: Rather than enforcing point-wise feature alignment, which is sensitive to local noise and spatial misalignment, we adopt a relation-based consistency constraint to align the structural relationships among spatial locations. This design enables robust distillation of local geometric and boundary information from the short-term local teacher $f_\theta^{T_l}$ to the student f_θ^S.

Specifically, we extract the deepest encoder feature maps from $f_\theta^{T_l}$ and f_θ^S, denoted as f_T^l and f_S, respectively, and then normalize them along the channel dimension to obtain F_T^{loc} and F_S^{loc}.

Next, we compute the self-correlation matrix between positions to characterize the cosine-similarity relations across all spatial positions:

$$M_S^{loc} = {F_S^{loc}}^\top F_S^{loc}, \quad M_T^{loc} = {F_T^{loc}}^\top F_T^{loc}. \tag{15}$$

The local relation consistency loss is defined as the mean squared error between the student's and teacher's region-based correlation:

$$\mathcal{L}_{loc} = (M_S^{loc} - M_T^{loc})^2. \tag{16}$$

Then, the total feature-level loss is: $\mathcal{L}_{Feat} = \mathcal{L}_{glb} + \mathcal{L}_{loc}$.

3.4 The Overall Training Objective

Inspired by previous SSMIS studies [31], we first conduct a supervised pre-training stage using labeled data only. Specifically, we perform bidirectional copy-paste on labeled samples (x_i^l, y_i^l) and (x_j^l, y_j^l) sampled from D_L, yielding mixed inputs and labels: $X_{mix}^l = x_i^l \odot M + x_j^l \odot (1-M)$, $Y_{mix}^l = y_i^l \odot M + y_j^l \odot (1-M)$, where M denotes a binary mask. The pre-training objective is defined as:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{DiceCE}(f_\theta(X_{mix}^l), Y_{mix}^l). \tag{17}$$

After pre-training, we initialize the models using the pretrained weights obtained from Eq. 17. Then, we jointly leverage both labeled and unlabeled data for self-training, and the overall training objective is formulated as:

$$\mathcal{L}_{self} = \mathcal{L}_{LG} + \mu\mathcal{L}_{RQG} + \varphi\mathcal{L}_{Feat}, \tag{18}$$

where $\mathcal{L}_{LG} = \mathcal{L}_{mix}^l + \mathcal{L}_{mix}^u$ represents the label-level loss induced by labeled-guided, and $\mathcal{L}_{RQG} = \mathcal{L}_{mix}^{Q1} + \mathcal{L}_{mix}^{Q2}$ corresponds to the replay-guided loss. μ and φ represent the weights for the replay-guided loss and feature-level loss, respectively.

In the test phase, only the student model f_θ^S is used. Given a test sample X_{test}, the predicted probability map is obtained as $P_{test} = f_\theta^S(X_{test})$, and the final segmentation is produced through $Y_{test} = argmax(\sigma(P_{test}))$.

4 Experiments and Results

The proposed method was comprehensively evaluated on three widely adopted public datasets for SSMIS: LA [34], PROMISE12 [16], and BraTS19 [22]. We evaluate the model using four widely-used metrics: Dice Score (%), Jaccard Score (%), 95% Hausdorff Distance (95HD) in voxel, and Average Surface Distance (ASD) in voxel. The first two metrics assess the volumetric overlap of segmented regions, while the latter two respectively reflect the maximum and average values of boundary error. Due to space constraints, further details regarding the datasets, method implementation, and additional experimental results such as parameter selection can be found in the Appendix.

4.1 Comparison with State-of-the-art Methods

LA dataset: We compare our method with ten SoTA approaches under two settings with 5% and 10% labeled data, and the quantitative results are reported in Table 1. As shown, our method significantly outperforms the SoTA methods and achieves the best performance under both labeled ratios. In particular, with 10% labeled data, our method attains a Dice score of 91.98 and reduces 95HD to 4.74; in comparison, the recent method [12] reports a Dice of 91.07 and 95HD of 4.96.

PROMISE12 dataset: Our method also achieves outstanding performance on the 2D PROMISE12 dataset. As shown in Table 2, under both the 10% and 20% labeled settings, it attains the best

Table 1: Quantitative comparison with SoTA methods on LA dataset using two labeled data settings (5% and 10%). The best and second-best results are denoted in red and blue.

Method	Volumes used in **LA**		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	95HD↓	ASD↓
VNet (SupOnly)	4 (5%)	0	52.55	39.60	47.05	9.87
VNet (SupOnly)	8 (10%)	0	82.74	71.72	13.35	3.26
UA-MT [37]	4 (5%)	76 (95%)	82.26	70.98	13.71	3.82
DTC [20]			81.25	69.33	14.90	3.99
SS-Net [32]			86.33	76.15	9.97	2.31
MCF [30]			86.52	77.43	9.12	2.40
BCP [2]			88.02	78.72	7.90	2.15
CML [31]			87.63	—	8.92	2.23
AD-MT [41]			89.63	81.28	6.56	1.85
GapMatch [11]			88.30	—	—	1.77
UCSeg [10]			90.36	82.48	6.35	1.76
Kumari et al. [12]			89.55	81.18	5.48	1.99
Ours			**90.90**	**83.37**	**4.91**	**1.73**
UA-MT [37]	8 (10%)	72 (90%)	86.28	76.11	18.71	4.63
DTC [20]			87.51	78.17	8.23	2.36
SS-Net [32]			88.55	79.62	7.49	1.90
MCF [30]			88.71	80.41	6.32	1.90
BCP [2]			89.62	81.31	6.81	1.76
CML [31]			90.36	—	6.06	1.68
AD-MT [41]			90.55	82.79	5.81	1.70
GapMatch [11]			91.00	—	—	1.46
UCSeg [10]			90.90	83.41	6.50	1.62
Kumari et al. [12]			91.07	83.67	4.96	1.65
Ours			**91.98**	**85.20**	**4.74**	**1.45**

Table 2: Quantitative comparison with SoTA methods on PROMISE12 dataset using two labeled data settings (10% and 20%). The best and second-best results are denoted in red and blue.

Method	Volumes used in **PROMISE12**		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	95HD↓	ASD↓
UNet (SupOnly)	4 (10%)	0	51.72	39.02	43.30	11.81
UNet (SupOnly)	7 (20%)	0	65.63	52.44	12.36	2.98
URPC [21]	4 (10%)	31 (90%)	52.96	39.93	37.53	11.43
DTC [20]			57.87	43.80	81.54	25.38
SS-Net [32]			61.10	47.07	23.73	7.44
SCP-Net [40]			66.21	—	—	11.56
BCP [2]			76.11	62.07	47.08	11.02
BMT-DCCL [15]			74.92	61.59	12.52	4.08
MCMCCL [5]			81.39	68.88	5.57	1.36
Ours			**84.44**	**73.26**	**2.59**	**1.09**
URPC [21]	7 (20%)	28 (80%)	67.04	54.01	11.54	2.11
DTC [20]			72.03	58.32	11.48	2.65
SS-Net [32]			71.56	59.35	14.38	3.03
SCP-Net [40]			77.06	—	—	3.52
BCP [2]			78.42	65.33	39.65	9.21
ABD [6]			82.06	—	—	1.33
AD-MT [41]			79.82	—	—	1.77
BMT-DCCL [15]			82.25	68.32	6.32	2.17
β-FFT [8]			83.75	—	—	1.20
MCMCCL [5]			84.16	73.00	3.68	1.17
Ours			**86.15**	**75.80**	**2.42**	**0.95**

results on all four metrics, yielding a clear improvement over the next-best method MCMCCL [5]. Meanwhile, Fig. 4 further provides qualitative comparisons on both LA and PROMISE12, where our method produces more accurate and complete segmentation with smoother boundaries and better preservation of anatomical details. Overall, these results highlight the robustness and generalization ability of the proposed approach.

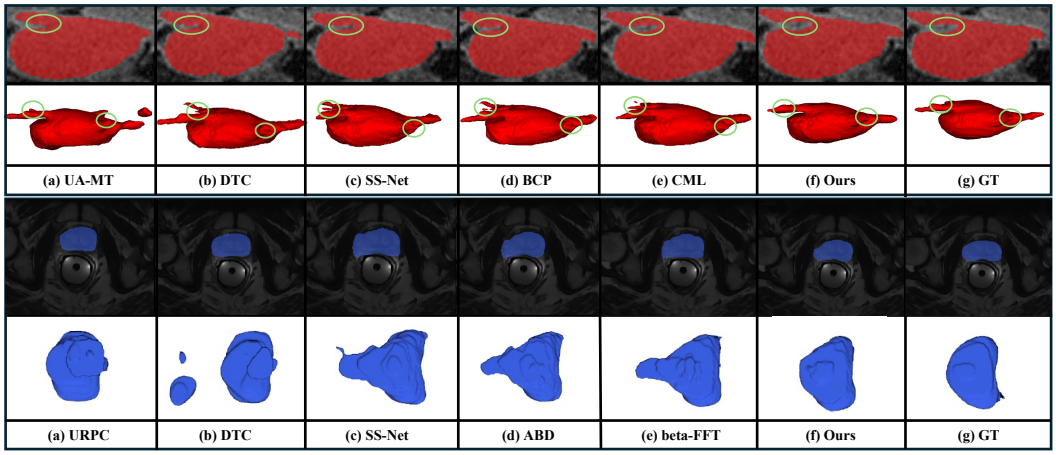

Figure 4: Qualitative comparisons on the LA and PROMISE12 datasets (2D/3D segmentation visualization). The top shows results on LA with 10% labeled data, and the bottom shows results on PROMISE12 with 20% labeled data.

4.2 Ablation Studies

We conduct ablation experiments using LA (10% labeled) and PROMISE12 (20% labeled) datasets to validate the impact of each component in the proposed method.

EMA Coefficient Analysis in Mean Teacher: To investigate how different EMA update coefficients α affect the model's ability to capture semantic information, we conduct experiments on the LA dataset with 10% labeled data under the standard MT framework. We select α from $\{0.999, 0.99, 0.9, 0.85, 0.8\}$ and report the results in Table 3. It can be observed that larger α consistently yields higher Dice and Jaccard scores, indicating better capture of regional structures. In contrast, for boundary metrics 95HD and ASD within $\alpha \in [0.9, 0.999]$, smaller α tends to perform

better, suggesting improved sensitivity to local details. However, when α further decreases to 0.85 or 0.8, performance drops, since an overly small α weakens EMA stabilization and causes unstable teacher predictions, introducing more uncertainty and degrading performance.

Effectiveness of Dual-EMA Teachers Framework: To validate the effectiveness of the dual-EMA design, we compare the Long-term Global Teacher (LGT), Short-term Local Teacher (SLT), and Student on PROMISE12 at the checkpoint where the student performs best. As shown in Fig. 5(a), LGT obtains higher Dice/Jaccard, suggesting stronger global region modeling, while SLT achieves lower 95HD/ASD, indicating better boundary delineation. This supports our expected global–local complementarity. To further investigate their individual contributions, we conduct an ablation study on the LA and PROMISE12 datasets for dual-EMA teachers, including w/o SLT (removing SLT), w/o LGT (removing LGT), and the full model, with results reported in Table 5. It should be noted that removing one teacher is functionally equivalent in terms of predictions to assigning identical EMA coefficients to both teachers (i.e., $\alpha_g = \alpha_l = 0.99$ or $\alpha_g = \alpha_l = 0.90$). Under identical initialization and synchronized updates, the two teachers maintain identical parameters and predictions throughout training, degenerating into a single Mean Teacher without additional prediction diversity.

Table 3: Model's performance using different update coefficients under the MT framework on LA dataset using 10% labeled data.

α	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
0.999	88.56	78.47	8.23	2.01
0.99	87.61	77.62	8.16	1.98
0.9	85.98	76.07	8.09	1.91
0.85	85.54	75.35	9.02	2.25
0.8	84.65	74.31	10.51	2.41

Table 4: Ablation studies on loss components on the LA dataset with 10% labeled data, validating the effectiveness of replay queue for pseudo-supervision reuse (RQ) and feature-level consistency (FC).

Method	$\mathcal{L}_{LG}$	$\mathcal{L}_{RQG}$	$\mathcal{L}_{Feat}$	Dice(%)↑	Jaccard↑	95HD↓	ASD↓
Dual-EMA (base)	✓			90.87	83.50	5.75	1.75
Dual-EMA + RQ	✓	✓		91.54	84.51	5.28	1.60
Dual-EMA + FC	✓		✓	91.16	84.12	5.42	1.68
Full Method	✓	✓	✓	**91.98**	**85.20**	**4.74**	**1.45**

As shown in Table 5, removing the SLT and using only the LGT leads to a decline in performance on boundary metrics, with ASD increasing from 0.95 to 2.16 on the PROMISE12 dataset. In contrast, removing the LGT and using only the SLT makes the model more sensitive to local details; thus, it performs better on boundary metrics but worse on region-level metrics, with Dice decreasing from 91.98 to 90.88 on the LA dataset. More importantly, the full model consistently outperforms setup using only a single teacher, achieving simultaneous improvements in both regional and boundary metrics. These findings indicate that the student can effectively absorb the complementary semantic knowledge captured by the dual-EMA teachers and translate it into improved segmentation performance.

Furthermore, these results also suggest that the performance gains cannot be simply attributed to an increased number of teachers or general ensemble stabilization. To further investigate this, we evaluate a dual-EMA teacher configuration with closer EMA coefficients ($\alpha_g = 0.99$, $\alpha_l = 0.95$), which achieves a Dice of 84.26, a Jaccard of 73.54, a 95HD of 3.86, and an ASD of 1.65 on PROMISE12. Although this configuration introduces some temporal-scale discrepancy and benefits from prediction fusion, it still underperforms our proposed configuration ($\alpha_g = 0.99$, $\alpha_l = 0.90$), which achieves 86.15, 75.80, 2.42, and 0.95 on the same metrics, respectively. These findings suggest that the improvements stem not merely from ensemble stabilization, but more importantly from the complementary global and local semantic knowledge induced by distinct EMA update rates. Moreover, our method also outperforms AD-MT [41], which similarly employs a dual-teacher framework, achieving a Dice of 91.98 vs 90.55 and an ASD of 1.45 vs 1.70 on LA. These findings further support the effectiveness of our proposed dual-EMA design.

Table 5: Ablation studies for the dual-EMA teachers framework on LA dataset and PROMISE12 dataset.

Model	LA dataset (10% labeled)				PROMISE12 dataset (20% labeled)			
	Dice(%)↑	Jaccard↑	95HD↓	ASD↓	Dice(%)↑	Jaccard↑	95HD↓	ASD↓
w/o Long-term Global Teacher	90.88	83.68	5.45	1.66	82.11	71.89	5.28	2.01
w/o Short-term Local Teacher	91.06	84.01	5.64	1.72	82.72	72.28	5.52	2.16
Full Method	**91.98**	**85.20**	**4.74**	**1.45**	**86.15**	**75.80**	**2.42**	**0.95**

Ablation on Replay Queue and Feature Consistency: To further validate the effectiveness of our proposed FIFO-queue-based pseudo-supervision reuse strategy and feature-level global–local consistency loss, we ablate the loss components and report the results in Table 4 and Fig. 5(b).

Specifically, $\mathcal{L}_{LG}$ supervises the samples mixed from labeled and unlabeled data, $\mathcal{L}_{RQG}$ performs label-level supervision on samples constructed by mixing high-confidence unlabeled pairs from the replay queue with the current unlabeled data, and $\mathcal{L}_{Feat}$ provides feature-level supervision, including the global semantic and the local relation consistency loss.

The results clearly show that removing any loss leads to varying degrees of degradation in model performance, demonstrating that each component is beneficial. Specifically, $\mathcal{L}_{LG}$ provides foundational labeled supervision, ensuring the model's baseline performance. $\mathcal{L}_{RQG}$ further improves performance by leveraging historical high-confidence pseudo-labeled pairs stored in a replay queue to introduce more useful training signals. Concretely, on the PROMISE12 dataset, $\mathcal{L}_{RQG}$ increases the Dice score by 2.58 and reduces the ASD by 0.63. Additionally, $\mathcal{L}_{Feat}$ enhances robustness through feature-level consistency, mitigating the impact of noisy pseudo labels.

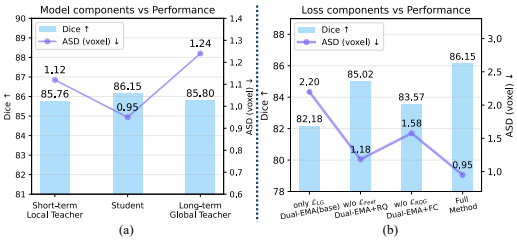

Figure 5: Ablation studies on PROMISE12 dataset using 20% labeled data. (a) Performance comparison of the three models in the proposed framework. (b) Ablation for loss components.

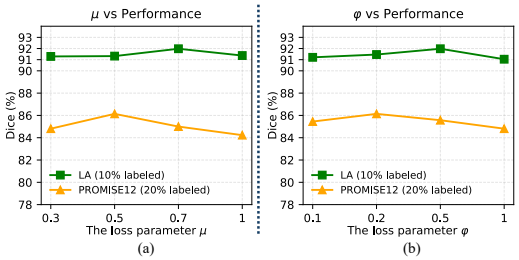

Figure 6: The loss weight sensitivity analysis for μ and φ on LA dataset (10% labeled) and PROMISE12 dataset (20% labeled).

Sensitivity Analysis of Loss Weights: We conducted a sensitivity analysis on the loss weight parameters μ and φ, as shown in Fig. 6, where μ and φ control the contributions of the label-level loss provided by replay-guided supervision and the feature-level consistency loss, respectively. The results indicate that the proposed method exhibits low sensitivity when μ ranges from 0.3 to 0.5 and φ ranges from 0.1 to 0.2. Based on these experimental findings, we set $\mu = 0.5$ and $\varphi = 0.2$ for PROMISE12, and $\mu = 0.7$ and $\varphi = 0.5$ for LA. Overall, these results demonstrate that the proposed framework exhibits no excessive sensitivity to the precise setting of loss weights, which validates the robustness of the proposed method.

5 Conclusion

In this work, we interpret EMA from a signal processing perspective, revealing its frequency-selective behavior. Based on this insight, we propose a dual-EMA teachers framework for SSMIS, where a long-term teacher focuses on global structures and a short-term teacher emphasizes local details. To distill this complementary guidance into the student, we further introduce a FIFO replay queue to reuse reliable pseudo-labeled pairs and mitigate sampling bias, together with feature-level global–local consistency to suppress pseudo-label noise propagation and improve robustness. Extensive experiments on three public medical image datasets demonstrate consistent improvements over SoTA methods.

6 Acknowledgment

This work was supported in part by National Key Research and Development Program of China (No. 2024YFC2310801) and National Natural Science Foundation of China (No. 62476052).

References

[1] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11514–11524, 2023.

[3] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19786–19797, June 2023.

[4] Daniel Morales Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research Journal*, 2024.

[5] Xunhang Cao, Xiaoxin Guo, Guangqi Yang, Qi Chen, and Hongliang Dong. Multi-view cross-consistency and multi-scale cross-layer contrastive learning for semi-supervised medical image segmentation. *Expert Systems with Applications*, 277:127223, 2025.

[6] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4070–4080, 2024.

[7] Ning Gao, Sanping Zhou, Le Wang, and Nanning Zheng. Pmt: Progressive mean teacher via exploring temporal consistency for semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pages 144–160. Springer, 2024.

[8] Ming Hu, Jianfu Yin, Zhuangzhuang Ma, Jianheng Ma, Feiyu Zhu, Bingbing Wu, Ya Wen, Meng Wu, Cong Hu, Bingliang Hu, et al. beta-fft: Nonlinear interpolation and differentiated training strategies for semi-supervised medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30839–30849, 2025.

[9] Huimin Huang, Yawen Huang, Shiao Xie, Lanfen Lin, Tong Ruofeng, Yen-wei Chen, Yuexiang Li, and Yefeng Zheng. Semi-supervised convolutional vision transformer with bi-level uncertainty estimation for medical image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5214–5222, 2023.

[10] Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, and Xiao-Jun Wu. Uncertainty-aware cross-training for semi-supervised medical image segmentation. *IEEE Transactions on Image Processing*, 2025.

[11] Wei Huang, Lei Zhang, Zizhou Wang, and Yan Wang. Gapmatch: Bridging instance and model perturbations for enhanced semi-supervised medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17458–17466, 2025.

[12] Suruchi Kumari and Pravendra Singh. Leveraging fixed and dynamic pseudo-labels in cross-supervision framework for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3530–3540, June 2025.

[13] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2021. doi: 10.1109/TNNLS.2020.2995319.

[14] Yuxin Li, Zihao Zhu, Yuxiang Zhang, Yifan Chen, and Zhibin Yu. Boost the inference with co-training: A depth-guided mutual learning framework for semi-supervised medical polyp segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10394–10403, 2025.

[15] Zhikang Lin, Zhenteng Li, Jiawei Su, Lei Li, and Sheng Lian. Enhancing semi-supervised medical image segmentation with boundary awareness and competitive contrastive learning. *Biomedical Signal Processing and Control*, 103:107450, 2025.

[16] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18 (2):359–373, 2014.

[17] Yuyuan Liu, Yu Tian, Chong Wang, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Translation consistent semi-supervised segmentation for 3d medical images. *IEEE Transactions on Medical Imaging*, 2024.

[18] Shanfu Lu, Ziye Yan, Wei Chen, Tingting Cheng, Zijian Zhang, and Guang Yang. Dual consistency regularization with subjective logic for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 170:107991, 2024.

[19] Yayuan Lu, Zheyu Zhang, Yueyi Zhang, Yunwei Ou, and Xiaoyan Sun. Semi-supervised medical image segmentation via dynamic pseudo-label refinement. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[20] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8801–8809, 2021.

[21] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–329. Springer, 2021.

[22] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[23] Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

[24] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[25] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[26] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 94: 103111, 2024.

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[28] Bing Wang, Taifeng Huang, Shuo Yang, Ying Yang, Junhai Zhai, and Xin Zhang. Dual-decoder data decoupling training for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 100:106984, 2025.

[29] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5): 1243–1267, 2022.

[30] Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. Mcf: Mutual correction framework for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15651–15660, 2023.

[31] Song Wu, Xiaoyu Wei, Xinyue Chen, Yazhou Ren, Jing He, and Xiaorong Pu. Cross-view mutual learning for semi-supervised medical image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9253–9261, 2024.

[32] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 34–43. Springer, 2022.

[33] Jinyi Xiang, Peng Qiu, and Yang Yang. Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–491. Springer, 2022.

[34] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.

[35] 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. *Medical Image Analysis*, 88:102880, 2023.

[36] Sicheng Yang, Zhaohu Xing, and Lei Zhu. Vq-seg: Vector-quantized token perturbation for semi-supervised medical image segmentation. *Advances in Neural Information Processing Systems*, 38:22148–22172, 2026.

[37] 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 *International conference on medical image computing and computer-assisted intervention*, pages 605–613. Springer, 2019.

[38] Yichi Zhang and Jicong Zhang. Dual-task mutual learning for semi-supervised medical image segmentation. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4*, pages 548–559. Springer, 2021.

[39] Yichi Zhang, Rushi Jiao, Qingcheng Liao, Dongyang Li, and Jicong Zhang. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. *Artificial Intelligence in Medicine*, 138:102476, 2023. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2022.102476. URL `https://www.sciencedirect.com/science/article/pii/S0933365722002287`.

[40] Zhenxi Zhang, Ran Ran, Chunna Tian, Heng Zhou, Xin Li, Fan Yang, and Zhicheng Jiao. Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–201. Springer, 2023.

[41] Zhen Zhao, Zicheng Wang, Longyue Wang, Dian Yu, Yixuan Yuan, and Luping Zhou. Alternate diverse teaching for semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pages 227–243. Springer, 2024.

[42] Xinghu Zhou, Guanghan Wang, Yuanzhi Cheng, Zixuan Wang, Xin Wang, Guohua Wang, and Shinichi Tamura. Joint learning of confidence fusion, semantic alignment and group-guided reliability: A novel semi-supervised learning framework for 3d medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

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 contributions presented in the abstract and introduction accurately reflect the paper's contributions and scope of research, and highlight the methodological innovations and validation results.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 limitations of our proposed method are discussed in the Appendix G.

   Guidelines:

   - The answer [N/A] 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: [Yes]

Justification: We provided a general theoretical explanation in the Method section 3.1 of the main text and presented a detailed discussion in the appendix F.

Guidelines:

- The answer [N/A] 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 information needed to reproduce the main experimental results is introduced in Appendix A, B, and the supplementary material.

   Guidelines:

   - The answer [N/A] 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 instructions on how to obtain the code and dataset in the supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details can be seen in Appendix B and the supplementary material.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: Our experiment did not include error bars, confidence intervals, or other statistical significance tests.

   Guidelines:

   - The answer [N/A] 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: The computer resources required to complete the experiment are detailed in Appendix B and E.

   Guidelines:

   - The answer [N/A] 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, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: We talk about the broader impacts in Appendix G.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not deal with this aspect of the problem.

    Guidelines:

    - The answer [N/A] 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 assets (datasets, code, models) used in the paper are open source, and we cite the original paper.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

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

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

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

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.

Appendix

A Datasets

LA dataset [34] is specifically designed for medical image segmentation tasks, originating from the 2018 Atrial Segmentation Challenge. It contains 100 3D GE-MRI scans of the left atrium. Each scan is provided with an isotropic spatial resolution of $0.625 \times 0.625 \times 0.625 mm^3$. In the context of semi-supervised learning, it is standard practice to partition the 100 volumes into 80 training scans and 20 testing scans.

PROMISE12 Dataset [16] was introduced as part of MICCAI 2012 Prostate MR Image Segmentation. This dataset comprises 50 MRI images acquired from multiple centers using different MRI vendors and various scanning protocols, resulting in significant diversity in image contrast and resolution. Following standard protocols, the dataset is partitioned into 35 volumes for training, 5 for validation, and 10 for independent testing.

BraTS19 Dataset [22] (Brain Tumor Segmentation Challenge 2019) is a premier benchmark for evaluating automated brain tumor segmentation algorithms. It comprises 335 pre-operative scans of patients diagnosed with either High-Grade Glioma (HGG) or Low-Grade Glioma (LGG). Each case provides four modalities (FLAIR, T1, T1ce, and T2) at an isotropic resolution of $1mm^3$. In this work, we split the dataset into 250 training cases, 25 validation cases, and 60 test cases.

B Implementation Details

All experiments are implemented using the PyTorch framework. For the LA dataset, we adopt VNet as the backbone network. All volumetric images are randomly cropped to a size of $112 \times 112 \times 80$ as input to the model, and the batch size is set to 8. Training includes 3k pre-training and 10k self-training iterations. Following the setting in [2], we set the mask ratio β to $\frac{2}{3}$ for both stages, and the loss weights $\mu = 0.7, \varphi = 0.5$. The replay queue size is $K = 6$. During inference, a sliding window of size $18 \times 18 \times 4$ is used to obtain full segmentation results. For the PROMISE12 dataset, we adopt UNet as the backbone. The input images are cropped to a size of 224×224, and the batch size is set to 16. The pre-training and self-training stages are conducted for 6k and 30k iterations, respectively. We set $\beta = \frac{2}{3}$, $\mu = 0.5$, and $\varphi = 0.2$, with a replay queue size of 4. For BraTS19 dataset, following previous work [35], we adopt 3D-UNet as the backbone. All images are cropped to a size of $96 \times 96 \times 96$ as input to the model. We train for 6k iterations in the pre-training stage and 30k iterations in the self-training stage. The mask ratio is kept the same as BCP [2] $\beta = \frac{2}{3}$, and the replay queue size is set to 4. Based on the parameter sensitivity analysis, we set $\mu = 0.5$ and $\varphi = 0.2$.

The Implementation Details of the Feature Map FFT analysis: In Section 3.1 of the main text, to verify the complementarity of the dual-EMA teacher, we conducted a frequency-domain analysis on the deep feature maps extracted from both the long-term global teacher model and short-term local teacher model. Specifically, we randomly sample 30 3D volumetric images from the LA dataset and forward them through both models to obtain their respective feature maps. We then apply a 3D Fast Fourier Transform (FFT) to convert the spatial feature maps into the frequency domain. To obtain the Power Spectrum Density (PSD), we compute the squared magnitude of the centered Fourier coefficients and average them across the channel dimension. Because the 3D frequency space is isotropic, we collapse the 3D PSD into a 1D radial profile by averaging the power values within concentric spherical shells of radius r (representing the spatial frequency, or the distance from the DC component). As illustrated in Fig. 2(b), we plot the logarithm of the averaged 1D PSD against the frequency sequence. The results clearly demonstrate that the long-term teacher preserves significantly higher energy in the low-frequency bands, indicating a stronger capacity to capture global anatomical structures. In contrast, the short-term teacher exhibits a relatively higher response in the high-frequency regions, suggesting that its rapid weight updates make it more sensitive to structural boundaries and local details.

C Pseudo code of our proposed method

To facilitate a better understanding of the overall method, we provide the pseudo code in Algorithm 1.

Algorithm 1: The pseudo code of proposed method

Input: labeled set $\mathcal{D}_L$, unlabeled set $\mathcal{D}_U$; student f_θ^S; long-term global teacher $f_\theta^{T_g}$; short-term local teacher $f_\theta^{T_l}$; queue size K; binary mask M

Initialization: FIFO queue $\mathcal{Q} \leftarrow \emptyset$

for *each iteration* $t = 1, 2, \ldots$ **do**

Sample labeled pairs $(x_a^l, y_a^l), (x_b^l, y_b^l) \sim \mathcal{D}_L$;

Sample unlabeled pairs $x_a^u, x_b^u \sim \mathcal{D}_U$;

`// Teacher predictions and fusion`

$P_a^g \leftarrow f_\theta^{T_g}(x_a^u)$, $P_a^l \leftarrow f_\theta^{T_l}(x_a^u)$, $P_b^g \leftarrow f_\theta^{T_g}(x_b^u)$, $P_b^l \leftarrow f_\theta^{T_l}(x_b^u)$;

$P_a \leftarrow \frac{1}{2}(P_a^g + P_a^l)$, $P_b \leftarrow \frac{1}{2}(P_b^g + P_b^l)$;

$\hat{y}_a^u \leftarrow \arg\max P_a$, $\hat{y}_b^u \leftarrow \arg\max P_b$;

`// Confidence estimation`

$C_a^{g/l} \leftarrow \frac{1}{|\Omega|}\sum_{p\in\Omega} \max_k \mathrm{Softmax}(P_a^{g/l})_k$, $C_b^{g/l} \leftarrow \frac{1}{|\Omega|}\sum_{p\in\Omega} \max_k \mathrm{Softmax}(P_b^{g/l})_k$;

$C_a = 0.5 \times C_a^g + 0.5 \times C_a^l$, $C_b = 0.5 \times C_b^g + 0.5 \times C_b^l$;

`// FIFO replay queue update`

if $C_a \geq C_b$ **then**

$\mathcal{Q}_t \leftarrow \mathrm{FIFO_Push}(\mathcal{Q}_{t-1}, (x_a^u, \hat{y}_a^u), K)$;

else

$\mathcal{Q}_t \leftarrow \mathrm{FIFO_Push}(\mathcal{Q}_{t-1}, (x_b^u, \hat{y}_b^u), K)$;

`// Labeled-guided supervision`

`// Bidirectional copy-paste`

$x_{mix}^l \leftarrow M \odot x_a^u + (1-M) \odot x_b^l$, $\hat{y}^l \leftarrow M \odot \hat{y}_a^u + (1-M) \odot y_b^l$;

$x_{mix}^u \leftarrow M \odot x_a^l + (1-M) \odot x_b^u$, $\hat{y}^u \leftarrow M \odot y_a^l + (1-M) \odot \hat{y}_b^u$;

$\mathcal{L}_{LG} \leftarrow \mathcal{L}_{mix}^l + \mathcal{L}_{mix}^u$;

`// Replay-guided supervision`

$\mathcal{L}_{RQG} \leftarrow 0$;

if $|\mathcal{Q}| \geq 3$ **then**

Randomly sample two distinct items $(x^{Q_1}, \hat{y}^{Q_1}), (x^{Q_2}, \hat{y}^{Q_2})$ from $\mathcal{Q}_t$;

$\tilde{x}^{Q1} \leftarrow M \odot x_a^u + (1-M) \odot x^{Q_1}$, $\hat{y}_{mix}^{Q1} \leftarrow M \odot \hat{y}_a^u + (1-M) \odot \hat{y}^{Q_1}$;

$\hat{x}_{mix}^{Q2} \leftarrow M \odot x^{Q_2} + (1-M) \odot x_b^u$, $\hat{y}_{mix}^{Q2} \leftarrow M \odot \hat{y}^{Q_2} + (1-M) \odot \hat{y}_b^u$;

$\mathcal{L}_{RQG} \leftarrow \mathcal{L}_{mix}^{Q1} + \mathcal{L}_{mix}^{Q2}$;

`// feature consistency losses introduced in Sec. 3.3`

$\mathcal{L}_{Feat} = \mathcal{L}_{glb} + \mathcal{L}_{loc}$;

`// Total loss`

$\mathcal{L} \leftarrow \mathcal{L}_{LG} + \mu \times \mathcal{L}_{RQG} + \varphi \times \mathcal{L}_{Feat}$;

Update student parameters:$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$;

Update teachers: $\theta^{T_g} \leftarrow \alpha_g \theta^{T_g} + (1-\alpha_g)\theta$, $\theta^{T_l} \leftarrow \alpha_l \theta^{T_l} + (1-\alpha_l)\theta$;

D More Experiments

D.1 Results on BraTS19 Dataset

On the BraTS19 dataset, we conduct experiments using 10% and 20% labeled data and compare our method with nine SoTA methods: UA-MT [37], URPC [21], DTC [20], BCP [2], AC-MT [35], MLRPL [26], TraCoCo [17], [42], and [12]. The results are reported in Table 6.

On the BraTS19 dataset, our method consistently achieves the best performance under both 10% and 20% labeled settings. With 10% labeled volumes, we obtain 87.16 Dice and 78.29 Jaccard, while reducing 95HD and ASD to 7.52 and 1.34, respectively. When increasing the labeled set to 50 volumes (20% labeled), our proposed approach achieves a Dice score of 88.97 and a Jaccard score of 80.78, outperforming the second-best method by 2.28 and 3.01, respectively, also with the lowest 95HD (6.95) and ASD (1.14). These results demonstrate that our approach yields more stable and finer-grained structural segmentation performance.

Table 6: Quantitative comparison with SoTA methods on BraTS19 dataset using two labeled data settings (10% and 20%).

Method	Volumes used in **BraTS19**		Metrics			
	Labeled	Unlabeled	Dice (%)↑	Jaccard (%)↑	95HD↓	ASD↓
3D-UNet(SupOnly)	25 (10%)	0	78.23	66.93	21.26	6.12
3D-UNet(SupOnly)	50 (20%)	0	80.16	71.55	22.68	3.43
UA-MT [37]	25 (10%)	225 (90%)	82.46	72.85	21.26	2.68
URPC [21]			82.03	71.03	17.36	4.87
DTC [20]			83.77	74.34	15.64	2.73
BCP [2]			83.15	73.14	12.81	3.57
AC-MT [35]			83.77	73.96	11.35	1.93
MLRPL [26]			84.29	74.74	9.57	2.55
TraCoCo [17]			85.71	76.39	9.20	2.27
Zhou et al. [42]			85.52	76.82	8.90	2.59
Kumari et al. [12]			85.71	76.35	8.15	1.83
Ours			**87.16**	**78.29**	**7.52**	**1.34**
UA-MT [37]	50 (20%)	200 (80%)	83.12	73.01	9.87	2.30
URPC [21]			82.93	72.57	15.93	4.19
DTC [20]			84.43	73.56	14.77	2.34
BCP [2]			82.71	72.72	9.99	1.86
AC-MT [35]			84.63	74.39	9.50	2.11
MLRPL [26]			85.47	76.32	7.76	2.00
TraCoCo [17]			86.69	77.69	8.04	1.93
Zhou et al. [42]			86.56	77.77	7.72	2.48
Kumari et al. [12]			86.65	77.72	7.32	1.70
Ours			**88.97**	**80.78**	**6.95**	**1.14**

D.2 Extreme Low-Labeled Setting

To further evaluate the robustness of the proposed method under more challenging scenarios, an additional experiment was conducted on the LA dataset with an even smaller amount of labeled data. Specifically, we conducted an experiment using only two labeled samples (2.5%), with results shown in Table 7. Aside from a reduction in the number of labeled data, we followed the same experimental settings without making any other adjustments.

The experimental results demonstrate that our method still achieves competitive performance with only 2 labeled samples. Specifically, when using 2 labeled samples, we achieve a Dice score of 90.26 and a Jaccard score of 82.32, with 95HD of 6.02 and ASD of 1.72. More importantly, our method outperforms the vast majority of SoTA methods trained with 4 labeled samples across these four metrics, and even surpasses some methods trained with 8 labeled samples. This result further demonstrates the effectiveness of our approach.

D.3 Ablation Studies for Dual-EMA Teachers

Effectiveness of Dual-EMA Teachers Framework:

In the main text, we have already presented the performance of the short-term local teacher (SLT) and long-term global teachers (LGT) when the student model achieves its best results, using 20% of the labeled data for training on the PROMISE12 dataset. We will now present the results obtained under the same settings using 10% of the labeled data for training on the LA dataset, as shown in Table 8. As can be seen, the results are consistent with those observed on the PROMISE12 dataset: LGT performs better on the Dice and Jaccard metrics, while SLT performs better on the boundary metrics 95HD and ASD, further illustrating the semantic complementarity of the dual-EMA framework.

Table 8: Performance comparison of the three models in the proposed framework on LA dataset using 10% labeled data.

Model	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
Long-term Global Teacher	91.97	85.18	4.89	1.52
Short-term Local Teacher	91.91	85.09	4.76	1.47
Student	**91.98**	**85.20**	**4.74**	**1.45**

Effect of the Dual-EMA Teachers Update Coefficient: To further examine the sensitivity to the EMA weighting coefficients, we perform a grid search over LGT's coefficient α_g and SLT's

Table 7: Quantitative comparison of proposed method (2 labeled) vs. SoTA methods (4 labeled and 8 labeled) on LA dataset.

Method	Volumes used in **LA**		Metrics			
	Labeled	Unlabeled	Dice (%)↑	Jaccard (%)↑	95HD↓	ASD↓
VNet(SupOnly)	4 (5%)	0	52.55	39.60	47.05	9.87
VNet(SupOnly)	8 (10%)	0	82.74	71.72	13.35	3.26
Ours	2 (2.5%)	78 (97.5%)	**90.26**	**82.32**	**6.02**	**1.72**
UA-MT [37]	4 (5%)	76 (95%)	82.26	70.98	13.71	3.82
DTC [20]			81.25	69.33	14.90	3.99
SS-Net [32]			86.33	76.15	9.97	2.31
MCF [30]			86.52	77.43	9.12	2.40
BCP [2]			88.02	78.72	7.90	2.15
CML [31]			87.63	—	8.92	2.23
AD-MT [41]			89.63	81.28	6.56	1.85
GapMatch [11]			88.30	—	—	1.77
UCSeg [10]			90.36	82.48	6.35	1.76
[12]			89.55	81.18	5.48	1.99
Ours			**90.90**	**83.37**	**4.91**	**1.73**
UA-MT [37]	8 (10%)	72 (90%)	86.28	76.11	18.71	4.63
DTC [20]			87.51	78.17	8.23	2.36
SS-Net [32]			88.55	79.62	7.49	1.90
MCF [30]			88.71	80.41	6.32	1.90
BCP [2]			89.62	81.31	6.81	1.76
CML [31]			90.36	—	6.06	1.68
AD-MT [41]			90.55	82.79	5.81	1.70
GapMatch [11]			91.00	—	—	1.46
UCSeg [10]			90.90	83.41	6.50	1.62
[12]			91.07	83.67	4.96	1.65
Ours			**91.98**	**85.20**	**4.74**	**1.45**

coefficient α_l on PROMISE12 dataset using 20% labeled data. The results are visualized in Fig. 7. The results indicate that the combination of $\alpha_g = 0.99$ and $\alpha_l = 0.9$ achieves the best overall performance, achieving an optimal trade-off between segmentation accuracy and the stability of teacher predictions. Specifically, the larger α_g fully exploits EMA's smoothing effect, providing stable global semantic guidance, while $\alpha_l = 0.9$ maintains a moderate responsiveness to recent updates, which helps capture fine-grained shapes and boundary details. However, using a smaller update coefficient for the SLT (e.g. $\alpha_l = 0.8$) makes the EMA teacher overly sensitive to short-term noise and random fluctuations, undermining the fundamental stabilization effect of EMA. This leads to more fluctuating teacher predictions and lower-quality pseudo labels, ultimately degrading overall performance. Additionally, we conduct an experiment using two EMA teachers with the same update coefficient. As expected, this setting does not improve performance, further confirming that the effectiveness of our approach stems from the semantic complementarity of dual teachers rather than simply increasing the number of teachers.

D.4 Ablation Studies for Feature-level Global–Local Consistency Loss

To further validate the effectiveness of feature-level consistency supervision, we conduct a more detailed ablation study: (1) w/o $\mathcal{L}_{glb}$: removing $\mathcal{L}_{glb}$; (2) w/o $\mathcal{L}_{loc}$: removing $\mathcal{L}_{loc}$; and (3) the complete method, which includes both global semantic and local relation feature-level consistency constraints. The results are illustrated in Table 9.

The results show that removing either consistency loss degrades performance, indicating that both global and local consistencies are necessary. Removing the global semantic consistency loss $\mathcal{L}_{glb}$ degrades region-overlap performance; on the PROMISE12 dataset, the Dice score drops to 85.61 and the Jaccard score decreases to 74.11. In contrast, removing the local relation consistency loss $\mathcal{L}_{loc}$ mainly harms boundary accuracy, leading to larger 95HD/ASD, which suggests weaker delineation of local details. With both enabled, the full model achieves the best results across all metrics.

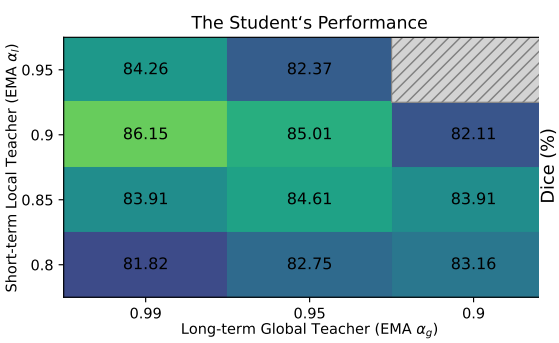

Figure 7: The results of grid search over EMA update coefficients α_g and α_l on PROMISE12 dataset using 20% labeled data.

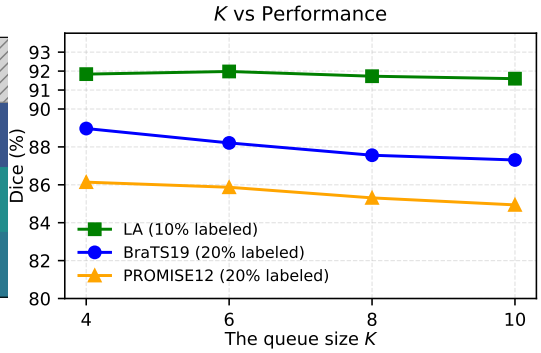

Figure 8: Performance of the model with varying replay queue sizes K across three datasets.

Table 9: Ablation studies for feature-level global-local consistency loss on LA dataset and PROMISE12 dataset.

Feature-level Loss	LA dataset (10% labeled)				PROMISE12 dataset (20% labeled)			
	Dice(%)↑	Jaccard↑	95HD↓	ASD↓	Dice(%)↑	Jaccard↑	95HD↓	ASD↓
w/o $\mathcal{L}_{glb}$	91.59	84.76	5.03	1.52	85.61	74.11	2.76	1.09
w/o $\mathcal{L}_{loc}$	91.72	84.93	5.17	1.58	85.88	74.79	2.92	1.16
Full Method	**91.98**	**85.20**	**4.74**	**1.45**	**86.15**	**75.80**	**2.42**	**0.95**

D.5 Selection of Replay Queue's Size

To analyze the impact of the replay queue size K, we conducted an ablation study with $K \in \{4, 6, 8, 10\}$ while keeping all other settings constant. As illustrated in Fig. 8, experimental results show that as K increases from smaller values, the model's performance first improves before declining: small K values struggle to provide sufficiently diverse high-confidence pseudo-supervision signals, while moderate K values strike a good balance between sample diversity and pseudo-label reliability, enabling the model to better learn from the teachers' complementary guidance. In contrast, an overly large K introduces more "outdated" pseudo-supervision signals, causing mismatch between historical pseudo labels and the current model distribution, undermining stability and ultimately degrading performance. Therefore, after jointly considering overall performance and stability, we set $K = 6$ for the LA dataset and $K = 4$ for PROMISE12 and BraTS19.

E Computational cost and resource efficiency

It is worth noting that introducing a second EMA teacher does not double the GPU memory consumption. This is because the teacher models are only used to generate pseudo labels and feature-level supervision, and their forward passes are performed under $no_grad()$. Therefore, no activation maps need to be stored for backpropagation, and no additional gradient memory is introduced. Specifically, under the same training setting on the LA dataset, the single-teacher baseline consumes 3814.36 MiB of GPU memory, while our dual-teacher framework consumes 4190.24 MiB, corresponding to only a 9.8% increase. Moreover, since only the student model is retained during inference, the proposed framework introduces no additional test-time computational or memory overhead.

F Theoretical Analysis

In this section, we provide a theoretical analysis of the proposed method. Accordingly, our theoretical analysis focuses on two aspects: (i) EMA performs temporal low-pass filtering on the parameter trajectory and different EMA coefficients preserve different frequency components; (ii) dual-EMA teacher prediction fusion can reduce pseudo-label noise;

F.1 Dual-EMA as Temporal Frequency-Selective Filtering

Let $\{\theta_S^{(t)}\}_{t=0}^{\infty}$ denote the parameter trajectory of the student model during training. An EMA teacher with smoothing coefficient α is updated by:

$$\theta_T^{(t)} = \alpha\theta_T^{(t-1)} + (1-\alpha)\theta_S^{(t)}. \tag{19}$$

This recursion can be viewed as a first-order IIR low-pass filter along the training-iteration dimension. Its transfer function is:

$$H_\alpha(z) = \frac{1-\alpha}{1-\alpha z^{-1}}. \tag{20}$$

Evaluating it on the unit circle $z = e^{j\omega}$ gives the magnitude response:

$$|H_\alpha(e^{j\omega})| = \frac{1-\alpha}{\sqrt{1+\alpha^2-2\alpha\cos\omega}}, \quad \omega \in [0,\pi]. \tag{21}$$

Proposition F.1 (Monotonicity of EMA frequency response)**.** *For any non-zero temporal frequency $\omega \in (0,\pi]$, the magnitude response $|H_\alpha(e^{j\omega})|$ is monotonically decreasing with respect to α.*

Proof. Taking the logarithm of the magnitude response, we have

$$\log|H_\alpha(e^{j\omega})| = \log(1-\alpha) - \frac{1}{2}\log(1+\alpha^2-2\alpha\cos\omega). \tag{22}$$

Differentiating it with respect to α gives

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log|H_\alpha(e^{j\omega})| &= -\frac{1}{1-\alpha} - \frac{\alpha-\cos\omega}{1+\alpha^2-2\alpha\cos\omega} \\ &= -\frac{(1+\alpha)(1-\cos\omega)}{(1-\alpha)(1+\alpha^2-2\alpha\cos\omega)}.\end{aligned} \tag{23}$$

Since $\alpha \in (0,1)$ and $\omega \in (0,\pi]$, we have $1-\cos\omega > 0$ and $1+\alpha^2-2\alpha\cos\omega > 0$. Therefore,

$$\frac{\partial}{\partial\alpha}\log|H_\alpha(e^{j\omega})| < 0. \tag{24}$$

Thus, $|H_\alpha(e^{j\omega})|$ decreases as α increases, which completes the proof. □

Corollary F.1 (Temporal complementarity of dual-EMA teachers)**.** *Given two EMA teachers with $\alpha_g > \alpha_l$, the long-term teacher with α_g suppresses non-zero temporal-frequency components more strongly, while the short-term teacher with α_l preserves more rapidly changing parameter variations.*

Proof. By Proposition F.1, for any $\omega \in (0,\pi]$, $|H_\alpha(e^{j\omega})|$ is monotonically decreasing with respect to α. Therefore, if $\alpha_g > \alpha_l$, then

$$|H_{\alpha_g}(e^{j\omega})| < |H_{\alpha_l}(e^{j\omega})|. \tag{25}$$

This means that the teacher with the larger EMA coefficient is more strongly smoothed, whereas the teacher with the smaller EMA coefficient better tracks short-term parameter variations. □

Discussion. The above analysis shows that different EMA coefficients induce different temporal filtering behaviors on the student parameter trajectory. This theoretically motivates our dual-EMA design: the long-term teacher with a larger α captures stable long-term trends, while the short-term teacher with a smaller α preserves more short-term variations. In the main text, we further empirically verify their semantic complementarity through feature-map FFT and boundary-confidence analyses.

F.2 Reliability of Dual-EMA Pseudo Labels

We first analyze why the proposed dual-EMA teachers can provide more reliable pseudo labels. The key intuition is that the long-term and short-term teachers are generated by different EMA coefficients and thus tend to have complementary prediction behaviors. When their prediction errors are not perfectly correlated, averaging their predictions can reduce pseudo-label noise.

Definition F.1 (Ideal prediction and teacher errors)**.** For an unlabeled sample x, let $p^*(\cdot|x)$ denote the ideal predictive distribution associated with the underlying ground-truth segmentation. The predictions of the two EMA teachers are written as

$$p_g(\cdot|x) = p^*(\cdot|x) + \epsilon_g(x), \qquad p_l(\cdot|x) = p^*(\cdot|x) + \epsilon_l(x), \tag{26}$$

where $\epsilon_g(x)$ and $\epsilon_l(x)$ represent the prediction errors of the long-term global teacher and short-term local teacher, respectively.

The fused dual-EMA prediction is defined as:

$$p_f(\cdot|x) = \frac{1}{2}p_g(\cdot|x) + \frac{1}{2}p_l(\cdot|x). \tag{27}$$

The corresponding hard pseudo label is $\hat{y}_f(x) = \arg\max p_f(\cdot|x)$.

Proposition F.2 (Error reduction effect of dual-EMA fusion)**.** *The fused prediction satisfies*

$$\mathbb{E}\|p_f - p^*\|_2^2 = \frac{1}{4}\left(\mathbb{E}\|\epsilon_g\|_2^2 + \mathbb{E}\|\epsilon_l\|_2^2 + 2\mathbb{E}\langle\epsilon_g, \epsilon_l\rangle\right). \tag{28}$$

Moreover,

$$\mathbb{E}\|p_f - p^*\|_2^2 = \frac{1}{2}\left(\mathbb{E}\|\epsilon_g\|_2^2 + \mathbb{E}\|\epsilon_l\|_2^2\right) - \frac{1}{4}\mathbb{E}\|\epsilon_g - \epsilon_l\|_2^2. \tag{29}$$

Therefore, the fused prediction has no larger expected squared error than the average error of the two individual teachers, and the improvement becomes larger when the two teacher errors are more different.

Proof. By definition,

$$\begin{aligned} p_f - p^* &= \frac{1}{2}p_g + \frac{1}{2}p_l - p^* \\ &= \frac{1}{2}(p_g - p^*) + \frac{1}{2}(p_l - p^*) \\ &= \frac{1}{2}(\epsilon_g + \epsilon_l). \end{aligned} \tag{30}$$

Thus,

$$\begin{aligned} \mathbb{E}\|p_f - p^*\|_2^2 &= \mathbb{E}\left\|\frac{1}{2}(\epsilon_g + \epsilon_l)\right\|_2^2 \\ &= \frac{1}{4}\mathbb{E}\left[\|\epsilon_g\|_2^2 + \|\epsilon_l\|_2^2 + 2\langle\epsilon_g, \epsilon_l\rangle\right]. \end{aligned} \tag{31}$$

Using the identity

$$2\langle\epsilon_g, \epsilon_l\rangle = \|\epsilon_g\|_2^2 + \|\epsilon_l\|_2^2 - \|\epsilon_g - \epsilon_l\|_2^2, \tag{32}$$

we obtain

$$\mathbb{E}\|p_f - p^*\|_2^2 = \frac{1}{2}\left(\mathbb{E}\|\epsilon_g\|_2^2 + \mathbb{E}\|\epsilon_l\|_2^2\right) - \frac{1}{4}\mathbb{E}\|\epsilon_g - \epsilon_l\|_2^2. \tag{33}$$

Since $\mathbb{E}\|\epsilon_g - \epsilon_l\|_2^2 \geq 0$, the fused prediction has no larger expected squared error than the average error of the two individual teachers. □

Corollary F.2 (Equal-variance case)**.** *Assume that the two teachers have comparable error magnitudes:*

$$\mathbb{E}\|\epsilon_g\|_2^2 = \mathbb{E}\|\epsilon_l\|_2^2 = \sigma^2, \tag{34}$$

and define the error correlation coefficient as $\rho = \frac{\mathbb{E}\langle\epsilon_g, \epsilon_l\rangle}{\sigma^2}$*. Then,*

$$\mathbb{E}\|p_f - p^*\|_2^2 = \frac{1+\rho}{2}\sigma^2. \tag{35}$$

If $\rho < 1$*, the fused prediction has a smaller expected error than either individual teacher.*

Proof. Substituting $\mathbb{E}\|\epsilon_g\|_2^2 = \mathbb{E}\|\epsilon_l\|_2^2 = \sigma^2$ and $\mathbb{E}\langle\epsilon_g, \epsilon_l\rangle = \rho\sigma^2$ into Proposition F.2 gives

$$\begin{aligned} \mathbb{E}\|p_f - p^*\|_2^2 &= \frac{1}{4}\left(\sigma^2 + \sigma^2 + 2\rho\sigma^2\right) \\ &= \frac{1+\rho}{2}\sigma^2. \end{aligned} \tag{36}$$

When $\rho < 1$, we have $\frac{1+\rho}{2}\sigma^2 < \sigma^2$, which completes the proof. □

Discussion. The above analysis indicates that dual-EMA prediction fusion can alleviate pseudo-label noise by exploiting the complementary behaviors of the two teachers. Specifically, the long-term teacher and the short-term teacher are generated with different EMA coefficients, which makes them preserve different temporal dynamics and produce non-identical prediction errors. When their errors are not perfectly correlated, averaging their predictions reduces the expected squared prediction error compared with relying on a single teacher. Therefore, dual-EMA fusion provides more reliable pseudo labels for unlabeled samples and improves the quality of student supervision.

G Limitations and Broader Impacts

One limitation of our method is that our framework benefits from the complementarity between the long-term and short-term teachers. Specifically, the short-term teacher can provide more informative guidance for local structures such as boundaries, while the long-term teacher offers more stable global semantic supervision. Although our feature map FFT and boundary confidence analyses support this design, the complementarity may be weakened in extremely noisy images or cases with blurred lesion boundaries and severe artifacts. Under such conditions, the short-term teacher may provide less reliable local guidance. Future work will investigate artifact-robust adaptation to further enhance the robustness of dual-EMA framework.

Moreover, our work holds significant potential for positive societal impact by drastically reducing the annotation burden required for medical image analysis. By effectively leveraging scarce labeled data alongside abundant unlabeled scans, this approach can accelerate the development and deployment of computer-aided diagnosis (CAD) systems, particularly in resource-constrained healthcare settings where expert annotations are prohibitively expensive. However, there are potential risks and negative impacts to consider. If the unlabeled dataset contains inherent demographic, hardware, or institutional biases, the self-training process risks amplifying these disparities, leading to unequal model performance across different patient populations. To mitigate these risks, future clinical deployments must ensure diverse, representative data collection and incorporate human-in-the-loop verification for critical diagnostic decisions. In future research, we plan to further improve the domain adaptation capability of the proposed method.