CSGE: Break the SSL Bottleneck in Medical Image Segmentation via Collaborative Semantic-Geometric Experts

Yufei Wan[1*], Hangbei Cheng[1*], Danchen Cui[1], Junran Li[1], Jiahui Jiang[1], Junxin Chen[2], Mingqiang Wei[1], Xueyu Liu[1(✉)], and Yongfei Wu[1(✉)]

[1] Taiyuan University of Technology, Taiyuan, Shanxi, 030024, China
liuxueyu@tyut.edu.cn, wuyongfei@tyut.edu.cn
[2] Dalian University of Technology, Dalian, Liaoning, 116024, China
*These authors contributed equally to this work.

Abstract. Semi-supervised learning for medical image segmentation, particularly within the prevalent Teacher-Student architecture, is frequently hindered by recursive error accumulation and the inherent rigidity of hard-threshold pseudo-labeling. These limitations lead to three critical bottlenecks: semantic-geometric misalignment, chronic confirmation bias, and significant loss of boundary information. To address these, we propose CSGE, a novel paradigm shifting from pseudo-label filtering to collaborative semantic-geometric refinement. CSGE synergistically leverages multimodal medical knowledge and geometric priors. By exploiting semantic cues from the textual modality, BiomedCLIP provides robust semantic anchoring, while SAM serves as a geometric refinement expert. We introduce a Conflict-aware Iterative Co-evolution module that detects discrepancies via XOR operations to drive iterative pseudo-label refinement. Furthermore, we replace hard thresholds with a spatial trust-aware soft-weighted consistency loss to maintain gradient flow within uncertain regions. Evaluated across three modalities (1%–30% labels), CSGE outperforms state-of-the-art methods, notably achieving a 4.9% average Dice improvement under the extreme 1% setting to demonstrate its robustness. Our code is available at `https://github.com/yufei050201/CSGE`.

Keywords: Semi-supervised Learning · Medical Image Segmentation · Multi-expert Collaboration.

1 Introduction

Medical image segmentation (MIS) is a key technology in modern clinical workflows, offering important quantitative support for disease diagnosis and treatment planning. However, its development has long been limited by high annotation costs and inevitable label noise. Semi-supervised learning (SSL) alleviates the scarcity of labeled data via large unlabeled datasets [27,3]. Existing SSL paradigms, particularly the Teacher-Student architecture [18], rely on

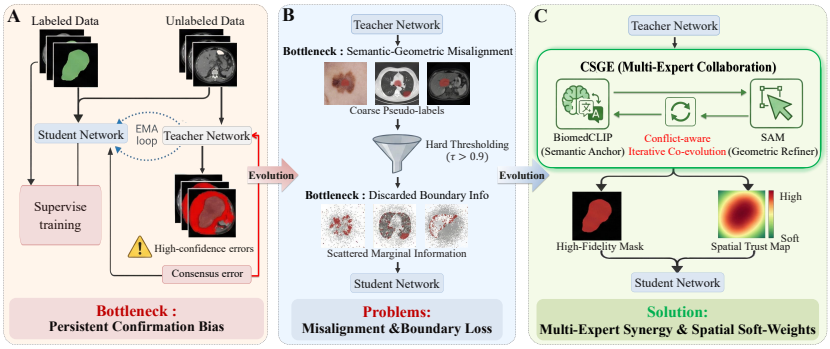

Fig. 1: Comparison of semi-supervised MIS paradigms. A-B reveal bottlenecks in traditional paradigms; C showcases our model's targeted improvements.

confidence-based pseudo-label filtering mechanisms [13]. Though functional, this strategy assumes a single model can simultaneously achieve reliable semantic localization and accurate geometric delineation, an assumption unfeasible in complex clinical scenarios, thus limiting the improvement of pseudo-label quality.

This traditional paradigm faces **three core bottlenecks** limiting performance enhancement. First, the inherent semantic-geometric misalignment, termed the "Scissors Gap"; CNNs and Transformers capture semantic information via large receptive fields [16,19,6], but their boundary localization is often impaired by downsampling and resolution constraints [10], causing high internal lesion confidence but low edge confidence and thus "shrinking" or blurred pseudo-labels. Second, SSL's recursive mechanism tends to induce confirmation bias [20] (Fig. 1(A)): in the classic Mean Teacher framework [18], the teacher's erroneous high-confidence predictions are transferred to the student and reinforced via EMA. Third, hard-threshold filtering causes information loss (Fig. 1(B)): methods like Polite Teacher and SSDT discard pixels with slightly sub-threshold confidence [13], yet key boundary details often lie in these uncertain regions [22,23], hindering fine-grained topological learning.

Despite these bottlenecks, foundation models like BiomedCLIP [25] and SAM [11] offer transformative opportunities to transcend traditional SSL limitations. While pioneering efforts like SemiSAM [26] and MedCLIP-SAMv2 [12] demonstrate potential in low-annotation scenarios, they often lack a dynamic semantic-geometric synergy, leaving issues like confirmation bias and boundary information loss unresolved. To bridge this gap, we propose CSGE (Fig. 1(C)), a novel paradigm that shifts from passive pseudo-label filtering to active collaborative refinement. Our **main contributions** are summarized as follows.

1. We introduce CSGE, a teacher-student evolution framework that synergizes semantic and geometric experts. It addresses the semantic-geometric misalignment bottleneck by re-framing segmentation as a multi-stage collaborative process of global localization, local refinement, and boundary alignment.

2. We introduce the Conflict-aware Iterative Co-evolution module, corrects boundaries via conflict-aware prompt resampling, and mitigates confirmation bias through iterative geometric feedback.
3. We introduce a Spatial Trust Awareness optimization that abandons hard thresholds to preserve gradient flow in uncertain areas, guiding the network to optimal boundaries while enhancing noise robustness.

2 Methodology

2.1 Overview of CSGE

The architecture of our CSGE framework (Fig. 2) synergizes domain-specific semantic anchoring with universal geometric refinement to transcend the limitations of internal knowledge distillation. In this paradigm, the student network performs end-to-end learning while the teacher provides initial pseudo-labels via EMA updates. To rectify the inherent semantic-geometric misalignment, we employ BiomedCLIP to provide robust category-level guidance, filtering background noise that often misleads traditional teachers and integrate SAM to refine boundaries through iterative prompt sampling, compensating for BiomedCLIP's lack of fine-grained structural awareness. Finally, a spatially-aware soft weighting strategy constructs a dynamic weight map $W(x, y)$, bridging the gap between heterogeneous expert outputs by fusing global consistency with local boundary enhancement to adaptively optimize student training.

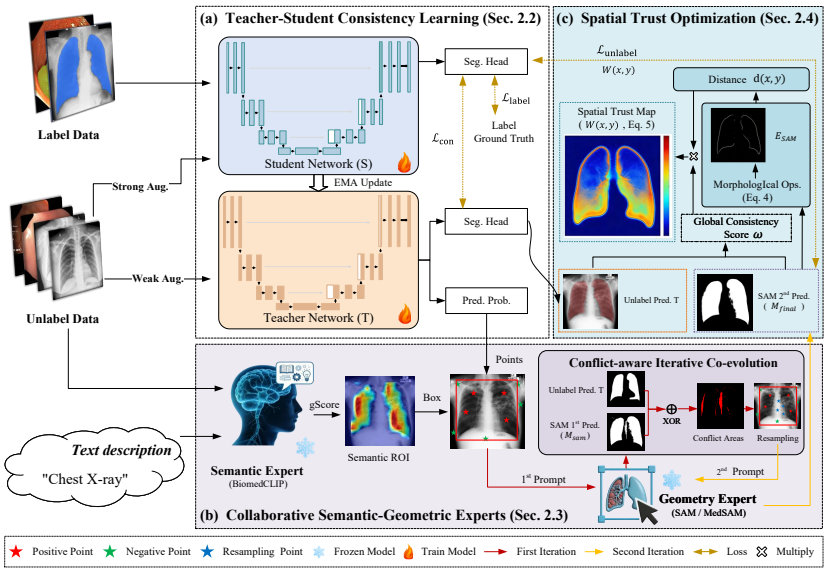

Fig. 2: Schematic diagram of the CSGE framework.

The framework comprises three progressive and collaborative core stages: (1) Teacher-Student Consistency Learning; (2) Collaborative Semantic-Geometric Experts; (3) Spatial Trust Optimization. These stages enable the evolution of pseudo-labels from simple filtering to high-quality generation.

2.2 Teacher-Student Consistency Learning

Let $\mathcal{D}_l = \{(I_i, Y_i)\}_{i=1}^{N}$ be the set of medical images labeled and $\mathcal{D}_u = \{I_j\}_{j=1}^{M}$ be the set without label with $M \gg N$, reflecting the scarce medical annotations. The framework uses a teacher-student architecture, where teacher parameters θ_t are updated by student parameters θ_s via EMA. For labeled data (I_i, Y_i) ($I_i \in \mathbb{R}^{H\times W\times C}$, $Y_i \in \mathbb{R}^{H\times W\times 1}$), the supervised loss is the equal-weight combination of Dice similarity coefficient and Binary Cross-Entropy:

$$\mathcal{L}_{\text{label}} = \frac{1}{2} \cdot \text{Dice}(f_{\theta_s}(I_i), Y_i) + \frac{1}{2} \cdot \text{BCE}(f_{\theta_s}(I_i), Y_i), \tag{1}$$

where $P_s = f_{\theta_s}(I_i)_{x,y} \in [0, 1]$ is the student network's soft prediction probability at pixel (x, y) for image I_i. For unlabeled data I_j, L2 consistency loss is used to constrain the alignment between the student and teacher network outputs, and the loss expression is:

$$\mathcal{L}_{\text{con}} = \frac{1}{H \times W} \| f_{\theta_s}(I_j) - f_{\theta_t}(I_j) \|_2^2, \tag{2}$$

where $P_t = f_{\theta_t}(I_j)$ is the soft prediction result of the teacher network for unlabeled image I_j, and the normalization term $1/(H \times W)$ ensures the loss scale is independent of image resolution.

2.3 Collaborative Semantic-Geometric Experts

This section designs a pseudo-label refinement strategy combining multi-modal semantic guidance and SAM geometric refinement; decoupling semantic localization and geometric boundaries improves pseudo-label quality.

Specifically, we use the biomedical vision-language pre-trained model BiomedCLIP to serve as a multi-modal semantic alignment expert. Let f_I be the image feature embedding of input medical image I_j from BiomedCLIP, and f_T the text feature embedding for task-related medical text (e.g., "Chest X-ray"). Their semantic similarity activates relevant regions. To eliminate training overhead, gScoreCAM [4] maps are pre-computed offline to constrain refinement to the target ROI. This static macroscopic guidance suffices for consistent anatomy, while SAM and subsequent iterations handle dynamic refinements.

Within these regions, SAM acts as a geometric expert. Specifically, global point prompts are generated by thresholding the teacher's soft prediction P_t: high-confidence pixels are sampled as positive prompts (foreground), and low-confidence ones as negative prompts (background), yielding the initial geometric candidate mask M_{sam}. To resolve potential discrepancies between semantic

guidance and geometric perception, we initialize our Conflict-Aware Iterative Co-Evolution Module by detecting inconsistent regions via an XOR operation $\text{Area}_{\text{diff}} = \mathbb{I}(P_t) \oplus M_{\text{sam}}$, where $\mathbb{I}(\cdot)$ denotes the binarization indicator function and $\oplus$ is the logical XOR. The resulting $\text{Area}_{\text{diff}}$ highlights pixels with conflicting predictions, isolating the core regions requiring pseudo-label refinement.

To resolve this inconsistency and synergistically fuse semantic and geometric information, we employ a single-step prompt sampling strategy. Specifically, we sample key prompts from the conflict regions $\text{Area}_{\text{diff}}$ and feed them back into SAM for a one-time structural refinement. This targeted update to the geometric candidate mask effectively breaks the confirmation bias loop, yielding the high-precision pseudo-label M_{final}.

2.4 Spatial Trust Optimization

To fully utilize valid information of pseudo-labels and suppress interference from unreliable regions, we design a Spatial Confidence-Aware Optimization to achieve spatially adaptive weighting in model training.

First, to quantify the reliability of the final pseudo-label M_{final}, we evaluate its agreement with the teacher's soft prediction P_t. The global trust weight is directly derived as $w = \text{Sigmoid}(\text{Dice}(P_t, M_{\text{final}}))$. To accurately extract the boundary regions for spatially adaptive weighting, we compute the contour mask E_{SAM} from M_{final} via morphological operations:

$$E_{\text{SAM}} = \text{Dilation}(M_{\text{final}}) - \text{Erosion}(M_{\text{final}}), \tag{3}$$

where the dilation and erosion utilize a standard 3×3 structuring element. Building upon the global trust weight w, we formulate the spatial trust map $W(x, y)$ to assign higher optimization weights near these boundaries:

$$W(x, y) = w \cdot \exp\left(-\frac{d(x, y)^2}{2\sigma^2}\right), \tag{4}$$

where $d(x, y)$ denotes the Euclidean distance from pixel (x, y) to the nearest pixel in E_{SAM}. A small d prioritizes the fine geometric contours derived from SAM, while a large d relies on the teacher model's semantics to maintain regional consistency. Based on the spatial trust map $W(x, y)$, we formulate a weighted unsupervised loss:

$$\mathcal{L}_{\text{unlabel}} = \sum_{x,y} W(x, y) \, \| f_{\theta_s}(I_j)(x, y) - M_{\text{final}}(x, y) \|_2^2 . \tag{5}$$

This objective adaptively scales pixel-wise gradients, directing the model's focus toward reliable boundary regions while suppressing noise to mitigate error accumulation. Finally, the training loss balances the supervisory contributions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{label}} + \lambda \mathcal{L}_{\text{unlabel}} + \mathcal{L}_{\text{con}}, \tag{6}$$

where λ weights the unsupervised loss. Strictly complementary, $\mathcal{L}_{\text{con}}$ aligns global semantics while $\mathcal{L}_{\text{unlabel}}$ refines local boundaries.

3 Experiments

3.1 Datasets and Implementation Details

Datasets: We evaluated CSGE (using a standard UNet backbone) on three fully de-identified, IRB-approved public medical datasets, strictly adhering to their ethical usage guidelines: Shenzhen Chest X-ray [8] (662 radiographs), Kvasir-SEG [9] (1000 colonoscopy images), and ISIC 2018 [5] (2500 dermoscopic images). We evaluate using Dice and IoU as they highly sensitize to the macroscopic overlap degradation caused by SSL bottlenecks, effectively validating CSGE's improvements in boundary alignment.
Implementation Details: All experiments were conducted on ten NVIDIA Tesla V100 GPUs, with labeled data randomly sampled at 1%, 5%, 10%, 20%, and 30% ratios. Models were trained for 100 epochs using the AdamW optimizer (lr = 1e-4, weight decay = 5e-5, batch size 8) and an EMA momentum $\eta = 0.999$. Final Dice and IoU scores are averaged over three independent runs with distinct random seeds.

3.2 Main Experimental Results

Comparison Results: Compared with SOTA semi-supervised methods, CSGE demonstrates two clear advantages (Tab. 1). Under extremely low annotation ratios, it achieves the largest performance gains across all datasets, for example, at 1% supervision, CSGE reached 0.915 Dice on Lung (+3.0% over the strongest competitor), 0.803 on Kvasir (+7.1%), and 0.825 on ISIC (+4.5%), indicating its semantic–geometric synergy effectively mitigates annotation scarcity. As the labeled ratio increases, although all methods improve, CSGE consistently maintains the best performance across all supervision levels without saturation degradation, which can be attributed to the proposed spatial trust mechanism that stabilizes boundary learning and preserves structural refinement.
Visualization Analysis: We visualize pseudo-label evolution in Fig. 3 to analyze model behavior. SAM refines coarse initial boundaries, while conflict-driven optimization further enhances mask integrity and smoothness. Compared with other methods (the last three columns of Fig. 3), CSGE precisely distinguishes boundaries between different tissues, reduces missed detections and over-segmentation, and generates sharper, more anatomically consistent contours at texture transition regions. These qualitative results align with quantitative metrics, validating our collaborative mechanism under limited supervision.
Extension Analysis: Our framework is architecture-agnostic and can be combined with different segmentation networks. We evaluated MedSAM [14] and SAM3 [2], which perform competitively on single-modality datasets. However, our method yields more consistent cross-dataset improvements, with average gains of 3.2% in Dice and 2.8% in IoU.In addition, for clinical deployment, BiomedCLIP and SAM are only activated during the training phase (peak 16GB, V100). Inference only uses a lightweight student network, taking <17 ms and ∼1.82 GB for 256×256 inputs, ensuring scalable deployment.

Table 1: Performance Comparison with 1%–30% Labeled Data

Method	Year/Source	1%		5%		10%		20%		30%	
		Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Lung Dataset											
Ours		**0.915**	**0.885**	**0.962**	**0.926**	**0.964**	**0.930**	**0.967**	**0.922**	**0.965**	**0.931**
U-Net [16]	MICCAI 2015	0.764	0.666	0.874	0.784	0.917	0.851	0.940	0.889	0.949	0.903
FixMatch [17]	NeurIPS 2020	0.812	0.752	0.908	0.842	0.930	0.872	0.944	0.894	0.950	0.904
UA-MT [24]	MICCAI 2019	0.782	0.799	0.933	0.879	0.926	0.868	0.936	0.885	0.940	0.891
SemiSAM [26]	MICCAI 2023	0.843	0.746	0.936	0.883	0.941	0.891	0.941	0.890	0.940	0.889
DyCON [1]	CVPR 2025	0.882	0.812	0.955	0.914	0.957	0.919	0.960	0.923	0.961	0.924
DuCiSC [21]	TMI 2025	0.882	0.818	0.956	0.918	0.958	0.921	0.958	0.921	0.958	0.921
SynFoC [15]	CVPR 2025	0.874	0.803	0.951	0.910	0.954	0.916	0.956	0.919	0.958	0.921
β-FFT [7]	CVPR 2025	0.885	0.820	0.957	0.917	0.959	0.921	0.961	0.923	0.962	0.925
Kvasir Dataset											
Ours		**0.803**	**0.763**	**0.862**	**0.835**	**0.904**	**0.846**	**0.932**	**0.874**	**0.945**	**0.900**
U-Net	MICCAI 2015	0.701	0.602	0.833	0.727	0.881	0.794	0.913	0.845	0.928	0.870
FixMatch	NeurIPS 2020	0.720	0.619	0.846	0.741	0.894	0.810	0.924	0.861	0.938	0.885
UA-MT	MICCAI 2019	0.714	0.613	0.839	0.735	0.887	0.802	0.918	0.855	0.934	0.880
SemiSAM	MICCAI 2023	0.727	0.626	0.854	0.752	0.900	0.819	0.928	0.870	0.941	0.892
DyCON	CVPR 2025	0.725	0.624	0.852	0.749	0.898	0.815	0.926	0.867	0.940	0.890
DuCiSC	TMI 2025	0.732	0.632	0.857	0.755	0.902	0.821	0.929	0.872	0.943	0.894
SynFoC	CVPR 2025	0.723	0.621	0.849	0.746	0.895	0.812	0.925	0.865	0.939	0.887
β-FFT	CVPR 2025	0.729	0.629	0.855	0.754	0.901	0.820	0.929	0.871	0.942	0.893
ISIC Dataset											
Ours		**0.825**	**0.705**	**0.833**	**0.720**	**0.887**	**0.797**	**0.918**	**0.846**	**0.934**	**0.873**
U-Net	MICCAI 2015	0.751	0.642	0.802	0.682	0.857	0.750	0.893	0.809	0.913	0.839
FixMatch	NeurIPS 2020	0.769	0.657	0.816	0.699	0.873	0.772	0.906	0.830	0.925	0.858
UA-MT	MICCAI 2019	0.761	0.653	0.811	0.694	0.867	0.764	0.901	0.822	0.920	0.852
SemiSAM	MICCAI 2023	0.776	0.668	0.825	0.710	0.882	0.785	0.913	0.840	0.932	0.868
DyCON	CVPR 2025	0.775	0.566	0.822	0.705	0.879	0.781	0.911	0.837	0.930	0.865
DuCiSC	TMI 2025	0.780	0.671	0.826	0.712	0.884	0.789	0.915	0.842	0.932	0.870
SynFoC	CVPR 2025	0.772	0.663	0.820	0.703	0.876	0.778	0.909	0.834	0.928	0.862
β-FFT	CVPR 2025	0.779	0.670	0.826	0.711	0.883	0.787	0.914	0.842	0.932	0.869

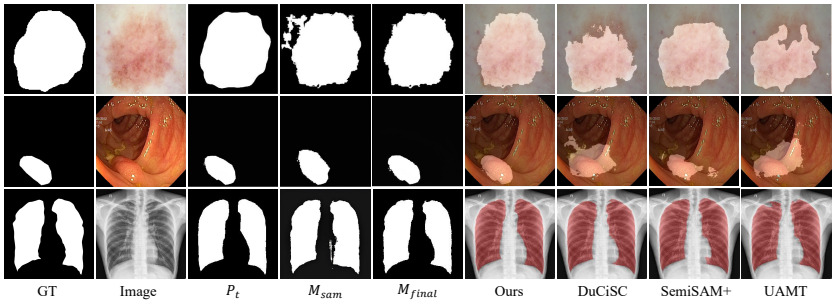

Fig. 3: Qualitative segmentation results over three medical imaging modalities using 1% labeled data.

3.3 Ablation Experiments

Parameter Selection: We systematically analyzed the key inherent parameters of CSGE (Tab. 2): the unsupervised loss weight λ (Eq. 6) and the spatial trust Gaussian kernel parameter σ (Eq. 4). Our model achieved optimal performance at $\lambda = 1.0$ and $\sigma = 0.5$ across three datasets, which aligns with our design by fully exploiting unsupervised supervision while avoiding interference from noisy pseudo-labels and eliminating the need for dataset-specific tuning. Experimental results demonstrate that our method maintains robust performance across a reasonable range of hyperparameter settings.

Table 2: Impact of Key Parameters (1% Labeled Data, the Lung Dataset).

No.	Unsupervised Loss Weight (λ)	Spatial Trust Gaussian Kernel (σ)	Dice (%)	IoU (%)
1	0.1	0.5	87.5	83.5
2	0.5	0.5	90.2	85.4
3	1.0	0.5	**91.5**	**88.5**
4	1.0	0.3	90.8	86.8
5	1.0	0.7	89.1	83.7

Ablation Experiments: To systematically evaluate CSGE, we first analyze isolated experts and then perform a cumulative ablation study to validate each component(Tab. 3). While the baseline model and isolated experts perform poorly, consistency learning (+teacher) enables stable pseudo-label initialization on unlabeled data. Building on this foundation, our semantic-geometric integration (+SAM/BiomedCLIP) combined with SAM and BiomedCLIP further boosts the Dice score to 86.3%. The proposed Conflict Iterative module (+Conflict Iterative) effectively resolves semantic–geometric discrepancies, bringing an additional 3.6% gain. Finally, the spatial trust maximally preserves boundary information (+Spatial Trust Map, Dice+1.6%). The overall 11.7% improvement over the baseline highlights the effectiveness of the proposed collaborative design.

Table 3: Ablation Experiment (1% Labeled Data, the Lung Dataset).

Configuration		Teacher	SAM	BiomedCLIP	Conflict Iterative	Spatial Trust Map	Dice (%)	IoU (%)
Only SAM		✗	✓	✗	✗	✗	80.1	70.8
Only BiomedCLIP		✗	✗	✓	✗	✗	67.9	58.2
Consistency Learning	+Model(Baseline)	✗	✗	✗	✗	✗	79.8	66.9
	+Teacher	✓	✗	✗	✗	✗	81.2	66.9
semantic-geometric	+SAM	✓	✓	✗	✗	✗	83.7	72.6
	+BiomedCLIP	✓	✓	✓	✗	✗	86.3	79.5
Conflict Iterative module	+ConflictIterative	✓	✓	✓	✓	✗	89.9	84.8
Spatial Trust Mechanism	+SpatialTrustMap	✓	✓	✓	✓	✓	91.5	88.5
Full Model		✓	✓	✓	✓	✓	**91.5**	**88.5**

4 Conclusion

In conclusion, we present CSGE, a Teacher–Student SSL framework that synergizes the semantic capabilities of BiomedCLIP with the geometric precision of SAM. By leveraging conflict-aware co-evolution and spatial reliability-aware optimization, CSGE mitigates pseudo-label noise. This ensures robust segmentation stability even under extremely limited annotations. Furthermore, the framework ensures lightweight inference, making it an efficient and scalable solution for clinical deployment. Ultimately, this work highlights the power of large-model-assisted training for lightweight networks, reducing annotation dependency from 10% to a mere 1%; however, optimizing their synergy remains an ongoing pursuit. Looking ahead, we plan to incorporate next-generation foundation models and advanced optimization techniques, and attempt to extend our framework to the 3D medical segmentation domain to tackle more complex and diverse medical imaging modalities.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62572339 and 61901292), the Natural Science Foundation of Shanxi Province (Grant No. 202303021211082), and the Key Research and Development Program of Shanxi Province (Grant No. 202402020101008).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Assefa, M., Naseer, M., Ganapathi, I.I., Ali, S.S., Seghier, M.L., Werghi, N.: Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30850–30860 (2025)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press (2006)
4. Chen, P., Li, Q., Biaz, S., Bui, T., Nguyen, A.: gscorecam: What objects is clip looking at? In: Proceedings of the Asian Conference on Computer Vision. pp. 1959–1975 (2022)
5. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
7. Hu, M., Yin, J., Ma, Z., Ma, J., Zhu, F., Wu, B., Wen, Y., Wu, M., Hu, C., Hu, B., et al.: beta-fft: Nonlinear interpolation and differentiated training strategies for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30839–30849 (2025)

8. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery **4**(6), 475 (2014)
9. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling. pp. 451–462. Springer (2019)
10. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ben Ayed, I.: Boundary loss for highly unbalanced segmentation. Medical Image Analysis **67**, 101851 (2021)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
12. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Medclip-samv2: Towards universal text-driven medical image segmentation. Medical Image Analysis p. 103749 (2025)
13. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
15. Ma, Q., Zhang, J., Li, Z., Qi, L., Yu, Q., Shi, Y.: Steady progress beats stagnation: Mutual aid of foundation and conventional models in mixed domain semi-supervised medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5175–5185 (2025)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
17. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
18. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks **145**, 90–106 (2022)
21. Wu, H., Wang, C., Cui, Z.: Dual cross-image semantic consistency with self-aware pseudo labeling for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2025)
22. Xiao, Z., Su, Y., Deng, Z., Zhang, W.: Efficient combination of cnn and transformer for dual-teacher uncertainty-guided semi-supervised medical image segmentation. Computer Methods and Programs in Biomedicine **226**, 107099 (2022)
23. Xu, M., Hu, X., Gupta, S., Abousamra, S., Chen, C.: Semi-supervised segmentation of histopathology images with noise-aware topological consistency. In: European Conference on Computer Vision. pp. 271–289. Springer (2024)

24. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 605–613. Springer (2019)
25. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
26. Zhang, Y., Yang, J., Liu, Y., Cheng, Y., Qi, Y.: Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3982–3986. IEEE (2024)
27. Zhu, X.: Semi-supervised learning literature survey. Tech. Rep. TR-1530, Computer Sciences, University of Wisconsin-Madison (2005), last modified on July 19, 2008