Hyperbolic Vision-Language Interaction for Semi-supervised Medical Image Segmentation

Qiuchi He, Shaocheng Jin, Rui Wang, Tianyang Xu, Xiao-Jun Wu, Tao Zhou(✉)

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, Jiangsu, China.
taozhou.ai@jiangnan.edu.cn

Abstract. Semi-supervised medical image segmentation alleviates the burden of costly expert annotation. Textual information provides complementary guidance to enhance visual feature learning under limited supervision. However, existing vision-language methods operate in Euclidean spaces, which fail to capture inherent tree-like anatomical hierarchies despite the benefits of textual cues, leading to geometric distortion. In this paper, we propose **HyperSemi**, a hyperbolic vision-language interaction network for semi-supervised medical image segmentation. Our framework is specifically designed to accommodate anatomical hierarchies through geometry-aware cross-modal learning. Specifically, we present a Hyperbolic Semantic Embedding (HSE) module, which projects both visual features and text embeddings into the Poincaré ball to enforce hierarchical semantic alignment, effectively mitigating geometric distortion. To enable geometry-consistent cross-modal interaction, we further propose a Geometry-aware Semantic Integration (GSI) module that leverages hyperbolic distance weighting and tangent-space aggregation to adaptively modulate encoder features. Finally, to mitigate confirmation bias introduced by pseudo-labels, we propose an Uncertainty-aware Region Mixing (URM) strategy that selectively identifies high-uncertainty regions for mixed supervision, improving model robustness. Extensive experiments on three benchmark datasets demonstrate that our model consistently outperforms state-of-the-art semi-supervised segmentation approaches. Our code is available at https://github.com/taozh2017/HyperSemi.

Keywords: Semi-supervised medical segmentation, HSE, GSI, URM

1 Introduction

Semi-supervised medical image segmentation [1,17,33] alleviates manual annotation burdens by leveraging extensive unlabeled data alongside limited labeled samples. Existing methods fall into two main categories: pseudo-labeling [18,31,16,2], and consistency regularization [30,4,24,20]. However, these approaches rely primarily on pixel-level visual cues, neglecting the rich contextual information embedded in anatomical concepts or textual descriptions.

To incorporate textual cues into visual representations, Vision-Language Models (VLMs) such as CLIP [14] have recently been introduced to medical imaging. For instance, semi-supervised frameworks like Text-SemiSeg [5] and VCLIPSeg [7] leverage textual cues to enhance segmentation predictions. However, these approaches largely constrain cross-modal interaction within a flat Euclidean space. This assumption is problematic for medical anatomy, which inherently exhibits complex and tree-like hierarchical structures [13]. Euclidean space lacks the capacity to embed such hierarchies without substantial distortion [12]. In contrast, hyperbolic geometry offers an intrinsic capacity to represent hierarchical data through its exponential volume expansion, making it mathematically ideal for embedding textual concepts with entailment relationships. Thus, how can we construct a vision-language framework that inherently accommodates the hierarchical nature of medical anatomy?

To this end, we propose HyperSemi, a novel semi-supervised medical image segmentation framework that intrinsically accommodates the hierarchical nature of medical anatomy via hyperbolic geometry. Specifically, we propose a Hyperbolic Semantic Embedding (HSE) module, which projects visual features and textual embeddings onto the Poincaré ball to perform hierarchical semantic alignment, thereby mitigating the geometric distortion of Euclidean embeddings. Additionally, we present a Geometry-aware Semantic Integration (GSI) module that combines hyperbolic structural awareness with tangent-space aggregation to fuse visual features and textual embeddings. To further mitigate the confirmation bias inherent in pseudo-labeling, we propose an Uncertainty-aware Region Mixing (URM) strategy, which leverages dual-network discrepancy to dynamically identify high-uncertainty regions for robust consistency training. **Contributions are as follows: i)** We propose a hyperbolic vision-language interaction network, introducing a novel geometric perspective for semi-supervised medical image segmentation. **ii)** We propose the HSE module to explicitly model anatomical hierarchies by aligning cross-modal features within the Poincaré ball, yielding more discriminative representations than conventional Euclidean approaches. **iii)** We present the GSI module for deeply fusing visual features and textual cues, achieving mathematically grounded and precise cross-modal semantic injection through tangent-space linearity. **iv)** Results on multiple datasets show that our model outperforms state-of-the-art semi-supervised segmentation approaches.

2 Proposed Method

Overview. In the semi-supervised setting, the training set $\mathcal{D}$ consists of a labeled subset $\mathcal{D}_L = \{(x_i^l, y_i)\}_{i=1}^{N_l}$ and an unlabeled subset $\mathcal{D}_U = \{x_i^u\}_{i=1}^{N_u}$, with $N_u \gg N_l$. $x_i^l, x_i^u \in \mathbb{R}^{W\times H\times D}$ represent the input volume and the $y_i \in \{0,1\}^{W\times H\times D}$ is the ground-truth map. Fig. 1 overviews the proposed framework. We employ a dual-network architecture comprising Network A and Network B, with V-Net [11] and ResVNet [19] as their respective backbones. Specifically, the input volumes $\{x_i^l, x_i^u\}$ are processed by the encoder to extract visual embeddings. To capture anatomical hierarchies, the HSE module projects these representations

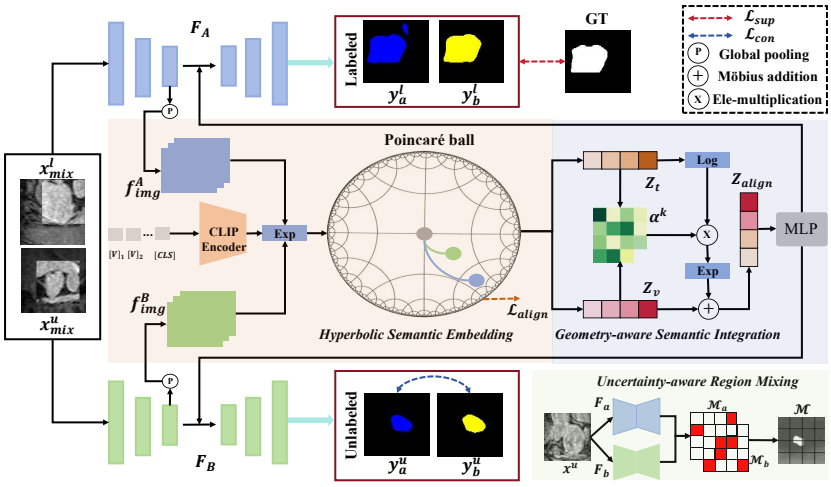

Fig. 1. Overview of our framework, with dual segmentation branches and a text encoder. Visual and textual features are projected into the hyperbolic space for semantic alignment, and the GSI is used to fuse the two features. Labeled predictions are directly supervised, while unlabeled ones are regularized with a consistency constraint.

into the Poincaré ball for alignment via hyperbolic distance. Subsequently, the GSI module performs cross-modal fusion via tangent space projection, ensuring effective semantic enhancement while preserving the hyperbolic structure. Finally, we leverage the URM strategy to mitigate confirmation bias by dynamically selecting high uncertainty regions for mixed supervision.

2.1 Hyperbolic Semantic Embedding

We first provide an introduction of Hyperbolic space. Hyperbolic space is a Riemannian manifold with constant negative curvature. It admits several equivalent isometric representations [25], among which the Poincaré ball model is the most widely used. In this work, we adopt the Poincaré ball model to characterize hyperbolic geometry. The Poincaré ball model $(\mathbb{D}_c^n, g^{\mathbb{D}_c})$ with radius $1/\sqrt{c}$ and negative curvature $-c$ $(c > 0)$ is a Riemannian manifold defined as $\mathbb{D}_c^n := \{\boldsymbol{x} \in \mathbb{R}^n : c\|\boldsymbol{x}\|^2 < 1\}$, equipped with the metric $g_x^{\mathbb{D}} = \lambda_x^2 g^E$, where $\lambda_x = \frac{2}{1-\|x\|^2}$ is the conformal factor, and $g^E = I_n$ denotes the Euclidean metric tensor. For $x, y \in \mathbb{D}_c^n$, the Möbius addition operation is defined as:

$$x \oplus_c y := \frac{(1 + 2c\langle x, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y\rangle + c^2\|x\|^2\|y\|^2}. \tag{1}$$

The distance of $x, y \in \mathbb{D}_c^n$ is defined as follows:

$$\mathrm{d}_{\mathbb{P}}^c(x, y) = \frac{2}{\sqrt{c}} \tanh^{-1}\left(\sqrt{c}\,\|-x \oplus_c y\|\right). \tag{2}$$

The tangent space $T_x\mathbb{D}_c^n$ at a point $x \in \mathbb{D}_c^n$ is the first-order approximation of $\mathbb{D}_c^n$, and it is an n-dimensional Euclidean space $\mathbb{R}^n$. The mappings between Euclidean and hyperbolic spaces [3] are established via the exponential and logarithmic maps at $x = 0$, which are defined for $v \in T_0\mathbb{D}_c^n$ and $y \in \mathbb{D}_c^n$ as follows:

$$\mathrm{Exp}_0^c(v) = \tanh\left(\sqrt{c} \cdot \|v\|\right) \cdot \frac{v}{\sqrt{c} \cdot \|v\|}, \quad \mathrm{Log}_0^c(y) = \tanh^{-1}\left(\sqrt{c} \cdot \|y\|\right) \cdot \frac{y}{\sqrt{c} \cdot \|y\|}. \tag{3}$$

Vision-language models like CLIP, pre-trained on 2D natural images, face a domain gap when applied to volumetric medical data. Additionally, the hierarchical structure of medical anatomy is poorly captured in Euclidean space. To address this, the HSE module introduces learnable semantic prompts and projects both visual and textual features into the Poincaré ball, enabling alignment within a geometry-aware manifold that respects anatomical hierarchies. For textual embeddings, hand-crafted prompt templates (e.g., "A photo of liver") are often insufficient to capture the fine-grained semantics and domain-specific nuances inherent to medical imagery. To address this limitation, we adopt Context Optimization (CoOp) [32] to learn continuous, data-driven prompts. Specifically, the input prompt $\mathcal{P}_k$ for the k-class is formulated by

$$\mathcal{P}_k = [\mathbf{v}]_1[\mathbf{v}]_2 \dots [\mathbf{v}]_M[\mathrm{CLS}]_k, \tag{4}$$

where $[\mathbf{v}]$ denotes a set of learnable context vectors shared across all classes, and $[\mathrm{CLS}]_k$ represents the class-specific token. Following [32], the set of prompts $\{\mathcal{P}_k\}_{k=1}^K$ is processed by the frozen CLIP text encoder to generate text features $f_t \in \mathbb{R}^{K \times C}$, where K and C denote the number of classes and channels. Correspondingly, we extract the global image feature $f_{img}^{A/B}$ from the encoder's deepest layer via global average pooling. To foster geometry-aware alignment, we project both feature representations into the Poincaré ball. By the exponential map defined in Eq.(3), the hyperbolic embeddings are obtained by

$$\mathcal{Z}_v^{\mathrm{A/B}} = \mathrm{Exp}_0^c(f_{img}^{\mathrm{A/B}}), \mathcal{Z}_t = \mathrm{Exp}_0^c(f_t). \tag{5}$$

To enforce semantic consistency, we minimize the hyperbolic distance between the visual feature and its textual embedding using the distance metric $\mathrm{d}_{\mathbb{P}}^c(\cdot,\cdot)$ defined in Eq.(2). Then, the alignment loss for each network is defined by

$$\mathcal{L}_{\mathrm{align}}^{\mathrm{A/B}} = \mathcal{L}_{ce}(-\mathrm{d}_{\mathbb{P}}^c(\mathcal{Z}_v^{\mathrm{A/B}}, \mathcal{Z}_t), y). \tag{6}$$

2.2 Geometry-aware Semantic Integration

We propose the GSI module to model hyperbolic relations and tangent-space aggregation for providing anatomically consistent guidance during decoding. Specifically, the hyperbolic visual embedding $\mathcal{Z}_v^{\mathrm{A/B}}$ is treated as the query, while textual embeddings $\mathcal{Z}_t^k$ serve as keys and values. The attention weights are defined based on the hyperbolic distance as follows:

$$\alpha_{\mathrm{A/B}}^k = \sigma\Big(\frac{1}{1 + \log(1 + \mathrm{d}_{\mathbb{P}}^c(\mathcal{Z}_v^{\mathrm{A/B}}, \mathcal{Z}_t^k))}\Big), \tag{7}$$

where $\sigma(\cdot)$ is the softmax activation function. The textual embeddings are aggregated in the tangent space and then fused with the visual feature $\mathcal{Z}_v^{\mathrm{A/B}}$ via a hyperbolic residual connection. The aligned representation $\mathcal{Z}_{align}^{\mathrm{A/B}}$ is derived by

$$\mathcal{Z}_{align}^{\mathrm{A/B}} = \mathcal{Z}_v^{\mathrm{A/B}} \oplus_c \mathrm{Exp}_0^c \left(\sum\nolimits_{k=1}^{K} \alpha_{\mathrm{A/B}}^k \cdot \mathrm{Log}_0^c(\mathcal{Z}_t^k) \right), \tag{8}$$

where $\mathrm{Log}_0^c(\cdot)$ and $\mathrm{Exp}_0^c(\cdot)$ denote the logarithmic and exponential maps defined in Eq. (3). Finally, this aligned representation guides the refinement of the deepest encoder representation $F_{enc}^{\mathrm{A/B}}$. The aligned hyperbolic embedding $\mathcal{Z}_{align}^{\mathrm{A/B}}$ is first mapped to the tangent space via the logarithmic map, and subsequently transformed by a lightweight MLP to produce channel-wise modulation weights. The refined feature $F_{out}^{\mathrm{A/B}}$ is passed to the decoder, which is obtained by

$$F_{out}^{\mathrm{A/B}} = F_{enc}^{\mathrm{A/B}} + F_{enc}^{\mathrm{A/B}} \otimes \mathrm{Sigmoid}\big(\mathrm{MLP}(\mathrm{Log}_0^c(\mathcal{Z}_{align}^{\mathrm{A/B}}))\big), \tag{9}$$

where $\otimes$ denotes channel-wise multiplication along the spatial dimensions.

2.3 Uncertainty-aware Region Mixing

We propose the URM strategy to mitigate the confirmation bias and ineffective background sampling inherent in stochastic CutMix [1,9]. Unlike random cropping, URM leverages prediction discrepancies between dual networks to dynamically identify high-uncertainty regions, enabling more targeted and effective consistency training. We formalize this process by generating a binary mask $\mathcal{M} \in \{0,1\}^{W \times H \times D}$ to identify the local region with the highest uncertainty, where 1 represents the selected highly uncertain region and 0 indicates the remaining areas. The uncertainty is computed as the mean squared difference between the predictions. Therefore, the mask can be obtained by

$$\mathcal{M} = \mathbf{1}[p \in \arg\max_{\Omega} \sum\nolimits_{q \in \Omega} \frac{1}{K} \sum\nolimits_{k=1}^{K} \big(P_1^{(k)}(q) - P_2^{(k)}(q)\big)^2], \tag{10}$$

where $\mathbf{1}[\cdot]$ is the indicator function, $P_1^{(k)}(q)$ and $P_2^{(k)}(q)$ denote the probability maps for class k at voxel q, predicted by the two networks respectively, and Ω denotes a candidate region with a preset size of $\gamma W \times \gamma H \times \gamma D$, where $\gamma \in (0,1)$. Then, we use this mask to control the mixing of unlabeled and labeled samples, obtaining enhanced mixed samples and corresponding labels:

$$x_{\mathrm{mix}} = \mathcal{M} \odot x^u + (1 - \mathcal{M}) \odot x^l, \quad y_{\mathrm{mix}} = \mathcal{M} \odot \hat{y}^u + (1 - \mathcal{M}) \odot y^l, \tag{11}$$

where $\odot$ denotes element-wise multiplication. The mixed sample x_{mix} is then fed into the two networks to obtain the predictions $\hat{y}_{\mathrm{mix}}^{\mathrm{A}}$ and $\hat{y}_{\mathrm{mix}}^{\mathrm{B}}$, respectively.

2.4 Overall Loss Function

Our segmentation loss is defined as $\mathcal{L}_{\mathrm{seg}}(\hat{y}, y) = \mathcal{L}_{\mathrm{dice}}(\hat{y}, y) + \mathcal{L}_{\mathrm{ce}}(\hat{y}, y)$, where $\hat{y}$ and y denote the predicted map and the corresponding ground truth, respectively. Therefore, the supervised loss is formulated by

$$\mathcal{L}_{\mathrm{sup}}^{\mathrm{A/B}} = \mathcal{L}_{\mathrm{seg}}(\hat{y}_{\mathrm{mix}}^{\mathrm{A/B}}, y_{\mathrm{mix}}) \odot (1 - \mathcal{M}) + \beta \mathcal{L}_{\mathrm{seg}}(\hat{y}_{\mathrm{mix}}^{\mathrm{A/B}}, y_{\mathrm{mix}}) \odot \mathcal{M}, \tag{12}$$

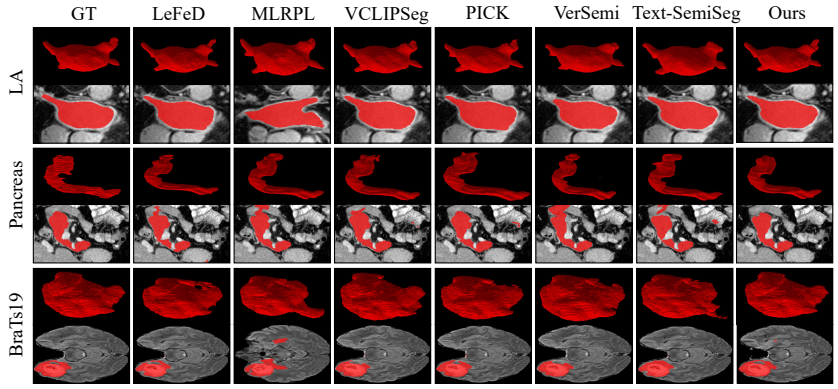

Fig. 2. Visualization results on the LA, Pancreas, and BraTS-2019 datasets.

where β controls the contribution of the unlabeled regions in the supervised loss. To encourage prediction consistency between the dual networks, we utilize a consistency loss by

$$\mathcal{L}_{\text{con}} = \left[\mathcal{L}_{\text{seg}}(\hat{y}^A_{\text{mix}}, \hat{y}^B_{\text{mix}}) + \mathcal{L}_{\text{seg}}(\hat{y}^B_{\text{mix}}, \hat{y}^A_{\text{mix}})\right] \odot \mathcal{M}. \tag{13}$$

Finally, the overall training loss function can be formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\text{A/B}}_{\text{sup}} + \mu\mathcal{L}^{\text{A/B}}_{\text{align}} + \lambda\mathcal{L}_{\text{con}}, \tag{14}$$

where μ and λ are trade-off parameters.

3 Experiments and Results

3.1 Experimental Setup

Datasets. In this study, we conduct comparison experiments on three commonly used datasets. • The LA dataset [22] is a commonly used 3D medical-image semi-supervised segmentation benchmark comprising 100 3D MRI volumes with a fixed resolution of $0.625 \times 0.625 \times 0.625$ mm. Following [26], 80 volumes are used for training, while the remaining 20 ones are reserved for validation. • The Pancreas dataset [15] comprises 82 contrast-enhanced 3D CT scans. Each scan has an in-plane size of 512×512 pixels, with slice thickness varying between 1.5 and 2.5 mm. Following previous work [8], 62 scans are used for training, and the remaining 20 ones for testing. • The BraTS-2019 dataset [10] contains 335 labeled brain tumor MRI volumes of size 240×240×155, each comprising four modalities: T1, T1-ce, T2, and FLAIR. Following [23], 250, 25, and 60 volumes are used for training, validation, and testing, respectively.

Implementation Details. All experiments are implemented using PyTorch 1.9.1 with CUDA 11.1 and conducted on a single NVIDIA RTX 3090 GPU.

Table 1. Quantitative results on the LA, Pancreas, and BraTS-2019 datasets.

Dataset	Method	10% / 8 labeled data				20% / 16 labeled data			
		Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
LA	UA-MT [26] (MICCAI'19)	86.28	76.11	18.71	4.63	88.74	79.94	8.39	2.32
	DTC [8] (AAAI'21)	87.51	78.17	8.23	2.36	89.52	81.22	7.07	1.96
	MC-Net [21] (MICCAI'21)	87.50	77.98	11.28	2.30	90.12	82.12	8.07	1.99
	MCF [19] (CVPR'23)	86.65	77.41	10.95	2.78	88.71	80.41	6.32	1.90
	BCP [1] (CVPR'23)	89.62	81.31	6.81	1.76	91.26	84.01	5.76	1.86
	LeFeD [29] (TMI'24)	90.12	81.78	7.02	1.74	91.39	84.17	6.15	1.47
	MLRPL [16] (MIA'24)	89.86	81.68	6.91	1.85	91.02	83.62	5.78	1.66
	VCLIPSeg [7] (MICCAI'24)	90.59	82.87	6.22	1.61	91.15	83.82	5.86	1.49
	PICK [27] (IJCV'25)	88.58	79.69	9.74	2.32	90.85	83.37	5.70	1.74
	VerSemi [28] (TMI'25)	89.01	80.52	9.03	2.57	90.65	83.12	5.95	1.82
	Text-SemiSeg [5] (MICCAI'25)	90.26	82.25	6.48	1.94	91.54	84.61	5.99	1.67
	Ours	**91.56**	**84.51**	**5.13**	**1.54**	**92.49**	**86.08**	**4.15**	**1.40**
	Method	10% / 6 labeled data				20% / 12 labeled data			
Pancreas	UA-MT [26] (MICCAI'19)	66.44	52.02	17.04	3.03	76.10	62.62	10.84	2.43
	DTC [8] (AAAI'21)	66.58	51.79	15.46	4.16	76.27	62.82	8.70	2.20
	MC-Net [21] (MICCAI'21)	69.07	54.36	14.53	2.28	78.17	65.22	6.90	1.55
	MCF [19] (CVPR'23)	67.97	53.03	16.98	2.78	75.00	61.27	11.59	3.27
	BCP [1] (CVPR'23)	74.89	61.23	10.45	2.98	82.91	70.97	6.43	2.25
	LeFeD [29] (TMI'24)	75.35	61.38	11.47	3.42	82.17	71.17	5.25	1.40
	MLRPL [16] (MIA'24)	75.93	62.12	9.07	1.54	81.53	69.35	6.81	1.33
	VCLIPSeg [7] (MICCAI'24)	74.77	60.88	11.02	1.82	80.56	68.13	6.75	1.49
	PICK [27] (IJCV'25)	73.76	62.21	10.72	3.03	80.14	67.65	6.12	1.55
	VerSemi [28] (TMI'25)	78.08	64.82	8.05	2.33	83.27	71.68	5.33	1.40
	Text-Semiseg [5] (MICCAI'25)	81.27	68.78	5.98	1.44	83.50	71.94	4.52	1.26
	Ours	**83.68**	**72.21**	**5.36**	**1.25**	**84.80**	**73.81**	**4.29**	**1.10**
	Method	10% / 25 labeled data				20% / 50 labeled data			
BraTS-2019	UA-MT [26] (MICCAI'19)	79.49	69.22	11.93	1.93	79.72	69.46	11.26	1.97
	DTC [8] (AAAI'21)	77.54	66.91	12.16	2.84	82.38	72.15	11.00	2.16
	MC-Net [21] (MICCAI'21)	81.52	71.89	13.26	4.56	83.79	73.69	9.65	1.76
	MCF [19] (CVPR'23)	78.83	68.49	12.25	1.92	80.07	69.55	10.65	2.00
	BCP [1] (CVPR'23)	78.64	68.59	12.88	2.81	80.20	69.66	10.52	2.08
	LeFeD [29] (TMI'24)	85.05	75.29	9.80	2.89	86.14	76.84	8.35	2.44
	MLRPL [16] (MIA'24)	84.29	74.74	9.57	2.55	85.47	76.32	7.76	2.00
	VCLIPSeg [7] (MICCAI'24)	83.12	73.61	10.40	1.65	84.05	74.38	9.75	1.35
	PICK [27] (IJCV'25)	85.30	75.69	9.90	2.66	86.27	76.96	8.10	2.18
	VerSemi [28] (TMI'25)	85.14	75.37	9.70	2.72	86.52	77.38	8.54	2.46
	Text-Semiseg [5] (MICCAI'25)	85.27	75.83	8.77	**1.35**	86.69	77.83	7.65	**1.24**
	Ours	**86.12**	**76.53**	**7.83**	1.51	**87.21**	**78.23**	**6.92**	1.38

Models are optimized using Adam with an initial learning rate of 1×10^{-3}. The batch size is 4, which contains 2 labeled data volumes and 2 unlabeled volumes. The input volumes are randomly cropped to 112×112×80 for the LA dataset, and 96×96×96 for the Pancreas and BraTS-2019 datasets. The number of learnable context tokens in CoOp is set to 4. We set $\beta = 0.5$, $\mu = 0.1$, $\gamma = 2/3$ and use a Gaussian warm-up schedule $\lambda(t) = 0.1 \cdot e^{-5(1-t/t_{\max})^2}$ [6] for λ. The curvature parameter of Poincaré ball is fixed to 0.1 in all experiments.

3.2 Comparison with State-of-the-arts

We compare our model with 11 state-of-the-art semi-supervised medical image segmentation methods, namely UA-MT [26], DTC [8], MC-Net [21], MCF [19], BCP [1], LeFeD [29], MLRP [16], VCLIPSeg [7], PICK [27], VerSemi [28], and Text-SemiSeg [5]. As shown in Table 1, our method consistently outperforms

Table 2. Ablation study on the key components using 10% labeled data.

Settings			LA (10% labeled data)				Pancreas (10% labeled data)			
HSE	GSI	URM	Dice ↑	IoU ↑	95HD ↓	ASD ↓	Dice ↑	IoU ↑	95HD ↓	ASD ↓
			88.20	80.21	6.95	3.80	77.53	65.12	9.32	4.16
✓			89.05	81.32	6.23	2.83	78.51	67.63	7.62	3.70
	✓		88.93	80.81	6.45	3.12	79.32	68.38	6.51	2.99
		✓	90.24	82.12	6.06	2.66	81.86	70.29	6.02	2.69
✓	✓		90.38	82.33	5.99	2.12	81.48	69.75	6.32	2.47
✓		✓	91.11	83.68	5.32	1.76	82.93	71.51	5.66	2.03
	✓	✓	90.84	82.95	5.84	1.82	82.14	71.07	5.87	1.57
✓	✓	✓	**91.56**	**84.51**	**5.13**	**1.54**	**83.68**	**72.21**	**5.36**	**1.25**

Table 3. Effect of HSE strategy applied to other semi-supervised methods.

Settings	LA (8 labeled data)				Pancreas (6 labeled data)			
	Dice ↑	IoU ↑	95HD ↓	ASD ↓	Dice ↑	IoU ↑	95HD ↓	ASD ↓
MCF [19]	86.65	77.41	10.95	2.78	67.97	53.03	16.98	2.78
+HSE	**88.12**	**78.82**	**7.63**	**2.15**	**71.01**	**56.02**	**13.23**	**2.55**
BCP [1]	89.62	81.31	6.81	1.76	74.89	61.23	10.45	2.98
+HSE	**90.92**	**83.40**	**6.28**	**1.52**	**78.63**	**65.45**	**8.66**	**2.20**

competitors on the LA, Pancreas, and BraTS-2019 datasets under limited annotations. Compared to the second-best methods, our approach improves Dice scores by 1.30%, 2.41%, and 0.85% under the 10% labeled setting, respectively. Fig. 2 illustrates its superior boundary accuracy and anatomical delineation, which are further evidenced by consistently lower 95HD and ASD metrics.

3.3 Ablation Studies

Effectiveness of Key Components: Table 2 shows the ablative results under 10% labeled data, where each module improves performance over the baseline. HSE and GSI enhance hierarchical semantic alignment and geometry-consistent fusion, while URM brings more pronounced gains by alleviating confirmation bias in pseudo-label learning. Notably, our full model with all components achieves the best performance, indicating their complementary effects.

Fairness Discussion: The existing semi-supervised approaches primarily depend on visual consistency, while the exploration of VLMs in 3D contexts remains limited. To verify both the necessity and validity of incorporating textual cues, we integrate our proposed HSE module into existing vision-only frameworks, including MCF [19] and BCP [1]. As shown in Table 3, the integration of our HSE module with textual guidance consistently improves segmentation performance across different methods.

Prompt and Complexity Analysis: Table 4 presents the performance of different prompt strategies alongside their computational costs on the LA dataset. It is observed that CoOp outperforms the fixed prompt template "A photo of [CLS]", indicating that learnable prompts can adapt textual semantics more effectively to medical images, thus providing more suitable semantic guidance for segmentation. Furthermore, the complexity analysis reveals that HyperSemi incurs only moderate computational overhead compared to its

Table 4. Prompt and complexity analysis on the LA dataset with 10% labeled data.

Settings	Prompt Strategy				Settings	Complexity		
	Dice ↑	IoU ↑	95HD ↓	ASD ↓		Time ↓	Mem. ↓	Lat. ↓
A photo of [CLS]	90.72	83.12	6.33	1.77	w/o Hyperbolic	**53.77**	**7.29**	**2.109**
CoOp	**91.56**	**84.51**	**5.13**	**1.54**	HyperSemi	76.37	8.31	2.533

non-hyperbolic counterpart, demonstrating a reasonable trade-off between performance and efficiency.

4 Conclusion

We propose a hyperbolic vision-language interaction network for semi-supervised medical image segmentation, offering a novel geometry-aware perspective that leverages cross-modal learning to model anatomical hierarchies. Experimental results on diverse datasets show that our method consistently outperforms existing state-of-the-art semi-supervised medical image segmentation approaches.

Acknowledgement . This work was in part supported by the National Natural Science Foundation of China (No. 62576153).

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

References

1. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: CVPR. pp. 11514–11524 (2023)
2. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proc. CVPR. pp. 19786–19797 (2023)
3. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic neural networks. Proc. NeurIPS 31 (2018)
4. Huang, H., Huang, Y., Xie, S., Lin, L., Tong, R., Chen, Y.W., Li, Y., Zheng, Y.: Combinatorial cnn-transformer learning with manifold constraints for semi-supervised medical image segmentation. In: Proc. AAAI. vol. 38, pp. 2330–2338 (2024)
5. Huang, K., Zhou, Y., Fu, H., Zhang, Y., Gong, C., Zhou, T.: Text-driven multiplanar visual interaction for semi-supervised medical image segmentation. In: Proc. MICCAI. pp. 604–613. Springer (2025)
6. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
7. Li, L., Lian, S., Luo, Z., Wang, B., Li, S.: Vclipseg: Voxel-wise clip-enhanced model for semi-supervised medical image segmentation. In: Proc. MICCAI. pp. 692–701. Springer (2024)
8. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI. vol. 35, pp. 8801–8809 (2021)

9. Ma, Q., Zhang, J., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In: Proc. CVPR. pp. 11642–11651 (2024)
10. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 34(10), 1993–2024 (2014)
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571. IEEE (2016)
12. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. AProc. NeurIPS 30 (2017)
13. Peng, W., Varanka, T., Mostafa, A., Shi, H., Zhao, G.: Hyperbolic deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(12), 10023–10044 (2021)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
15. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: Proc. MICCAI. pp. 556–564. Springer (2015)
16. Su, J., Luo, Z., Lian, S., Lin, D., Li, S.: Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. Medical Image Analysis 94, 103111 (2024)
17. Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C., Desrosiers, C.: Self-paced and self-consistent co-training for semi-supervised image segmentation. Medical Image Analysis 73, 102146 (2021)
18. Wang, X., Yuan, Y., Guo, D., Huang, X., Cui, Y., Xia, M., Wang, Z., Bai, C., Chen, S.: SSA-Net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning. Medical Image Analysis 79, 102459 (2022)
19. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: MCF: Mutual correction framework for semi-supervised medical image segmentation. In: Proc. CVPR. pp. 15651–15660 (2023)
20. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Medical Image Analysis 81, 102530 (2022)
21. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: Proc. MICCAI. pp. 297–306. Springer (2021)
22. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., 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)
23. Xu, Z., Wang, Y., Lu, D., Yu, L., Yan, J., Luo, J., Ma, K., Zheng, Y., Tong, R.K.y.: All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. IEEE Journal of Biomedical and Health Informatics 26(7), 3174–3184 (2022)
24. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proc. CVPR. pp. 7236–7246 (2023)

25. Yang, M., Zhou, M., Li, Z., Liu, J., Pan, L., Xiong, H., King, I.: Hyperbolic graph neural networks: A review of methods and applications. arXiv preprint arXiv:2202.13852 (2022)
26. 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: Proc. MICCAI. pp. 605–613. Springer (2019)
27. Zeng, Q., Lu, Z., Xie, Y., Xia, Y.: Pick: Predict and mask for semi-supervised medical image segmentation. International Journal of Computer Vision 133(6), 3296–3311 (2025)
28. Zeng, Q., Xie, Y., Lu, Z., Lu, M., Wu, Y., Xia, Y.: Segment together: A versatile paradigm for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging 44(7), 2948 – 2959 (2025)
29. Zeng, Q., Xie, Y., Lu, Z., Lu, M., Zhang, J., Xia, Y.: Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging 44(1), 44–56 (2024)
30. Zhang, F., Liu, H., Wang, J., Lyu, J., Cai, Q., Li, H., Dong, J., Zhang, D.: Cross co-teaching for semi-supervised medical image segmentation. Pattern Recognition 152, 110426 (2024)
31. Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., Ooi, B.C.: Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In: Proc. CVPR. pp. 20666–20676 (2022)
32. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision 130(9), 2337–2348 (2022)
33. Zhou, T., Gu, Y., Huang, K., Fu, H., Wu, X.j., Kittler, J.: Cross-sample consistency learning for semi-supervised medical image segmentation. International Journal of Computer Vision 134(5), 218 (2026)