Dual-Teacher Knowledge Distillation with SAM and Learnable Prototype Guidance for Semi-Supervised 3D Medical Segmentation

Shuaiguang Huo, Zifeng Sang, Xu Luo, Jinghua Zhu*, and Heran Xi**

Heilongjiang University, Harbin, 150080, China
{zhujinghua, xiheran}@hlju.edu.cn

Abstract. 3D medical image segmentation is critical for disease diagnosis and treatment planning, yet requires costly large-scale fine-grained annotations. Semi-Supervised Learning (SSL), combining limited labeled data with massive unlabeled data, offers an ideal solution. However, mainstream Mean Teacher (MT) methods suffer from two core limitations: (1) The single-teacher paradigm is prone to confirmation bias during iteration, causing error accumulation especially in low-contrast challenging regions; (2) Limited labels fail to provide sufficient anatomical priors. To this end, we propose a dual-teacher knowledge distillation framework for semi-supervised 3D medical segmentation, integrating an internal teacher, an external teacher and a learnable prototype guidance mechanism. An EMA-updated internal teacher provides stable consistency supervision, while a frozen SAM-Med3D external teacher generates high-quality segmentation maps via prompt-driven generation to distill anatomical shape and boundary priors. We introduce a learnable prototype module to calibrate internal pseudo-labels via residual fusion, and design an entropy-guided loss to mine hard examples in high-uncertainty regions to boost feature discriminability. Experiments on LA, Pancreas-CT and BraTS19 benchmarks demonstrate the proposed method achieves state-of-the-art performance on most metrics.The code is available at https://github.com/Aguang094/Dualteacher-with-SAM.

Keywords: 3D Medical Image Segmentation · SAM-Med3D · Learnable Prototypes

1 Introduction

3D medical image segmentation is essential for disease diagnosis, surgical planning, and treatment evaluation [14]. Recent Transformer architectures and large-scale foundation models [24] have significantly advanced 3D segmentation performance. However, these methods rely on extensive voxel-level annotations, the

* Corresponding author
** Corresponding author

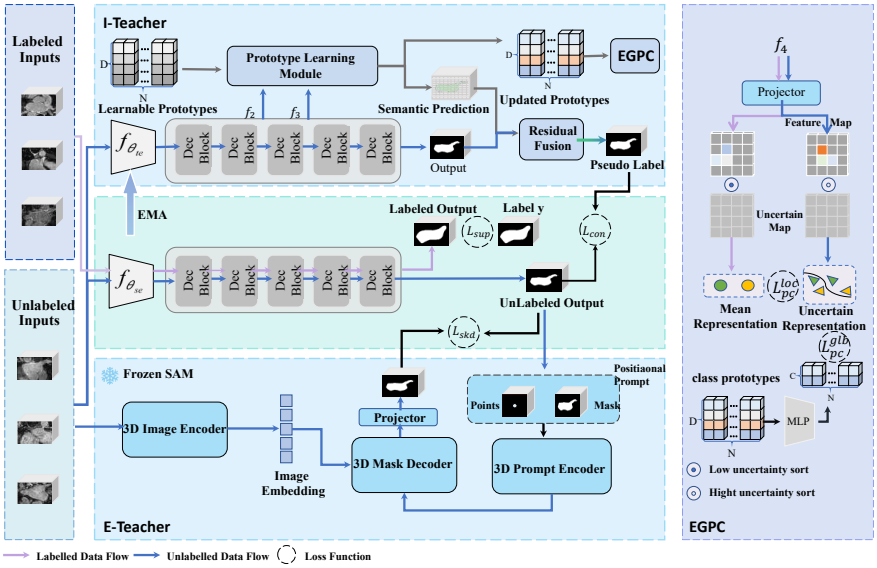

Fig. 1. Dual-teacher framework combining an internal EMA teacher and an external SAM-Med3D teacher. A learnable prototype module corrects pseudo-labels via residual fusion; an entropy-guided prototype calibration loss enhances feature discriminability.

acquisition of which demands substantial expert labor and prolonged annotation cycles. SSL alleviates this burden by leveraging unlabeled data [15], yet effectively mining information from unlabeled samples while avoiding noisy supervision remains a core challenge.

Most semi-supervised methods follow the MT framework and its variants [20], which build a teacher via exponential moving average (EMA) of student parameters and use consistency regularization for self-distillation. Some studies also enhance robustness through dual-task consistency or uncertainty guidance. However, this single-teacher paradigm still suffers from two inherent limitations. First is confirmation bias [1]: the teacher's knowledge originates entirely from the student's historical states, and low-quality pseudo-labels tend to accumulate errors during training. Second is insufficient anatomical priors: medical images contain numerous low-contrast tissue boundaries and small lesions with variable morphologies. Models trained on limited labeled data struggle to generalize to complex anatomical structures.

To address these issues, we propose a Dual-Teacher Knowledge Distillation framework with SAM and learnable prototype guidance, retaining the EMA model as internal teacher for training stability while introducing a frozen SAM-Med3D as external teacher. As a foundational segmentation model, SAM contains rich general visual knowledge and has shown promising performance on various medical image tasks [14]. By prompting the frozen SAM-Med3D to generate high-quality probability maps and aligning the student's predictions with

them via distillation loss, SAM's anatomical shape priors are transferred to the student.

However, dual-teacher supervision may cause signal conflicts. We further design a learnable prototype guidance module: learnable class prototypes in the decoder are iteratively updated via clustering, attention, and interaction with decoder features, learning robust class-specific semantic representations. These prototypes not only reconstruct predictions to correct EMA pseudo-labels and mitigate confirmation bias, but also serve as semantic anchors for feature space calibration [23]. Combined with the proposed entropy-guided prototype calibration loss, they impose contrastive constraints on high-uncertainty regions in feature space, improving segmentation accuracy on fuzzy boundaries.The main contributions are summarized as follows:

- A dual-teacher distillation framework that injects external SAM priors into semi-supervised learning for more robust supervision;
- A learnable prototype module that corrects pseudo-label noise via prototype-induced semantic prediction and residual fusion, mitigating confirmation bias;
- An entropy-guided prototype calibration loss that mines hard samples and improves accuracy on fuzzy boundaries and small anatomical structures.

2 Methodology

We propose a Dual-Teacher Knowledge Distillation framework with an internal EMA teacher and an external SAM-Med3D teacher, where learnable prototypes calibrate pseudo-labels and enhance feature discrimination. Fig. 1 shows the overall architecture.

2.1 Mean Teacher Baseline

In semi-supervised segmentation, we have a small labeled set $\mathcal{D}_L = \{(x_i, y_i)\}$ and a large unlabeled set $\mathcal{D}_U = \{x_j\}$. The MT framework [20] trains a student f_θ by minimizing:

$$\min_{\theta} \left(\sum_{(x_i,y_i)\in\mathcal{D}_L} \mathcal{L}_{\text{sup}}\big(f(x_i;\theta), y_i\big) + \sum_{x\in\mathcal{D}_U} \mathcal{L}_{\text{mt-cons}}\big(f(x;\theta), \tilde{y}(x)\big) \right), \tag{1}$$

where $\mathcal{L}_{\text{sup}}$ combines cross-entropy and Dice loss, and $\mathcal{L}_{\text{mt-cons}}$ enforces consistency with pseudo-labels $\tilde{y}(x) = \arg\max f_{\theta_t}(x)$. The teacher is updated via EMA: $\theta_t^T = \alpha\theta_{t-1}^T + (1-\alpha)\theta_t^S$.

2.2 Dual-Teacher Collaborative Distillation Framework

We propose a dual-teacher framework with an internal EMA teacher and a frozen SAM-Med3D teacher [21], plus learnable prototype calibration and entropy-guided feature enhancement.

Pseudo-Label Calibration Guided by Learnable Prototypes We maintain L learnable query prototypes $\mathbf{Q} \in \mathbb{R}^{B\times L\times d_q}$, where B is the batch size, L the number of prototypes, and d_q the feature dimension. Given a decoder feature map $\mathbf{F} \in \mathbb{R}^{B\times F\times H'\times W'\times D'}$ with F channels and spatial size $H'\times W'\times D'$, we compute pixel embeddings $\mathbf{E}_x = \mathrm{Norm}_{\ell_2}(\varphi(\mathbf{F}))$ via a learnable projection $\varphi(\cdot)$ and ℓ_2-normalization, and project prototypes to $\mathbf{K} = \psi(\mathbf{Q})$ with a learnable 1D convolution $\psi(\cdot)$. Cosine similarity logits are:

$$\mathbf{S}_{b,l,h,w,d} = \frac{\sum_{f=1}^{F} \mathbf{E}_{x,b,f,h,w,d} \cdot \mathbf{K}_{b,l,f}}{\|\mathbf{E}_{x,b,:,h,w,d}\|_2 \cdot \|\mathbf{K}_{b,l,:}\|_2}, \tag{2}$$

and hard assignments are $\mathbf{A}_{b,l,h,w,d} = \mathbb{I}[\, l = \arg\max_{l'} \mathbf{S}_{b,l',h,w,d}\,]$.

A classification head h_{cls} predicts class distributions $\mathbf{P} = \mathrm{Softmax}(h_{\mathrm{cls}}(\mathbf{Q})) \in [0,1]^{B\times L\times (C+1)}$, where C is the number of foreground classes and the $+1$ accounts for a background class. Prototypes are aggregated using hard assignments and updated with momentum:

$$\mathbf{Q}^{\mathrm{agg}} = \frac{\sum_s \mathbf{A}_{:,:,s} \cdot \mathbf{V}_{:,:,s}}{\sum_s \mathbf{A}_{:,:,s}}, \quad \mathbf{Q} \leftarrow (1-\mu)\mathbf{Q} + \mu\mathbf{Q}^{\mathrm{agg}}, \tag{3}$$

with separately projected value features $\mathbf{V}$. Queries are then refined via multi-head self-attention (MHSA) and a feed-forward network (FFN): $\mathbf{Q}' = \mathbf{Q} + \mathrm{MHSA}(\mathrm{LN}(\mathbf{Q}))$, $\mathbf{Q}'' = \mathbf{Q}' + \mathrm{FFN}(\mathrm{LN}(\mathbf{Q}'))$.

The reconstructed prediction is $\mathbf{Y}^{\mathrm{proto}}_{b,c,s} = \sum_l \mathbf{P}_{b,l,c}\mathbf{A}_{b,l,s}$. It is fused with the main UNet prediction $\mathbf{Y}^{\mathrm{main}}$ via a residual module:

$$\mathbf{R} = g_\phi([\mathbf{Y}^{\mathrm{main}}; \mathbf{Y}^{\mathrm{proto}}]), \quad \mathbf{Y}^{\mathrm{corr}} = \mathbf{Y}^{\mathrm{main}} + \mathbf{R}, \tag{4}$$

where g_ϕ consists of two $1\times1\times1$ convolutions (BN-ReLU). This calibration is applied at the 3rd and 2nd decoder levels, which have 64 and 32 channels respectively, sharing the same set of query prototypes across scales.

External Teacher Distillation Based on SAM-Med3D For an unlabeled image $\mathbf{X}_u$, the EMA teacher produces a pseudo-mask $\tilde{\mathbf{M}}_u = \arg\max_c\ f_{\theta^T}(\mathbf{X}_u)_c$. We sample 10 point prompts per volume, 5 foreground and 5 background, denoted as $\mathbf{q}_u$, and feed them into frozen SAM-Med3D F_Ω to obtain $\mathbf{P}_u^{\mathrm{sam}} = \sigma(F_\Omega(\mathbf{X}_u, \mathbf{q}_u))$. The student's foreground probability $\mathbf{P}_u^{\mathrm{stu}} = \mathrm{Softmax}(\hat{\mathbf{Y}}_u)_{c=1}$ is aligned via MSE:

$$\mathcal{L}_{\text{sam-dist}} = \left\|\mathbf{P}_u^{\mathrm{stu}} - \mathbf{P}_u^{\mathrm{sam}}\right\|_2^2, \tag{5}$$

activated after a 5-epoch warm-up.

2.3 Prototype Construction and Entropy-Guided Feature Calibration

We further enhance feature discriminability by aligning high-uncertainty voxels with class prototypes. For a voxel with prediction $\mathbf{p}_i \in [0,1]^C$, entropy is

$e_i = -\sum_c p_{i,c} \log(p_{i,c})$. On labeled data, we apply 3D erosion with a 3^3 kernel to ground-truth masks to obtain reliable class interiors, then average features of low-entropy voxels: $\mathbf{p}_{b,c}^{\text{loc}} = \frac{1}{|\mathcal{S}_{b,c}^{\text{loc}}|} \sum_{i \in \mathcal{S}_{b,c}^{\text{loc}}} \mathbf{z}_{b,:,i}^{l}$, where $\mathcal{S}_{b,c}^{\text{loc}}$ denotes selected low-entropy voxels for class c, and $\mathbf{z}_{b,:,i}^{l} \in \mathbb{R}^{F_r}$ are feature vectors from the representation branch.

Global prototypes are aggregated from projected query prototypes $\tilde{\mathbf{Q}}$:

$$\mathbf{p}_{b,c}^{\text{glb}} = \frac{\sum_{q=1}^{L} w_{b,q,c}\ \tilde{\mathbf{Q}}_{b,q,:}}{\sum_{q=1}^{L} w_{b,q,c} + \epsilon}, \quad w_{b,q,c} = \text{Softmax}(\mathbf{L}_{b,q,:C})_c, \tag{6}$$

where $\mathbf{L}$ are prototype classification logits for foreground classes only, and $\epsilon = 10^{-6}$.

For unlabeled data, we select high-entropy anchor voxels $\mathcal{A}_b$. For an anchor with feature $\mathbf{a}_{b,n}$ and prediction $\mathbf{p}_{b,n}^{u}$, cosine similarity to class prototypes, scaled by temperature τ, gives $\boldsymbol{\ell}_{b,n,c} = \frac{\cos(\mathbf{a}_{b,n}, \mathbf{p}_{b,c})}{\tau}$ and a prototype-assigned distribution $\boldsymbol{\pi}_{b,n} = \text{Softmax}(\boldsymbol{\ell}_{b,n,:})$. We align the student's re-normalized prediction $\tilde{\mathbf{p}}_{b,n}^{u}$ with $\boldsymbol{\pi}_{b,n}$ via KL divergence:

$$\mathcal{L}_{\text{pc}}^{(\cdot)} = \frac{1}{B} \sum_{b} \frac{1}{|\mathcal{A}_b|} \sum_{n \in \mathcal{A}_b} \text{KL}\big(\tilde{\mathbf{p}}_{b,n}^{u} \parallel \boldsymbol{\pi}_{b,n}\big)\,, \tag{7}$$

where $(\cdot)$ denotes local or global prototypes, and $\tilde{\mathbf{p}}_{b,n}^{u}$ is the re-normalized student prediction. For high-confidence anchors $\mathcal{A}_b^{\text{hard}}$, we additionally minimize hard cross-entropy:

$$\mathcal{L}_{\text{hard}} = \frac{1}{B} \sum_{b} \frac{1}{|\mathcal{A}_b^{\text{hard}}|} \sum_{n \in \mathcal{A}_b^{\text{hard}}} \text{CE}\Big(\boldsymbol{\ell}_{b,n,:},\ \arg\max_{c} \mathbf{p}_{b,n,c}^{u}\Big)\,. \tag{8}$$

The total prototype calibration loss is:

$$\mathcal{L}_{\text{pc}} = \lambda_{\text{loc}} \mathcal{L}_{\text{pc}}^{\text{loc}} + \lambda_{\text{glb}} \mathcal{L}_{\text{pc}}^{\text{glb}} + \lambda_{\text{hard}} \mathcal{L}_{\text{hard}}. \tag{9}$$

The overall objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{sam}} \mathcal{L}_{\text{sam-dist}} + \lambda_{\text{pc}} \mathcal{L}_{\text{pc}}, \tag{10}$$

with weights given in Sec. 3.1.

3 Experiments and Results

3.1 Datasets and Implementations

Comprehensive experiments are conducted on three representative 3D medical segmentation benchmarks spanning brain tumors, cardiac structures, and abdominal organs, covering both MRI and CT imaging modalities. These tasks

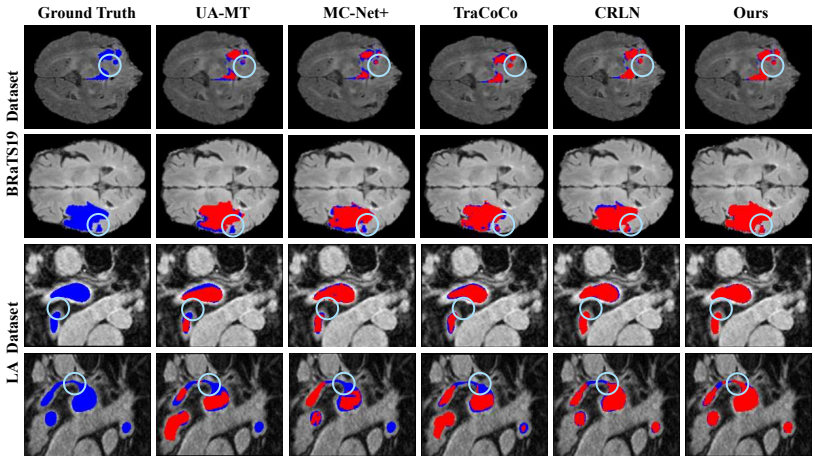

Fig. 2. Visualization comparison on LA and BRaTS19 datasets with 20% labeled ratio. (Blue: ground truth, Red: prediction results)

collectively encompass critical challenges including significant anatomical shape variations, scale discrepancies, and heterogeneous tissue contrast characteristics.

Dataset. The **LA** dataset [26] is a standard benchmark for semi-supervised 3D medical image segmentation. It consists of 100 3D MRI volumes with isotropic resolution $0.625 \times 0.625 \times 0.625$ mm. Following [11], each volume is cropped to $112 \times 112 \times 80$ around the heart region. Data are split into 80/20 for training and validation. The **BraTS19** dataset [16] includes 335 MRI scans with expert-labeled tumor annotations. We conduct whole tumor segmentation on the T2-FLAIR modality, adopting the standard 250/25/60 train/val/test split. All volumes are randomly cropped into $96 \times 96 \times 96$ sub-volumes for efficient training, matching the setup in [11]. The **Pancreas-CT** dataset [5] includes 82 contrast-enhanced abdominal CT scans. Preprocessing follows [11]: voxel values are clipped to $[-125, 275]$ Hounsfield Units, resampled to isotropic $1.0 \times 1.0 \times 1.0$ mm, and cropped to $96 \times 96 \times 96$. Data are split into 62/20 for training and validation.

Metrics. Reported metrics include the Dice coefficient, Jaccard index, Average Surface Distance (ASD), and 95% Hausdorff Distance (95HD).

Implementation Details. We evaluate on three benchmarks — LA, Pancreas-CT, and BraTS 2019 under both 10% and 20% labeled data settings. VNet [18] is used for LA and Pancreas-CT, and 3D U-Net [4] for BraTS19. All models are trained for 200 epochs using stochastic gradient descent with 0.9 momentum, 5×10^{-4} weight decay, and an initial learning rate of 2.5×10^{-3} that decays polynomially with a power of 0.9, with a batch size of 4 comprising 2 labeled and 2 unlabeled examples. Augmentation includes random scaling, rotation, flipping, intensity jitter, Gaussian noise, and ClassMix. EMA teacher decay: 0.99. Proto-

Table 1. Comparative results on three datasets with 10% and 20% labeled scans. Best results are in bold.

Dataset	Method	10% labeled				20% labeled			
		Dice(%)	Jaccard(%)	ASD	95HD	Dice(%)	Jaccard(%)	ASD	95HD
LA	UA-MT [27]	84.25	73.48	3.36	13.84	88.78	80.05	2.27	7.41
	LG-ER [6]	85.24	74.69	3.77	14.99	89.50	81.10	2.08	7.33
	URPC [13]	85.01	74.36	3.96	15.37	88.74	79.93	3.66	12.73
	SASSNet [10]	86.65	76.69	4.17	14.69	89.02	80.45	2.98	8.83
	ASE-Net [8]	87.83	78.45	2.17	9.86	90.29	82.76	1.64	7.18
	TAC [3]	84.73	74.38	2.72	11.45	87.75	78.60	2.04	9.45
	DCR [12]	89.05	80.27	2.84	9.72	91.21	83.54	1.92	8.03
	CAC4SSL [9]	89.92	81.77	2.04	6.38	90.55	82.84	1.71	5.98
	RCPS [28]	90.71	83.10	2.04	7.91	91.19	83.87	1.78	6.34
	TraCoCo [11]	89.78	81.58	1.98	6.59	91.44	84.30	1.79	5.62
	CRLN [22]	91.62	83.65	1.76	4.91	91.95	85.15	**1.45**	**4.33**
	Ours	**91.91**	**83.81**	**1.72**	**4.78**	**92.25**	**85.66**	1.46	4.46
Pancreas CT	UA-MT [27]	66.44	52.02	3.03	17.04	72.10	62.62	2.43	10.84
	SASSNet [10]	68.97	54.29	1.96	18.83	76.39	63.17	1.42	11.06
	URPC [13]	73.53	59.44	7.85	22.57	80.02	67.30	1.98	8.51
	MCCauSSL [17]	72.89	58.06	4.37	14.19	80.92	68.26	1.53	8.11
	CAC4SSL [9]	74.19	59.82	3.01	15.36	80.43	67.85	2.25	12.05
	RCPS [28]	76.81	63.07	2.90	15.43	81.62	69.25	2.00	7.20
	MC-Net+ [25]	69.44	55.16	3.74	16.72	78.95	66.37	1.79	8.78
	MLRPL [19]	75.93	62.12	**1.54**	9.07	81.53	69.95	1.33	6.81
	CRLN [22]	81.24	68.84	3.15	8.38	83.52	69.22	1.36	5.68
	Ours	**81.64**	**69.02**	2.92	**8.33**	**83.65**	**69.57**	**1.28**	**5.01**
BraTS19	UA-MT [27]	84.64	74.76	2.36	10.47	85.32	75.93	1.98	8.68
	SASSNet [10]	84.73	74.89	2.44	9.88	85.64	76.33	2.04	9.17
	LG-ER [6]	84.75	74.97	2.21	9.56	85.67	76.36	1.99	8.92
	URPC [13]	84.53	74.60	2.55	9.79	85.38	76.14	1.87	8.36
	MC-Net+ [25]	84.96	75.14	2.36	9.45	86.02	76.98	1.98	8.74
	SDC-SSL [7]	84.76	75.11	1.95	11.29	85.45	76.23	1.96	7.61
	MLRPL [19]	84.29	74.74	2.55	9.57	85.47	76.32	2.00	7.76
	BCP [2]	85.14	76.01	2.88	9.89	86.13	77.24	2.06	8.99
	TraCoCo [11]	85.79	76.49	1.72	7.03	86.69	77.69	1.93	8.04
	CRLN [22]	86.73	77.85	1.67	7.15	87.17	78.19	1.62	6.79
	Ours	**86.77**	**77.86**	**1.47**	**5.61**	**87.71**	**78.57**	**1.20**	**4.97**

type settings: $L = 16$, $d_q = 64$, $\tau = 0.07$. Loss weights: $\lambda_{\text{cons}} = 0.5$, $\lambda_{\text{sam}} = 0.1$, $\lambda_{\text{pc}} = 0.1$, with $\lambda_{\text{loc}} = 1.0$, $\lambda_{\text{glb}} = 0.5$, $\lambda_{\text{hard}} = 0.5$ for prototype calibration.

3.2 Comparison with State-of-the-Art Methods

Table 1 summarises quantitative results on LA, Pancreas-CT, and BraTS19 datasets under 10% and 20% labeled settings. Our method consistently achieves top-ranked Dice and Jaccard scores across all benchmarks. On LA, it sets new state-of-the-art performance at both labeling ratios, surpassing CRLN [22] by 0.29% Dice (10% labels) and notably reducing surface distances. On the challenging Pancreas-CT dataset, our approach yields substantial gains over previous methods, improving Dice by 12.2% over MC-Net+ [25] under 10% labels, demonstrating the advantage of prototype guidance for complex anatomies. On BraTS19, our method achieves the best Dice, Jaccard, and boundary error measures; against CRLN (20% labels), it improves Dice by 0.54% and reduces HD95 by 1.82 mm, indicating superior handling of tumor heterogeneity. Qualitative

Table 2. Ablation study on LA and BraTS19 datasets (20% labeled). Components: SAM external teacher, Prototype Calibration (Proto), Prototype Contrastive Learning (Contr.), Entropy Guidance (Ent.).

Setting	Components				Metrics (LA / BraTS19)			
	SAM	Proto	Contr.	Ent.	Dice(%)↑	Jaccard(%)↑	ASD↓	95HD↓
Left Atrium (LA)								
Baseline (MT)	×	×	×	×	86.24	78.43	3.42	12.79
+ SAM	✓	×	×	×	89.62	81.41	1.86	6.04
+ Proto	✓	✓	×	×	91.94	85.14	1.62	4.85
+ Contr.	✓	✓	✓	×	92.02	84.58	1.51	4.62
+ Ent. (Full)	✓	✓	✓	✓	**92.25**	**85.66**	**1.46**	**4.46**
BraTS19								
Baseline (MT)	×	×	×	×	85.29	75.26	2.78	8.54
+ SAM	✓	×	×	×	86.75	76.82	1.82	7.84
+ Proto	✓	✓	×	×	87.22	77.73	1.45	5.25
+ Contr.	✓	✓	✓	×	86.62	77.01	1.24	5.02
+ Ent. (Full)	✓	✓	✓	✓	**87.71**	**78.57**	**1.20**	**4.97**

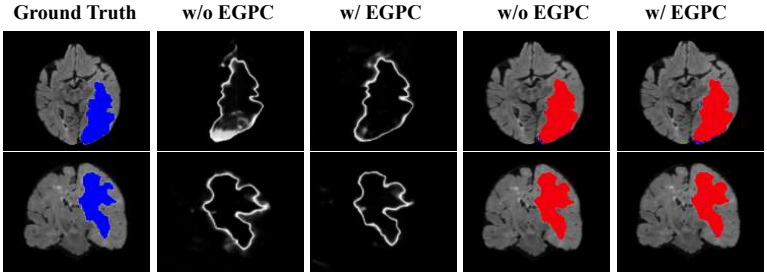

Fig. 3. Visual comparison of the EGPC module on the BraTS19 dataset with 20% labeled data. Compared to the baseline, introducing EGPC sharpens confidence maps and yields more precise boundary predictions.

examples in Fig. 2 further confirm fewer false positives and better structural preservation.

3.3 Ablation Studies

We conduct systematic ablations on LA and BraTS19 with 20% labeled data, using the MT baseline. Table 2 shows progressive improvements. Adding the frozen SAM-Med3D teacher (+SAM) improves Dice on LA from 86.24% to 89.62% and nearly halves ASD, confirming SAM's strong shape priors [14]. The prototype calibration module (+Proto) further boosts Dice to 91.94% on LA and 87.22% on BraTS19, calibrating pseudo-labels via residual fusion to mitigate confirmation bias [1]. Prototype contrastive learning (+Contr.) reduces ASD to 1.51 on LA and 1.24 on BraTS19, further sharpening boundaries. However, it slightly lowers Jaccard on LA and more noticeably Dice on BraTS19, as strict feature

constraints may suppress prediction diversity on heterogeneous tumor regions. The full entropy guidance (+Ent.) achieves the best results: 92.25% Dice on LA and 87.71% on BraTS19. Fig. 3 shows that entropy guidance sharpens confidence maps and produces more precise boundaries.

4 Conclusion

This paper proposes a unified Dual-Teacher Knowledge Distillation framework for semi-supervised 3D medical image segmentation, aiming to address two key challenges: confirmation bias and insufficient anatomical prior knowledge. By integrating an internal EMA-based teacher, an external prior teacher, and a learnable prototype-guided calibration module, the method significantly improves segmentation accuracy, especially in challenging regions with low contrast and ambiguous boundaries.

Acknowledgments. Supported by the Natural Science Foundation of Heilongjiang Province of China (PL2025F041).

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

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2020)
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11514–11524 (2023)
3. Chen, J., Zhang, J., Debattista, K., Han, J.: Semi-supervised unpaired medical image segmentation through task-affinity consistency. IEEE Transactions on Medical Imaging **42**(3), 594–605 (2022)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
5. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging **26**(6), 1045–1057 (2013)
6. Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K.S., Qin, J.: Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 562–571. Springer (2020)
7. Lei, T., Liu, H., Wan, Y., Li, C., Xia, Y., Nandi, A.K.: Shape-guided dual consistency semi-supervised learning framework for 3-d medical image segmentation. IEEE Transactions on Radiation and Plasma Medical Sciences **7**(7), 719–731 (2023)

8. Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K.: Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE transactions on medical imaging **42**(5), 1265–1277 (2022)
9. Li, L., Lian, S., Luo, Z., Wang, B., Li, S.: Contour-aware consistency for semi-supervised medical image segmentation. Biomedical Signal Processing and Control **89**, 105694 (2024)
10. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 552–561. Springer (2020)
11. Liu, Y., Tian, Y., Wang, C., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Translation consistent semi-supervised segmentation for 3d medical images. IEEE Transactions on Medical Imaging (2024)
12. Lu, S., Yan, Z., Chen, W., Cheng, T., Zhang, Z., Yang, G.: Dual consistency regularization with subjective logic for semi-supervised medical image segmentation. Computers in Biology and Medicine **170**, 107991 (2024)
13. Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D.N., Zhang, S.: Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Medical Image Analysis **80**, 102517 (2022)
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, S., Du, H., An, Y., Wang, J., Wang, Q., Wu, H., Lawlor, A., Dong, R.: Deep learning approaches for medical imaging under varying degrees of label availability: A comprehensive survey. arXiv preprint arXiv:2504.11588 (2025)
16. 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)
17. Miao, J., Chen, C., Liu, F., Wei, H., Heng, P.A.: Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21426–21437 (2023)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
19. 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)
20. 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)
21. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: Towards general-purpose segmentation models for volumetric medical images. arxiv 2023. arXiv preprint arXiv:2310.15161
22. Wang, Y., Song, K., Liu, Y., Ma, S., Yan, Y., Carneiro, G.: Leveraging labelled data knowledge: A cooperative rectification learning network for semi-supervised 3d medical image segmentation. Medical Image Analysis **101**, 103461 (2025)
23. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Boundary-aware prototype in semi-supervised medical image segmentation. IEEE Transactions on Image Processing (2024)
24. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Med-

ical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
25. 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)
26. 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)
27. 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)
28. Zhao, X., Qi, Z., Wang, S., Wang, Q., Wu, X., Mao, Y., Zhang, L.: Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation. IEEE Journal of Biomedical and Health Informatics **28**(1), 251–261 (2023)