Texture-Refined Probabilistic Prototype Learning for Semi-Supervised Medical Image Segmentation

Liangjie Wang*, Ning Jiang*,✉, Xiaoqi Kuang*, Wanli Dong*
*School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China
✉ Email: jiangning@swust.edu.cn

***Abstract*—Given the scarcity of expert annotations, semi-supervised learning has become standard practice for medical image segmentation. The scarcity of expert annotations in medical imaging necessitates semi-supervised learning approaches. In semi-supervised medical image segmentation, existing methods typically impose point-wise consistency regularization or contrastive learning on individual voxels. However, such voxel-centric paradigm inherently lacks global structural constraints over class distributions, thus failing to disambiguate ambiguous boundaries with intra-class heterogeneity. While prototype learning provides distribution-level guidance, its semi-supervised application remains constrained by the reliability challenge of learning prototypes from noisy pseudo-labels. To this end, we present TR-PC, a framework that performs feature purification and semantic decoupling in probabilistic projection space. It first introduces a Texture-Refined Probabilistic Head (TR-Head) to suppress feature noise in 3D convolutions via local contextual modeling and dimension-balanced projection, thereby mapping raw voxels into high-fidelity multivariate Gaussian distributions. Afterwards, leveraging predictive uncertainty as a dynamic discriminator, it explicitly decouples each class center into cohesive core prototypes and dispersed boundary prototypes, and enforces divide-and-conquer likelihood matching. Therefore, the model is encouraged to infer and delineate highly challenging ambiguous regions via these decoupled core-boundary distributional conditions. TR-PC achieves state-of-the-art performance on three authoritative SSMIS benchmarks (LA, ACDC, and Pancreas-CT), with particularly significant improvements in boundary metrics (HD95).**

***Index Terms*—semi-supervised segmentation, data uncertainty, probabilistic representation, prototype learning**

I. Introduction

Medical image segmentation plays a crucial role in computer-aided diagnosis [1], [4]. While fully supervised methods achieve high precision, they require large amounts of annotated data. However, due to the high cost of pixel-level annotation and the demand for specialized expertise, acquiring large-scale high-quality annotated datasets remains extremely challenging. Facing the challenge of label scarcity, semi-supervised learning (SSL) has become a research hotspot by combining a small amount of annotated data with large numbers of unlabeled medical images [2], [18].

Existing semi-supervised medical image segmentation methods can be broadly categorized into several types. Pseudo-labeling methods generate pseudo-labels for unlabeled data and iteratively train models [3], but are prone to confirmation bias, leading to error accumulation. Co-training methods leverage multiple models for complementarity [29], yet maintaining inter-model diversity remains challenging. Consistency-based methods constrain the model to produce consistent outputs under perturbations [13]. However, limited transformation applicability in medical imaging leads to restricted perturbation design space, causing distribution shift and imprecise segmentation [19]. Contrastive methods treat voxels as positive or negative instances and minimize embedding distance between same-class samples [9]. However, intrinsic ambiguity in medical imaging poses challenges due to unclear boundaries and expert variability [31], while forcibly mapping voxels into single coordinates in feature space inherently lacks global structural constraints over class distributions. Consequently, this voxel-centric pattern may lack contextual information and be insufficient in segmenting ambiguous regions [5], [7].

Traditional probabilistic modeling approaches (e.g., variational inference, Monte Carlo Dropout) can quantify uncertainty but rely on deterministic features and lack distribution-level semantic modeling [8], [17], [21]. In recent years, prototypical learning naturally provides global class guidance [20], [30]. However, its application in semi-supervised settings remains constrained by learning prototypes from noisy pseudo-labels [27]. Furthermore, convolutional operations in 3D medical images often introduce feature noise [6], undermining prototype quality. More critically, traditional single-prototype representations cannot distinguish between core regions and ambiguous boundaries within each class [12], leading to inference failure in high-uncertainty regions. To address these issues, we propose TR-PC, which achieves feature purification and semantic decoupling via texture-refined probabilistic representation and uncertainty-driven dual-prototype learning.

The main contributions of this paper are summarized as follows:

- We propose a Texture-Refined Probabilistic Head (TR-Head) that maps voxels into high-fidelity Gaussian distributions via local contextual modeling, effectively suppressing feature noise in 3D convolutions.
- We design an uncertainty-driven dual-prototype decoupling strategy that decomposes each class into cohesive core and dispersed boundary prototypes, enabling divide-and-conquer likelihood matching for intra-class heterogeneity.
- Extensive experiments on three SSMIS benchmarks (LA, ACDC, Pancreas-CT) demonstrate that TR-PC achieves state-of-the-art performance with significant improvements in boundary metrics (HD95).

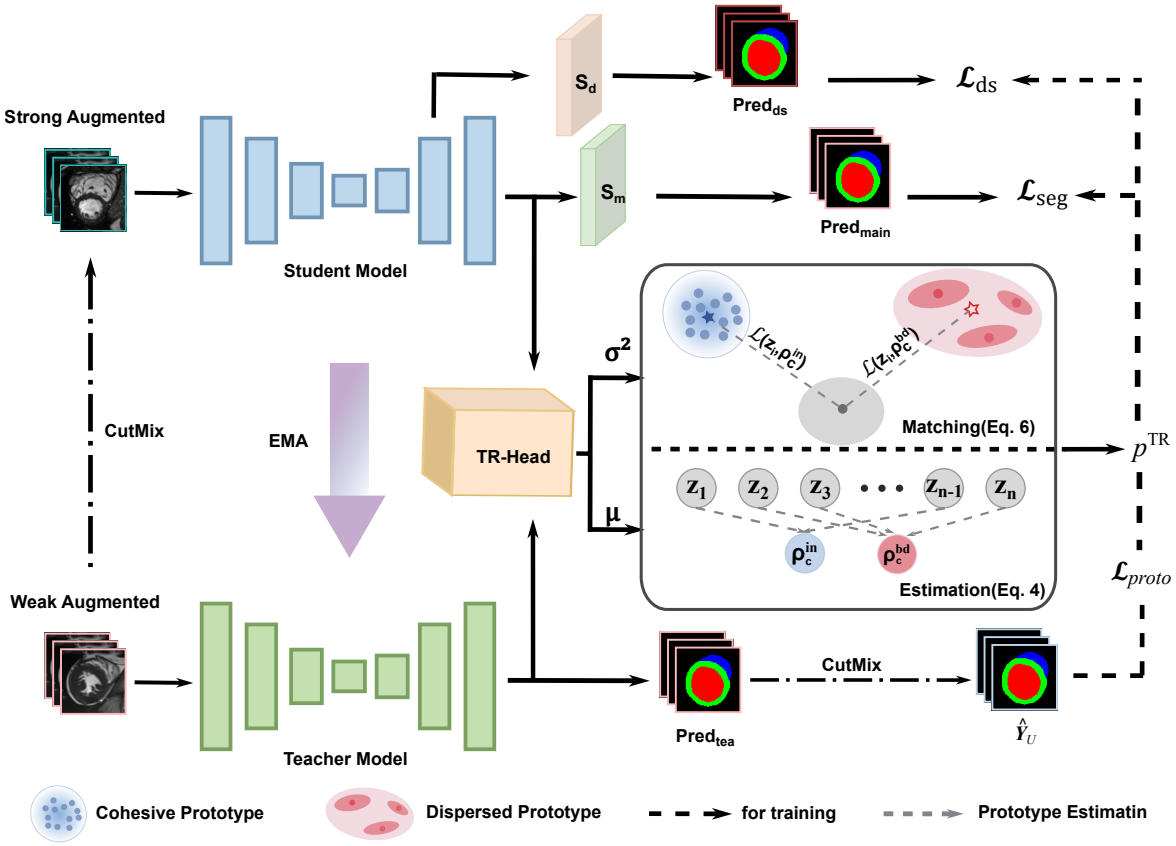

Fig. 1. Illustration of the TR-PC framework. Labeled and unlabeled patches are processed by a shared backbone to extract multi-scale features. The TR-Head refines these features into probabilistic embeddings, which are then adaptively partitioned into cohesive core and dispersed boundary prototypes using uncertainty-driven decoupling. The model is encouraged to infer challenging boundaries via maximum likelihood response (p^{TR}). Consistency regularization ($\mathcal{L}_{con}$) and multi-scale deep supervision ($\mathcal{L}_{ds}$) are jointly optimized to ensure robust convergence under sparse supervision.

II. Methodology

A. Model Overview

Given labeled data $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_l}$ and unlabeled data $\mathcal{D}_u = \{\mathbf{u}_i\}_{i=1}^{N_u}$ with $N_u \gg N_l$, where $\mathbf{x}_i \in \mathbb{R}^{H\times W\times D}$ denotes the 3D medical image and $\mathbf{y}_i \in \{0,1\}^{C\times H\times W\times D}$ denotes the ground-truth mask, we propose a semi-supervised framework as illustrated in Fig. 1. The framework adopts the Mean-Teacher architecture with a student model and a teacher model constrained by Exponential Moving Average (EMA). Training is driven by supervised loss on labeled data and consistency regularization on unlabeled data with strong augmentation.

The input 3D images are fed into an encoder-decoder network to extract multi-scale features, which are then passed to two complementary branches: (1) the conventional segmentation branch with a main head S_m and deep supervision heads S_d, and (2) the probabilistic prototype branch for uncertainty-aware prototype learning, detailed in the following subsections.

B. Texture-Refined Probabilistic Representation

In existing methods, voxels are typically mapped into deterministic point embeddings, which cannot explicitly quantify predictive uncertainty and are susceptible to high-frequency noise from 3D convolutions. Therefore, we model voxel representations as probabilistic distributions to capture uncertainty and suppress feature noise.

We model voxel representations as multivariate Gaussian distributions to explicitly describe predictive uncertainty in medical images:

$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)), \tag{1}$$

where $\mathrm{diag}(\boldsymbol{\sigma}_i^2)$ denotes the diagonal covariance matrix, the mean vector $\boldsymbol{\mu}_i$ represents the semantic position of the voxel in feature space, and the variance vector $\boldsymbol{\sigma}_i^2$ quantifies the aleatoric uncertainty.

To produce distribution parameters, we introduce the Texture-Refined Probabilistic Head (TR-Head). Its mapping process is defined as:

$$(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2) = \Psi_{\mathrm{TR}}(\mathbf{f}_i; \theta), \tag{2}$$

where θ are the learnable parameters. TR-Head utilizes $1\times1\times1$ convolutions for dimension-balanced projection, followed by $3\times3\times3$ convolutional blocks for local contextual modeling, ensuring spatial continuity in uncertainty estimation. The variance branch adopts Softplus activation to ensure strict positive-definiteness and smooth gradients.

When estimating class prototypes, traditional pixel-agnostic averaging strategies are susceptible to interference from high-uncertainty pixels. To this end, we adopt Bayesian estimation to compute the prototype distribution center $\mathbf{P}_c$ for class c:

$$\mathbf{P}_c = \frac{\sum_{i\in\Omega_c} w_i \boldsymbol{\mu}_i}{\sum_{i\in\Omega_c} w_i}, \quad w_i = \left(\frac{1}{K}\sum_{k=1}^{K}\sigma_{i,k}^2 + \epsilon\right)^{-1}, \tag{3}$$

where Ω_c denotes the set of voxels belonging to class c, w_i is the confidence weight computed based on variance, and ϵ is a small constant. This mechanism assigns greater weights to high-confidence pixels, establishing robust feature beacons.

C. Uncertainty-Driven Dual-Prototype Learning

Traditional single-prototype representations model each class as a single distribution, ignoring intra-class heterogeneity in medical images. Specifically, organ core regions exhibit stable features and high confidence, while boundary regions show high uncertainty due to anatomical transitions. Therefore, we decompose class centers into cohesive core prototypes and dispersed boundary prototypes.

We establish a discrimination criterion based on the average variance $\bar{\sigma}_i^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_{i,k}^2$ for each voxel, where K is the feature dimension. Using the median threshold τ within the current batch, we partition the voxel set Ω_c into a core subset $\Omega_c^{\text{in}} = \{i \in \Omega_c | \bar{\sigma}_i^2 \le \tau\}$ and a boundary subset $\Omega_c^{\text{bd}} = \{i \in \Omega_c | \bar{\sigma}_i^2 > \tau\}$. Following the Bayesian estimation framework in PPC [30], we estimate each sub-prototype distribution as $\mathbf{P}_c^m \sim \mathcal{N}(\boldsymbol{\mu}_c^m, \text{diag}((\boldsymbol{\sigma}_c^m)^2))$ for $m \in \{\text{in}, \text{bd}\}$:

$$\boldsymbol{\mu}_c^m = \frac{\sum_{i\in\Omega_c^m} w_i \boldsymbol{\mu}_i}{\sum_{i\in\Omega_c^m} w_i}, \quad (\boldsymbol{\sigma}_c^m)^2 = \left(\sum_{i\in\Omega_c^m}\frac{1}{\boldsymbol{\sigma}_i^2 + \epsilon}\right)^{-1}, \tag{4}$$

where $w_i = (\boldsymbol{\sigma}_i^2 + \epsilon)^{-1}$ is the precision-based confidence weight, and ϵ is a small constant for numerical stability. This formulation assigns higher weights to high-confidence voxels, while the prototype variance $(\boldsymbol{\sigma}_c^m)^2$ inversely reflects the aggregated precision of observations.

To measure the proximity between voxel distributions and prototype distributions, we define a log-likelihood matching operator. Given a voxel-level random variable $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2))$ and a sub-prototype $\mathbf{P}_c^m$, the similarity is computed as:

$$\mathcal{L}(\mathbf{z}_i, \mathbf{P}_c^m) = -\frac{1}{2}\sum_{k=1}^{K}\left[\frac{(\mu_{i,k} - \mu_{c,k}^m)^2}{\sigma_{i,k}^2 + (\sigma_{c,k}^m)^2} + \ln\left(\sigma_{i,k}^2 + (\sigma_{c,k}^m)^2\right)\right], \tag{5}$$

where $\mu_{i,k}$, $\sigma_{i,k}^2$ are the voxel distribution parameters in the k-th feature dimension, and $\mu_{c,k}^m$, $(\sigma_{c,k}^m)^2$ are the corresponding prototype parameters. Based on this, we define the similarity between voxel i and class c as the Maximum Likelihood Response (MLR):

$$S_i^c = \max_{m\in\{\text{in},\text{bd}\}} \mathcal{L}(\mathbf{z}_i, \mathbf{P}_c^m). \tag{6}$$

Finally, the predictive probability that voxel $\mathbf{x}_i$ belongs to class c is given by:

$$p^{\text{TR}}(y_i = c|\mathbf{x}_i) = \frac{\exp(S_i^c/T)}{\sum_{j=1}^{C}\exp(S_i^j/T)}, \tag{7}$$

where T is the temperature hyperparameter.

D. Overall Objective

For each sample from $\mathcal{D}_l$ or $\mathcal{D}_u$, we obtain multi-branch predictions through the shared backbone network and functional heads:

$$\{\text{Pred}_{\text{main}}, \text{Pred}_{\text{ds}}, P_p\} = \mathcal{F}(\mathbf{x}_i; \theta), \tag{8}$$

where $\text{Pred}_{\text{main}}$ is the main segmentation output, Pred_{ds} denotes the deep supervision output for multi-scale gradient guidance, and P_p is the dual-prototype probabilistic prediction. For labeled data, we define the supervised loss as:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{seg}}(\text{Pred}_{\text{main}}, \mathbf{y}_i) + \mathcal{L}_{\text{ds}}(\text{Pred}_{\text{ds}}, \mathbf{y}_i) + \mathcal{L}_{\text{proto}}(P_p, \mathbf{y}_i), \tag{9}$$

where $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{ds}}$ are the segmentation and deep supervision losses, and $\mathcal{L}_{\text{proto}}$ aggregates the negative log-likelihood in (5) over all voxels. For unlabeled samples, we formulate the unsupervised loss using a reliability mask based on average predictive variance $\bar{\sigma}^2 = \frac{1}{HWD}\sum_i \|\boldsymbol{\sigma}_i^2\|$:

$$\mathcal{L}_{\text{unsup}} = \mathbb{1}(\bar{\sigma}^2 < \gamma) \cdot \left[\ell(\text{Pred}_{\text{main}}, \hat{\mathbf{y}}_u) + \ell(P_p, \hat{\mathbf{y}}_u)\right], \tag{10}$$

where $\mathbb{1}(\cdot)$ is the indicator function with threshold γ, $\ell(\cdot)$ is the average of cross-entropy and Dice loss, and $\hat{\mathbf{y}}_u$ denotes the pseudo-labels generated by the teacher model. Finally, the overall optimization objective is:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda(t)\mathcal{L}_{\text{unsup}}, \tag{11}$$

where $\lambda(t)$ is the time-dependent consistency weight that gradually increases during training, with specific schedule determined empirically for each dataset.

III. Experiments

A. Datasets and Implementation Details

To evaluate the effectiveness of the proposed TR-PC, we conduct experiments on three widely used medical image segmentation datasets: LA [14], Pancreas-CT [23], and ACDC [15]. The data splits strictly follow the common experimental protocols [9], [13], [25].

For 3D datasets (LA and Pancreas-CT), we use V-Net [4] as the backbone; for the 2D dataset (ACDC), we use U-Net [1]. We employ deep supervision with auxiliary heads at three decoder levels with loss weights of 0.5, 0.25, and 0.125. We train for 18k iterations on 3D datasets and 30k iterations on the 2D dataset, using the SGD optimizer with an initial learning rate of 0.01 and a Poly learning rate scheduler. Random flipping, rotation, and cropping are used as weak augmentation, while CutMix [16] is applied as strong augmentation. During training, LA scans are cropped to $112\times112\times80$, and Pancreas-CT scans to $96\times96\times96$. For ACDC, we perform slice-by-slice

TABLE I
QUANTITATIVE COMPARISON ON THE LA DATASET. V-NET (ALL) DENOTES THE FULLY SUPERVISED UPPER BOUND.

Method	Scans used (Labeled / Unlabeled)							
	8 (10%) / 72 (90%)				16 (20%) / 64 (80%)			
	Dice↑	Jc↑	95HD↓	ASD↓	Dice↑	Jc↑	95HD↓	ASD↓
V-Net (All)	91.14	83.82	5.75	1.52	91.14	83.82	5.75	1.52
V-Net [4]	79.99	68.12	21.11	5.48	86.50	76.30	10.20	3.50
DTC [9]	84.55	73.91	13.80	3.69	84.55	73.91	13.80	3.69
URPC [19]	83.37	71.99	17.91	4.41	83.37	71.99	17.91	4.41
SS-Net [12]	86.56	76.61	12.76	3.02	86.56	76.61	12.76	3.02
CAML [22]	89.62	81.28	8.76	2.02	89.62	81.28	8.76	2.02
UniMatch [24]	89.09	80.47	12.50	3.59	89.09	80.47	12.50	3.59
BCP [25]	89.55	81.22	7.10	**1.69**	89.62	81.31	6.81	1.76
TraCoCo [26]	89.29	80.82	6.92	2.28	90.94	83.47	5.49	1.79
MLRPL [27]	89.86	81.68	6.91	1.85	91.02	83.62	5.78	1.66
Ours	**90.70**	**83.06**	**6.37**	1.72	**91.77**	**84.60**	**5.42**	**1.59**

TABLE II
QUANTITATIVE COMPARISON ON THE ACDC DATASET. U-NET (ALL) DENOTES THE FULLY SUPERVISED UPPER BOUND.

Method	Scans used (Labeled / Unlabeled)			
	7 (10%) / 63 (90%)			
	Dice↑	Jc↑	95HD↓	ASD↓
U-Net (All)	91.44	84.59	4.30	0.99
U-Net [1]	79.99	68.11	9.35	2.70
URPC [19]	83.10	72.41	4.84	1.53
SS-Net [12]	86.78	77.67	6.07	1.40
DMD [28]	87.52	78.62	4.81	1.60
BCP [25]	88.84	80.62	3.98	1.17
PPC [30]	89.20	80.99	2.83	0.90
URCA [31]	87.86	-	4.21	1.36
Ours	**89.56**	**82.03**	**2.30**	**0.73**

inference with crops of 256×256 and stack the 2D predictions to reconstruct the 3D volume.

For pseudo-label generation on unlabeled data, TR-Head maps features to a 64-dimensional probabilistic space. The percentile threshold γ for the reliability mask is set to 95%. The time-dependent consistency weight $\lambda(t)$ increases from 0 to 2 via a sigmoid ramp-up schedule. Prototypes are frozen after 10,000 iterations to prevent feature drift. During inference, a Gaussian-weighted sliding window strategy is used, with stride fixed at $18 \times 18 \times 4$ for LA and $16 \times 16 \times 4$ for Pancreas-CT. Evaluation metrics include Dice coefficient (Dice), Jaccard index (Jac), 95% Hausdorff distance (95HD), and average surface distance (ASD). All experiments are conducted on an NVIDIA 3090 GPU with a fixed random seed.

B. Comparison with State-of-the-Art Methods

We comprehensively compare the proposed TR-PC with a series of state-of-the-art semi-supervised segmentation methods [11], [13], [25].

LA dataset (Table I). With 10% labeled data, TR-PC achieves 90.70% Dice and 6.37 95HD, outperforming MLRPL [27] and BCP [25]. With 20% labeled data, its Dice further improves to 91.77%, 0.83% higher than the 2024 work TraCoCo [26]. The performance gain is attributed to the dual-prototype decoupling strategy's targeted modeling of uncertain regions.

TABLE III
QUANTITATIVE COMPARISON ON THE PANCREAS-CT DATASET. V-NET (ALL) DENOTES THE FULLY SUPERVISED UPPER BOUND.

Method	Scans used (Labeled / Unlabeled)			
	6 (10%) / 56 (90%)			
	Dice↑	Jc↑	95HD↓	ASD↓
V-Net (All)	77.84	64.78	8.92	3.73
V-Net [4]	42.56	36.71	42.61	11.68
DTC [9]	59.54	45.61	16.53	3.12
MC-Net [10]	69.07	54.36	14.53	2.28
MC-Net+ [11]	70.00	55.66	16.03	3.87
UniMatch [24]	69.90	55.13	12.94	3.56
RCPS [29]	76.62	-	16.32	3.01
TraCoCo [26]	79.22	66.04	8.46	**2.57**
Ours	**80.31**	**67.54**	**7.34**	2.77

TABLE IV
ABLATION STUDY OF THE CORE COMPONENTS IN THE PROPOSED TR-PC FRAMEWORK. ✓ INDICATES THE MODULE IS UTILIZED. TH: TEXTURE-REFINED HEAD, DP: DUAL-PROTOTYPE DECOUPLING, DS: DEEP SUPERVISION.

Module			Metrics			
TH	DP	DS	Dice↑	Jac↑	95HD↓	ASD↓
✓			88.14	79.79	6.89	1.95
✓	✓		88.95	81.95	7.64	1.79
✓		✓	89.43	81.48	6.73	1.81
	✓	✓	89.11	80.93	7.55	1.92
✓	✓	✓	**90.70**	**83.06**	**6.37**	**1.72**

ACDC dataset (Table II). In the 2D multi-class task, TR-PC achieves 89.56% Dice and 2.30 95HD. The significant reduction in 95HD and ASD metrics demonstrates the smoothing effect of the texture refinement mechanism in handling complex inter-slice boundaries, reducing isolated mis-segmentation regions.

Pancreas-CT dataset (Table III). In the low-contrast pancreas segmentation task, TR-PC achieves 80.31% Dice with 10% labeled data, a 1.09% improvement over TraCoCo [26], validating that uncertainty-driven representation learning can enhance the model's semantic discrimination capability at highly challenging tissue boundaries.

C. Further Analysis

Ablation study of TR-PC components. We first evaluate the effectiveness of each core component in TR-PC as shown in Table IV. Here, **TH**, **DP**, and **DS** denote the Texture-Refined Head, the Dual-Prototype Decoupling strategy, and the Multi-scale Deep Supervision, respectively. With **TH** alone, the model achieves 88.14% Dice on the LA dataset by effectively suppressing convolutional noise. When **TH** is combined with

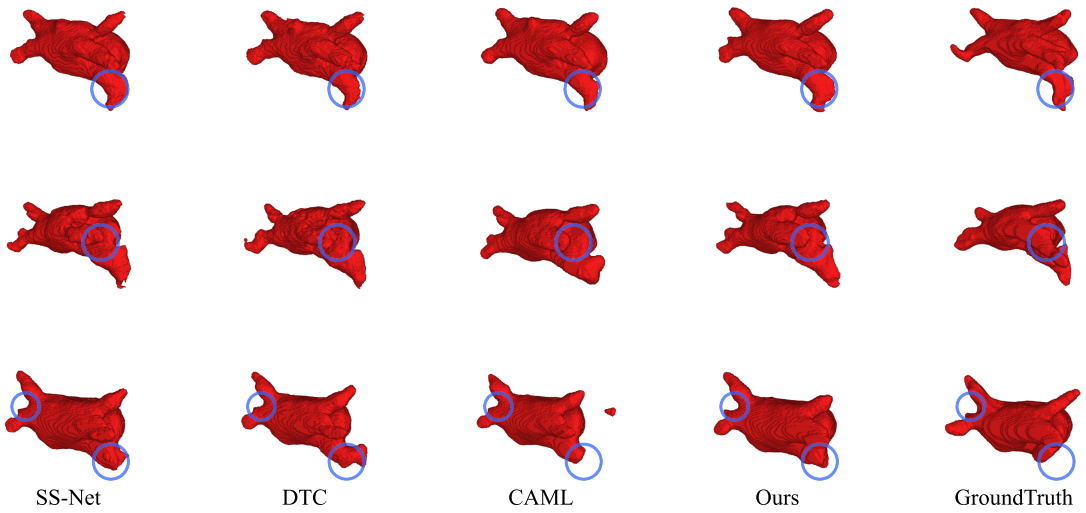

Fig. 2. Visualization examples of several semi-supervised segmentation methods with 10% labeled data and ground truth on the LA dataset.

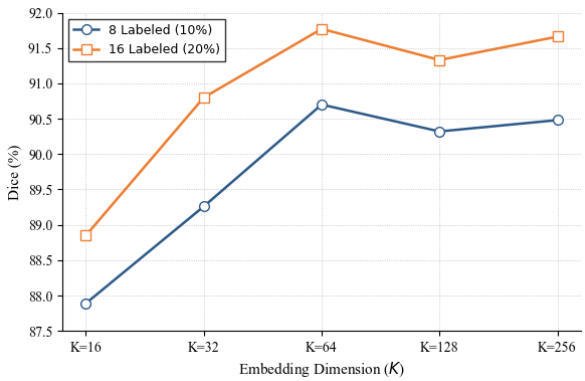

Fig. 3. Impact of the embedding dimension K on the segmentation performance (Dice) under 5% and 10% labeled settings on the LA dataset.

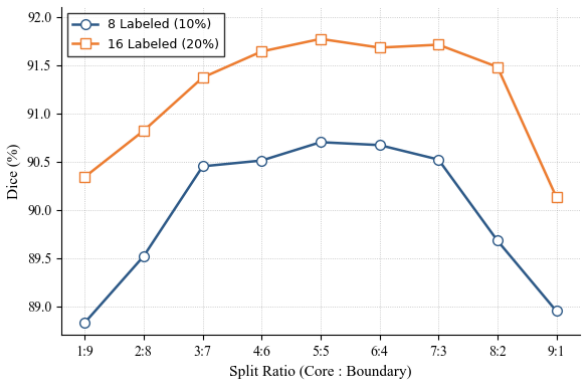

Fig. 4. Impact of different core-boundary split ratios on the segmentation performance (Dice) under 5% and 10% labeled settings on the LA dataset.

DP, the Dice score improves by 0.81%, validating the advantage of the "divide-and-conquer" mechanism in modeling intra-class heterogeneity. Furthermore, the integration of **DS** provides additional multi-scale gradient guidance, which significantly stabilizes the training process and facilitates convergence. The combination of all three components yields the best comprehensive performance, demonstrating their synergistic effect in capturing both global anatomical structures and local boundary details.

Qualitative segmentation results. We present qualitative comparisons with state-of-the-art methods on the LA dataset in Fig. 2. The blue circles highlight challenging regions where TR-PC demonstrates clear advantages. Specifically, compared with DTC [9] and SS-Net [12], our method reduces incomplete segmentation and fragmented regions (rows 1-2). Furthermore, compared with CAML [22], TR-PC achieves superior boundary delineation in ambiguous regions (row 3). These improvements validate the effectiveness of uncertainty-driven dual-prototype learning in modeling intra-class heterogeneity and handling challenging anatomical structures.

Impact of feature dimensionality. As TR-Head maps features to the probabilistic space, we investigate 5 dimensional choices: 16, 32, 64, 128, 256 dimensions. Results are shown in Fig. 3. Compared with 16 dimensions, all other dimensions significantly outperform it, highlighting the importance of sufficient representation capability. 64 and 128 dimensions exhibit similar performance, with 64 dimensions slightly better, demonstrating the robustness of the proposed method. Considering the balance between computational efficiency and representation depth, all experiments adopt the 64-dimensional strategy.

Impact of different split ratios. We evaluate the impact of different core-boundary split ratios in Fig. 4. Overall, TR-PC is relatively stable for the split ratio. For ratio settings,

only extreme splits (3:7 or 7:3) lead to slight Dice drops. This may be attributed to extreme imbalance: the under-sampled prototype becomes unstable under limited supervision, while the over-sampled one absorbs excessive heterogeneous samples beyond its semantic scope, both degrading modeling. The model achieves the highest Dice at 5:5, and we use this value in all experiments. This stability is attributed to the reliability mask pre-filtering noisy voxels, ensuring robustness of prototype estimation.

IV. Conclusion

In this work, we propose TR-PC to enhance the robustness of semi-supervised medical image segmentation through texture-refined probabilistic representation and dual-prototype decoupling. TR-Head suppresses high-frequency noise in 3D convolutions via local context modeling, while the dual-prototype strategy explicitly decomposes semantic categories into cohesive and dispersed prototypes, achieving a "divide-and-conquer" approach for stable structures and ambiguous boundaries. Extensive experiments on three authoritative benchmark datasets (LA, ACDC, Pancreas-CT) demonstrate that TR-PC significantly outperforms existing state-of-the-art methods in metrics such as Dice and 95HD, fully validating its effectiveness and robustness in extremely sparse annotation scenarios.

V. ACKNOWLEDGMENT

This paper is supported by the Central Guidance for Local Science and Technology Development Projects (No. 2023ZYDF074).

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[2] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[3] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. ICML Workshop Challenges Represent. Learn.*, vol. 3, 2013, p. 896.

[4] M. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2016, pp. 565–571.

[5] J. Tang, S. Chen, G. Niu, M. Sugiyama, and C. Gong, "Distribution shift matters for knowledge distillation with webly collected images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 17470–17480.

[6] J. Tang, N. Jiang, H. Zhu, J. T. Zhou, and C. Gong, "Learning student network under universal label noise," *IEEE Trans. Image Process.*, 2024.

[7] J. Tang, S. Chen, G. Niu, H. Zhu, J. T. Zhou, C. Gong, and M. Sugiyama, "Direct distillation between different domains," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 154–172.

[8] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1530–1538.

[9] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 10, 2021, pp. 8801–8809.

[10] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2021, pp. 297–306.

[11] Y. Wu, Z. Ge, D. Zhang, M. Xu, L. Zhang, Y. Xia, and J. Cai, "Mutual consistency learning for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 81, p. 102530, 2022.

[12] Y. Wu, Z. Wu, Q. Wu, Z. Ge, and J. Cai, "Exploring smoothness and class-separation for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2022, pp. 34–43.

[13] L. Yu, S. Wang, X. Li, C. W. Fu, and P. A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019, pp. 605–613.

[14] Z. Xiong et al., "Atrial segmentation challenge 2018: LA segmentation from Gd-enhanced MRI," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart (STACOM)*, 2018.

[15] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2018, pp. 635–643.

[16] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[17] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.

[18] Y. Gao, B. Liu, G. Zhao, Y. Zhang, L. Shi, and M. He, "MCC-Net: Mamba based consistency constraints network for semi-supervised 3D medical image segmentation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2025, pp. 1–8.

[19] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, D. N. Metaxas, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Med. Image Anal.*, vol. 80, p. 102517, 2022.

[20] L. Zhang, R. Wang, Y. Yu, and D. Shen, "LUNA: Learning uncertainty via neighborhood aggregation for medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[21] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016.

[22] S. Gao, Z. Zhang, J. Ma, Z. Li, and S. Zhang, "Correlation-aware mutual learning for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2023, pp. 98–108.

[23] H. R. Roth, L. Lu, A. Farag, H. C. Shin, J. Liu, and S. Turkbey, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 556–564.

[24] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7236–7246.

[25] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 11514–11524.

[26] Y. Liu, Y. Tian, C. Wang, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Translation consistent semi-supervised segmentation for 3D medical images," *IEEE Trans. Med. Imag.*, 2024.

[27] J. Su, Z. Luo, S. Lian, D. Lin, and S. Li, "Mutual learning with reliable pseudo label for semi-supervised medical image segmentation," *Med. Image Anal.*, 2024.

[28] Y. Xie, Y. Yin, Q. Li, and Y. Wang, "Deep mutual distillation for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2023, pp. 540–550.

[29] X. Zhao, Z. Qi, S. Wang, Q. Wang, X. Wu, Y. Mao, and L. Zhang, "Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 1, pp. 251–261, 2024.

[30] Y. Yuan, X. Wang, X. Yang, et al., "Effective semi-supervised medical image segmentation with probabilistic representations and prototype learning," *IEEE Trans. Med. Imag.*, 2024.

[31] C. Qin, Y. Wang, and J. Zhang, "Urca: Uncertainty-based region clipping algorithm for semi-supervised medical image segmentation," *Comput. Methods Programs Biomed.*, 2024.