Multi-Prototype-based Embedding Refinement for Medical Image Segmentation

Yali Bi[1*], Enyu Che[1*], Yinan Chen[1], Yuanpeng He[2], and Jingwei Qu[1†]
[1]College of Computer and Information Science, Southwest University, Chongqing, P.R. China
[2]School of Computer Science, Peking University, Beijing, P.R. China

***Abstract*—Medical image segmentation aims to identify anatomical structures at the voxel-level. Segmentation accuracy relies on distinguishing voxel differences. Compared to advancements achieved in studies of the inter-class variance, the intra-class variance receives less attention. Moreover, traditional linear classifiers, limited by a single learnable weight per class, struggle to capture this finer distinction. To address the above challenges, we propose a Multi-Prototype-based Embedding Refinement method for semi-supervised medical image segmentation. Specifically, we design a multi-prototype-based classification strategy, rethinking the segmentation from the perspective of structural relationships between voxel embeddings. The intra-class variations are explored by clustering voxels along the distribution of multiple prototypes in each class. Next, we introduce a consistency constraint to alleviate the limitation of linear classifiers. This constraint integrates different classification granularities from a linear classifier and the proposed prototype-based classifier. In the thorough evaluation on two popular benchmarks, our method achieves superior performance compared with state-of-the-art methods. Code is available at https://github.com/Briley-byl123/MPER.**

***Index Terms*—Medical image segmentation, multi-prototype, semi-supervised learning**

I. Introduction

Medical image segmentation aims to accurately identify anatomical structures like organs and tumors [1], and has been utilized in various fields, such as disease diagnosis [2]. It is more challenging than natural image segmentation since the data labeling requires medical expertise.

Semi-supervised frameworks are increasingly applied in medical image segmentation due to their advantages in exploiting unlabeled data. These methods explore different optimization strategies for predictions of unlabeled data, including consistency regularization [3, 4], entropy minimization [5, 6], and pseudo-labeling [7, 8]. Nonetheless, the progress in studying the intra-class voxel variance remains limited.

The quality of segmentation highly depends on capturing variations between voxels, as investigated in previous studies. However, these studies focus on the inter-class variance, overlooking the finer distinctions between voxels belonging to the same class. In fact, the intra-class variance offers a more granular perspective for precise segmentation. Moreover, traditional linear classifiers learn a single weight for each class. The class-level weights have limited capacity to distinguish semantic ambiguities between voxels of the same class, thereby also hindering deeper exploration of the intra-class variance.

*Equal contribution. †Corresponding author: qujingwei@swu.edu.cn.

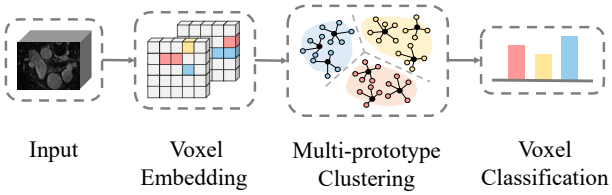


Fig. 1: Multi-prototype-based classification. Voxel embeddings extracted from input medical images are assigned to the most similar prototypes. The resulting clusters help generate more precise segmentation through finer classification.

To address the above challenges, we propose a Multi-Prototype-based Embedding Refinement method for medical image segmentation. Prototype learning selects representative samples to summarize critical features and inherent structures of original data. It has been introduced into semi-supervised segmentation frameworks [9, 10]. A single prototype is generated for each object class via masked average pooling over voxel embeddings. However, a single prototype struggles to describe the inner variations in a class.

Our solution incorporates a multi-prototype-based classification strategy (Fig. 1) into a semi-supervised framework, which rethinks medical image segmentation from the perspective of structural relationships between voxel embeddings. Multiple prototypes form several sub-centers of a class in the voxel embedding space. Their spatial distribution depicts inner structures of the class. Voxels cluster along this distribution based on their feature similarity with prototypes. This cluster approach implicitly reveals the structural relationships among voxels, simultaneously capturing their intra-class variations. To alleviate the limitation of linear classifiers, we design a consistency constraint between a linear classifier and the proposed prototype-based classifier, enhancing segmentation quality by combining their classification granularities. Experiments on two benchmarks demonstrate our method's superior performance, both quantitatively and qualitatively, compared to state-of-the-art (SOTA) methods.

II. Methodology

We implement our method based on the classical Mean Teacher model [11] (Fig. 2). Our model is trained by a two-stage strategy: pre-training and self-training. During the pre-training, we utilize labeled images to train a supervised feature extractor. The Copy-Paste augmentation [12] is adopted to

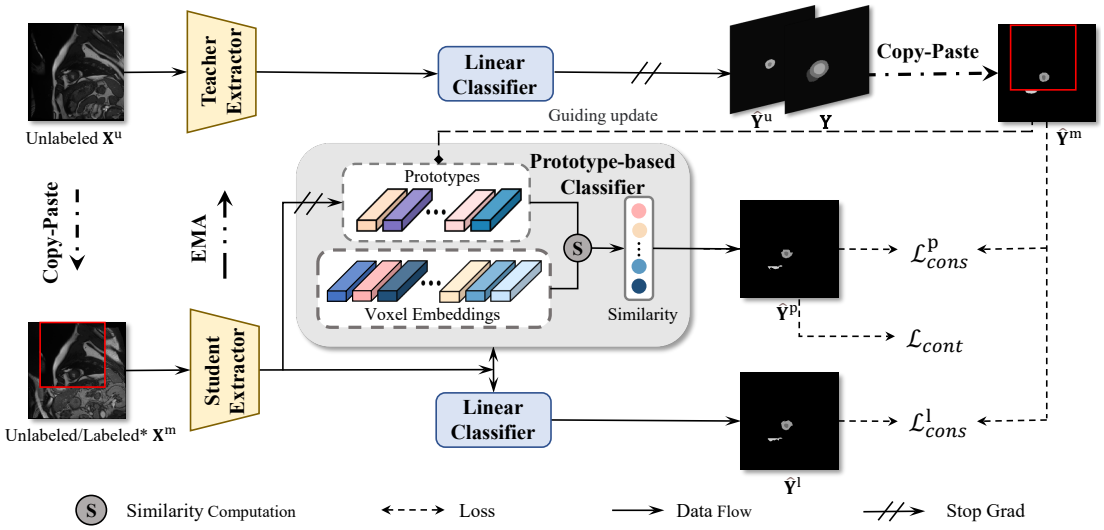


Fig. 2: Overview of our method.

enrich the training data. One image is randomly cropped and pasted onto another to obtain a mixed image.

In the self-training stage, we introduce the multi-prototype-based classification strategy and the consistency constraint. The extractors of the Teacher and Student networks are both initialized using the pre-trained feature extractor. The Teacher network processes an unlabeled image $\mathbf{X}^{\mathrm{u}} \in \mathbb{R}^{H\times W\times D}$ to generate a pseudo-label $\hat{\mathbf{Y}}^{\mathrm{u}} \in \mathbb{R}^{H\times W\times D}$, with weights updated through Exponential Moving Average (EMA) of the Student network. W, H, and D denote the width, height, and depth of the image, respectively. The Student network receives a mixed image $\mathbf{X}^{\mathrm{m}}$ generated by bidirectional Copy-Paste augmentation [13], where the unlabeled image $\mathbf{X}^{\mathrm{u}}$ is randomly cropped and pasted onto a labeled image $\mathbf{X}^{\mathrm{l}}$, or vice versa. For clarity, only the former case is shown in Fig. 2. This augmentation strategy enables the semi-supervised model to better incorporate information from unlabeled data. The embedding $\mathbf{Z} \in \mathbb{R}^{H\times W\times D\times d}$ of the input image $\mathbf{X}^{\mathrm{m}}$ is obtained by the Student extractor, where d denotes the embedding dimensions. Then, the prototype-based classifier predicts the label $\hat{\mathbf{Y}}^{\mathrm{p}}$ based on the embedding $\mathbf{Z}$. Analogously, another label $\hat{\mathbf{Y}}^{\mathrm{l}}$ is estimated by a linear classifier. In addition, we combine the ground-truth label $\mathbf{Y} \in \{0, 1, \ldots, C-1\}^{H\times W\times D}$ of the labeled image $\mathbf{X}^{\mathrm{l}}$ and the pseudo-label $\hat{\mathbf{Y}}^{\mathrm{u}}$ to obtain a mixed label $\hat{\mathbf{Y}}^{\mathrm{m}}$, where C is the number of object classes. This is done similarly to how the input $\mathbf{X}^{\mathrm{m}}$ is processed. $\hat{\mathbf{Y}}^{\mathrm{m}}$ provides the supervision information for the Student network.

A. Multi-Prototype-based Classification

A prototype is defined as a representative of a class in the embedding space. Built upon this, multiple prototypes of the class form different sub-centers, and implicitly divide the embedding space at a finer granularity level. Formally, we define K prototypes for each class. The k-th prototype of the c-th class is denoted as $\mathbf{p}_c^k \in \mathbb{R}^d$, where $c \in \{0, 1, \ldots, C-1\}$ and $k \in \{0, 1, \ldots, K-1\}$.

Prototype-based Classifier. To estimate the category of each voxel, we compute its posterior probability distribution by measuring its similarity with all prototypes. Specifically, the cosine similarities between the voxel embedding $\mathbf{z} \in \mathbb{R}^d$ and all prototypes $\left\{\mathbf{p}_c^k\right\}_{c,k=0}^{C-1,K-1}$ are computed. Next, the highest similarity with K prototypes of each class forms a similarity vector, which is further transformed into the probability distribution $\hat{\mathbf{y}}^{\mathrm{p}} \in \mathbb{R}^C$ by the softmax function:

$$\hat{\mathbf{y}}_c^{\mathrm{p}} = \frac{\exp\left(\mathrm{sim}\left(\mathbf{z}, \mathbf{p}_c^{k^*}\right)/\tau_1\right)}{\sum_{c'=0}^{C-1}\exp\left(\mathrm{sim}\left(\mathbf{z}, \mathbf{p}_{c'}^{k^*}\right)/\tau_1\right)} \tag{1}$$

where $\hat{\mathbf{y}}_c^{\mathrm{p}}$ denotes the probability that the voxel belongs to the c-th class, $\mathrm{sim}(\cdot,\cdot)$ is defined as the cosine similarity function between two vectors, k^* is the index of the prototype with the highest similarity in each category, and τ_1 is a temperature parameter [14]. The class to which the prototype with the highest similarity belongs is the predicted class of the voxel.

Prototype Initialization and Update. Prototypes are initialized as the clustering results of the voxel embeddings through the MiniBatchKMeans algorithm [15]. To update prototypes, we design a momentum-based strategy under the guidance of the mixed label generated from the Teacher network. First, the voxels of the input image $\mathbf{X}^{\mathrm{m}}$ are classified according to the mixed label $\hat{\mathbf{Y}}^{\mathrm{m}}$. Second, the voxels in each class further cluster into different groups based on their similarities with the multiple prototypes, *i.e.*, each voxel is assigned to the prototype with the highest similarity. Finally, the embeddings of voxels in each cluster are utilized to update the prototype by the momentum-based strategy:

$$\mathbf{p}_{t+1} = \eta\mathbf{p}_t + (1-\eta)\,\bar{\mathbf{z}}_t \tag{2}$$

where $\bar{\mathbf{z}}_t$ denotes a mean embedding of the cluster at the t-th iteration, and $\eta \in (0, 1)$ is a momentum coefficient controlling the prototype update rate. To compute the mean embedding, we only keep the embeddings of unlabeled voxels whose pseudo-label confidence exceeds a given threshold α. These are then used together with the embeddings of labeled

voxels to compute the mean embedding. This ensures reliable embeddings are used, effectively integrating information from both labeled and unlabeled data.

B. Consistency Constraint & Training Objective

The traditional linear classifier shows robustness against noise and data incompleteness during the early training phases, thus enhancing training stability and accelerating model convergence, as investigated in previous studies. Specifically, it predicts the probability distribution $\hat{\mathbf{y}}^{\mathrm{l}}$ of each voxel by a linear layer with the softmax function:

$$\hat{\mathbf{y}}_c^{\mathrm{l}} = \frac{\exp\left(\mathbf{w}_c^\top \mathbf{z}\right)}{\sum_{c'=0}^{C-1} \exp\left(\mathbf{w}_{c'}^\top \mathbf{z}\right)} \tag{3}$$

where $\hat{\mathbf{y}}_c^{\mathrm{l}}$ is the probability that the voxel belongs to the c-th class, and $\mathbf{w}_c \in \mathbb{R}^d$ is the learnable weight of the c-th class.

Consistency Constraint. However, the class-level weights limit deeper exploration of the intra-class variance. Therefore, we build a consistency constraint to integrate different classification granularities from the linear classifier and the prototype-based classifier. The constraint is established by two consistency losses used for training our model: linear consistency loss $\mathcal{L}_{cons}^{\mathrm{l}}$ and prototype consistency loss $\mathcal{L}_{cons}^{\mathrm{p}}$. They are both implemented by Cross-Entropy loss $\mathcal{L}_{ce}$ [16] and Dice loss $\mathcal{L}_{dice}$ [17]:

$$\mathcal{L}_{cons}^{*} = \mathcal{L}_{ce}\left(\hat{\mathbf{Y}}^{*}, \hat{\mathbf{Y}}^{\mathrm{m}}\right) + \mathcal{L}_{dice}\left(\hat{\mathbf{Y}}^{*}, \hat{\mathbf{Y}}^{\mathrm{m}}\right) \tag{4}$$

where $*$ denotes the superscript l or p. The mixed label $\hat{\mathbf{Y}}^{\mathrm{m}}$ from the Teacher network acts as a bridge, aligning the prediction results $\hat{\mathbf{Y}}^{\mathrm{p}}$ and $\hat{\mathbf{Y}}^{\mathrm{l}}$ of the two classifiers.

To further optimize the matching relationship between voxels and prototypes, we design a contrastive loss $\mathcal{L}_{cont}$ based on InfoNCE [18]. Each voxel embedding $\mathbf{z}$ is treated as a query vector, and all prototypes $\left\{\mathbf{p}_c^k\right\}_{c,k=0}^{C-1,K-1}$ are viewed as key samples. It is assumed that the embedding $\mathbf{z}$ and its closest prototype $\mathbf{p}^*$ form a positive pair, with the remaining prototypes considered as negative samples for $\mathbf{z}$:

$$\mathcal{L}_{cont} = -\log \frac{\exp\left(\mathbf{z}^\top \mathbf{p}^* / \tau_2\right)}{\sum_{c,k=0}^{C-1,K-1} \exp\left(\mathbf{z}^\top \mathbf{p}_c^k / \tau_2\right)} \tag{5}$$

where τ_2 denotes a temperature parameter that adjusts the sensitivity of the softmax function.

Finally, we combine the three losses nonlinearly to guide voxels to the expected classes:

$$\mathcal{L} = \mathcal{L}_{cons}^{\mathrm{l}} + \lambda(t)\left(\mathcal{L}_{cons}^{\mathrm{p}} + \gamma \mathcal{L}_{cont}\right) \tag{6}$$

where γ is a scaling factor controlling the relative importance of the contrastive loss $\mathcal{L}_{cont}$. $\lambda(t)$ is a dynamic weighting factor governed by a Sigmoid Ramp-up function, increasing the weight of prototype-associated losses from 0 to 1 over the training period. This dynamic weighting strategy ensures initial training stability through linear predictions while progressively enhancing the role of prototype-based predictions.

TABLE I: Comparison of segmentation quality on LA.

Method	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
V-Net [17]	4(5%)	0	52.55	39.60	47.05	9.87
	8(10%)	0	82.74	71.72	13.35	3.26
	80(All)	0	91.47	84.36	5.48	1.51
UA-MT [19]	4(5%)	76(95%)	82.26	70.98	13.71	3.82
SASSNet [20]			81.60	69.63	16.16	3.58
DTC [21]			81.25	69.33	14.90	3.99
URPC [22]			82.48	71.35	14.65	3.65
MC-Net [23]			83.59	72.36	14.07	2.70
SS-Net [24]			86.33	76.15	9.97	2.31
MCF [25]			82.56	71.19	16.05	4.97
BCP [13]			88.02	78.72	7.90	2.15
Ours			**88.82**	**80.00**	**7.40**	**2.02**
UA-MT [19]	8(10%)	72(90%)	87.79	78.39	8.68	2.12
SASSNet [20]			87.54	78.05	9.84	2.59
DTC [21]			87.51	78.17	8.23	2.36
URPC [22]			86.92	77.03	11.13	2.28
MC-Net [23]			87.62	78.25	10.03	1.82
SS-Net [24]			88.55	79.62	7.49	1.90
MCF [25]			88.71	80.41	6.32	1.90
BCP [13]			89.62	81.31	6.81	1.76
Ours			**90.01**	**81.94**	**6.51**	**1.74**

III. Experiments

A. Experimental Settings & Implementation Details

We evaluate our method against several SOTA methods on two popular benchmarks: LA [26] and ACDC [27]. Following the general evaluation protocol [24], we adopt 3D V-Net [17] and 2D U-Net [28] as the feature extractors for the two datasets, respectively. For the LA dataset, training data consists of randomly cropped image patches of size $112 \times 112 \times 80$, with the zero-value region of mask defined as $74 \times 74 \times 53$, and data augmentation techniques (e.g., rotations, flipping) are applied. For the ACDC dataset, the randomly cropped zero-value area of the mask is defined as 170×170. For all experiments, we set $\eta = 0.999$, $\alpha = 0.9$, $\tau_1 = \tau_2 = 0.1$, and $K = 3$. Other hyperparameters are dataset-specific: (1) the scaling factor $\gamma = 0.02, 0.1$; (2) the batch size being 8 and 24; (3) the pre-training stage including 2k and 10k iterations; (4) the self-training stage including 15k and 30k iterations. All experiments are run on an NVIDIA GeForce RTX 4090 GPU.

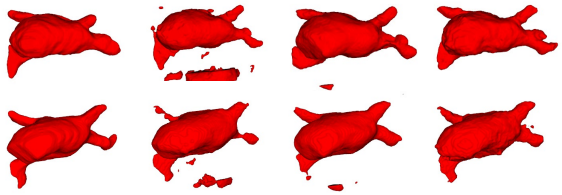
B. Evaluation Results

Quantitative Results. The quantitative results in Tabs. I and II show that our method surpasses SOTA methods on both benchmarks. For the LA dataset, our method consistently leads across all metrics, especially with limited labeled data. At 5% labeling, it achieves Dice coefficient of 88.82% and Jaccard index of 80.0%, with the lowest 95HD (7.4) and ASD (2.02) values. Similarly, for the ACDC dataset, our method performs strongly across varying amounts of labeled data. These results suggest that our method effectively utilizes unlabeled data. Notably, with only 10% of annotations, our method shows a minimal Dice coefficient reduction of just 1.75% compared to a fully labeled U-Net model.

Qualitative Results. The predicted segmentation results of several examples on the two datasets under 10% labeling are visualized in Figs. 3 and 4. Our method produces more precise segmentation than other methods on both 2D and 3D cases.

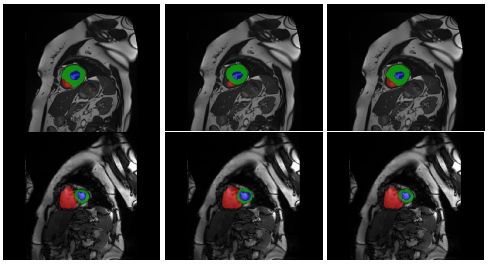
TABLE II: Comparison of segmentation quality on ACDC.

Method	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
U-Net[28]	3(5%)	0	47.83	37.01	31.16	12.62
	7(10%)	0	79.41	68.11	9.35	2.70
	70(All)	0	91.44	84.59	4.30	0.99
UA-MT [19]	3(5%)	67(95%)	46.04	35.97	20.08	7.75
SASSNet [20]			57.77	46.14	20.05	6.06
DTC [21]			56.90	45.67	23.36	7.39
URPC [22]			55.87	44.64	13.60	3.74
MC-Net [23]			62.85	52.29	7.62	2.33
SS-Net [24]			65.83	55.38	6.67	2.28
Co-BioNet [29]			87.46	77.93	1.11	1.11
BCP [13]			87.59	78.67	1.90	0.67
Ours			**88.64**	**80.21**	**1.51**	**0.60**
UA-MT [19]	7(10%)	63(90%)	81.65	70.64	6.88	2.02
SASSNet [20]			84.50	74.34	5.42	1.86
DTC [21]			84.29	73.92	12.81	4.01
URPC [22]			83.10	72.41	4.84	1.53
MC-Net [23]			86.44	77.04	5.50	1.84
SS-Net [24]			86.78	77.67	6.07	1.40
Co-BioNet [29]			88.49	79.76	3.70	1.14
BCP [13]			88.84	80.62	3.98	1.17
Ours			**89.69**	**81.88**	**1.65**	**0.56**



(a) Ground truth (b) UA-MT (c) MCF (d) Ours

Fig. 3: 3D segmentation visualization on LA.



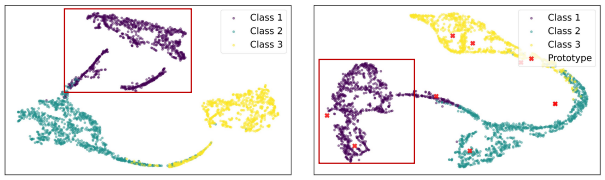
(a) Ground truth (b) BCP (c) Ours

Fig. 4: 2D segmentation visualization on ACDC.

C. Ablation Study

Key Components. To evaluate the effect of the prototype-based classifier, we implement a baseline model without it. Using 20% labeled data on the ACDC dataset, t-SNE visualizations (Fig. 5) show that our method achieves more compact embeddings for Class 1. We then train another baseline model, excluding the prototype consistency loss $\mathcal{L}_{cons}^{\mathrm{p}}$ and the contrastive loss $\mathcal{L}_{cont}$. Experiments with 5% labeled data on ACDC (Tab. III) reveal performance improvements as each loss is added. These results demonstrate the effectiveness of the prototype-based classifier and consistency constraint in modeling intra-class variance.

Prototype Update. In the proposed prototype update strategy, both labeled and unlabeled voxels are used. To assess its effect on the segmentation quality, we compare three alternative approaches: no update, using only labeled voxels, and using only unlabeled voxels. Experiments with 10% labeled data on the LA dataset show that all three strategies degrade performance (Tab. IV). This demonstrates that our update strategy builds representative prototypes in each category, and can adapt to potential shifts in data distribution.



(a) Baseline (b) Ours

Fig. 5: t-SNE visualization on ACDC.

TABLE III: Ablation studies of losses on ACDC.

$\mathcal{L}_{cons}^{\mathrm{l}}$	$\mathcal{L}_{cons}^{\mathrm{p}}$	$\mathcal{L}_{cont}$	Dice↑	Jaccard↑	95HD↓	ASD↓
✓			87.18	78.02	3.73	1.07
✓	✓		88.05	79.26	2.73	0.80
✓	✓	✓	**88.64**	**80.21**	**1.51**	**0.60**

TABLE IV: Ablation studies of prototype update on LA.

Labeled	Unabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
		89.55	81.25	6.83	1.73
✓		89.76	81.53	6.79	1.80
	✓	89.57	81.25	7.10	1.79
✓	✓	**90.01**	**81.94**	**6.51**	**1.74**

TABLE V: Ablation studies of prototype quantity on ACDC.

K	Dice↑	Jaccard↑	95HD↓	ASD↓
1	87.10	77.93	3.32	1.05
2	87.52	78.45	2.59	0.87
3	**88.64**	**80.21**	**1.51**	**0.60**
4	88.11	79.43	2.87	0.83
5	87.98	79.18	3.00	0.76

Prototype Quantity. Table V shows the performance of our method with different prototype numbers. Experiments with 5% labeled data on ACDC reveal that when $K = 1$, the prototype degrades into the mean embedding of each class. Performance improves as K increases from 1 to 3, but decreases when K exceeds 3. We speculate that too many sub-centers per class disrupt the compactness. Thus, to balance segmentation accuracy and computational cost, we set $K = 3$.

IV. Conclusion

We propose a Multi-Prototype-based Embedding Refinement method for semi-supervised medical image segmentation. By leveraging a multi-prototype classification strategy, our method captures intra-class voxel variations and reframes segmentation through structural relationships between voxel embeddings. Furthermore, a consistency constraint integrates different classification granularities. Experiments validate the effectiveness of our approach.

Acknowledgment

This work was supported by Natural Science Foundation of Chongqing, China (No. CSTB2023NSCQ-MSX0881) and Fundamental Research Funds for the Central Universities (No. SWU-KR22032).

References

[1] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Medical Image Analysis*, vol. 55, pp. 88–102, 2019.

[2] B. Ghoshal, A. Tucker, B. Sanghera, and W. Lup Wong, "Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection," *IEEE Comput. Intell. Mag.*, vol. 37, no. 2, pp. 701–734, 2021.

[3] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. of ICCV*, 2019, pp. 6022–6031.

[4] L. L. C.-M. P. W. J. Z. J. Yuanpeng He, Yali Bi, "Mutual evidential deep learning for semi-supervised medical image segmentation," in *Proc. of BIBM*, 2024.

[5] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, "A simple baseline for semi-supervised semantic segmentation with strong data augmentation*," in *Proc. of ICCV*, 2021, pp. 8209–8218.

[6] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. of NeurIPS*, 2004, pp. 529–536.

[7] L. Ran, Y. Li, G. Liang, and Y. Zhang, "Semi-supervised semantic segmentation based on pseudo-labels: A survey," *CoRR*, vol. abs/2403.01909, 2024.

[8] X. Kong, Z. Ren, and L. Liu, "Semi-supervised volumetric medical image segmentation via class prototype guided distribution-aligned representation learning," in *Proc. of ICASSP*, 2024, pp. 1931–1935.

[9] H. Mai, R. Sun, T. Zhang, Z. Xiong, and F. Wu, "DualRel: Semi-supervised mitochondria segmentation from a prototype perspective," in *Proc. of CVPR*, 2023, pp. 19 617–19 626.

[10] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. of CVPR*, 2022, pp. 4238–4247.

[11] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Proc. of NeurIPS*, vol. 30, 2017.

[12] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. of CVPR*, 2021.

[13] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *Proc. of CVPR*, 2023, pp. 11 514–11 524.

[14] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. of CVPR*, 2018, pp. 3733–3742.

[15] D. Sculley, "Web-scale k-means clustering," in *Proc. of WWW*, 2010, pp. 1177–1178.

[16] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Proc. of NeurIPS*, vol. 31, 2018.

[17] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. of 3DV*, 2016, pp. 565–571.

[18] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[19] L. Yu, S. Wang, X. Li, C. Fu, and P. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *Proc. of MICCAI*, 2019, pp. 605–613.

[20] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *Proc. of MICCAI*, 2020, pp. 552–561.

[21] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. of AAAI*, 2021, pp. 8801–8809.

[22] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Proc. of MICCAI*, 2021, pp. 318–329.

[23] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *Proc. of MICCAI*, 2021, pp. 297–306.

[24] Y. Wu, Z. Wu, Q. Wu, Z. Ge, and J. Cai, "Exploring smoothness and class-separation for semi-supervised medical image segmentation," *CoRR*, vol. abs/2203.01324, 2022.

[25] Y. Wang, B. Xiao, X. Bi, W. Li, and X. Gao, "MCF: mutual correction framework for semi-supervised medical image segmentation," in *Proc. of CVPR*, 2023.

[26] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 67, p. 101832, 2021.

[27] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, 2018.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, 2015, pp. 234–241.

[29] H. Peiris, M. Hayat, Z. Chen, G. F. Egan, and M. Harandi, "Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation," *Nature Machine Intelligence*, vol. 5, pp. 724–738, 2023.