UM-SAM: Unsupervised Medical Image Segmentation using Knowledge Distillation from Segment Anything Model

Jia Fu[1], He Li[1], Tao Lu[2], Shaoting Zhang[1,3], and Guotai Wang[1,3]

[1] University of Electronic Science and Technology of China, Chengdu, 611731, China
[2] Department of Radiology, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, 610072, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, 200030, China
guotai.wang@uestc.edu.cn

Abstract. Despite the success of deep learning in automatic medical image segmentation, it heavily relies on manual annotations for training that are time-consuming to obtain. Unsupervised segmentation approaches have shown potential in eliminating manual annotations, while they often struggle to capture distinctive features for low-contrast and inhomogeneous regions, limiting their performance. To address this, we propose UM-SAM, a novel unsupervised medical image segmentation framework that harnesses Segment Anything Model (SAM)'s capabilities for pseudo-label generation and segmentation network training. Specifically, class-agnostic pseudo-labels are generated via SAM's everything mode, followed by a shape prior-based filtering strategy to select valid pseudo-labels. Given SAM's lack of class information, a shape-agnostic clustering technique based on ROI pooling is proposed to identify target-relevant pseudo-labels based on their proximity. To reduce the impact of noise in pseudo-labels, a triple Knowledge Distillation (KD) strategy is proposed to transfer knowledge from SAM to a lightweight task-specific segmentation model, including pseudo-label KD, class-level feature KD, and class-level contrastive KD. Extensive experiments on fetal brain and prostate segmentation tasks demonstrate that UM-SAM significantly outperforms existing unsupervised and prompt-based methods, achieving state-of-the-art performance without requiring manual annotations.

Keywords: Segment Anything Model · ROI feature clustering · Contrastive learning · Knowledge distillation · Unsupervised segmentation.

1 Introduction

Automatic anatomical structure segmentation from medical images is essential for effective diagnosis and treatment planning [32]. While deep learning techniques have made significant advances in this area [10,12], they typically rely on large-scale annotated datasets for training [20]. However, obtaining accurate annotations for medical images is both time-consuming and labor-intensive, often

requiring specialized expertise [30]. To address this, unsupervised learning [19,22] has gained much attention for eliminating manual annotations.

Traditional unsupervised segmentation methods, such as DeepCluster [3] and level set [13], group pixels based on feature similarity or low-level edge information. Despite the efficiency, these methods struggle with medical images that often have low contrast and inhomogeneous target regions. To improve the performance, graph neural networks have been combined with clustering for this task [1], but are computationally intensive and noise-sensitive. Recently, self-supervised learning techniques [6,14,36] have been explored to extract low- and high-level semantic features through pretext tasks, such as image reconstruction and contrastive learning. Additionally, generative models like Variational Autoencoder (VAE) [25] and Generative Adversarial Network (GAN) [29] have been developed to learn feature representation. However, they still face challenges in capturing distinctive pixel-level feature representations for segmentation.

Recently, Segment Anything Model (SAM) [16], a prompt-driven foundation model for natural image segmentation, has shown impressive zero-shot performance and strong transferability across various downstream tasks [34,18]. Some studies [2,17,23] tried to apply SAM to assist unsupervised medical image segmentation. For example, SaLIP [2] uses SAM for part-based segmentation and employs CLIP [26] to retrieve segments containing the target object. However, its performance is limited by the significant domain gap between natural and medical images. Another approach, MedCLIP-SAM [17], combines BiomedCLIP [33] and gScoreCAM [5] for target localization, generating bounding box prompts for SAM to produce pixel-level pseudo-labels to train a segmentation network. While it shows encouraging results in breast and brain tumor segmentation, it primarily focuses on pseudo-label generation and underutilizes SAM's robust representation capability to boost the downstream segmentation model.

To address these limitations, we propose a novel framework, UM-SAM, for unsupervised medical image segmentation. The main contributions are threefold: 1) We propose UM-SAM, a SAM-guided unsupervised medical image segmentation method that leverages SAM for high-quality pseudo-label generation at the logit level and robust segmentation model training via feature-level knowledge distillation. 2) To derive target-specific pseudo-labels from SAM's class-agnostic raw output, we introduce a hybrid pseudo-label filtering method that combines prior knowledge with feature clustering based on ROI pooling, effectively rejecting irrelevant segments and ensuring high-quality pseudo-labels. 3) We design a triple KD-based learning framework to train a downstream segmentation model. It integrates pseudo-label KD, class-level feature KD, and class-level contrastive KD to transfer SAM's strong feature representation ability, significantly improving the downstream model's performance. UM-SAM achieved a mean DSC of 90.15% and 80.44% for 2D fetal brain and prostate segmentation, respectively, outperforming state-of-the-art unsupervised segmentation methods.

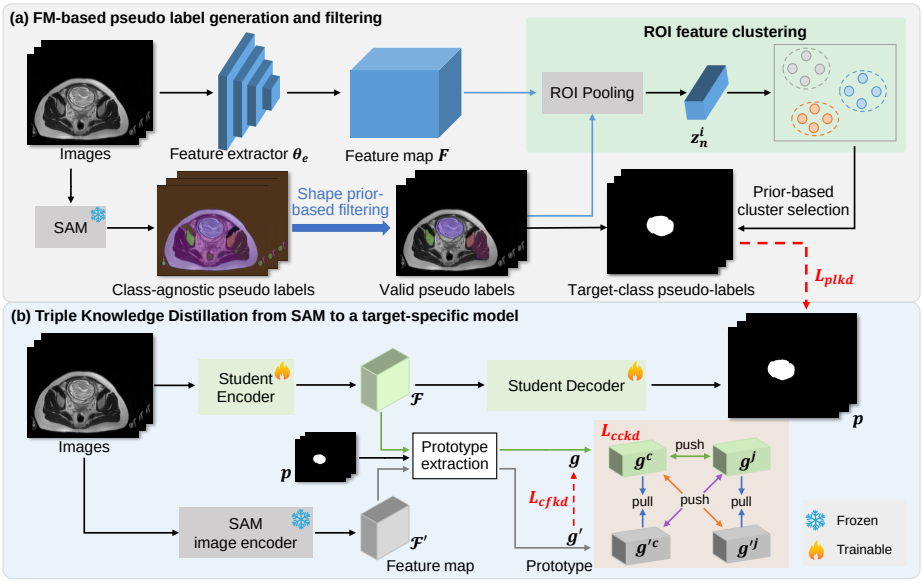


Fig. 1. Overview of the proposed UM-SAM framework for unsupervised segmentation. $\mathcal{L}_{plkd}$, $\mathcal{L}_{cfkd}$ and $\mathcal{L}_{cckd}$ denotes the pseudo-label KD loss, class-level feature KD loss and class-level contrastive KD loss, respectively.

2 Method

Fig. 1 illustrates an overview of the proposed UM-SAM. First, each input image is processed through SAM's everything mode to produce class-agnostic pseudo-labels, which are subsequently filtered via shape priors to select valid pseudo-labels. A feature clustering method based on ROI pooling is further used to identify target-class pseudo-labels. To eliminate the impact of noise in pseudo-labels, we propose a triple KD strategy to distill knowledge from SAM to a lightweight task-specific model for robust feature representations.

2.1 SAM-based pseudo-label generation and filtering

Class-agnostic segments generation. Given the lack of manual annotation, we leverage SAM in a fully automatic manner (everything mode) rather than a semi-automatic manner (point or bounding box prompt) to segment an input image into different class-agnostic segments. Let N denote the number of training images, and X_n be the n-th training image. Grid-wise potential prompt locations $P \in \mathcal{R}^{m^2}$ are sampled across X_n, where m denotes the number of points along each side of the image. The raw pseudo-label (mask map) $\mathcal{M}_n$ generated by SAM's everything mode SAM_{EM} is expressed as:

$$\mathcal{M}_n = SAM_{EM}(X_n, P) = \{M_n^i\}_{i=1}^{K_n} \in \{0,1\}^{H \times W}, \tag{1}$$

where M_n^i is the i^{th} class-agnostic segment (i.e., ROI) generated from image X_n, and K_n denotes the number of segments. The set of generated class-agnostic segments for all the training images is denoted as Ω_0:

$$\Omega_0 = \{M_n^i;\ \text{for } n = 1, ..., N \text{ and } i = 1, ..., K_n\}, \tag{2}$$

Shape prior-based filtering. Typically, the segment instances in Ω_0 mainly contain non-target tissues including the background, with only a small set containing the target class. Thus, we need to filter out irrelevant segments before using it to train a segmentation model. To achieve this, we first filter valid pseudo-labels based on general shape priors of the target, such as aspect ratio and size. The set of valid pseudo-labels Ω_v is:

$$\Omega_v = \{M_n^i | M_n^i \in \Omega_0 \text{ and } V^i \in [V_{min}, V_{max}] \text{ and } A^i \in [A_{min}, A_{max}]\}, \tag{3}$$

where V^i is the ratio of M_n^i's area in the entire image, and A^i is the aspect ratio of segment M_n^i. V_{min} and V_{max} are the minimal and maximal segment area ratio, and A_{min} and A_{max} are lower and upper bounds for aspect ratio. These basic shape priors can effectively filter out irregular segments that are too small/large or have an uncommon aspect ratio.

Class-specific pseudo-labels based on ROI feature clustering. Although the prior-based filtering can reject a large number of irrelevant segments, the valid pseudo-label set Ω_v still contains some background ROIs that have shape and size similar to the target, making it desirable to identify the exact class of each ROI. Although CLIP can be used for this purpose [2], its performance is limited by the lack of domain-specific knowledge in medical imaging, especially for fine-grained tissue classes. To overcome this, we propose a shape-robust ROI clustering method for selecting target segments.

Inspired by Fast R-CNN [9], instead of extracting feature for each ROI, we extract the feature map for each training image via a feature extractor in a single forward pass. Due to SAM's lack of semantics [18], we used an on-the-shelf pre-trained encoder, DINO [4], as the feature extractor θ_e. The feature $z_n^i \in R^D$ of an ROI M_n^i is obtained by ROI pooling from $F_n \in R^{D \times H' \times W'}$, where D is the dimension of features. The set of features based on Ω_v is denoted as:

$$\mathcal{F}_v = \{z_n^i | M_n^i \in \Omega_v\} \quad \text{with} \quad z_n^i = Pool_{roi}(F_n, M_n^i), \tag{4}$$

where $Pool_{roi}(F_n, M_n^i)$ is the ROI pooling operation based on F_n and M_n^i, i.e., the features for pixels in M_n^i are averaged.

Then, the ROI instances in Ω_v are clustered based on $\mathcal{F}_v$. For simplicity, we used the K-means [15] clustering method. After clustering, the cluster with the minimal discrepancy with the prior attribute of the target is selected as the target-class pseudo-labels. Let M_k^j denote the j-th ROI in the k-th cluster $\mathcal{C}_k$, and $\mathcal{A}(M_k^j)$ denotes its attribute, such as size and aspect ratio. The cluster-level attribute is denoted as $\mathcal{A}_k = \sum_j \mathcal{A}(M_k^j)/|\mathcal{C}_k|$. The average prior attribute for the target is denoted as $\hat{\mathcal{A}}$. The ROI cluster corresponding to the target class is

$\mathcal{C}_{k^*} = \arg\min_k |\mathcal{A}_k - \hat{\mathcal{A}}|,\ i \in [1, K]$, where $|\cdot|$ is the L1 loss. For each unlabeled image X_n, the pseudo-label Y_n is set as the mask of M_n^i that in $\mathcal{C}_{k^*}$, otherwise empty if none of its M_n^i is in $\mathcal{C}_{k^*}$. The training set with target-class pseudo-labels is therefore denoted as:

$$\mathcal{D}_p = \{(X_n, Y_n)\}_{n=1}^N \quad \text{with} \quad Y_n = \begin{cases} M_n^i, \text{ if } \exists M_n^i \in \mathcal{C}_{k^*} \\ 0, \quad \text{Otherwise} \end{cases}, \tag{5}$$

2.2 Triple Knowledge Distillation from SAM to task-specific model

Given SAM's high computational cost and reliance on manual prompts, we train a lightweight task-specific model θ_s for automatic and efficient segmentation. As a baseline method, pseudo-label Y_n from $\mathcal{D}_p$ is used to supervise θ_s, where Y_n from SAM serves as logit-level distillation. The pseudo-label KD loss $\mathcal{L}_{plkd}$ is:

$$\mathcal{L}_{plkd} = \frac{1}{N} \sum_{n=1}^{N} \big(L_{ce}(P_n, Y_n) + L_{dice}(P_n, Y_n) \big), \tag{6}$$

where P_n denotes the downstream model θ_s's prediction for X_n. L_{ce} and L_{dice} are standard Cross-Entropy (CE) loss and Dice loss for segmentation, respectively. As pseudo-label Y_n is noisy, using $\mathcal{L}_{plkd}$ to supervise θ_s can yield suboptimal results. To avoid this, we propose two novel objectives: class-level feature KD that aligns feature of θ_s with that of SAM for each class, and class-level contrastive KD that encourages inter-class feature dissimilarity for both θ_s and SAM.

First, unlike traditional approaches [18] that rely on hard segmentation results, we leverage soft predictions from θ_s to obtain class prototypes for robustness against noise. Let $P_{\mathbf{x}}^c$ denote the probability of pixel $\mathbf{x}$ being class c obtained by θ_s. The feature maps obtained by the bottleneck of θ_s and SAM are denoted as $\mathcal{F}$ and $\mathcal{F}'$, respectively. The prototypes for class c obtained by θ_s and SAM are denoted as g^c and g'^c. The class-level feature KD loss $\mathcal{L}_{cfkd}$ is defined as:

$$\mathcal{L}_{cfkd} = \sum_{c=0}^{C-1} ||g^c - g'^c||_2^2 = \sum_{c=0}^{C-1} || \frac{\sum_{\mathbf{x}} \mathcal{F}_{\mathbf{x}} \cdot P_{\mathbf{x}}^c}{\sum_{\mathbf{x}} P_{\mathbf{x}}^c} - \frac{\sum_{\mathbf{x}} \mathcal{F}'_{\mathbf{x}} \cdot P_{\mathbf{x}}^c}{\sum_{\mathbf{x}} P_{\mathbf{x}}^c} ||_2^2, \tag{7}$$

where C is the number of classes for the segmentation task. $||\cdot||_2$ is the $L2$ norm.

Second, to further enforce intra-class consistency and inter-class separability, we introduce a class-level contrastive KD loss L_{cckd}, which encourages the prototypes of the same class to be close while those from different classes to be further apart. We set (g^c, g'^c) as a positive pair, and (g^c, g'^j) and (g^c, g^j) as negative pairs, where $j \neq c$. L_{cckd} is formulated with the InfoNCELoss [24] as:

$$\mathcal{L}_{cckd} = -\sum_{c=0}^{C-1} log \frac{e^{sim(g^c, g'^c)/\tau}}{e^{sim(g^c, g'^c)/\tau} + \sum_{j \neq c} e^{sim(g^c, g'^j)/\tau} + \sum_{j \neq c} e^{sim(g^c, g^j)/\tau}}. \tag{8}$$

where τ is the temperature, and $sim(\cdot, \cdot)$ represent cosine similarity. Overall, the downstream segmentation model θ_s is trained with the joint loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{plkd} + \lambda_1 \mathcal{L}_{cfkd} + \lambda_2 \mathcal{L}_{cckd}, \tag{9}$$

where λ_1 and λ_2 are hyper-parameters for the weights of the feature KD losses.

3 Experiments and Results

3.1 Datasets and Implementation Details

In this work, we evaluated UM-SAM and compared it with several previous works on two medical image segmentation datasets: 1) Fetal Brain (FB) [7,8] dataset that contains 115 Half-Fourier Acquisition Single-shot Turbo spin-Echo (HASTE) sequences from pregnant women in the second trimester, which randomly divided into 80, 10 and 25 cases for training, validation and testing, respectively. 2) Promise12 dataset [21] that consists of 100 transverse T2-weighted Magnetic Resonance (MR) images collected from four medical centers. We used 50 training cases, 30 test cases, and 20 live challenge cases from this dataset for training, validation, and testing, respectively. For quantitative evaluation, we adopted the commonly used Dice Similarity Coefficient (DSC) and Average Symmetric Surface Distance (ASSD) in 3D space.

For preprocessing, the intensity of each volume was clipped to the 1st and 99th percentile of the values and normalized to [0, 1]. For SAM, we use the officially released version with the ViT-H model for mask generation and its image encoder's offline features for efficient feature KD. To balance the recognition of targets and the number of segments, m was 32 and 48 for the FB and Promise12 datasets, respectively. Given the characteristics of the fetal brain and prostate, we set V_{min} and V_{max} as (0.01, 0.25) and (0.01, 0.2), A_{min} and A_{max} as (2/3, 1.5) and (0.5, 2) for the FB and Promise12 datasets, respectively. K was 8 and 10 for the FB and Promise12 datasets using the Sum of Squared Error (SSE) method [31]. The attribute $\mathcal{A}$ was the aspect ratio for the FB dataset ($\hat{\mathcal{A}} = 1.0$) and ROI size for Promise12 ($\hat{\mathcal{A}} = 1200\ mm^2$ according to the normal size of prostate in a 2D slice). We trained UNet [28] as θ_s for 200 epochs with a batch size of 32 using a Stochastic Gradient Descent (SGD) optimizer, where the momentum was 0.9, and the weight decay was 5×10^{-4}. Data augmentation methods include random cropping with a size of 256x256, random rotation and noising. Following [11], τ in Eq. (8) was set as 0.5 for both datasets. Based on the best results on the validation set, λ_1 and λ_2 in Eq. (9) were set as 0.1 and 0.01 for the FB dataset, and 0.5 and 0.01 for the Promise12 dataset, respectively. During inference, we used θ_s to obtain segmentation results.

3.2 Comparison with state-of-the-art methods

Comparison with prompt-based segmentation methods. To evaluate the effectiveness and efficiency of our method, we first compared it with prompt-based foundation models: 1) SAM [16], which leverages foreground points (fg), background points (bg) or bounding box ($bbox$) for each positive slice as prompt; 2) SAM2 [27], which treats 3D volumes as videos and employs a foreground point or bbox prompt for target tracking; 3) CryoSAM [35], which expands a given point prompt for segmentation by matching features based on their similarity with extracted target features. Results are listed in Table 1, where $\alpha bbox$ denotes a bbox expanded by α times that of ground truth bbox for simulating user inputs.

Table 1. Quantitative comparison of several existing prompt-based and unsupervised methods on FB and Promise12 datasets.

Method	Prompt type	FB		Promise12	
		DSC (%)	ASSD (mm)	DSC (%)	ASSD (mm)
SAM [16]	1 *fg*+1 *bg*/slice	86.78±8.27	16.11±29.87	61.89±14.04	12.18±11.05
SAM [16]	1.3 *bbox*/slice	89.19±3.27	2.73±1.52	82.33±3.59	1.80±0.85
SAM [16]	1.4 *bbox*/slice	85.52±3.79	5.29±2.88	78.00±3.66	3.02±1.34
SAM2 [27]	1 *fg*/volume	46.70±27.77	108.69±133.22	32.03±22.53	43.13±56.55
SAM2 [27]	1 *bbox*/volume	60.21±17.47	35.58±22.86	60.82±22.95	7.56±8.11
CryoSAM [35]	1 *fg*/volume	44.64±20.86	32.78±56.74	39.45±22.41	16.77±10.34
Kmeans (intensity) [15]	/	59.45±12.04	33.39±61.20	39.94±9.96	27.71±22.63
Kmeans (DINO) [4]	/	68.90±12.33	8.79±25.10	63.82±7.57	4.01±3.19
SaLIP [2]	/	76.63±14.20	75.96±129.91	32.32±18.33	84.51±61.37
MedCLIP-SAM [17]	/	60.56±21.53	11.94±21.69	10.62±10.10	68.93±45.90
Ours	/	**90.15±4.87**	**0.76±0.88**	**80.44±9.60**	**1.04±0.76**
FullySup	dense label	96.26±2.75	0.19±0.10	88.34±3.22	0.48±0.18

It shows that the performance of prompt-based methods is highly dependent on user inputs. In contrast, our method achieved comparable or even superior results without requiring manual annotations during inference, obtaining a mean DSC of 90.15% and 80.44% on the FB and Promise12 datasets, respectively.

Comparison with unsupervised segmentation methods. Additionally, we compared our method with several unsupervised methods, including 1) K-means (intensity) and K-means (DINO), which perform K-means clustering [15] based on image intensity and features from DINO [4], respectively; 2) SaLIP [2] and 3) MedCLIP-SAM [17]. For K-means-based methods, we took the cluster with maximal overlap with the ground truth as the segmentation result. We provide the class label for each slice with SaLIP and MedCLIP-SAM, as they require information about the existence of the target. Results in Table 1 demonstrate that SaLIP [2] obtained a satisfactory Dice (76.63%) for the fetal brain, but the value is very poor (32.32%) for the prostate, which is mainly caused by the missing information for the prostate from CLIP. Similarly, MedCLIP-SAM [17] also obtained a poor performance that is even worse than K-means-based methods. Compared with them, our method improved the average DSC by over 13.52 and 16.62 percentage points for the fetal brain and prostate, respectively. Fig. 2 visually shows that compared to these methods, our method achieved more accurate segmentation results, with boundaries closely aligned to the ground truth.

3.3 Ablation study

Effectiveness of pseudo-label filtering strategies. To evaluate our filtering strategies, we employ an ROI-level precision score to measure the quality of the generated pseudo-label sets, which is defined as the fraction of true positive pseudo-labels (IOU with ground truth $>$ threshold T) among all retrieved pseudo-labels. Fig. 3(a) shows the precision scores across different IOU thresholds (0.1 to 0.9) for the class-agnostic pseudo-label set Ω_0, the valid pseudo-label

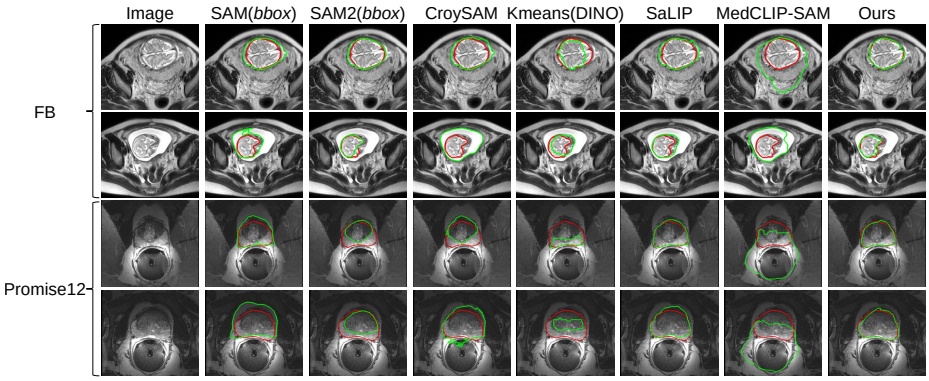
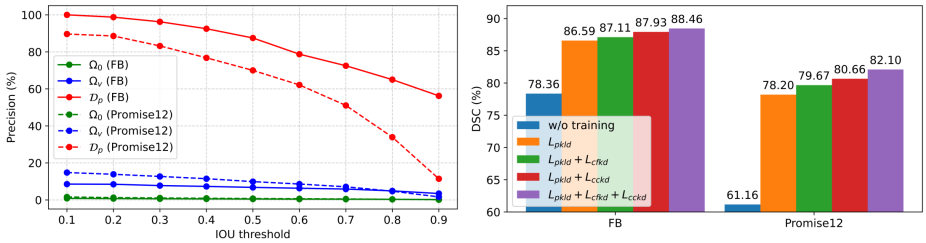


Fig. 2. Visual comparison of our method with existing methods. Red and green contours indicate the boundary of ground truth and prediction, respectively.



(a) Ablation study on pseudo-label filtering methods (b) Ablation study on the triple KD strategy

Fig. 3. Ablation study on pseudo-label filtering methods and the triple KD strategy.

set Ω_v, and the target-class pseudo-label set $\mathcal{D}_p$. The precision scores for Ω_0 were consistently low due to the large amount of irrelevant segments from SAM's everything mode. After applying shape prior-based filtering and ROI feature clustering, the precision scores for Ω_v and $\mathcal{D}p$ improved significantly.

Effectiveness of our triple knowledge distillation strategy. Fig. 3(b) summarizes the results of each component in the triple KD strategy on the validation sets of the FB and Promise12 datasets. The baseline is applying Eq. 5 to testing images to obtain target-class labels without training θ_s. Compared with it, training with L_{plkd} significantly improved the mean DSC values from 78.36% and 61.16% to 86.59% and 78.20% on FB and Promise12 datasets, respectively. Introducing $\mathcal{L}_{cfkd}$ or $\mathcal{L}_{cckd}$ to L_{plkd} improved the segmentation model's performance on both datasets. Combining the three KD losses further enhanced results, indicating the effectiveness of the proposed triple KD loss.

4 Conclusion

This work proposes an unsupervised medical image segmentation method, UM-SAM, based on the foundation model SAM. To generate pixel-wise pseudo-labels,

we leverage SAM to automatically partition images into multiple class-agnostic segments, followed by shape prior-based filtering and ROI feature clustering for target retrieval. A triple KD strategy is proposed to enhance the performance of segmentation model in the presence of noisy labels. Experimental results on the FB and Promise12 datasets show that UM-SAM outperformed existing unsupervised methods, and achieved comparable performance to promptable methods with tight prompts. It is of interest to extend our method to other SAM variants and apply it to multi-class and 3D segmentation tasks in the future.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (62271115), and in part by the Chengdu Science and Technology Program (2024-YF05-01160-SN).

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

References

1. Aflalo, A., Bagon, S., Kashti, T., Eldar, Y.: DeepCut: Unsupervised segmentation using graph neural networks clustering. In: ICCV. pp. 32–41 (2023)
2. Aleem, S., Wang, F., Maniparambil, M., Arazo, E., Dietlmeier, J., Curran, K., Connor, N.E., Little, S.: Test-time adaptation with SaLIP: A cascade of SAM and CLIP for zero-shot medical image segmentation. In: CVPR. pp. 5184–5193 (2024)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV. pp. 132–149 (2018)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
5. Chen, P., Li, Q., Biaz, S., Bui, T., Nguyen, A.: gscorecam: What objects is clip looking at? In: PACCV. pp. 1959–1975 (2022)
6. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: CVPR. pp. 16794–16804 (2021)
7. Fu, J., Lu, T., Zhang, S., Wang, G.: UM-CAM: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised fetal brain segmentation. In: MICCAI. pp. 315–324 (2023)
8. Fu, J., Wang, G., Lu, T., Yue, Q., Vercauteren, T., Ourselin, S., Zhang, S.: Um-cam: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised segmentation. Pattern Recognition p. 111204 (2024)
9. Girshick, R.: Fast R-CNN. In: ICCV. p. 1440–1448 (2015)
10. Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Transactions on Medical Imaging **40**(2), 699–711 (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)

12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
13. Jeon, M., Alexander, M., Pedrycz, W., Pizzi, N.: Unsupervised hierarchical image segmentation with level set and additive operator splitting. Pattern Recognition Letters **26**(10), 1461–1469 (2005)
14. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: ICCV. pp. 9865–9874 (2019)
15. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 881–892 (2002)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
17. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Medclip-sam: Bridging text and image towards universal medical image segmentation. In: MICCAI. pp. 643–653 (2024)
18. Kweon, H., Yoon, K.J.: From SAM to CAMs: Exploring segment anything model for weakly supervised semantic segmentation. In: CVPR. pp. 19499–19509 (2024)
19. Li, B.N., Chui, C.K., Chang, S., Ong, S.H.: Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. Computers in Biology and Medicine **41**(1), 1–10 (2011)
20. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
21. Litjens, G., Toth, R., Van De Ven, W., Hoeks, C., Kerkstra, S., Van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical Image Analysis **18**(2), 359–373 (2014)
22. Liu, L., Aviles-Rivero, A.I., Schönlieb, C.B.: Contrastive registration for unsupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems **36**(1), 147–159 (2023)
23. Ma, X., Fu, J., Liao, W., Zhang, S., Wang, G.: CLISC: Bridging clip and sam by enhanced cam for unsupervised brain tumor segmentation. In: ISBI. pp. 1–5. IEEE (2025)
24. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
25. Pinaya, W.H., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Medical Image Analysis **79**, 102475 (2022)
26. 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 (2021)
27. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)

29. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis **54**, 30–44 (2019)
30. Shen, W., Peng, Z., Wang, X., Wang, H., Cen, J., Jiang, D., Xie, L., Yang, X., Tian, Q.: A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9284–9305 (2023)
31. Su, T., Dy, J.: A deterministic method for initializing k-means clustering. In: ICTAI. pp. 784–786 (2004)
32. Van Ginneken, B., Schaefer-Prokop, C.M., Prokop, M.: Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology **261**(3), 719–732 (2011)
33. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 **2**(3), 6 (2023)
34. Zhao, X., Li, P., Luo, X., Yang, M., Chang, S., Li, Z.: Sam-driven weakly supervised nodule segmentation with uncertainty-aware cross teaching. In: ISBI. pp. 1–5. IEEE (2024)
35. Zhao, Y., Bian, H., Mu, M., Uddin, M.R., Li, Z., Li, X., Wang, T., Xu, M.: Cryosam: Training-free CryoET tomogram segmentation with foundation models. In: MICCAI. pp. 124–134 (2024)
36. Zhou, Y., Gu, R., Zhang, S., Wang, G.: Self-supervised pre-training based on contrastive complementary masking for semi-supervised cardiac image segmentation. In: ISBI. pp. 1–5. IEEE (2024)