Journal of Science and Technology – Smart Systems and Devices
Volume 36, Issue 2, May 2026, 128–136

Ovarian Ultrasound Image Segmentation with Limited Training Data

Thanh-Phuc Dao[1], Sy-Thien Dinh[1], Hoang-Son Bui[1], Thi-Loan Pham[1,3], Thi Hong Thien Dang[2], Van-Thang Nguyen[2], Phuong-Thao Nguyen[2], Hai Vu[1], Thanh-Hai Tran[1], Duy-Hai Vu[1], Thi-Lan Le[1,*]

[1]*School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Ha Noi, Vietnam*
[2]*National Hospital of Obstetrics and Gynecology, Ha Noi, Vietnam*
[3]*Hai Duong University, Hai Phong, Vietnam*
[*]*Corresponding author email: lan.lethi1@hust.edu.vn*

Abstract

Ultrasound imaging is pivotal for ovarian tumor diagnosis, yet it poses significant segmentation challenges due to severe speckle noise, low contrast, and high inter-patient morphological variability. The limited availability of annotated medical data, making few-shot segmentation a practical and necessary solution. Although recent universal few-shot models such as UniverSeg demonstrate promising cross-task generalization, their performance is highly sensitive to stochastic support set sampling, often leading to unstable and inferior predictions. To address this limitation, we propose a CLIP-guided support retrieval strategy that replaces random sampling with deterministic similarity-based selection in the semantic embedding space of the Contrastive Language–Image Pre-training model. By retrieving morphologically consistent support samples for each query image, our method enhances structural alignment and reduces support-query mismatch without requiring additional training or model modification. Extensive experiments on two ovarian ultrasound datasets, OvaTUS and OTU_2D, demonstrate that our approach consistently outperforms the baseline UniverSeg and other few-shot methods. Specifically, on the OvaTUS dataset, our method achieves a Dice Coefficient of 75.84% and Intersection over Union of 64.87%, surpassing the random selection baseline by 2.19% and 2.69%, respectively. Furthermore, our approach shows superior robustness in extreme few-shot settings ($N = 1$), improving the Dice score by over 8% compared to the baseline. Code will be publicly released upon acceptance.

Keywords: CLIP, few-shot segmentation, medical image analysis, ovarian ultrasound, UniverSeg.

1. Introduction

Ultrasound imaging (US) is widely adopted in clinical practice due to its non-invasive nature, affordability, and real-time imaging capability. In gynecological applications, US plays a crucial role in the detection, characterization, and longitudinal monitoring of ovarian tumors. Despite its clinical importance, accurate analysis of ovarian ultrasound images remains challenging and highly operator-dependent. These challenges stem from intrinsic characteristics of ultrasound imaging, including severe speckle noise, low contrast, ambiguous lesion boundaries, and substantial inter-patient variability in tumor morphology and size. These limitations highlight the need for robust and automated segmentation methods tailored to ovarian ultrasound imaging.

In recent years, deep learning approaches have achieved great success in medical image segmentation. CNN-based architectures [1–3] can automatically learn discriminative features and deliver high segmentation accuracy when trained on large, fully annotated datasets. However, in real-world scenarios, gathering and annotating medical imaging data is expensive, time-consuming, and requires the expertise of experienced sonographers. As a result, supervised segmentation models often struggle to generalize to new tasks or rare datasets, limiting their scalability and practical deployment.

Few-shot segmentation has emerged as an alternative learning paradigm to address data scarcity, enabling segmentation on unseen images using only a small number of annotated support samples. Instead of retraining models for each new task, few-shot approaches emphasize rapid adaptation and cross-task generalization. Within this framework, UniverSeg [4] was proposed as a universal medical image segmentation model capable of handling diverse datasets under a few-shot setting without task-specific retraining. By leveraging a unified architecture and inference mechanism, UniverSeg demonstrates promising cross-dataset generalization performance.

Despite these advantages, UniverSeg suffers from a critical limitation: its support set is typically selected randomly from a support pool. Such stochastic selection does not guarantee morphological or structural similarity between support and query images. As a result, support–query misalignment may occur, leading

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.190.ssad.2026.36.2.14
Received: Feb 25, 2026; Revised: Mar 7, 2026;
Accepted: Mar 29; Online: Apr 17, 2026.

to unstable predictions and degraded segmentation accuracy, particularly in challenging modalities such as US.

To address this issue, we propose a CLIP-guided support retrieval strategy that enhances inference without modifying the underlying network architecture or retraining the model. Specifically, we employ a pretrained CLIP model [5] to extract global semantic embeddings for both query and candidate support images. Cosine similarity in the embedding space is then used to retrieve morphologically consistent support samples for each query instance. By replacing stochastic sampling with deterministic similarity-based selection, our approach improves structural alignment and reduces inference instability.

The primary contribution of this work is the proposal and evaluation of a CLIP-based guided support selection strategy for UniverSeg [4], along with a detailed analysis of its impact on segmentation accuracy and result stability. Experimental results demonstrate that the proposed method consistently outperforms random support selection, offering a practical and effective solution for medical image segmentation under limited annotation scenarios.

2. Related Work

In recent years, deep learning has achieved remarkable success in medical image analysis, becoming the de facto standard for segmentation tasks. CNNs, particularly the UNet architecture, have demonstrated state-of-the-art performance across various modalities, including CT, MRI, and ultrasound [6, 7]. These fully supervised models rely on an architecture to capture both semantic context and spatial details [8]. However, their primary limitation lies in their data-hungry nature; achieving high diagnostic accuracy typically requires large-scale, pixel-level annotated datasets. In clinical scenarios, particularly ovarian tumor segmentation, acquiring such datasets is often prohibitively expensive and time-consuming due to the scarcity of expert annotators and the high variability in tumor morphology.

To reduce annotation demands, transfer learning pretrains models on large natural image datasets (e.g., ImageNet [9]) and fine-tunes them for medical tasks. Although this improves convergence on small datasets, domain shift limits feature transferability, and fine-tuning still requires sufficient labeled data, making it less suitable for rare conditions.

Consequently, the research focus has shifted towards few-shot segmentation (FSS), a paradigm designed to generalize to new classes using only a handful of labeled support examples [10]. Unlike traditional supervised learning, FSS models learn a similarity metric between a query image and a support set, allowing them to segment unseen objects without weight updates. Early works such as SE-Net [11] and PANet [12] utilized prototype learning to represent class features, yet they often required task-specific episodic training and struggled to generalize across entirely different anatomical domains.

More recently, universal frameworks like UniverSeg aim to handle diverse medical segmentation tasks without retraining. UniverSeg [4] represents a significant breakthrough in this direction. Unlike conventional FSS methods restricted to specific organs, UniverSeg is trained on a massive collection of diverse medical datasets (MegaMedical) to learn a task-agnostic segmentation function. It employs a novel CrossBlock mechanism within a UNet-like architecture to dynamically interact features from the query image with those from the support set. This design allows UniverSeg to perform segmentation on completely unseen datasets by simply swapping the support set, eliminating the need for fine-tuning. Despite its robust generalization capabilities, the standard UniverSeg framework typically selects support images randomly. This stochastic selection can lead to suboptimal performance, particularly in heterogeneous datasets like ovarian ultrasound, where the morphological similarity between the chosen support samples and the query image plays a critical role in segmentation accuracy.

3. Proposed Method

3.1. Problem Formulation and Overall Framework

We formulate the task of ovarian tumor segmentation in ultrasound images as a few-shot segmentation problem.

Let $\mathbf{D}_{\text{pool}} = \{(x_i, y_i)\}_{i=1}^{M}$ denote a large pool of available annotated medical images, where M is the total number of samples.

In a fully supervised setting, a segmentation model is trained on the large fully annotated dataset $\mathbf{D}_{\text{pool}}$ and learns a global mapping $f_\theta : x \to y$ that generalizes to unseen image x_q. This paradigm assumes abundant labeled data that sufficiently captures morphological variability, which is often unrealistic in clinical practice as obtaining dense segmentation masks for all ultrasound images is expensive and requires expert effort.

In contrast, few-shot segmentation aims to generalize to a target class using only a limited number of annotated examples. Given a query image $x_q \in \mathbb{R}^{H\times W}$, the objective is to predict its binary mask $\hat{y}_q \in \{0,1\}^{H\times W}$ conditioned on a small support set:

$$\mathbf{S} = \{(x_{s_j}, y_{s_j})\}_{j=1}^{N}, \tag{1}$$

where N is typically small (e.g., $N \in \{1,\ldots,16\}$). Instead of relying solely on a large static training set, the model performs segmentation by leveraging task-specific information provided by $\mathbf{S}$ at inference time. The central challenge is to construct an optimal support set $\mathbf{S}$ that maximizes segmentation accuracy under severe annotation scarcity.

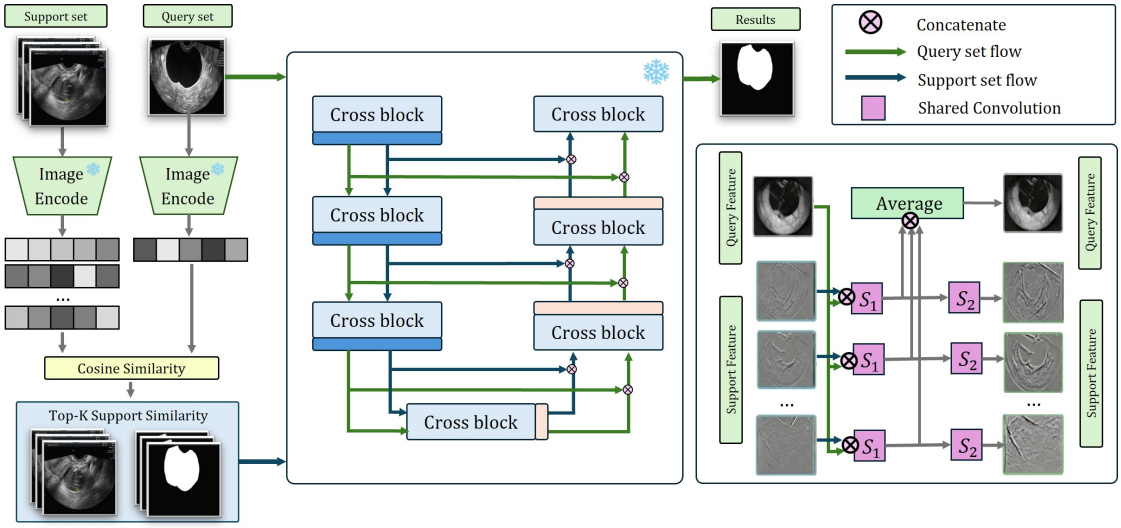

Fig. 1. Overview of the proposed CLIP-guided UniverSeg architecture, which consists of a frozen CLIP encoder for support set retrieval and a cross-attention mechanism for few-shot segmentation

The model learns a task-agnostic mapping function f_θ and then predicts the binary mask:

$$\hat{y}_q = f_\theta(x_q, \mathbf{S}), \tag{2}$$

where θ represents the pre-trained parameters of the network. This function generates the segmentation mask through a single forward pass without gradient updates.

UniverSeg [4] represents a significant breakthrough in few-shot segmentation. UniverSeg employs a novel CrossBlock mechanism within a U-Net-like architecture to dynamically interact features from the query image with those from the support set. However, UniverSeg typically selects support images randomly. This stochastic selection can lead to suboptimal performance. In this paper, we improve UniverSeg by introducing a new support set selection mechanism. Fig. 1 illustrates the proposed method consisting of two distinct phases: (A) **CLIP-Guided Support Set Selection**, where a frozen CLIP encoder embeds the support pool and the query to select the top-N morphologically similar samples; and (B) **UniverSeg Pre-trained**, where the UniverSeg backbone processes the query and the selected support set to produce the final prediction by leveraging (C) **Cross Block** for dual-path learning from both support and query set.

3.2. CLIP-Guided Support Set Selection

UniverSeg typically employs stochastic support sampling, where $\mathbf{S}$ is drawn randomly from $\mathbf{D}_{\text{pool}}$. In heterogeneous domains like ovarian ultrasound, random selection often results in a *semantic gap*, where the selected support images differ significantly from the query in terms of morphology and echogenicity. This discrepancy introduces noise into the network, leading to performance degradation.

To mitigate this, we propose a mechanism to select the support set leveraging the Contrastive Language–Image Pre-training (CLIP) framework [5]. We hypothesize that support samples which are visually similar to the query in the embedding space will provide the most effective guidance for segmentation.

We utilize the ViT-B/16 variant of CLIP as the feature extractor g_ϕ. This model employs a Vision Transformer backbone with a patch size of 16×16, capable of capturing fine-grained global dependencies essential for characterizing subtle textural variations in medical images.

The retrieval process is divided into an offline indexing phase and an online query phase:

1) **Offline Indexing:** We pre-compute the embeddings for all images in the support pool $\mathbf{D}_{pool}$. Let x_{s_j} be a candidate image; its embedding is $z_{s_j} = g_\phi(x_{s_j}) \in \mathbb{R}^{512}$. These embeddings are stored in a lookup table.

2) **Online Retrieval:** For a new query image x_q, we compute its embedding $z_q = g_\phi(x_q)$. We then calculate the cosine similarity between the z_q and all pooled embeddings by the following equation:

$$c_j = \frac{z_q \cdot z_{s_j}}{\|z_q\| \|z_{s_j}\|}. \tag{3}$$

The support set $\mathbf{S}$ is constructed by selecting the top-N samples with the highest similarity scores c_j. This approach ensures structural consistency between the support and query, reducing the need for the segmentation network to resolve large morphological conflicts.

3.3. Segmentation

The segmentation backbone employs a parallel U-Net-like encoder–decoder structure with shared weights. The network consists of an encoder $\mathscr{E}$ that extracts multi-scale feature maps and a decoder $\mathscr{D}$ that reconstructs the segmentation mask.

To preserve spatial details inherent in ultrasound images, such as speckle patterns and subtle tissue interfaces, we utilize bilinear interpolation for all downsampling and upsampling operations. Unlike max-pooling, which discards spatial information to achieve translation invariance, bilinear interpolation maintains smoother feature transitions and mitigates aliasing artifacts, which is crucial for defining the often indistinct boundaries of ovarian tumors.

3.3.1. Cross-convolution

The core mechanism enabling the transfer of knowledge from the support set to the query is the Cross-Convolution layer. Let $u^l \in \mathbb{R}^{h\times w\times c}$ denote the feature map of the query image at layer l, and v_i^l denote the feature map of the i-th support image at the same resolution. The interaction is modeled by concatenating the query features with each support feature map, followed by a convolutional block:

$$z_i^l = \sigma(\mathrm{Conv}_{1\times 1}([u^l \parallel v_i^l]; \theta_z)), \tag{4}$$

where $[\cdot \parallel \cdot]$ denotes channel-wise concatenation, θ_z are learnable parameters, and $\sigma(\cdot)$ is a non-linear activation function (LeakyReLU). This operation computes a pixel-wise similarity and feature correlation map z_i^l between the query and the i-th support sample. Crucially, the weights θ_z are shared across all support samples, ensuring that the feature extraction is permutation invariant.

3.3.2. CrossBlock interaction

To aggregate information from the entire support set, we employ the CrossBlock mechanism as shown in Fig. 1. This block updates the representations of both the query and support branches by aggregating the interaction features calculated in the previous step. The update rule is defined as:

$$u' = \frac{1}{N}\sum_{i=1}^{N} \mathrm{A}(z_i), \tag{5}$$

$$v_i' = \mathrm{A}(\mathrm{Conv}_{3\times 3}(z_i; \theta_v)), \tag{6}$$

where $\mathrm{A}(\cdot)$ denotes an activation layer. Equation (5) is particularly significant: by averaging the interaction features over N, the model constructs a single, robust representation of the query that incorporates "consensus" information from all support examples. These Cross Blocks are embedded at multiple stages of the encoder, allowing the model to facilitate the propagation of task-specific information across diverse semantic scales, from low-level texture details to high-level semantic shapes.

3.4. Efficient Inference Strategy

The baseline UniverSeg model typically employs a K-ensemble strategy (often with $K = 50$) to stabilize predictions. This involves randomly sampling 50 different support sets, running the forward pass 50 times, and averaging the output logits. While this reduces the variance caused by random sampling, it increases the computational cost linearly by a factor of K.

In our framework, the CLIP-based selection yields a *deterministic* and highly relevant support set. Since the support samples are explicitly chosen to be the "best matches" for the query, the variance associated with random sampling is effectively eliminated. Consequently, the marginal gain from ensemble averaging is negligible. We therefore adopt a single-pass inference strategy ($K = 1$). This strategy reduces the inference time by approximately 98% compared to the standard $K = 50$ ensemble, making the framework highly suitable for real-time clinical deployment where low latency is required.

4. Experimental Results

This section presents the datasets, preprocessing pipeline, inference configuration, and evaluation metrics used to assess the performance of the proposed CLIP-guided UniverSeg framework.

4.1. Datasets

Experiments are conducted on two ovarian ultrasound datasets OvaTUS [13] and OTU_2D [14]. Information of these dataset are shown in Table 1.

Table 1. Class distribution of OvaTUS and OTU_2D datasets

Dataset	Class	Images per Class
OvaTUS	Solid tumor	119
	Multilocular cyst	151
	Unilocular cyst	107
	Dermoid cyst	116
	Multilocular-solid cyst	59
	Unilocular-solid cyst	31
OTU_2D	Chocolate cyst	336
	Serous cystadenoma	219
	Teratoma	336
	Theca cell tumor	88
	Simple cyst	66
	Normal ovary	267
	Mucinous cystadenoma	104
	High grade serous	53

OvaTUS, introduced in [13], comprises 583 two-dimensional ultrasound images acquired from 293 patients. Each image is annotated with expert-verified polygonal masks delineating ovarian tumors and cystic structures. The dataset contains six tumor categories and exhibits notable class imbalance, closely reflecting real-world clinical distributions. All annotations are validated by clinicians based on diagnostic records, ensuring high labeling reliability.

OTU_2D is a publicly available subset of the MMOTU dataset [14], consisting of 1,469 2D ultrasound images collected from 294 patients across eight ovarian tumor categories. Annotations are provided by an experienced radiologist and independently reviewed by another specialist. OTU_2D presents pronounced class imbalance, with several categories containing fewer than 100 samples. In addition, tumors typically occupy less than 2% of the image area, posing substantial challenges for accurate segmentation.

To prevent data leakage, both datasets are split at the patient level into a support pool and a query set, as summarized in Table 2. All experiments are conducted in an inference-only few-shot setting without any fine-tuning of the pretrained segmentation backbone.

Table 2. Dataset split configuration

Dataset	Split	Images	Patients	Ratio
OvaTUS	Support	458	234	80%
	Query	125	59	20%
OTU_2D	Support	1000	171	68%
	Query	469	76	32%

All images undergo identical preprocessing. Images are resized to 128×128 pixels and normalized to the range $[0,1]$. RGB inputs are converted to single-channel grayscale via channel averaging. Polygon annotations are rasterized into binary masks. Segmentation is performed in a class-wise manner, where each category is processed independently.

4.2. Inference Strategy and Post-Processing

Although UniverSeg originally adopts a K-ensemble strategy to reduce randomness from support sampling, the proposed CLIP-guided selection yields deterministic support sets. Therefore, we set $K = 1$ during inference in all experiments to reduce computational cost.

The output probability map is binarized using a fixed threshold $T = 0.45$, chosen empirically based on the Precision–Recall curve to balance precision and recall under severe foreground–background imbalance. Finally, we combine the binary masks from each class to get the final multiclass segmentation.

4.3. Evaluation Metrics

Segmentation performance is evaluated using standard pixel-wise metrics, including Dice coefficient (DSC), Intersection over Union (IoU), Precision, and Recall. These complementary metrics provide a comprehensive evaluation of segmentation accuracy and robustness, particularly under class-imbalanced conditions.

5. Experimental Results

In this section, we present a comprehensive evaluation of the proposed CLIP-guided support set selection strategy integrated into the UniverSeg framework. We conduct quantitative comparisons against the baseline random selection strategy and state-of-the-art few-shot segmentation methods.

5.1. Effectiveness of CLIP-Guided Support Selection

To demonstrate the efficacy of our proposed method, we compare the segmentation performance of the baseline UniverSeg (Random Selection) against our UniverSeg + CLIP approach across varying support set sizes ($N \in \{1,2,4,8,16,32,64\}$). Table 3 summarizes the results for both OvaTUS and OTU_2D datasets.

The results indicate that our CLIP-based strategy consistently outperforms the random selection baseline across both datasets, particularly in low-shot scenarios.

On OvaTUS dataset, in the extreme few-shot setting ($N = 1$), our method achieves a Dice score of 44.89% and IoU of 30.77%, significantly surpassing the baseline (36.72% Dice, 23.94% IoU). This improvement suggests that when support data is scarce, retrieving a semantically similar image provides critical guidance for the segmentation network. As N increases ($N = 8$), the proposed method maintains its advantage (Dice 71.91% vs. 71.28%). At $N = 64$, the performance gap narrows but remains stable for our method (Dice 75.84%). Notably, Recall reaches its peak at 89.16%, reflecting the model's ability to minimize false negatives, which is a crucial factor in medical diagnosis.

Regarding OTU_2D dataset, a similar trend is observed on the larger OTU_2D dataset. The 1-shot performance is boosted from 32.41% to 39.47%. By $N = 64$, our method achieves a Dice score of 75.23% compared to 72.92% for the baseline. The consistent improvement in Recall (reaching 84.38% at $N = 64$) demonstrates that support selection mechanism helps the model capture the entire tumor region more effectively, reducing under-segmentation errors.

5.2. Class-Wise Performance Analysis

To further assess the model's generalization capability, we analyze the segmentation performance on specific ovarian tumor subclasses at $N = 64$.

5.2.1. OvaTUS dataset

Table 4 details the performance across different tumor types. The model performs exceptionally well on multilocular cysts and unilocular cysts, achieving Dice scores of 88.53% and 84.82%, respectively. This is attributed to the relatively clear boundaries of fluid regions in ultrasound.

However, performance drops for solid components. Solid Tumors present the biggest challenge, with a Dice score of 55.97% and Precision of 48.39%. Despite the low Precision, the Recall remains high (85.57%), indicating that the model prioritizes coverage of the tumor, leading to some over-segmentation but ensuring the tumor is not missed.

Table 3. Performance comparison between the baseline UniverSeg and the proposed CLIP-guided method across varying support set sizes (N) on OvaTUS and OTU_2D, with best results per metric highlighted in bold

N	OvaTUS Dataset								OTU_2D Dataset							
	Baseline (Random)				*Ours (CLIP-Guided)*				*Baseline (Random)*				*Ours (CLIP-Guided)*			
	Dice	**IoU**	**Precision**	**Recall**	**Dice**	**IoU**	**Precision**	**Recall**	**Dice**	**IoU**	**Precision**	**Recall**	**Dice**	**IoU**	**Precision**	**Recall**
1	36.72	23.94	52.88	33.51	44.89	30.77	55.42	44.58	32.41	20.88	39.73	37.62	39.46	27.22	46.54	43.11
2	58.79	45.46	62.31	64.08	61.94	48.25	63.76	68.17	45.72	33.34	51.58	51.74	56.83	45.06	60.31	61.59
4	68.35	55.97	65.22	79.81	68.82	56.31	67.58	78.74	55.68	43.49	59.04	62.91	65.57	54.28	66.49	72.06
8	71.28	59.26	66.34	85.93	71.91	60.14	69.27	82.65	62.23	50.47	62.19	71.86	69.85	59.08	69.17	77.25
16	72.74	61.32	67.55	87.76	73.68	62.39	69.41	85.97	67.04	55.58	65.98	77.21	72.56	62.09	71.48	80.87
32	74.22	62.83	69.31	88.45	75.16	64.28	70.15	88.19	70.47	59.76	69.35	80.52	74.29	64.25	72.86	83.38
64	73.65	62.18	70.96	86.05	**75.84**	**64.87**	**70.79**	**89.16**	72.88	62.61	71.93	82.26	**75.22**	**65.37**	**73.75**	**84.42**

Table 4. Class-wise segmentation performance (%) on the OvaTUS dataset ($N = 64$), with best results per metric highlighted in bold

Class	**Dice**	**IoU**	**Precision**	**Recall**
Multilocular Cyst	**88.53**	**80.64**	**87.46**	92.27
Unilocular Cyst	84.82	75.83	80.24	92.06
Solid Multilocular	77.54	65.67	68.73	**93.25**
Solid Unilocular	75.08	64.86	70.57	89.14
Dermoid Cyst	72.95	60.18	69.54	83.23
Solid Tumor	55.97	42.68	48.39	85.57

5.2.2. OTU_2D dataset

Table 5 shows the results obtained for OTU_2D dataset classes. Results illustrate strong performance on Serous Cystadenoma (Dice 89.27%) and Mucinous Cystadenoma (Dice 90.37%), likely due to their distinct morphological features. Conversely, the Ovary Normal class yields the lowest performance (Dice 50.15%), as distinguishing normal ovarian tissue from surrounding background noise without a defined tumor boundary is inherently difficult.

Table 5. Class-wise segmentation performance (%) on the OTU_2D dataset ($N = 64$), with best results per metric highlighted in bold

Class Name	**Dice**	**IoU**	**Precision**	**Recall**
Chocolate Cyst	77.51	68.19	74.81	87.84
Serous Cystadenoma	89.27	82.22	**87.89**	93.18
Teratoma	76.99	65.62	73.94	86.23
Theca Cell Tumor	81.49	72.05	76.48	91.29
Simple Cysts	79.98	71.06	79.01	86.30
Ovary Normal	50.15	38.33	54.40	64.29
Mucinous Cystadenoma	**90.37**	**83.46**	87.83	**94.22**
High-grade Serous	77.19	65.52	71.97	85.28

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

Table 6 presents the comparative results our proposed method against other few-shot segmentation models (ALPNet, PANet).

On the OvaTUS dataset, our method outperforms traditional few-shot approaches by a large margin. Specifically, we achieve a Dice score of 75.84%, compared to 63.54% for ALPNet and 58.92% for PANet. The significant boost in IoU (64.87% vs 49.82% for ALPNet) confirms that our model produces segmentation masks that overlap much more accurately with the ground truth.

On the OTU_2D dataset, the superiority of the proposed method is maintained. Our model surpasses ALPNet and PANet by over 20% in Dice score. The fact that our method narrows the gap while offering the flexibility of zero-training adaptation highlights its potential for clinical applications where annotated data is scarce.

5.4. Qualitative Evaluation

Fig. 2 presents a visualization of the segmentation results obtained by our CLIP-guided UniverSeg model on the OvaTUS test set. The figure illustrates the model's predictions (green) overlaid with the ground truth annotations (red).

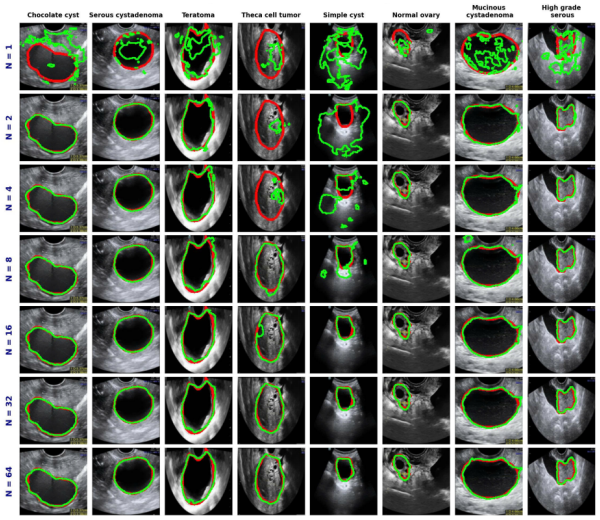

Fig. 2. Segmentation results of the proposed framework on the OTU_2D. The predicted contours are shown in green, while the ground truth contours are shown in red

As observed in Fig. 3, the proposed framework demonstrates robust segmentation capabilities across various tumor morphologies. The visualization clearly illustrates the progressive refinement of the segmentation masks as the support set size (N) increases. At extreme few-shot settings ($N = 1$ or $N = 2$), the predicted contours often exhibit fragmentation, leakage,

Table 6. Quantitative comparison of segmentation performance on the OvaTUS and OTU_2D datasets across different few-shot models, with best results highlighted in **bold** (%)

Dataset	Method	Dice	IoU	Recall	Precision
OvaTUS	ALPNet	63.54	49.82	81.03	58.61
	PANet	58.92	43.86	88.24	48.19
	UniverSeg	73.71	62.28	86.05	**71.02**
	CLIP-Guided (Ours)	**75.84**	**64.87**	**89.21**	70.79
OTU_2D	ALPNet	50.38	36.64	70.87	46.92
	PANet	45.91	31.86	66.85	43.77
	UniverSeg	75.18	65.37	84.36	73.84
	CLIP-Guided (Ours)	**75.22**	**65.41**	**84.42**	**73.89**

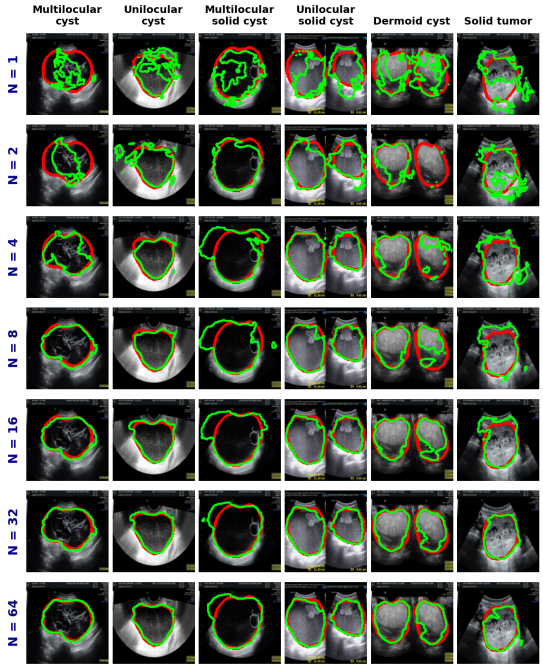

Fig. 3. Segmentation results of the proposed framework on OvaTUS, where predicted contours are shown in green and ground truth contours in red

or struggle to capture the full extent of complex lesions, as seen in the Simple cyst and Teratoma examples. However, as N scales up to 16, 32, and 64, the model successfully overcomes significant speckle noise and shadowing artifacts, yielding smooth and precise boundaries that tightly adhere to the ground truth annotations across all classes. Notably, with a sufficient support set, the framework effectively delineates the heterogeneous internal structures and irregular margins typical of Chocolate cysts, Theca cell tumors, and High grade serous lesions, which are common areas of failure of traditional segmentation methods.

This qualitative performance indicates that the semantic guidance provided by CLIP-selected support images enables the network to effectively distinguish the tumor from the surrounding ovarian. By anchoring the segmentation process to an increasing number of morphologically relevant prototypes, the model achieves a high degree of spatial consistency. This minimizes both the under-segmentation of the tumor body and leakage into background tissues that typically occur when support data is severely limited, ultimately producing the highly accurate contours seen at $N = 64$. These visual results strongly correlate with the the results reported in Table 3, confirming the practical applicability of our approach for ovarian tumor delineation.

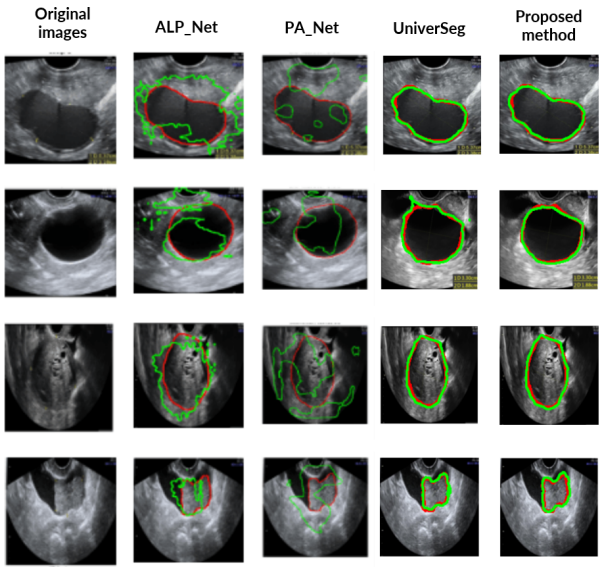

Fig. 4. Qualitative comparison of segmentation results with state-of-the-art few-shot methods, where predicted contours are shown in green and ground truth contours in red

Fig. 4 provides several examples of segmentation results produced by our proposed method in comparison with state-of-the-art few-shot segmentation models, including ALPNet, PANet, and the baseline UniverSeg. As observed, traditional prototype-based methods (ALPNet and PANet) struggle with speckle noise and low contrast, frequently resulting in under-segmentation and fragmented masks, particularly in regions with acoustic shadowing.

While the baseline UniverSeg improves performance, its random support selection introduces instability; the retrieval of morphologically disparate images often leads to hallucinations and imprecise boundary predictions. In contrast, our CLIP-Guided

Table 7. Performance of the proposed framework using BioMedCLIP as the feature encoder for support set retrieval

N	OvaTUS (BiomedCLIP)				OTU_2D (BiomedCLIP)			
	Dice	**IoU**	**Recall**	**Precision**	**Dice**	**IoU**	**Recall**	**Precision**
1	43.33	29.59	43.47	53.67	41.94	29.21	45.56	48.39
2	60.27	46.75	66.26	63.18	57.91	46.38	62.99	59.87
4	67.16	54.58	77.47	66.03	67.14	56.53	73.05	66.71
8	69.77	57.96	80.20	67.61	71.85	61.38	79.55	70.33
16	72.49	61.04	84.78	68.47	74.47	64.32	82.98	72.47
32	73.75	62.71	87.32	68.86	75.89	66.22	84.23	74.05
64	74.41	63.44	88.65	69.46	76.07	66.37	84.94	74.37

strategy consistently generates the most accurate masks. By retrieving morphologically consistent support samples, the model reinforces structural priors, resulting in smoother contours that closely align with the Ground Truth across both complex cystic and solid tumor structures.

5.5. Feature Encoder Comparison

Although the proposed framework employs CLIP as the feature encoder for support retrieval, it is not restricted to a specific visual encoder. In principle, any encoder can be integrated into the framework. To further investigate this aspect, we additionally conduct an experiment using BioMedCLIP [15] as an alternative encoder. BioMedCLIP is a vision–language model specifically designed for the biomedical domain, trained on large-scale biomedical image–text pairs, which enables it to capture domain-specific visual semantics more effectively. To ensure a fair comparison, we keep the retrieval pipeline unchanged.

Results are reported in Table 7. From these results, we observe that the impact of the encoder varies across datasets. On the OTU_2D dataset, BioMedCLIP slightly improves segmentation performance across most support sizes. For instance, at $N = 64$, BioMedCLIP achieves a Dice score of 76.07%, compared with 75.22% obtained by CLIP. This suggests that biomedical pretraining helps capture morphological patterns that are more relevant to ovarian tumor structures in this dataset.

However, on the OvaTUS dataset, BioMedCLIP does not outperform CLIP encoder and yields slightly lower Dice and IoU scores. One possible explanation is that the visual characteristics of OvaTUS images, including stronger speckle noise and higher intra-class variability, may benefit more from the broader visual representations learned by CLIP from large-scale natural images.

Overall, these observations suggest that while domain-specific encoders such as BioMedCLIP can improve retrieval quality in certain scenarios, the effectiveness of support retrieval remains dataset-dependent. Nevertheless, both encoders consistently outperform random support selection, confirming that similarity-guided retrieval plays a critical role in improving the robustness of few-shot segmentation.

6. Conclusions and Future Works

This paper proposes a CLIP-guided support selection strategy to improve few-shot ovarian tumor segmentation within the UniverSeg framework. Instead of random sampling, semantic retrieval based on CLIP embeddings selects more relevant support examples, providing stronger structural priors under limited supervision. Experiments on OvaTUS and OTU_2D show consistent improvements over the baseline across different support sizes, with the largest gains in extreme low-shot settings. The method also achieves competitive or superior performance compared to ALPNet and PANet. Quantitative and qualitative results demonstrate better boundary delineation, reduced under-segmentation, and improved robustness to ultrasound artifacts such as speckle noise and low contrast. However, the framework relies on CLIP pretrained on natural images, which may not fully capture ultrasound-specific characteristics. Future work will investigate domain-adapted vision-language models, more efficient multi-class inference, and learnable support selection to further enhance robustness under extremely limited supervision.

Acknowledgment

This research is funded by the Ministry of Science and Technology (MOST) under grant number KC-4.0-45/19-25 "Research and development of a computer-aided support system for ovarian cancer diagnosis using ultrasound images".

References

[1] H.-S. Bui, S.-H. Tran, T.-B. Nguyen, T.-H. Tran, H. Vu, and T.-L. Le, Marker-aware ovarian tumor segmentation from ultrasound images, in 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2024, pp. 1–6.

[2] S. Wei, Z. Hu, and L. Tan, Res-ECA-UNet++: an automatic segmentation model for ovarian tumor ultrasound images based on residual networks and channel attention mechanism, Frontiers in Medicine, vol. Volume 12 - 2025, 2025. https://doi.org/10.3389/fmed.2025.1589356

[3] H.-P. Luong, H.-S. Bui, N.-K. Nguyen, T.-L. Pham, G.-M. Pham, S.-H. Tran, T.-H. Tran, and T.-L. Le, SovaSeg-net: scale invariant ovarian tumors segmentation from ultrasound images, in 2024 IEEE International Conference on Image Processing (ICIP), 2024, pp. 2081–2087.

[4] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, Universeg: Universal medical image segmentation, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21438–21451.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in International Conference on Machine Learning, 2021.

[6] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241.

[7] T.-L. Pham, H.-S. Bui, V.-H. Le, T.-L. Le, V.-T. Nguyen, C.-M. Pham, H.-T. Dang, T.-A. Nguyen, D.-H. Vu, H. Vu, and T.-H. Tran, BKSeg-Net: Segmentation of Ovarian Tumors from Ultrasound Images with Boundary Keypoints Loss, in 2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2024, pp. 1–6.

[8] T.-P. Dao, H.-T. To, H.-S. Bui, and T.-L. Le, Data Augmentation-Driven Segmentation of Ovarian Tumor Ultrasound Images Using Vision Mamba, in 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2025, pp. 2299–2304.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in 2009 IEEE conference on computer vision and pattern recognition, 2009, pp. 248–255.

[10] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, One-Shot Learning for Semantic Segmentation, ArXiv, vol. abs/1709.03410, 2017. https://api.semanticscholar.org/CorpusID:23237949

[11] A. Guha Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, 'Squeeze excite' guided few-shot segmentation of volumetric images, Medical Image Analysis, vol. 59, pp. 101587, 2020. https://doi.org/10.1016/j.media.2019.101587

[12] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, Panet: Few-shot image semantic segmentation with prototype alignment, in proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9197–9206.

[13] N.-K. Nguyen, H.-S. Bui, T.-L. Pham, T. T. Thao Nguyen, V. Hai, T. T. Hai Tran, V.-T. Nguyen, C.-M. Pham, H.-T. Dang, and T.-L. Le, A method for ovarian tumor segmentation based on segment anything model, in 2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2024, pp. 1–6.

[14] S. Lyu, Q. Zhao, W. Bai, L. Cai, G. Cheng, G. Cui, M. Yang, L. Chen, and H. Zhou, Unsupervised cross-domain semantic segmentation on multi-modality ovarian tumor ultrasound data, Pattern Recognition, vol. 171, pp. 112311, 2026. https://doi.org/10.1016/j.patcog.2025.112311

[15] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon, A multimodal biomedical foundation model trained from fifteen million image–text pairs, NEJM AI, vol. 2, no. 1, 2024. https://doi.org/10.1056/AIoa2400640