# UniOCTSeg: Towards Universal OCT Retinal Layer Segmentation via Hierarchical Prompting and Progressive Consistency Learning

Jian Zhong[1], Li Lin[1,2⋆], Chaoran Miao[1], Kenneth K. Y. Wong[2], and Xiaoying Tang[1,3]

[1] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
tangxy@sustech.edu.cn
[2] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China
[3] Jiaxing Research Institute, Southern University of Science and Technology, Jiaxing, China

**Abstract.** Accurate segmentation and quantitative thickness analysis of retinal layers in optical coherence tomography (OCT) are crucial for early diagnoses of ocular disorders. To address the clinical needs of diagnosing various ocular and systemic diseases, numerous multi-granularity OCT datasets are constructed. While deep learning achieves impressive results in retinal layer segmentation, general training paradigms require separate models for datasets with different annotation granularities. Universal models are developed to unify diverse datasets and tasks via advanced techniques such as prompt learning, but they overlook across-granularity information and struggle to generalize to new granularities. In this paper, we propose a universal OCT segmentation model, named UniOCTSeg, which builds its basis upon Hierarchical Prompting Strategy (HPS) and Progressive Consistency Learning (PCL). HPS employs a granularity-merging strategy to construct prompts at various granularities, based on the finest-grained prompts, and develops a universal segmentation model that utilizes these hierarchical prompts. Meanwhile, PCL leverages an Exponential Moving Average teacher model to generate pseudo-supervision signals, guiding the student model through easy-to-hard progression to ensure consistency across hierarchical levels. Extensive experiments across eight publicly available OCT datasets involving six distinct granularity levels demonstrate UniOCTSeg's superior performance compared with state-of-the-art methods, while also illustrating its high flexibility and strong generalizability. Our code and data are available at https://github.com/Halcyon1010/UniOCTSeg.

**Keywords:** Universal model · Hierarchical prompting · Progressive consistency learning · Retinal layer segmentation.

⋆ J. Zhong and L. Lin contributed equally to this work.

## 1 Introduction

Certain ocular [1,2,3] and systemic [4] diseases may exhibit characteristic changes in retinal layer thickness in their early stages. Precise segmentation and analysis of retinal layers' thickness changes on optical coherence tomography (OCT) are thus important, which can help early diagnosis and disease monitoring. To address diverse clinical needs, a variety of OCT retinal layer segmentation datasets with annotations of different layer granularities are constructed (Fig. 1a).

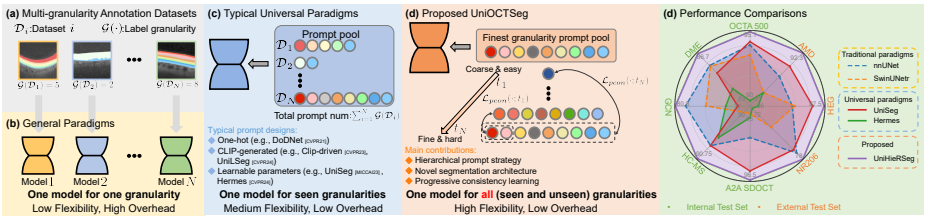


Fig. 1: **(a)** Multi-granularity annotation datasets constructed for diverse clinical needs. **(b)** and **(c)** represent general and universal segmentation paradigms, respectively. **(d)** Our proposed UniOCTSeg and its strengths. **(e)** Performance comparisons between different segmentation paradigms and UniOCTSeg.

Recently, deep learning-based retinal layer segmentation in OCT is extensively studied, demonstrating remarkable performance. Existing methods [6,7,8,9] mainly focus on model architecture design, loss function optimization, and utilization of topological information. These methods are nevertheless confined to single-granularity datasets under general training paradigms, requiring specialized models for distinct tasks and preventing combined use of datasets with different annotation granularities (Fig. 1b). This results in inflexibility, high computational costs, and limited performance.

Universal segmentation models [10,11,12,13,14,15] draw significant attention as a promising solution, leveraging prompt learning to adapt to across-dataset tasks. They mainly employ: (1) one-hot prompts [10]; (2) CLIP-driven prompts [11,13]; or (3) learnable prompts [12,14,15]. For datasets with varying granularities, existing universal models can achieve multi-granularity segmentation at a relatively low training cost by assigning unique prompts to each granularity (Fig. 1c), demonstrating a certain degree of flexibility. However, they tend to overlook the inherent correlations among tasks involving varying granularities. Moreover, existing universal segmentation methods are limited to granularities encountered during training. When new annotation schemes arise, the model must be retrained or its architecture gets expanded, thus limiting its flexibility.

In such context, this paper proposes a universal OCT segmentation model, namely UniOCTSeg, as shown in Fig. 1d. In UniOCTSeg, Hierarchy Prompt Strategy (HPS) is designed to generate multi-granularity prompts based on hierarchical priors of the retinal layers, leveraging the finest-grained prompts.

Meanwhile, Progressive Consistency Learning (PCL) is incorporated to establish coherent consistency among tasks of different granularities, so as to achieve more accurate and robust segmentation.

This work's main contributions are three-fold: (1) To the best of our knowledge, we introduce the first universal segmentation framework for OCT retinal layer analysis, which leverages hierarchical dependencies and combinatorial consistency across retinal layers. (2) We propose HPS to achieve comprehensive compatibility with both seen and unseen multi-granularity datasets through finest-grained prompts and a hierarchical merging strategy. We also develop PCL to bridge task-related granularity gaps, thereby improving the consistency across tasks and enhancing segmentation robustness. (3) Extensive experiments on eight publicly available OCT datasets across six distinct granularies demonstrate the superiority of UniOCTSeg over other state-of-the-art (SOTA) methods, as shown in Fig. 1e.

## 2 Method

**Problem definition.** Given a set of $N$ datasets with retinal layer segmentation of different granularities, denoted as $\bigcup_{i=1}^{N} \mathcal{D}_i(x, y)$, where $\mathcal{D}_i$ is the *i-th* dataset, $x$ is an OCT image with the width and height of $W$ and $H$, and $y$ is the corresponding ground truth retinal layer segmentation. The granularity level of $\mathcal{D}_i$ is defined as $\mathcal{G}(\mathcal{D}_i)$, namely the number of annotated layers in that dataset. The proposed UniOCTSeg (illustrated in Fig. 2) tackles hierarchical retinal layer segmentation through a novel universal segmentation network architecture (**Sec.** 2.1), an HPS module (**Sec.** 2.2), and a PCL module (**Sec.** 2.3).

### 2.1 Universal Segmentation Network Architecture

As illustrated in Fig. 2a, our network has two key components: 1) an architecture with a vision encoder and a pixel decoder that extracts multi-scale image features; 2) a prompt decoder which captures relationships among the finest-grained/basic prompts and the multi-stage outputs from the pixel decoder.

**Vision Encoder and Pixel Decoder.** UniOCTSeg employs a hybrid vision encoder to extract features from OCT images, integrating a convolutional neural network (CNN) encoder and a transformer encoder. The CNN encoder consists of five blocks, each comprising two convolution layers followed by instance normalization and ReLU activation. After processing, the output features $\mathcal{F}_{cnn} \in \mathbb{R}^{C \times \frac{W}{32} \times \frac{H}{32}}$, where $C$ is the number of channels, are reshaped into $\mathcal{F}_T \in \mathbb{R}^{\frac{WH}{32 \times 32} \times C}$ before being sent to the transformer encoder which adopts the ViT-B-16 architecture and pretrained on the ImageNet-21k dataset. Finally, the outputs of the transformer encoder are shaped as $\mathcal{F} \in \mathbb{R}^{C \times \frac{W}{32} \times \frac{H}{32}}$. These outputs are then fed into both the pixel decoder and the prompt decoder. The pixel decoder, mirroring the CNN encoder's five-block structure, utilizes skip connections to upsample multi-scale features from the vision encoder, thereby generating the final decoding features $\mathcal{F}_p \in \mathbb{R}^{C_{out} \times W \times H}$, where $C_{out}$ is the number of

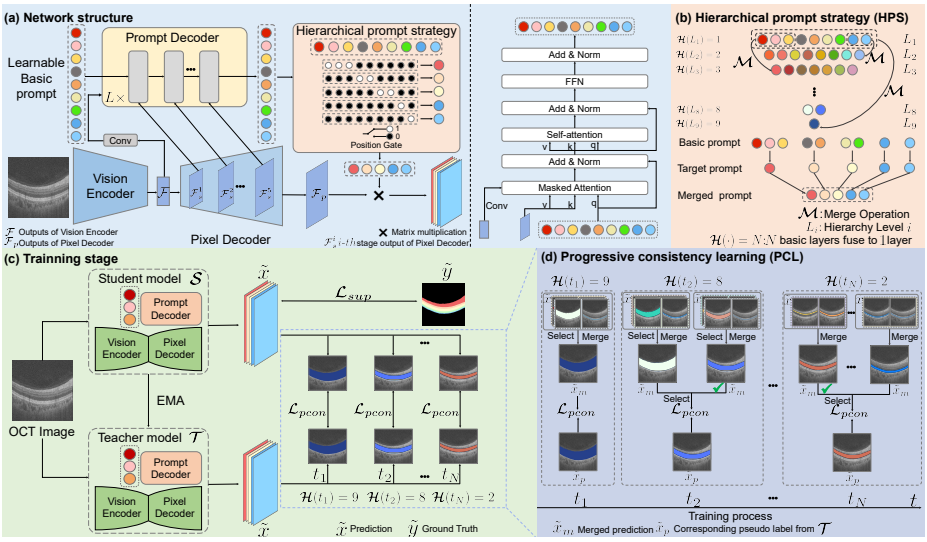


Fig. 2: Overview of UniOCTSeg. (a) UniOCTSeg's network architecture. **(b)** HPS for multi-granularity prompt generation from basic prompts. **(c)** Training procedure of UniOCTSeg, where $\mathcal{T}$ is incorporated to achieve pseudo-supervised consistency. **(d)** PCL enforcing across-granularity consistency.

output channels. Concurrently, multi-stage outputs $\{\mathcal{F}_s^1, \mathcal{F}_s^2, \ldots, \mathcal{F}_s^5\}$ from the pixel decoder are forwarded to the prompt decoder to facilitate prompt learning.

**Prompt Decoder.** The prompt decoder is constructed using $L$ specially designed transformer decoder layers that incorporate masked attention and self-attention operations [16]. Additionally, nine learnable basic prompts are introduced, denoted as $\bigcup_{l=1}^{9} \mathcal{P}_b^l \in \mathbb{R}^{\frac{W}{32} \times \frac{H}{32}}$, corresponding to nine basic retinal layers. The basic retinal layers are represented as $R_b = \{R_b^l, l = 1, 2, \cdots, 9\}$, which are detailed in [21]. The basic prompts are fed into the prompt decoder to get updated. With this configuration, the prompt decoder effectively integrates the pixel decoder's multi-stage features and captures the interrelationships among the basic prompts. After being processed by the prompt decoder, the basic prompts are integrated through our proposed HPS, which generates multi-granularity prompts. This design ensures a comprehensive understanding of the retinal layers, enhancing a model's ability to generate detailed and contextually relevant outputs.

## 2.2 Hierarchy Prompt Strategy (HPS)

To flexibly representing retinal segmentation tasks involving various granularities, we design HPS (see Fig. 2b). This strategy considers the retinal segmentation task may correspond to a total of nine hierarchical levels, with $L_i$ representing the *i-th* level. Each level is characterized by a distinct combination of the

basic prompts, with the exception of $L_1$ which directly corresponds to the basic retinal layers defined by [21]. To assure the prompt position information aligns with the retinal layers, we construct a position gate using 1 and 0 to respectively represent turn on and turn off. When representing a target prompt, the basic prompts are selected to be turned on or off by the position gate and fed into a convolution layer where they are merged into the target prompt of interest. This merging method, denoted as $\mathcal{M}$, is expressed as:

$$\mathcal{M} : \text{Conv}(\mathcal{P}_b^i \odot G^i) = \mathcal{P}^T \tag{1}$$

where $\text{Conv}(\cdot)$ denotes a convolution layer, $\odot$ represents the Hadamard product, $\mathcal{P}^T$ is the target prompt, $i$ represents the $i\text{-}th$ basic prompt and $G$ is the corresponding position gate. Specifically, the hierarchical prompt at each level consists of $N$ basic prompts, denoted as $\mathcal{H}(\cdot) = N$. In this way, multi-granularity prompts are uniformly represented using learnable basic prompts. The basic prompts, updated by the prompt decoder, are merged into prompts corresponding a target granularity, thereby forming prompts suitable for representing the desired granularity. For example, the merged result of $\mathcal{P}_b^1$, $\mathcal{P}_b^2$ and $\mathcal{P}_b^3$ is given as:

$$\mathcal{P}_M = \mathcal{M}(\mathcal{P}_b^1, \mathcal{P}_b^2, \mathcal{P}_b^3) \tag{2}$$

where $\mathcal{P}_M$ is the merged prompts, representing the retinal layer combination of $R_b^1$, $R_b^2$ and $R_b^3$. Finally, $\mathcal{P}_M$ gets multiplied with $\mathcal{F}_p$, outputting the retinal layer segmentation of the target granularity. Theoretically, HPS can generalize to retinal layer segmentation tasks of any granularity. Therefore it facilitates scalable and efficient representations of segmentation tasks across granularities and datasets, thereby enhancing the model's adaptability and versatility.

### 2.3 Progressive Consistency Learning (PCL)

To strengthen the correlation among hierarchical segmentation tasks, we utilize a novel training method termed PCL (see Fig. 2d), which is based on a teacher-student training paradigm as shown in Fig. 2c. The paradigm generates pseudo-supervised signals by aligning the merged results of a student model $\mathcal{S}$'s hierarchical outputs with the target pseudo-label produced by an exponential moving average (EMA) teacher model $\mathcal{T}$, thereby facilitating consistency learning across multi-granularity tasks.

Specifically, two distinct hierarchical outputs, $\tilde{x}_i$ and $\tilde{x}_j$, are randomly selected from $\mathcal{S}$ and merged according to Eq. 3. The merged result, denoted as $\tilde{x}_m$, is then aligned with the pseudo-label $\tilde{x}_p$ from $\mathcal{T}$ by applying the progressive consistency loss $\mathcal{L}pcon$ as given in Eq. 4, where $t_N$ represents the current training iteration, and the granularity of $\tilde{x}_m$ matches that of $\tilde{x}_p$. Other than $\mathcal{L}_{pcon}$, a supervised loss $\mathcal{L}_{sup}$ is also employed, which is the combination of the Dice loss and the binary cross-entropy loss.

$$\tilde{x}_m = \text{Argmax}(\text{Concat}(\tilde{x}_i, \tilde{x}_j)), (i \neq j). \tag{3}$$

$$L_{pcon}(\tilde{x}_m, \tilde{x}_p; t_N) = 1 - \frac{2 \cdot (\tilde{x}_m \tilde{x}_p)}{\tilde{x}_m + \tilde{x}_p}. \quad (4)$$

The proposed progressive training strategy ensures stable convergence through gradual difficulty incrementing scheduling, and enables hierarchical consistency across granularities. Initially, the target pseudo-label would be chosen from a coarse granularity, providing robust initialization for consistency alignment. As training progresses, the proposed UniOCTSeg escalates to fine-grained tasks, introducing controlled complexity via randomized sub-level merging tasks and more difficult alignments. The evolving consistency constraints establish dependencies across different segmentation levels, enabling the model to develop representations that are more aligned with the retinal layers' anatomy.

## 3 Experiments

**Datasets.** For this study, we collect five publicly available OCT retinal segmentation datasets as the internal sets (A2A SDOCT [17], OCTA500 [18], DME [19], GCN [20], HC-MS [21]), respectively containing {384, 500, 10, 244, 35} samples with {2, 5, 7, 8, 8} granularity levels. We also collect three other publicly available datasets (AMD [22], HEG [23], NR206 [24]), which involve {20, 100, 206} samples with {2, 7, 8} granularities, respectively, as the external test sets. Note that both HC-MS [21] and GCN [20] contain seven basic layers and one merging layer. However, due to the difference in their two merging layers, the training set includes all nine basic layers, as defined in our method. We divide the internal training, validation and test sets in a 7:1:2 ratio, after which they are preprocessed to normalized $512 \times 512$ slices. Data augmentation is applied to the internal training sets, including horizontal flipping, random Gaussian noise addition, and random brightness and contrast.

**Implementation Details.** Experiments are implemented in PyTorch using four A6000 GPUs. Optimization is performed for 80,000 iterations with the Adam optimizer and a batch size of 24. The learning rate starts at 0.0001 and gets annealed using polynomial decay with $power = 0.9$ at each iteration.

**Comparison with SOTAs.** To comprehensively evaluate the performance of UniOCTSeg, we conduct extensive comparisons with both advanced general models including UNet [25], SwinUnetr [26], nnUNet [27], and retinal segmentation-specific models, including YNet [7], LightReSeg [9], and TCCT [8]. We also compare with universal models, namely UniSeg [12], UniLSeg [13], and Hermes [14]. All methods are evaluated by two metrics: the Dice Similarity Coefficient (DSC) to assess region accuracy, and the 95% Hausdorff Distance (HD95) to quantify boundary shape agreement.

As shown in Table 1, UniOCTSeg outperforms other models, achieving higher DSC and lower HD95 across multi-granularity tasks. Quantitatively, it surpasses the second-best models by 0.29%, 0.66%, 0.82%, 0.51%, and 0.43% in DSC on all five datasets, and by 0.08 and 0.05 in HD95 on A2A SDOCT [17] and HC-

Table 1: Comparisons with general and universal models on the internal test sets, as evaluated by DSC (%) and HD95. The best results are in **bold** while the second best ones are underlined.

| Method | A2A SDOCT | | OCTA 500 | | DME | | GCN | | HC-MS | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DSC ↑ | HD95 ↓ | DSC ↑ | HD95 ↓ | DSC↑ | HD95 ↓ | DSC ↑ | HD95 ↓ | DSC ↑ | HD95 ↓ | DSC ↑ | HD95 ↓ |
| general models | | | | | | | | | | | | |
| UNet | 94.07±0.06 | 2.19±2.79 | 93.21±0.08 | 3.24±12.40 | 84.06±0.07 | 2.85±3.65 | 78.06±0.10 | 3.51±3.34 | 88.85±0.06 | 3.80±4.71 | 87.65±0.07 | 3.12±5.38 |
| SwinUNetr | 94.25±0.05 | 1.64±2.39 | 93.46±0.08 | 2.28±10.62 | 85.75±0.04 | 2.52±3.08 | 79.23±0.10 | 3.21±3.51 | 89.21±0.05 | 1.95±3.66 | 88.38±0.06 | 2.4±4.65 |
| nnUNet | 95.03±0.04 | 0.87±0.59 | 94.12±0.07 | 1.76±9.22 | 85.81±0.03 | 2.04±1.20 | 80.26±0.11 | 2.77±2.52 | 90.30±0.04 | 1.16±3.20 | 89.10±0.06 | 1.72±3.35 |
| specially designed models for retinal layers segmentation | | | | | | | | | | | | |
| YNet | 93.85±0.06 | 2.07±3.73 | 93.57±0.08 | 2.61±13.87 | 83.49±0.05 | 6.71±6.47 | 78.49±0.10 | 4.00±7.29 | 89.17±0.05 | 2.25±3.25 | 87.71±0.07 | 3.53±6.92 |
| LightReSeg | 94.46±0.04 | 1.88±1.67 | 93.23±0.08 | 2.51±12.37 | 85.66±0.04 | 2.68±3.18 | 78.37±0.12 | 3.26±3.26 | 90.18±0.04 | 1.07±1.30 | 88.38±0.06 | 2.28±4.36 |
| TCCT | 94.88±0.05 | 1.39±1.19 | 93.79±0.08 | 2.12±11.80 | 85.64±0.04 | 2.54±1.91 | 78.90±0.11 | 3.49±4.16 | 89.00±0.06 | 2.08±2.65 | 88.44±0.06 | 2.12±0.72 |
| universal models | | | | | | | | | | | | |
| UniSeg | 94.38±0.04 | 0.96±2.15 | 90.36±0.10 | 2.62±12.20 | 84.21±0.06 | 3.45±2.16 | 78.22±0.12 | 5.29±3.52 | 90.04±0.05 | 1.28±4.24 | 87.44±0.07 | 2.72±4.85 |
| Hermes | 96.11±0.02 | 0.58±1.78 | 94.35±0.07 | 1.43±8.13 | 85.29±0.05 | 2.13±1.27 | 77.62±0.12 | 3.27±2.64 | 90.21±0.05 | 1.05±1.33 | 88.71±0.06 | 1.69±3.03 |
| UniLSeg | 88.14±0.08 | 2.25±3.32 | 93.97±0.07 | **0.82±3.15** | 84.20±0.04 | **1.50±0.73** | 79.48±0.10 | **1.74±2.16** | 86.58±0.05 | 1.02±0.64 | 86.47±0.07 | 1.47±2.00 |
| Ours | **96.40±0.02** | **0.50±1.55** | **95.01±0.06** | 1.06±5.35 | **86.62±0.04** | 1.92±1.14 | **80.77±0.10** | 2.71±2.54 | **90.73±0.04** | **0.92±0.67** | **89.93±0.05** | **1.42±2.25** |

MS [21]. Notably, UniOCTSeg exhibits consistent improvements across varying tasks, highlighting its robustness across granularities.

To assess cross-domain and cross-granularity adaptability, we conduct evaluations on external test sets. While AMD [22] and NR206 [24] share annotation granularities with the training data, HEG [23] presents previously unseen granularity levels. Consequently, **both general and universal models cannot be directly applied and evaluated on HEG [23]**. We restructure HC-MS [21] by combining the $6^{th}$ and $7^{th}$ retinal layers to align with the granularity of tasks in HEG [23] for fair comparison. General models are trained on the reconstructed HC-MS [21], and universal models require prompt redesign by adding new task-specific prompts for tasks involving unseen granularities. In this way, fair performance assessments can be performed. External test results are listed in Table 2; our UniOCTSeg surpasses the second best model by 0.32%, 0.39% and 0.45% in DSC and 0.09, 0.02 and 0.05 in HD95 without retraining nor structure modification, highlighting its strong adaptability and superiority compared to other models on unseen datasets and granularities. To assess computational efficiency, we also report the number of parameters and floating point operations (FLOPs) in Table 2. Although proposed UniOCTSeg has a larger per-model footprint, its ability to handle multi-granularity tasks within a single framework

Table 2: Comparisons with general and universal models on the external test sets, as evaluated by DSC (%), HD95, Param (M) and Flops (G).

| Method | AMD | | HEG | | NR206 | | Avg. | | Param | Flops | Flexibility |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | DSC ↑ | HD95 ↓ | DSC↑ | HD95 ↓ | DSC↑ | HD95 ↓ | DSC↑ | HD95 ↓ | | | |
| general models | | | | | | | | | | | |
| UNet | 91.27±0.05 | 2.03±6.55 | 86.48±0.04 | 1.80±1.73 | 40.57±0.09 | 132.85±41.53 | 72.77±0.06 | 45.56±16.60 | 142.5 [23.75 * 6] | 591.00 [98.50 * 6] | **Low** |
| SwinUNetr | 91.33±0.04 | 1.52±2.17 | 86.52±0.04 | 1.73±1.15 | 69.96±0.09 | 21.20±35.40 | 82.60±0.05 | 8.15±12.91 | 176.40 [29.40 * 6] | 30.12 [5.02 * 6] | **Low** |
| nnUNet | 91.72±0.04 | 1.15±1.10 | 86.44±0.04 | 1.90±2.20 | 76.90±0.04 | 8.76±17.45 | 85.02±0.04 | 3.94±6.92 | 123.84 [20.64 * 6] | 234.06 [39.01 * 6] | **Low** |
| specially designed models for retinal layers segmentation | | | | | | | | | | | |
| YNet | 91.25±0.04 | 1.43±2.81 | 86.40±0.05 | 4.87±4.90 | 51.96±0.10 | 15.33±18.40 | 76.54±0.06 | 7.21±8.70 | 70.80 [11.80 * 6] | 90.66 [15.11 * 6] | **Low** |
| LightReSeg | 89.16±0.06 | 3.59±3.37 | 86.81±0.04 | 2.33±3.49 | 67.33±0.14 | 22.91±49.83 | 81.10±0.08 | 9.61±18.90 | 60.90 [10.15 * 6] | 26.34 [4.39 * 6] | **Low** |
| TCCT | 90.71±0.05 | 2.27±2.14 | 86.98±0.05 | 4.37±11.77 | 50.10±0.16 | 61.82±41.20 | 75.93±0.09 | 22.82±18.37 | 91.62 [15.27 * 6] | 50.70 [8.45 * 6] | **Low** |
| universal models | | | | | | | | | | | |
| UniSeg | 91.32±0.04 | 1.32±1.68 | 85.28±0.04 | 1.46±0.78 | 65.56±0.06 | 3.40±0.62 | 80.72±0.05 | 2.06±1.02 | 42.05 | 27.13 | **Median** |
| Hermes | 91.97±0.04 | 1.07±1.17 | 87.13±0.04 | 1.39±0.55 | 76.17±0.05 | 2.57±0.64 | 85.09±0.04 | 1.68±0.79 | 8.14 | 25.67 | **Median** |
| UniLSeg | 85.24±0.08 | 2.41±2.50 | 85.10±0.04 | **1.32±0.37** | 68.16±0.06 | 2.50±1.23 | 79.50±0.06 | 2.08±1.37 | 68.42 | 25.02 | **Median** |
| Ours | **92.29±0.04** | **0.98±1.32** | **87.52±0.04** | 1.34±0.60 | **77.35±0.05** | **2.45±0.94** | **85.72±0.04** | **1.59±0.95** | 92.53 | 53.10 | **High** |

**Low**, **Median**, and **High** represent a model's range of solvable granularities.

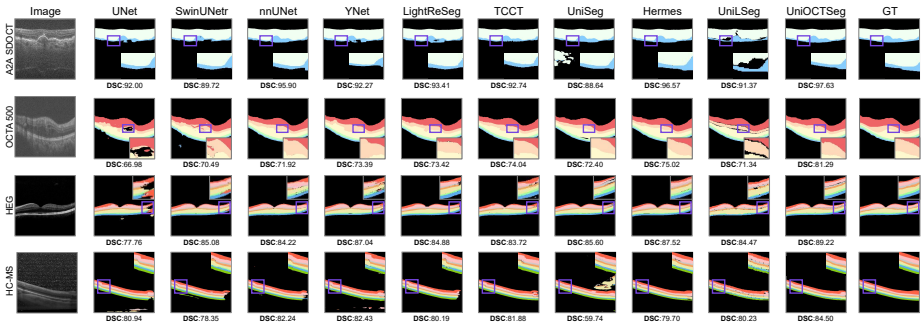


Fig. 3: Visualization results from UniOCTSeg and other SOTA methods.

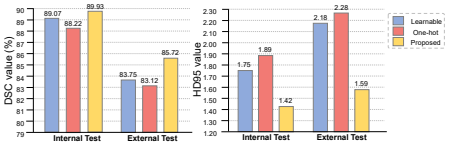


Fig. 4: Ablation results of prompt methods.

Table 3: Ablation results of PCL's key components.

| Consistency learning | Progressive training | internal test DSC ↑ | internal test HD95↓ | external test DSC↑ | external test HD95 ↓ | avg. DSC↑ | avg. HD95↓ |
|---|---|---|---|---|---|---|---|
| | | 89.15 ±0.06 | 1.93 ±4.18 | 83.53 ±0.05 | 1.85 ±0.87 | 86.34 ±0.06 | 1.89 ±2.53 |
| ✓ | | 88.97 ±0.06 | 1.77 ±3.57 | 80.99 ±0.05 | 5.76 ±4.70 | 84.98 ±0.06 | 3.77 ±4.14 |
| ✓ | ✓ | **89.93 ±0.05** | **1.42 ±2.25** | **85.72 ±0.04** | **1.59 ±0.95** | **87.83 ±0.05** | **1.51 ±1.6** |

leads to a more efficient overall solution compared to methods requiring retraining or separate models for each granularity level. The qualitative segmentation results of different methods are illustrated in Fig. 3 for visual comparison.

**Ablation Study.** To verify the effectiveness of our proposed innovations, we conduct ablation experiments to test HPS and PCL. To validate the hierarchical prompt design, we compare three different configurations under our proposed network architecture: 1) $\mathcal{G}(\sum_{i=1}^{9} L_i) = 45$ **task-specific learnable prompts** are employed, which is the total number of distinct retinal layer granularities; 2) utilizing predefined **fixed prompts** in our proposed strategy; 3) our proposed HPS. The comparative analysis of the prompt methods, as illustrated in Fig. 4, reveals segmentation performance enhancement with learnable prompts over fixed prompts. Notably, the proposed prompt strategy achieves the best results in both DSC and HD95, underscoring its effectiveness in capturing inter-layer dependencies in multi-granularity retinal layer segmentation.

For PCL, we evaluate two key components: consistency learning and progressive training. Three configurations are designed: 1) neither consistency learning nor progressive training is employed; 2) consistency learning is implemented exclusively by simultaneously computing the consistency loss across tasks of all granularities; 3) consistency learning is applied by computing the consistency loss with a progressively changing target (from coarse to fine). As demonstrated in Table 3, consistency learning leads to a decrease in segmentation performance across multi-granularity datasets, compared to the baseline without it. However, the full PCL setting achieves the highest gains, demonstrating that progressive

training stabilizes the training process and facilitates the adaptation of learnable prompts and the model to complex tasks and enhances segmentation accuracy.

## 4 Conclusion

This study proposes UniOCTSeg, a universal framework for retinal layer segmentation in OCT that addresses key limitations of existing methods, particularly their constrained flexibility and computational inefficiency. We propose three essential innovations: a novel universal network architecture, HPS, and PCL. We perform comprehensive evaluations of UniOCTSeg across eight publicly available OCT datasets on six distinct granularities and the results demonstrate its superior segmentation accuracy, high flexibility, and robust across-granularity adaptability. In future work, we will explore strategies for coordinating multiple data distributions and utilizing unlabeled data, while extending our model to other domains, such as hierarchical brain region segmentation [28], to meet broader clinical and research needs.

**Acknowledgments.** This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (T2422012, 62071210); the Guangdong Basic and Applied Basic Research (2024B1515020088); the Shenzhen Science and Technology Program (RCYX20210609103056042); the High Level of Special Funds (G030230001, G03034K003).

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

## References

1. Cai, Z., Lin, L., He, H., Tang, X.: Corolla: An efficient multi-modality fusion framework with supervised contrastive learning for glaucoma grading. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–4 (2022). https://doi.org/10.1109/ISBI52829.2022.9761712
2. Moradi, M., Chen, Y., Du, X., Seddon, J.M.: Deep ensemble learning for automated non-advanced amd classification using optimized retinal layer segmentation and sd-oct scans. Computers in Biology and Medicine **154**, 106512 (2023)
3. Raja, H., Akram, M.U., Hassan, T., Ramzan, A., Aziz, A., Raja, H.: Glaucoma detection using optical coherence tomography images: a systematic review of clinical and automated studies. IETE Journal of Research **69**(11), 7958–7978 (2023)
4. Elsharkawy, M., Sharafeldeen, A., Soliman, A., Khalifa, F., Ghazal, M., El-Daydamony, E., Atwan, A., Sandhu, H.S., El-Baz, A.: A novel computer-aided diagnostic system for early detection of diabetic retinopathy using 3d-oct higher-order spatial appearance model. Diagnostics **12**(2), 461 (2022)
5. Soundara Pandi, S.P., Winter, H., Smith, M.R., Harkin, K., Bojdo, J.: Preclinical retinal disease models: Applications in drug development and translational research. Pharmaceuticals **18**(3), 293 (2025)

6. Cai, Z., Lin, L., He, H., Cheng, P., Tang, X.: Uni4eye++: A general masked image modeling multi-modal pre-training framework for ophthalmic image classification and segmentation. IEEE Transactions on Medical Imaging **43**(12), 4419–4429 (2024). https://doi.org/10.1109/TMI.2024.3422102
7. Farshad, A., Yeganeh, Y., Gehlbach, P., Navab, N.: Y-net: A spatiospectral dual-encoder network for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 582–592. Springer (2022)
8. Tan, Y., Shen, W.D., Wu, M.Y., Liu, G.N., Zhao, S.X., Chen, Y., Yang, K.F., Li, Y.J.: Retinal layer segmentation in oct images with boundary regression and feature polarization. IEEE Transactions on Medical Imaging **43**(2), 686–700 (2023)
9. He, X., Song, W., Wang, Y., Poiesi, F., Yi, J., Desai, M., Xu, Q., Yang, K., Wan, Y.: Light-weight retinal layer segmentation with global reasoning. IEEE transactions on instrumentation and measurement (2024)
10. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1195–1204 (2021)
11. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21152–21164 (2023)
12. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y.: Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 508–518. Springer (2023)
13. Liu, Y., Zhang, C., Wang, Y., Wang, J., Yang, Y., Tang, Y.: Universal segmentation at arbitrary granularity with language instruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3459–3469 (2024)
14. Gao, Y.: Training like a medical resident: Context-prior learning toward universal medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11194–11204 (2024)
15. Lin, L., Liu, Y., Wu, J., Cheng, P., Cai, Z., Wong, K.K., Tang, X.: Fedlppa: learning personalized prompt and aggregation for federated weakly-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2024)
16. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
17. Farsiu, S., Chiu, S.J., O'Connell, R.V., Folgar, F.A., Yuan, E., Izatt, J.A., Toth, C.A., Group, A.R.E.D.S..A.S.D.O.C.T.S., et al.: Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. Ophthalmology **121**(1), 162–172 (2014)
18. Li, M., Huang, K., Xu, Q., Yang, J., Zhang, Y., Ji, Z., Xie, K., Yuan, S., Liu, Q., Chen, Q.: Octa-500: a retinal dataset for optical coherence tomography angiography study. Medical image analysis **93**, 103092 (2024)
19. Chiu, S.J., Allingham, M.J., Mettu, P.S., Cousins, S.W., Izatt, J.A., Farsiu, S.: Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomedical optics express **6**(4), 1172–1194 (2015)

20. Li, J., Jin, P., Zhu, J., Zou, H., Xu, X., Tang, M., Zhou, M., Gan, Y., He, J., Ling, Y., et al.: Multi-scale gcn-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary oct images. Biomedical Optics Express **12**(4), 2204–2220 (2021)
21. He, Y., Carass, A., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L.: Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls. Data in brief **22**, 601 (2018)
22. Chiu, S.J., Izatt, J.A., O'Connell, R.V., Winter, K.P., Toth, C.A., Farsiu, S.: Validated automatic segmentation of amd pathology including drusen and geographic atrophy in sd-oct images. Investigative ophthalmology & visual science **53**(1), 53–61 (2012)
23. Tian, J., Varga, B., Somfai, G.M., Lee, W.H., Smiddy, W.E., Cabrera DeBuc, D.: Real-time automatic segmentation of optical coherence tomography volume data of the macular region. PloS one **10**(8), e0133908 (2015)
24. He, X., Wang, Y., Poiesi, F., Song, W., Xu, Q., Feng, Z., Wan, Y.: Exploiting multi-granularity visual features for retinal layer segmentation in human eyes. Frontiers in Bioengineering and Biotechnology **11**, 1191803 (2023)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
26. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
27. 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)
28. Lyu, J., Xu, P., Nasrallah, F., Tang, X.: Learning ontology-based hierarchical structural relationship for whole brain segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 385–394. Springer (2023)