CAMEL: Confidence-Aware Multi-task Ensemble Learning with Spatial Information for Retina OCT Image Classification and Segmentation

Juho Jung[1*] Migyeong Yang[1*] Hyunseon Won[1] Jiwon Kim[1]
Jeong Mo Han[2] Joon Seo Hwang[3] Daniel Duck-Jin Hwang[4,5†] Jinyoung Han[1†]
[1]Sungkyunkwan University, Seoul, South Korea
[2]Seoul Bombit Clinic, Seoul, South Korea [3]Seoul Plus Eye Clinic, Seoul, South Korea
[4]Hangil Eye Hospital, Incheon, South Korea [5]Lux Mind, Incheon, South Korea
{jhjeon9, mgyang, dprth1014, jjeon416}@g.skku.edu,
{joehan712, poppn78, hallelu7}@gmail.com, jinyounghan@skku.edu

Abstract

Precise retina Optical Coherence Tomography (OCT) image classification and segmentation are important for diagnosing various retinal diseases and identifying specific regions. Alongside comprehensive lesion identification, reducing the predictive uncertainty of models is crucial for improving reliability in clinical retinal practice. However, existing methods have primarily focused on a limited set of regions identified in OCT images and have often faced challenges due to aleatoric and epistemic uncertainty. To address these issues, we propose **CAMEL** *(***C**onfidence-**A**ware **M**ulti-task **E**nsemble **L**earning*), a novel framework designed to reduce task-specific uncertainty in multi-task learning.* CAMEL *achieves this by estimating model confidence at both pixel and image levels and leveraging confidence-aware ensemble learning to minimize the uncertainty inherent in single-model predictions.* CAMEL *demonstrates state-of-the-art performance on a comprehensive retinal OCT image dataset containing annotations for nine distinct retinal regions and nine retinal diseases. Furthermore, extensive experiments highlight the clinical utility of* CAMEL*, especially in scenarios with minimal regions, significant class imbalances, and diverse regions and diseases. Our code is publicly available at:* https://github.com/DSAIL-SKKU/CAMEL.

1. Introduction

Optical Coherence Tomography (OCT) is a non-invasive imaging method, which is renowned for its high-resolution and three-dimensional capabilities in ophthalmology applications [32, 48]. As OCT shows notable features, including high speed, real-time imaging, and rich information [32], OCT has been widely applied in monitoring the progression of retinal diseases like diabetic macular edema (DME) or age-related macular degeneration (AMD) [24]. With these advantages, the development of OCT-based automated diagnostic systems for classification [11, 20–22, 26] and segmentation tasks [23, 32, 52] has received significant attention. As image-level segmentation can be viewed as pixel-level classification [46], multi-task learning has emerged as a viable solution for handling both classification and segmentation tasks simultaneously [38, 42]. This approach has been validated as useful in improving performance by utilizing shared task-specific knowledge and providing clinicians with valuable information on disease identification and region localization [37, 46]. In this way, collaborative efforts on integrating these two objectives within a unified framework have become a matter of importance [16].

Research on multi-task learning for OCT image classification and segmentation can be broadly categorized into two approaches [46, 60]: (i) parameter sharing between models and (ii) applying a cascaded architecture. Specifically, the parameter sharing approach improves the performances of both tasks by finding the optimal convergence point of each task through supervised learning [12, 33]. In contrast, the cascaded architecture approach utilizes segmentation or classification results to guide the other task [10, 42]. Although these approaches have led to improved performance for applying to actual clinical practice, they can be limited in that they tend to focus a small number of regions out of an entire OCT image [32, 34, 51]; e.g., only one or two regions, such as epiretinal membrane (ERM) and intraretinal fluid (IRF) were covered in prior works [16, 23]. Note that since the definition of the disease is based on entire regions with different sizes and positions, a comprehensive identification across various regions is essential for an accurate diagnosis [50].

*Equal contribution.
†Corresponding authors.

While multi-task learning shows improved performance, enhancing reliability and reducing uncertainty are necessary for clinical practice as model errors can directly impact patient health [5, 36]. Since evaluating confidence levels and quantifying predictive uncertainty can enhance model reliability and performance in the retinal field [29], various efforts have been made to adapt uncertainty estimation in retinal imaging [46, 53, 54]. However, this approach relies on a single model, which can suffer from the inherent uncertainty of a single model [25, 31].

To address these challenges, we introduce ***CAMEL*** (***C****onfidence-****A****ware* ***M****ulti-task* ***E****nsemble* ***L****earning*) to enhance reliability and performance in multi-task learning for comprehensive retina OCT image classification and segmentation. *CAMEL* employs task-specific uncertainty estimation and confidence-aware ensemble learning, assessing uncertainties at both image and pixel levels, and addressing epistemic uncertainty inherent in single models. Specifically, weighting based on the reliability score of each model in each task, *CAMEL* dynamically incorporates confidence levels from multiple models to achieve optimal results. Furthermore, we integrate adaptable individual branches for particular diseases that require specialized analysis, thereby improving the model's versatility. OCT images are preprocessed to adapt their spatial information, ensuring precise pixel matching at lesion boundaries. Unlike other images, medical images can be characterized by the definition of tissue, lesions, and organ shapes and positions [49]. Based on the well-structured spatial information, we model the medical and geometric relationships among various lesions, incorporating post-processing adjustments to generate refined resized mask images.

To evaluate our proposed methods, we use a new retina OCT dataset, suitable for multi-task learning, containing original OCT images and masks for nine retinal regions and nine retinal diseases, which three retina specialists manually annotate. As publicly available datasets typically contain fewer than four types of regions or diseases [7, 17, 57], utilizing a diverse range of datasets can provide valuable insight into assessing both segmentation and classification tasks. We also validate *CAMEL* with the publicly available dataset for showcasing its generalizability. Extensive experimental evaluation demonstrates the effectiveness and potential clinical use of our proposed methods.

2. Related Work

2.1. Multi-task learning

Many researchers have explored techniques to enhance retinal image classification [11, 20–22, 26, 59] and segmentation [23, 32, 52]. Taking a step further, multi-task learning, also known as joint learning, has been proposed as an effective solution for simultaneously addressing OCT classification and segmentation tasks [38, 42]. Research in multi-task learning for OCT images can be broadly categorized into two approaches [46, 60]: (i) parameter sharing between models (parallel architecture) and (ii) cascaded architecture.

Diao et al. [12] introduced cascaded multi-task learning using dual guidance networks, employing guiding masks for classifying Drusen and CNV and class activation maps for segmenting these lesions. Gende et al. [16] used output segmentation maps as inputs for classifying the Epiretinal Membrane (ERM). In the parallel architecture, Asgari et al. [4] used a shared encoder with a multi-decoder architecture for multi-task drusen segmentation, while Cao et al. [8] integrated distance maps of retinal layer surfaces and employed task-specific attention modules to fuse segmentation and classification features. However, these studies tend to focus on only a few lesions and diseases and lack uncertainty estimation, limiting model reliability. In this study, we propose a model that segments nine possible lesions and simultaneously classifies nine diseases commonly found in OCT images. Additionally, we emphasize the significance of conducting research based on OCT, noting the prevalence of fundus images in retina imaging.

2.2. Ensemble learning

Ensemble learning aims to achieve better predictions by combining multiple models [13, 41, 43]. This strategy has been successfully applied to OCT image classification and segmentation, enhancing performance and robustness. For instance, some researchers have shown that ensemble learning improves retinal disease diagnosis, such as DME [27, 40, 45]. Moradi et al. [40] used majority voting across multiple classifiers, which take OCT scans and fundus images as inputs, emphasizing the importance of ensemble learning in OCT image classification. Additionally, ensemble learning has also been shown to enhance OCT segmentation performance [2, 9, 44]. Cazañas-Gordón et al. [9] compared three ensemble prediction schemes—majority voting, weighted averaging, and stacking—that aggregate the results of multiple independent classifiers. They highlighted the need for simple and effective aggregation methods in OCT segmentation, where modeling the layer structure is crucial. While some studies have utilized ensemble learning in multi-task pipelines [28, 30], they primarily rely on simple soft voting of multiple CNN-based models without fully integrating ensemble learning within a multi-task framework. To address this, we propose *CAMEL*, which calculates each model's reliability score based on confidence and enhances classification, segmentation, and specific disease detection through confidence-aware ensemble predictions.

2.3. Calibration in learning

Enhancing reliability and reducing uncertainty in OCT image classification and segmentation models is crucial for

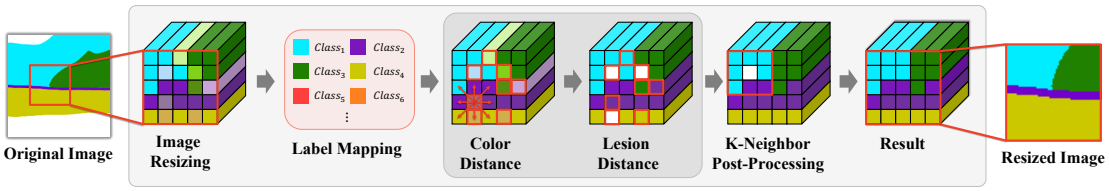

Figure 1. Overall architecture of our data pre-processing method.

clinical application, as model errors can directly impact patient health [5, 36]. Evaluating confidence levels and quantifying predictive uncertainty can improve model reliability and performance in retinal imaging [29]. Various efforts have been made to incorporate uncertainty estimation in retinal imaging. For instance, Seeböck *et al.* [54] proposed Bayesian U-Net with test-time *Monte Carlo Dropout* (MC Dropout) [15] for capturing epistemic uncertainty in OCT segmentation. Sedai *et al.* [53] introduced uncertainty-guided semi-supervised learning using a student-teacher approach, where the student model is updated based on the teacher model's uncertainty. Recently, Ren *et al.* [46] explored multi-task learning for image classification and segmentation by utilizing pixel-wise and image-level uncertainty. However, relying on a single model can lead to inherent uncertainty issues [25, 31]. Thus, we propose using model confidence to calculate a reliability score, which serves as the weight in the ensemble aggregation process. This approach aims to improve performance and reduce prediction uncertainty in both classification and segmentation. To the best of our knowledge, no previous work has applied confidence-aware ensemble learning to retina OCT classification and segmentation.

3. OCT Image Dataset

3.1. Data Annotation

To evaluate our method, we used spectral-domain optical coherence tomography (SD-OCT) (Heidelberg Spectralis, Heidelberg Engineering, Heidelberg, Germany) images from patients treated at Hangil Eye Hospital between July 2014 and June 2021. The study focused on eyes with conditions such as CRVO, CSC, DM, ERM, MH, wetAMD, RAP, PCV, and Normal. Each condition included 100 patients, except CRVO, which had 110 due to image quality. For each patient, 25 OCT images were collected, including those taken 5 scans before and after the primary lesion area, resulting in 2,730 images from 910 patients.

Three retinal specialists manually annotated each OCT scan. They first reviewed a sample of images to establish annotation guidelines, refining them iteratively to ensure consistency. Once consensus was reached, each specialist annotated the images using Autodesk SketchBook on an iPad Pro. They outlined eight lesion structures with an Apple Pencil, marking uncolored areas as background. The annotated structures included EpiRetinal Membrane (ERM), Retina, IntraRetinal Fluid (IRF), SubRetinal Fluid (SRF), Subretinal HyperReflective Material (SHRM), Retinal Pigment Epithelium (RPE), retinal Pigment Epithelial Detachment (PED), and Choroid. As a result, the final OCT image segmentation masks were nine-color images with each specified lesion color as illustrated in Table 1.

Table 1. Annotated regions with their corresponding colors and RGB values in our dataset.

Region	R	G	B	color
ERM (EpiRetinal Membrane)	255	89	0	
Retina	0	236	255	
IRF (IntraRetinal Fluid)	255	163	162	
SRF (SubRetinal Fluid)	32	128	0	
SHRM (Subretinal HyperReflective Material)	204	102	0	
RPE (Retinal Pigment Epithelium)	112	0	204	
PED (retinal Pigment Epithelial Detachment)	255	62	62	
Choroid	204	200	0	
Background	255	255	255	

Subsequently, one of the three retinal specialists reviewed all the initially annotated images to ensure that the annotations adhered to a consistent standard. Any necessary adjustments were made by the specialist at this stage to meet the established criteria, and these adjustments were then cross-checked by the other two retinal specialists. The final dataset consists of 2,730 OCT images, each with a resolution of 768 $\times$ 496, along with their corresponding ground truth masks. Our study complied with the guidelines of the Declaration of Helsinki, and the research plan was approved by the Ethics Committee of Hangil Eye Hospital (IRB-21018). Due to the retrospective observational nature of the study, the committee waived the requirement for informed consent.

3.2. Data Pre-processing

As illustrated in Figure 1, we resize the image and post-process the blurry boundaries between regions using the following three steps: (1) *Pixel Label Mapping*, (2) *Pixel Remapping*, and (3) *K-Neighbor Post-Processing*. In *Pixel Label Mapping*, inaccurate pixels are identified by comparing the pixels of the resized mask to those of the original mask. Next, during *Pixel Remapping*, pixels are remapped based on their color distance and lesion distance from each other, considering the medical relationships between lesions and their class colors. Finally, in *K-Neighbor Post-Processing*, adjustments are made by referencing neighboring pixels. For more details, see the supplementary material (Section A.1).

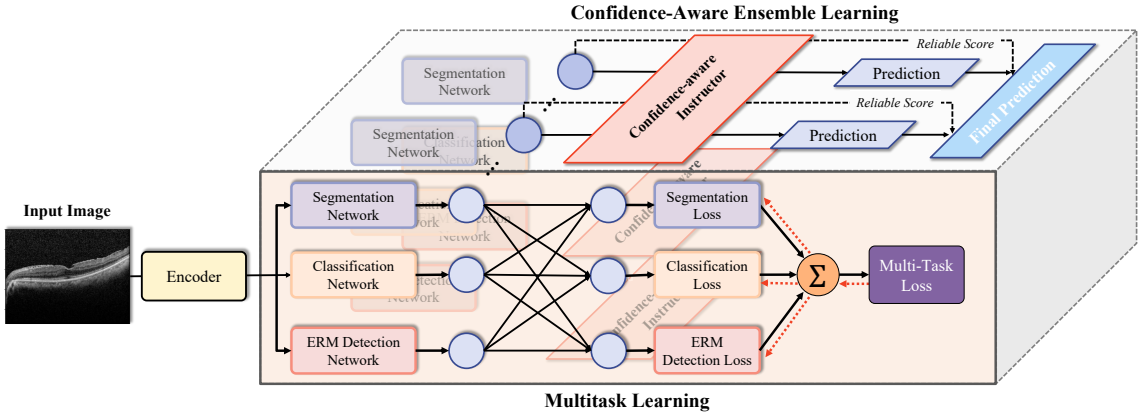

Figure 2. Illustration of confidence-aware ensemble learning in the proposed *CAMEL* method. Predictions (classification, segmentation, detection) and predictive confidences are obtained from multiple models. The *Confidence-aware Instructor* calculates the *reliability score* based on each model's confidence for each task, determining the weight of each model's prediction during ensemble learning.

4. CAMEL: Confidence-Aware Multi-task Ensemble Learning

Figure 2 illustrates the design of our proposed method, *CAMEL*. *CAMEL* employs a multi-task learning framework to perform segmentation, classification, and specific disease detection simultaneously. Additionally, it enhances each task with confidence-aware ensemble predictions.

4.1. Multi-task Learning: Classification, Segmentation, and Detection

Since semantic segmentation can be considered a pixel-level classification, we consider the multiple disease classification, legion segmentation, and specific disease detection problems as N-class classification problems. We measure confidence using the probability of the model's predictions at both the image and pixel levels [18]. Following the loss setting introduced in [36], both Cross-Entropy loss and Dice loss are used for training the network. Specifically, Cross-Entropy loss (L_{CE}) is calculated as follows:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p(\hat{y}_i = c|x_i,\theta)), \quad (1)$$

where $p(\hat{y}_i = c|x_i,\theta)$ represents the probability that image or pixel i belongs to class c, and $\hat{y}_i$ is the predicted class based on the input x_i and model parameters θ. The binary indicator $y_{i,c}$ denotes whether class c is the correct label for the ith pixel or image. Since the goal is to segment nine regions and classify nine diseases, the total number of classes, C, is set to nine, making this a multi-class classification problem. Moreover, Dice loss (L_{DS}) is also employed to guide the model's parameter selection towards maximizing the similarity with the ground truth segmentation mask:

$$L_{DS} = 1 - \sum_{c=1}^{C}\frac{2\sum_{i=1}^{N} y_{i,c}\times p_{i,c}}{\sum_{i=1}^{N} y_{i,c} + \sum_{i=1}^{N} p_{i,c}}, \quad (2)$$

where C is the total number of classes, and N represents the number of pixels in each mini-batch. $y_{i,c}$ is the binary indicator for whether pixel i belongs to class c, and $p_{i,c}$ is the predicted probability for pixel i belonging to class c.

Notably, several studies have highlighted the negative impact of Dice loss on calibration quality [6, 19], while cross-entropy loss has been shown to provide better-calibrated predictions and uncertainty estimations [18]. These findings emphasize the importance of combining both losses [36]. As such, we defined the following loss functions: Categorical Cross-Entropy for classification, a balanced combination of Cross-Entropy and Dice loss for segmentation, and binary Cross-Entropy for ERM detection. It is worth noting that we treated ERM detection as a separate branch because ERM is crucial both as a lesion and a disease [14]. Additionally, OCT has become the most useful single auxiliary test for diagnosing ERM, showing higher sensitivity compared to clinical examinations alone [55]. This detection branch can be adapted for more significant or harder-to-detect diseases depending on the clinical context.

The total loss, denoted as L_{MTL}, is a combination of three individual task losses, each weighted by learnable parameters as follows:

$$L_{MTL} = \alpha L_{Cls} + \beta L_{Seg} + \gamma L_{ERM} \quad (3)$$

Here, L_{Cls} represents the classification task loss for predicting one of nine diseases, L_{Seg} represents the segmentation task loss for classifying each pixel of medical images, and L_{ERM} denotes the ERM detection task loss.

Each task loss is associated with pixel-level and image-level uncertainty estimation based on the model's confidence. The task reflection ratio, α and β, are learnable parameters during the training process, while the reflection ratio γ is a pre-specified constant. These ratios control the

Table 2. Comparison of segmentation IoU between baseline models and *CAMEL*.

	ERM	Retina	IRF	SRF	SHRM	RPE	PED	Choroid	B.G.	Mean IoU
UML [46]	0.3664	0.9438	0.2143	0.7212	0.2483	0.5876	0.3778	0.8546	**0.9784**	0.5882
TCCT-BP [56]	0.3294	0.9306	0.1969	0.6919	0.1587	0.5580	0.4440	0.7963	0.9567	0.5625
Attention-based U-Net [39]	0.3303	0.9358	0.4987	0.7795	0.3968	0.5598	0.6698	0.8174	0.9734	0.6624
CAMEL	**0.8186**	**0.9439**	**0.8285**	**0.9218**	**0.9544**	**0.8558**	**0.9275**	**0.8801**	0.9738	**0.9005**

Table 3. Comparison of classification accuracy between baseline models and *CAMEL*.

	CRVO	CSC	DME	ERM	MH	Normal	PCV	RAP	wetAMD	Total Cls.
UML [46]	0.9400	0.9459	0.9167	0.9315	**0.8923**	**1.0000**	0.8906	0.9437	**0.7937**	0.9175
MedViT-S [35]	0.9400	0.9189	0.9028	0.9041	**0.8923**	0.9516	0.9219	0.8873	**0.7937**	0.9007
VGG-19-based model [20]	0.8400	0.9730	0.9028	0.9041	0.8769	**1.0000**	0.7031	0.9296	0.6984	0.8737
CAMEL	**1.0000**	**1.0000**	**0.9700**	**1.0000**	0.8900	0.9800	**1.0000**	**0.9900**	0.2300	**0.9288**

contribution of each task to the overall loss, allowing the model to optimize the performance of individual tasks. The training process aims to find optimal values for α, β, and γ to enhance *CAMEL*'s learning and predictive capabilities across all tasks.

4.2. Confidence-Aware Ensemble Learning

To reduce uncertainty within a single model, we implement confidence-aware ensemble learning, as illustrated in Figure 2. We first employ n distinct Multi-Task Learning (MTL) models, each utilizing different backbone encoders and parameter configurations. We then assess each model's confidence in classification, segmentation, and detection independently using Expected Calibration Error (ECE), defined as:

$$ECE = \frac{1}{N}\sum_{m=1}^{M} \left| acc(b_m) - conf(b_m) \right|, \tag{4}$$

where N, M, and b_m represent the total dataset size, the number of bins, and the collection of predictions belonging to the mth bin, respectively. Also, $acc(b_m)$ and $conf(b_m)$ denote the accuracy and confidence of the mth bin, respectively. Since the confidence of each task can be linked to the uncertainty of the models at the image and pixel levels [36], we used weighted voting in the ensemble to make the final prediction for each task.

To elaborate, as depicted in Figure 2, the *Confidence-aware Instructor* evaluates a *reliability score* of each task branch based on the ECE of each model, and assigns weights to their predictions as follows:

$$reliability\ score_m = 1 - \frac{ECE_m}{\sum_{i=1}^{n} ECE_i} \tag{5}$$

This *reliability score* is then used to weight the predictions of each task (i.e., three tasks) branch for each model (i.e., five models) during ensemble learning. This ensures that more reliable models have a greater influence on the final prediction as follows:

$$\hat{y} = \arg\max_{c} \left(\frac{\sum_{i=1}^{N} w_i \times p_{i,c}}{\sum_{i=1}^{N} w_i} \right) \tag{6}$$

where $p_{i,c}$ is the prediction of i-th model for class c, w_i is the weight (in this case, the $reliability\ score$) for i-th model, and $\hat{y}$ is the final ensemble prediction, calculated as the weighted sum of individual model predictions.

5. Experimental Settings

5.1. Dataset

To evaluate our method, we utilized our dataset as explained in Section 3.1. We divided the dataset into training, validation, and test sets in a 7:1:2 ratio. Additionally, we evaluated the proposed model and baselines using the publicly available OCT5k dataset [3]. The OCT5k dataset includes manual labels for three disease classes (AMD, DME, and Normal for classification) and six-layer classes (ILM, OPL-Henle, IS/OS junction, IBRPE, OBRPE, and background for segmentation) based on retinal OCT images. We conducted experiments using all 1,672 images, splitting them into training, validation, and test sets in a 5:1:4 ratio, as provided by the dataset.

5.2. Baselines

To evaluate the performance of *CAMEL* in comparison to baseline models, we consider three types of baseline models in retinal imaging: (i) multi-task learning, (ii) segmentation, and (iii) classification. Due to the limited availability of multi-task models in retinal imaging, we only include UML [46] as the multi-task learning baseline. Additionally, we compare multi-class OCT image segmentation models: (i) TCCT-BP [56] and (ii) Attention-based U-Net [39]. For classification, we compare the following recent OCT image classification models: (i) MedViT-S [35] and (ii) VGG-19-based model [20]. Note that our pre-processing method was implemented prior to applying the specified methods for each baseline model (See Section B.1 for more details).

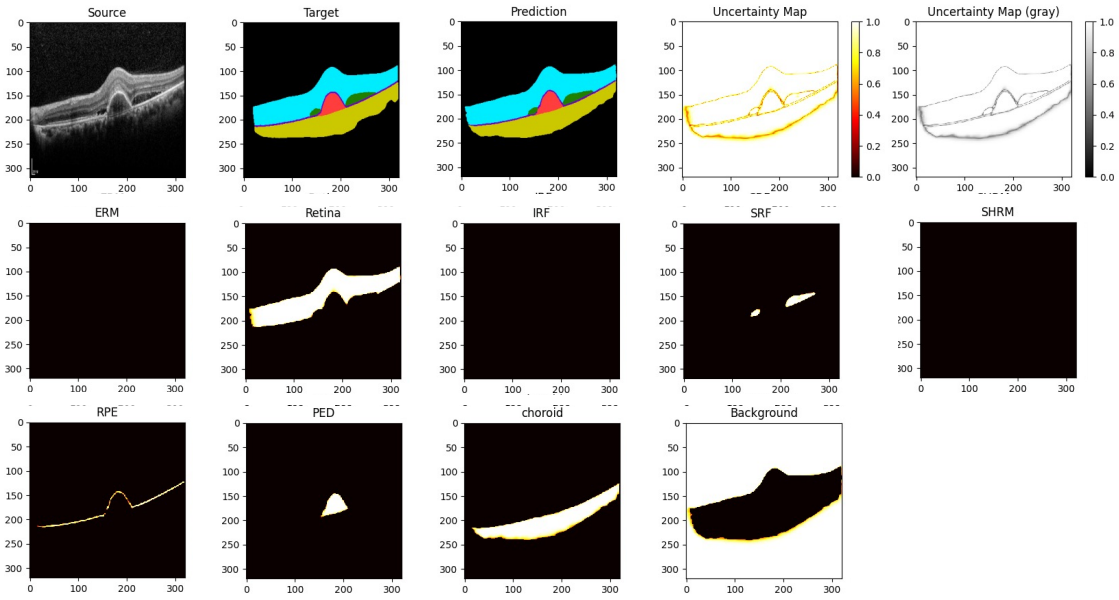

Figure 3. Visualization of region-specific uncertainty maps generated by *CAMEL*'s uncertainty estimation.

Table 4. Comparison of classification accuracy, segmentation IoU, and speed (sec.) between baselines and *CAMEL* on the OCT5k dataset.

	AMD	DME	Normal	Total Cls.	ILM	OPL-Henles	IS/OS junction	IBRPE	OBRPE	B.G.	Mean IoU	Computation Time	
												Training	Inference
MedViT-S [35]	0.6214	0.9868	0.6494	0.6636	-	-	-	-	-	-	-	1567.84	4.11
VGG-19-based model [20]	0.6004	**1**	0.6897	0.6598	-	-	-	-	-	-	-	**388.95**	**1.94**
TCCT-BP [56]	-	-	-	-	0.93	0.9153	0.7979	0.7719	0.9932	0.9657	0.8957	1637.10	7.96
Attention-based U-Net [39]	-	-	-	-	0.9165	0.8982	0.7819	0.7592	0.9928	0.987	0.8893	2196.07	15.20
UML [46]	0.3461	0.8421	0.6667	0.4670	0.8446	0.7944	0.6298	0.5966	0.975	0.9868	0.8045	9557.49	49.14
CAMEL-Ens	0.5679	1	0.7471	0.6507	0.8998	0.8728	0.7214	0.7002	0.9923	0.9881	0.8624	1239.96	8.25
CAMEL	**0.6998**	**1**	**0.7241**	**0.7348**	**0.9482**	**0.9375**	**0.8044**	**0.7943**	**0.9980**	**0.9876**	**0.9117**	1983.23	14.62

5.3. Implementation Details

CAMEL was implemented using Tensorflow [1] with U-Net [47] as a backbone model. We employed five independently trained models for confidence-aware ensemble learning. ResNet101, EfficientNetB0, and EfficientNetB7 were selected as the backbone encoder of the U-Net model based on their superior performance in empirical experiments. ResNet101 was employed with reflection ratios of 0.1, 0.2, and 0.5, while both EfficientNetB0 and EfficientNetB7 used 0.2 for the ERM detection task. The reflection ratio for the segmentation and classification tasks was learned during training, as detailed in Section 4.1.

6. Results

6.1. Overall Performance

Table 2 shows the segmentation performances (IoU; Intersection over Union) of the baseline models and *CAMEL*. *CAMEL* significantly outperformed all baselines across all classes in the segmentation task. Notably, our model achieved substantially higher performance on lesions requiring precise and detailed segmentation, such as IRF and SHRM. Among the baseline models, the Attention-based U-Net [39] demonstrated the best performance in segmenting small lesions such as IRF, SRF, SHRM, and PED. It seems that the soft attention mechanism of this model helped complement the insufficient information in small lesions by assigning weights. For other regions, the UML [46] showed the highest segmentation performance. This highlights the importance of using a joint learning framework and leveraging both image-level and pixel-wise confidence scores in medical image segmentation. By combining these strengths with ensemble learning, *CAMEL* achieved the highest performance. Furthermore, by dedicating a separate branch to ERM, a critical lesion and disease, the model demonstrated even greater performance improvements.

The classification performances of the baseline models and *CAMEL* for the predicted class labels are shown in Table 3. *CAMEL* demonstrates the highest overall accuracy in retinal disease classification compared to other baseline models. Although our method shows lower accuracy in the wetAMD class, this is attributed to the fact that PCV, RAP,

Table 5. Analysis of the *CAMEL* components. B.G., Seg., Cls., and Det. represent Background, Segmentation, Classification, and Detection, respectively. *MTL*, *CA*, and *Ens* refer to Multi-task Learning, Confidence-Aware, and Ensemble Learning. In *MTL*, classification, segmentation, and detection tasks are integrated. The table presents the Dice score for each configuration.

Tasks	ERM	Retina	IRF	SRF	SHRM	RPE	PED	Choroid	B.G.	Total Seg.	Total Cls.	Total ERM Det.
Area Ratio	**0.02%**	**13.63%**	**0.39%**	**0.31%**	**0.09%**	**0.42%**	**0.43%**	**9.07%**	**75.26%**	-	-	-
Cls.	-	-	-	-	-	-	-	-	-	-	0.6833	-
ERM.	-	-	-	-	-	-	-	-	-	-	-	0.7533
Seg.	0.7926	0.9616	0.6687	0.9238	0.8917	0.6110	0.8994	0.9007	0.9879	0.9700	-	-
Cls.+ERM.	-	-	-	-	-	-	-	-	-	-	0.7667	0.8233
Cls.+Seg.	0.7751	0.9525	0.3114	0.9133	0.7229	0.5921	0.7793	0.8636	0.9821	0.9600	0.8717	-
ERM.+Seg.	0.7619	0.9472	0.3011	0.9039	0.7037	0.5983	0.7739	0.8603	0.9811	0.9569	-	0.9217
MTL	**0.8192**	**0.9657**	**0.8553**	**0.9272**	**0.9236**	**0.6545**	**0.9221**	**0.9164**	**0.9895**	**0.9738**	**0.9010**	**0.9272**
MTL+CA	0.8295	0.9707	0.8630	0.9373	0.9495	0.7186	0.9339	0.9257	0.9899	0.9866	0.9117	0.9288
MTL+Ens	0.8425	0.9732	**0.8713**	0.9431	0.9610	0.8261	0.9375	0.9402	0.9910	0.9872	0.9214	0.9291
MTL+CA+Ens (*CAMEL*)	**0.8789**	**0.9768**	0.8610	**0.9503**	**0.9721**	**0.9336**	**0.9593**	**0.9731**	**0.9960**	**0.9880**	**0.9288**	**0.9301**

and wetAMD all fall under the broader category of nAMD (neovascular Age-related Macular Degeneration). Nevertheless, *CAMEL* achieved higher or comparable scores in the other classes.

6.2. Visual Results of CAMEL

Figure 3 illustrates the segmentation results produced by *CAMEL*. The predictions closely match the target regions, and the model effectively verifies reliability at region boundaries without overconfidence. As shown in the "*Uncertainty Map*," despite designing CAMEL to reduce uncertainty through its confidence-aware ensemble learning, some uncertainty remains in specific locations, particularly at lesion sites and boundary areas. It seems that, as Mehrtash et al. [36] strongly emphasize, higher levels of uncertainty at boundaries are still evident. Nevertheless, medical professionals can evaluate comprehensive results and examine uncertainty maps to identify areas of low confidence, thereby enhancing the decision-making process in clinical practice.

6.3. Validating CAMEL's Generalizability

To evaluate the generalizability of *CAMEL* on other OCT datasets, we used the publicly available OCT5k [3] dataset and compared the model's performance with baseline models. As shown in Table 4, *CAMEL* outperformed all baseline models in both classification and segmentation tasks. Notably, while our dataset focuses on labeling small lesion areas, the OCT5k dataset is annotated based on the layers seen in OCT images. Despite this difference, *CAMEL* demonstrated strong performance on the OCT5k dataset, highlighting its ability to generalize across different types of annotations.

6.4. Computational Complexity

While model performance is crucial, computational efficiency is also important for real-world clinical applications. We compared the training and inference times of *CAMEL* with baseline models using the OCT5k [3] dataset. As shown in Table 4, *CAMEL* took 1983.23 seconds for training and 14.62 seconds for inference, which is slightly higher than classification-only baselines but comparable to segmentation-only models and faster than multi-task models. Removing ensemble learning could reduce time complexity, though it would lower performance, highlighting the trade-off between efficiency and accuracy in clinical settings.

6.5. Single-task vs. Multi-task

Table 5 presents a comparison between single-task and multi-task approaches for *CAMEL*, which addresses three primary tasks: Segmentation, Classification, and ERM detection. Experiments were conducted for each task individually, in pairwise combinations, and jointly for all three tasks (see Tasks in Table 5). As shown, multi-task learning significantly improves performance, especially for classification and ERM detection accuracy. While segmentation performance (Dice score) slightly decreased when combined with classification or ERM detection alone, it notably improved when all tasks were integrated, particularly for small and sparsely distributed regions like IRF. This highlights the advantage of multi-task learning in *CAMEL*, demonstrating that disease-specific branches (e.g., ERM) enhance the detection of small regions alongside classification and segmentation.

6.6. Ablation Study

Table 5 presents the results of the ablation study for each component of *CAMEL*. Multi-task learning (*MTL*) alone performed worse than when combined with Confidence-Aware (*CA*) or Ensemble methods (*Ens*). The best performance was achieved when all methods were used together. The first and second rows in Figure 4 compare the results with and without ensemble learning (*MTL* vs. *MTL+ Ens*),

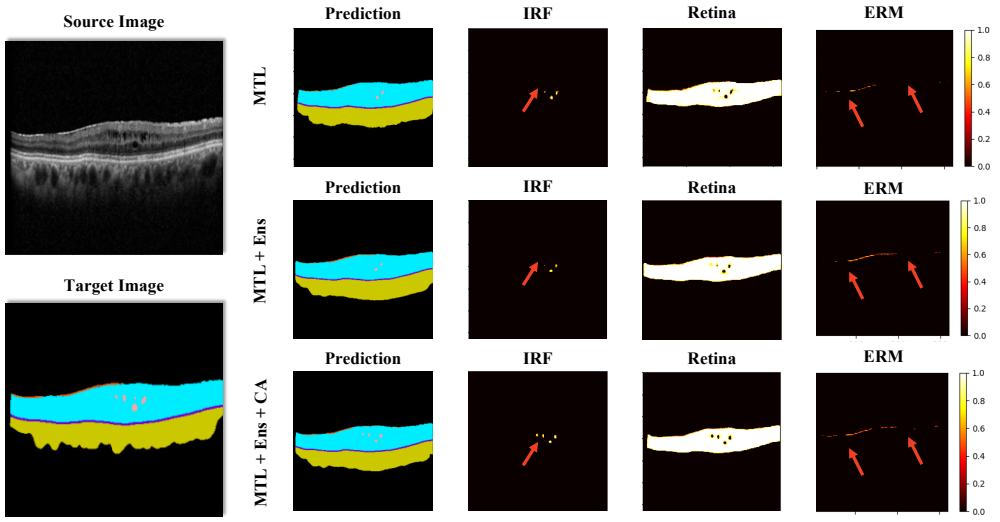

Figure 4. The impact of each component in *CAMEL*."MTL" denotes Multi-task learning, while "$MTL + Ens$" denotes Multi-task learning integrated with Ensemble learning. "$MTL + Ens + CA$" refers to Confidence-aware Multi-task Ensemble learning, which incorporates *CAMEL*. In the confidence map, values range from 0 to 1 (100% confidence), representing the model's prediction confidence for each pixel.

Table 6. Performance comparison between our data pre-processing method and other resizing techniques on OCT segmentation. B.G. denotes Background, and Avg. denotes Average. The value of the table presents the Dice score.

Method	ERM	Retina	IRF	SRF	SHRM	RPE	PED	Choroid	B.G.	Avg.
Bicubic	0.6432	0.9527	0.5092	0.7952	0.5913	0.5637	0.5539	0.8728	0.9659	0.9408
Linear	0.6611	0.9537	0.5151	0.7819	0.5986	0.5683	0.5573	0.8744	0.9663	0.9414
Lanczos4	0.6498	0.9488	0.5150	0.8139	0.5761	0.5216	0.5324	0.8638	0.9649	0.9380
Nearest	0.6741	0.9524	0.5117	0.8018	0.5811	0.5531	0.5416	0.8704	0.9654	0.9407
Our pre-processing method	**0.8789**	**0.9768**	**0.8610**	**0.9503**	**0.9721**	**0.9336**	**0.9593**	**0.9731**	**0.9960**	**0.9880**

where applying ensemble methods improved prediction accuracy, as seen in the ERM and IRF calibration maps. The second and third rows in Figure 4 further illustrate the benefits of adding Confidence-aware instructor (*MTL+ Ens* vs. *MTL+ Ens+ CA*), showing that calibration enhances performance by more accurately capturing uncertainty without overconfidence.

6.7. Comparison with Interpolation Methods in OCT Image Pre-processing

Table 6 compares the effects of different image resizing methods on the segmentation performance of *CAMEL*. In this study, images and masks were resized to 320 $\times$ 320 dimensions. Notably, using the resized masks generated by our proposed pre-processing method resulted in the highest Dice score, demonstrating the effectiveness of the proposed image resizing techniques for segmentation tasks in *CAMEL*.

7. Conclusion

In this paper, we introduced *CAMEL* (*Confidence-Aware Multi-task Ensemble Learning*), a model designed for accurate and comprehensive classification and segmentation of retinal OCT images. By calculating model confidence at both the pixel and image levels and employing confidence-aware ensemble learning, *CAMEL* effectively reduces task-specific uncertainty, improving both performance and reliability. Experiments on our retina OCT dataset, manually annotated by three retina specialists, show that *CAMEL* outperforms single-task models, achieving a Dice score of 0.9880 on the test set, especially in scenarios involving small regions and severe class imbalances. Additionally, validation on a publicly available dataset confirms the model's generalizability for retinal OCT image classification and segmentation. Future work will explore expanding our approach to other medical imaging domains.

Acknowledgements This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C2007625) and in part by the MSIT (Ministry of Science, ICT), Korea, under the Global Scholars Invitation Program (RS-2024-00459638) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *Osdi*, volume 16, pages 265–283. Savannah, GA, USA, 2016. 6

[2] BN Anoop, Rakesh Pavan, GN Girish, Abhishek R Kothari, and Jeny Rajan. Stack generalized deep ensemble learning for retinal layer segmentation in optical coherence tomography images. *Biocybernetics and Biomedical Engineering*, 40(4):1343–1358, 2020. 2

[3] Mustafa Arikan, James Willoughby, Sevim Ongun, Ferenc Sallo, Andrea Montesel, Hend Ahmed, Ahmed Hagag, Marius Book, Henrik Faatz, Maria Vittoria Cicinelli, et al. Oct5k: A dataset of multi-disease and multi-graded annotations for retinal layers. *bioRxiv*, pages 2023–03, 2023. 5, 7, 13, 14

[4] Rhona Asgari, José Ignacio Orlando, Sebastian Waldstein, Ferdinand Schlanitz, Magdalena Baratsits, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Multiclass segmentation as multitask learning for drusen segmentation in retinal optical coherence tomography. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 192–200. Springer, 2019. 2

[5] Christian F Baumgartner et al. Phiseg: Capturing uncertainty in medical image segmentation. In *MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 119–127. Springer, 2019. 2, 3

[6] Jeroen Bertels, David Robben, Dirk Vandermeulen, and Paul Suetens. Optimization with soft dice can lead to a volumetric bias. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5*, pages 89–97. Springer, 2020. 4

[7] Hrvoje Bogunović et al. Retouch: The retinal oct fluid detection and segmentation benchmark and challenge. *IEEE transactions on medical imaging*, 38(8):1858–1874, 2019. 2

[8] Jun Cao, Xiaoming Liu, Ying Zhang, and Man Wang. A multi-task framework for topology-guaranteed retinal layer segmentation in oct images. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3142–3147. IEEE, 2020. 2

[9] Alex Cazañas-Gordón, Esther Parra-Mora, and Luís A Da Silva Cruz. Ensemble learning approach to retinal thickness assessment in optical coherence tomography. *IEEE Access*, 9:67349–67363, 2021. 2

[10] Jeffrey De Fauw et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018. 1

[11] Joaquim de Moura, Gabriela Samagaio, Jorge Novo, Pablo Almuina, María Isabel Fernández, and Marcos Ortega. Joint diabetic macular edema segmentation and characterization in oct images. *Journal of Digital Imaging*, 33:1335–1351, 2020. 1, 2

[12] Shengyong Diao, Jinzhu Su, Changqing Yang, Weifang Zhu, Dehui Xiang, Xinjian Chen, Qing Peng, and Fei Shi. Classification and segmentation of oct images for age-related macular degeneration based on dual guidance networks. *Biomedical Signal Processing and Control*, 84:104810, 2023. 1, 2

[13] Thomas G Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2(1):110–125, 2002. 2

[14] Adrian T Fung, Justin Galvin, and Tuan Tran. Epiretinal membrane: a review. *Clinical & Experimental Ophthalmology*, 49(3):289–308, 2021. 4

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 3

[16] Mateo Gende, Joaquim de Moura, Jorge Novo, and Marcos Ortega. End-to-end multi-task learning approaches for the joint epiretinal membrane segmentation and screening in oct images. *Computerized Medical Imaging and Graphics*, 98:102068, 2022. 1, 2

[17] Peyman Gholami, Priyanka Roy, Mohana Kuppuswamy Parthasarathy, and Vasudevan Lakshminarayanan. Octid: Optical coherence tomography image database. *Computers & Electrical Engineering*, 81:106532, 2020. 2

[18] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007. 4

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 4

[20] Jinyoung Han, Seong Choi, Ji In Park, Joon Seo Hwang, Jeong Mo Han, Junseo Ko, Jeewoo Yoon, and Daniel Duck-Jin Hwang. Detecting macular disease based on optical coherence tomography using a deep convolutional network. *Journal of Clinical Medicine*, 12(3):1005, 2023. 1, 2, 5, 6, 13

[21] Jinyoung Han, Seong Choi, Ji In Park, Joon Seo Hwang, Jeong Mo Han, Hak Jun Lee, Junseo Ko, Jeewoo Yoon, and Daniel Duck-Jin Hwang. Classifying neovascular age-related macular degeneration with a deep convolutional neural network based on optical coherence tomography images. *Scientific Reports*, 12(1):2232, 2022. 1, 2

[22] Jeong Mo Han, Jinyoung Han, Junseo Ko, Juho Jung, Ji In Park, Joon Seo Hwang, Jeewoo Yoon, Jae Ho Jung, and Daniel Duck-Jin Hwang. Anti-vegf treatment outcome prediction based on optical coherence tomography images in neovascular age-related macular degeneration using a deep neural network. *Scientific Reports*, 14(1):28253, 2024. 1, 2

[23] Bilal Hassan, Shiyin Qin, and Ramsha Ahmed. Seadnet: Deep learning driven segmentation and extraction of macular fluids in 3d retinal oct scans. In *2020 IEEE ISSPIT*, pages 1–6. IEEE, 2020. 1, 2

[24] David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, et al. Optical coherence tomography. *science*, 254(5035):1178–1181, 1991. 1

[25] Siddhartha Jain, Ge Liu, Jonas Mueller, and David Gifford. Maximizing overall diversity for improved uncertainty estimates in deep ensembles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4264–4271, 2020. 2, 3

[26] Juho Jung, Jinyoung Han, Jeong Mo Han, Junseo Ko, Jeewoo Yoon, Joon Seo Hwang, Ji In Park, Gyudeok Hwang, Jae Ho Jung, and Daniel Duck-Jin Hwang. Prediction of neovascular age-related macular degeneration recurrence using optical coherence tomography images with a deep neural network. *Scientific Reports*, 14(1):5854, 2024. 1, 2

[27] Jongwoo Kim and Loc Tran. Ensemble learning based on convolutional neural networks for the classification of retinal diseases from optical coherence tomography images. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 532–537. IEEE, 2020. 2

[28] K Susheel Kumar and Nagendra Pratap Singh. Retinal disease prediction through blood vessel segmentation and classification using ensemble-based deep learning approaches. *Neural Computing and Applications*, 35(17):12495–12511, 2023. 2

[29] Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof von Kalle, and Titus Josef Brinker. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8):e36427, 2022. 2, 3

[30] Siyu Li, Zhen Li, Limin Guo, and Gui-Bin Bian. Glaucoma detection: Joint segmentation and classification framework via deep ensemble network. In *2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 678–685. IEEE, 2020. 2

[31] Jeremiah Liu, John Paisley, Marianthi-Anna Kioumourtzoglou, and Brent Coull. Accurate uncertainty estimation and decomposition in ensemble learning. *Advances in neural information processing systems*, 32, 2019. 2, 3

[32] Wen Liu, Yankui Sun, and Qingge Ji. Mdan-unet: multiscale and dual attention enhanced nested u-net architecture for segmentation of optical coherence tomography images. *Algorithms*, 13(3):60, 2020. 1, 2

[33] Xiaoming Liu, Yingjie Bai, Jun Cao, Junping Yao, Ying Zhang, and Man Wang. Joint disease classification and lesion segmentation via one-stage attention-based convolutional neural network in oct images. *Biomedical Signal Processing and Control*, 71:103087, 2022. 1

[34] Donghuan Lu, Morgan Heisler, Sieun Lee, Gavin Weiguang Ding, Eduardo Navajas, Marinko V Sarunic, and Mirza Faisal Beg. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. *Medical image analysis*, 54:100–110, 2019. 1

[35] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023. 5, 6, 13

[36] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020. 2, 3, 4, 5, 7

[37] Sachin Mehta et al. Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 893–901. Springer, 2018. 1

[38] Kristen M Meiburger, Massimo Salvi, Giulia Rotunno, Wolfgang Drexler, and Mengyang Liu. Automatic segmentation and classification methods using optical coherence tomography angiography (octa): a review and handbook. *Applied Sciences*, 11(20):9734, 2021. 1, 2

[39] Martina Melinščak. Attention-based u-net: Joint segmentation of layers and fluids from retinal oct images. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, pages 391–396. IEEE, 2023. 5, 6, 13

[40] Mousa Moradi, Yu Chen, Xian Du, and Johanna M Seddon. 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. 2

[41] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *Ieee Access*, 10:66467–66480, 2022. 2

[42] G Muni Nagamani and Eswaraiah Rayachoti. Deep learning network (dl-net) based classification and segmentation of multi-class retinal diseases using oct scans. *Biomedical Signal Processing and Control*, 88:105619, 2024. 1, 2

[43] Robi Polikar. Ensemble learning. *Ensemble machine learning: Methods and applications*, pages 1–34, 2012. 2

[44] Mohammad Rahil, BN Anoop, GN Girish, Abhishek R Kothari, Shashidhar G Koolagudi, and Jeny Rajan. A deep ensemble learning-based cnn architecture for multiclass retinal fluid segmentation in oct images. *IEEE Access*, 11:17241–17251, 2023. 2

[45] Reza Rasti, Hossein Rabbani, Alireza Mehridehnavi, and Fedra Hajizadeh. Macular oct classification using a multi-scale convolutional neural network ensemble. *IEEE transactions on medical imaging*, 37(4):1024–1034, 2017. 2

[46] Kai Ren, Ke Zou, Xianjie Liu, Yidi Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. Uncertainty-informed mutual learning for joint medical image classification and segmentation. *arXiv preprint arXiv:2303.10049*, 2023. 1, 2, 3, 5, 6, 12

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[48] Abhijit Guha Roy et al. Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical optics express*, 8(8):3627–3642, 2017. 1

[49] Berkman Sahiner, Aria Pezeshk, Lubomir M Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H Cha, Ronald M

Summers, and Maryellen L Giger. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019. 2

[50] DilipKumar Jang Bahadur Saini et al. Convolution neural network model for predicting various lesion-based diseases in diabetic macula edema in optical coherence tomography images. *Biomedical Signal Processing and Control*, 86:105180, 2023. 1

[51] Loza Bekalo Sappa, Idowu Paul Okuwobi, Mingchao Li, Yuhan Zhang, Sha Xie, Songtao Yuan, and Qiang Chen. Retfluidnet: Retinal fluid segmentation for sd-oct images using convolutional neural network. *Journal of Digital Imaging*, 34(3):691–704, 2021. 1

[52] Thomas Schlegl et al. Fully automated detection and quantification of macular fluid in oct using deep learning. *Ophthalmology*, 125(4):549–558, 2018. 1, 2

[53] Suman Sedai et al. Uncertainty guided semi-supervised segmentation of retinal layers in oct images. In *MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 282–290. Springer, 2019. 2, 3

[54] Philipp Seeböck et al. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE transactions on medical imaging*, 39(1):87–98, 2019. 2, 3

[55] William Stevenson, Claudia M Prospero Ponce, Daniel R Agarwal, Rachel Gelman, and John B Christoforidis. Epiretinal membrane: optical coherence tomography-based diagnosis and classification. *Clinical ophthalmology*, pages 527–534, 2016. 4

[56] Yubo Tan, Wen-Da Shen, Ming-Yuan Wu, Gui-Na Liu, Shi-Xuan Zhao, Yang Chen, Kai-Fu Yang, and Yong-Jie Li. Retinal layer segmentation in oct images with boundary regression and feature polarization. *IEEE Transactions on Medical Imaging*, 2023. 5, 6, 12

[57] Ignacio A Viedma, David Alonso-Caneiro, Scott A Read, and Michael J Collins. Deep learning in retinal optical coherence tomography (oct): A comprehensive survey. *Neurocomputing*, 2022. 2

[58] Jason R Wilkins, Carmen A Puliafito, Michael R Hee, Jay S Duker, Elias Reichel, Jeffery G Coker, Joel S Schuman, Eric A Swanson, and James G Fujimoto. Characterization of epiretinal membranes using optical coherence tomography. *Ophthalmology*, 103(12):2142–2151, 1996. 12

[59] Jeewoo Yoon, Jinyoung Han, Junseo Ko, Seong Choi, Ji In Park, Joon Seo Hwang, Jeong Mo Han, Kyuhwan Jang, Joonhong Sohn, Kyu Hyung Park, et al. Classifying central serous chorioretinopathy subtypes with a deep neural network using optical coherence tomography images: a cross-sectional study. *Scientific Reports*, 12(1):422, 2022. 2

[60] Yan Zhao, Xiuying Wang, Tongtong Che, Guoqing Bao, and Shuyu Li. Multi-task deep learning for medical image computing and analysis: A review. *Computers in Biology and Medicine*, 153:106496, 2023. 1, 2