entropy

IMPACT FACTOR 2.1

Indexed in: PubMed

CITESCORE 4.9

Article

# An Efficient Retinal Fluid Segmentation Network Based on Large Receptive Field Context Capture for Optical Coherence Tomography Images

Hang Qi, Weijiang Wang, Hua Dang, Yueyang Chen, Minli Jia and Xiaohua Wang

MDPI

https://doi.org/10.3390/e27010060

entropy

MDPI

Article

# An Efficient Retinal Fluid Segmentation Network Based on Large Receptive Field Context Capture for Optical Coherence Tomography Images

**Hang Qi [1], Weijiang Wang [1,2], Hua Dang [1], Yueyang Chen [1], Minli Jia [1] and Xiaohua Wang [1,2,*]**

[1] School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China; qihang@bit.edu.cn (H.Q.); wangweijiang@bit.edu.cn (W.W.); danghuabit@163.com (H.D.); yyc@bit.edu.cn (Y.C.); 7420220059@bit.edu.cn (M.J.)

[2] BIT Chongqing Institute of Microelectronics and Microsystems, Chongqing 401332, China

[*] Correspondence: xh_wong@bit.edu.cn

**Abstract:** Optical Coherence Tomography (OCT) is a crucial imaging modality for diagnosing and monitoring retinal diseases. However, the accurate segmentation of fluid regions and lesions remains challenging due to noise, low contrast, and blurred edges in OCT images. Although feature modeling with wide or global receptive fields offers a feasible solution, it typically leads to significant computational overhead. To address these challenges, we propose LKMU-Lite, a lightweight U-shaped segmentation method tailored for retinal fluid segmentation. LKMU-Lite integrates a Decoupled Large Kernel Attention (DLKA) module that captures both local patterns and long-range dependencies, thereby enhancing feature representation. Additionally, it incorporates a Multi-scale Group Perception (MSGP) module that employs Dilated Convolutions with varying receptive field scales to effectively predict lesions of different shapes and sizes. Furthermore, a novel Aggregating-Shift decoder is proposed, reducing model complexity while preserving feature integrity. With only 1.02 million parameters and a computational complexity of 3.82 G FLOPs, LKMU-Lite achieves state-of-the-art performance across multiple metrics on the ICF and RETOUCH datasets, demonstrating both its efficiency and generalizability compared to existing methods.

**Keywords:** optical coherence tomography; large kernel attention; multi-scale perception; lightweight; retinal fluid segmentation; deep learning

check for updates

Academic Editor: Amelia Carolina Sparavigna

Received: 28 November 2024
Revised: 8 January 2025
Accepted: 9 January 2025
Published: 11 January 2025

**Citation:** Qi, H.; Wang, W.; Dang, H.; Chen, Y.; Jia, M.; Wang, X. An Efficient Retinal Fluid Segmentation Network Based on Large Receptive Field Context Capture for Optical Coherence Tomography Images. *Entropy* **2025**, *27*, 60. https://doi.org/10.3390/e27010060

## 1. Introduction

Optical Coherence Tomography (OCT) is a non-invasive imaging modality that provides high-resolution cross-sectional views of retinal structures at a microscopic scale [1]. OCT utilizes a broadband light source that is split into two paths: one beam is directed at the sample to detect backscattered light, while the other is sent to a reference system. Through low-coherence interference, signals with an optical path difference within the coherence length are detected, allowing depth information to be extracted and enabling the generation of 2D (B-scan) or 3D structural images [2]. This exceptional capability to visualize the retina's intricate layer-wise architecture has established OCT as an indispensable tool in ophthalmology, particularly for the diagnosis and monitoring of conditions such as Macular Edema [3] and other retinal pathologies. The accurate segmentation of OCT images is crucial for identifying fluid accumulations, lesions, and structural abnormalities, as quantitative analysis of abnormal fluid aids in formulating targeted treatment plans.

Recent advancements in Artificial Intelligence (AI)-driven automatic segmentation have transformed diagnostic support systems, providing robust and adaptable solutions tailored to diverse clinical scenarios [4–8]. By predicting the majority of fluid regions, algorithms can significantly reduce clinicians' workload and minimize the risk of misdiagnosis, thereby enabling more efficient and reliable diagnostic interventions.

The Fully Convolutional Network (FCN) [9] is widely regarded as the foundational work in pixel-level dense prediction, playing a vital role in driving the rapid advancement of intelligent segmentation technology. As a significant evolution, UNet [10] introduces a U-shaped encoder–decoder architecture with skip connections, delivering exceptional performance and establishing itself as the de facto standard for medical image segmentation [11]. UNet implements a process analogous to information transmission, as described in information theory, through its symmetric encoder–decoder structure. In this process, the encoder compresses the input image by extracting features, while the decoder restores the image resolution to its original size to reconstruct the predicted information. The skip connections serve as an additional pathway in the information transmission process, helping to retain more low-level semantic information. This structure effectively reduces potential information loss during transmission. The U-shaped architecture has inspired a series of variants, including UNet++ [12], Att-UNet [13], and others. However, the principle of OCT imaging introduces significant challenges to lesion edge detection, as it results in high noise intensity, low contrast, and blurred edge structures. Consequently, directly extracting the edge information becomes difficult. Additionally, the varying shapes and sizes of fluid regions further complicate the segmentation task.

A series of measures have been proposed to improve the segmentation performance of OCT images. For example, Lu et al. [14] utilize graph cuts for preprocessing and apply random forest classifiers to eliminate incorrectly labeled liquid areas. RetiFluidNet [15] introduces self-adaptive attention and deep supervision strategies to enable hierarchical learning, while Rahil et al. [16] employ deep ensemble learning architectures to enhance multi-category segmentation. However, these methods are mainly based on U-shaped architectures with fixed-size convolution kernels and do not fully explore the potential of wide-range receptive fields. Blurred edge regions often exhibit subtle gradient changes, making them challenging to accurately identify using a single narrow-view convolution kernel (e.g., 3 × 3 kernel [17]). Effective boundary segmentation often necessitates the capability to model wide-range dependencies, as boundary cues typically rely on broader contextual information. This process integrates data from neighboring regions, enhancing the model's ability to infer potential boundary locations. Expanding the receptive field, as demonstrated by the use of Dilated Convolution [18] in DeepLab [19] and the multi-scale receptive field integration in PSPNet [20], is a widely adopted and effective strategy in computer vision tasks. However, despite significant advancements, the inherent local inductive bias of Convolutional Neural Networks (CNNs) can result in performance bottlenecks, as CNNs are less effective at capturing global long-range dependencies [21]. These global contextual dependencies are crucial for understanding correlations between scattered but related regions in an image, such as dispersed retinal fluid regions.

Recently, global modeling approaches have emerged as promising solutions to overcome the limitations of CNNs. Among these, Vision Transformer (ViT) [22], derived from the self-attention mechanism [23] originally developed in Natural Language Processing (NLP), has gained significant attention. While this approach effectively captures the global receptive field, it incurs significant computational costs due to its quadratic complexity [24] and high parameter requirements. This limitation restricts the model's training efficiency on small-scale datasets, particularly in medical application scenarios. Conversely, capturing local patterns remains essential, as it corresponds to the local edge and texture information in OCT images. However, Transformers demonstrate inherent limitations in this capability [25]. Therefore, some approaches aim to design CNN–Transformer hybrid models [26–28] to balance structural information and long-range relationships while optimizing computational efficiency. In the context of customized models for OCT tasks, Cao et al. [29] propose a vertical self-attention module at the bottleneck of the U-shaped network, tailored to the characteristics of retinal cross-sectional images. FNeXter [30] combines ConvNeXt [31] and Transformer architectures to capture both local details and global features. Furthermore, large kernel approaches [32–34] have garnered attention, as they are regarded as effective methods for balancing global and local feature representation. Although considerable effort has been devoted to developing resource-efficient and accurate methods, these approaches generally incur higher overhead compared to naive CNN-based models, making them less suitable for devices with limited computing resources, particularly mobile auxiliary diagnostic devices.

Motivated by the aforementioned challenges, we propose a lightweight U-shaped segmentation method, named LKMU-Lite, which integrates a Decoupled Large Kernel Attention module and multi-scale receptive field perception strategy for retinal fluid segmentation in OCT imaging. The primary contributions of this paper are summarized as follows:

1. We introduce a Decoupled Large Kernel Attention (DLKA) module that leverages the equivalent large kernel convolution receptive field to generate an attention map. This map integrates both local patterns and long-range dependencies, enhancing the salient features of the original input through a gated mechanism.
2. The Multi-scale Group Perception (MSGP) module is proposed to integrate multi-granular information by utilizing Dilated Convolutions with different receptive field scales in a grouped configuration, allowing for effective capture of features across variable lesion shapes and sizes.
3. We propose a parameter-free spatial-shift operation integrated with point-wise convolution, combined with a multi-path feature aggregation strategy, resulting in the lightweight Aggregating-Shift decoder. This innovative approach retains effective feature representation while substantially minimizing the parameters.
4. The final design of LKMU-Lite features only 1.02 M parameters. Comprehensive experiments on two OCT datasets demonstrate its state-of-the-art (SOTA) performance across multiple metrics compared to existing methods, confirming the efficiency and robustness of the proposed approach.

## 2. Related Works

UNet [10] has gained widespread adaption in medical image segmentation due to its elegant design. In this architecture, the encoder is responsible for capturing texture details and contextual features, progressively extracting meaningful semantic information as the network deepens. Simultaneously, the decoder utilizes skip connections to incrementally upsample feature maps for accurate mask prediction. Consequently, the U-shaped architecture has become a fundamental framework in numerous medical image segmentation models. For instance, Att-UNet [13] employs a gated mechanism to emphasize critical features across various spatial regions. UNet++ [12], on the other hand, enhances multi-scale feature representation by incorporating nested skip connections and establishing dense linkages between encoder and decoder layers. ResUNet++ [35] leverages attention modules alongside Atrous Spatial Pyramid Pooling, improving the fusion of information. Meanwhile, CENet [36] introduces an innovative context encoder module, designed to effectively capture and transmit contextual knowledge between the encoder and decoder components. Despite these remarkable achievements, a standard convolutional kernel in a

CNN processes only a small, localized portion of the target image, leading to a restricted receptive field and limiting its capability to capture long-range dependencies. Although there are numerous attempts in the literature to address this challenge, including approaches like Dilated Convolution [18,19] and the Pyramid Pooling module [20], these methods fall short of completely solving the problem.

Leveraging the robust representation capabilities for capturing global context information inherent in self-attention mechanisms [23], ViT [22] represents a groundbreaking effort to adapt the Transformer architecture for image recognition, achieving performance comparable to state-of-the-art models. Building on this innovation, TransUNet [37] became the first to incorporate Transformer layers within a U-shaped network, enabling effective global information modeling. Extending this idea further, TransFuse [26] combines CNN and Transformer branches, leveraging their strengths to simultaneously capture global relationships and fine-grained spatial features in a compact design. Nonetheless, the self-attention mechanism in vision tasks presents a computational challenge, with complexity increasing quadratically as image resolution rises. This arises from the division of images into fixed-size patches, which are processed as sequences to enable the direct modeling of relationships between any pair of patches. To alleviate this problem, certain lightweight models have been designed, such as MedT [38], which introduces the axial attention mechanism [39] and a global–local strategy. However, the hardware requirements for training remain high.

In recent years, large kernel convolutional methods [32–34] have gained attention for their ability to achieve high performance while balancing computational complexity, as well as their capacity to effectively balance wide-range feature interactions and local pattern modeling. Inspired by similar concepts, many recent medical image segmentation model designs have subtly incorporated this approach. For example, ConvUNeXt [31] and CMUNeXt [40] incorporate ConvNeXt [41] principles by using larger convolution kernels to enhance feature representation capabilities, while AMSUNet [42] designs an attention module using atrous multi-scale (AMS) convolution to capture larger receptive fields, and DCSAU-Net [43] enhances model compactness and efficiency by introducing the Compact Split-Attention module and improving primary feature retention. EMCAD [44] primarily employs efficient multi-scale convolutional attention modules and a designed Large-kernel Grouped Attention Gate to create a new decoder. While these methods have achieved a certain trade-off between performance and computational cost, there is still room for further improvement in developing an even more lightweight model.

## 3. Methods

Figure 1 illustrates the framework of the proposed intelligent retinal fluid segmentation method based on lightweight LKMU-Lite, comprising both training and testing stages. During the training phase, the parameters of LKMU-Lite are optimized for the target task. In the testing phase, the trained network predicts previously unseen data. Once trained, the network has the potential to be deployed on resource-constrained devices, enabling the efficient and accurate segmentation of retinal fluid.

The following sections will elaborate on the macrostructure and key components of LKMU-Lite.

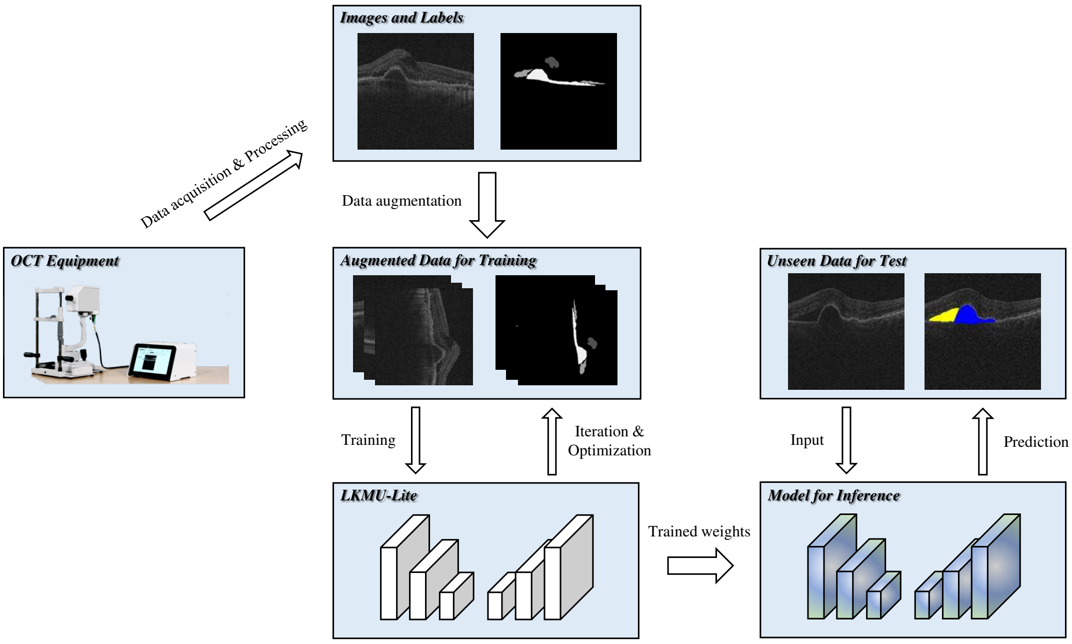


**Figure 1.** Framework of the proposed intelligent retinal fluid segmentation method.

### 3.1. Overall Architecture

The designed LKMU-Lite adopts the widely used U-shaped structure [10,11], consisting of an encoder, a decoder, and skip connections, organized into five stages, as shown in Figure 2. The number of channels in each stage is set to {32, 64, 128, 160, and 256}. In the encoder, a pooling operation is applied after each stage, except the last, to extract salient information and reduce the spatial dimensions of the feature maps. Conversely, the decoder incrementally reconstructs the output through stage-by-stage upsampling combined with convolutional modules. Skip connections link the encoder stages to their corresponding decoder stages, facilitating effective information aggregation.

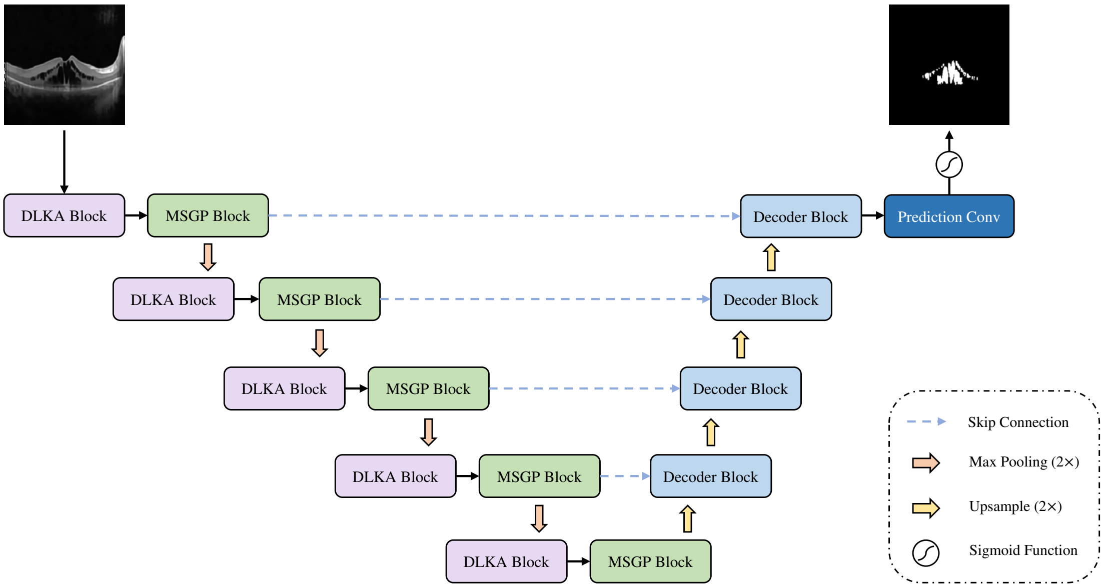


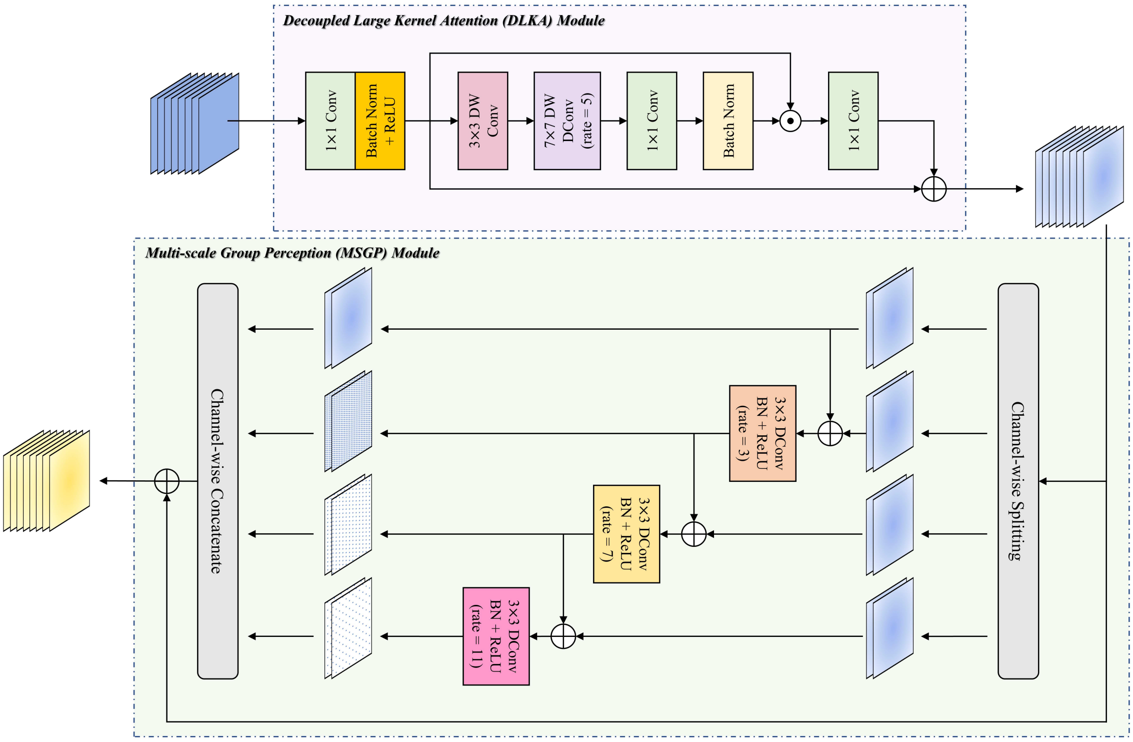
**Figure 2.** The architecture of LKMU-Lite.

Within the encoder, LKMU-Lite incorporates a Decoupled Large Kernel Attention (DLKA) module cascaded with a Multi-scale Group Perception (MSGP) module at each stage. These two modules are responsible for generating attention weights from a wide-area receptive field and for integrating features extracted from receptive fields at multiple scales, respectively. Furthermore, we propose a lightweight decoder, termed the Aggregating-Shift decoder, which combines a multi-path aggregation strategy and spatial-shift operation with point-wise convolution to achieve a balance between parameter efficiency and feature representation quality. The integration of these designs results in LKMU-Lite, which is both lightweight and high-performance.

### *3.2. Decoupled Large Kernel Attention Module*

In various vision tasks, the attention mechanism functions as a selective process that emphasizes relevant, information-rich regions while suppressing irrelevant features, making it widely applicable [45]. For instance, the squeeze-excitation (SE) module [46] focuses on modeling attention at the channel level. Its variants, the CBAM [47] and BAM [48] modules, extend this capability by integrating attention across both spatial and channel dimensions. While these attention modules are well structured and easily integrated into various backbones, they rely on fixed-size convolution kernels with limited local receptive fields, making them less effective for long-range modeling. Vision self-attention mechanisms [22], on the other hand, operate as non-local processes, addressing this limitation. However, their quadratic computational complexity introduces significant overhead and high parameter counts, increasing device requirements and potentially leading to suboptimal performance in clinical scenarios with limited data.

Building on the successful application of large kernel convolutions and their variants [32–34,49], we integrate the Decoupled Large Kernel Attention (DLKA) module into the encoder of LKMU-Lite, as depicted in the purple area of Figure 3. This module addresses the aforementioned limitations and achieves an optimal balance between capturing local patterns and modeling long-range dependencies.



Decoupled Large Kernel Attention (DLKA) Module
1×1 Conv
Batch Norm + ReLU
3×3 DW Conv
7×7 DW DConv (rate = 5)
1×1 Conv
Batch Norm
1×1 Conv
Multi-scale Group Perception (MSGP) Module
Channel-wise Concatenate
3×3 DConv BN + ReLU (rate = 3)
3×3 DConv BN + ReLU (rate = 7)
3×3 DConv BN + ReLU (rate = 11)
Channel-wise Splitting

**Figure 3.** The flowchart of the Decoupled Large Kernel Attention (DLKA) module and Multi-scale Group Perception (MSGP) module. DW Conv, DWD Conv, and DConv represent Depth-wise Convolution, Depth-wise Dilated Convolution, and standard Dilated Convolution, respectively.

Specifically, given the input $\mathbf{X}$, a point-wise convolution is first applied to expand the channel dimensions of the feature:

$$\mathbf{X}' = \rho(BN(PW(\mathbf{X}))), \tag{1}$$

where $PW(\cdot)$ denotes the point-wise convolution operation, while $BN(\cdot)$ and $\rho(\cdot)$ represent the Batch Normalization and ReLU activation function, respectively.

The core component of DLKA involves three cascaded convolutional layers, which equivalently achieve the wide receptive field of a large convolution kernel and generate the attention map $\mathbf{Att}$ based on this large receptive field:

$$\mathbf{Att} = BN\Big(PW\Big(DWD_7^5\big(DW_3(\mathbf{X}')\big)\Big)\Big). \tag{2}$$

Among these three decoupled convolution operations, $DW_3(\cdot)$ and $DWD_7^5(\cdot)$ represent depth-wise separable convolution with a kernel size of 3 × 3 and depth-wise separable Dilated Convolution with a kernel size of 7 × 7 (rate = 5), respectively. The formula for the equivalent receptive field $R$ of these two cascaded convolutional layers is given as follows:

$$R = K_1 + [K_2 + (K_2 - 1) \times (r - 1)] - 1, \tag{3}$$

where $K_1$ represents the kernel size of the Depth-wise Convolution, while $K_2$ and $r$ denote the kernel size and dilation rate of the Depth-wise Dilated Convolution, respectively. In the context of Dilated Convolution, the dilation rate defines the spacing between kernel elements during convolution operations on input data. Unlike standard convolution, where the kernel elements are contiguous, Dilated Convolution introduces gaps between kernel elements. This approach allows the receptive field of the convolutional filter to expand without increasing the number of trainable parameters or computational complexity. Consequently, it can be calculated that $DW_3(\cdot)$ and $DWD_7^5(\cdot)$ together capture a receptive field of size 33 × 33. These operations are designed to capture long-range dependencies in the spatial dimension while accounting for local patterns. Subsequently, a point-wise convolution $PW(\cdot)$ is applied to model features at the channel level.

Finally, the Hadamard product operation ($\odot$) is applied to the original feature $\mathbf{X}'$ and the attention map $\mathbf{Att}$ to emphasize important feature regions in a gating manner. This process is accompanied by a residual connection [50], which improves training convergence and enhances feature flow, resulting in the final output:

$$\mathbf{F} = PW\big(\mathbf{Att} \odot \mathbf{X}'\big) + \mathbf{X}'. \tag{4}$$

In addition, changes in the channel dimension at each stage of the encoder occur exclusively within the DLKA module. Specifically, the increase in channels is confined to the point-wise convolution operation in the first step of the DLKA module.

### *3.3. Multi-Scale Group Perception Module*

In the retinal fluid segmentation scenario for OCT imaging, receptive field information at different scales is essential. Larger receptive fields primarily capture the overall structure of lesions and their relationship with the background, while smaller and medium-sized receptive fields focus on details of the lesion area, such as edge information. Moreover, integrating receptive field information across different scales enhances the ability to process objects of varying sizes and types. Building on this foundation, we design a lightweight Multi-scale Group Perception (MSGP) module to combine and complement the DLKA module, as illustrated in the green area of Figure 3.

After receiving the output **F** from the DLKA module, the feature map is split into four parts along the channel dimension. Three sub-features extract multi-scale receptive field information using Dilated Convolutions with 3 × 3 kernels and different dilation rates, while one sub-feature retains the original semantics. Additionally, inspired by Res2Net [51], the MSGP module incorporates adders between adjacent paths to enhance feature interaction and reuse. Finally, a concatenation operation is applied along the channel dimension to restore the feature map's original size, followed by a residual connection to aggregate the output of the DLKA module. The above process can be represented by the following equations:

$$\mathbf{f_1}, \mathbf{f_2}, \mathbf{f_3}, \mathbf{f_4} = Split(\mathbf{F}), \tag{5}$$

$$\mathbf{k_i} = \begin{cases} \mathbf{f_i} & \text{if } i = 1, \\ \rho(BN(D_3(\mathbf{f_i} + \mathbf{f_{i-1}}))) & \text{if } i = 2, \\ \rho(BN(D_7(\mathbf{f_i} + \mathbf{f_{i-1}}))) & \text{if } i = 3, \\ \rho(BN(D_{11}(\mathbf{f_i} + \mathbf{f_{i-1}}))) & \text{if } i = 4, \end{cases} \tag{6}$$

$$\mathbf{Out} = Concat(\mathbf{k_1}, \mathbf{k_2}, \mathbf{k_3}, \mathbf{k_4}) + \mathbf{F}. \tag{7}$$

Here, $Split(\cdot)$ represents the division of the input feature map along the channel dimension, and $Concat(\cdot)$ denotes the concatenating operation. $D_3(\cdot)$, $D_7(\cdot)$, and $D_{11}(\cdot)$ indicate standard Dilated Convolutions with dilation rates of 3, 7, and 11, respectively.

### *3.4. Lightweight Aggregating-Shift Decoder*

The 3 × 3 convolution is commonly employed as a fundamental component of the decoder module in vanilla UNet [10] and its variations due to its effective balance between model performance and computational cost. However, the presence of skip connections introduces additional challenges: each module within a decoder stage must process feature maps from both the preceding decoder stage and the encoder branch within the same stage. As tensor channels increase due to the concatenation of multiple features, the parameters of convolution kernels also grow accordingly. Hence, this raises a natural question: how can we design a decoder with low parameters yet high representation while still preserving the capability to handle multi-channel features effectively?

A straightforward method to reduce parameters is the use of point-wise convolution. However, this strategy limits the decoder's ability to learn effectively due to its fixed receptive field and lack of local feature aggregation with neighboring pixels. Drawing inspiration from the spatial-shift MLP [52], we propose a parameter-free spatial-shift operation [53] combined with point-wise convolution to enable interaction among adjacent spatial pixels. As illustrated in Figure 4, the spatial-shift operation consists of two main steps: First, features are uniformly divided along the channel dimension into eight groups, and each group is shifted in a specific direction. The empty positions resulting from the shifts are padded with zeros. Subsequently, the original pixel at any given location is replaced with visual information from an adjacent pixel. Finally, a point-wise convolution is applied to extract relevant features. By utilizing eight distinct shift directions, each channel group effectively corresponds to a different neighboring pixel. When integrated with point-wise convolution, this method achieves an equivalent receptive field to a 3 × 3 convolution while greatly reducing both parameters and computational complexity.

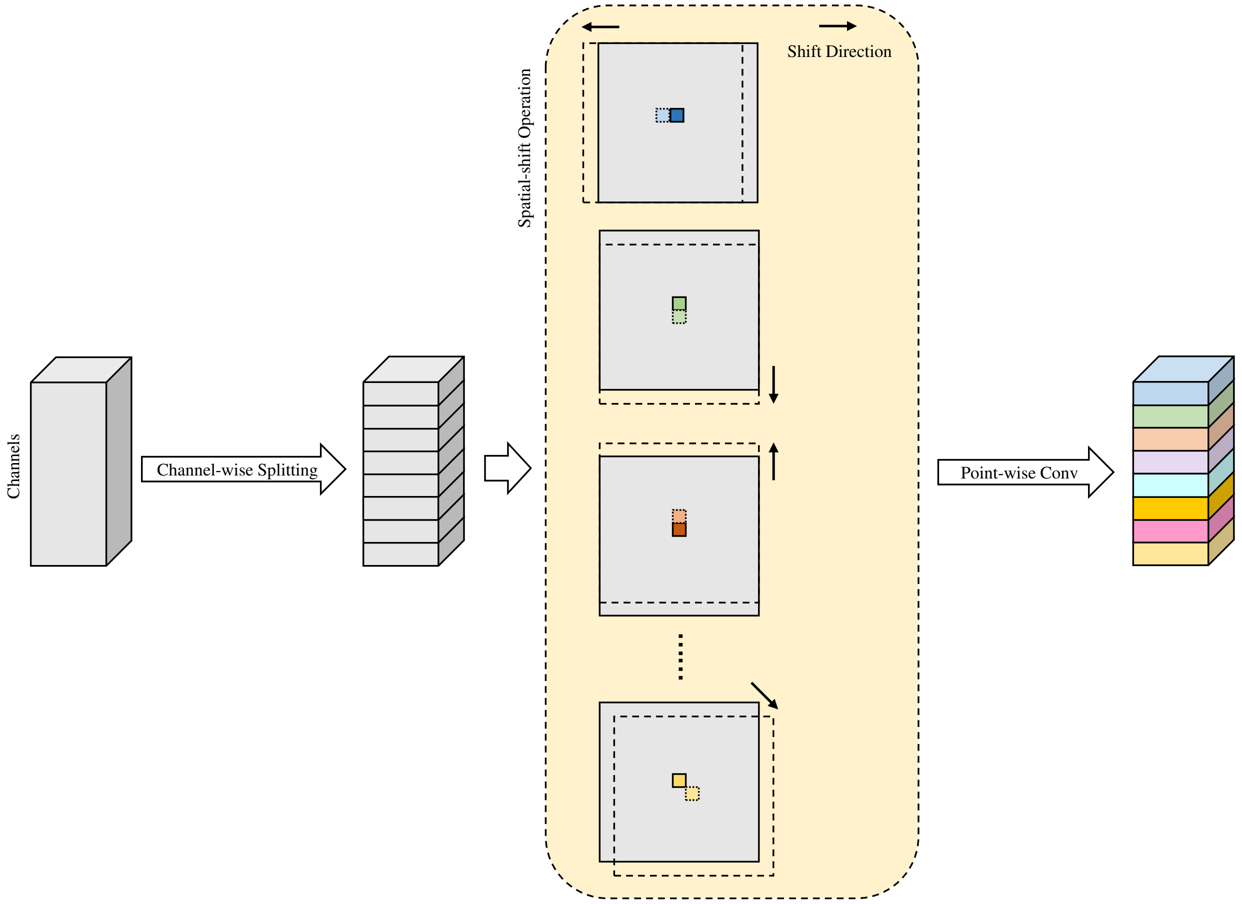


**Figure 4.** Detailed diagram of spatial-shift operation followed by point-wise convolution.

To maximize the rich semantic information from the preceding decoder and skip connection while reducing computational cost, a lightweight multi-path feature aggregation strategy is introduced for processing the concatenated feature maps. Given the input $\mathbf{X_d}$ of the decoder and the output channel $C_0$ of the current stage, the output $\mathbf{Y_d}$ of the multi-path feature aggregation module can be expressed as follows:

$$\mathbf{Y_d}|_{out=C_0} = Concat\left\{ SPPW(\mathbf{X_d})|_{out=\frac{C_0}{2}}, GC(\mathbf{X_d})|^{group=4}_{out=\frac{C_0}{4}}, DC(\mathbf{X_d})|^{rate=3}_{out=\frac{C_0}{4}} \right\}, \tag{8}$$

where $SPPW(\cdot)$, $GC(\cdot)$, and $DC(\cdot)$ denote the spatial-shift followed by point-wise convolution, $3 \times 3$ group-wise convolution with $groups = 4$, and $3 \times 3$ Dilated Convolution with $rate = 3$, respectively. The multi-path feature aggregation is followed by two spatial-shift operations combined with point-wise convolutions, integrated with a residual connection, resulting in the lightweight Aggregating-Shift decoder, as depicted in Figure 5. The final output of the decoder module is obtained as follows:

$$\mathbf{Y} = SPPW(SPPW(\mathbf{Y_d})) + \mathbf{Y_d}. \tag{9}$$

After sequential processing by the decoder at each stage, the features pass through a prediction head with a point-wise convolutional layer. The number of output channels corresponds to the number of categories.

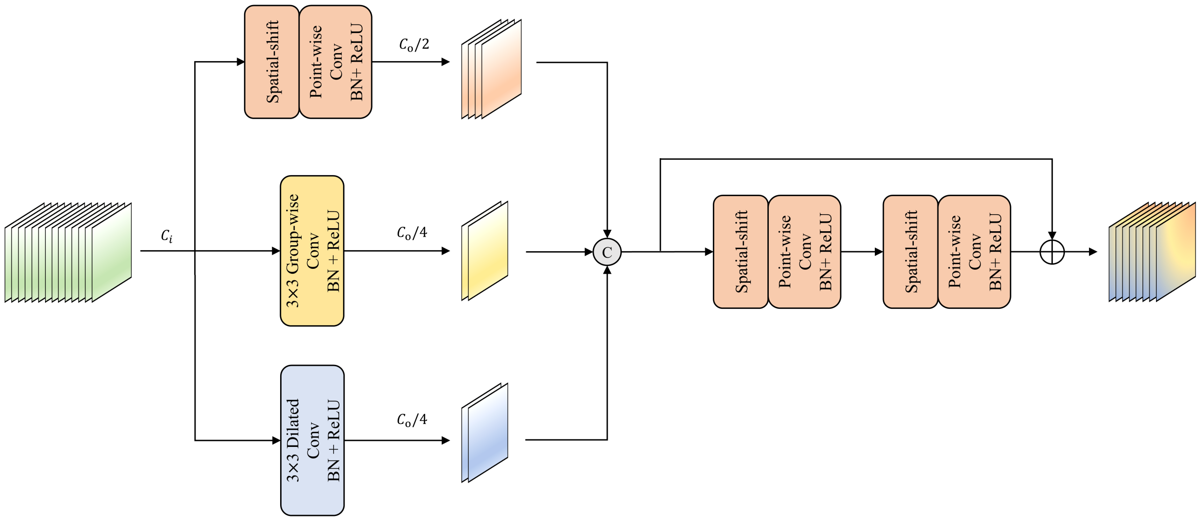


**Figure 5.** Architecture of the lightweight Aggregating-Shift decoder.

## 4. Experiments and Discussion

### *4.1. Datasets*

All experiments are conducted using two public retinal OCT datasets: Intraretinal Cystoid Fluid (ICF) [54] and the Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge (RETOUCH) [55].

The ICF dataset is derived from OCT scans of patients with Diabetic Macular Edema (DME) and includes 1006 images with corresponding annotations for a single category: Cystoid Macular Edema (CME). These OCT images are meticulously annotated and selected by experts, featuring multiple CME regions with diverse shapes and sizes. Furthermore, all images in the ICF dataset have been preprocessed by the publisher, uniformly resized to 320 × 320, and their image quality enhanced using techniques such as morphological filtering, denoising, deconvolution, and contrast enhancement.

The RETOUCH dataset comprises 6936 image slices from 70 patient cases, including scans obtained from various imaging devices such as Spectralis, Cirrus, and Topcon, ensuring diversity in imaging characteristics. The image resolutions captured by the three devices vary. This dataset focuses on three primary fluid types: Intraretinal Fluid (IRF), Subretinal Fluid (SRF), and Pigment Epithelial Detachment (PED), with expert annotations provided for each. Unlike the ICF dataset, this dataset does not undergo deliberate quality enhancement steps. RETOUCH is widely used for benchmarking algorithms addressing retinal diseases like Age-related Macular Degeneration (AMD) and DME, serving as a critical resource for advancing OCT image analysis.

Visualization examples from the RETOUCH and ICF datasets are presented in Figure 6, both datasets consist of raw grayscale images. However, the more complex scene morphology and the influence of multi-brand devices make the RETOUCH dataset more challenging than the ICF dataset. Specifically, the RETOUCH dataset comprises images captured from three different devices, each exhibiting significantly different contrast and noise levels. This multi-brand nature requires addressing distributional differences between devices during model training, increasing the dataset's complexity. Moreover, the annotated regions in the slice samples vary widely in size and display relatively complex category differences. For instance, multiple lesion categories may appear simultaneously in the same sample, or in some cases, only partially. We evaluate our method on two datasets with varying magnitudes, image quality, and classification difficulty to comprehensively assess its robustness. Each dataset is randomly partitioned into training, validation, and test sets with an 8:1:1 ratio.

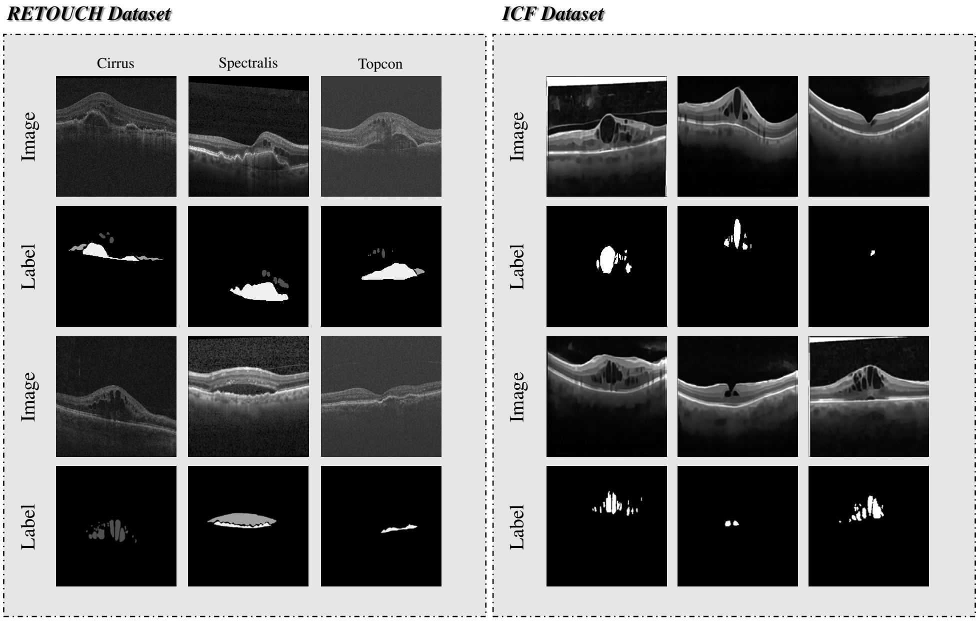


**Figure 6.** Visualizations of the RETOUCH and ICF datasets, showing both the raw images and corresponding visual labels. Compared to the ICF dataset, the RETOUCH dataset exhibits more complex scene morphology and distinct characteristics of multi-brand devices.

### *4.2. Evaluation Metrics*

The performance of the proposed LKMU-Lite on the ICF dataset is evaluated using five widely adopted metrics: Intersection over Union (IoU), Dice Similarity Coefficient (DSC), the 95th percentile Hausdorff Distance (HD95), Precision (Pre), and Sensitivity (Sen). IoU and DSC are overlap-based metrics that quantify the agreement between the predicted segmentation and the ground truth, with higher values indicating better overlap. HD95 measures the spatial distance between segmentation boundaries, emphasizing worst-case alignment errors and providing insights into boundary accuracy, where lower values signify better alignment. Precision evaluates the proportion of correctly predicted positive pixels among all predicted positives, offering an indication of the false positive rate. Sensitivity, also referred to as recall, evaluates the model's ability to capture all true positives, highlighting its effectiveness in detecting target regions. Together, these metrics provide a comprehensive evaluation of the model's segmentation performance, addressing region overlap, boundary precision, and classification accuracy.

For the RETOUCH dataset, IoU and DSC are used to evaluate the segmentation accuracy across three categories. Additionally, the mean IoU and mean DSC are computed to offer a comprehensive assessment of performance.

### *4.3. Implementation Details*

All experiments are conducted utilizing the PyTorch framework on a workstation equipped with an Nvidia RTX 3090 GPU (24 GB). The Adam optimizer is employed, initialized with a learning rate of $1 \times 10^{-3}$ and a weight decay of $1 \times 10^{-4}$. Each model is trained over 150 epochs with a batch size of eight. The model training employs a polynomial learning rate decay strategy with $power = 0.9$, where the learning rate $\eta$ at each epoch is adjusted according to the following formula:

$$\eta(\text{epoch}) = \eta_{\text{init}} \left(1 - \frac{\text{epoch}}{\text{num\_epochs}}\right)^{power}. \tag{10}$$

This approach ensures a gradual and smooth reduction in the learning rate, promoting stable convergence and enhancing performance. The model achieving the lowest validation loss is selected for generating predictions on the test set.

All images are normalized to a standard normal distribution using the mean and standard deviation derived from the pixel distribution statistics of the ImageNet dataset and resized to a resolution of 256 × 256. Data augmentation techniques such as horizontal flipping, vertical flipping, and random rotation are employed. To ensure a fair comparison, all experiments are conducted under consistent settings, with all models trained from scratch.

The loss function of our method is a composite of Binary Cross-entropy (BCE) loss and Dice Loss. Cross-entropy is used to measure the difference between two probability distributions and encode the average amount of information required for the real distribution based on the predicted distribution. BCE Loss minimizes prediction uncertainty, enabling the model to concentrate the output probability distribution on the true category as much as possible. However, it primarily focuses on pixel-level classification and is sensitive to imbalanced category ratios. Dice Loss, on the other hand, plays a complementary role by directly optimizing the overlap between the target and predicted areas. It is more robust to category imbalances and is particularly effective for segmenting small and medium-sized targets. The descriptions of these two losses are as follows:

$$L_{BCE}(P,G) = -\frac{1}{N}\sum_{i=1}^{N}[g_i \cdot \log(p_i) + (1-g_i) \cdot \log(1-p_i)], \tag{11}$$

$$L_{Dice}(P,G) = 1 - (2\sum_{i=1}^{N} p_i \cdot g_i + \epsilon)/(\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \epsilon), \tag{12}$$

where $P$ represents the predicted segmentation, and $G$ denotes the ground truth. Here, $p_i$ and $g_i$ represent individual pixels in $P$ and $G$, respectively. The total number of pixels is denoted by $N$, and $\epsilon$ serves as a smoothing term. With the specific weights often determined based on empirical experience, the comprehensive loss function is expressed as follows:

$$L = 0.5 \cdot L_{BCE}(P,G) + L_{Dice}(P,G). \tag{13}$$

### *4.4. Comparison with State-of-the-Art Models*

We compare the proposed LKMU-Lite with a series of state-of-the-art medical image segmentation models. Specifically, the compared methods include UNet [10] and its common variants [12,13,35,36], segmentation models utilizing Vision Transformers [26,37,38], and several methods proposed within the past two years [31,40,42–44].

Tables 1 and 2 show the quantitative results of various baselines on the ICF and RETOUCH datasets, respectively. Furthermore, to assess efficiency, we present the parameters (Params) and computational complexity of LKMU-Lite alongside other methods in Table 1. The model complexity is evaluated based on Floating-Point Operations (FLOPs), with both FLOPs and parameters determined using an input image of size 256 × 256 and a batch size of one. According to the results, LKMU-Lite outperforms all compared methods across every metric on the ICF dataset. The RETOUCH dataset presents a greater challenge due to the lack of preprocessing to enhance image quality, its complex scene morphology, and the influence of multi-brand devices inherent to the task. However, apart from the IoU score for SRF segmentation, which is slightly lower than that of TransFuse-S, LKMU-Lite still achieves the most competitive results across other metrics. Notably, in the IRF and PED categories, LKMU-Lite demonstrates a significant performance advantage, resulting in the best mean IoU and DSC scores.

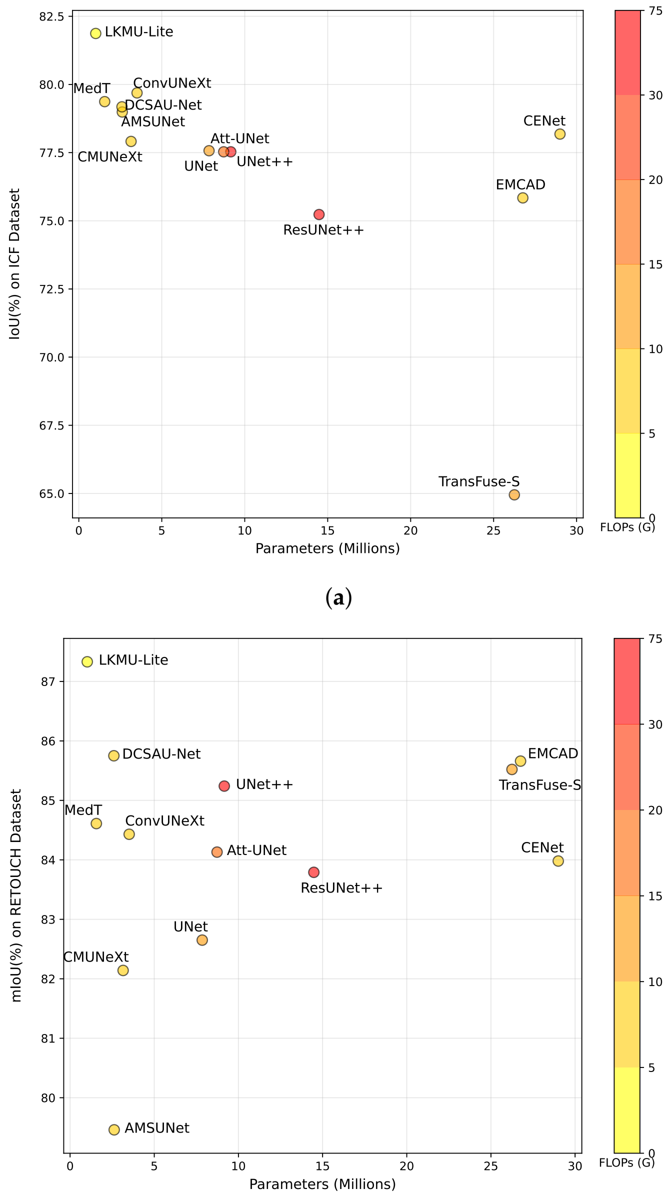
**Table 1.** Quantitative evaluation results on the ICF dataset compared to existing SOTA methods. Parameters and computational complexity for each model are also provided. The symbol "↑" indicates that higher values are preferred, whereas "↓" indicates that lower values are better. The best results are highlighted in bold.

| **Model** | **Year** | **Metrics** | | | | | **Resource Usage** | |
|---|---|---|---|---|---|---|---|---|
| | | **IoU (%) ↑** | **DSC (%) ↑** | **HD95 ↓** | **Pre (%) ↑** | **Sen (%) ↑** | **Params ↓** | **FLOPs ↓** |
| UNet [10] | 2015 | 77.57 | 86.90 | 3.64 | 88.82 | 86.72 | 7.85 M | 14.01 G |
| UNet++ [12] | 2018 | 77.53 | 86.81 | 4.50 | 89.55 | 85.98 | 9.16 M | 34.66 G |
| Att-UNet [13] | 2018 | 77.53 | 86.92 | 4.73 | 89.36 | 86.18 | 8.73 M | 16.76 G |
| CENet [36] | 2019 | 78.18 | 87.31 | 3.74 | 89.32 | 86.66 | 29.00 M | 7.16 G |
| ResUNet++ [35] | 2019 | 75.23 | 85.13 | 4.46 | 87.07 | 85.21 | 14.48 M | 70.77 G |
| TransUNet [37] | 2021 | 78.87 | 87.68 | 4.39 | 89.19 | 87.49 | 105.32 M | 38.53 G |
| TransFuse-S [26] | 2021 | 64.95 | 76.23 | 9.43 | 82.20 | 73.41 | 26.25 M | 11.53 G |
| MedT [38] | 2021 | 79.37 | 88.12 | 3.94 | 89.94 | 87.45 | 1.56 M | 5.62 G |
| ConvUNeXt [31] | 2022 | 79.69 | 88.26 | 2.66 | 90.69 | 87.34 | 3.51 M | 7.18 G |
| AMSUNet [42] | 2023 | 78.99 | 87.87 | 3.60 | 89.51 | 87.73 | 2.62 M | 6.06 G |
| DCSAU-Net [43] | 2023 | 79.18 | 87.96 | 4.05 | 89.43 | 88.02 | 2.60 M | 6.72 G |
| CMUNeXt [40] | 2024 | 77.91 | 87.12 | 2.95 | 90.16 | 85.84 | 3.15 M | 7.32 G |
| EMCAD [44] | 2024 | 75.84 | 85.79 | 3.38 | 87.13 | 85.80 | 26.76 M | 5.83 G |
| LKMU-Lite (ours) | 2024 | **81.87** | **89.66** | **2.11** | **91.13** | **89.16** | **1.02 M** | **3.82 G** |

**Table 2.** Quantitative evaluation results on the RETOUCH dataset. The best results are highlighted in bold.

| **Model** | **Year** | **IoU (%) ↑** | | | | **DSC (%) ↑** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **IRF** | **SRF** | **PED** | **Mean** | **IRF** | **SRF** | **PED** | **Mean** |
| UNet [10] | 2015 | 80.31 | 90.32 | 77.33 | 82.65 | 82.86 | 91.89 | 76.62 | 83.79 |
| UNet++ [12] | 2018 | 80.28 | 89.47 | 85.97 | 85.24 | 82.81 | 90.91 | 87.20 | 86.97 |
| Att-UNet [13] | 2018 | 79.05 | 88.57 | 84.76 | 84.13 | 81.81 | 89.46 | 85.96 | 85.74 |
| CENet [36] | 2019 | 78.12 | 89.33 | 84.50 | 83.98 | 81.05 | 91.02 | 85.92 | 86.00 |
| ResUNet++ [35] | 2019 | 79.46 | 89.58 | 82.33 | 83.79 | 81.98 | 90.88 | 83.15 | 85.34 |
| TransUNet [37] | 2021 | 79.84 | 90.90 | 87.42 | 86.05 | 82.57 | 92.22 | 88.89 | 87.89 |
| TransFuse-S [26] | 2021 | 79.07 | **91.08** | 86.42 | 85.52 | 81.85 | 92.72 | 88.05 | 87.54 |
| MedT [38] | 2021 | 78.96 | 89.66 | 85.21 | 84.61 | 81.76 | 91.06 | 86.60 | 86.47 |
| ConvUNeXt [31] | 2022 | 78.72 | 89.16 | 85.42 | 84.43 | 81.46 | 90.28 | 86.66 | 86.13 |
| AMSUNet [42] | 2023 | 75.10 | 80.72 | 82.56 | 79.46 | 76.96 | 80.52 | 83.80 | 80.43 |
| DCSAU-Net [43] | 2023 | 80.19 | 90.29 | 86.77 | 85.75 | 83.01 | 91.60 | 88.57 | 87.73 |
| CMUNeXt [40] | 2024 | 77.89 | 83.91 | 84.61 | 82.14 | 80.77 | 84.69 | 85.68 | 83.71 |
| EMCAD [44] | 2024 | 79.54 | 90.85 | 86.59 | 85.66 | 82.30 | 92.24 | 87.59 | 87.38 |
| LKMU-Lite (ours) | 2024 | **81.97** | 90.98 | **89.04** | **87.33** | **85.31** | **92.78** | **91.15** | **89.75** |

In addition to its performance advantages, LKMU-Lite is highly efficient in terms of computing resources. For instance, compared to the vanilla UNet, our model has 6.7 times fewer parameters and 2.67 times lower computational complexity. While the Transformer-based TransUNet, which features a global receptive field modeling capability, achieves competitive performance on both datasets, it comes at the cost of requiring 102.25 times more parameters and 9.09 times higher complexity compared to LKMU-Lite. According to the scatter plot in Figure 7, scatter points closer to the upper left corner indicate a better balance between efficiency and performance. This suggests that LKMU-Lite achieves an optimal trade-off between these two factors.



(**a**)

(**b**)

**Figure 7.** Scatter plots comparing various methods based on parameters, computational complexity, and mean IoU (mIoU) performance on (**a**) the ICF dataset and (**b**) the RETOUCH dataset. Note that TransUNet, which has significantly higher parameters, is excluded to provide a clearer visualization of the performance differences among the compared models.

Figure 8 presents a comparative analysis of segmentation visualizations, with predicted regions overlaid using non-transparent colors on the original image to represent different target categories. Other methods exhibit more evident mis-segmentation or even missed segmentation in the slice examples shown in the figure. In contrast, LKMU-Lite, leveraging the combination of the large kernel attention mechanism and multi-scale receptive field integration, provides a more accurate depiction of retinal fluid areas of varying shapes and sizes.

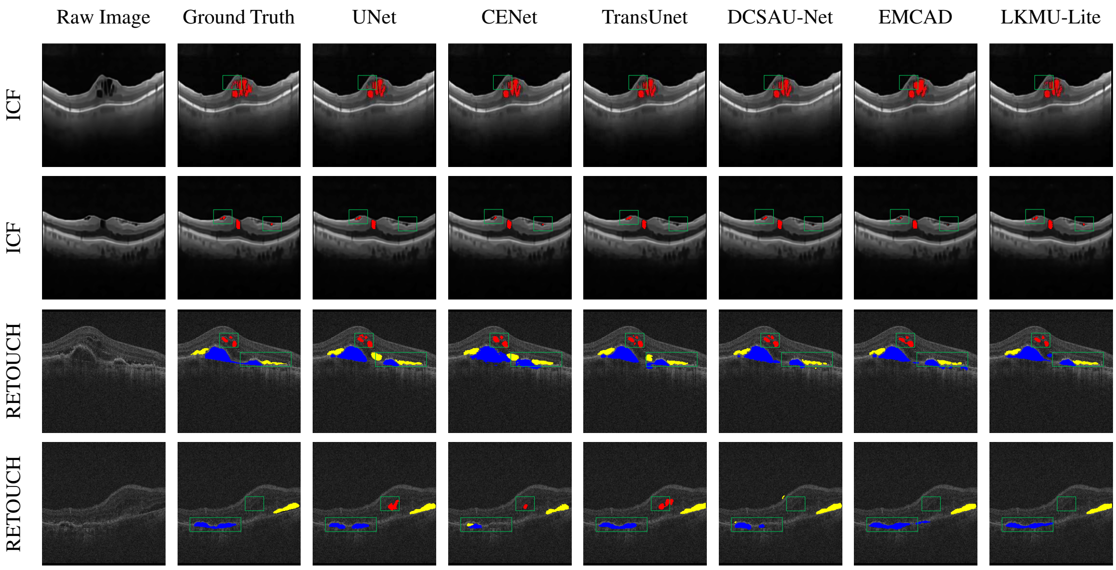


**Figure 8.** Comparison of visualized predictions between LKMU-Lite and selected baseline models on the test sets of the ICF and RETOUCH datasets. The red color represents Intraretinal Fluid (IRF), while the Cystoid Macular Edema (CME) in the ICF dataset is a specific form of IRF. The yellow and blue colors represent Subretinal Fluid (SRF) and Pigment Epithelial Detachment (PED), respectively. The areas highlighted by green boxes emphasize the improvements of our method over other models in mitigating mis-segmentation and missed segmentation issues.

### *4.5. Ablation Analysis*

To validate the effectiveness of the individual components in LKMU-Lite, we conduct ablation studies on both the ICF and RETOUCH datasets. The experimental configurations are categorized into three groups as follows:

(1) To determine the most suitable equivalent Receptive Field Size for the DLKA module, which is primarily controlled by the Depth-wise Dilated Convolution (DWD Conv) kernel size, we evaluate four models with different configurations: A (DWD Conv with a 3 × 3 kernel), B (DWD Conv with a 5 × 5 kernel), C (LKMU-Lite), and D (DWD Conv with a 9 × 9 kernel).

(2) The effectiveness of the proposed DLKA and MSGP modules are evaluated, respectively. Specifically, the DLKA module is replaced with a 3 × 3 convolutional block from VGG [17], while the MSGP module remains unchanged, resulting in a model named 3 × 3 Conv + MSGP. Conversely, when the MSGP module is replaced with a 3 × 3 convolutional block, the modified model is denoted as DLKA + 3 × 3 Conv. Additionally, replacing both the DLKA and MSGP modules with 3 × 3 convolutional blocks results in an encoder backbone similar to that of a vanilla UNet [10], and the modified model is named 3 × 3 Conv + 3 × 3 Conv.

(3) To evaluate the proposed Aggregating-Shift decoder, we introduce two variants of the decoder module: one composed entirely of 3 × 3 convolutions, forming a standard convolutional decoder with a residual connection, and another that replaces all spatial-shift point-wise convolutions with 3 × 3 convolutions while retaining the multi-path feature aggregation structure. The modified models are named 3 × 3 Conv Decoder and 3 × 3 Conv Decoder-mod, respectively.

The quantitative results under different equivalent receptive field settings are presented in Tables 3 and 4. The optimal DWD Conv kernel size is 7 × 7 rather than 9 × 9. The experimental results indicate that an excessively large receptive field is not necessary for achieving good segmentation performance. This is because increasing the receptive field too much may introduce unnecessary features, leading to information redundancy and negatively impacting the model's performance when handling lesions with rich local details. Therefore, at the feature modeling level, a balance between wide-range dependencies and local patterns should be considered.

**Table 3.** Evaluation on the ICF dataset under different equivalent receptive field settings, where "RF size" stands for Receptive Field Size.

| **Settings** | **DW Kernel** | **DWD Kernel** | **RF Size** | **Metrics** | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | **IoU (%) ↑** | **DSC (%) ↑** | **HD95 ↓** | **Pre (%) ↑** | **Sen (%) ↑** |
| A | 3 × 3 | 3 × 3 | 13 × 13 | 81.61 | 89.54 | 2.64 | 89.76 | **90.13** |
| B | 3 × 3 | 5 × 5 | 23 × 23 | 80.17 | 88.59 | **2.01** | 90.10 | 87.99 |
| C (LKMU-Lite) | 3 × 3 | 7 × 7 | 33 × 33 | **81.87** | **89.66** | 2.11 | **91.13** | 89.16 |
| D | 3 × 3 | 9 × 9 | 43 × 43 | 79.84 | 88.44 | 2.50 | 89.20 | 88.76 |

**Table 4.** Evaluation on the RETOUCH dataset under different equivalent receptive field settings.

| **Settings** | **DW Kernel** | **DWD Kernel** | **RF Size** | **IoU (%) ↑/DSC (%) ↑** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **IRF** | **SRF** | **PED** | **Mean** |
| A | 3 × 3 | 3 × 3 | 13 × 13 | 80.84/84.14 | 90.88/92.63 | 87.82/89.78 | 86.51/88.85 |
| B | 3 × 3 | 5 × 5 | 23 × 23 | 81.46/84.75 | 91.20/92.93 | 87.41/89.16 | 86.69/88.95 |
| C (LKMU-Lite) | 3 × 3 | 7 × 7 | 33 × 33 | **81.97/85.31** | 90.98/92.78 | **89.04/91.15** | **87.33/89.75** |
| D | 3 × 3 | 9 × 9 | 43 × 43 | 81.10/84.30 | **91.64/93.37** | 88.44/90.12 | 87.06/89.26 |

The analysis of the effectiveness of the DLKA and MSGP modules is shown in Tables 5 and 6. For 3 × 3 Conv + 3 × 3 Conv, which stacks the encoder using a fixed-size kernel convolution, the performance is similar to that of UNet, as shown in Tables 1 and 2. This further demonstrates that the limited receptive field capture range struggles to adequately address the variable lesion feature patterns in OCT imaging. On this basis, the introduction of either the DLKA or MSGP modules not only improves performance but also helps reduce parameters and computational complexity. Furthermore, the combination of the large kernel attention mechanism (DLKA) and the multi-scale receptive field integration strategy (MSGP) demonstrates the most competitive performance.

**Table 5.** Ablation study of the DLKA and MSGP modules on the ICF dataset.

| **Settings** | **Metrics** | | | | | **Resource Usage** | |
|---|---|---|---|---|---|---|---|
| | **IoU (%) ↑** | **DSC (%) ↑** | **HD95 ↓** | **Pre (%) ↑** | **Sen (%) ↑** | **Params ↓** | **FLOPs ↓** |
| 3 × 3 Conv + 3 × 3 Conv | 79.24 | 87.92 | 3.68 | 89.99 | 87.20 | 2.15 M | 5.73 G |
| 3 × 3 Conv + MSGP | 81.29 | 89.37 | 2.57 | **91.17** | 88.62 | 1.33 M | 3.94 G |
| DLKA + 3 × 3 Conv | 80.22 | 88.68 | **2.03** | 90.70 | 87.79 | 1.85 M | 5.61 G |
| DLKA + MSGP (LKMU-Lite) | **81.87** | **89.66** | 2.11 | 91.13 | **89.16** | **1.02 M** | **3.82 G** |

**Table 6.** Ablation study of the DLKA and MSGP modules on the RETOUCH dataset.

| **Settings** | **IoU (%) ↑** | | | | **DSC (%) ↑** | | | |
|---|---|---|---|---|---|---|---|---|
| | **IRF** | **SRF** | **PED** | **Mean** | **IRF** | **SRF** | **PED** | **Mean** |
| 3 × 3 Conv + 3 × 3 Conv | 80.19 | **91.14** | 76.36 | 82.56 | 82.93 | 92.49 | 75.73 | 83.72 |
| 3 × 3 Conv + MSGP | 80.67 | 90.93 | 87.71 | 86.44 | 83.94 | 92.27 | 89.36 | 88.52 |
| DLKA + 3 × 3 Conv | 80.82 | 90.83 | 87.91 | 86.52 | 83.81 | 92.58 | 89.88 | 88.76 |
| DLKA + MSGP (LKMU-Lite) | **81.97** | 90.98 | **89.04** | **87.33** | **85.31** | **92.78** | **91.15** | **89.75** |

For the evaluation of the decoder, by comparing the Aggregating-Shift decoder with the 3 × 3 Conv decoder-mod from Table 7 and 8, it is evident that the spatial-shift combined with point-wise convolution significantly reduces computational overhead, specifically lowering parameters by approximately 55% and complexity by around 62%, without sacrificing feature representation. Additionally, there is a significant performance gap between the pure 3 × 3 Conv decoder and the modified 3 × 3 Conv decoder-mod. This demonstrates that introducing the multi-path feature aggregation method enriches the gradient return path without increasing overhead, which is beneficial for accurate segmentation.

**Table 7.** Ablation study of the Aggregating-Shift decoder on the ICF dataset.

| Settings | Metrics | | | | | Resource Usage | |
|---|---|---|---|---|---|---|---|
| | IoU (%) ↑ | DSC (%) ↑ | HD95 ↓ | Pre (%) ↑ | Sen (%) ↑ | Params ↓ | FLOPs ↓ |
| 3 × 3 Conv decoder | 79.85 | 88.45 | 2.67 | 89.71 | 88.28 | 2.45 M | 10.99 G |
| 3 × 3 Conv decoder-mod | 81.22 | 89.24 | 2.48 | **91.77** | 88.14 | 2.25 M | 9.95 G |
| Aggregating-Shift decoder | **81.87** | **89.66** | **2.11** | 91.13 | **89.16** | **1.02 M** | **3.82 G** |

**Table 8.** Ablation study of the Aggregating-Shift decoder on the RETOUCH dataset.

| Settings | IoU (%) ↑ | | | | DSC (%) ↑ | | | |
|---|---|---|---|---|---|---|---|---|
| | IRF | SRF | PED | Mean | IRF | SRF | PED | Mean |
| 3 × 3 Conv decoder | 77.48 | 87.49 | 84.69 | 83.22 | 79.94 | 88.76 | 85.33 | 84.68 |
| 3 × 3 Conv decoder-mod | 81.26 | 90.42 | 88.15 | 86.61 | 84.51 | 91.92 | 89.80 | 88.74 |
| Aggregating-Shift decoder | **81.97** | **90.98** | **89.04** | **87.33** | **85.31** | **92.78** | **91.15** | **89.75** |

## 5. Conclusions

In this paper, we introduce LKMU-Lite, a lightweight U-shaped segmentation model specifically designed for retinal fluid segmentation. Concretely, we incorporate a Decoupled Large Kernel Attention (DLKA) module to capture both local patterns and long-range dependencies, enhancing feature representation through a gated mechanism. Additionally, a Multi-scale Group Perception (MSGP) module is designed to capture multi-granular features using Dilated Convolutions at varying receptive field scales, enabling the effective segmentation of different lesion shapes and sizes. Finally, the lightweight Aggregating-Shift decoder reduces computational overhead while preserving segmentation accuracy by combining spatial-shift operations and point-wise convolutions. Extensive experiments on the ICF and RETOUCH datasets demonstrate the superiority of LKMU-Lite over existing models while requiring significantly fewer parameters and maintaining low complexity. Additionally, LKMU-Lite holds potential for application in medical image segmentation tasks across other imaging modalities. In future works, we will integrate cross-validation strategies to further explore the performance of LKMU-Lite in other tasks and make appropriate improvements to develop a more real-time and versatile segmentation architecture.

**Author Contributions:** Conceptualization, H.Q. and W.W.; methodology, H.Q. and X.W.; software, H.Q.; validation, H.Q. and H.D.; formal analysis, Y.C. and M.J.; investigation, H.Q. and M.J.; resources, W.W. and H.D.; data curation, Y.C. and M.J.; writing—original draft preparation, H.Q.; writing—review and editing, Y.C. and X.W.; visualization, H.Q.; supervision, W.W. and H.D.; project administration, X.W.; funding acquisition, H.D. and X.W. All authors have read and agreed to the published version of this manuscript.

**Funding:** This work was supported in part by the Natural Science Foundation of Chongqing Municipality, China (Grant No. cstc2021jcyj-msxmX1090) and the Key Laboratory of Medical Electronics and Digital Health of Zhejiang Province, China (Grant No. MEDC202304).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data can be accessed upon reasonable request to the corresponding authors.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| OCT | Optical Coherence Tomography |
| AI | Artificial Intelligence |
| FCN | Fully Convolutional Network |
| CNN | Convolutional Neural Network |
| ViT | Vision Transformer |
| NLP | Natural Language Processing |
| DLKA | Decoupled Large Kernel Attention |
| MSGP | Multi-scale Group Perception |
| SOTA | state-of-the-art |
| SE | squeeze-excitation |
| DW Conv | Depth-wise Convolution |
| DWD Conv | Depth-wise Dilated Convolution |
| DConv | Dilated Convolution |
| ICF | Intraretinal Cystoid Fluid |
| RETOUCH | Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge |
| DME | Diabetic Macular Edema |
| CME | Cystoid Macular Edema |
| IRF | Intraretinal Fluid |
| SRF | Subretinal Fluid |
| PED | Pigment Epithelial Detachment |
| AMD | Age-related Macular Degeneration |
| IoU | Intersection over Union |
| DSC | Dice Similarity Coefficient |
| HD95 | 95th percentile Hausdorff Distance |
| Pre | Precision |
| Sen | Sensitivity |
| BCE | Binary Cross-entropy |
| Params | parameters |
| FLOPs | Floating-Point Operations |
| mIoU | mean Intersection over Union |
| RF Size | Receptive Field Size |

## References

1. Huang, D.; Swanson, E.A.; Lin, C.P.; Schuman, J.S.; Stinson, W.G.; Chang, W.; Hee, M.R.; Flotte, T.; Gregory, K.; Puliafito, C.A.; et al. Optical coherence tomography. *Science* **1991**, *254*, 1178–1181. [CrossRef] [PubMed]
2. Walther, J.; Gaertner, M.; Cimalla, P.; Burkhardt, A.; Kirsten, L.; Meissner, S.; Koch, E. Optical coherence tomography in biomedical research. *Anal. Bioanal. Chem.* **2011**, *400*, 2721–2743. [CrossRef] [PubMed]
3. Dysli, M.; Rückert, R.; Munk, M.R. Differentiation of underlying pathologies of macular edema using spectral domain optical coherence tomography (SD-OCT). *Ocul. Immunol. Inflamm.* **2019**, *27*, 474–483. [CrossRef]
4. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [CrossRef]

5. Zhu, W.; Liu, C.; Fan, W.; Xie, X. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 673–681.
6. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Cham, Switzerland, 2020; pp. 263–273.
7. Cao, K.; Xia, Y.; Yao, J.; Han, X.; Lambert, L.; Zhang, T.; Tang, W.; Jin, G.; Jiang, H.; Fang, X.; et al. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nat. Med.* **2023**, *29*, 3033–3043. [CrossRef]
8. Wang, Y.R.; Yang, K.; Wen, Y.; Wang, P.; Hu, Y.; Lai, Y.; Wang, Y.; Zhao, K.; Tang, S.; Zhang, A.; et al. Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. *Nat. Med.* **2024**, *30*, 1471–1480. [CrossRef]
9. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
11. Azad, R.; Aghdam, E.K.; Rauland, A.; Jia, Y.; Avval, A.H.; Bozorgpour, A.; Karimijafarbigloo, S.; Cohen, J.P.; Adeli, E.; Merhof, D. Medical image segmentation review: The success of u-net. *arXiv* **2022**, arXiv:2211.14830. [CrossRef]
12. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018*; Proceedings 4; Springer: Cham, Switzerland, 2018; pp. 3–11.
13. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [CrossRef]
14. Lu, D.; Heisler, M.; Lee, S.; Ding, G.W.; Navajas, E.; Sarunic, M.V.; Beg, M.F. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. *Med. Image Anal.* **2019**, *54*, 100–110. [CrossRef]
15. Rasti, R.; Biglari, A.; Rezapourian, M.; Yang, Z.; Farsiu, S. RetiFluidNet: A self-adaptive and multi-attention deep convolutional network for retinal OCT fluid segmentation. *IEEE Trans. Med. Imaging* **2022**, *42*, 1413–1423. [CrossRef]
16. Rahil, M.; Anoop, B.; Girish, G.; Kothari, A.R.; Koolagudi, S.G.; Rajan, J. A deep ensemble learning-based CNN architecture for multiclass retinal fluid segmentation in oct images. *IEEE Access* **2023**, *11*, 17241–17251. [CrossRef]
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Yu, F. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
19. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
21. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. MLP-Mixer: An all-MLP Architecture for Vision. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2021; Volume 34, pp. 24261–24272.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**.
24. Huang, T.; Huang, L.; You, S.; Wang, F.; Qian, C.; Xu, C. Lightvit: Towards light-weight convolution-free vision transformers. *arXiv* **2022**, arXiv:2207.05557.
25. Azad, R.; Heidari, M.; Wu, Y.; Merhof, D. Contextual attention network: Transformer meets u-net. In *International Workshop on Machine Learning in Medical Imaging, Proceedings of the 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, 18 September 2022*; Springer: Cham, Switzerland, 2022; pp. 377–386.
26. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2021, Proceedings of the 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Springer: Cham, Switzerland, 2021; pp. 14–24.
27. Xie, Y.; Zhang, J.; Shen, C.; Xia, Y. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Proceedings of the 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Proceedings, Part III 24; Springer: Cham, Switzerland, 2021; pp. 171–180.

28. Yuan, F.; Zhang, Z.; Fang, Z. An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognit.* **2023**, *136*, 109228. [CrossRef]
29. Cao, G.; Wu, Y.; Peng, Z.; Zhou, Z.; Dai, C. Self-attention CNN for retinal layer segmentation in OCT. *Biomed. Opt. Express* **2024**, *15*, 1605–1617. [CrossRef]
30. Niu, Z.; Deng, Z.; Gao, W.; Bai, S.; Gong, Z.; Chen, C.; Rong, F.; Li, F.; Ma, L. FNeXter: A Multi-Scale Feature Fusion Network Based on ConvNeXt and Transformer for Retinal OCT Fluid Segmentation. *Sensors* **2024**, *24*, 2425. [CrossRef]
31. Han, Z.; Jian, M.; Wang, G.G. ConvUNeXt: An efficient convolution neural network for medical image segmentation. *Knowl.-Based Syst.* **2022**, *253*, 109512. [CrossRef]
32. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
33. Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; Shan, Y. UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio Video Point Cloud Time-Series and Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 5513–5524.
34. Chen, H.; Chu, X.; Ren, Y.; Zhao, X.; Huang, K. PeLK: Parameter-efficient Large Kernel ConvNets with Peripheral Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 5557–5567.
35. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; Lange, T.D.; Halvorsen, P.; Johansen, H.D. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019.
36. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
37. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
38. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2021, Proceedings of the 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Springer: Cham, Switzerland, 2021; pp. 36–46.
39. Ho, J.; Kalchbrenner, N.; Weissenborn, D.; Salimans, T. Axial attention in multidimensional transformers. *arXiv* **2019**, arXiv:1912.12180.
40. Tang, F.; Ding, J.; Quan, Q.; Wang, L.; Ning, C.; Zhou, S.K. CMUNEXT: An Efficient Medical Image Segmentation Network Based on Large Kernel and Skip Fusion. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 27–30 May 2024; pp. 1–5.
41. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
42. Yin, Y.; Han, Z.; Jian, M.; Wang, G.G.; Chen, L.; Wang, R. AMSUnet: A neural network using atrous multi-scale convolution for medical image segmentation. *Comput. Biol. Med.* **2023**, *162*, 107120. [CrossRef]
43. Xu, Q.; Ma, Z.; HE, N.; Duan, W. DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation. *Comput. Biol. Med.* **2023**, *154*, 106626. [CrossRef]
44. Rahman, M.M.; Munir, M.; Marculescu, R. EMCAD: Efficient Multi-Scale Convolutional Attention Decoding for Medical Image Segmentation. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 11769–11779.
45. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
46. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
47. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
48. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. A simple and light-weight attention module for convolutional neural networks. *Int. J. Comput. Vis.* **2020**, *128*, 783–798. [CrossRef]
49. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
51. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef]
52. Yu, T.; Li, X.; Cai, Y.; Sun, M.; Li, P. S2-mlp: Spatial-shift mlp architecture for vision. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 297–306.

53. Wu, B.; Wan, A.; Yue, X.; Jin, P.; Zhao, S.; Golmant, N.; Gholaminejad, A.; Gonzalez, J.; Keutzer, K. Shift: A zero flop, zero parameter alternative to spatial convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9127–9135.
54. Ahmed, Z.; Panhwar, S.Q.; Baqai, A.; Umrani, F.A.; Ahmed, M.; Khan, A. Deep learning based automated detection of intraretinal cystoid fluid. *Int. J. Imaging Syst. Technol.* **2021**, *32*, 902–917. [CrossRef]
55. Bogunovic, H.; Venhuizen, F.; Klimscha, S.; Apostolopoulos, S.; Bab-Hadiashar, A.; Bagci, U.; Beg, M.F.; Bekalo, L.; Chen, Q.; Ciller, C.; et al. RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge. *IEEE Trans. Med. Imaging* **2019**, *38*, 1858–1874. [CrossRef] [PubMed]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.