OphFusionNet: Uncertainty-Driven Multi-Scale Multimodal Feature Fusion Network for Ophthalmic Diseases Classification

Bo Wu, Weifang Zhu, Dehui Xiang, Xinjian Chen, Tao Peng, Chenwei Gui, Qing Peng, and Fei Shi

***Abstract*— Multimodal imaging has become an essential tool in clinical ophthalmology, offering complementary perspectives for disease diagnosis. However, current automated diagnostic approaches often fail to fully exploit the rich, complementary information provided by different imaging modalities. In this paper, to advance automated ophthalmic disease diagnosis through effective multimodal data integration, we propose the OphFusionNet, a novel multimodal learning framework based on uncertainty-driven multi-scale multimodal feature fusion. Drawing inspiration from clinical observations that ophthalmic lesions often appear at multiple spatial scales, we design a multi-scale feature fusion module with sparse self-attention (MSFF-SSA). This module captures hierarchical representations while suppressing redundancy, thereby enhancing both the expressiveness and efficiency of the extracted features. To further improve multimodal fusion, we introduce an uncertainty-aware multimodal fusion module with a game-theoretic selection strategy (UMF-GTSS). This component estimates the uncertainty associated with different features and adaptively weights them based on their relative reliability, yielding more robust and trustworthy diagnostic outcomes. To mitigate the tendency to over-rely on dominant modalities and underutilize the informative potential of subordinate ones, we propose a modality distillation strategy (MDS), which leverages multimodal features to guide and refine the learning of single-modal representations. This strategy enhances generalization and boosts the discriminative capacity of each individual modality. The OphFusionNet was evaluated on four publicly available ophthalmic datasets. Extensive experiments demonstrate that our approach achieves superior multimodal integration, resulting in state-of-the-art (SOTA) performance in multimodal ophthalmic disease diagnosis. The code will be available at: https://github.com/wb66715/OphFusionNet**

***Index Terms*— Multimodal learning, multimodal fusion, uncertainty estimation, ophthalmic disease diagnosis.**

I. Introduction

OPHTHALMIC diseases represent one of the leading causes of vision impairment and blindness worldwide [1]. With the global population aging, the incidence of such conditions is steadily rising. Common ocular disorders include age-related macular degeneration (AMD) [2], glaucoma [3], and diabetic macular edema (DME) [4]. These diseases significantly compromise patients' quality of life and place a substantial burden on public health systems and the global economy [5], [6]. Consequently, early diagnosis and accurate screening are essential to improving clinical outcomes and mitigating the societal and economic impacts of vision loss.

With the rapid advancements in deep learning, artificial intelligence (AI) has demonstrated substantial potential in the field of medical image analysis [7], [8]. AI-powered automated screening systems for ocular diseases can significantly improve diagnostic efficiency while alleviating the workload of ophthalmologists and healthcare professionals [9]. In particular, the integration of multimodal imaging modalities such as color fundus photography (CFP), optical coherence tomography (OCT), fluorescein fundus angiography (FFA), and indocyanine green angiography (ICGA) enables AI systems to leverage diverse and complementary data sources, facilitating more comprehensive and accurate diagnostic assessments.

Each ophthalmic imaging modality possesses distinct advantages and limitations. CFP is a critical tool for disease screening, offering a wide field of view with effective visualization of the morphology of retinal vessels and optic nerve head as well as many types of lesions [10]. However, it only shows the enface image of the retina, and is not sensitive to early and subtle pathological changes. In contrast, OCT offers high-resolution cross-sectional images that reveal the microscopic architecture of the retinal layers and

Received 7 November 2025; revised 18 January 2026 and 4 April 2026; accepted 15 April 2026. Date of publication 20 April 2026; date of current version 9 July 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62271337, Grant 62371326, and Grant 62371328; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231310; in part by China Postdoctoral Science Foundation under Grant 2023M742568; and in part by the Natural Science Foundation of Jiangsu Higher Education Institutions of China under Grant 24KJB510042. Recommended by Associate Editor H. Bogunović. *(Corresponding author: Fei Shi.)*

Bo Wu, Weifang Zhu, Dehui Xiang, Xinjian Chen, and Fei Shi are with the MIPAV Laboratory, School of Electronics and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: wubo_66715@163.com; wfzhu@suda.edu.cn; xiangdehui@suda.edu.cn; xjchen@suda.edu.cn; shifei@suda.edu.cn).

Tao Peng is with the School of Future Science and Engineering, Soochow University, Suzhou 215222, China (e-mail: tpeng@suda.edu.cn).

Chenwei Gui and Qing Peng are with the Department of Ophthalmology, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai 200072, China (e-mail: gcw31vv@163.com; pengqing@tongji.edu.cn).

Digital Object Identifier 10.1109/TMI.2026.3685304

1558-254X © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

lesions [11]. Nevertheless, cross-sectional OCT images do not directly provide information on retinal vasculature. Therefore, both as non-invasive structural imaging, CFP and OCT often complement each other. Additionally, FFA and ICGA are a complementary pair of invasive means for detailed assessment of the morphology and function of vessels. FFA excels in imaging retinal blood vessels, including vessel morphology, leakage, and hemorrhage, making it crucial for diagnosing retinal vascular diseases like diabetic retinopathy. In comparison, ICGA specializes in visualization of choroidal vessels, particularly in the context of choroidal neovascularization. This makes ICGA especially advantageous for diagnosing conditions such as AMD and various choroidal disorders. Although individual modalities can perform well in particular diagnostic scenarios, they often fall short in capturing the full complexity of ocular diseases, which are characterized by intricate structural and functional alterations. This limitation can result in reduced diagnostic accuracy.

Integrating multimodal imaging data for comprehensive analysis offers a promising solution to the limitations inherent in single-modality approaches, thereby enhancing both the accuracy and reliability of ophthalmic disease diagnosis. In recent years, deep learning-based methods leveraging multimodal data have gained increasing attention in medical image analysis [12], [13], [14]. However, current multimodal fusion technologies still face several challenges that impede their effectiveness and reliability in practical applications.

Firstly, existing methods struggle to comprehensively extract multi-scale features from complex lesion areas. Ophthalmic diseases are highly complex, with lesions that vary considerably in morphology, size, and spatial distribution. Current multi-scale feature extraction techniques typically combine features from different scales through direct concatenation or weighted fusion. While this enhances the model's expressive power, it often introduces substantial redundant information. For instance, features at different scales frequently contain repetitive low-level information (e.g., edges, textures). Moreover, in non-adaptive fusion methods, the contribution of features from different scales cannot be adjusted flexibly based on specific task requirements.

Secondly, existing methods fail to effectively and dynamically integrate multimodal features. Although multimodal images provide rich diagnostic information, the integration of features from diverse modalities poses substantial challenges. This is primarily due to the variability in relevance and interplay of these features across various clinical cases. Most approaches rely on simple concatenation or linear combinations, lacking the ability to dynamically adjust the feature fusion process. This limitation prevents the model from capturing the intricate interactions between features from different modalities, ultimately undermining the efficacy of the fused representations in delivering comprehensive diagnostic insights.

Lastly, modality imbalance is a pervasive issue in multimodal methods. For specific cases, certain modalities may contain more diagnostic information, while others may offer comparatively less. This imbalance can lead to models that are biased toward the dominant modality and overlook certain disease features provided by other modalities, which may result in wrong predictions. Consequently, this imbalance reduces the generalizability and robustness of the model.

To address the aforementioned challenges, we propose the OphFusionNet, which represents a novel uncertainty-based multimodal ophthalmic disease diagnostic framework. This method incorporates three key components: a multi-scale feature fusion module with sparse self-attention (MSFF-SSA), an uncertainty-aware multimodal fusion module with a game-theoretic strategy selection (UMF-GTSS), and a modality distillation strategy (MDS). Each of these modules provides an innovative solution to overcome the specific limitations of existing methods.

First, the MSFF-SSA module integrates the ReLU activation function with a self-attention mechanism to reduce feature redundancy and improve feature fusion within each modality. By introducing a gating mechanism, it further adjusts the contribution of multi-scale features, enhancing adaptability. This dynamic feature fusion process enables the model to focus on the most relevant multi-scale features, effectively eliminating redundant or noisy information.

Second, the UMF-GTSS module quantifies the uncertainty of features using payoff and utility functions derived from game theory. By dynamically adjusting the contribution of features from two fusion strategies based on their associated uncertainty, this module effectively captures the complex interactions between different modalities, thereby enhancing the overall fusion of multimodal features.

Finally, the MDS module transfers knowledge from the fused multimodal representations to guide the feature learning of individual modalities. By serving as a supervisory signal, the fused multimodal representation provides a regularizing influence that helps the model enhance the feature extraction process for weaker modalities. This ensures that no modality is neglected during training, thereby improving the overall performance and generalization capability of the model.

In summary, the three proposed modules work synergistically in effectively extracting multi-scale and balanced modality-specific features and in adaptively obtaining expressive multimodal features, which ultimately contribute to superior classification performance.

To validate the effectiveness of our proposed method, we conducted extensive experiments across four publicly available datasets, involving some dominant types of ophthalmic diseases, including AMD, glaucoma, DME, etc. The results from these experiments demonstrate the reliability and robustness of OphFusionNet in multimodal ophthalmic disease diagnosis.

In summary, the key contributions of this paper are as follows:

1) We propose a novel uncertainty-driven multimodal diagnostic approach for ophthalmic diseases, offering a new perspective on effective and reliable multimodal image classification in ophthalmology.
2) We design a multi-scale feature fusion module integrated with sparse self-attention, effectively reducing redundancy and enhancing multi-scale feature fusion.

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

3) We introduce an uncertainty-aware multimodal fusion module with a game-theoretic strategy selection, ensuring flexible feature fusion.
4) We propose a modality distillation strategy that boosts weaker modalities, enhancing the model's overall performance and generalization.

II. Related Work

A. Ophthalmic Disease Screening Based on Multimodal Imaging

In recent decades, deep learning-based methods for eye disease screening have advanced rapidly, particularly with the transition from single-modal to multimodal approaches [12], [14]. Multimodal image fusion techniques can be classified into early, middle, and late-stage fusion [13]. Currently, much of the research focuses on middle-stage fusion methods, which integrate information from different modalities at the intermediate layers of the network.

Yoo et al. [16] were the first to attempt multimodal image fusion for diagnosing AMD. They used a pre-trained VGG model to extract features from CFP and OCT images, concatenated these features, and diagnosed AMD using a random forest classifier. He et al. [17] utilized a multi-scale attention module to extract CFP features and a region-guided attention module to extract OCT features. These features were then concatenated, and a multimodal retinal image classification network was trained. Wang et al. [18] proposed a dual-stream convolutional neural network (CNN) that integrates CFP and OCT information using a spatially invariant fusion technique for AMD diagnosis. E et al. [19] introduced a two-stage knowledge-driven model for fine-grained wet AMD classification. In the first stage, ten types of lesions on CFP and OCT images were learned as prior knowledge, while in the second stage, fine-grained classification of wet AMD was performed using this prior knowledge.

However, the multimodal feature fusion methods mentioned above primarily rely on simple concatenation, which lacks adaptability and therefore does not fully leverage the complementary information across modalities. Liu et al. [20] addressed this limitation by introducing a cross-modal attention module, which adaptively fuses multimodal features. Wang et al. [21] proposed a geometry-correspondence-based multimodal learning network, which takes into account the geometric correspondence between OCT slices and CFP regions, enabling the network to learn correlated features from both modalities and achieve robust fusion. Despite these advancements, both approaches primarily focused on enhancing one modality with the features from the other and overlooked the need to model the collaboration and conflict between modalities effectively. Additionally, both methods ignored the exploit of multi-scale features from each modality.

In contrast, the method proposed in this study utilizes the UMF-GTSS module, which combines maximum fusion and multiplicative fusion strategies to facilitate deeper interaction between multimodal features. Additionally, we introduce a feature selection approach based on game theory, which allows for more effective modeling of collaboration and conflict between features. Our method also considers the diversity of lesion sizes in each modality and the varying discriminative power of different modalities in disease diagnosis.

B. The Application of Uncertainty Estimation in Ophthalmic Image Classification

Uncertainty estimation plays a key role in fields like machine learning, decision analysis, and risk assessment. Traditionally, methods based on classical probability theory, such as Bayesian estimation and Monte Carlo simulation, have been extensively used to address practical problems [22]. However, in recent years, scholars have advocated new uncertainty estimation approaches that go beyond classical probability theory. Various alternative theoretical frameworks have been proposed, including Dempster-Shafer theory [23], fuzzy logic [24], and probabilistic game theory [25]. These methods each have their own advantages and limitations [22]. In practice, the choice of estimation method should depend on the specific characteristics of the problem. Classical probability methods are typically more suitable when data is abundant and conditions are well-defined, while in complex decision-making environments, methods such as Dempster-Shafer theory and probabilistic game theory offer more flexible solutions.

In recent years, uncertainty estimation has gained significant attention in medical image classification tasks [26], aiming to improve the reliability and robustness of model predictions. Various approaches have also been proposed to quantify uncertainty in the context of ophthalmic disease diagnosis. For instance, Leibig et al. [27] utilized a dropout-based Bayesian uncertainty measurement technique for diagnosing diabetic retinopathy (DR) from fundus images, allowing the model to assess the variability in its predictions. Ayhan et al. [28] explored a test-time data augmentation method to quantify uncertainty in DR diagnosis, leveraging multiple augmented samples for more reliable predictions. Jaskari et al. [29] employed entropy-based uncertainty estimation, calculating the entropy of the posterior predictive distribution for DR classification to assess the confidence in predictions.

Most existing methods apply uncertainty to assist decision-making, with few effectively utilizing it for deep fusion of multimodal features. Zhou et al. [30] introduced an MC Dropout layer to generate varied prediction probabilities, calculated feature uncertainty using the mean prediction entropy, and used it to calculate weights for multimodal fusion. However, the calculation of uncertainty required multiple forward passes, which added to the computational complexity. Furthermore, Zou et al. [31] extended the deep evidence regression model to ophthalmic multimodal fusion tasks, using a Normal-Inverse Gamma (NIG) prior to capture modality uncertainty. They transformed the NIG prior into Student's t-distribution, proposing a confidence-aware fusion strategy based on a mixture of Student's t-distribution model. However, the use of the univariate NIG limited its application to binary classification tasks.

In the propose method, we quantify the uncertainty through game-theoretic strategy selection which is efficiently

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

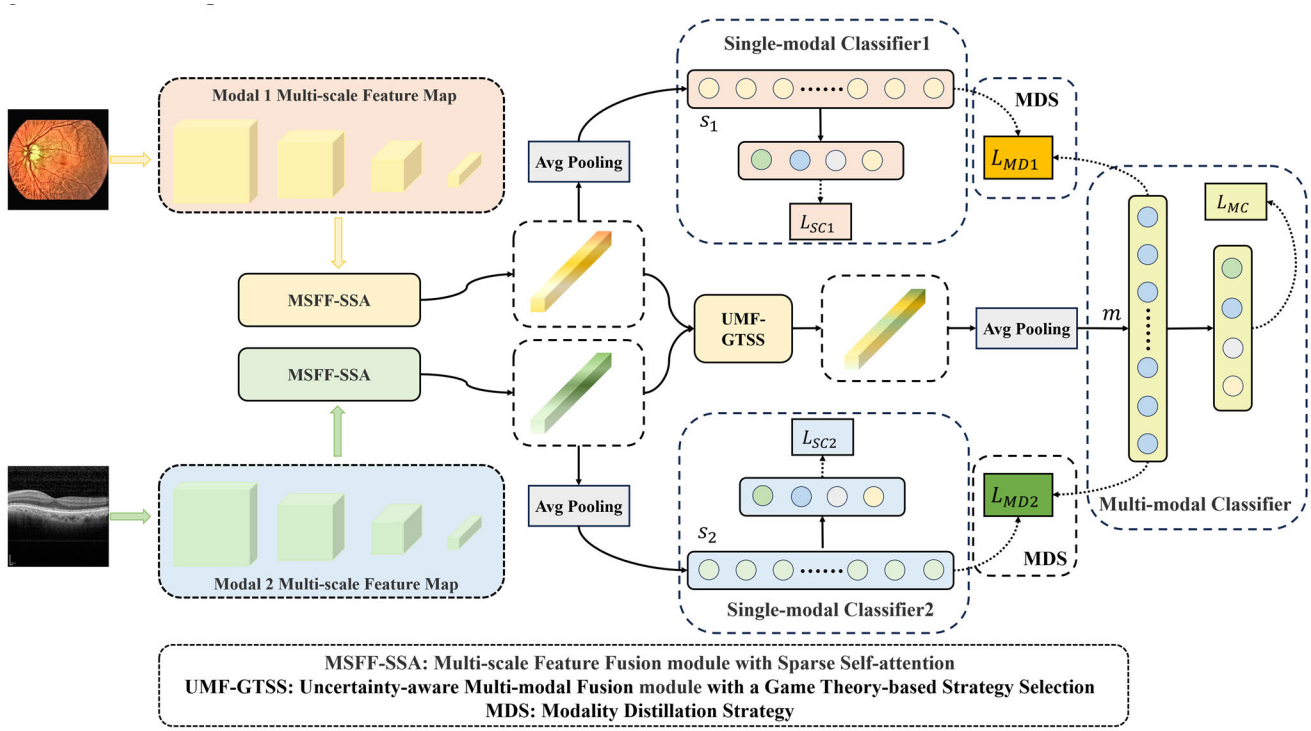


Fig. 1. The framework of uncertainty-driven multi-scale multimodal feature fusion network.

integrated into the feature fusion process, and utilize it for adaptive weighting of multiple fused features.

C. Cross-Modal Distillation

Knowledge distillation has garnered significant attention in multimodal learning, particularly in tasks that involve transferring knowledge from dominant modalities to less effective ones. For example, distillation strategies are frequently employed between RGB images and depth images to enhance the performance of each modality [32], [33]. Cross-modal distillation has also been applied to images and audio in tasks like speaker recognition [34]. In the domain of medical imaging, cross-modal distillation techniques have achieved remarkable success in tasks such as computed tomography (CT) and magnetic resonance image (MRI) segmentation. For instance, Jiang et al. [35] proposed a method in which a student network for CT image was guided by a teacher network trained on MRI data.

Based on these advancements, Song et al. [36] applied cross-modal distillation in ophthalmology, proposing a method to transfer knowledge from multimodal OCT and visual field networks to a single-modal OCT network, resulting in a significant improvement in glaucoma detection performance. Likewise, Wang et al. [37] decomposed disease attributes into refined concepts and conducted knowledge distillation for different disease features, leading to a marked increase in diagnostic accuracy for retinal diseases.

However, existing cross-modal distillation methods typically necessitated the training of multiple independent networks, which incurred high computational and storage costs. Inspired by self-distillation strategies [38], we propose an innovative modality distillation approach. This method integrates multimodal features to guide the training of single-modal features, thereby effectively enhancing the performance of weaker modalities. Importantly, it avoids excessive reliance on dominant modalities, promoting more efficient inter-modality synergy in multimodal learning.

III. The Proposed Method

A. Overall Framework

The overall framework of OphFusionNet is illustrated in Fig. 1. It consists of two parallel ResNet34 [15] networks that process the eye-pair multimodal images separately. Each network extracts multi-scale features from its respective modality. These features are then fused using the MSFF-SSA module to form single-modal multi-scale fused features.

Subsequently, the two single-modal features are imported to two single-modal classifiers respectively. The UMF-GTSS module combines the features from both modalities, and the fused multimodal features are imported to a multimodal classifier. Ultimately, this multimodal classifier generates the final classification results, leveraging information from both modalities to achieve accurate and robust diagnoses. Additionally, the MDS is employed to optimize the single-modal features by leveraging the guidance provided by the multimodal features.

B. Multi-Scale Feature Fusion Module With Sparse Self-Attention

This section presents a multi-scale feature fusion module with sparse self-attention (MSFF-SSA), illustrated in Fig. 2, designed to tackle the challenge of varying lesion sizes in ophthalmic images. Traditional multi-scale fusion methods often

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

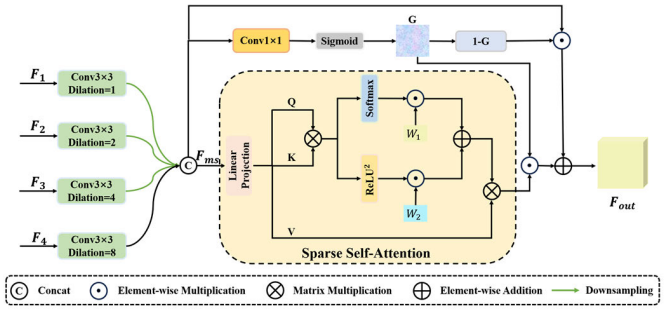


Fig. 2. The framework of multi-scale feature fusion module with sparse self-attention.

struggle with information redundancy, resulting in inefficient feature integration. MSFF-SSA overcomes this limitation by utilizing the sparse self-attention mechanism, which enables the efficient extraction of high-level semantic information across feature maps at different scales while simultaneously suppressing redundant information during the fusion process.

Initially, dilated convolutions are utilized to process features at multiple scales. Let the input feature maps be denoted as (F_1, F_2, F_3, F_4), corresponding to different stages of the feature extraction network. To enlarge the receptive field, dilated convolutions with varying dilation factors d_i (set to 1, 2, 4, and 8) are applied to extract multi-scale features as follows:

$$F_i^{\text{dilated}} = \text{Conv}\left(F_i, \text{kernel} = 3, \text{dilation} = d_i\right) \tag{1}$$

Next, to effectively integrate features from different scales, the feature maps from larger scales are downsampled to match the size of the smallest-scale feature map. Subsequently, the feature maps from different scales are fused through concatenation to form the fused multi-scale feature map F_{ms}.

To capture long-range dependencies between different regions of the feature map and minimize redundant information, a sparse self-attention mechanism is introduced. Initially, F_{ms} is projected into query (Q), key (K), and value (V) vectors:

$$Q, K, V = \text{LinearProjection}(F_{ms}) \tag{2}$$

The self-attention scores $(Attn)$ are then computed as:

$$Attn = \frac{QK^T}{\sqrt{d_k}} + B \tag{3}$$

where d_k denotes the dimension of each attention head, and B is the relative positional bias learned during training.

Traditional self-attention mechanisms consider all tokens within a feature map. This may introduce redundant or irrelevant features, resulting in inefficiencies in computation. In recent years, several studies [39], [40] have attempted to incorporate Top-K selection operations into self-attention mechanisms to focus on the most relevant features. However, the choice of parameter K is often highly sensitive to different tasks, potentially affecting the model's generalization ability. Meanwhile, other research [41] has proposed to replace the Softmax activation function in the self-attention mechanism with the ReLU activation function to increase the sparsity of the feature map. However, the use of ReLU alone can lead to excessively sparse feature maps, which may result in the loss of crucial information [42].

To address the aforementioned issues, we utilize the ReLU activation function in conjunction with a parallel Softmax activation function. This combination helps retain essential information while effectively preventing the feature map from becoming overly sparse. The Softmax function normalizes the attention scores, emphasizing higher values to enhance the representation of significant features while minimizing redundant ones. The ReLU function sets all negative attention scores to 0, retaining only positive values, which eliminates irrelevant or negatively correlated features, thereby increasing the sparsity of the attention map. To further emphasize the significance of relevant features, we perform a square operation on the ReLU output [43], which increases the impact of positive attention scores, further reducing the influence of the less important or redundant features. The results of these operations are then combined through weighted averaging to generate the final sparse attention map $Attn_{sparse}$:

$$Attn_{sparse} = \text{Softmax}(Attn) \cdot w_1 + \text{ReLU}^2(Attn) \cdot w_2 \tag{4}$$

where w_1 and w_2 are learned weight coefficients such that $w_1 + w_2 = 1$.

To enable flexible feature fusion, a gating mechanism is introduced. The fused multi-scale feature map F_{ms} is passed through a convolutional layer, followed by a sigmoid activation function to generate the gating signal G:

$$G = \text{Sigmoid}(\text{Conv}(F_{ms}, \text{kernel} = 1)) \tag{5}$$

Finally, the output F_{out} is a weighted combination of the fused features and the attention enhanced features:

$$F_{out} = G \cdot Attn_{sparse}V + (1 - G) \cdot F_{ms} \tag{6}$$

C. Uncertainty-Aware Multimodal Fusion Module With a Game-Theoretic Strategy Selection

In this section, we introduce an innovative multimodal feature fusion method: the uncertainty-aware multimodal fusion module with a game-theoretic strategy selection (UMF-GTSS), illustrated in Fig. 3. This method integrates features from two distinct input modalities while incorporating uncertainty quantification through game theory-based strategy selection. By combining payoff and utility functions, it seeks to account for the impact of features from different fusion modes and their associated uncertainties.

Initially, the input features of two modalities (x_1, x_2) are processed independently through convolutional layers, followed by batch normalization and ReLU activation to refine features for each modality:

$$x_i' = ReLU\left(BN\left(Conv\left(x_i\right)\right)\right), i = 1, 2 \tag{7}$$

Subsequently, we apply two fusion strategies: maximum fusion and multiplicative fusion, which result in fused features x_{max} and x_{mul}, respectively.

$$x_{max} = Max\left(x_1', x_2'\right) \tag{8}$$

$$x_{mul} = x_1' \odot x_2' \tag{9}$$

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

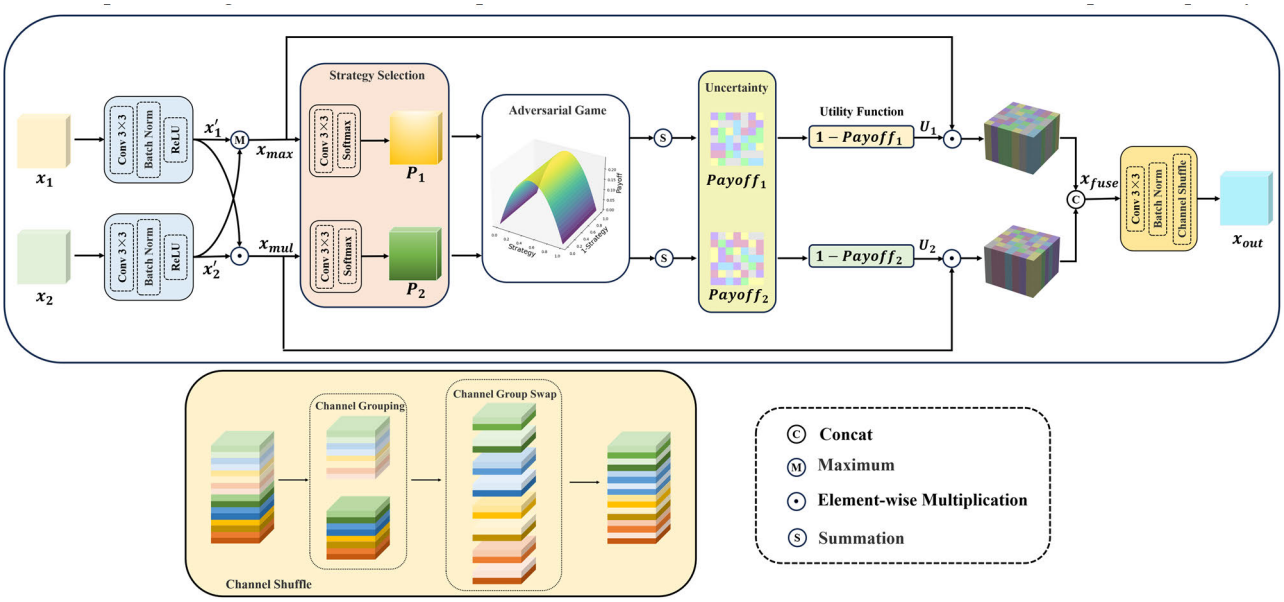


Fig. 3. The framework of uncertainty-aware multimodal fusion module with a game-theoretic strategy selection.

The maximum fusion operation highlights the most significant features, enabling the model to capture the most informative aspects of each modality. In contrast, the multiplicative fusion involves element-wise multiplication between the features x_1' and x_2', capturing the complementarity between the two modalities and enhances the fusion effect when the features are strongly aligned.

During the feature fusion process, we leverage the concept of mixed strategies from game theory [44], where different strategies are associated with probabilities, in order to introduce flexibility and diversity in the system's response to uncertainty. The outputs of the convolutional layers are processed through the Softmax function, which estimates the selection probabilities P_1 and P_2 for the two fused features. Specifically,

$$P_1 = \text{Softmax}\left(\text{Conv}\left(x_{\text{max}}\right)\right), P_2 = \text{Softmax}\left(\text{Conv}\left(x_{\text{mul}}\right)\right) \tag{10}$$

Here, P_i is associated with a mixed strategy that "focus on feature i" with probability P_i while "not focus on feature i" with probability $1 - P_i$. This dynamic adjustment of feature selection prevents over-reliance on any single one, thereby enhancing the system's adaptability.

Subsequently, we introduce the payoff function from game theory to assess the gain of different strategies, i.e., different selection probabilities. In game theory, the payoff function quantifies the reward or outcome a player receives based on their selected strategy. Considering that in reality, high rewards and risks are often associated with high uncertainty, we define the payoff function as follows:

$$Payoff_i = \sum_{c=1}^{C} P_i^c \cdot \left(1 - P_i^c\right), i = 1, 2 \tag{11}$$

where C represents the number of channels. This payoff function approaches zero as P_i nears 0 or 1, indicating a deterministic strategy selection. Conversely, when P_i assumes intermediate values, $Payoff_i$ increases, reflecting a strategy with higher uncertainty.

Further, we define the utility function, which represents the player's preference for different strategy combinations in the game. Given that our payoff function is structured to measure uncertainty, we design the strategic utility function U_i to measure the reliability and effectiveness for the two fused features, respectively:

$$U_i = 1 - Payoff_i, i = 1, 2 \tag{12}$$

Here, a higher value of U_i signifies greater stability and lower uncertainty, thereby enhancing the confidence in the feature selection.

Inspired by game theory's weighting mechanism, we perform weighted fusion by element-wise multiplying the utility function U_i with their respective fused features x_{max} and x_{mul}:

$$x_{fuse} = Concat\left(\left(x_{\text{max}} \cdot U_1\right), \left(x_{\text{mul}} \cdot U_2\right)\right) \tag{13}$$

Therefore, the utility functions guide the model during feature fusion towards more beneficial features for classification.

Finally, the channel shuffle operation is applied to rearrange the order of channels, facilitating the feature fusion across channels. This operation enhances the network's expressive power without increasing the number of parameters.

$$x_{out} = ChannelShuffle\left(\text{ReLU}\left(\text{BN}\left(\text{Conv}\left(x_{fuse}\right)\right)\right)\right) \tag{14}$$

Throughout multiple iterations of the training process, strategies are continuously refined and adjusted to converge towards the optimal solution. Similar to repeated gameplay, the model fine-tunes its parameters and adapts its approach based on feedback. In the Nash equilibrium of this iterative process, feature selection stabilizes, and no unilateral change in strategy can further improve the model's utility [44].

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

D. Modality Distillation Strategy

In multimodal learning, different modalities exhibit varying performance during the optimization process, posing significant challenges for effective learning from all modalities under a unified joint training objective [45]. To address this issue, we propose to use both single-modal and multimodal classifiers, while introduce a modality distillation strategy to apply constraints between them. This strategy aims to maintain the performance of the dominant modality while simultaneously enhancing the performance of the weaker modality, ultimately leading to improved multimodal feature representations.

It is reasonable to believe that the fused multimodal features possess superior discriminative power compared to single-modal features. By utilizing multimodal features to guide the learning of single-modal features, the quality of the single-modal representations can be substantially enhanced. To achieve this, we propose the modality distillation loss function which employs cosine similarity to measure the similarity between each single-modal feature vector and the multimodal feature vector, respectively.

$$\mathrm{cs}\left(a_1, a_2\right) = \frac{\sum_{i=1}^{n} a_{1_i} a_{2_i}}{\sqrt{\sum_{i=1}^{n} a_{1_i}^2}\sqrt{\sum_{i=1}^{n} a_{2_i}^2}} \tag{15}$$

$$\begin{aligned} L_{MD}\left(s_1, s_2, m\right) &= L_{MD1} + L_{MD2} \\ &= \alpha \mathbb{E}\left[\mathrm{cs}\left(s_1, m\right)^2\right] + \beta \mathbb{E}\left[\mathrm{cs}\left(s_1, m\right)^2\right] \end{aligned} \tag{16}$$

where s_1 and s_2 represent the multi-scale feature vectors input into the two single-modal classifiers, and m denotes the multimodal feature vector input into the multimodal classifier, and α and β are hyperparameters.

By employing this modality distillation strategy, we are able to enhance the feature representations of the weaker modality while preserving the information from the dominant modality. This facilitates the collaborative optimization of multimodal features and leads to a significant improvement in the overall performance of the model.

E. Loss Function

Considering the class imbalance in the datasets, in this study, we employ the focal loss [46] as the loss function for the two single-modal classifiers and the multimodal classifier, which are denoted as L_{SC1}, L_{SC2} and L_{MC}, respectively.

The total loss of OphFusionNet is composed of these three classification losses and the modality distillation loss L_{MD}. The total loss function can be expressed as:

$$L_{\mathrm{total}} = L_{\mathrm{SC1}} + L_{\mathrm{SC2}} + L_{\mathrm{MC}} + L_{\mathrm{MD}} \tag{17}$$

IV. Experimental Study

A. Datasets

We assessed the performance of our method using four publicly available datasets: MMC-AMD [18], GAMMA [47], APTOS-2021 [48], and APTOS-2023 [49]. These datasets contain ophthalmic multimodal images that are paired based on the same eye. Importantly, this study relies on the availability of the full set of paired multimodal data for each patient. The detailed data splits of the training, validation and test sets are provided in Table I, and the splits are on patient-level. This means that images from the same patient are allocated to only one subset.

TABLE I
Composition of Datasets

Dataset	Modality	Class	Train	Val.	Test
MMC-AMD	CFP OCT	Dry AMD	27	35	38
		Normal	156	20	20
		PCV	268	44	47
		Wet AMD	486	38	38
GAMMA	CFP OCT	Normal	50	50	51
		Early-stage	26	26	25
		Late-stage	24	24	24
APTOS-2021	IR OCT	Wet AMD	395	102	122
		PCV	280	60	77
		DME	351	78	101
APTOS-2023	FFA ICGA	CSCR	143	13	17
		Dry AMD	160	21	18
		MNV	472	64	62
		Normal	343	40	44

The MMC-AMD dataset includes images from two modalities: CFP and OCT. Specifically, CFP images were obtained using a Topcon fundus camera, while OCT images were captured using a Topcon OCT scanner or a Heidelberg OCT scanner. It is categorized into four classes: dry AMD, polypoidal choroidal vasculopathy (PCV), wet AMD, and normal. The dataset partition for training, validation, and test sets follows the official settings. Some images in the training set are not paired, and we only used the paired ones.

The GAMMA dataset is the first publicly available multimodal glaucoma image dataset, comprising 300 samples. These samples are split equally, with 100 used for training, 100 for validation, and 100 for testing. The dataset includes both CFP and OCT modalities and is divided into three categories: normal, early-stage glaucoma, and late-stage glaucoma. The CFP images were captured using the KOWA camera or the Topcon TRCNW400 camera, while the OCT images were obtained using the Topcon DRI OCT Triton scanner. During training, the three most central OCT slices were selected and paired with corresponding CFP images, forming three pairs per sample. For validation and testing, only the central OCT slice was paired with the CFP image for each sample.

The APTOS-2021 dataset was released during the 2021 Asia Pacific Ophthalmology Association Big Data Competition. For each eye, a set of images are provided, each containing a macula-centered enface infrared (IR) image with overlying radial or horizontal scanning lines, and an OCT B-scan corresponding to the single highlighted scanning line. All images in the dataset were acquired using a Heidelberg scanner. Following [20], we selected three categories, namely wet AMD, PCV, and DME, from the dataset and followed the official data partition. For each eye, a slice near the center of the OCT volume containing lesions was selected. The pair of IR and OCT images were cropped from the same original image. Additionally, the scanning lines on the IR images

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

TABLE II
PERFORMANCE COMPARISON OF MODELS ON MMC-AMD, APTOS-2021, AND APTOS-2023 DATASETS

Dataset	Method	Acc	F1	Sen	Spe	Pre	Rec	Auc	Kappa
MMC-AMD	MM_CNN†	80.4	83.5	83.1	93.2	85.9	83.1	/	82.2
	MM_CNN_da†	86.7	88.7	88.5	95.4	89.7	88.5	/	87.7
	MSAN	83.9	86.1	86.1	94.3	86.1	86.1	94.5	80.1
	CRD-Net	85.3	87.0	87.9	94.9	86.7	87.9	97.0	86.6
	Tformer	88.1	90.0	89.9	95.9	90.7	89.9	98.4	87.1
	TDF-Net	87.4	89.1	88.8	95.6	90.1	88.8	94.8	79.9
	OphFusionNet	**92.3**	**93.4**	**93.5**	**97.3**	**93.4**	**93.5**	**98.9**	**92.9**
APTOS-2021	MM_CNN	77.7	75.4	75.3	88.3	76.6	75.3	90.5	79.1
	MSAN	78.3	77.3	77.2	88.8	77.7	77.2	90.4	75.0
	CRD-Net	81.0	80.1	80.0	90.3	80.1	80.0	91.8	81.2
	Tformer	84.0	83.2	82.7	91.6	84.1	82.7	93.0	82.7
	TDF-Net	80.0	79.0	78.6	89.6	79.7	78.6	92.0	79.3
	OphFusionNet	**85.7**	**84.5**	**84.1**	**92.6**	**85.3**	**84.1**	**93.3**	**87.2**
APTOS-2023	MM_CNN	79.4	73.6	73.3	92.9	76.0	73.3	90.1	64.8
	MSAN	80.1	66.5	64.7	92.3	74.2	64.7	87.2	60.7
	CRD-Net	82.3	76.2	76.9	94.1	75.8	76.9	**93.3**	70.6
	Tformer	83.7	74.2	74.4	94.1	76.6	74.4	92.6	75.9
	TDF-Net	84.4	78.2	77.7	94.5	79.5	77.7	92.8	73.0
	OphFusionNet	**86.5**	**82.5**	**82.3**	**95.4**	**83.5**	**82.3**	92.2	**78.5**

† indicates published results from reference [18].

were preserved, as they occupied a small portion of the image while indicated the spatial correspondence between the two modalities, and therefore could potentially benefit the model learning. To ensure the integrity of the training dataset, a few images with excessive noise were excluded.

The APTOS-2023 dataset was released during the 2023 Asia Pacific Ophthalmology Association Big Data Competition and contains images from two modalities: FFA and ICGA. All images in the dataset were acquired using a Heidelberg fundus camera. We selected four classes, central serous chorioretinopathy (CSCR), dry AMD, macular neovascularization (MNV), and normal. Moreover, we curated the dataset by removing samples with multiple disease labels. Subsequently, the dataset was divided into training, validation, and test sets with a ratio of 8:1:1.

TABLE III
PERFORMANCE COMPARISON OF MODELS ON GAMMA DATASET

Method	Acc	Kappa
SmartDSP†	81.0	85.5
VoxelCloud†	82.0	85.0
EyeStar†	82.0	84.8
HZL†	83.0	84.0
IBME†	78.0	82.6
CRD-Net	80.0	83.3
GeCoMNet†	/	86.0
GeCoMNet (Ensemble)†	/	88.4
OphFusionNet	**87.0**	**90.5**

† indicates published results from corresponding references.

B. Experimental Details

We performed model training using the PyTorch framework on an RTX 4060Ti GPU equipped with 16GB of memory. The input image dimensions were standardized to 224×224 pixels, with a batch size of 8. The training process utilized the Adam [50] optimizer, initialized with a learning rate of 1×10^{-4}, which was progressively decayed using a factor of 1×10^{-2}. The total number of epochs was set to 150, and the cosine annealing was applied to adjust the learning rate throughout the training process.

To augment the diversity of the input data, several widely used data augmentation techniques were employed. These included horizontal flipping, rotation, color adjustments, and contrast enhancement.

For the evaluation of classification performance, we adopted a comprehensive set of standard metrics, including accuracy, F1 score, precision, sensitivity, specificity, recall, area under the curve (AUC), and the weighted quadratic Kappa coefficient. These metrics offer a multidimensional assessment of the model's performance, facilitating a thorough evaluation of its effectiveness in classification tasks.

C. Comparison With State-of-the-Art Methods

To assess the effectiveness of our proposed method, we conducted comparative analyses with state-of-the-art techniques across four publicly available datasets. For the MMC-AMD, APTOS-2021, and APTOS-2023 datasets, we compared our approach with five leading multimodal classification methods, as described below:

1) MM-CNN [18]: A dual-stream CNN that integrates information from the CFP and OCT streams using spatially invariant fusion techniques.

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
PERFORMANCE COMPARISON OF MULTI-SCALE FEATURE FUSION METHODS

Dataset	Method	Acc	F1	Sen	Spe	Pre	Rec	Auc	Kappa
MMC-AMD	GPG	88.1	90.0	90.5	96.0	92.3	90.5	98.4	87.6
	HFF	84.6	87.0	87.0	94.7	89.1	87.0	94.7	77.8
	MSCA	86.0	88.3	88.7	95.2	90.2	88.7	97.4	88.9
	MSC	87.4	89.3	88.8	95.4	90.3	88.8	96.7	92.4
	MSFF-SSA	**92.3**	**93.4**	**93.5**	**97.3**	**93.4**	**93.5**	**98.9**	**92.9**
GAMMA	GPG	84.0	79.5	79.6	92.7	79.9	79.6	91.3	87.7
	HFF	82.0	76.6	75.6	90.6	78.2	75.6	91.4	86.4
	MSCA	84.0	79.9	80.4	92.9	79.5	80.4	92.5	86.1
	MSC	83.0	77.3	77.1	91.3	78.8	77.1	**93.1**	87.9
	MSFF-SSA	**87.0**	**83.0**	**83.1**	**94.0**	**83.0**	**83.1**	91.9	**90.5**
APTOS-2021	GPG	83.7	82.8	82.5	91.5	83.3	82.5	92.8	83.1
	HFF	84.0	83.6	83.5	91.8	83.8	83.5	93.6	83.6
	MSCA	85.0	84.2	83.5	92.1	**85.5**	83.5	**94.3**	84.7
	MSC	83.0	82.2	81.7	91.1	83.0	81.7	93.7	80.8
	MSFF-SSA	**85.7**	**84.5**	**84.1**	**92.6**	85.3	**84.1**	93.3	**87.2**
APTOS-2023	GPG	85.8	80.9	79.9	95.0	82.9	79.9	**93.7**	75.0
	HFF	83.7	78.0	76.9	94.1	80.1	76.9	91.9	70.2
	MSCA	83.7	77.2	77.1	94.3	77.5	77.1	92.0	74.3
	MSC	85.1	80.3	79.6	94.6	81.2	79.6	92.0	73.6
	MSFF-SSA	**86.5**	**82.5**	**82.3**	**95.4**	**83.5**	**82.3**	92.2	**78.5**

2) MSAN [17]: A modality-specific attention network for multimodal retinal image classification. This method utilizes multi-scale attention modules to extract both local and global features from fundus images and incorporates a region-guided attention module to encode features from different retinal layers.
3) CRD-Net [20]: A retinal disease classification model that employs a cross-modal attention module to adaptively focus on lesion-relevant features across different modalities, while fusing modality-specific information through various loss functions.
4) Tformer [51]: A model with a Transformer-based feature extraction backbone that designs a series of dual-branch, hierarchical multimodal transformers, progressively integrating image information from different modalities, originally proposed for skin lesion diagnosis.
5) TDF-Net [52]: A trustworthy dynamic feature fusion network that captures modality-specific information through a dual-branch feature extraction module, originally proposed for breast cancer diagnosis using multimodal ultrasound images. It employs a Dirichlet distribution to quantify the contribution of each modality to the diagnostic result, thereby enabling dynamic fusion of multimodal features.

Table II presents a comparative performance analysis of different methods across three datasets: MMC-AMD, APTOS-2021, and APTOS-2023. OphFusionNet outperforms all other methods across all three datasets, particularly excelling in key metrics such as accuracy, F1 score, AUC, and the Kappa coefficient, thereby demonstrating its exceptional classification capability and stability. Specifically, OphFusionNet achieves the highest accuracy, with 92.3% on MMC-AMD, 85.7% on APTOS-2021, and 86.5% on APTOS-2023. Moreover, OphFusionNet also leads in F1 score, with values of 93.4%, 84.5%, and 82.5% for the respective datasets, along with high Kappa coefficients of 92.9%, 87.2%, and 78.5%. In contrast, while CRD-Net shows a slight advantage over OphFusionNet in terms of AUC (93.3% vs. 92.2%) on the APTOS-2023 dataset, its overall performance does not exceed that of OphFusionNet. Additionally, Tformer demonstrates strong performance on specific datasets, particularly APTOS-2021 and APTOS-2023, with accuracy values of 84.0% and 83.7%, respectively. However, its overall performance still lags behind OphFusionNet.

For the GAMMA dataset, we compared our method with the top five methods from the GAMMA challenge [47], as well as two advanced methods developed specifically for the dataset: CRD-Net [20] and GeCoMNet [21]. GeCoMNet improves the robustness of feature fusion by incorporating the geometric correspondence between OCT slices and CFP regions, thereby allowing the model to effectively learn complementary features from both modalities. GeCoMNet (Ensemble) uses different random seeds to train multiple model instances. It then selects the three models with the highest AUC scores from the individual training runs. The final predictions are generated by averaging the predictions from these three selected models.

As shown in Table III, OphFusionNet achieves exceptional performance on the GAMMA dataset, attaining an accuracy of 87.0%, significantly surpassing that of the other comparison methods. When compared to the top five methods from the GAMMA challenge, OphFusionNet leads in accuracy, with a particularly remarkable Kappa value of 90.5%. Although the ensemble approach of GeCoMNet closely approaches OphFusionNet in terms of the Kappa value, OphFusionNet still demonstrates superior overall performance.

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

TABLE V
PERFORMANCE COMPARISON OF MULTIMODAL FEATURE FUSION METHODS

Dataset	Method	Acc	F1	Sen	Spe	Pre	Rec	Auc	Kappa
MMC-AMD	AF	86.0	88.0	87.8	95.1	88.3	87.8	96.5	83.9
	MF	84.6	86.7	86.4	94.5	87.1	86.4	96.2	84.3
	CMA	85.3	87.5	87.6	95.0	90.6	87.6	97.3	75.9
	TDF	86.0	87.8	87.3	95.0	89.0	87.3	95.8	79.3
	CFAM	87.4	89.4	89.5	95.6	90.2	89.5	96.4	89.1
	UMF-GTSS	**92.3**	**93.4**	**93.5**	**97.3**	**93.4**	**93.5**	**98.9**	**92.9**
GAMMA	AF	81.0	74.2	74.4	90.2	76.2	74.4	91.3	86.7
	MF	82.0	77.4	77.0	90.9	78.0	77.0	92.1	84.7
	CMA	83.0	78.0	77.6	91.8	78.6	77.6	91.6	87.1
	TDF	80.0	74.2	75.1	91.2	75.3	75.1	89.4	82.4
	CFAM	83.0	78.3	77.6	91.3	79.2	77.6	**93.4**	83.1
	UMF-GTSS	**87.0**	**83.0**	**83.1**	**94.0**	**83.0**	**83.1**	91.9	**90.5**
APTOS-2021	AF	83.7	82.4	82.3	91.7	82.7	82.3	92.9	86.4
	MF	83.3	81.9	81.3	91.2	83.5	81.3	91.4	84.5
	CMA	84.0	83.7	84.0	92.0	83.6	84.0	92.2	84.8
	TDF	84.3	82.9	82.4	91.8	84.4	82.4	93.1	85.8
	CFAM	84.3	83.6	83.4	92.0	83.9	83.4	92.9	84.8
	UMF-GTSS	**85.7**	**84.5**	**84.1**	**92.6**	**85.3**	**84.1**	**93.3**	**87.2**
APTOS-2023	AF	83.0	75.8	74.0	93.9	78.4	74.0	92.7	69.4
	MF	81.6	74.7	73.9	93.5	76.1	73.9	**93.6**	65.5
	CMA	84.4	78.2	77.0	94.6	80.8	77.0	93.1	74.1
	TDF	84.4	78.5	77.4	94.5	80.1	77.4	91.7	74.9
	CFAM	84.4	78.8	78.2	94.5	79.8	78.2	**93.6**	71.7
	UMF-GTSS	**86.5**	**82.5**	**82.3**	**95.4**	**83.5**	**82.3**	92.2	**78.5**

D. Performance Comparison of Multi-Scale Feature Fusion Methods

In this section, we compare the proposed MSFF-SSA module with some existing multi-scale feature fusion modules. The global pyramid guidance (GPG) module [53] employs convolutions with varying dilation rates to extract global contextual information from feature maps at different levels. The hierarchical feature fusion (HFF) module [54] integrates high-level features and refines the fused features using a channel attention mechanism. The multi-scale cross-axis attention (MSCA) module [55] embeds multiscale local features into an axial attention network to effectively address the issue of local information loss and enhance the modeling of global dependencies. The multi-scale sparse cross-attention (MSC) module [40] cross-fuses high and low-level features to efficiently extract key information while using a Top-K sparse computation method to reduce the interference of irrelevant information.

We replaced the proposed MSFF-SSA module with these modules respectively, and the results on the four datasets are listed in Table IV. The experimental results demonstrate that MSFF-SSA outperforms other methods on most of the evaluation metrics, particularly excelling in accuracy, sensitivity, and the Kappa coefficient. For instance, on the MMC-AMD dataset, MSFF-SSA achieved an accuracy of 92.3% and a Kappa coefficient of 92.9%, far surpassing methods such as MSCA and HFF. These results indicate the superiority of MSFF-SSA in effectively enhance the extraction and integration of multi-level information.

E. Performance Comparison of Multimodal Feature Fusion Methods

In this section, we further compare the proposed UMF-GTSS module with some existing multimodal fusion methods, including additive fusion (AF), multiplicative fusion (MF), cross-attention-based fusion (CAM) [20], uncertainty-driven dynamic fusion (TDF) [52] and coordinate-attention-based fusion (CFAM) [56]. The results obtained by replacing UMF-GTSS with these modules are listed in Table V.

When compared to the aforementioned multimodal feature fusion methods, our proposed UMF-GTSS method demonstrates clear advantages on most performance metrics. Specifically, the UMF-GTSS outperforms other methods on the MMC-AMD dataset, achieving superior results in accuracy (92.3%), F1 score (93.4%), and sensitivity (93.5%). The method also exhibits exceptional performance on the GAMMA, APTOS-2021, and APTOS-2023 datasets, particularly excelling in accuracy and Kappa values. These results further validate the effectiveness of UMF-GTSS in handling complex multimodal medical image data.

F. Ablation Tests

In this study, we conducted ablation experiments across four datasets to evaluate the impact of MSFF-SSA, UMF-GTSS, and MDS on model performance. The results are summarized in Table VI. We began with ResNet34 as the feature extractor, concatenating the deepest features and passing them into the classifier to obtain the final classification results. Subsequently, we progressively integrated MSFF-SSA, UMF-GTSS, and

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
ABLATION STUDY RESULTS ON MMC-AMD, GAMMA, APTOS-2021, AND APTOS-2023 DATASETS

Dataset	MSFF-SSA	UMF-GTSS	MDS	Acc	F1	Sen	Spe	Pre	Rec	Auc	Kappa
MMC-AMD				80.4	83.3	83.6	93.3	85.1	83.6	96.0	72.9
	√			86.0	88.3	88.5	95.2	90.8	88.5	97.7	82.6
		√		86.0	88.3	88.5	95.1	89.0	88.5	97.7	90.7
	√	√		89.5	91.0	90.8	96.3	91.5	90.8	97.4	86.4
	√	√	√	**92.3**	**93.4**	**93.5**	**97.3**	**93.4**	**93.5**	**98.9**	**92.9**
GAMMA				73.0	64.8	64.7	86.6	65.2	64.7	88.2	73.8
	√			79.0	72.1	72.2	90.5	72.1	72.2	89.1	84.7
		√		78.0	71.1	71.6	90.1	71.6	71.6	90.5	82.7
	√	√		84.0	78.4	78.3	92.5	79.0	78.3	90.0	88.5
	√	√	√	**87.0**	**83.0**	**83.1**	**94.0**	**83.0**	**83.1**	**91.9**	**90.5**
APTOS-2021				77.7	77.0	77.9	88.9	76.6	77.9	90.8	76.5
	√			82.3	81.8	82.3	91.2	81.5	82.3	92.9	81.8
		√		80.7	79.4	79.3	90.1	79.6	79.3	91.2	83.1
	√	√		84.7	83.7	83.4	92.1	84.1	83.4	93.1	85.2
	√	√	√	**85.7**	**84.5**	**84.1**	**92.6**	**85.3**	**84.1**	**93.3**	**87.2**
APTOS-2023				77.3	67.6	68.0	92.5	67.3	68.0	91.6	59.8
	√			80.9	71.3	70.1	93.0	78.0	70.1	**92.2**	61.8
		√		81.6	74.3	73.5	93.4	75.4	73.5	90.5	63.3
	√	√		85.8	78.1	77.4	94.9	80.5	77.4	91.2	75.2
	√	√	√	**86.5**	**82.5**	**82.3**	**95.4**	**83.5**	**82.3**	**92.2**	**78.5**

TABLE VII
EVALUATION OF ACTIVATION FUNCTION COMBINATIONS IN MSFF-SSA ACROSS MMC-AMD, GAMMA, APTOS-2021, AND APTOS-2023 DATASETS

Dataset	Softmax	Relu^2	Acc	F1	Sen	Spe	Pre	Rec	Auc	Kappa
MMC-AMD	√		85.3	87.6	88.6	95.0	91.1	88.6	98.3	90.9
		√	84.6	87.1	87.3	94.7	88.8	87.3	97.5	85.3
	√	√	**92.3**	**93.4**	**93.5**	**97.3**	**93.4**	**93.5**	**98.9**	**92.9**
GAMMA	√		83.0	77.3	76.9	91.8	77.8	76.9	91.5	87.6
		√	81.0	74.7	74.2	90.6	75.6	74.2	90.6	85.7
	√	√	**87.0**	**83.0**	**83.1**	**94.0**	**83.0**	**83.1**	**91.9**	**90.5**
APTOS-2021	√		84.0	83.5	83.7	92.0	83.3	83.7	93.2	84.9
		√	84.3	83.3	82.8	91.8	84.4	82.8	92.8	83.8
	√	√	**85.7**	**84.5**	**84.1**	**92.6**	**85.3**	**84.1**	**93.3**	**87.2**
APTOS-2023	√		83.7	77.9	77.6	94.3	78.4	77.6	90.0	71.2
		√	85.1	80.1	80.6	95.2	80.8	80.6	**94.2**	78.1
	√	√	**86.5**	**82.5**	**82.3**	**95.4**	**83.5**	**82.3**	92.2	**78.5**

MDS, resulting in a significant improvement in the model's performance across various metrics. The results demonstrate that our methods exhibit strong adaptability across different datasets and show robust generalization capability.

To evaluate the contributions of individual components within MSFF-SSA, an ablation study was conducted to assess the impact of Softmax and ReLU^2 activations on model performance across four datasets. The results, summarized in Table VII, demonstrate that the integration of Softmax and ReLU^2 consistently outperformed their individual applications on all four datasets. These findings confirm the effectiveness of the combined activation functions, highlighting their complementary roles in enhancing the overall performance of the proposed framework.

To evaluate the effectiveness of UMF-GTSS in multimodal feature fusion, we compared the classification performance of the two single-modal classifiers and the multimodal classifier across four datasets. The experimental results, presented in Table VIII, indicate that fusing multimodal features significantly improved classification performance on all four datasets, thereby validating the effectiveness and reliability of UMF-GTSS across different datasets.

To evaluate the effectiveness of single-modal classifiers within the OphFusionNet, we conducted a comprehensive analysis comparing the performance of the model with and without these classifiers across various datasets. As detailed in Table IX, the experimental results consistently indicate performance improvements, highlighting the critical importance

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
EVALUATION OF UMF-GTSS MODULE IN MULTIMODAL FUSION ACROSS MMC-AMD, GAMMA, APTOS-2021, AND APTOS-2023 DATASETS

Dataset	Modality	Acc	F1	Sen	Spe	Pre	Rec	Auc	Kappa
MMC-AMD	OCT	81.8	84.4	83.9	93.6	85.8	83.9	96.5	75.3
	CFP	74.8	77.3	78.8	91.4	79.5	78.8	91.9	60.1
	Multimodal	**92.3**	**93.4**	**93.5**	**97.3**	**93.4**	**93.5**	**98.9**	**92.9**
GAMMA	OCT	74.0	68.4	68.1	86.8	70.2	68.1	87.9	72.1
	CFP	76.0	69.0	68.7	88.2	69.4	68.7	87.6	73.4
	Multimodal	**87.0**	**83.0**	**83.1**	**94.0**	**83.0**	**83.1**	**91.9**	**90.5**
APTOS-2021	OCT	78.3	76.9	76.7	88.8	77.5	76.7	89.7	77.7
	IR	79.0	77.5	77.7	89.3	77.9	77.7	89.4	77.0
	Multimodal	**85.7**	**84.5**	**84.1**	**92.6**	**85.3**	**84.1**	**93.3**	**87.2**
APTOS-2023	FFA	75.2	66.4	64.9	91.0	69.3	64.9	88.6	56.4
	ICGA	80.1	72.6	71.4	92.9	74.5	71.4	90.1	53.7
	Multimodal	**86.5**	**82.5**	**82.3**	**95.4**	**83.5**	**82.3**	**92.2**	**78.5**

TABLE IX
EVALUATION OF SINGLE-MODAL CLASSIFIERS ACROSS MMC-AMD, GAMMA, APTOS-2021, AND APTOS-2023 DATASETS

Dataset	Method	Acc	F1	Sen	Spe	Pre	Rec	Auc	Kappa
MMC-AMD	w/o single-modal classifiers	90.2	91.6	91.5	96.6	91.9	91.5	98.0	89.8
	w single-modal classifiers	**92.3**	**93.4**	**93.5**	**97.3**	**93.4**	**93.5**	**98.9**	**92.9**
GAMMA	w/o single-modal classifiers	83.0	77.9	77.6	91.8	78.4	77.6	90.5	85.0
	w single-modal classifiers	**87.0**	**83.0**	**83.1**	**94.0**	**83.0**	**83.1**	**91.9**	**90.5**
APTOS-2021	w/o single-modal classifiers	84.7	83.6	83.5	92.1	84.0	83.5	92.9	85.2
	w single-modal classifiers	**85.7**	**84.5**	**84.1**	**92.6**	**85.3**	**84.1**	**93.3**	**87.2**
APTOS-2023	w/o single-modal classifiers	84.4	78.1	77.1	94.5	80.3	77.1	**92.5**	68.3
	w single-modal classifiers	**86.5**	**82.5**	**82.3**	**95.4**	**83.5**	**82.3**	92.2	**78.5**

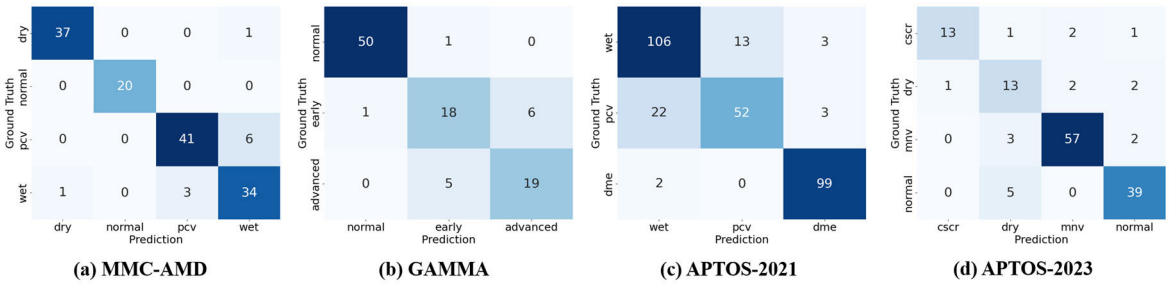


Fig. 4. Confusion matrices of OphFusionNet across MMC-AMD, GAMMA, APTOS-2021, and APTOS-2023 datasets.

of integrating single-modal classifiers into the OphFusionNet. The primary rationale for these performance enhancements can be attributed to the auxiliary supervision provided by the single-modal classifiers, which significantly aids in the accurate extraction of features from each modality.

G. Complexity Analysis

In the context of real-world clinical deployment, the feasibility of deep learning models largely depends on their inference speed and computational resource requirements. Table X compares our method with several SOTA methods in terms of frame rate, parameter count, and the total floating point operations (FLOPs).

OphFusionNet has 83.42 million parameters and a computational complexity of 11.92 GFlops, and achieves a frame rate of 85.01 FPS. The relatively high computational demands primarily stem from the use of high-dimensional feature maps. Although the OphFusionNet does not compare favorably with

TABLE X
PERFORMANCE COMPARISON OF RUNTIME AND MODEL SIZE

Method	Frame Rate (fps)	Params (M)	FLOPs (G)
MM_CNN	**189.27**	**22.88**	**3.65**
MSAN	112.87	23.76	3.84
CRD-Net	109.26	82.71	25.71
Tformer	39.45	77.62	13.03
TDF-Net	32.69	43.87	154.38
OphFusionNet	85.01	83.42	11.92

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

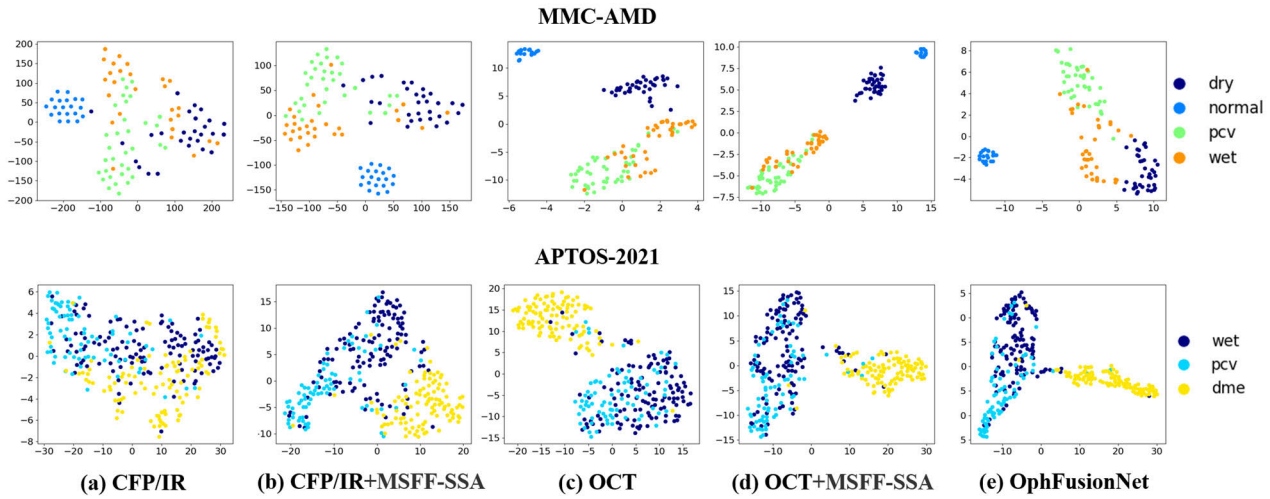


Fig. 5. T-SNE visualization on MMC-AMD and APTOS-2021 datasets.

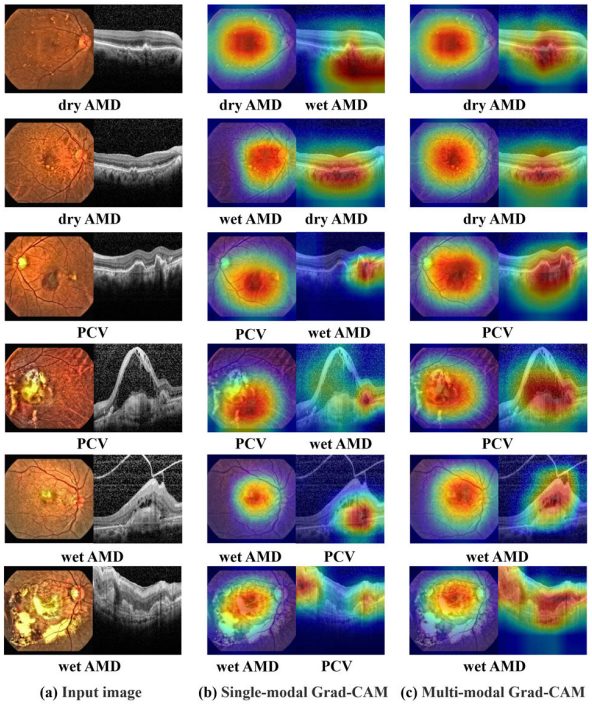


Fig. 6. Grad-CAM visualizations on MMC-AMD dataset. The images in (a) are annotated with the ground truth labels from the dataset. (b) and (c) show the Grad-CAM visualizations with the predicted labels from the model.

some existing methods, it is easy to be deployed on a computer with moderate GPU configuration, and is able to achieve real-time decision-making in clinical environments. Furthermore, the performance gain of OphFusionNet justifies the trade-off between computational cost and accuracy.

H. Visualization and Discussion

To visually demonstrate the performance of OphFusionNet across four datasets, we present the confusion matrices in Fig. 4. As illustrated in the figure, our method excels at classifying minority class samples, particularly in the CSCR and dry AMD classification tasks within the APTOS-2023 dataset, where it effectively distinguishes between these categories. Furthermore, for more challenging categories, such as wet AMD and PCV, OphFusionNet achieves favorable classification results on both the MMC-AMD and APTOS-2021 datasets. Additionally, the model demonstrates robust classification performance for both early and advanced stages of glaucoma. These results demonstrate the impressive performance and robustness of OphFusionNet in multimodal ophthalmic disease diagnosis.

Additionally, we utilized t-SNE [57] for feature visualization to showcase the features extracted from different layers of the network. The incorporation of the MSFF-SSA module results in better differentiation between wet AMD and PCV in the CFP modality of the MMC-AMD dataset, as shown in Fig. 5(a) and Fig. 5(b). It also enhances the distinction between DME and AMD in the IR modality of the APTOS-2021 dataset, as demonstrated in Fig. 5(a) and Fig. 5(b). Moreover, in both the MMC-AMD and APTOS-2021 datasets, the model's ability to distinguish the more challenging cases of wet AMD and PCV is significantly improved, as illustrated in Fig. 5(c) and Fig. 5(d). Furthermore, the inclusion of the UMF-GTSS module further increases the distance between features from different categories, as depicted in Fig. 5(e), thereby improving the model's classification performance.

We employed gradient-weighted class activation mapping (Grad-CAM) [58] to visualize the regions of focus that the model attends to when making predictions on input images. Several sample pairs containing lesion regions from the MMC-AMD test set were selected for this analysis. The visualizations were generated for the OphFusionNet, as well as for two single-modal models, each trained using ResNet34 with CFP and OCT images, respectively. As shown in Fig. 6, the single-modal models either insufficiently focus on the lesion regions or incorrectly focus on non-lesion areas, resulting in poor prediction performance. In contrast, the multimodal

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

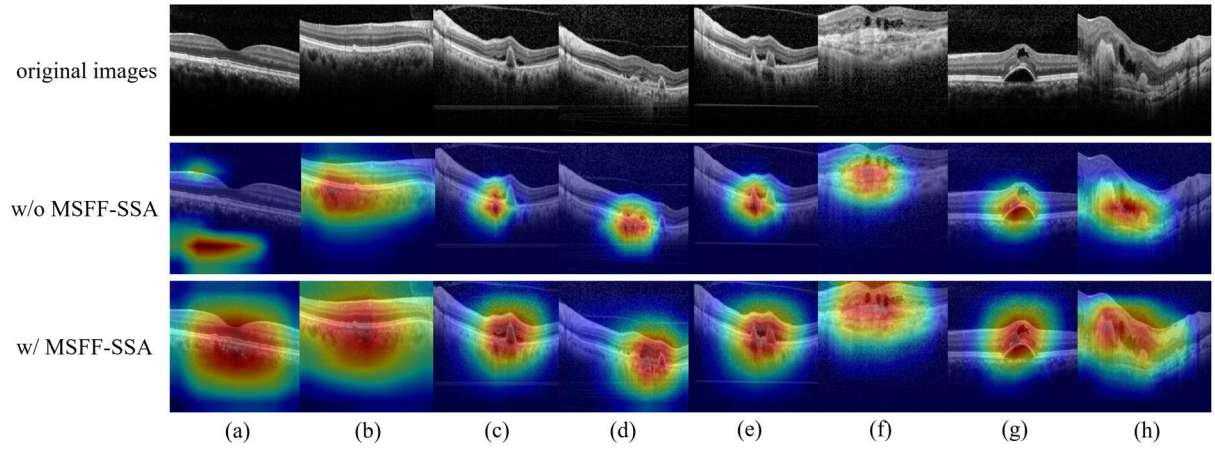


Fig. 7. Grad-CAM visualization results evaluating the effectiveness of MSFF-SSA on the MMC-AMD dataset.

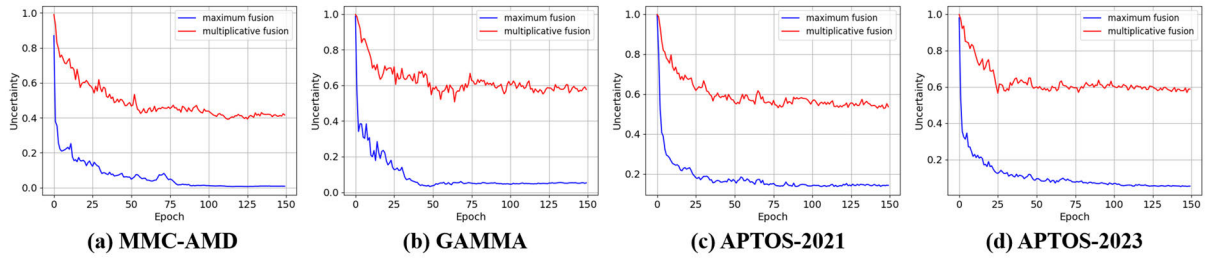


Fig. 8. Evolution of uncertainty during training.

model effectively focuses on lesion regions at different scales, allowing it to capture both large and small pathological features. For example, in diagnosing dry AMD, the multimodal model accurately identifies drusen, a key feature visible at a small scale. For PCV, it detects various features, including pigment epithelial detachment, digitiform protrusions, and retinal cystoid edema, which vary in size from large to small lesions. Similarly, when diagnosing wet AMD, the model highlights both small-scale features, such as subretinal exudates, and larger-scale features, such as retinal edema.

To validate the effectiveness of the proposed MSFF-SSA module, we generated the Grad-CAM maps with and without the inclusion of this module, analyzing the model's attention to the lesions. We chose the OCT images from the MMC-AMD dataset for the Grad-CAM comparisons, as this dataset specifically focuses on AMD, and OCT images enable the direct visualization of multi-scale lesions typical of AMD. As shown in Fig. 7(a) and (b), without the MSFF-SSA module, the model failed to effectively detect small lesions such as drusens. What's more, Fig. 7(c)-(e) illustrate that when the module is absent, the model shows insufficient attention to some middle-sized lesions. Fig. 7(f)-(h) further demonstrate that, although the model can focus on larger lesions like choroidal neovascularization and retinal pigment epithelium detachment, some smaller lesions, such as macular cystic edema and subretinal fluid, are missed. In comparison, after incorporating the MSFF-SSA module, the Grad-CAM maps exhibit broader and more detailed attention areas, indicating that the model effectively focuses on lesions of various sizes across different scales. This suggests that the MSFF-SSA module not only significantly enhances the model's ability to recognize prominent lesions but also plays a crucial role in detecting small abnormalities, greatly improving the diagnostic accuracy.

To visually demonstrate how uncertainty evolves during the training process, we compute the mean of all elements in the two uncertainty maps, $Payoff_1$ and $Payoff_2$, from the UMF-GTSS, representing the overall uncertainty of fusion strategies. From Fig. 8, it is evident that as training progresses, the uncertainty gradually decreases and stabilizes. This suggests that the multimodal feature fusion is stabilizing and that the model is approaching its optimal solution. The decreasing uncertainty indicates that the network is refining its decision-making process and becoming more confident in its predictions as it learns from the data. This trend further implies that the model is effectively integrating the multimodal information to enhance its performance.

V. Conclusion

In this study, we propose the OphFusionNet designed to enhance the accuracy of ophthalmic disease diagnosis. To achieve this, we develop MSFF-SSA, which efficiently integrates multi-scale information from each modality. This approach effectively addresses the challenge of significant lesion size disparities across ophthalmic multimodal imaging.

Additionally, we introduce UMF-GTSS. By quantifying the uncertainty within the features, this strategy enables more

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

precise fusion of multimodal features, fully leveraging the correlations and complementarities across the modalities. As a result, it improves the diagnostic accuracy of the model.

To further optimize the multimodal learning process, we incorporate MDS, where the multimodal features guide the learning of the single-modal features. This approach effectively addresses the common issue of insufficient optimization for underperforming modality features during multimodal learning.

We conducted extensive experiments across four publicly available datasets, achieving superior performance compared to existing SOTA methods. Our approach demonstrates significant improvements in tasks such as fine-grained classification of AMD, glaucoma grading, and DME diagnosis. These results validate the effectiveness of our proposed method in enhancing both diagnostic accuracy and robustness.

The proposed framework does have some limitations. A key challenge is its reliance on paired multimodal data for training, which may not always be readily available, limiting its clinical applicability. In future work, we plan to explore the use of unpaired multimodal imaging data for model training. Specifically, we will investigate the use of techniques such as loose pairing [18], image synthesis [18] and domain adaptation [59], enabling the framework to better leverage all available multimodal data.

Another important challenge is the issue of missing modalities in the inference stage, as the absence of certain modalities can significantly impact the model's performance. We will investigate techniques for handling missing modalities, such as modality masking and imputation [60], efficient model adaptation [61], and knowledge distillation [62].

Additionally, we intend to explore the integration of large pre-trained models, which have shown great potential in enhancing the performance of multimodal systems. By combining these models with multimodal data, we expect to improve the framework's generalization capabilities, making it more adaptable to real-world clinical scenarios, even in the presence of missing or incomplete data.

REFERENCES

[1] S. R. Flaxman et al., "Global causes of blindness and distance vision impairment 1990-2020: A systematic review and meta-analysis," *Lancet Glob. Health*, vol. 5, no. 12, pp. e1221–e1234, Dec. 2017.

[2] P. Mitchell, G. Liew, B. Gopinath, and T. Y. Wong, "Age-related macular degeneration," *Lancet*, vol. 392, no. 10153, pp. 1147–1159, Sep. 2018.

[3] H. Jayaram, M. Kolko, D. S. Friedman, and G. Gazzard, "Glaucoma: Now and beyond," *Lancet*, vol. 402, no. 10414, pp. 1788–1801, Nov. 2023.

[4] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," *Lancet*, vol. 376, no. 9735, pp. 124–136, Jul. 2010.

[5] M. J. Burton et al., "The lancet global health commission on global eye health: Vision beyond 2020," *Lancet Glob. Health*, vol. 9, no. 4, pp. e489–e551, Apr. 2021.

[6] A. P. Marques et al., "The economics of vision impairment and its leading causes: A systematic review," *EClinicalMedicine*, vol. 46, Apr. 2022, Art. no. 101354.

[7] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 221–248, Jun. 2017.

[8] X. Chen et al., "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102444.

[9] K. Jin and J. Ye, "Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives," *Adv. Ophthalmol. Pract. Res.*, vol. 2, no. 3, Nov. 2022, Art. no. 100078.

[10] N. Panwar et al., "Fundus photography in the 21st century—A review of recent technological advances and their implications for worldwide healthcare," *Telemedicine E-Health*, vol. 22, no. 3, pp. 198–208, Mar. 2016.

[11] W. Drexler and J. G. Fujimoto, "State-of-the-art retinal optical coherence tomography," *Prog. Retin. Eye Res.*, vol. 27, no. 1, pp. 45–88, Jan. 2008.

[12] S. El-Ateif and A. Idri, "Multimodality fusion strategies in eye disease diagnosis," *J. Imag. Informat. Med.*, vol. 37, no. 5, pp. 2524–2558, Apr. 2024.

[13] Y. Li et al., "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Comput. Biol. Med.*, vol. 177, Jul. 2024, Art. no. 108635.

[14] R. Ma et al., "Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses," *Npj Digit. Med.*, vol. 8, no. 1, p. 64, Jan. 2025.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] T. K. Yoo, J. Y. Choi, J. G. Seo, B. Ramasubramanian, S. Selvaperumal, and D. W. Kim, "The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: A preliminary experiment," *Med. Biol. Eng. Comput.*, vol. 57, no. 3, pp. 677–687, Mar. 2019.

[17] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1591–1602, Jun. 2021.

[18] W. Wang et al., "Learning two-stream CNN for multi-modal age-related macular degeneration categorization," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4111–4122, Aug. 2022.

[19] E. Haihong et al., "KFWC: A knowledge-driven deep learning model for fine-grained classification of wet-AMD," *Comput. Methods Programs Biomed.*, vol. 229, Feb. 2023, Art. no. 107312.

[20] Z. Liu et al., "Cross-modal attention network for retinal disease classification based on multi-modal images," *Biomed. Opt. Exp.*, vol. 15, no. 6, pp. 3699–3714, May 2024.

[21] Y. Wang et al., "Geometric correspondence-based multimodal learning for ophthalmic image analysis," *IEEE Trans. Med. Imag.*, vol. 43, no. 5, pp. 1945–1957, May 2024.

[22] F. Cuzzolin, "Uncertainty measures: A critical survey," *Inf. Fusion*, vol. 114, Feb. 2025, Art. no. 102609.

[23] R. R. Yager, "Generalized Dempster–Shafer structures," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 3, pp. 428–435, Mar. 2019.

[24] R. I. John and P. R. Innocent, "Modeling uncertainty in clinical diagnosis using fuzzy logic," *IEEE Trans. Syst., Man Cybern., B, Cybern.*, vol. 35, no. 6, pp. 1340–1350, Dec. 2005.

[25] G. El Rahi, S. R. Etesami, W. Saad, N. B. Mandayam, and H. V. Poor, "Managing price uncertainty in prosumer-centric energy trading: A prospect-theoretic Stackelberg game approach," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 702–713, Jan. 2019.

[26] L. Huang, S. Ruan, Y. Xing, and M. Feng, "A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods," *Med. Image Anal.*, vol. 97, Oct. 2024, Art. no. 103223.

[27] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, no. 1, p. 17816, Dec. 2017.

[28] M. S. Ayhan, L. Kühlewein, G. Aliyeva, W. Inhoffen, F. Ziemssen, and P. Berens, "Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101724.

[29] J. Jaskari et al., "Uncertainty-aware deep learning methods for robust diabetic retinopathy classification," *IEEE Access*, vol. 10, pp. 76669–76681, 2022.

[30] Q. Zhou, T. Chen, H. Zou, and X. Xiao, "Uncertainty-aware incomplete multimodal fusion for few-shot central retinal artery occlusion classification," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102200.

[31] K. Zou et al., "Confidence-aware multi-modality learning for eye disease screening," *Med. Image Anal.*, vol. 96, Aug. 2024, Art. no. 103214.

[32] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 103–118.

[33] F. Hafner et al., "Cross-modal distillation for RGB-depth person re-identification," 2018, *arXiv:1810.11641*.

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.

[34] Y. Jin, G. Hu, H. Chen, D. Miao, L. Hu, and C. Zhao, "Cross-modal distillation for speaker recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 11, 2023, pp. 12977–12985.
[35] J. Jiang, A. Rimner, J. O. Deasy, and H. Veeraraghavan, "Unpaired cross-modality educed distillation (CMEDL) for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 5, pp. 1057–1068, May 2022.
[36] D. Song et al., "Asynchronous feature regularization and cross-modal distillation for OCT based glaucoma diagnosis," *Comput. Biol. Med.*, vol. 151, Dec. 2022, Art. no. 106283.
[37] L. Wang et al., "MultiEYE: Dataset and benchmark for OCT-enhanced retinal disease recognition from fundus images," *IEEE Trans. Med. Imag.*, vol. 44, no. 4, pp. 1711–1722, Apr. 2025.
[38] L. Zhang, C. Bao, and K. Ma, "Self-distillation: Towards efficient and compact neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4388–4403, Aug. 2022.
[39] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5896–5905.
[40] J. Ma, W. Jiang, X. Tang, X. Zhang, F. Liu, and L. Jiao, "Multiscale sparse cross-attention network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025.
[41] B. Zhang, I. Titov, and R. Sennrich, "Sparse attention with linear units," 2021, *arXiv:2104.07012*.
[42] K. Shen, J. Guo, X. Tan, S. Tang, R. Wang, and J. Bian, "A study on ReLU and softmax in transformer," 2023, *arXiv:2302.06461*.
[43] D. So, W. Mańke, H. Liu, Z. Dai, N. Shazeer, and Q. V. Le, "Searching for efficient transformers for language modeling," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 6010–6022.
[44] J. F. Nash Jr., "Equilibrium points in n-person games," *Proc. Nat. Acad. Sci. USA*, vol. 36, no. 1, pp. 48–49, 1950.
[45] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12692–12702.
[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[47] J. Wu et al., "GAMMA challenge: Glaucoma grading from multi-modality images," *Med. Image Anal.*, vol. 90, Dec. 2023, Art. no. 102938.
[48] APTOS 2021. (2021). *Prediction of Anti-VEGF Treatment Outcomes in Patients With Diabetic Maculopathies*. [Online]. Available: https://tianchi.aliyun.com/specials/promotion/APTOS?spm=a2c22.12281978.0.0
[49] APTOS 2023. (2023). *Report Generation From Fundus Fluorescein Angiography Images*. [Online]. Available: https://tianchi.aliyun.com/dataset/170128
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[51] Y. Zhang, F. Xie, and J. Chen, "TFormer: A throughout fusion transformer for multi-modal skin lesion diagnosis," *Comput. Biol. Med.*, vol. 157, May 2023, Art. no. 106712.
[52] P. Yan et al., "TDF-Net: Trusted dynamic feature fusion network for breast cancer diagnosis using incomplete multimodal ultrasound," *Inf. Fusion*, vol. 112, Dec. 2024, Art. no. 102592.
[53] S. Feng et al., "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.
[54] T. Zhou, Y. Zhou, G. Li, G. Chen, and J. Shen, "Uncertainty-aware hierarchical aggregation network for medical image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7440–7453, Aug. 2024.
[55] H. Shao, Q. Zeng, Q. Hou, and J. Yang, "MCANet: Medical image segmentation with multi-scale cross-axis attention," *Mach. Intell. Res.*, vol. 22, no. 3, pp. 437–451, Jun. 2025.
[56] X. Lu, Y. Jiang, H. Hong, Q. Sun, and C. Zhuo, "DCAFuse: Dual-branch diffusion-CNN complementary feature aggregation network for multi-modality image fusion," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 1524–1533.
[57] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, Nov. 2008.
[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.
[59] J. Jiang et al., "Self-derived organ attention for unpaired CT-MRI deep domain adaptation based MRI segmentation," *Phys. Med. Biol.*, vol. 65, no. 20, Oct. 2020, Art. no. 205001.
[60] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou, "Unified multi-modal image synthesis for missing modality imputation," *IEEE Trans. Med. Imag.*, vol. 44, no. 1, pp. 4–18, Jan. 2025.
[61] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 2, pp. 742–754, Feb. 2025.
[62] H. Wang et al., "Learnable cross-modal knowledge distillation for multi-modal learning with missing modality," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2023, pp. 216–226.

Authorized licensed use limited to: Soochow University. Downloaded on October 15,2026 at 07:03:20 UTC from IEEE Xplore. Restrictions apply.