Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2024.058647

ARTICLE

AMSFuse: Adaptive Multi-Scale Feature Fusion Network for Diabetic Retinopathy Classification

Chengzhang Zhu[1,2], Ahmed Alasri[1], Tao Xu[3], Yalong Xiao[1,2,*], Abdulrahman Noman[1], Raeed Alsabri[1], Xuanchu Duan[4] and Monir Abdullah[5]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2]School of Humanities, Central South University, Changsha, 410083, China
[3]Department of Human Anatomy and Histology & Embryology, Basic Medical Sciences, Changsha Health Vocational College, Changsha, 410100, China
[4]Glaucoma Institute, Changsha Aier Eye Hospital, Changsha, 410000, China
[5]Department of Computer Science and Artificial Intelligence, College of Computing and Information Technology, University of Bisha, Bisha, 67714, Saudi Arabia

* Corresponding Author: Yalong Xiao. Email: ylxiao@csu.edu.cn

Received: 17 September 2024; Accepted: 28 November 2024; Published: 06 March 2025

ABSTRACT: Globally, diabetic retinopathy (DR) is the primary cause of blindness, affecting millions of people worldwide. This widespread impact underscores the critical need for reliable and precise diagnostic techniques to ensure prompt diagnosis and effective treatment. Deep learning-based automated diagnosis for diabetic retinopathy can facilitate early detection and treatment. However, traditional deep learning models that focus on local views often learn feature representations that are less discriminative at the semantic level. On the other hand, models that focus on global semantic-level information might overlook critical, subtle local pathological features. To address this issue, we propose an adaptive multi-scale feature fusion network called (AMSFuse), which can adaptively combine multi-scale global and local features without compromising their individual representation. Specifically, our model incorporates global features for extracting high-level contextual information from retinal images. Concurrently, local features capture fine-grained details, such as microaneurysms, hemorrhages, and exudates, which are critical for DR diagnosis. These global and local features are adaptively fused using a fusion block, followed by an Integrated Attention Mechanism (IAM) that refines the fused features by emphasizing relevant regions, thereby enhancing classification accuracy for DR classification. Our model achieves 86.3% accuracy on the APTOS dataset and 96.6% RFMiD, both of which are comparable to state-of-the-art methods.

KEYWORDS: Diabetic retinopathy; multi-scale feature fusion; global features; local features; integrated attention mechanism; retinal images

1 Introduction

Diabetic retinopathy (DR) is a condition that affects the retina as a result of diabetes complications, resulting in permanent eye damage and, in some cases, vision loss. If left untreated, this form of issue among patients has a high priority of people going blind. DR is a leading cause of vision impairment and blindness among working-age adults globally. Early detection and management are crucial in preventing irreversible damage.

According to recent statistics, around 537 million adults globally have diabetes as of 2023, with the number expected to increase to 783 million by 2045 and potentially 1.3 billion by 2050. This increase emphasizes the rising worldwide impact of diabetes [1]. As shown in Fig. 1, DR is classified into five stages, reflecting the progression of the disease:

1. Normal: where there are no retinopathy symptoms.
2. Mild Non-Proliferative Retinopathy: characterized by the presence of microaneurysms. At this stage, patients may not experience symptoms, but microaneurysms indicate the onset of retinal damage. Regular monitoring is essential to prevent progression.
3. Moderate Non-Proliferative Retinopathy: where some blood vessels become distorted and swollen. Blockages in retinal vessels begin to occur, potentially leading to noticeable visual changes. Early intervention can help manage the condition.
4. Severe Non-Proliferative Retinopathy: involving significant blockage of blood vessels. A larger number of vessels are blocked, signaling the retina to grow new vessels, which can lead to more severe complications if not addressed promptly.
5. Proliferative Retinopathy: an advanced stage marked by the growth of new blood vessels. These fragile new vessels can bleed into the vitreous, causing vision loss or blindness. Immediate treatment is critical at this stage to prevent severe outcomes.

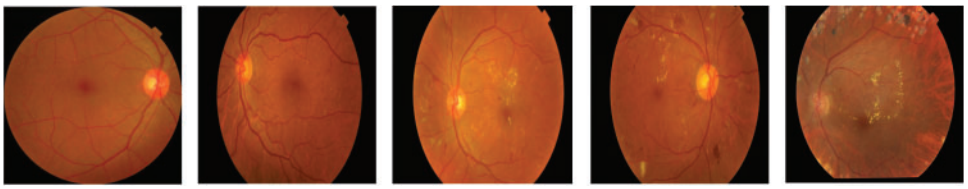

Figure 1: Retinal fundus images with various stages of DR

Each stage of DR has distinct characteristics and features that doctors may dismiss, potentially resulting in misdiagnosis. Accurate classification of these stages is vital for determining appropriate treatment strategies, such as laser therapy or surgical interventions, to prevent vision deterioration. This difficulty emphasizes the necessity for an automated approach to assist in the correct detection and classification of DR. It is estimated that timely and appropriate treatment and regular screening of the eyes can reduce the incidence of new cases of this disease by at least 56% [2]. With the goal of early detection, regular screening is advised by the International Council of Ophthalmology, and the American Academy of Ophthalmology [3]. These organizations recommend screening intervals of 12 to 24 months for patients without or with mild DR. While regular screening is vital for preventing blindness, the anticipated rise in diabetes patients poses a significant challenge for the screening and follow-up processes. To address this growing demand, automated DR detection systems leveraging artificial intelligence offer a scalable and efficient solution. Unlike traditional manual screening methods, AI approaches can analyze large volumes of retinal images rapidly, ensuring timely diagnosis and reducing the burden on healthcare professionals.

The diagnosis of DR has advanced significantly in the last ten years in the field of deep learning [4]. Convolutional Neural Networks (CNNs) have been proposed in a number of different forms to automate the grading of DR [5]. Recently, Vision Transformers (ViTs) have become a powerful tool that enhances deep learning models' capabilities [6]. ViTs have demonstrated their effectiveness with DR classification, such as [7] and computed tomography [8].

However, in these models, images are represented as one-dimensional token sequences, which leads to ignoring either their actual local or global structures. Combining multi-scale global and local features which are necessary for tasks like segmentation and image classification is the first step toward solving this problem [9]. Although significant progress has been made by recent works, including ViTAE [10], StoHisNet [11], Transfuse [12], CMT [13], and Comformer [14], which integrated features derived from convolutional and self-attention processes [15], challenges remain in adaptively fusing multi-scale features without compromising their individual strengths.

To address this issue, we introduced AMSFuse, which integrates global and local blocks to concurrently extract and adapt global and local features. These are followed by fusion blocks that merge these features at various semantic scales and then comes an Integrated Attention Mechanism (IAM) that refines and weights fused features to enhance representation. This approach enhances our model's performance compared to earlier classification techniques that primarily depend on either transformer or convolutional models. Results show that our approach achieves better results on common DR datasets. Our contributions can be listed as follows:

1. We proposed an AMSFuse network that integrates global and local features at multiple scales. This network adaptively captures both semantic information and local spatial features through global and local feature blocks.
2. We introduced an integrated attention mechanism to refine the fused global and local features, highlighting critical pathological regions for enhanced diagnostic precision.
3. The proposed model achieves state-of-the-art performance with 86.3% accuracy on the APTOS dataset and 96.6% on the RFMiD dataset, demonstrating competitive or superior performance compared to existing methods.

2 Related Work

2.1 DR Classification Based on CNN

Due to the complexity of eye fundus images, traditional approaches for classifying diabetic retinopathy have faced difficulties, which have limited early-stage identification. Convolutional neural networks (CNNs), one deep learning technique, has demonstrated potential in tackling these issues by leveraging artificial intelligence for classification [3]. Deep learning methods have shown great potential for diagnosing and categorizing DR, which is a leading cause of blindness in patients with diabetes. These techniques are particularly beneficial for the early detection and classification of DR, several studies have suggested using deep learning models, including MobileNetV2 [4], VGG16 [5], DenseNet121 [16], and ResNet50 [16]. These models have shown to be useful in automating the diagnosis process, showing good results when identifying DR based on the retinal fundus.

2.2 DR Classification Based on ViT

ViT has shown promising results in medical image classification tasks [6], including DR classification. With new findings, researchers have investigated the application of transformer-based models such as BEiT [7], DeiT [8], CaiT [17], TransMIL [18], and MIL-ViT [19] for automated comprehension of DR severity from fundus images. Transformer models have achieved significant success in the field of computer vision applications, despite their superior performance in natural language processing [20]. Recent models such as Mvitv2 [21], ViT-CoMer [22], and EfficientViT [22] have enhanced the performance of the ViT architecture, facilitating the extraction of global features. Furthermore, compared to conventional convolution-based techniques, the use of pre-trained transformers that have been optimized for use on DR datasets.

2.3 Multi-Scale Feature Fusion

Multi-scale feature fusion is vital for accurately classifying medical images. To tackle the limitation of insufficient local features, DeiT [8] introduced a distillation token to combine CNN features with ViT. Additionally, T2TViT [23] proposed a designed method to improve the Vision Transformer's ability to capture local features. Methods such as CMT [13], Conformer [14], VitAE [10], and StoHis [11] demonstrate that integrating local features with global representations significantly improves the Transformer's ability to notice small local features. Furthermore, models such as MFFM [24] and SegR-Net [25] Although these models perform exceptionally well on common datasets like ImageNet and various downstream tasks, they do not achieve the same success in the medical imaging domain. This shortfall is due to the insufficient datasets of medical images, where pathological features are more scattered and challenging to detect compared to ordinary images.

In order to improve DR lesion classification, [26] offers multi-scale feature fusion. However, it primarily focuses on local features, limiting its ability to capture the global feature context, which is necessary for accurate DR classification.

Therefore, we decided to leverage adaptive features along with multi-scale global-local characteristics. Our novel approach, AMSFuse, presents an adaptive parallel fusion network that is developed to maintain a parallel structure and prevent interference between local and global features. Convolutional layers specialize in detecting and extracting local patterns and features within images, such as edges and textures. On the other hand, ViT utilizes a transformer architecture, enabling it to analyze the entire image at once, capturing global context at different parts of the image simultaneously without the need for sequential processing. These features are efficiently integrated within this network by an adaptive feature fusion block, which maintains the integrity of both global and local blocks. Furthermore, we incorporate IAM to refine the fused features, emphasizing relevant regions and enhancing the overall classification accuracy for DR.

3 Proposed Method

3.1 Architecture Overview

As shown in Fig. 2, the AMSFuse network, introduced as a new method for DR classification, aims to adaptively capture both local spatial details and global semantic representations of fundus images at varying scales. This is achieved by employing a parallel structure that extracts global and local information through distinct feature blocks. The features are then adaptively fused using the AMSFuse block, and a downsampling step is performed. This process is repeated across four stages. Finally, the output from the last AMSFuse Block is processed through an integrated attention mechanism, followed by a Global Average Pooling layer, a Layer Norm layer, and a Linear layer to produce the classification result. This network facilitates a comprehensive integration of multi-scale global and local features, significantly enhancing the accuracy of DR classification.

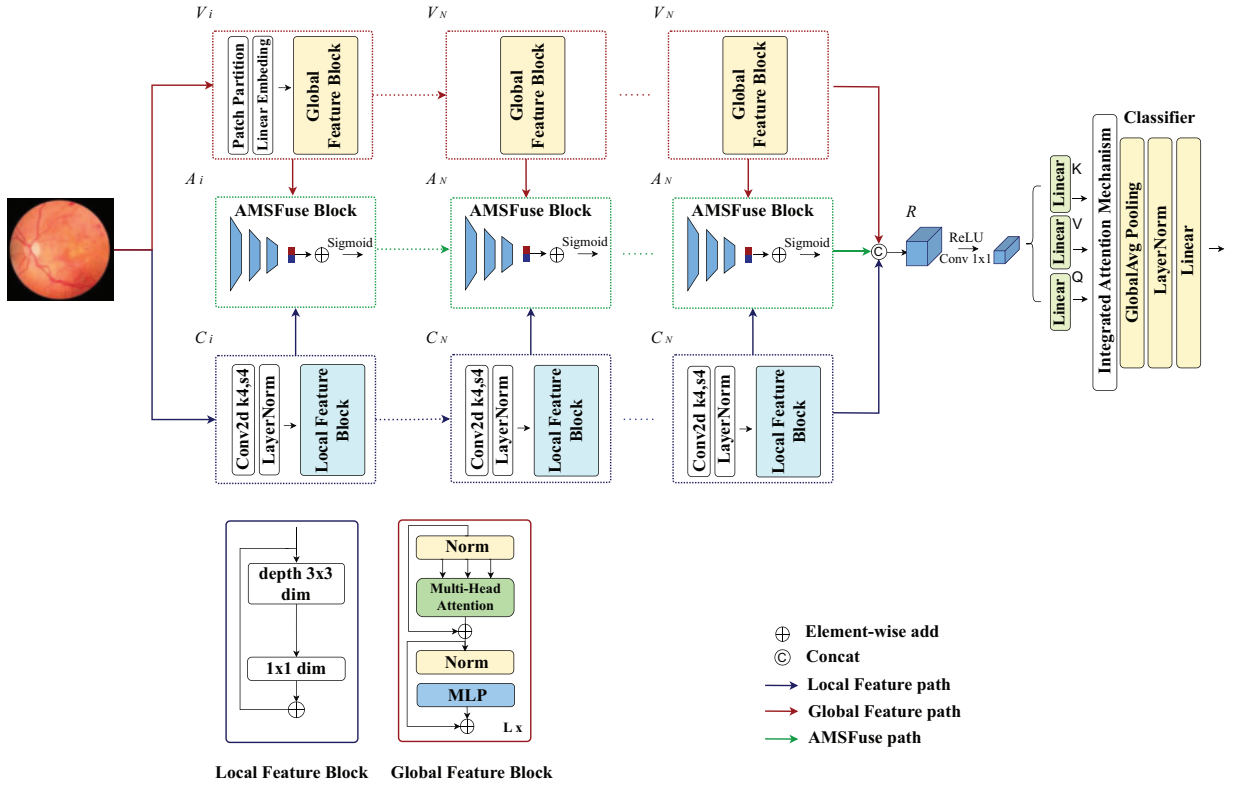

Figure 2: Proposed AMSFuse network

3.2 Local Feature Block

Local spatial features in fundus images are crucial. The local feature block, depicted in (1), utilizes a 3 × 3 depthwise convolution, a specific type of grouped convolution where each group contains a single channel. This depthwise convolution significantly reduces the FLOPs of the network. Following this, cross-channel information is integrated through linear layers [9]. The extracted local features are then fed into the AMSFuse block, as shown in (1). The output features, denoted by C_i, where $i = 1, 2, \ldots, N$ and N represents the number of blocks, are computed through a combination of two operations:

- **Depthwise Convolution:** A 3 × 3 depthwise convolution operation $f^{\text{depth3}\times3}$ is applied to the previous layer's output features C_{i-1}. This operation extracts spatial features from C_{i-1} while preserving the number of channels.
- **Linear Transformation and Addition:** The result of the depthwise convolution is passed through a linear transformation (f_1 × 1) with a 1 × 1 kernel. Finally, this transformed output is added element-wise to the previous layer's features C_{i-1}.

$$C_i = f^{1\times1}(\text{LN}(f^{\text{depth3}\times3}(C_{i-1}))) + C_{i-1} \tag{1}$$

$f^{\text{depth3}\times3}$ is the depthwise convolution function, while LN is the layer normalization applied to the convolved features. The linear transformation $f^{1\times1}$ integrates cross-channel information to provide efficient extraction and aggregation of local features before input into the AMSFuse block.

3.3 Global Feature Block

The global feature block denoted by V_i for $i = 1, 2, \ldots, N$, where N is the number of blocks utilizes a pre-trained ViT model [6]. This model processes the image by first dividing it into patches and embedding them into a higher-dimensional space. These embeddings are then fed into transformer encoder blocks, which consist of multi-head self-attention layers and feed-forward networks. The self-attention layers allow the model to focus on informative parts of the image by learning relationships between different patches. The feed-forward networks introduce non-linearity, the output from the last encoder block, which captures high-level semantic information about the image content. The final feature representation is passed as input to the AMSFuse block for further processing.

3.4 AMSFuse Block

The AMSFuse block, indicated by A_i for $i = 1, 2, \ldots, N$, where N represents the number of blocks, is designed to effectively integrate local and global features extracted from various layers. This integration encompasses different representations and semantic information, enabling a more comprehensive feature fusion. It is designed to enhance fusion features by adaptively focusing on important feature channels. AMSFuse block receives two sets of input features: global features extracted by the global Feature Block (denoted as V_i) and local features extracted by the local Feature Block (denoted as C_i). These features are obtained at different semantic scales and carry complementary information about the retinal image. To ensure that the features from both sources are compatible for fusion, the adaptive feature pooling step is employed. This step adjusts the spatial dimensions of the global features to match those of the local features. It achieves this by applying adaptive average pooling, which ensures that the global features are downsampled or upsampled appropriately without losing crucial semantic information. V_i' denotes the global feature map after it has been resized using adaptive average pooling the process is depicted in (2):

$$V_i' = \text{AdaptiveAvgPool}(V_i, \text{size} = C_i.\text{shape}[-2:]), \tag{2}$$

after matching the spatial dimensions, the global and local features are combined through element-wise addition. This operation effectively merges the high-level semantic information from the global features with the fine-grained details from the local features, providing a richer and more comprehensive feature representation the process is depicted in (3):

$$A_i = V_i' + C_i, \tag{3}$$

where (A_i) the combined features are then passed through the fusion block, which consists of several stages, each incorporating a 3×3 depthwise convolution followed by a 1×1 convolution. The depthwise convolution extracts spatial information, while the 1×1 convolution integrates cross-channel information. This dual-convolution approach ensures that the fused features retain both spatial and channel-wise information from the original global and local features the process is depicted in (4):

$$A_i' = f^{\text{depthwise } 3\times3}(A_i), \tag{4}$$

A_i' represents the feature map after applying the 3×3 depthwise convolution to A_i. as depicted in (5), A_i'' represents the feature map after applying the 1×1 convolution to A_i'.

$$A_i'' = f^{1\times1}(A_i'). \tag{5}$$

3.5 Integrated Attention Mechanism (IAM)

To further refine the fused features, we propose an Integrated Attention Mechanism. This mechanism integrates features from various stages of the model, including the global features block, local features block, and adaptively fused features from the AMSFuse block, depicted in (6), the reduced feature map, indicated as R, is created by concatenating the global features (V_i), local features (C_i), and adaptively fused features (A_i) from the AMSFuse block, followed by a 1×1 convolution to reduce the number of channels.

$$R = \text{Conv2d}(\text{cat}(V_i, C_i, A_i)). \tag{6}$$

After reducing the feature map, it is passed through the IAM. The query, key, and value tensors are obtained from the input feature map R using 1×1 convolutions with adaptable weight matrices W_Q, W_K, and W_V. These tensors are then reshaped and permuted to facilitate the dot product operation. The dot product of Q and K is scaled by $\sqrt{d_k}$ and passed through the softmax function to obtain attention weights. These weights are used to compute a weighted sum of the value tensor V, resulting in the final attention-weighted output. The query, key, and value are calculated as follows (7):

$$Q = W_Q \times R, \quad K = W_K \times R, \quad V = W_V \times R. \tag{7}$$

The attention mechanism calculates the output by taking the dot product of queries Q and keys K, scaling it by the square root of the key dimension d_k using the softmax function and then multiplying the result by the values V depicted in (8).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V, \tag{8}$$

the final output from the IAM, after applying the attention mechanism is computed as (9).

$$O_i = \sigma(\text{Conv2}(\text{ReLU}(\text{Conv1}(A_i)))), \tag{9}$$

where O_i is the final output of the Integrated Attention Mechanism (IAM) after applying a set of transformations to the fused feature map A_i. To obtain O_i, first apply a convolutional layer (Conv1) to the input A_i, followed by a ReLU activation function, another convolutional layer (Conv2), and finally a sigmoid activation function (σ). The transformations produce a refined representation of the input feature map, which is then passed into the classifier to generate the final classification result.

The Integrated Attention Mechanism (IAM) effectively leverages the combined feature map, by dynamically focusing on different parts of the input feature map, (IAM) enhances the model's ability to capture complex dependencies and relationships within the data, leading to improved performance. The final output fed to the classifier to obtain the classification.

3.6 Algorithm Summary

Algorithm 1 provides a comprehensive overview of the AMSFuse network. The methodology begin with the extraction of local and global features through dedicated feature extraction blocks. These features are subsequently fused within the AMSFuse block using an adaptive fusion strategy, ensuring a seamless integration of multi-scale information. The Integrated IAM further refines the fused features by prioritizing critical regions and suppressing irrelevant information, enhancing the model's focus on diagnostically significant areas. This systematic approach leverages both global contextual and localized spatial information, resulting in a robust and highly accurate classification of DR severity.

Algorithm 1: AMSFuse algorithm for diabetic retinopathy classification

Input: Retinal fundus images
Output: Diabetic retinopathy classification
Step 1: Feature Extraction
Extract global features V_i using a pre-trained Vision Transformer (ViT) and local features C_i using depthwise convolutions (Eq. (1)).
Step 2: Adaptive Fusion
Adaptively fuse V_i and C_i with the AMSFuse block (Eqs. (2), (3)).
Step 3: Integrated Attention Mechanism (IAM)
Refine the fused features using IAM (Eqs. (6)–(9)).
Step 4: Classification
Pass the refined features through Global Average Pooling, Layer Normalization, and a Linear layer to produce class probabilities. The class with the highest probability is selected as the final prediction.

4 Experimental Results and Visualization

The AMSFuse model is trained using the cross-entropy loss function, which is commonly used for classification tasks. This loss function measures the difference between the predicted class probabilities and the actual class labels, and its formula is as follows:

$$\text{Loss} = -\sum_{i=1}^{n} y_i \log \hat{y}_i. \tag{10}$$

In Eq. (10), y_i represents the ground truth label of sample i, and $\hat{y}_i$ represents the predicted probability of sample i being the target label.

We ran a series of experiments with two datasets to evaluate the AMSFuse model's effectiveness and reliability. The outcomes of these trials show that, in terms of classification performance, our method performs better than earlier state-of-the-art network models. In the subsequent sections, we will offer an in-depth analysis of the datasets used, outline the evaluation metrics utilized, and describe the experimental setup. Following this, we will present the results of our experiments across all datasets, including a comprehensive examination of our model's performance through a series of ablation studies conducted on the APTOS dataset.

4.1 Dataset

We utilized two publicly available datasets for evaluating the performance of the AMSFuse model: APTOS2019 and RFMiD.

The **APTOS2019 Blindness Detection Dataset** [27] includes 5590 fundus images with five DR classifications labeled into various stages of DR: No DR (label 0), Mild (label 1), Moderate (label 2), Severe (label 3), and Proliferative (label 4). Only the annotations of the training set (3662 photos) are publicly accessible. These annotations were randomly divided into 70% for training, 15% for validation, and 15% for testing.

The **RFMiD Retinal Fundus Multi-disease Dataset** [28] contains a total of 1900 fundus images, and we exclusively utilize it for diagnosing diabetic retinopathy. In both datasets, we followed standardized protocols commonly used in the research community to ensure consistency and fairness in our study. All utilized images are of high quality, with clear visibility of retinal features necessary for accurate DR classification.

4.2 Pre-Trained Model

A ViT model trained on a dataset of 345,271 fundus samples is used as the pre-trained model for retinal disease diagnosis tasks. This dataset includes samples annotated as normal (208,733), diabetic retinopathy (38,284), age-related macular degeneration (21,962), glaucoma (24,082), and cataract (67,230) [19].

4.3 Evaluation Metrics and Implementation Information

We chose ACC, F1, Precision, and Kappa as the main classification metrics, all derived from the confusion matrix. This matrix consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) counts. Thus, Accuracy (ACC) is calculated using the following formula Eq. (11), which represents the proportion of correctly identified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{11}$$

Precision, reflecting the model's prediction accuracy, is computed using Eq. (12), representing the ratio of true positive samples among those predicted to be positive:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{12}$$

Recall is calculated using the Eq. (13):

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{13}$$

this equation calculates the recall rate, which measures the percentage of actual positive cases that the model accurately identifies.

The F1 score, defined by Eq. (14), balances Precision and the proportion of actual positives:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}. \tag{14}$$

The Kappa coefficient, used to measure classifier consistency, ranges from 0 (random agreement) to 1 (perfect agreement), and is calculated by Eq. (15):

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{15}$$

where P_o represents the observed agreement, which is the actual accuracy of the model, calculated as the proportion of instances where the model's predictions agree with the ground truth. P_e is the expected agreement by chance, calculated based on the marginal frequencies of the classes in the dataset.

4.4 Parameter Tuning and Analysis

To optimize AMSFuse, we conducted a grid search for key hyperparameters. The learning rate was set to $1e^{-4}$ after testing values from $1e^{-6}$ to $1e^{-3}$, ensuring a balance between convergence speed and stability. AMSFuse block and IAM parameters were fine-tuned to enhance multi-scale feature fusion and emphasize critical image regions. For comparison methods, we used the parameters reported in their original publications for fairness. Additionally, the model was trained for a total of 100 epochs, determined based on preliminary experiments to ensure sufficient learning without overfitting. A batch size of 32 was

selected to balance computational efficiency and model performance. The optimizer used was AdamW with a momentum of 0.9 and a weight decay of 0.01 to regularize the model and improve generalization. The minimum learning rate was set to $1e^{-6}$ to allow for fine-tuning during the later stages of training, shown in Table 1.

Table 1: Selected hyperparameters and training configuration for AMSFuse

Hyperparameter/Configuration	Value
Framework	PyTorch
GPU	NVIDIA A100-SXM4-40 GB
Optimizer	AdamW
Learning rate	$1e^{-4}$
Min learning rate	$1e^{-6}$
Optimizer momentum	0.9
Batch size	32
Number of epochs	100
Weight decay	0.01
Number of layers	6
Image size	384 × 384

4.5 Comparison with Other Advanced Models

DR Classification with APTOS Dataset. The comparison on the APTOS2019 dataset, shown in Table 2, clearly shows that the AMSFuse network outperforms several state-of-the-art methods for diabetic retinopathy (DR) grading. AMSFuse stands out with an impressive accuracy of 86.3%, an F1 score of 86.2%, and a Kappa coefficient of 92.8%, which are the highest among the models compared. This indicates that AMSFuse not only accurately identifies DR cases but also provides reliable and balanced performance. The improvements over other models, including DLI [29], BiFormer-B [30], EfficientViT [31], and even advanced architectures like ViT-CoMer [22] and MIL-ViT [19]. Fig. 3 shows the experimental results visualization.

Table 2: Comparison with state-of-the-art DR classification methods on APTOS2019

Method	Acc (%)	F1 (%)	Kappa (%)
DANIL [32]	83.8	67.2	–
GREEN-ResNet50 [33]	84.4	83.6	90.8
GREEN-SE-ResNext50 [33]	85.7	85.2	91.2
UniFormer-B [34]	80.2 ± 0.4	79.5 ± 0.3	88.0 ± 0.2
BiFormer-B [30]	81.0 ± 0.5	80.5 ± 0.4	89.0 ± 0.3
TransMIL [18]	82.5	82.9	91.5
Mvitv2 [21]	82.6 ± 0.3	82.6 ± 0.4	90.7 ± 0.2
ViT-CoMer [22]	83.2 ± 0.4	83.0 ± 0.5	91.2 ± 0.3
EfficientViT [22]	84.1 ± 0.3	84.4 ± 0.2	91.9 ± 0.4
MIL-ViT [19]	85.8 ± 0.2	85.5 ± 0.3	92.3 ± 0.2
AMSFuse (Ours)	**86.3±0.2**	**86.2±0.3**	**92.8±0.2**

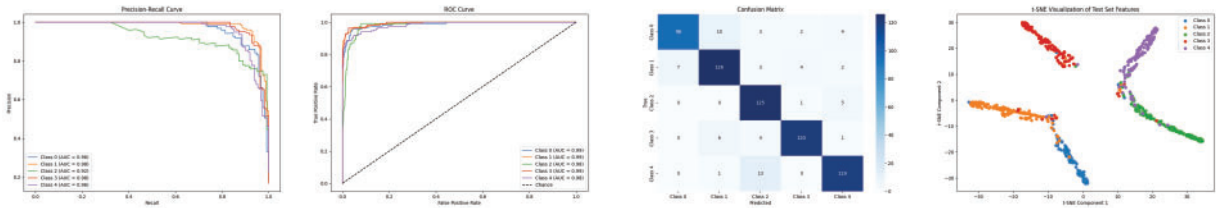

Figure 3: PR curve, ROC curve, confusion matrix, and t-SNE visualization on APTOS

DR Classification with RFMiD Dataset. In the comparison of the RFMiD2020 dataset, shown in Table 3, the AMSFuse network outperforms various leading diabetic retinopathy (DR) classification methods. AMSFuse has the highest metrics, with an accuracy of 96.6%, F1 score of 96.1%, precision of 97.1%, and recall of 96.5%. These results demonstrate the model's performance to accurately and reliably categorize DR instances. Other models, such as ResNeXt50 [35] show lower performance metrics, with ResNet34 obtaining 87.8% accuracy and ResNeXt50 achieving 89.8%. Mvitv2 [21] and BiFormer-B [34] perform well, but fall short of AMSFuse, with accuracies of 88.0% and 88.5%, respectively. Advanced models such as TransMIL [18], EfficientViT [31], and ViT-CoMer [22] demonstrate improvements, with CSA-DMIL reaching 90.6% accuracy. However, the AMSFuse model shows better results compared to these models. Notably, the MIL-ViT [19] model achieves competitive results with 93.2% accuracy, but it does not surpass the performance of AMSFuse. Fig. 4 shows the experimental results visualization.

Table 3: Comparison with state-of-the-art DR classification methods on RFMiD2020

Method	Acc (%)	F1 (%)	Precision (%)	Recall (%)
ResNeXt50 [35]	89.8	93.6	92.9	94.4
Swin [36]	88.5	92.8	92.2	93.4
UniFormer-B [34]	88.8 ± 0.3	89.1 ± 0.2	88.7 ± 0.3	89.6 ± 0.2
BiFormer-B [30]	89.1 ± 0.2	89.5 ± 0.3	89.1 ± 0.2	90.2 ± 0.3
CSA-DMIL [37]	90.6	93.9	94.0	93.8
LG-DMIL [38]	89.8	93.7	94.1	94.4
TransMIL [18]	91.1	94.5	95.5	93.6
Mvitv2 [21]	91.3 ± 0.3	94.3 ± 0.2	95.8 ± 0.3	94.1 ± 0.2
ViT-CoMer [22]	91.6 ± 0.2	95.1 ± 0.3	95.4 ± 0.2	94.6 ± 0.3
EfficientViT [31]	92.2 ± 0.3	96.4 ± 0.2	94.2 ± 0.4	93.1 ± 0.3
MIL-ViT [19]	93.2 ± 0.2	95.8 ± 0.3	96.8 ± 0.2	94.9 ± 0.2
AMSFuse (Ours)	**96.6±0.2**	**96.1±0.2**	**97.1±0.2**	**96.5±0.3**

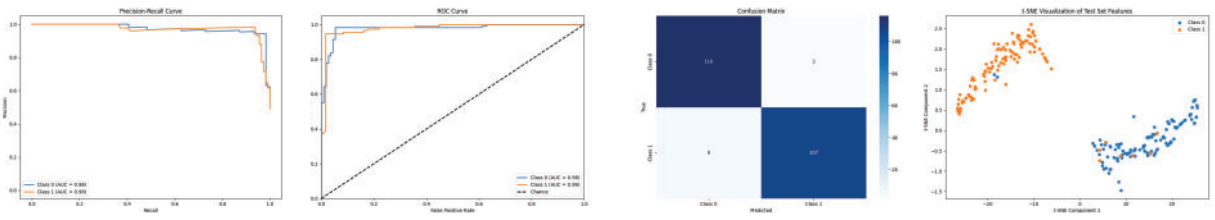

Figure 4: PR curve, ROC curve, confusion matrix, and t-SNE visualization on RFMiD

4.6 Ablation Study

We conducted an ablation study to evaluate the impact of the different branches of our network on the APTOS dataset. The study is divided into three levels that examine the contributions of the local block, global block, AMSFuse, and IAM.

Effectiveness of Local Block, Global Block, and AMSFuse. We started by comparing the performance of running the local and global blocks individually using AMSFuse. The results demonstrate that the global block outperforms the local block, and the application of AMSFuse considerably increases performance. The local block obtained 75.0% accuracy, 60.0% F1 score, and 65.0% Kappa, whereas the global block increased these metrics to 78.0%, 68.0%, and 70.0%. The integration of AMSFuse led to significant gains, obtaining the greatest metrics with an accuracy of 86.3%, F1 score of 86.2%, and Kappa of 92.8%, shown in Table 4.

Table 4: Experimental results of the effectiveness of local block and global block on the APTOS dataset

Branch	Acc (%)	F1 (%)	Kappa (%)
- Local Block	75.0	60.0	65.0
- Global Block	78.0	68.0	70.0
- AMSFuse	86.3	86.2	92.8

Impact of AMSFuse Block and IAM. An additional experiment was conducted to further assess the influence of the AMSFuse Block and IAM in the performance of the model. For this ablation experiment, three configurations were tested:

- Model with the IAM disabled: with only global and local branches without AMSFuse block.
- Model with AMSFuse block disabled: with only global and local branches with IAM.
- Model with AMSFuse and IAM disabled.
- Model with AMSFuse and IAM enabled.

Table 5 the results indicate that the AMSFuse block, along with IAM, significantly contributes to improved performance. The model with concat and the IAM disabled achieved 79.4% accuracy, 79.6% F1 score, and 83.3% Kappa. On the other hand, the model without the AMSFuse block obtained 80.1% accuracy, 80.5% F1 score, and 92.0% Kappa. In contrast, the model that integrates both blocks exhibited remarkable improvements with an accuracy of 86.3%, F1 score of 86.2%, and Kappa of 92.8%, which demonstrates the benefits of adaptively fusing multi-scale features without compromising their individual representation.

Table 5: Ablation study results on AMSFuse block and IAM on APTOS dataset

Branch	Acc (%)	F1 (%)	Kappa (%)
- Global + Local (Concat)	79.4	79.6	83.3
- Global + Local + IAM	80.1	80.5	92.0
- AMSFuse − IAM	85.8	85.5	92.3
- AMSFuse + IAM	86.3	86.2	92.8

The ablation study results indicate that both the AMSFuse Block and IAM play crucial roles in enhancing the performance of the proposed network. The global block consistently outperforms the local block when

used individually, and the integration of these blocks via AMSFuse further boosts accuracy. The IAM further refines the feature fusion, leading to significant gains in classification metrics.

5 Discussion

While AMSFuse has demonstrated performance in grading DR, there are several constraints to consider. Variability in datasets, such as differences in image quality and patient demographics, may influence how well the model performs in different clinical contexts.

To better understand AMSFuse's effectiveness, it's helpful to compare its performance with current clinical standards. For instance, research has shown that deep-learning algorithms can sometimes match the accuracy of manual grading by retinal specialists, particularly in terms of sensitivity and specificity across various stages of DR classification [39]. In this regard, AMSFuse's results suggest that it has the potential to improve conventional diagnostic techniques in healthcare settings.

Interpretability is crucial for clinical acceptance. Making the model's decisions more transparent will help to build trust among healthcare professionals. Future work will focus on integrating more interpretable components. With the potential to adapt AMSFuse for related tasks, such as medical image segmentation, we expect that our approach could improve segmentation accuracy compared to traditional methods, and we plan to investigate this further. By capturing global context and local details through adaptive multi-scale feature fusion, and highlighting important regions with the IAM, we hope to enhance the model's ability to focus on critical anatomical features. This could potentially lead to better segmentation performance, but further research is needed to confirm this. Furthermore, AMSFuse's computational demands and the need for high-performance hardware may limit its use in resource-constrained environments. By addressing these areas, we aim to make AMSFuse a reliable and versatile tool for clinical use.

6 Conclusion

We introduced AMSFuse, an Advanced Multi-Scale Fusion framework designed to improve the classification of diabetic retinopathy (DR) from retinal fundus images. The core innovation lies in the AMSFuse block, which combines multi-scale feature fusion with an Integrated Attention Mechanism (IAM) to capture both global and local features. This approach enables the model to focus on critical regions in the images, enhancing classification accuracy across different DR stages. Our experiments on the APTOS2019 and RFMiD2020 datasets demonstrated that AMSFuse outperforms existing state-of-the-art models, achieving higher accuracy, precision, recall, and F1 scores. These results highlight the effectiveness of our multi-scale fusion strategy and attention mechanisms in capturing the complex patterns associated with DR. By improving the accuracy of DR classification, AMSFuse has significant implications for medical image analysis and automated DR screening. It can assist ophthalmologists in early detection and treatment planning, potentially reducing the risk of vision loss among diabetic patients. Its scalability and efficiency make it suitable for integration into clinical workflows, meeting the growing demand for automated screening solutions. For future work, we plan to extend our framework to other retinal diseases, explore additional modalities like optical coherence tomography (OCT) images, and assess the generalizability of AMSFuse in diverse clinical settings.

Acknowledgement: The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

Funding Statement: This work is supported by the National Natural Science Foundation of China (No. 62376287), the International Science and Technology Innovation Joint Base of Machine Vision and Medical Image Processing in Hunan Province (2021CB1013), the Natural Science Foundation of Hunan Province (Nos. 2022JJ30762, 2023JJ70016).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Chengzhang Zhu, Ahmed Alasri; data collection: Tao Xu, Abdulrahman Noman, Raeed Alsabri; analysis and interpretation of results: Ahmed Alasri, Xuanchu Duan, Monir Abdullah, Raeed Alsabri; draft manuscript preparation: Ahmed Alasri, Yalong Xiao, Chengzhang Zhu, Raeed Alsabri; supervision and project administration: Chengzhang Zhu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: In this study, we used public datasets, which can be downloaded from the website if needed.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. International Diabetes Federation. Diabetes facts and figures. IDF Diabetes Atlas. 2023. p. 1–12. Available from: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. [Accessed 2024 Oct 27].
2. American Diabetes Association Professional Practice Committee. 12. Retinopathy, neuropathy, and foot care: standards of medical care in diabetes—2022. Diabetes Care. 2021;45:S185–94.
3. Bhulakshmi D, Rajput S. A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images. PeerJ Comput Sci. 2024;10(11):e1947. doi:10.7717/peerj-cs.1947.
4. Dong K, Zhou C, Ruan Y, Li Y. MobileNetV2 model for image classification. In: 2nd International Conference on Information Technology and Computer Application (ITCA); 2020. p. 476–80.
5. Rocha Da DA, Ferreira FMF, Peixoto ZMA. Diabetic retinopathy classification using VGG16 neural network. Res Biomed Eng. 2022;38(2):761–72. doi:10.1007/s42600-022-00200-8.
6. Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
7. Bao H, Dong L, Piao S, Wei F. BEiT: BERT pre-training of image transformers. arXiv:2106.08254. 2021.
8. Touvron H, Cord M, Jégou H. DeiT III: revenge of the ViT. In: European Conference on Computer Vision; 2022. p. 516–33.
9. Huo X, Sun G, Tian S, Wang Y, Yu L, Long J, et al. HiFuse: hierarchical multi-scale feature fusion network for medical image classification. Biomed Signal Process Control. 2024;87(7660):105534. doi:10.1016/j.bspc.2023.105534.
10. Xu Y, Zhang Q, Zhang J, Tao D. ViTAE: vision transformer advanced by exploring intrinsic inductive bias. Adv Neural Inf Process Syst. 2021;34:28522–35.
11. Fu B, Zhang M, He J, Cao X. StoHisNet: a hybrid multi-classification model with CNN and transformer for gastric pathology images. Comput Methods Programs Biomed. 2022;221(6):106924. doi:10.1016/j.cmpb.2022.106924.
12. Zhang Y, Liu H, Hu Q. TransFuse: fusing transformers and CNNs for medical image segmenta- tion. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021; 2021. p. 14–24.
13. Guo J, Han K, Wu H, Tang Y, Chen X. CMT: convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 12175–85.
14. Peng Z, Huang W, Gu S. Conformer: local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 367–76.
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. USA: Curran Associates, Inc.; 2017. Vol. 30, p. 1–12.
16. Zhang J, Xie B, Wu X, Ram R, Liang D. Classification of diabetic retinopathy severity in fundus images with DenseNet121 and ResNet50. arXiv:2108.08473. 2021.
17. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 32–42.
18. Yang F, Yang H, Fu J. Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 5791–800.

19. Bi Q, Sun X, Yu S, Ma K, Bian C, Ning M, et al. MIL-ViT: a multiple instance vision transformer for fundus image classification. J Vis Commun Image Represent. 2023;97(12):103956. doi:10.1016/j.jvcir.2023.103956.
20. Zhou SK, Greenspan H, Davatzikos C, Duncan J. Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE. 2021;109(5):820–38. doi:10.1109/JPROC.2021.3054390.
21. Li Y, Wu C-Y, Fan H. MViTv2: improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 4804–14.
22. Xia C, Wang X. ViT-CoMer: vision transformer with convolutional multi-scale feature interaction for dense predictions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024. p. 5493–502.
23. Yuan L, Chen Y, Wang T. Tokens-to-token ViT: training vision transformers from scratch on ImageNet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 558–67.
24. Chen F, Ma S, Hao J, Liu W. Dual-path and multi-scale enhanced attention network for retinal diseases classification using ultra-wide-field images. IEEE Access. 2023;11(1106):45405–15. doi:10.1109/ACCESS.2023.3273613.
25. Ryu J, Rehman MU, Nizami IF, Chong KT. SegR-Net: a deep learning framework with multi-scale feature fusion for robust retinal vessel segmentation. Comput Biol Med. 2023;163(2):107132. doi:10.1016/j.compbiomed.2023.107132.
26. Fan R, Liu Y, Zhang R. Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification. Electronics. 2021;10(12):1369. doi:10.3390/electronics10121369.
27. Karthik M, Sohier D. Aptos 2019 blindness detection. 2019. p. 1–12. Available from: https://kaggle.com/competitions/aptos2019-blindness-detection. [Accessed 2024 Oct 27].
28. Pachade S, Porwal P, Thulkar M, Deshmukh G. Retinal fundus multi-disease image dataset (RFMiD): a dataset for multi-disease detection research. Data. 2021;6(2):14. doi:10.3390/data6020014.
29. Rakhlin A. Diabetic retinopathy detection through integration of deep learning classification framework. bioRxiv. 2018;225508. doi:10.1101/225508.
30. Zhu L, Wang X, Ke Z, Zhang W, Lau RW. BiFormer: vision transformer with bi-level routing attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 10323–33.
31. Liu X, Peng H, Zheng N, Yang Q. EfficientViT: memory efficient vision transformer with cascaded group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 14420–30.
32. Gong L, Ma K, Zheng Y. Distractor-aware neuron intrinsic learning for generic 2D medical image classifications. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2020. p. 591–601.
33. Liu S, Gong L, Ma K, Zheng Y. GREEN: a graph residual re-ranking network for grading diabetic retinopathy. In: Medical Image Computing and Computer Assisted Intervention; 2020. p. 585–94.
34. Li K, Wang Y, Zhang J, Liu Y, Li H, Qiao Y. UniFormer: unifying convolution and self-attention for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2023;45(10):12581–600. doi:10.1109/TPAMI.2023.3282631.
35. Xie S, Girshick P, Ross R, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 1492–500.
36. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 10012–22.
37. Bi Q, Qin K, Li Z, Zhang H, Xu K, Xia G-S. A multiple-instance densely-connected con- vnet for aerial scene classification. IEEE Trans Image Process. 2020;29:4911–26. doi:10.1109/TIP.2020.2975718.
38. Bi Q, Yu S, Ji W, Bian C, Gong, Liu. Local-global dual perception based deep multiple instance learning for retinal disease classification. In: Medical image computing and computer assisted intervention–MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science; 2021; Cham: Springer. p. 55–64. doi:10.1007/978-3-030-87237-3_6.
39. Gulshan V, Rajan RP, Widner K, Wu P, Rhodes J. Performance of a deep-learning algorithm *vs*. manual grading for detecting diabetic retinopathy. JAMA Ophthalmol. 2019;137(9):987–93. doi:10.1001/jamaophthalmol.2019.2004.