mathematics

MDPI

Article

FMA-Net: Fusion of Multi-Scale Attention for Grading Cervical Precancerous Lesions

Zhuoran Duan [1,2], Chao Xu [1,2,*], Zhengping Li [1,2], Bo Feng [1,2] and Chao Nie [1,2]

[1] School of Integrated Ciruits, Anhui University, Hefei 230601, China; p21301239@stu.ahu.edu.cn (Z.D.); 04173@ahu.edu.cn (Z.L.); p18101008@stu.ahu.edu.cn (B.F.); wb22101003@stu.ahu.edu.cn (C.N.)
[2] Anhui Engineering Laboratory of Agro-Ecological Big Data, Hefei 230601, China
* Correspondence: xchao@ahu.edu.cn; Tel.: +86-133-3919-9368

Abstract: Cervical cancer, as the fourth most common cancer in women, poses a significant threat to women's health. Vaginal colposcopy examination, as the most cost-effective step in cervical cancer screening, can effectively detect precancerous lesions and prevent their progression into cancer. The size of the lesion areas in the colposcopic images varies, and the characteristics of the lesions are complex and difficult to discern, thus heavily relying on the expertise of the medical professionals. To address these issues, this paper constructs a vaginal colposcopy image dataset, ACIN-3, and proposes a Fusion Multi-scale Attention Network for the detection of cervical precancerous lesions. First, we propose a heterogeneous receptive field convolution module to construct the backbone network, which utilizes combinations of convolutions with different structures to extract multi-scale features from multiple receptive fields and capture features from different-sized regions of the cervix at different levels. Second, we propose an attention fusion module to construct a branch network, which integrates multi-scale features and establishes connections in both the spatial and channel dimensions. Finally, we design a dual-threshold loss function and introduce positive and negative thresholds to improve sample weights and address the issue of data imbalance in the dataset. Multiple experiments are conducted on the ACIN-3 dataset to demonstrate the superior performance of our approach compared to some classical and recent advanced methods. Our method achieves an accuracy of 92.2% in grading and 94.7% in detection, with average AUCs of 0.9862 and 0.9878. Our heatmap illustrates the accuracy of our approach in focusing on the locations of lesions.

Keywords: cervix; precancerous lesions; multi-scale; attention; medical image analysis; deep learning

MSC: 68T07

check for updates

Citation: Duan, Z.; Xu, C.; Li, Z.; Feng, B.; Nie, C. FMA-Net: Fusion of Multi-Scale Attention for Grading Cervical Precancerous Lesions. *Mathematics* **2024**, *12*, 958. https://doi.org/10.3390/math12070958

Academic Editor: Vince Grolmusz

Received: 27 February 2024
Revised: 19 March 2024
Accepted: 21 March 2024
Published: 23 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cervical cancer is the fourth most common cancer among women worldwide, with 604,000 new cases reported in 2020 [1]. Out of the 342,000 deaths caused by cervical cancer, approximately 90% occurred in low- and middle-income countries [1,2]. The World Health Organization (WHO) classifies cervical cancer precursor lesions into two categories: low-grade squamous intraepithelial lesions (LSIL) and high-grade squamous intraepithelial lesions (HSIL). While most low-grade lesions tend to regress spontaneously, high-grade lesions have the potential to progress to invasive cancer. Therefore, early identification of high-grade lesions is crucial for timely intervention and the prevention of cervical cancer. Deep learning techniques can accurately and rapidly detect and classify the severity of lesions from acetic acid test images, aiding in the early detection of high-grade lesions and facilitating timely intervention and treatment. Thus, computer-assisted diagnosis of cervical cancer plays a significant role in effectively preventing the occurrence of cancer, making it of great importance in clinical practice.

At present, cervical cancer screening mainly includes TCT and HPV testing, colposcopy, and cervical biopsy. Colposcopy is a crucial step in the screening process and

is one of the most commonly used endoscopic techniques in gynecology, primarily used for the auxiliary diagnosis and assessment of cervical intraepithelial neoplasia, cervical cancer, and other early lower genital tract lesions. Colposcopy essentially involves using a high-magnification microscope to observe the structures and vascular distribution of the cervix and vagina, among other parts, and using a series of tests to determine the presence of cervical lesions and their possible locations. Importantly, colposcopy is a very economical and widely applicable examination method. In clinical practice, colposcopy has become an important means of cervical cancer screening and early diagnosis. Compared to other expensive screening methods, such as HPV testing, colposcopy is more affordable and can provide accurate results in a shorter period of time. Therefore, in many medical institutions, colposcopy has become the preferred means of cervical cancer screening and early diagnosis.

In colposcopy, the observation of cervical lesions using 3–5% acetic acid staining is the core content of the examination [3]. Acetic acid can induce reversible coagulation of nuclear proteins and cell keratin in cervical epithelial tissues. By observing the degree of white epithelial thickening with acetic acid staining and its regression over time, the location of the lesion can be determined. Generally, the more severe the cervical lesion, the more obvious the whitening change and the longer the duration of whitening, which is usually higher than that of normal epithelial tissue and accompanied by irregular blood vessels. However, the whitening patterns of the lesions are very similar to those of some other diseases, such as psoriasis, Candida vulvovaginitis, and lichen planus. In addition, false-positive areas caused by the reflection of light on the surface of a normal cervix can also pose difficulties in classification.

Our laboratory collaborated with the first affiliated hospital of Wannan Medical College, Yijishan Hospital to collect colposcopy images from precancerous cervical lesions detected by acetic acid tests. The images were annotated by professional physicians based on HPV diagnosis reports, and our colposcopy dataset, named ACIN-3, was constructed. Figure 1 shows a partial display of our dataset.



Figure 1. ACIN-3 dataset presentation.

Since the extent of cervical cancer lesions is the main basis for grading cervical precancerous lesions, the position information of lesion invasion in the epithelial layer is crucial for network classification. In this paper, we propose a novel network called Fusion of Multi-scale Attention Networks (FMA-Net). The network is composed of heterogeneous receptive field convolutional (HRFC) modules, extracting features at different scales in space. Moreover, for different scales of feature information in the branch network, we re-designed an information fusion attention module (IFAM) to strengthen the information exchange between spatial and channel domains. Finally, we devised a dual-threshold loss function, setting positive and negative thresholds as hyperparameters, to address the issue

of sample imbalance and effectively mitigate the impact of labeling errors in samples on the network performance.

Our main contributions are as follows:

- We propose the Fusion of Multi-scale Attention Networks (FMA-Net) for cervical precancerous lesion detection, utilizing colposcopy images from the acetic acid test as the basis for classifying lesions of different degrees.
- Based on the positional characteristics of the lesions, a heterogeneous receptive field convolutional module (HRFC) is proposed to extract features at multiple scales, effectively utilizing spatial information.
- We propose an Information Fusion Attention Module (IFAM) that combines spatial and channel information through block-based hyper/subscale transformations, enhancing the efficiency of feature communication.
- We design a dual-threshold loss function for our dataset with positive and negative probability thresholds (t^+, t^-), addressing the issue of sample imbalance and improving the classification performance of the network.

2. Related Work

In recent years, significant progress has been made in the application of neural networks to medical imaging, becoming important tools in medical image diagnosis, disease prediction, and treatment selection [4]. For example, Dahou et al. [5] presented an alternative skin cancer detection method and achieved excellent performance across multiple datasets. Obayya et al. [6] proposed an automated skin cancer classification and detection model utilizing ODL-SCDC technology, which extends the application scope to the Internet of Things (IoT) environment. Gulshan et al. [7] demonstrated the clinical potential of using deep learning algorithms for the automatic detection of diabetic retinopathy. In segmentation tasks, Ronneberger et al. proposed U-Net [8], a commonly used convolutional neural network structure widely applied in medical image segmentation tasks [9,10]. Rajpurkar et al. [11] combined reinforcement learning with multiple related tasks to achieve automation in chest X-ray pathology detection. Furthermore, neural networks have been applied in various aspects, such as generation and reconstruction in medical imaging [12,13], bringing about many innovations and breakthroughs in the field of medical imaging.

Most studies on the classification of cervical precancerous lesions are based on cervical tissue biopsies [14–16], while the research on colposcopic images is relatively limited. We believe that there are two main reasons for this phenomenon: first, the lack of relevant datasets and difficulties in collecting related images, and second, the similarity between precancerous cervical lesions and other diseases, making it challenging to distinguish them accurately. Currently, deep learning has made some progress in colposcopic images. Tao Zhang et al. [17] first performed ROI extraction and data augmentation preprocessing on colposcopic image data. They proposed a computer-aided diagnosis (CAD) method for the automatic classification of cervical precancerous lesions using a pre-trained dense convolutional neural network. This method was applied to identify CIN2 or higher-grade lesions in cervical images. Luo et al. [18] proposed a deep-learning-based approach that integrates multiple CNN decision features for the classification and diagnosis of cervical lesions. The aforementioned approach, although employed for cervical cancer classification, mostly relies on training and integration within existing frameworks, lacking specific optimizations based on image characteristics.

Attention mechanisms have been widely applied in deep learning in recent years, and significant results have been achieved. By assigning different weights to different parts of the input data, they enable the model to focus more on task-relevant information. The SENet proposed by Hu et al. [19] is a commonly used channel attention module which introduces global average pooling and fully connected layers to learn channel weights and applies them to the classic ResNet [20] model, resulting in significant performance improvements. Woo et al. proposed the CBAM module [21], which combines the channel attention mechanism and spatial attention mechanism to capture both channel correlations and

spatial correlations in images, thereby enhancing the model's performance. Fang et al. [22] proposed a deep reverse residual network for CIN grading based on the improved channel attention using ShuffleNet V1 [23]. Although the current attention mechanism enhances weights, it typically performs weight calculation learning only in a single dimension and is limited to learning single-scale features in one pass, lacking information exchange across multiple dimensions and scales.

3. Materials and Methods

3.1. Network Structure

The texture features and lesion location and size in the aceto-white epithelium are the primary indicators for classification. However, due to the high similarity between the malignant features and normal tissues, as well as other pathological features, the feature extraction capability of the network is highly demanding. Additionally, location information is crucial for distinguishing between low- and high-grade lesions. Therefore, we propose a multi-stage feature extraction network to extract features at different levels.

The network structure is shown in Figure 2, consisting of four main components: the initial feature extraction module, the A backbone network, the B branch network, and the classifier. In the initial feature extraction layer, we replaced the previous 7×7 convolutional layer with a stack of three 3×3 convolutional layers to improve the network's nonlinear feature extraction capability while maintaining the same receptive field. The backbone network is stacked with different numbers of MRE blocks to extract detailed features and location information at different receptive fields. The branch network, which includes downsampling and IRA modules, integrates the feature maps extracted from each stage of the backbone network through attention learning, constructing feature correlations at different scales. Finally, the features learned by the backbone network and the branch network are concatenated along the channel dimension and fed into the classifier for classification.

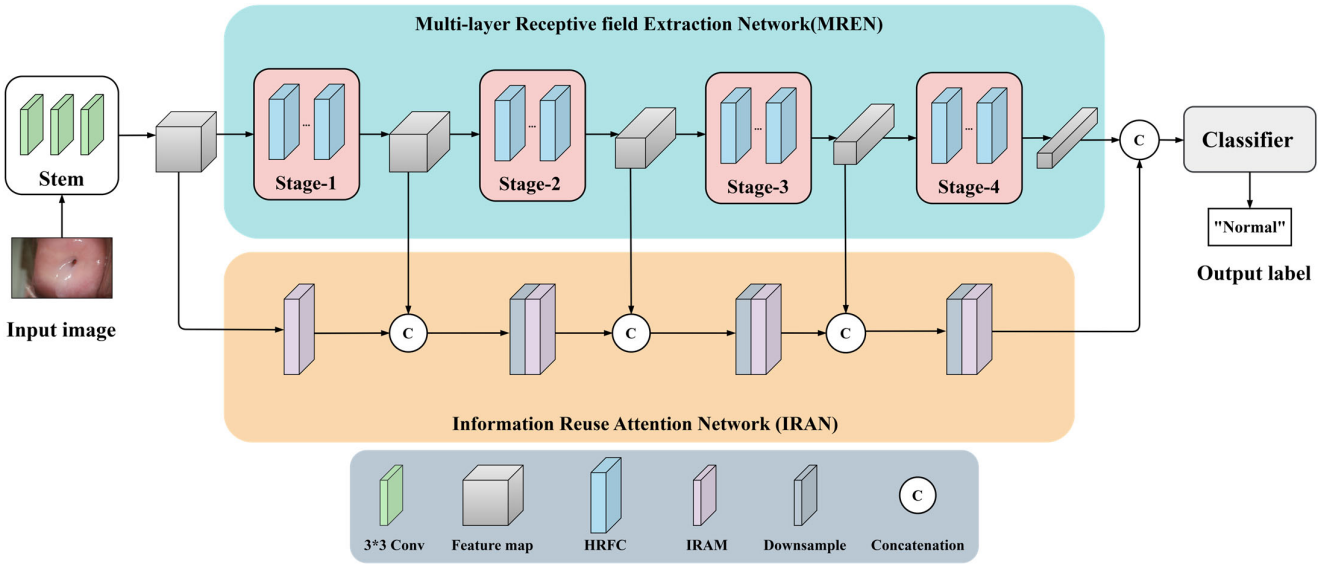


Figure 2. The proposed FMA-Net architecture consists of a multi-scale feature extraction backbone network (MREN) and an information fusion attention network branch network (IFAN).

3.2. Multi-Layer Receptive Filed Extraction Network

Multi-scale feature extraction can help models to comprehensively understand images, as well as improve the accuracy of tasks such as object detection and image classification by detecting and describing object features at different scales. Inspired by Inception [24–27], we designed a heterogeneous receptive field convolution (HRFC) module to better extract correlations between global information. By introducing dilated convolutions, the extraction range of the receptive field was further increased. Dilated convolutions can enlarge the

receptive fields of features without increasing additional parameters. Different convolution combinations for different feature groups contain different sizes of receptive fields. This enables us to extract the positional information of lesion features in different ranges, increases the global information extraction ability, and preserve the extraction of basic features. Our multi-scale feature extraction network consists of four stages, each containing 3, 3, 4, and 3 HRFC modules, with pooling and downsampling operations between each stage.

Our HRFC module structure, as shown in Figure 3, adopts a similar bottleneck structure as ResNet, except that the number of feature groups is set to 4. Finally, the features of different sub-channels are concatenated via channel concatenation.

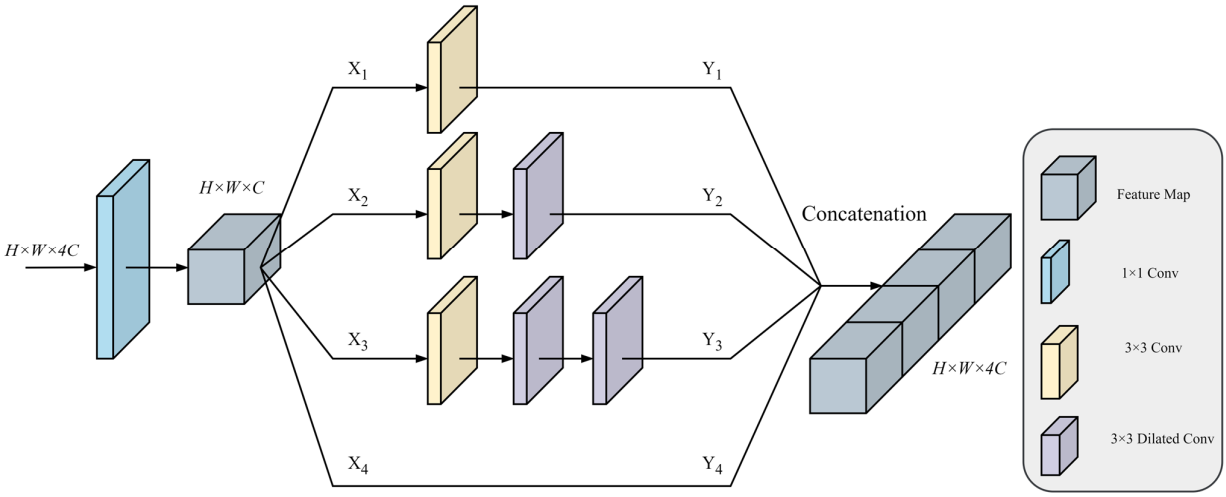


Figure 3. The architecture of the heterogeneous receptive field convolution module (HRFC) consists of four sub-channels, each responsible for extracting detailed features from different receptive fields.

Specifically, the input features are integrated using 1×1 convolutions and divided into 4 groups, denoted as X_1–X_4. The feature sizes of each sub-channel are the same. By using different numbers and sizes of convolutional kernels, the network's non-linear modeling capacity can be improved while reducing the number of parameters. For X_4, no operation is performed, and the resulting feature map is denoted as Y_4. This helps to maintain model accuracy while reducing the risk of overfitting.

For the computation of Y_1, Y_2, and Y_3, unlike previous structures, we first apply a 3×3 convolution operation, denoted as $f_c^{3\times3}$, to X_1, X_2, and X_3 for initial feature extraction. Additionally, for Y_1 and Y_3, dilated convolutions with different strides are applied, denoted as $f_{d1}^{3\times3}$ and $f_{d2}^{3\times3}$, respectively. Therefore, the expression for Y_i is shown as (1).

$$Y_i = \begin{cases} f_c^{3\times3}(X_i), & i = 1; \\ f_{d1}^{3\times3}\left(f_c^{3\times3}(X_i)\right), & i = 2; \\ f_{d2}^{3\times3}\left(f_{d1}^{3\times3}\left(f_c^{3\times3}(X_i)\right)\right), & i = 3; \\ X_i, & i = 4. \end{cases} \tag{1}$$

In the HRFC module, we adopt dilated convolutions to expand the receptive field, which allows Y_2 to have a wider receptive field and establishes broader global information connectivity for Y_3. This is advantageous compared to Inception, as it increases the receptive field of each sub-channel and enhances the non-linear feature extraction capabilities on each channel. Additionally, the global receptive field information is no longer concentrated in a single sub-channel. Our HRFC module can distribute this pressure across two sub-channels, Y_2 and Y_3.

Since dilated convolution is a sparse sampling method, some pixels may not be utilized when multiple atrous convolutions are stacked, leading to a loss of information continuity and correlation and resulting in the gridding effect [28]. Therefore, we use a regular

convolution layer (dilation factor = 1) for initial information extraction. For the design of the dilation factors in $f_{d1}^{3\times3}$ and $f_{d2}^{3\times3}$, we adopt larger dilation factors in stage 1–3 and relatively smaller ones in stage 4.

Specifically, in stage 1–3, the dilation factors for $f_{d1}^{3\times3}$ and $f_{d2}^{3\times3}$ are set to 3 and 5, respectively. In stage 4, the dilation factors are set to 2 and 3. Combined with our initial convolutional layer, the dilation rates of our convolutional kernel are finally set to [1,3,5] and [1–3], respectively, satisfying the condition $M_2 < K$, where M is the maximum distance between two nonzero values, calculated according to Formula (2).

$$M_i = max[M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i] \tag{2}$$

The final size of the receptive field for a sequence of convolution operations completely covers a square region without any holes or missing edges, thus avoiding the gridding effect and information waste.

The receptive field sizes and pixel coverage density of our four channels are illustrated in Figure 4. The more reddish the color, the higher the number of times the corresponding pixel is utilized.

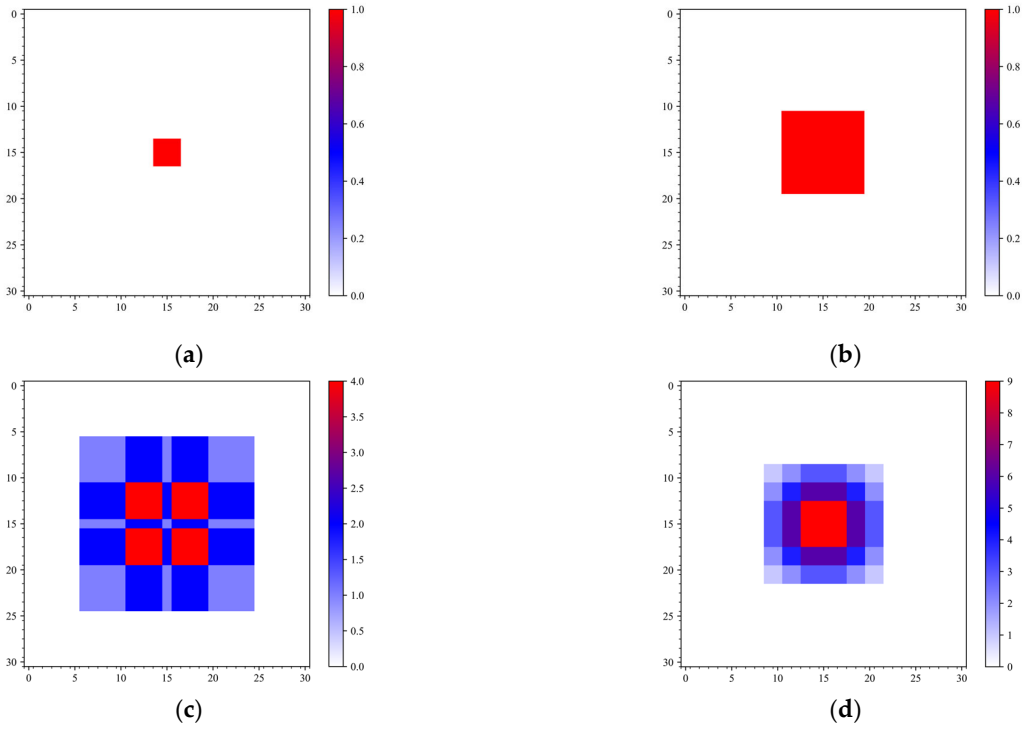


Figure 4. The receptive field sizes and pixel utilization frequency maps of feature maps are presented. (**a**,**b**) correspond to channels 1 and 2. (**c**,**d**) represent stage 1–3 and stage 4 within channel 3.

For the early stages, larger dilation factors can effectively extract global information, while in later stages, smaller dilation factors can make the network focus more on specific texture features as the feature maps become more abstract. This improves the network's utilization of information.

3.3. Information Fusion Attention Network

Traditional attention mechanisms based on convolutional neural networks focus more on analyzing the channel domain and are limited to considering the interaction between feature map channels. CBAM starts from two scopes, channel and spatial, and introduces two analysis dimensions, spatial attention and channel attention, to realize the sequential attention structure from channel to space. Assuming the input features are denoted as

F, the output features of channel attention as F', and the final output features as F'', the calculation is conducted as shown in Equation (3).

$$\begin{cases} F' = M_c(F) \times F \\ F'' = M_s(F') \times F' \end{cases} \tag{3}$$

where $\times$ denotes element-wise multiplication and M_s and M_c represent the channel and spatial attention weight matrices, respectively. During the multiplication process, the weights are assigned to the corresponding feature elements to enhance crucial features and weaken irrelevant ones.

Spatial attention can make the neural network focus more on the pixel regions in the image that have a decisive impact on classification, while ignoring irrelevant areas. Channel attention is used to handle the allocation relationship of feature map channels. Enhancing the attention mechanism for both dimensions improves the performance of the model.

The operations of spatial attention and channel attention do not utilize each other's information, which can cause some information loss. After each stage, the feature maps contain information from four scales. We establish a connection between the four dimensions by stacking pixels and optimize the spatial/channel attention separately to consider both channel and spatial information during the attention process.

The fusion of spatial attention is illustrated in Figure 5. For the input feature map F_1, it is divided into four groups along the channel dimension, corresponding to the four multi-scale features [X_1, X_2, X_3, X_4] output from each stage of our backbone network. These groups are then subjected to a block-based super-dimension transformation, denoted as super-block and represented by R_1. The transformed fusion feature map F_1' is generated.

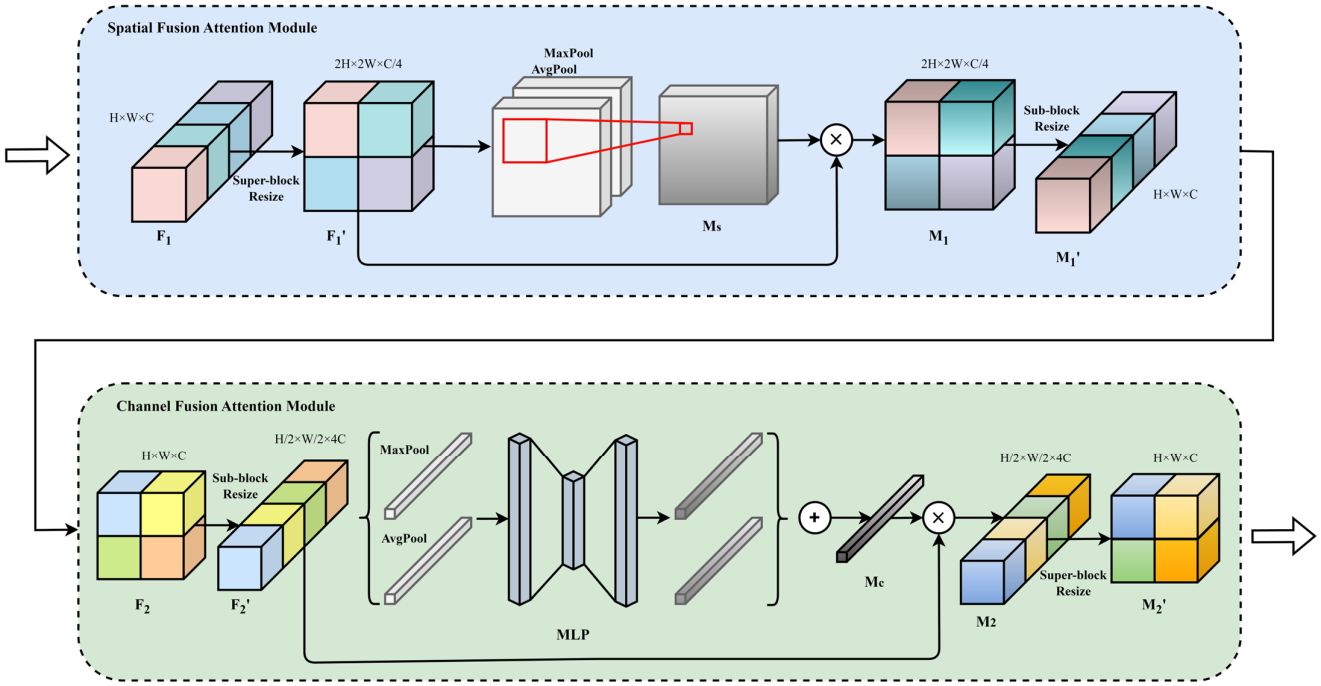


Figure 5. The architecture of the fusion attention module, which consists of a spatial fusion module and a channel-wise attention fusion module.

Subsequently, global average pooling and max pooling operations are applied along the channel dimension to obtain pooled feature maps. These pooled feature maps are then concatenated to generate a feature descriptor matrix, which effectively displays important information regions across multiple scales. Finally, a 7×7 convolutional kernel is employed

to generate the spatial attention mapping matrix M_S, with the specific calculation method shown in Equation (4).

$$M_S = f^{7\times7}\left(\left[AvgPool\left(F_1'\right); MaxPool\left(F_1'\right)\right]\right) \tag{4}$$

where $f^{7\times7}$ represents a convolution operation with a filter size of 7×7.

Finally, the mapping matrix M_S is normalized with the sigmoid function to obtain spatial weights. The fusion feature F_1' is multiplied element-wise with the spatial weights to generate the spatial attention feature M_S. Then, a block-based sub-dimension transformation, denoted as sub-block and represented by R_2, is applied to the spatial attention feature. This operation reduces the dimensionality and subsequently restores it to the original size. The calculation method for spatial fusion attention is described in Equation (5).

$$M_1' = R_2\left(Sigmoid(M_S) \times F_1'\right) \tag{5}$$

The operation of integrating channel attention is illustrated in Figure 5. The input feature F_2 is the output feature M_1' obtained from the spatial attention. The specific operation involves dividing the input feature map into four groups on the spatial dimension and applying a sub-block operation to each group along the channel dimension. The resulting features are stacked along the channel dimension to generate the feature F_2', which contains more spatial-scale information in the channel dimension. Next, global max pooling and global average pooling are applied to F_2' to aggregate spatial information. The pooled feature maps are then fed into a multi-layer perceptron network (MLP) that consists of two fully connected layers. The first layer reduces the number of channels C to $C/2$, while the second layer restores it to C. The outputs of the MLP are added element-wise to generate the channel attention mapping matrix M_C using the specific calculation method described in Equation (6).

$$M_C = MLP\left(AvgPool\left(F_2'\right)\right) + MLP\left(MaxPool\left(F_2'\right)\right) \tag{6}$$

Finally, the two feature maps are stacked along the channel dimension and activated using the sigmoid function. The normalized weights are multiplied element-wise with the feature F_2' to generate the channel attention feature M_2'. Then, the super-block operation is applied to restore the feature to its original size. The calculation method for channel fusion attention is described in Equation (7).

$$M_2' = R_1\left(Sigmoid(M_C) \times F_2'\right) \tag{7}$$

The CNN naturally achieves multi-scale analysis through its built-in hierarchical structure, capturing features at different levels in different layers, which enhances the efficiency and accuracy of the fusion process. We combine the multi-scale and fusion attention modules by concatenating feature maps of different scales along different dimensions. We perform downsampling operations on features of different sizes using 3×3 depthwise separable convolutions to reduce computational complexity while considering spatial and channel information and introducing non-linearity. Finally, the fused features, along with the output features from the backbone network, are fed into the classifier for the final prediction.

The benefits of multi-scale fusion technology are manifold. It improves the quality and resolution of images by integrating information from different scales, leading to richer image details and clearer edge information. This is particularly important in the field of medical imaging analysis. Our designed multi-scale fusion structure fully considers information from multiple scales and works in conjunction with the MRE block and IRA module to extract the most effective features, thereby effectively improving the network's recognition performance.

3.4. Optimization of the Loss Function

Due to the significant imbalance in the number of low-grade squamous intraepithelial lesions samples compared to other samples in our dataset, along with a relatively larger amount of normal samples, there exists an issue of sample quantity imbalance, which affects the improvement of recognition accuracy. In order to alleviate this problem, we introduce two additional hyperparameters, t^+ and t^-, based on the focal loss function [29], setting the positive shift probability P^+ and negative shift probability P^-. This is specified as shown in Formula (8):

$$\begin{cases} P^+ = min(p + t^+, 1) \\ P^- = max(1 - p - t^-, 0) \end{cases} \tag{8}$$

where t^+ and t^- are adjustable hyperparameters, and $\boldsymbol{p}$ represents the predicted probability. By incorporating the shift probability P into the focal loss function, we obtain the double-threshold probability loss (DTL), as shown in Formula (9).

$$DTL = \begin{cases} L_+ = (1 - P^+)^\gamma log(P^+) \\ L_- = (P^-)^\gamma log(1 - P^-) \end{cases} \tag{9}$$

When the probability of positive samples is high, the positive shift probability P^+ applies a threshold to handle positive samples, attenuating their contribution to the loss. When the predicted probability of positive samples is very high, the sample is automatically discarded, and the gradient update direction is focused on the more challenging samples. Similarly, the negative shift probability P^- handles very easy negative samples by applying a threshold. When the probability of negative samples is extremely low, those samples are also discarded. The setting of shift probabilities concentrates the learning direction on more valuable samples, attenuating the influence of a large number of easily learnable normal samples and effectively balancing the impact of different samples on gradients. Additionally, it allows incorrectly labeled samples to be discarded to prevent them from affecting the improvement of network performance.

3.5. Data Source and Processing

The ACIN-3 dataset was provided by The First Affiliated Hospital of Anhui Medical College. To ensure the rigor of the dataset, each image was annotated by professional physicians in conjunction with the corresponding patient's HPV examination report. The dataset comprises 1976 images of normal samples, 1064 images of LISI (CIN1) lesions, and 1423 images of HISI (CIN2+) lesions, totaling 4463 colposcopy images.

Conventional methods, including rotation and flipping, were used to augment the ACIN-3 dataset to verify the model's generalization ability. The ratio of the training set to the validation set was set to 8:2. The specific distribution is shown in Figure 6.

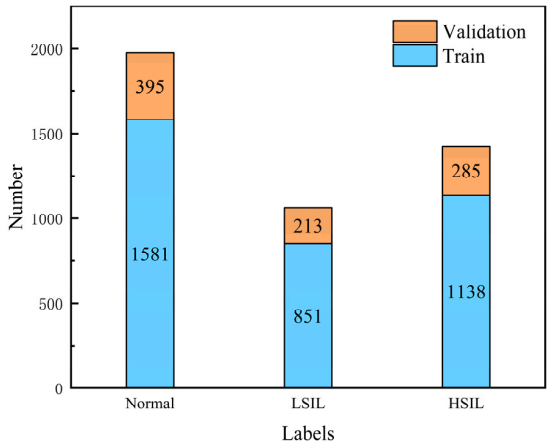


Figure 6. Distribution of the number of datasets trained and validated.

In the data preprocessing stage, we applied the following specific methods and steps to augment the input images:

(1) Resizing the image to 256 × 256;
(2) Randomly cropping the image to 224 × 224;
(3) Randomly rotating the image within a range of [−10, 10] degrees;
(4) Randomly flipping the image with a probability of 50%;
(5) To tensor and normalization.

4. Results

4.1. Experimental Conditions

Our algorithm was implemented in the PyTorch (1.11.0) coding framework. Model training and evaluation were conducted using Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz and NVIDIA RTX A5000 GPU (24 G). All programs were run on Ubuntu 20.04. For training, we employed the AdamW [30] optimizer, which is a stochastic gradient descent optimizer based on Adam, combined with L2 regularization to prevent overfitting. The initial learning rate was set to 0.0001. Training was stopped when the loss on the validation set no longer decreased or started to increase.

4.2. Evaluation Metrics

We evaluated our model based on four fundamental performance metrics: accuracy, precision, recall, and F1 score. Accuracy is the most basic metric, measuring the proportion of correct predictions among all predictions made by the model. Precision refers to the proportion of true positive samples among all predicted positive samples. Recall measures the proportion of true positive samples among all actual positive samples. The F1 score is a comprehensive metric that considers both precision and recall, calculated as the harmonic mean of the two. We calculated these metrics using a confusion matrix, and the specific calculation methods were as follows (10)–(13):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1 - scores = \frac{2TP}{2TP + FP + FN} \tag{13}$$

To thoroughly evaluate the performance of our model, we additionally employed the ROC curve (receiver operating characteristic curve) to assess its performance. The ROC curve displays the relationship between the true positive rate (TPR) and the false positive rate (FPR) of the model at different classification thresholds. The true positive rate (TPR) and false positive rate (FPR) were obtained from Equations (14) and (15).

$$TPR = \frac{TP}{TP + FN} \tag{14}$$

$$FPR = \frac{FP}{FP + TN} \tag{15}$$

The x-axis of the ROC curve represents the false positive rate, which is the proportion of negative samples that are incorrectly predicted as positive. The y-axis represents the true positive rate, which is the proportion of positive samples that are correctly predicted as positive. The area under the ROC curve (AUC) can be used as a quantitative measure to evaluate the performance of the model.

4.3. Experimental Results and Analyis

In order to evaluate and validate the effectiveness of the proposed network structure, we selected some classic classification networks and their corresponding versions, including ResNet34, GoogleNet(Inception V1), Inception V3, DenseNet121 [31], ConvNext V1-Tiny [32], ViT-B/16 [33], and Swin-Tiny [34], as well as advanced cervical cancer classification algorithms in recent years, including Zhang et al., Luo et al., Li et al. [35], and Fang et al. For fairness, we uniformly used our dataset for training and testing and conducted experiments in the same environment. The models were trained using transfer learning and pre-trained on the ImageNet dataset in advance. As shown in Table 1, this paper compared accuracy, precision, recall, and F1 score.

Table 1. The comparative results of accuracy, precision, recall, and F1 score for the CIN grading classification.

Module	Accuracy	Precision	Recall	F1-Score
ResNet34	0.887	0.893	0.931	0.928
GoogleNet	0.898	0.813	0.931	0.929
Inception-v3	0.908	0.841	0.932	0.935
DenseNet121	0.901	0.895	0.920	0.931
ViT-B/16	0.915	0.904	0.944	0.943
ConvNeXt-T	0.912	0.881	0.941	0.939
Zhang et al. [17]	0.901	0.895	0.920	0.931
Luo et al. [18]	0.916	0.897	0.947	0.941
Li et al. [35]	0.914	0.896	0.946	0.939
Fang et al. [22]	0.910	0.891	0.937	0.942
Ours	0.922	0.887	0.954	0.947

As shown in Table 2, our network achieved the highest accuracy of 0.922, recall of 0.954, and F1 score of 0.947 among the classic networks and the networks proposed for precancerous cervical lesion classification in recent years. In terms of precision, the vision transformer achieved the best result of 0.904.

Table 2. The comparative results of accuracy, precision, recall, and F1 score for recognizing CIN2+.

Module	Accuracy	Precision	Recall	F1-Score
ResNet34	0.914	0.946	0.936	0.906
GoogleNet	0.914	0.813	0.931	0.929
Inception-v3	0.923	0.935	0.918	0.914
DenseNet121	0.928	0.937	0.922	0.919
ViT-B/16	0.943	0.950	0.937	0.936
ConvNeXt-T	0.927	0.939	0.923	0.918
Zhang et al. [17]	0.934	0.948	0.935	0.927
Luo et al. [18]	0.944	0.849	0.936	0.937
Li et al. [35]	0.940	0.949	0.936	0.932
Fang et al. [22]	0.939	0.961	0.953	0.932
Ours	0.947	0.954	0.942	0.940

We also included an additional binary dataset by combining the normal and LSIL data to identify HSIL (CIN2+) or higher-grade cervical lesions. The experimental results are presented in Table 2.

In addition, to comprehensively evaluate our model, we plotted the ROC curve of the model as shown in Figure 7 and calculated the AUC (area under the curve). The AUC is calculated based on the ROC curve of the classification model at different decision thresholds. In the case of class imbalance, the model may be more prone to predicting the majority class. However, the AUC metric has a certain robustness against class imbalance issues, as it can comprehensively consider the true positive rate and false positive rate

without being affected by the imbalanced class distribution. Therefore, it can better evaluate the model's performance.

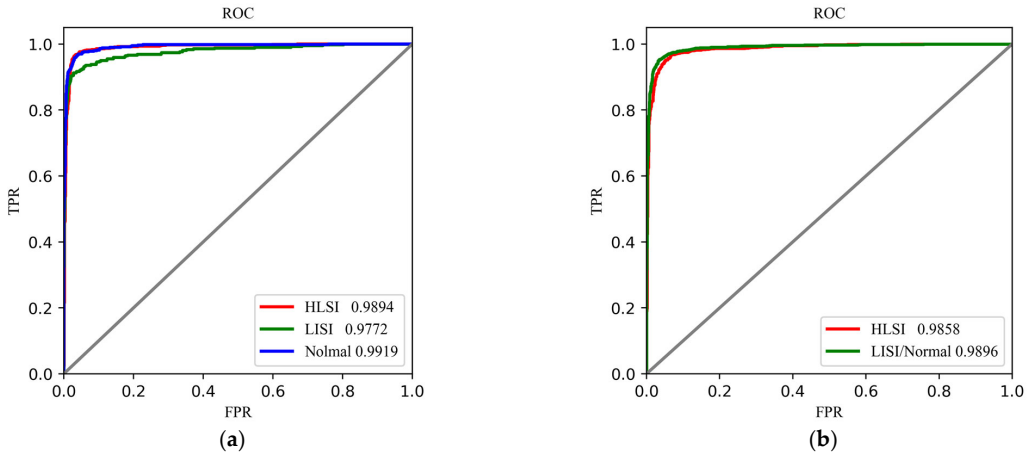


Figure 7. ROC curve generated by our network. (**a**) The ROC curve for the triple classification, with an average AUC of 0.9862. (**b**) The ROC curve for the binary classification, with an average AUC of 0.9878.

The results indicate that the ROC curves of all three labels are located in the upper-left corner of the figure, with particularly outstanding performance in the recognition of normal and HISI. The detection task also showed excellent performance, with all metrics exceeding 0.98. These findings suggest that FMA-Net exhibits good overall classification performance and performs well in predicting all labels.

We calculated the results of the AUC and compared them with other models. As shown in Table 3 our method demonstrated excellent performance in both classification and detection tasks, achieving improvements of 1.9% and 1.5% over the best-performing methods, respectively.

Table 3. The comparative results of the average AUC for the triple classification and binary classification are as follows.

Method	AUC (Triple Classification)	AUC (Binary Classification)
ResNet34	0.9064	0.9227
GoogleNet	0.9152	0.9240
Inception-v3	0.9376	0.9532
DenseNet121	0.9206	0.9443
ViT-B/16	0.9394	0.9598
ConvNeXt-T	0.9375	0.9521
Zhang et al. [17]	0.9253	0.9542
Luo et al. [18]	0.9672	0.9726
Li et al. [35]	0.9558	0.9674
Fang et al. [22]	0.9394	0.9542
Ours	0.9862	0.9878

4.4. Ablation Experimental Results and Analysis

4.4.1. Module Validation and Analysis

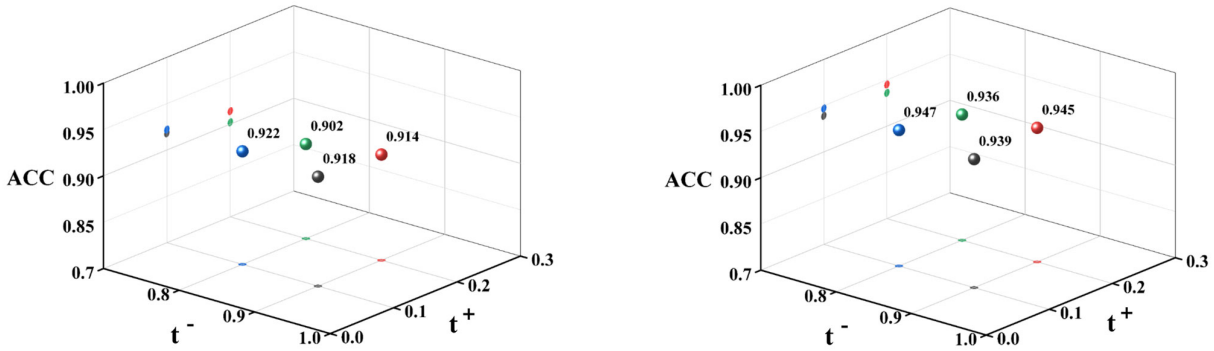
To validate the effectiveness of the module, we sequentially utilized one to two modules. Specifically, the base network replaced the heterogeneous receptive field convolution module with the structure of Inception v1, removed the fusion attention branch network, and replaced the double-threshold loss function with the cross-entropy loss function. The results of the metrics are presented in Table 4.

Table 4. Experimental results of different module combinations. (HRFC: heterogeneous receptive field convolution module. IRAM: information fusion attention module. Total: HRFC + IRAM + DFT).

Module	Triple Classification					Binary Classification				
	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score	AUC
Base Network	0.910	0.893	0.934	0.932	0.953	0.932	0.941	0.938	0.921	0.954
HRFC	0.916	0.897	0.947	0.941	0.973	0.937	0.954	0.944	0.930	0.967
IRAM	0.915	0.904	0.944	0.943	0.969	0.934	0.949	0.936	0.927	0.961
HRFC + IRAM	0.921	0.894	0.950	0.943	0.981	0.944	0.943	0.927	0.936	0.979
Total	0.922	0.887	0.954	0.947	0.986	0.947	0.954	0.942	0.940	0.988

4.4.2. Hyperparametric Experimental Analysis

For the two hyperparameters, t^+ and t^-, in the double-threshold probability loss function, we set four combinations of thresholds: [0.1, 0.9], [0.1, 0.8], [0.2, 0.9], and [0.2, 0.8]. The accuracy for each combination of thresholds is shown in Figure 8.



(**a**) (**b**)

Figure 8. Experimental results on hyperparameter accuracy. (**a**) The results of the triple classification. (**b**) The results of the binary classification.

The results indicate that the highest accuracy was achieved when the thresholds were set to [0.1, 0.8]. We believe that a relatively smaller positive probability threshold can effectively weaken the influence of some easily classified positive samples, whereas increasing this threshold would have a counterproductive effect. For negative samples, a relatively loose threshold was set, such that when the probability exceeded 0.8, it was considered an easily learnable negative sample. This allowed the network to focus more on the challenging aspects of negative sample classification and dynamically adjust its learning direction, thereby improving the overall classification capability of the network.

4.5. Visualization of Experimental Analysis

We employed the Grad-CAM [36] method to generate heatmaps highlighting the regions of focus for the network. Grad-CAM enables backpropagation through the model and computes gradient information for the feature maps. By weighting the channels of the feature maps using this gradient information, heatmaps are generated to depict the regions of interest for a specific class in colposcopic images. To validate the accuracy of the model's identification, we compared the generated heatmaps with lesion annotation maps provided by expert doctors in the dataset, as shown in Figure 9.

As shown in Figure 9, it can be observed that the corresponding lesion regions are activated, and the prominently highlighted areas in red represent the most attended parts in the model, which serve as the primary basis for diagnosis. Visually, the activated regions in the image align with the areas annotated by expert doctors. The highly concentrated red areas, which can be quickly identified by the naked eye, are indicative of the occurrence of lesions. Additionally, atypical vessels or visible small dot-like vessels are important

indicators for HISI diagnosis, as their distribution exhibits diversity. Our network also demonstrates attention towards these features in the generated heatmaps. In summary, even though the sizes and abnormal regions of the lesions may appear different, the FMA-Net is still capable of accurate identification.

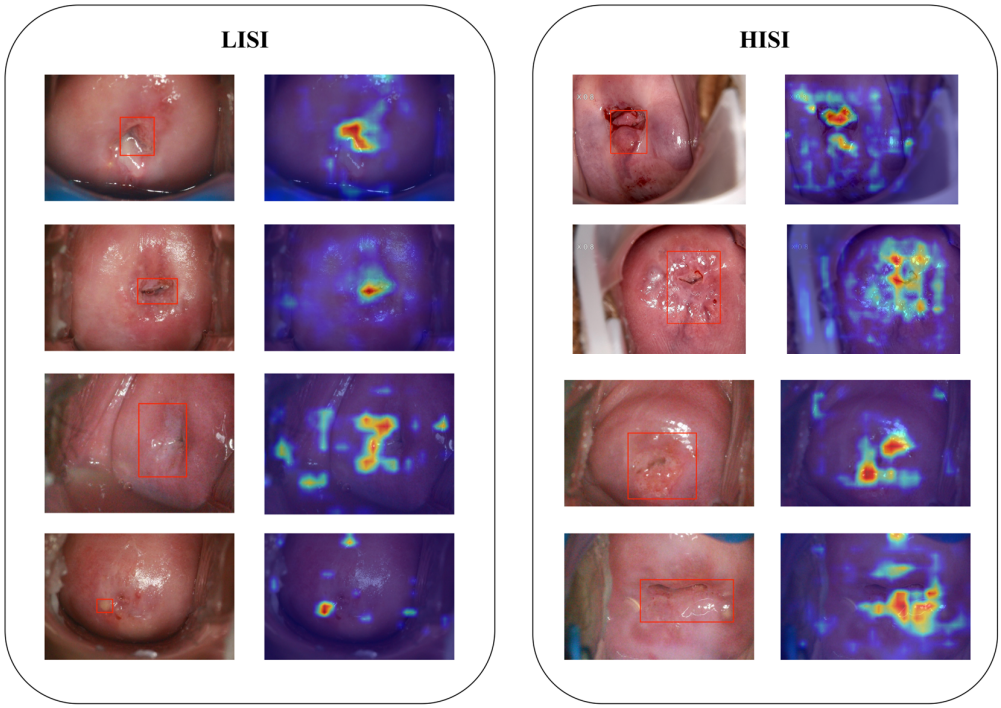


Figure 9. An example of a heatmap generated by FMA-Net is shown. The left image was manually annotated by a professional doctor, marking the basic factual boundary boxes of possible lesion regions. It can be observed that the network is very accurate in identifying lesions.

5. Discussion

Hysteroscopy, as the most economical and simple step in cervical cancer screening, possesses universal applicability. It is particularly beneficial for resource-constrained regions or economically disadvantaged areas, providing more convenient cervical cancer screening services and playing an important role in the early detection and treatment of cervical cancer. Currently, most existing methods adopt mature classification network architectures, which do not fully exploit the characteristics of cervical cancer lesions. We propose a heterogeneous receptive field convolutional module that utilizes convolutions with different dilation factors to learn from multiple receptive field features, taking into account the variable spatial characteristics of the lesion regions. Additionally, in conjunction with a fusion attention module, multiple receptive field features are subjected to attentional fusion learning, enabling a global perspective on the location information of the lesions. Addressing the issue of dataset imbalance, we propose a dual-threshold loss function to adjust the weights of positive and negative samples. As shown in Table 4, ablative experiments corroborate the effectiveness of the heterogeneous receptive field convolutional module, information fusion attention module, and dual-threshold loss function. The standalone fusion attention module exhibits a relatively marginal enhancement, whereas the combined use of the heterogeneous receptive field convolutional module and information fusion attention module yields a more significant improvement.

The proposed FMA-Net was comprehensively compared with classical classification networks and more advanced cervical cancer classification networks in our ACIN-3 dataset, yielding outstanding results. We conducted comparisons on both three-class classification for cervical cancer grading and two-class classification for detecting CIN2+, as shown in Table 1. The proposed FMA-Net achieved accuracy rates of 92.2% and 94.7% for the

three-class and two-class classifications, respectively, showing improvements of 0.6% and 0.3%. Additionally, we compared the AUC values, as shown in Table 1, which indicated that the average AUCs of FMA-Net for the three-class and two-class classifications were 98.62% and 98.78%, respectively, demonstrating improvements of 1.9% and 1.5%. These results confirm the effectiveness and generalizability of our model in discriminating precancerous cervical lesions.

The present study has certain limitations. Firstly, the dataset used in this research is relatively small. In future studies, we intend to continuously collect more data, including samples of cancer, to enhance the applicability of our network for the more accurate diagnosis of cervical cancer. Additionally, the vaginal colposcopy images contain some noise and exposure issues, especially in terms of exposure. These are inevitable during the acquisition of vaginal colposcopy images and can be observed to have a certain interference on the network model through heat map analysis. Therefore, it would be beneficial to employ image enhancement techniques to assist the network in classification. Lastly, the current annotation information of the dataset is relatively limited. In fact, the acetic acid test images in colposcopy examinations are continuous and dynamic. The timing of image acquisition at different time points may impact the final diagnosis. Therefore, it is suggested to introduce secondary label annotations for recording the timing of image capture and other related information in order to enrich the sample data input for deep learning networks.

Our model can also be applied to other medical classification domains, for instance, the classification and detection of lung CT scans and brain diseases, where the model methodology can be optimized according to different diseases to adapt to various tasks. In conclusion, our model exhibits broad potential for application in the field of medical classification, providing a powerful tool for healthcare professionals and researchers to improve the accuracy and efficiency of disease diagnosis and treatment.

6. Conclusions

In this paper, we constructed a vaginal colposcopy image dataset, ACIN-3, for the classification of precancerous cervical lesions. Meanwhile, we proposed a network, FMA-Net, for the classification of precancerous cervical lesions. FMA-Net was designed to determine the lesion grade from vaginal colposcopy images, and a multi-scale feature extraction network was introduced to extract features from regions with different sizes to account for the location characteristics of cervical precancerous lesions. By introducing a heterogeneous receptive field convolution module, the network aggregates multi-scale features. Additionally, the fusion attention module was introduced to establish attentional correlations on multi-scale features to help the network perceive regions of interest from multiple layers of features and learn relevant details regarding cervical cancer. Extensive qualitative and quantitative experiments demonstrated that FMA-Net outperforms the current state-of-the-art cervical precancer classification methods and some classical networks in the tasks of cervical precancer grading and detection, indicating that it is an efficient cervical cancer grading network.

Author Contributions: Conceptualization, Z.D. and C.X.; methodology, Z.D.; software, Z.D.; validation, Z.L., B.F. and C.N.; data curation, B.F.; writing—original draft preparation, Z.D.; writing—review and editing, C.X. and Z.L.; visualization, Z.D.; supervision, C.X.; project administration, C.X.; funding acquisition, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key Research and Development Program of China (No. 2019YFC0117800).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author (accurately indicate status).

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

References

1. WHO. Cervical Cancer. Available online: https://www.who.int/health-topics/cervical-cancer#tab=tab_1 (accessed on 1 January 2024).
2. Kornovski, Y.; Slavchev, S.; Kostov, S.; Ivanova, Y.; Yordanov, A. Precancerous lesions of the cervix—Aetiology, classification, diagnosis, prevention. *Oncol. Clin. Pract.* **2021**, *17*, 271–276. [CrossRef]
3. Mohamad, K.A.A.; Saad, A.S.; Murad, A.W.A.; Altraigy, A. Visual inspection after acetic acid (via) as an alternative screening tool for cancer cervix. *Apollo Med.* **2016**, *13*, 204–207. [CrossRef]
4. Kotei, E.; Thirunavukarasu, R. Visual attention condenser model for multiple disease detection from heterogeneous medical image modalities. *Multimed. Tools Appl.* **2023**, *10*, 30563–30585. [CrossRef]
5. Dahou, A.; Aseeri, A.O.; Mabrouk, A.; Ibranhim, R.A.; Al-Betar, M.A.; Elaziz, M.A. Optimal Skin Cancer Detection Model Using Transfer Learning and Dynamic-Opposite Hunger Games Search. *Diagnostics* **2023**, *13*, 1579. [CrossRef]
6. Obayya, M.; Arasi, M.A.; Almalki, N.S.; Alotaibi, S.S.; Al Sadig, M.; Sayed, A. Internet of Things-Assisted Smart Skin Cancer Detection Using Metaheuristics with Deep Learning Model. *Cancers* **2023**, *15*, 5016. [CrossRef]
7. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef] [PubMed]
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015.
9. Roth, H.R.; Lu, L.; Liu, J.M.; Yao, J.H.; Seff, A.; Cherry, K.; Kim, L.; Summers, R.M. Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation. *IEEE Trans. Med. Imaging* **2016**, *35*, 1170–1181. [CrossRef]
10. Dou, Q.; Chen, H.; Yu, L.Q.; Zhao, L.; Qin, J.; Wang, D.F.; Mok, V.C.T.; Shi, L.; Heng, P.A. Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks. *IEEE Trans. Med. Imaging* **2016**, *35*, 1182–1195. [CrossRef]
11. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
12. Yang, G.; Yu, S.; Dong, H.; Greg, S.; Pier, L.; Ye, X.; Liu, F.; Simon, A.; Jennifer, K.; Guo, Y.; et al. DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *IEEE Trans. Med. Imaging* **2018**, *37*, 1310–1321. [CrossRef]
13. Qian, Z.; Tang, H.; Gong, G.Z.; Yin, Y.; Huang, C.; Fan, W. CorGAN: Context aware Recurrent Generative Adversarial Network for Medical Image Generation. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM), Online, 16–19 December 2020.
14. Sholik, M.; Fatichah, C.; Amaliah, B. Classification of Cervical Cell Images into Healthy or Cancer Using Convolution Neural Network and Linear Discriminant Analysis. In Proceedings of the 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), Bali, Indonesia, 13–15 July 2023.
15. An, H.M.; Ding, L.Y.; Ma, M.Y.; Huang, A.H.; Gan, Y.; Sheng, D.L.; Jiang, Z.N.; Zhang, X. Deep Learning-Based Recognition of Cervical Squamous Interepithelial Lesions. *Diagnostics* **2023**, *13*, 1720. [CrossRef] [PubMed]
16. Tang, J.; Zhang, T.; Gong, Z.; Huang, X. High Precision Cervical Precancerous Lesion Classification Method Based on ConvNeXt. *Bioengineering* **2023**, *10*, 1424. [CrossRef] [PubMed]
17. Zhang, T.; Luo, Y.; Li, P.; Liu, P.; Du, Y.; Sun, P.; Dong, B.; Xue, H. Cervical precancerous lesions classification using pre-trained densely connected convolutional networks with colposcopy images. *Biomed. Signal Process. Control* **2020**, *55*, 101566. [CrossRef]
18. Luo, Y.; Zhang, T.; Li, P.; Liu, P.; Sun, P.; Dong, B.; Ruan, G. MDFI: Multi-CNN Decision Feature Integration for Diagnosis of Cervical Precancerous Lesions. *IEEE Access* **2020**, *8*, 29616–29626. [CrossRef]
19. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E.H. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016.
21. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
22. Fang, S.; Yang, J.; Wang, M.; Liu, C.; Liu, S. An Improved Image Classification Method for Cervical Precancerous Lesions Based on ShuffleNet. *Comput. Intell. Neurosci.* **2022**, *2022*, 9675628. [CrossRef]
23. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
24. Szegedy, C.; Liu, W.; Jia, Y.Q.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
25. Loffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015.
26. Szegedy, C.; Vanhoucke, V.; Loffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016.

27. Szegedy, C.; Loffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017.
28. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
29. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
30. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**, arXiv:1711.05101.
31. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
32. Liu, Z.; Mao, H.Z.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S.N. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
34. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.X.; Zhang, Z.; Lin, S.; Guo, B.N. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Online, 11–17 October 2021.
35. Li, Y.X.; Liu, Z.H.; Xue, P.; Chen, J.W.; Ma, K.; Qian, T.Y.; Zheng, Y.F.; Qiao, Y.L. GRAND: A large-scale dataset and benchmark for cervical intraepithelial Neoplasia grading with fine-grained lesion description. *Med. Image Anal.* **2021**, *70*, 102006. [CrossRef] [PubMed]
36. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]

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