Lightweight High Precision Sensitive Image Recognition Method Based on Improved Runge Kutta and Feature Guidance Fusion

Shuaina Huang [*,**,***], **Zhiyong Zhang** [*,**,***], **Song Bin** [*,**,***]

[*] *Information Engineering College,Henan University of Science and Technology, Luoyang, 471023, Henan, China(e-mail: 205400000024 @stu.haust.edu.cn, xidianzzy@126.com, songbin@haust.edu.cn)*
[**] *Henan International Joint Laboratory of Cyberspace Security Applications,Henan University of Science and Technology, Luoyang, 471023, Henan, China*
[***] *Henan Intelligent Manufacturing Big Data Development Innovation Laboratory, Henan University of Science and Technology, Luoyang, 471023, Henan, China*

Abstract: The dissemination of sensitive content like pornography and violence on social networks adversely affects the physical and mental well-being of young individuals and societal order. Identifying and categorizing these sensitive images accurately is crucial for public safety. However, challenges like varying lighting conditions, significant changes in sensitive areas, and background interference often hinder image recognition effectiveness. Additionally, existing methods often neglect computational efficiency. The current study proposes a novel approach named Improved Runge Kutta and Feature Guidance Fusion (IRFGN-Net) for sensitive image recognition in order to address these issues. IRFGN-Net aims to balance computational efficiency and classification accuracy. It introduces an enhanced Runge Kutta method to extract more detailed features, improving the discrimination of key characteristics while minimizing noise and irrelevant background impact. Furthermore, a novel feature guidance fusion module integrates texture and depth features of sensitive images, reducing illumination and noise influence on recognition accuracy. Comprehensive experiments validated the proposed method's effectiveness. Initially, IRFGN-Net was compared with several widely used convolutional neural network models. Subsequently, ablation experiments demonstrated the efficacy of the improved Runge Kutta method and feature guidance fusion module. The experimental results unequivocally show the superior performance of IRFGN-Net in sensitive image recognition tasks. It significantly improves recognition accuracy and has a strong inference speed.

Keywords: Sensitive Image Recognition, Lightweight, Runge Kuta, Attention, Feature Guidance Fusion

1. INTRODUCTION

The rapid development of social networks in recent years has led to a significant increase in sensitive content such as pornography and violence (Steely Smith and Ten-Bensel, 2024). Sensitive information seriously affects young people's physical and mental health as well as social security, creating significant regulatory pressure (Camilleri et al., 2021). Most existing sensitive image recognition methods rely primarily on manual design of behavioral features. By creating local or global feature descriptors, these methods extract temporal and spatial features of behaviors, and then uses support vector machines or random forest models to classify behaviors after removing redundant information. The recognition accuracy of these methods is low and often requires complex preprocessing and substantial manpower. Furthermore, these methods are easily disturbed by image noise, resolution and image background, making real-time and efficient detection challenging. Hence, efficiently and stably identifying sensitive images is an urgent issue that requires attention.

Traditional sensitive image recognition methods typically involve feature extraction followed by classification using handcrafted feature extractors such as SIFT and HOG, along with linear classifiers like SVM and logistic regression (Huang and Kong, 2016; Ulges and Stahl, 2011; Yang et al., 2004). These methods often require significant manual intervention and tuning, resulting in limited performance. On the other hand, deep learning-based approaches have the ability to automatically learn high-level feature representations and classifiers, resulting in more accurate and efficient recognition of sensitive image (Tran et al.,

* The work is supported by National Natural Science Foundation of China Grant No.61972133, Project of Leading Talents in Science and Technology Innovation in Henan Province Grant No.204200510021, Program for Henan Province Key Science and Technology under Grant No.222102210177, 242102211077 and Henan Province University Key Scientific Research Project under Grant No.23A520008.
Corresponding author: Zhiyong Zhang, e-mail: xidianzzy@126.com

2019). Deep learning models that are rountinely adopted include convolutional neural networks (CNNs) and recurrent neural networks (RNNs). They are characterised by various optimization techniques, such as model architecture optimization and feature fusion to enhance model performance and robustness (Xu et al., 2020). Despite recent advancements in image classification techniques, such as MobileNet (Torky et al., 2023), ShuffleNet (Zhang et al., 2023), SqueezeNet (Hu and Hu, 2023), and EfficientNet (Hu et al., 2023), applying them to sensitive image recognition remains challenging. These methods can be useful in some specific scenarios, but they suffer from low feature utilization and are susceptible to background interference, especially in real-time content filtering scenarios. These shortcomings reduce the model's generalization ability and consequently affect its reliability.

Given the pressing nature of the aforementioned issues, this study aims to explore novel approaches to address the trade-off between inference speed and detection accuracy in sensitive image recognition. This study proposes a convolutional neural network based on improved Runge Kutta and feature guidance fusion (IRFGN-Net). The improved Runge Kutta method introduces second-order Runge Kutta into the deep neural network and optimizes its channel and spatial dimensions. The recognition ability of sensitive features is enhanced, and the influence of noise and irrelevant background on sensitive images is reduced. The feature guidance fusion module combines texture features and depth features of sensitive images, which is able to effectively mitigate the negative influence of light and noise on the accuracy recognition of sensitive images. In addition, we have introduced depthwise separable convolution and ghost module into IRFGN-Net for more lightweight. Not only does the IRFGN-Net improves the model's ability to extract features but also demonstrates remarkable trainability, potentially leading to breakthroughs in lightweight image recognition tasks. The key contributions of the current study can be outlined as follows:

1) A novel lightweight end-to-end deep-learning framework called IRFGN-Net is proposed to improve sensitive image recognition accuracy. IRFGN-Net can obtain more abundant feature information under the condition of lower network layers, and shows remarkable accuracy and lightweight.
2) An improved Runge Kutta algorithm is designed, which combines deep and shallow features to reduce information loss by optimizing the numerical solutions of ordinary differential equations. In the meantime, the feature map's attention weight directs the model to prioritize sensitive features and minimize background interference.
3) The feature guidance fusion module, combining texture and depth features, reduces the influence of light and noise on recognition accuracy, providing more informative model inputs.

The structure of this paper is as follows: Section 2 provides the background on sensitive image recognition and lightweight neural networks. Section 3 describes the proposed method in detail. In Section 4, a comparative assessment of the proposed method with state-of-the-art techniques and the conduct of ablation experiments are presented. Finally, Section 5 contains concluding remarks of the study.

2. RELATED WORKS

The core technologies for traditional sensitive image recognition include skin area detection, texture detection, global or local feature detection, pose detection, visual words, and classification feature vectors (Ries and Lienhart, 2014). Jones et al. (Jones and Rehg, 2002) conducted in-depth research on statistical skin area models to detect inappropriate images primarily based on skin color information. Zeng et al. (Zeng et al., 2004) extracted various features from color and texture information and used SVM learning algorithms to train sensitive image classifiers. Lv et al. (Lv et al., 2011) proposed an improved Bag-of-Visual-Words (BoVW) model that filters pornographic images using advanced semantic features. However, these methods have high false positive rates and poor generalization ability. They also require significant time to adjust the parameters of the feature extractor. Chen and Hauptmann (2009) introduced optical flow histograms utilizing the scale-invariant feature transform (SIFT) and introduced a motion scale-invariant space transform (MoST) for local feature description of sensitive behaviors by extracting temporal and spatial characteristics. Building upon this work, Nievas et al. (Bermejo Nievas et al., 2011) utilized MoSIFT to extract appearance and motion features for sensitive behavior recognition. These methods have simple feature extraction processes suitable for small-scale datasets but require manual selection and design of features, which are often time-consuming for large-scale sensitive image recognition.

As deep learning advances (Boulkaboul et al., 2024; Liu and Zhang, 2024), an growing number of algorithms based on convolutional neural networks (CNNs) have been developed for detecting sensitive images. The advantage of CNNs is that they do not require manual extraction of image features, thus avoiding poor recognition results caused by improper feature selection. Huang and Kong (2016) proposed the fusion of multi-task convolutional neural networks to identify pornographic images by detecting specific sensitive areas. However, this method is not effective for the recognition of images with small sensitive parts or complex backgrounds. Connie et al. (2017) constructed eight different structures of convolutional neural networks for image classification and then used least squares regression to calculate their respective weights before merging them to obtain the final recognition result. Nevertheless, this approach performs poorly in identifying images with hidden sensitive features. Moustafa (2015) combined AlexNet and GoogleNet models for image classification identify sensitive images.

However, recent research lacks the recognition and detection of multi-class sensitive images. Moreover, the aforementioned studies focus solely on improving accuracy for identifying sensitive images while ignoring deployment requirements in real-world environments. More model parameters lead to longer training and loading times. Therefore, using lightweight CNNs would better meet the requirements in practical applications.

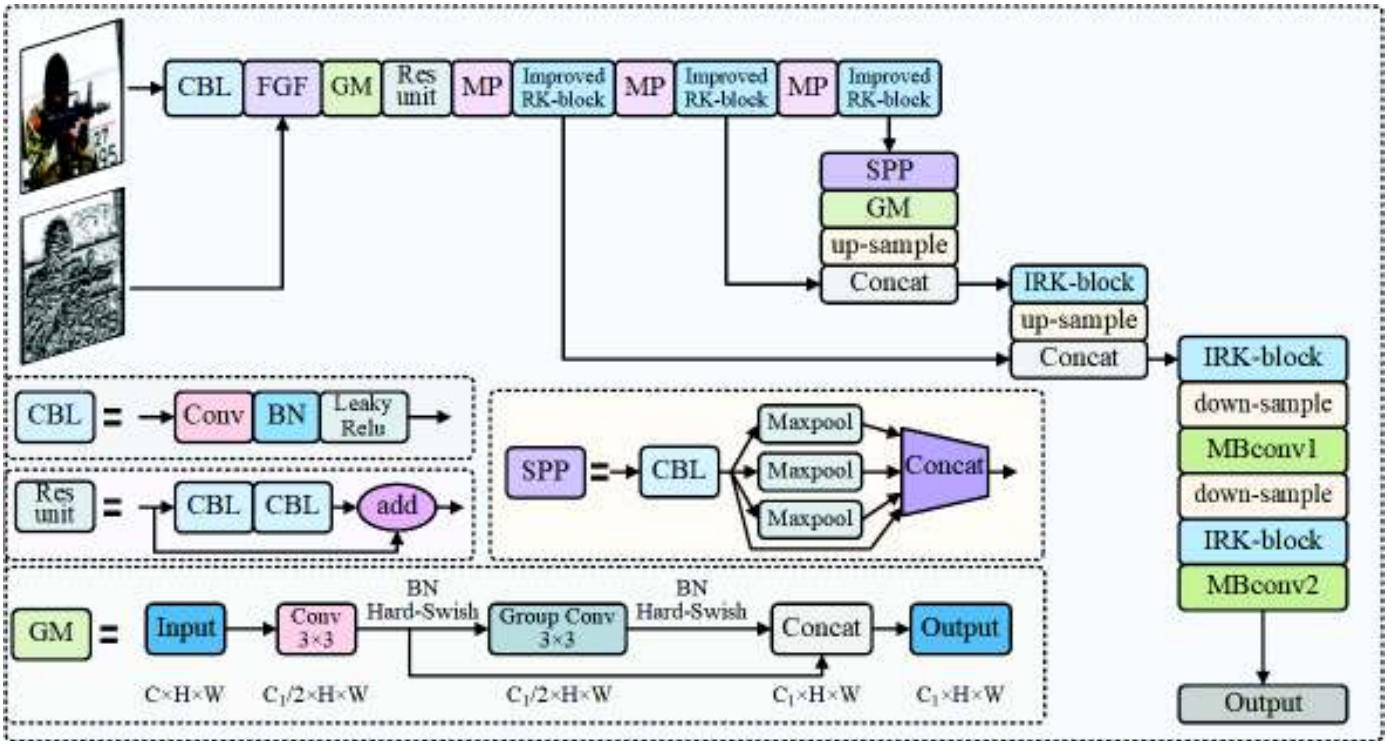

Fig. 1. Overall structure block diagram of the model

In various computer vision research fields, CNNs have been extensively adopted (Rezaee et al., 2021), but the traditional CNN models have high demands for computing and memory resources, making them difficult to deploy on mobile phones, embedded systems and IoT devices. Model lightweight research (Gugulothu and Balaji, 2023) refers to the design and development of neural network with smaller size and more compact model structures, which have lower computational requirements, smaller size, and lower memory footprint than traditional deep learning networks. The goal is to implement deep learning models on devices with restricted computational processing power and small memory, such as intelligent surveillance cameras, drones, medical and old-age equipment, etc. Researchers focus on reducing parameters and model inference time, leading to the emergence of many lightweight models.

MobileNet (Howard et al., 2017; Sandler et al., 2018) employs depthwise separable convolutions to construct neural networks, incorporating a series of lightweight enhancements. MobileNet employs depthwise separable convolution kernels to reduce computational demands and uses 1x1 convolution kernels to reduce channel dimensions, resulting in a substantial reduction in model parameters and computational load. MobileNet excels in compact size and rapid inference, making it well-suited for real-time image classification on embedded and mobile devices. EfficientNet (Kurt et al., 2023) represents a family of lightweight CNN models that leverage automatic network scaling to achieve optimal performance within specified computational resource constraints. It achieves this by balanced scaling across channels, depths, and resolutions. EfficientNet (Kurt et al., 2023) has demonstrated impressive efficacy in lightweight image classification tasks, delivering both high accuracy and a compact size. Nonetheless, balancing inference speed and recognition accuracy in sensitive image recognition tasks remains a persistent issue, necessitating the exploration of innovative solutions.

These lightweight methods bring new insights to the research of sensitive image recognition. Most current research on sensitive image recognition uses features that are not rich enough or are easily affected by light, sensitive locations, and background. Additionally, most studies build models based on networks such as ResNet and VGGNet, and the number of parameters in these models greatly affects the practical use. Finally, current models may lose information during feature extraction. In summary, current research has shortcomings in model structure and data sources, which need to be further improved to achieve higher accuracy and lighter model weight.

3. PROPOSED METHOD

IRFGN-Net is an end-to-end lightweight method designed to enhance the recognition performance of sensitive images. The main components of IRFGN-Net are detailed in the section.

3.1 Overview

This paper proposes an innovative lightweight convolutional neural network structure IRFGN-Net based on improved Runge Kutta and feature guidance fusion. Figure 1 illustrates the overall architecture of the proposed IRFGN-Net. The input image first undergoes feature extraction through a CBL module consisting of a 3×3 convolution, a batch normalization layer, and LeakyReLU. The extracted features are combined with the edge features of the input image and processed through the innovative Feature Guidance Fusion (FGF) module, detailed in

Section 3.3. The FGF module fuses texture features of sensitive images with depth features, effectively reducing the influence of illumination and noise on the recognition accuracy. These features are then enhanced by residual units and the improved Runge Kutta (IRK) block, detailed in Section 3.2. The innovative improved Runge Kutta high-order convolutional block replaces the traditional residual blocks. This method enhances feature extraction ability by introducing weighted higher-order information. Specifically, it replaces skip connections in each ResBlock with an improved higher-order Runge Kutta method, while paying adaptive attention to spatial and channel information for each branch, thereby improving background differentiation and feature representation quality, which is particularly beneficial for sensitive image recognition tasks.

In Figure 1, the multi-scale outputs from the three improved Runge Kutta modules are concatenated after spatial pyramid pooling (SPP) and upsampling. The SPP employs max-pooling operations (MP) with kernels of 5×5, 9×9, and 13×13 to extract and aggregate features from the feature map from different perspectives, enhancing the algorithm's robustness and accuracy. The IRFGN-Net introduces Depthwise Separable Convolution, the MBConv module, and the Ghost Module(GM), which greatly reduces the number of parameters in the network and achieving a balance between high performance and lightweight design. Depthwise Separable Convolution divides the convolution operation into two parts: Depthwise Convolution and Pointwise Convolution. In Depthwise Convolution, a single-channel convolution kernel is applied independently to each channel of the input feature map. Pointwise Convolution uses a 1×1 convolution kerne, matching the number of channels in the previous feature map to perform weighted combinations along the depth direction of the feature map. The MBconv module consists of a 1×1 expansion convolution, depth-wise separable convolution, SENet, 1×1 reduced dimension convolution, and a Dropout layer. The Ghost Module(GM) applies a linear transformation Φ to the intrinsic feature map to obtain a ghost map, which is then combined with the intrinsic map as the output. The activation function of GM uses the HardSwish function. The IRFGN-Net uses the cross-entropy loss function to solve for the model parameters.

3.2 Improved Runge Kutta Method

This section is split into two parts. In Section 3.2.1, the theoretical derivation of the relationship between the Runge Kutta method and convolutional neural networks is presented. An improved Runge Kutta method is introduced in Section 3.2.2.

3.2.1 Runge Kutta Method and Neural Networks

A study by Weinan (2017) established that neural networks can be understood as continuous dynamical systems. This insight led researchers to consider residual neural networks from the perspective of ordinary differential equations (ODEs). In 2018, research on interpreting neural networks using ODEs (Chen et al., 2018) progressed by replacing discrete DNNs with ODE solvers. The researchers used the ODE to describe the dynamical system, defined as:

$$\begin{cases} \frac{dy}{dx} = f(x, y) \\ y(x_0) = y_0 \end{cases} \tag{1}$$

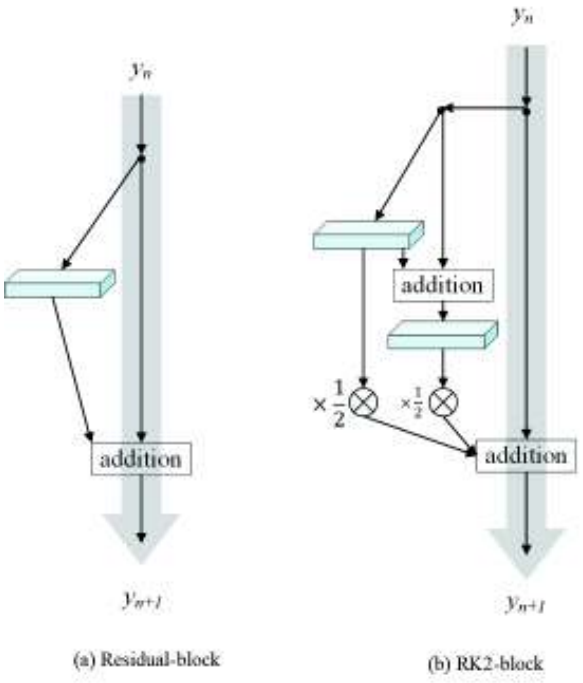

Fig. 2. Schematic diagram of Residual block and second-order Runge Kutta block

The numerical method of ordinary differential equation is to find the approximate value $y(x_n)$ of the exact solution $y(x)$ on the node.

In order to obtain the approximate solution, the difference formula of equation is established by numerical integration method on the interval $[x_n, x_{n+1}]$:

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x, y)dx \tag{2}$$

Applying the rectangle formula to the right-hand integral of the above formula, the Euler method as shown below can be obtained:

$$y_{n+1} - y_n = hf(x_n, y_n) \tag{3}$$

From the current state and the derivative of the current state, the representation of the state at the next time can be solved. where h is the interval and $f()$ is the derivative at x_n, which can be expressed as the incremental relationship between two states. Similarly, there are similar features in the residual module, which can be expressed as:

$$y_{n+1} = y_n + G(y_n) \tag{4}$$
where $G(y_n) = hf(x_n, y_n)$

There is the problem of supervised sensitive image recognition, involving the use of training data to learn a mapping function. Many steps are required before reaching the final state. Each of the steps corresponds to a CNN block. Adding more steps or refining each step helps in achieving the goal. This corresponds to adding more blocks and designing more elaborate blocks. Higher-order methods are considered to have advantages in obtaining exact solutions

of ordinary differential equations. For finer blocks, we can use other numerical methods.

The Runge Kutta (RK) method is an important implicit or explicit iterative method for solving the numerical solutions of ordinary differential equations. Moreover, the Runge Kutta method is high precision algorithm widely used in engineering. Its main idea is to select specific points near the function point x, and then combine the derivative values of these points linearly to construct a set of approximate formulas. The approximate formula is compared with the Taylor expansion of the solution of the ordinary differential equation. By aligning the leading terms of the approximate formula with those of the Taylor expansion, a numerical calculation formula with a specified precision is obtained. The trapezoidal formula is used in the following manner:

$$y_{n+1} = y_n + \frac{h}{2}\left(f\left(x_n, y_n\right) + f\left(x_{n+1}, \tilde{y}_{n+1}\right)\right) \tag{5}$$

Replacing $\tilde{y}_{n+1}$ with its first-order approximation, we get the following second order Runge Kutta equation:

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{2}\left(G_1 + G_2\right) \\ G_1 &= hf\left(x_n, y_n\right) \\ G_2 &= hf\left(x_n + h, y_n + G_1\right) \end{aligned} \tag{6}$$

To map these equations to the CNN block, we replace G_1, G_2 with the Depthwise Separable Convolution and PReLU combination module G. The structures of the corresponding residual block and second-order Runge Kutta block are shown in Figure 2.

3.2.2 Improved Runge Kutta Method Based on the Attention Mechanism

Sensitive images, such as those containing pornography and violence, are recognized based on sensitive features such as exposed body parts, guns, and blood. However, background information can interfere with recognition accuracy. To focus more on these sensitive features and reduce interference from redundant background features, an improved Runge Kutta method is designed in this section. As shown in Figure 3, the two branches of the improved Runge Kutta provide method adaptive weighting for spatial and channel features, respectively. This approach allows the model to focus its attention on sensitive areas while suppressing irrelevant background and redundant information.

The image processing flow of the second-order Runge Kutta method can be expressed by the following formula:

$$y_{n+1} = y_n + \frac{1}{2}\left(G\left(y_n\right) + G\left(y_{n+1}\right)\right) \tag{7}$$

$$G\left(y_n\right) = hf\left(x_n, y_n\right) \tag{8}$$

$$G\left(y_{n+1}\right) = hf\left(x_n + h, y_n + G\left(y_n\right)\right) \tag{9}$$

The improved Runge Kutta method enhances the output features of the two branches of the second-order Runge Kutta method in both the channel and spatial dimensions. This method dynamically adjusts the model's attention to input features, helping it better understand and process complex information. The improved Runge Kutta method utilizes a spatial attention mechanism in its first branch to represent the feature map's internal spatial relationships, focusing on the effective information locations. It emphasizes the positional information of sensitive features while reducing attention to background information. The global max pooling and global average pooling on the input feature map are employed for the current branch $F = G\left(y_n\right) \in R^{H \times W \times C}$, resulting in two feature maps of $H \times W \times 1$. Channel pooling is used to compress the channel size, facilitating the subsequent learning of spatial features. We then combined those two feature maps along the channel dimension to form a feature map of dimensions $H \times W \times 2$. The improved RK-block processes this combined feature map using a 7×7 ghost convolution operation producing a feature map of dimensions $H \times W \times 1$. The Sigmoid activation function is applied to this feature map to derive the spatial attention weight matrix M_s. Then, through the element-wise multiplication of M_s with the original feature map, we have a spatiallly weighted feature map:

$$M_s = \delta\left(f^{7*7}\left(\left[AvgPool\left(F\right)\right]; \left[MaxPool\left(F\right)\right]\right)\right) \tag{10}$$

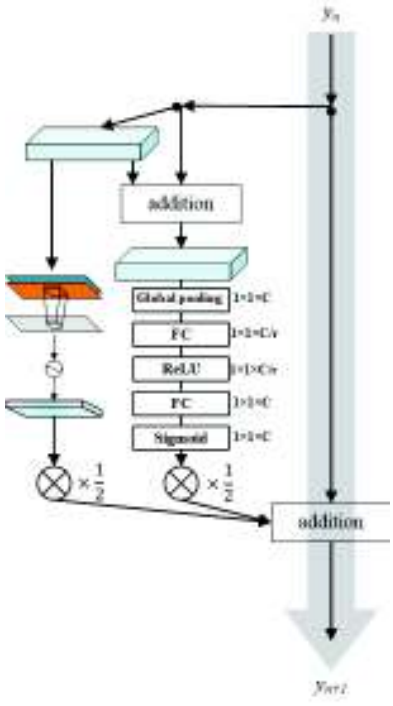

Fig. 3. Schematic illustration of Improved Runge Kutta block (Improved RK-block)

In the second branch of the improved Runge Kutta method, a channel attention mechanism is employed to adaptively adjust the weight of each channel in the feature map and explicitly model the correlation between feature channels. By performing global average pooling on the input feature map $U = G\left(y_{n+1}\right) \in R^{H \times W \times C}$, its dimensionality is changed from $H \times W \times C$ to $1 \times 1 \times C$:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c\left(i, j\right) \tag{11}$$

It conducts feature compression in the spatial dimension by converting each two-dimensional feature map into a

single real number. This process is akin to a pooling operation with a global receptive field, while maintaining the number of feature channels unchanged. The weight of each channel is determined using two fully connected layers and two activation functions. Following the first fully connected layer, we apply the ReLU activation function, resulting in an output dimension of $1 \times 1 \times C'$. Generally, C' is set to one-quarter of C. Subsequently, the Sigmoid function is applied after the second fully connected layer to normalize the value of each layer to $[0, 1]$ and represent the weight s of each channel. The output of the second fully connected layer is also $1 \times 1 \times C$. This weight is then multiplied with the feature map U to obtain the new feature map $\tilde{U}$ after channel weighting:

$$\tilde{U} = \sigma\left(W_2 \delta\left(W_1 z\right)\right) \cdot U \tag{12}$$

The improved Runge Kutta method adjusts feature weights based on the loss incurred throughout the network training process. This adjustment assigns greater weights to effective feature maps and smaller weights to less effective or ineffective ones. The improved Runge Kutta method enables the model to globally integrate features from different positions and channels, enhancing its focus on key features and improving its understanding of sensitive information. This method combines and enhances features, which enables the model to gain richer variant information and more discriminative features, leading to more accurate localization and identification of sensitive information.

3.3 Feature Guidance Fusion Module

Deep features extracted by neural networks are easily affected by noise such as lighting variations and displacements. The structure and texture details are often ignored. To overcome these challenges, this paper proposes an innovative feature guidance fusion (FGF) module. This module leverages local binary pattern (LBP) texture features to guide the fusion of depth features, thereby enhancing the recognition of sensitive images. The architecture of the FGF module is shown in Figure 4.

To enhance the integration of depth and texture features, we employ a feature guidance operation consisting of the following two primary steps. First, the depth feature $p_i^k \in \mathbb{R}^{H \times W \times C}$ is split into m_i ($m_i = C/g_i$) groups, along the channel dimension, where g_i represents the size of each group after splitting. Second, the LBP texture features are periodically interpolated between the grouped feature $p_{i,j}^k \in \mathbb{R}^{H \times W \times g_i}$. The new fused features for subsequent training is obtained. Subsequently, a residual operation is applied to generate the refined feature map p_{i+1}^k.

$$p_{i+1}^k = p_i^k + conv(concat(p_{i,j}^k, h_i^k)) \tag{13}$$

Textures serve as vital visual cues on object surfaces, encompassing information related to shape, size, brightness, and direction. Extracting texture information facilitates the identification and differentiation of various objects, scenes, and patterns. We utilize the LBP descriptor to extract textural features, benefiting from its robust grayscale invariant properties and its ability to mitigate the effects of displacements and uneven lighting. Assume a pixel in the image is represented by $f\left(x_c, y_c\right)$, and the eight points

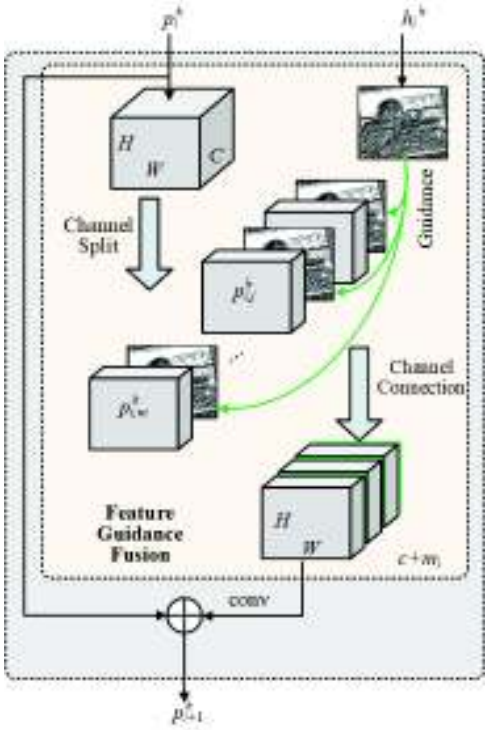

Fig. 4. Schematic diagram of Feature Guidance Fusion

within the 3×3 window with u_c as the center point are denoted as $u_0, u_1, \cdots, u_7$. $H = h\left(u_0, u_1, \cdots, u_7\right)$ is defined as the texture of the local area, and the non-center pixels are binarized within the window, with the threshold determined based on the grayscale of the center pixel. We compare the grayscale values of pixels within the window with the grayscale value of the center pixel. Pixels with a grayscale value lower than that of the center point are represented by 0, while those with a higher grayscale value are denoted by 1, as illustrated below:

$$H \approx h\left(s\left(u_0 - u_c\right), s\left(u_1 - u_c\right), \cdots, s\left(u_7 - u_c\right)\right) \tag{14}$$

$$s\left(u_i - u_c\right) = \begin{cases} 1, & u_i - u_c > 0 \\ 0, & u_i - u_c \leq 0 \end{cases} \quad i = 0, 1, \cdots, 7 \tag{15}$$

An 8-bit binary number can be derived in a clockwise manner, representing a binary pattern. This pattern encapsulates the characteristic value of the center pixel. Hence, each symbolic function gets converted into a decimal number thorough the application of the following formula:

$$\mathrm{LBP}\left(x_c, y_c\right) = \sum_{i=0}^{7} s\left(u_i - u_c\right) 2^i \tag{16}$$

The sensitive images to be identified using the LBP operator are scanned. An image's LBP texture vector map is deliniated as follows:

$$H = \sum_{x,y} I\left\{f_i(x, y) = i\right\}, i = 0, 1, 2, \cdots, n-1 \tag{17}$$

where n represents the number of LBP codes, and each LBP code within the 3×3 pattern is an 8-bit binary number with 256 levels. The resulting LBP code encompasses texture details such as edges and sensitive areas of the image, with each LBP code termed a micro-pattern.

Consequently, the vector comprising 256 LBP codes within a sensitive image constitutes an LBP texture feature.

4. EXPERIMENTAL RESULT AND DISCUSSION

4.1 Datasets and Experimental Setup

The absence of a unified public dataset for three-class sensitive image recognition requires the study to retrieve multiple binary classification public datasets and Kaggle datasets. Python programs were then employed to extract relevant images, thereby enriching our three-class sensitive dataset. Additionally, some images were obtained from browser keyword searches. The sensitive dataset used in this study has a total of 30,000 images, including 10,000 pornographic images, 10,000 violent images, and 10,000 normal images. The hardware setup for the experiments includes an RTX 4090 GPU, a Linux operating system, and PyTorch as the deep learning framework. The training process involves 300 epochs with a batch size of 64, a learning rate set at 5e-4, and a weight decay of 5e-2. No pretrained weights are employed, and we do not use frozen training strategies.

4.2 Evaluation Metrics and Heat Map Visualization

Accuracy, precision, recall and F1-score were adopted as the primary evaluation metrics for the assessment of the model's performance in sensitive image recognition tasks. Furthermore, we consider inference speed to evaluate the practical feasibility of the model. The formulas for these metrics are as follows:

$$Recall = \frac{TP}{TP + TN} \tag{18}$$

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{21}$$

Here, TP represents True Positive, the number of samples that the model correctly predicts as positive; FP represents False Positive, the number of samples that the model predicts falsely as positive; TN represents True Negative, the number of samples that the model predicts correctly as negative; and FN represents False Negative, the number of samples that the model predicts falsely as negative.

We obtained heat maps by weighting the class activation map of the last convolutional layer in each recognition model. The heat maps highlight the sensitive regions in the input image that are most relevant to the target class across different recognition models. The heat map visualization technique we used is Grad-CAM. Grad-CAM (Selvaraju et al., 2017) first performs backpropagation for the target class within the model output to compute the gradient of the feature map. Then, spatial average pooling is applied to the gradient through adaptive average pooling to obtain the weights of each feature map channel. The final heat map is obtained by multiplying the feature map and the weights channel by channel, summing the results, and then applying the ReLU activation function. The heat map is visualized on the original image, showing which high-level features extracted by the IRFGN network most impact final image recognition.

$$L^c_{\text{Grad-CAM}} = \text{Re}LU\left(\sum_k \alpha^c_k A^k\right) \tag{22}$$

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{23}$$

Here, A stands for a specific feature layer, k denotes the k-th channel within feature layer A, c denotes class c, A^k stands for the data in channel k of feature layer A, a^c_k represents the weight corresponding to A^k, y represents the network's predicted score for class c, $A_{i,j}$ represents the data at the (i, j) position in the feature layer A's channel k , and Z denotes the width times the height of the feature layer.

4.3 Comparison Experiments and Analysis

As to the effectiveness verification of IRFGN-Net, this study compared it with VggNet16 (Abundez et al., 2023), Resnext50 (Xie et al., 2017), DenseNet121 (Pisantanaroj et al., 2020), ConvNext_tiny (Liu et al., 2022), EfficientNet_v1_B0 (Tan and Le, 2019), GhosTNet (Han et al., 2020), AnimeNet (Tang, 2023), AqCONet (Gupta et al., 2022),LCSNet-72 (Zhang and Huang, 2024) , and AttM-CNN (Gangwar et al., 2021). Several comparison methods were retrained in the experimental environment with reference to the original training parameter settings. Table 1 contains detailed experimental results. We achieved the best results in accuracy, precision, recall, and f1-score indexes through the proposed novel approach. Furthermore, the proposed method has the fewest parameters. Compared with ConvNext_tiny, VggNet16 and DenseNet121, IRFGN-Net adds higher-order convolution and adaptive attention. These improvements increased the accuracy of IRFGN-Net by 0.038, 0.015, and 0.009, respectively. The accuracy of ResNet50 model is 0.931, which is 0.011 lower than that of the model proposed in this paper. The result demonstrates that the innovative improved Runge Kutta high-order convolution modules and FGF modules extract rich sensitive features and reduce the impact of noise on accuracy. The accuracy rates of the EfficientNet and AnimeNet reach 0.933 and 0.932, but are still lower than the models proposed in this paper. This result demonstrates the effectiveness of the proposed FGF module. Regarding model size, the proposed IRFGN-Net is only 3.64M, whereas other models exceed 4M. Although the number of parameters for GhostNet and EfficientNet is 4.21M and 4.0M, the accuracy of these models is not as high as that of IRFGN-Net. Additionally, LBP descriptors, Ghost modules, and depthwise separable convolution used by IRFGN-Net typically have fewer parameters, which can improve training and inference efficiency in scenarios with limited computational resources and large-scale datasets. Compared to the baseline models, the proposed IRFGN-Net has fewer parameters and higher accuracy.

The accuracy curves of various models during training and validation are illustrated in Figure 5. Figure 5(a) and (c) show that the increase in the number of validation epochs was followed by the continuous improvement of validation accuracy of this model. After 200 epochs, it stabilizes above other models, fully demonstrating excellent fitting

Table 1. Performance comparison of recognition results of the proposed model and other models.

Method	Accuracy	Precision	Recall	F1-score	Parameters
ConvNext_tiny	0.904	0.901	0.901	0.901	27.82M
VggNet16	0.927	0.923	0.922	0.922	134.28M
GhostNet	0.928	0.927	0.927	0.927	4.21M
Resnext50	0.931	0.930	0.928	0.929	22.99M
EfficientNet_v1_B0	0.933	0.931	0.930	0.930	4.01M
DenseNet121	0.933	0.931	0.930	0.930	6.96M
AnimeNet	0.932	0.928	0.927	0.927	33.48M
AqCONet	0.906	0.903	0.901	0.901	42.51M
LCSNet-72	0.909	0.907	0.904	0.905	4.0M
AttM-CNN	0.910	0.905	0.904	0.904	57.2M
IRFGN-Net	**0.942**	**0.941**	**0.942**	**0.941**	**3.64M**

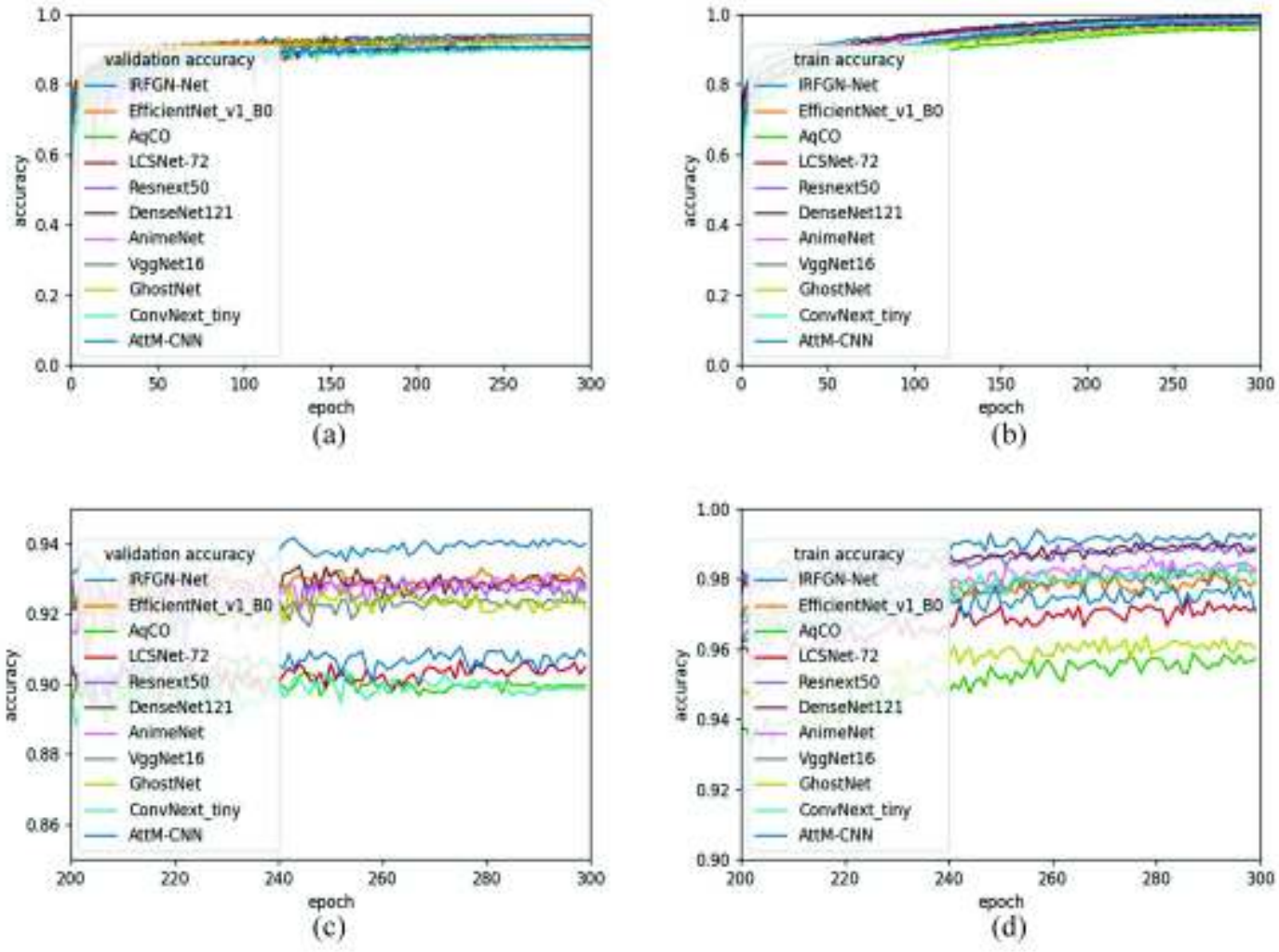

Fig. 5. Validation and training accuracy curves of the proposed IRFGN-Net and other recognition networks. (a) and (b) show the accuracy rates for 0-300 epochs, while (c) and (d) provide detailed accuracy rates for 200-300 epochs.

ability. As shown in Figure 5(b) and (d), in the majority of training sessions, the proposed model demonstrates higher training accuracy compared to the baseline model. It indicats that the proposed model effectively extracts and identifies sensitive features. Figure 6 illustrates that the proposed model's loss curve quickly reaches the minimum. This indicates a good initial network design enabling rapid convergence. The loss curve of the model tends to become smoother and stabilize at a faster rate as the accuracy improves. Comparing the accuracy and loss curves, the

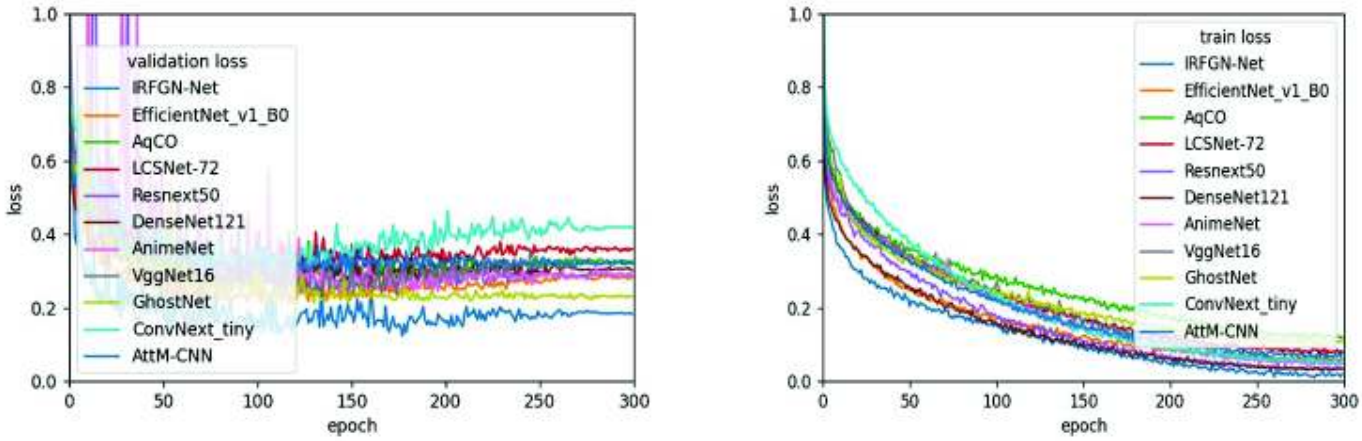

Fig. 6. Validation and training loss curves of the proposed IRFGN-Net and other recognition networks.

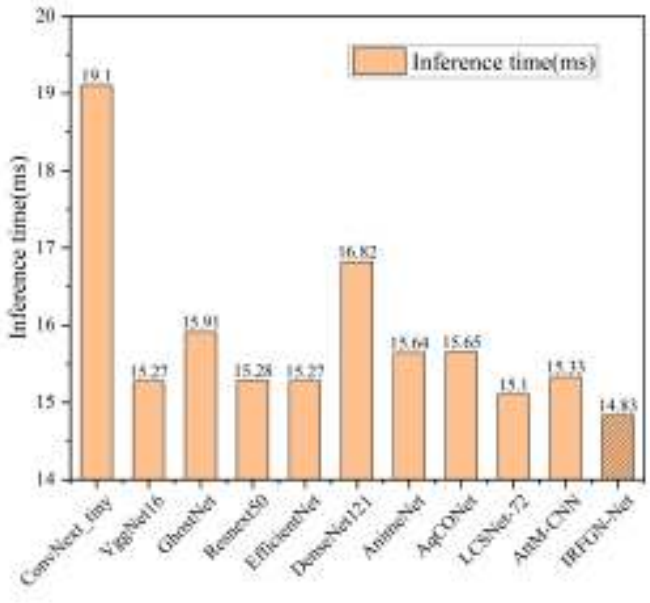

Fig. 7. Inference time comparison of different models

model's good network structure is evidenced. It embodies fast convergence and good fitting ability, superior to other deep learning methods.

The current study demonstrated the computational performance of the proposed network through the calculation of the inference time ConvNext_tiny, VggNet16, AttM-CNN, GhostNet, Resnext50, DenseNet121, and IRFGN-Net. The results are shown in Figure 7. Compared to the baseline models, IRFGN-Net has an inference time of 14.83ms, indicating that it requires minimal prediction time. This efficiency allows for practical implementation in real-world environments. While Resnet101 and DenseNet121 show notable performance on sensitive datasets, their processing speed is compromised due to the large number of hidden layers. IRFGN-Net outperforms baseline models in both inference time and detection accuracy. Therefore, IRFGN-Net achieves a balance between performance and computational efficiency, making it suitable for real-time detection on resource-constrained devices in natural environments.

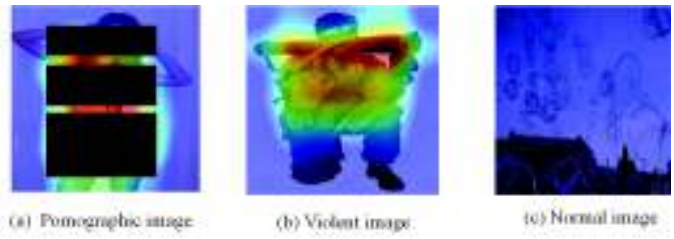

Fig. 8. Attention visualization results of the proposed model

To compare the details of different models in sensitive image recognition, heat maps are introduced to visualize the details. Figure 8 shows IRFGN-Net heat maps for pornographic, violent, and normal images. In pornographic image, IRFGN-Net focuses on private parts and large areas of exposed skin. In the violent image, IRFGN-Net focuses mainly on violent features such as guns. The normal image is a simple background image with no obvious pornographic or violent features. So IRFGN-Net does not give obvious attention. The heat maps show that IRFGN-Net learns rich sensitive features through the innovative IRK high-order convolution and FGF modules, effectively avoiding background interference. Figure 9 shows the heat maps of six models. IRFGN-Net focuses on the overall gun, with heat points on each part and attention to the edges of the gun. The VggNet16, Resnext50, and GhostNet model focus on the front part of the gun. The AnimeNet focuses on the body of the gun, and some heat points are scattered in the background image. EfficientNet focuses on small areas such as the body and stock of the gun. This result shows that the design of improved RK-block and FGF module enables the model to distinguish sensitive features from non-sensitive features well, and pay more comprehensive attention to sensitive features. The visualization results show that the proposed IRFGN-Net can recognize many kinds of sensitive images well and has good anti-interference performance.

4.4 Ablation Experiments and Analysis

Ablation experiments assess how much specific components contribute to the model's performance by remov-

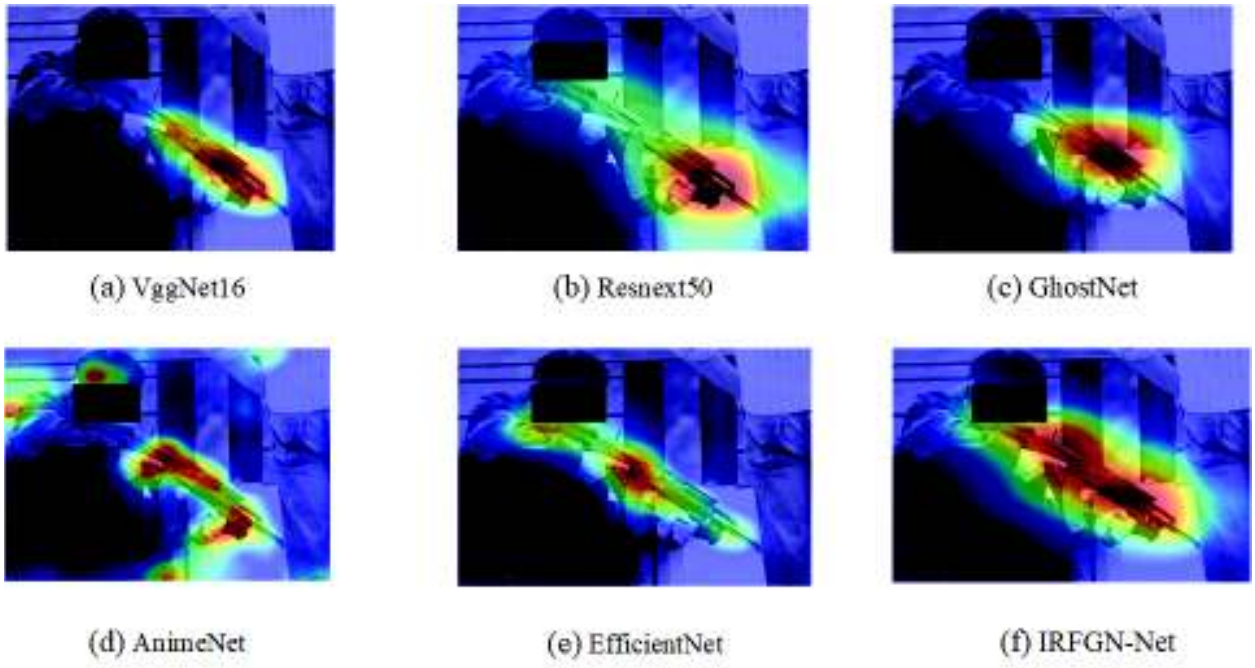

Fig. 9. Results on visualization of attention for each model

Table 2. Results of the ablation experiment.

Residual block	RK2-block	Improved RK-block	FGF-block	Accuracy	Precision	Recall	F1-score
✓				0.912	0.914	0.911	0.912
	✓			0.918	0.917	0.920	0.917
		✓		0.932	0.934	0.929	0.930
			✓	0.919	0.917	0.921	0.918
✓	✓	✓	✓	**0.942**	**0.941**	**0.942**	**0.941**

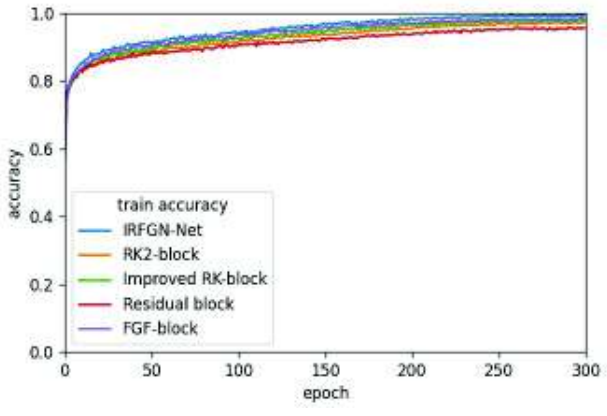

Fig. 10. Ablation experiment training accuracy curve

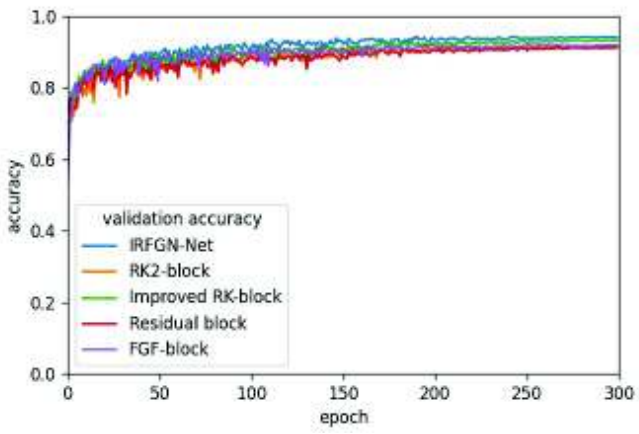

Fig. 11. Ablation experiment validation accuracy curve

ing or altering them in a systematic manner. This study conducted ablation experiments separately for the improved Runge Kutta block and the Feature Guidance Fusion (FGF-block). In the ablation experiments, the study initially replaced the improved RK-block with residual modules. Then we removed the attention feature fusion while keeping the training parameters consistent. Table 2 contains the experimental results which indicate that replacing the improved RK-block with ResNet's residual modules reduced the model's accuracy from 0.932 to 0.912. Precision, recall, and F1-score dropped to 0.914, 0.911 and 0.912. The higher-order convolution and attention weighting used by the improved RK-block are better than the Residual block in feature extraction. This study compared the improved RK-block with the RK2-block. The four evaluation indexes of the improved RK-block increased by 0.014, 0.017, 0.09 and 0.013 respectively. This result demonstrates that the adaptive attention weighting of

channels and spatial dimensions improves the attention to sensitive sites. Compared with the Residual block, the use of FGF-block increased the accuracy from 0.912 to 0.919, and the other three indicators improved. Attention to texture features improves the richness of input features and enhances the recognition of sensitive images.

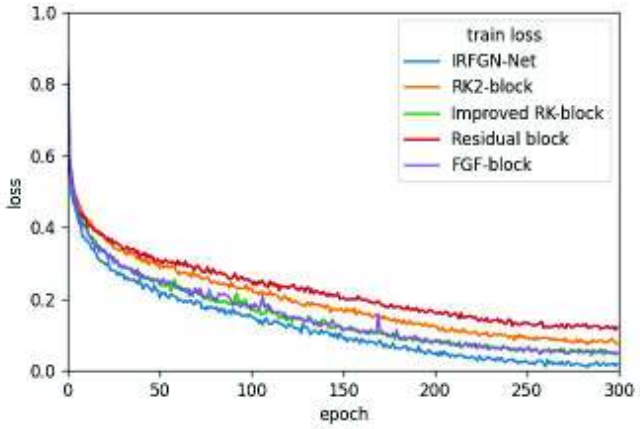

Fig. 12. Ablation experiment training loss curve

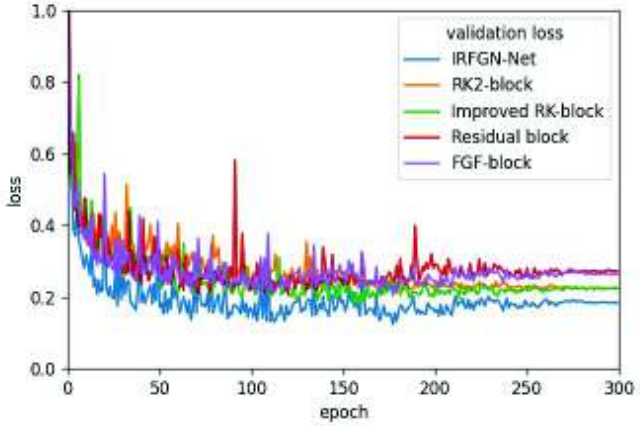

Fig. 13. Ablation experiment validation loss curve

Figure 10 shows the training accuracy curve of the ablation experiments, showing IRFGN-Net consistently maintaining the highest accuracy. Conversely, the curves for models without the improved RK-block and FGF exhibit varying degrees of decline. Figure 11 displays the validation accuracy curve, which follows a similar trend as the training accuracy curve. The training and validation loss curves are shown in Figures 12 and 13 with the removal of the improved RK-block and FGF leading to varying degrees of reduced convergence speed. The ablation experiments provided results showing that the improved RK-block provides stronger fitting capabilities than the residual modules. The improved Runge Kutta high-order convolutional module comprises multiple attention convolutional layers and non-linear activation functions, providing more robust nonlinear modeling capabilities. This enables the high-order convolutional module to better adapt to complex data distributions and feature representations, enhancing the model's expressive power. The loss curve shows that the addition of FGF module reduced the time needed for convergence of the model. The feature guidance fusion module integrates the texture and depth features of sensitive images, effectively improving the recognition accuracy of sensitive images.

5. CONCLUSION

In the digital era, the widespread dissemination of large-scale image and video content on the Internet has made the automatic identification of sensitive imagery imperative. This task has profound implications for the health and safety of users and society and extends to domains like child protection and social media content governance. The main challenge is balancing swift inference with high recognition accuracy. This study introduces an improved Runge Kuta method and an innovative feature guidance fusion module to solve this problem. The improved Runge Kutta high-order convolutional network module extracts more fine-grained sensitive features. By optimizing the numerical solution of ordinary differential equation and weighting the feature mapping, the method enhances the recognition of key characteristics while mitigating the noise and irrelevant background of sensitive images. The proposed FGF module fuses texture and depth features of sensitive images, effectively reducing the influence of light and noise on recognition accuracy of sensitive images.

These innovations enhance the model's feature representation capability and make it more lightweight. Through exhaustive experiments, the effectiveness of the proposed method is evidenced. Comparison with existing convolutional neural networks for sensitive image recognition show that IRFGN-Net significantly improves all evaluation indexes on multi-classification sensitive datasets, with optimal model parameters and inference time. While the proposed IRFGN-Net has made notable strides in sensitive image recognition, many directions beckon further research. IRFGN-Net modules and feature guidance fusion mechanisms can be applied to object detection, semantic segmentation, and similar computer vision tasks, to validate their broader effectiveness. Additionally, refining the model may include increasing network depth or introducing more data augmentation strategies to bolster robustness.

DATA AVAILABILITY

The data sets applied in this paper are available on the Internet. These data were derived from the following publicly available web resources:
https://www.kaggle.com/datasets/sakshamagrawal/terrorism
https://www.kaggle.com/datasets/abdulmananraja/real-life-violence-situations
https://www.kaggle.com/datasets/issaisasank/guns-object-detection
https://aistudio.baidu.com/datasetdetail/60943
https://github.com/alex000kim/nsfw_data_scraper

REFERENCES

Abundez, I., Miranda-Piña, G., Alejo, R., Granda-Gutiérrez, E., Cisniega, A., and Portillo-Rodríguez, O.

(2023). Empirical analysis on the effectiveness of pre-trained models in the identification of physical violence against women in videos for a multi-class approach. In *International Conference on Data Mining and Big Data*, 152–164. Springer.

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., and Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14*, 332–339. Springer.

Boulkaboul, S., Bouchama, S., Kasser, S., and Ali, B.A.S. (2024). D-resnet: Deep resnet based approach for ecg classification. *Journal of Control Engineering and Applied Informatics*, 26(1), 64–71.

Camilleri, C., Perry, J.T., and Sammut, S. (2021). Compulsive internet pornography use and mental health: A cross-sectional study in a sample of university students in the united states. *Frontiers in psychology*, 11, 613244.

Chen, M.y. and Hauptmann, A. (2009). Mosift: Recognizing human actions in surveillance videos. *Computer Science Department*, 929.

Chen, R.T., Rubanova, Y., Bettencourt, J., and Duvenaud, D.K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Connie, T., Al-Shabi, M., and Goh, M. (2017). Smart content recognition from images using a mixture of convolutional neural networks. In *IT Convergence and Security 2017: Volume 1*, 11–18. Springer.

Gangwar, A., González-Castro, V., Alegre, E., and Fidalgo, E. (2021). Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images. *Neurocomputing*, 445, 81–104.

Gugulothu, V.K. and Balaji, S. (2023). A novel ezs-msca and selu squeezenet-based lung tumor detection and classification. *Soft Computing*, 1–12.

Gupta, J., Pathak, S., and Kumar, G. (2022). Aquila coyote-tuned deep convolutional neural network for the classification of bare skinned images in websites. *International Journal of Machine Learning and Cybernetics*, 13(10), 3239–3254.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. (2020). Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1580–1589.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, H., Lv, Y., Yuan, R., Xu, S., and Zhu, W. (2023). A novel vibro-acoustic fault diagnosis approach of planetary gearbox using intrinsic wavelet integrated ge-efficientnet. *Measurement Science and Technology*, 35(2), 025131.

Hu, Z. and Hu, L. (2023). Visual loop closure detection based on squeezenet multi-layer feature fusion and adaptive range matching algorithm. *Journal of Intelligent & Robotic Systems*, 108(3), 55.

Huang, Y. and Kong, A.W.K. (2016). Using a cnn ensemble for detecting pornographic and upskirt images. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–7. IEEE.

Jones, M.J. and Rehg, J.M. (2002). Statistical color models with application to skin detection. *International journal of computer vision*, 46, 81–96.

Kurt, Z., Işık, Ş., Kaya, Z., Anagün, Y., Koca, N., and Çiçek, S. (2023). Evaluation of efficientnet models for covid-19 detection using lung parenchyma. *Neural Computing and Applications*, 35(16), 12121–12132.

Liu, Z. and Zhang, C. (2024). Multi-grained feature aggregation based on transformer for unsupervised person re-identification. *Journal of Control Engineering and Applied Informatics*, 26(1), 72–82.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Lv, L., Zhao, C., Lv, H., Shang, J., Yang, Y., and Wang, J. (2011). Pornographic images detection using high-level semantic features. In *2011 Seventh International Conference on Natural Computation*, volume 2, 1015–1018. IEEE.

Moustafa, M. (2015). Applying deep learning to classify pornographic images and videos. *arXiv preprint arXiv:1511.08899*.

Pisantanaroj, P., Tanpisuth, P., Sinchavanwat, P., Phasuk, S., Phienphanich, P., Jangtawee, P., Yakoompai, K., Donphoongpi, M., Ekgasit, S., and Tantibundhit, C. (2020). Automated firearm classification from bullet markings using deep learning. *IEEE Access*, 8, 78236–78251.

Rezaee, K., Mousavirad, S.J., Khosravi, M.R., Moghimi, M.K., and Heidari, M. (2021). An autonomous uav-assisted distance-aware crowd sensing platform using deep shufflenet transfer learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(7), 9404–9413.

Ries, C.X. and Lienhart, R. (2014). A survey on visual adult image recognition. *Multimedia tools and applications*, 69, 661–688.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Steely Smith, M.K. and Ten-Bensel, T. (2024). Sexual grooming behavior and processes of women who commit sexual offenses against children. *Journal of interpersonal violence*, 08862605241239450.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tang, Y. (2023). Animenet: A deep learning approach for detecting violence and eroticism in animated content. *Computers, Materials & Continua*, 77(1).

Torky, M., Dahy, G., and Hassanein, A.E. (2023). Gh2_mobilenet: Deep learning approach for predicting green hydrogen production from organic waste mixtures.

Applied Soft Computing, 138, 110215.

Tran, K., Nguyen, D., and Nguyen, H. (2019). Toward deep and handcraft features for detecting violent behaviors. In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, 422–427. IEEE.

Ulges, A. and Stahl, A. (2011). Automatic detection of child pornography using color visual words. In *2011 IEEE international conference on multimedia and expo*, 1–6. IEEE.

Weinan, E. (2017). A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5), 1–11.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xu, W., Parvin, H., and Izadparast, H. (2020). Deep learning neural network for unconventional images classification. *Neural Processing Letters*, 52(1), 169–185.

Yang, J., Fu, Z., Tan, T., and Hu, W. (2004). A novel approach to detecting adult images. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, 479–482. IEEE.

Zeng, W., Gao, W., Zhang, T., and Liu, Y. (2004). Image guarder: An intelligent detector for adult images. In *Asian conference on computer vision*, 1080–1084.

Zhang, X. and Huang, H. (2024). Lcsnet: Light-weighted convolution-based segmentation method with separable multi-directional convolution module for concrete crack segmentation in drones. *Electronics*, 13(7), 1307.

Zhang, Y., Xie, W., and Yu, X. (2023). Design and implementation of liveness detection system based on improved shufflenet v2. *Signal, Image and Video Processing*, 17(6), 3035–3043.