Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.038586
Article

Fine-Grained Pornographic Image Recognition with Multi-Instance Learning

Zhiqiang Wu* and Bing Xie

Department of Network Security, Henan Police College, Zhengzhou, 450000, China
*Corresponding Author: Zhiqiang Wu. Email: 20190035@hnp.edu.cn

Received: 20 December 2022; Accepted: 03 March 2023; Published: 26 May 2023

Abstract: Image has become an essential medium for expressing meaning and disseminating information. Many images are uploaded to the Internet, among which some are pornographic, causing adverse effects on public psychological health. To create a clean and positive Internet environment, network enforcement agencies need an automatic and efficient pornographic image recognition tool. Previous studies on pornographic images mainly rely on convolutional neural networks (CNN). Because of CNN's many parameters, they must rely on a large labeled training dataset, which takes work to build. To reduce the effect of the database on the recognition performance of pornographic images, many researchers view pornographic image recognition as a binary classification task. In actual application, when faced with pornographic images of various features, the performance and recognition accuracy of the network model often decrease. In addition, the pornographic content in images usually lies in several small-sized local regions, which are not a large proportion of the image. CNN, this kind of strong supervised learning method, usually cannot automatically focus on the pornographic area of the image, thus affecting the recognition accuracy of pornographic images. This paper established an image dataset with seven classes by crawling pornographic websites and Baidu Image Library. A weakly supervised pornographic image recognition method based on multiple instance learning (MIL) is proposed. The Squeeze and Extraction (SE) module is introduced in the feature extraction to strengthen the critical information and weaken the influence of non-key and useless information on the result of pornographic image recognition. To meet the requirements of the pooling layer operation in Multiple Instance Learning, we introduced the idea of an attention mechanism to weight and average instances. The experimental results show that the proposed method has better accuracy and F1 scores than other methods.

Keywords: Deep learning; multi-instance learning; pornographic image; multi-classification; residual network

1 Introduction

With the development of information technology, the Internet has become an indispensable part of people's life. While people enjoy the convenience of the Internet, they also suffer from the harassment of harmful information brought by it [1]. At present, there are a large number of pornographic images on the Internet. To create a clean and positive Internet environment, network enforcement agencies need an automatic and efficient pornographic image recognition tool.

Pornographic images have the characteristics of various forms and strong concealment. The most simple and quick way to recognize pornographic images is to audit them manually. Because there are tens of thousands of images on the Internet, it is unimaginable if we review them manually. It is an excellent choice to recognize pornographic images with the help of computers automatically.

As a challenging research topic in computer vision, in earlier years, various approaches have been proposed to solve the problem of pornographic image recognition [2]. It can be divided into two stages according to the development of technology. In the first stage, researchers mainly based on the region of interest (ROI) method. The skin color, shape, and posture of the human body are essential indicators [3–5]. These features are extracted and used as classifier inputs to output whether the image is pornographic. However, this method has many limitations and often needs to introduce other technologies, such as face recognition, human trunk contour extraction, geometric constraints on body parts, etc. These technologies increase the complexity of the system. With the infinite size of the image sample, these artificially selected image features will inevitably increase the misjudgment rate. After the rise of deep learning (DL), CNN achieved great image recognition success. It brings the recognition of pornographic images to a new stage. The method based on DL no longer extracts image features manually but learns them automatically through a large amount of training data, which has high accuracy and robustness in pornographic image recognition.

One of the essential works in pornographic image recognition is the extraction of image features. Many researchers have proposed various CNN-based pornographic image recognition methods [6,7], which have achieved good results due to CNN's excellent image feature representation ability. Nevertheless, the pornographic contents, such as genitals and breasts in images, usually lie in several small-sized local regions, which are not a large proportion of the image. CNN, this kind of strong supervised learning method, usually cannot automatically focus on the pornographic area of the image, thus affecting the recognition accuracy of pornographic images. Additionally, owing to the large number of parameters involved in CNN, it requires large-scale labeled training data that may be difficult to obtain in pornographic image recognition. Therefore, many studies treat pornographic image recognition as a binary classification task. When confronted with pornographic images of diverse characteristics in practical applications, the performance and recognition accuracy of the network models often decrease. To address this issue, we propose an innovative pornographic image recognition method based on MIL and multi-classification that learns more features under weak supervision. The main contributions of this paper are summarized as follows:

(1) Many previous studies have regarded pornographic image recognition as a binary classification task, ignoring the diversity of pornographic images. In order to address this issue, this paper combines the diverse characteristics of pornographic images, converts the binary pornographic image recognition task into a multi-classification task, and then judges whether the image is a pornographic image according to the output score.

(2) There are problems of low accuracy and poor network adaptability when using the original MIL algorithm to recognize pornographic images. In order to improve the accuracy of spatial and channel feature extraction, this paper proposes a multi-classification model for pornographic image recognition based on MIL and ResNet with a squeeze-and-excitation module (SE-ResNet-MIL).

(3) To increase accuracy and network adaptability, an attention mechanism is applied in the pooling operation to perform a weighted average operation on image instances.

The remainder of this article is organized as follows. Section 2 provides an overview of the related work. Section 3 introduces in detail the proposed method for pornographic image recognition. Section 4 explains the experimental methods and procedures. The fifth section analyses and discusses the results of the experiments. Finally, Section 6 concludes the paper.

2 Related Work

Research on the recognition of pornographic images has a long history, and researchers proposed various approaches that can be divided into two categories. These are ROI-based methods and DL-based methods.

2.1 ROI-Based Methods

Based on the fact that pornographic images tend to contain a large amount of naked skin color information [8], the pornographic image recognition method based on ROI takes skin and sensitive parts of the human body as regions of interest. It segments the image and determines pornographic images by defining several rules based on visual features. In 1996, Fleck et al. [9] first extracted the segmented skin color information and then judged whether there was a naked human body in the image according to the different postures of the human body. Subsequent studies have further improved upon this technique. For instance, Nugroho et al. [10] proposed a pornographic video negative content recognition method using skin segmentation in composite video frames, which combines RGB and YCbCr color spaces as the skin recognition algorithm to improve the accuracy of video class determination. Tian et al. [11] proposed a pornographic image recognition framework based on a sexual organ detector. This approach uses color attributes to describe the local color of sexual organs and combines them with shape features based on a directional gradient histogram to represent sexual organs.

The two key steps involved in these skin color-based recognition methods are image feature extraction and classifier design. Commonly used techniques for feature extraction include the skin pixel definition method, the Bayesian decision color statistical model, and the Gaussian model. Classifications include Support Vector Machines [12], C4.5 decision trees [13], and more. Researchers have also utilized multiple color spaces to define skin color models, enabling more accurate extraction of skin color regions by adding texture recognition [14]. They then use the pixel proportion of bare skin color regions, the number, position, and shape of skin color, and other features as input for the classifiers. The primary issue with these methods based on skin color features and human recognition is that the manually selected skin color models always have some degree of deviation. In reality, the skin color of different races differs, and the same skin color can show mixed results under different lighting [15]. Furthermore, objects are similar to skin color found in nature, resulting in this similarity-matching method based on low-level semantic features having a high false recognition rate.

2.2 DL-Based Methods

In 2012, Krizhevsky et al. proposed a CNN model called AlexNet, which won the ImageNet competition [16]. This success in image recognition sparked interest in applying it to pornographic image recognition. Alguliyev et al. [17] proposed a multi-layer deep neural network architecture of five convolutional blocks, referred to as ChildNet, which can recognize naked bodies in images and perform better than classical CNN. As most of the key pornographic information is present in the local area of an image, many researchers believe that local context information is useful for identifying such information. Cheng et al. [18] then proposed DCNN, a method that can use global and local information about an image to improve the accuracy of pornography recognition. As the critical pornographic content of an image is usually located in certain regions, recognition of sensitive organs became the bottleneck to improving the recall rate of pornographic image recognition. Zhu et al. [19] designed a sensitive organ recognition module and achieved good results based on residual networks. Wang et al. [20] integrated global classification and local sensitive area recognition into one network and proposed a local context-aware network based on the DL method to improve recognition accuracy. Surinta et al. [21] compared AlexNet, GoogleNet, and ResNet [22] as three typical CNN frameworks to identify pornographic images and found that ResNet could achieve higher accuracy.

In recent years, multi-instance learning as a typical weakly supervised learning method has gradually emerged alongside convolutional neural networks in computer vision research [23]. Ilse et al. [24] describe the multi-instance learning problem as the distribution of bag labels. They parameterized the probability of bag labels by the neural network. The test results on the histopathological dataset show that their proposed method has better results than other methods. Jin et al. [25] modeled each image as a group of regions and built a pornographic image recognition model based on neural networks and MIL methods. According to existing research, multi-instance learning has achieved good results in many applications [26–29] and has specific applications for pornographic image recognition. It is often used in conjunction with the neural network method.

As shown in Table 1, there are many methods for recognizing pornography images based on CNN. These methods generally regard the recognition of pornographic images as a binary classification problem without considering the diversity of pornographic image categories. Moreover, pornographic content is usually distributed in some areas of the images, accounting for a small proportion. It is often difficult for CNN's strong supervised learning method to automatically focus on pornographic areas in the image, thus affecting the recognition accuracy of pornographic images. In order to solve these problems, this paper designs a weakly supervised porn image recognition model based on MIL, which can automatically recognize pornography.

Table 1: Comparison table of previous methods

Reference	Methodology	Findings
Chen et al. [7]	Deep one-class classification with a visual attention mechanism	Accuracy of 98.419%.
Alguliyev et al. [17]	Multi-layer deep neural network	The proposed method can recognize naked bodies in images and perform better than classical CNN.
Cheng et al. [18]	Deep convolutional neural network (DCNN)	Accuracy of 96.6%.

(Continued)

Table 1: Continued

Reference	Methodology	Findings
Zhu et al. [19]	Residual networks with a sensitive organ recognition module	The method has good performance.
Wang et al. [20]	Local context-aware network based on the DL	The approach achieved the best classification accuracy compared with several methods investigated.
Jin et al. [25]	Multiple instance learning (MIL)	Accuracy of 97.5%.

3 Proposed Method

3.1 The Architecture for Pornographic Image Recognition

Pornographic image recognition based on DL consists of three key parts. First, we need to define a classifier model to determine the criteria for pornographic images. Second, a large-scale image dataset that is labeled is constructed, allowing the model to fully learn the features of image classification through training. Finally, the model parameters must be adjusted through consistent experiments to achieve the best performance on the evaluation criteria. In this process, the key steps are creating a network model for recognizing pornographic images and extracting the necessary features.

According to previous research results, ResNet, based on CNN, has demonstrated greater results compared to other algorithms in pornographic image recognition. Because CNN requires many parameters when used, it needs to rely on large-scale training data, which is not easy to obtain in pornographic image recognition. To obtain better recognition results in the case of a relative lack of labeled training data, the methods of weakly supervised learning could be a good choice. Multiple Instance Learning (MIL) is an excellent method for weakly supervised learning, capable of managing classification, clustering, and regression problems. Regarding pornographic image recognition, each image can be divided into several instances. An image can then be interpreted as a package that includes multiple instances.

The architecture model we applied for pornographic image recognition is shown in Fig. 1. Firstly, all images are uniformly scaled to 224 × 224, and each image is segmented into multiple sub-images. Next, the features of each sub-image must be extracted and clustered. Finally, depending on the scores for different categories, the recognition results are output using the criteria stated in Section 3.5. In pornographic image recognition, feature extraction of instances in the packet is vital and directly related to the recognition effect and network performance. The feature extraction method often used by scholars is CNN, and the classical CNN algorithm ResNet has achieved better results than other algorithms. We thoroughly use the relationship between channels to improve recognition accuracy while improving the spatial dimensions. Based on the experimental results, we propose to use 152 layers of the residual neural network containing SE modules to extract pornographic image features.

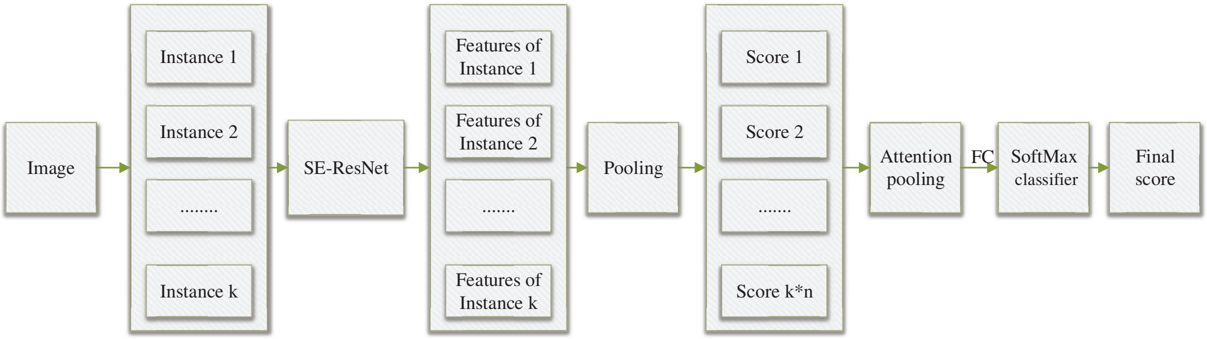

Figure 1: The architecture of the proposed method to recognize pornographic images

3.2 *Multi-Instance Learning Module*

The main challenge is to build an image into a multiple-instance bag form to recognize pornographic images using a multiple-instance learning framework. Currently, the most popular method is segmentation or selective search techniques to extract local regions from the image and then view each image as a bag. Inspired by the idea of Vision Transformer [30], the image is constructed into a multiple-instance bag by grid partition and linear projection.

Let X denote an image, which is automatically divided to form a bag with n sub-images. The mathematical representation is as Eq. (1).

$$X = \{x_1, x_2, \ldots, x_n\} \tag{1}$$

$$D = \{(X_{1,}\mathrm{y}_1), (X_{2,}\mathrm{y}_2), \ldots (X_{m,}\mathrm{y}_m)\} \tag{2}$$

Let $X = R^{\mathrm{d}}$ denotes a d-dimensional instance space and $\mathrm{K} = \{1, 2, \ldots, \mathrm{k}\}$ presents the k-class label space. The training dataset can be represented as Eq. (2). Where m denotes the labeled multiple instance training bags, the package $X_i(i = 1, 2, \ldots, m)$ contains n instances and y_i presents the label number of X_i. The task of multiple instance learning is to learn from D to obtain a classification function that can predict the label of any unlabeled bag X.

3.3 *Image Feature Extraction*

CNN comprises convolutional and pooling layers, which can effectively extract image features. Traditionally, researchers believed that the more convolutional and pooling layers, the more comprehensive image feature information could be acquired through learning, leading to better image classification results. However, experiments have shown that with the increasing number of convolutional and pooling layers, the learning effect does not necessarily become better and better, but due to the problems of gradient disappearance and degradation, the image classification effect may become worse. To combat this, in 2015, He et al. from Microsoft Research proposed ResNet, a residual neural network. Unlike the previous network structure, ResNet finds the optimal number of network layers by calculating whether data can complete identity mapping and ensuring that the output and input through the identity layer are identical. The residual neural network is composed of residual blocks, the overall structure of which is illustrated in Fig. 2, and the residual calculation is shown in Eq. (3).

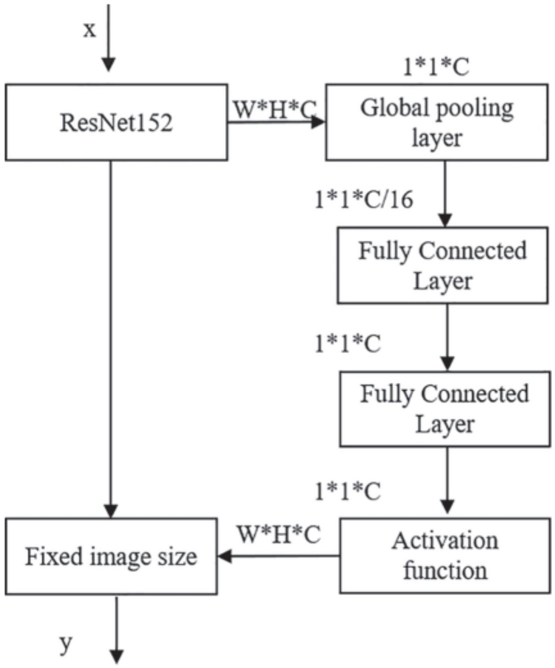

Figure 2: The constructed SE-ResNet model

$$y = F(x) + x \tag{3}$$

In this formula, x denotes the object to be classified, F(x) is a vector derived from the weight layers and activation functions, and y is the desired mapping. By changing the output into a combination of fitting and output, the network is more sensitive to variations between the output y and the input x. This paper constructed a 152-layer residual neural network with a Bottleneck structure, which reduced the number of parameters and improved the model training efficiency. As the network depth increases, the original convolutional neural network's training errors will initially decrease and then increase. In contrast, the training errors of the residual neural network will decrease as the depth increases.

ResNet improves network performance through spatial dimension aggregation of features obtained from various receptive fields. However, it lacks interdependence between feature channels. This paper introduces the SE module [31] to the channel-based ResNet to enhance image feature extraction accuracy. The constructed SE-ResNet model is illustrated in Fig. 2. The relationship between feature channels is used to build the model so that the target vector can more accurately represent instance information. Network training is used to obtain weight parameters of different feature channels, leveraging global information to strengthen important information and reduce the influence of non-key, useless information on pornographic image recognition results. The SE module mainly consists of two parts: Squeeze and Extraction.

The SE Module is a computing unit based on the convolution operator F_{tr}. It maps the input $X \in R^{W'*H'*C'}$ to a characteristic graph $U \in R^{W\times H\times C}$. The H denotes the height of the image, the W denotes the image's width, and the C means the image channel. The first step of the SE module is the convolution operation, as shown in Eq. (4). The $V = [v_{1,}v_2,\ldots,v_c]$ is a group of learned convolution kernels. The x^s denotes the sth input under the current convolution kernel coverage. The v_c represent the parameters of the cth convolution kernel. The input is $X = [x_{1,}x_2,\ldots,x_c]$. The output is

$U = [u_{1,}u_2, \ldots, u_c]$.

$$\mathrm{u_c} = v_c * X = \sum_{S=1}^{C'} v_c^s * x^s \tag{4}$$

The U obtained by convolution operation cannot use the context information of the receptive field. Squeeze employs average pooling to squeeze each channel's features as the channel's descriptor, utilizing Eq. (5) to average the features in the channel to yield an output vector representing the distribution of C feature mappings in this layer.

$$z_c = F_{Sq}\left(u_c\right) = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c\left(i,j\right) \tag{5}$$

To take advantage of the information aggregated in the squeeze operation, perform the excitation operation to capture the channel dependency completely. Firstly, the weight is generated for each feature channel through parameters. A fully connected layer operation is performed by multiplying the result z obtained from the previous squeeze operation. The dimension of W_1 is $C \times C/r$, and the r is a scaling parameter, its size is $1 \times 1 \times C$, which can reduce the number of channels and the amount of calculation. The result of $W_1 z$ is $1 \times 1 \times C/r$. Through a ReLU activation layer, the output dimension remains the same. Secondly, a fully connected operation is performed. Due to the size of W_2 is $C \times C/r$, the output size is $1 \times 1 \times C$. Finally, the weighting coefficient s of C feature vectors in a vector U can be obtained through the sigmoid function. The excitation operation can be expressed by Eqs. (6) and (7).

$$s = F_{ex}(z, W) = \delta(g(z, W)) = \delta(W_2 \delta(W_1 z)) \tag{6}$$

$$\widetilde{x}_c = F_{scale}\left(u_c, s_c\right) = s_c u_c \tag{7}$$

SE-ResNet is used to extract the features of images. Each image is divided into instances, and the images are output as feature vectors after passing through the network.

3.4 Attention Pooling

Pooling operation mainly reduces each feature dimension by compressing the number of data and parameters, maintains the vital information of the image, and reduces the overfitting problem. The pooling layer has no parameters. It performs the downsampling step of the results from the upper layers. Average pooling and maximum pooling are two commonly used downsampling methods.

Both of these two pooling operations have the problems of pre-definition and are untrainable. Maximum pooling is unsuitable for embedded-based multi-instance learning, and average pooling has disadvantages in clustering instance scores. In order to meet the requirement of pooling layer operation in multi-instance learning, this paper introduces the idea of an attention mechanism, and the instances are weighted and averaged so that the pooling layer can be flexibly applied. The neural network determines the weight parameters of the pooling layer, and the weight of the entire package is guaranteed to be added to 1. The scoring function $S\left(x\right) = \mathrm{g}\left(\sum_{x \in X} f\left(x\right)\right)$ of a group of instances in a bag is symmetric. Generally, the formula of pooling operation in MIL is shown in Eqs. (8) and (9). $B = \{b_1, b_2 \ldots, b_n\}$ denotes a bag, w and V are network weight parameters, $\tanh(\cdot)$ is a hyperbolic tangent nonlinear function. To achieve good convergence and reduce the difficulty of training, we replace $\tanh(\cdot)$ in traditional pooling operations with $\tanh(\mathrm{x}) \cdot \mathrm{sigm}(\mathrm{x})$. Sigmoid is a nonlinear activation function, which can reduce the error of $\tanh(\cdot)$ in feature extraction and improve the robustness of classification.

$$z = \sum_{n=1}^{N} a_n b_n \tag{8}$$

$$a_n = \frac{\exp\left\{w^T \tanh\left(V {b_n}^T\right)\right\}}{\sum_{j=1}^{N} \exp\left\{w^T \tanh\left(V {b_j}^T\right)\right\}} \tag{9}$$

3.5 Fine-Grained Classification and Recognition Method

The previous studies based on DL often consider pornographic image recognition as a binary classification problem, disregarding the various characteristics of pornographic images. When training network models, if all pornographic images are combined into one category during training, it may cause confusion when learning high-level semantic features of the network models, resulting in inaccurate classifications of certain images and impaired performance of the network model. Thus, based on the analysis of the varied features of pornographic images, this paper develops a pornographic image set with seven distinct classifications, as listed in Table 2. This results in more precise classification output than binary classifications.

Table 2: Pornographic image recognition category

Classification code	Category description
1	Normal images with a human body
2	Normal images without a human body
3	Images of male lower body sexual organ exposure
4	Image of female upper body breast exposure
5	Images of female lower body sexual organ exposure
6	Sexual behavior poses images containing a large number of naked skins.
7	Other vulgar, pornographic images, such as close-up images of sensitive parts of underwear

Because the SoftMax classifier works best when the features between categories are mutually exclusive, this paper uses the SoftMax classifier as the network's last layer. It could output the probability values that the image is classified into a specific type, and the probability values range from 0 to 1. First, the probability of various classifications is obtained through the network model. Second, find the classification with the highest probability and then determine whether the image is pornographic based on the classification number and probability. Categories 3, 4, 5, and 6 are the most apparent characteristics and principal components of pornographic images. After empirical analysis, it is found that the threshold for judging pornographic images is set to 0.7, which is suitable. Category 7 mainly includes some low-level images, and the boundary between them and pornographic images is fuzzy. In order to reduce false positives, the threshold for judging that pornographic images should be appropriately increased. After the experiment, it was set to 0.8. The specific calculation process is shown in Algorithm 1.

Algorithm1. Pornographic image recognition.

```
procedure Last layer
    Input: C, V. (The C represents the classification category number with the highest probability
            value, and the V represents the probability of output category C)
    Output: 0, 1 (the 0 means the input image is not pornographic, and the 1 represents the input
            image is pornographic)
    IF the C is less than or equal to two
        RETURN 0
    ELSE IF the C is less than seven and the V is greater than or equal to 0.7
        RETURN 1
    ELSE IF the C is equal to seven and the V is greater than or equal to 0.8
        RETURN 1
    ELSE
        RETURN 0
    END IF
end procedure
```

4 Experiments

4.1 Experimental Dataset

Compared with other types of images, pornographic images have certain particularities in terms of legal constraints, and there are very few publicly available experimental datasets. Many studies regard the pornographic image recognition task as a binary classification problem. In this paper, a high-quality dataset is constructed to check the performance of the proposed method. Pornographic images are crawled from some pornographic websites, while normal images are obtained from the Baidu image library. These images are then manually labeled according to Table 2's classification standards in order to improve the network's generalization ability by covering a variety of topics, such as animals, cars, landscapes, and buildings. The experimental dataset is outlined in Table 3.

Table 3: Data set details

Classification code	1	2	3	4	5	6	7
Total images	13523	7382	4212	4315	4428	4226	2623
Training images	10818	5906	3610	3452	3542	3381	2098
Testing images	2705	1476	902	863	886	845	525

4.2 Experimental Settings

This paper proposes a pornographic image recognition method based on MIL and residual neural networks, which can improve the network's performance in spatial dimensions and increase the interdependence between feature channels. An attention mechanism is also introduced. This performs a weighted average operation on instances to meet the requirements of MIL. In order to evaluate the effectiveness of the proposed method, ablation experiments are conducted using mathematical statistics and analysis. The experiments are carried out on an HP graphics workstation with an ubuntu 18.04 operating system, 64 GB memory, a Xeon W2223 processor, and Nvidia Quadro RTX3090 24G

GPU. The network model is constructed with the open-source DL tool Caffe, implemented in Python 3.8. The practical steps are as follows:

(1) Establish training and test image sets. All images are scaled to 224 × 224 pixels. 80% of images from the dataset are selected randomly to form the model training set, while the remaining images are used as the test set. To enhance the feature information and avoid overfitting phenomenon during network model training, horizontal image flipping, rotation, random noise, slight modifications in image brightness, Gaussian blur, and other processing methods are used. This process converts one image in the training data set into 32 images. In order to improve test speed, one image in the test dataset is randomly converted into two images using one of the data enhancement methods.
(2) As the constructed dataset is small, the model is first pre-trained on the ImageNet1000 dataset. Then transfer training is carried out on the training set [32]. Each image is segmented into 16 instances during the training. The initial learning rate is set to 0.01, the momentum to 0.9, and the weight decay to 0.0005. The weight decay is lowered by ten times every 50 epochs to optimize our models.
(3) To check the effect of adding an SE module to the pornographic image recognition method proposed in this paper based on MIL. VGG16, ResNet50, ResNet101, ResNet152, and their models combined with SE modules are used to conduct comparative experiments on our dataset.
(4) A set of comparison experiments are established to verify the effect of pooling operations with an attention mechanism. One experiment uses improved pooling operations, and the other uses traditional ones. All experiments are conducted on a dataset specially constructed for this purpose.
(5) To verify the influence of the proposed multi-classification method on the performance of pornographic image recognition, VGG16, ResNet50, ResNet101, and ResNet152 network models containing SE modules are constructed. The traditional binary classification algorithm and the multi-classification algorithm proposed in this paper are then used to train the constructed network model. Finally, the trained network model is used to conduct a comparative experiment of pornographic image recognition on the test image set.
(6) To compare the effectiveness of the pornographic image recognition methods proposed in this paper, we revise the dataset with traditional binary labels. We select four representative algorithms from prior studies: A [11], ROI-based recognition with color saliency features; B [7], One-Class classification model with a visual attention mechanism; C [18], a depth convolution neural network for classifying images as pornographic, sexy and normal; D [25], weighted MIL problem in a neural network framework. For convenience, these algorithms are denoted as A to D.

4.3 Experimental Evaluation Metrics

The results of pornographic image recognition can be divided into four types:

1) The image is pornographic and is predicted to be pornographic; TP represents this type.
2) The image is not pornographic and is predicted to be pornographic; FP represents this type.
3) The image is pornographic and is predicted not to be pornographic; FN represents this type.
4) The image is not pornographic and is predicted not to be pornographic; TN represents this type.

In the practical application of pornographic image recognition, TP and TN cannot simply be used to evaluate the algorithm's performance. Therefore, referring to previous research methods, this paper uses the accuracy rate and F1-score as the evaluation indicators for experimental results. The specific calculation method is shown in Eqs. (10) and (11).

$$Accuracy = \frac{TP + TN}{Total} \times 100\% \tag{10}$$

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{11}$$

In Formula 10, the Total represents the total number of test images. The evaluation standard of experimental results is that the higher the Accuracy and F1 score, the better.

5 Discussion

5.1 Experimental Analysis of SE Module

This paper introduces the SE module to improve the accuracy of image feature extraction. To verify the effectiveness of the SE module, four DL models, VGG16, ResNet50, ResNet101, and ResNet152, and their models combined with SE modules, are respectively used to conduct comparative experiments on our dataset. The experimental results are shown in Table 4. The experimental results are visualized to facilitate comparison, as shown in Figs. 3 and 4.

Table 4: Analysis of the impact of the SE module

Approaches	SE module	Accuracy (%)	F1-score
VGG16	Yes	90.73	0.912
VGG16	No	88.73	0.882
ResNet-50	Yes	92.28	0.923
ResNet-50	No	91.47	0.918
ResNet-101	Yes	94.67	0.947
ResNet-101	No	93.51	0.931
ResNet-152	Yes	96.58	0.964
ResNet-152	No	95.31	0.953

(1) Table 4 shows the accuracy and F1-score of four pornographic image recognition approaches based on DL with and without the SE module. These results demonstrate that the operation cost of residual networks with SE modules above 101 layers is better than that of VGG networks with 16 layers. The pornographic image recognition performance is excellent. ResNet-50, ResNet101, and ResNet152 use a Bottleneck structure that can adapt to the dimensions during network training and target vectors to those vectors that pass through the network, thus saving calculation costs. Table 3 illustrates that ResNet-152 has the best performance in pornographic image recognition. It demonstrates that the residual network improves its performance through spatial dimension with increasing network depth. It can aggregate features from various receptive fields to obtain better performance gains and extract image features better.

(2) It can be found that results with SE modules have higher accuracy and F1-score than those without, regardless of the depth of the network. The SE module can bring performance gains to deep learning by obtaining the weight parameters of different feature channels and using global information to strengthen critical information while weakening the impact of non-key useless information.

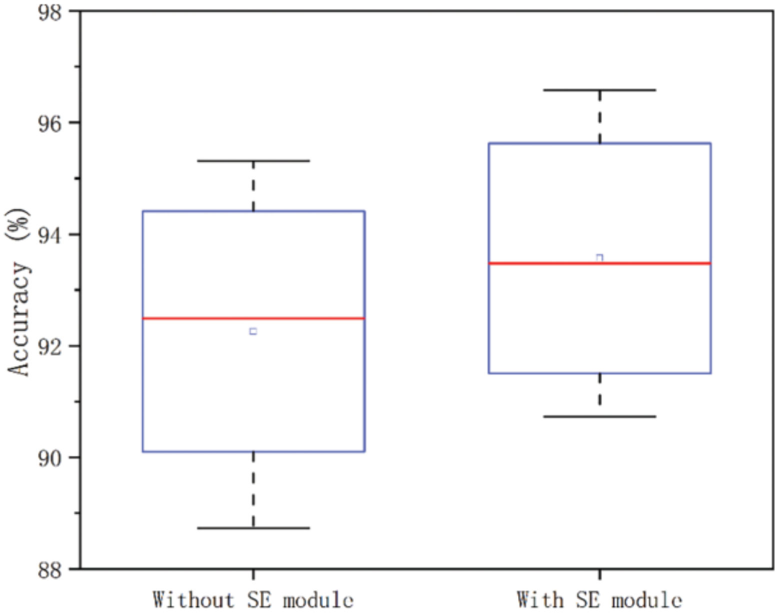

Figure 3: The accuracy distribution comparison of four pornographic image recognition approaches based on DL with and without the SE module

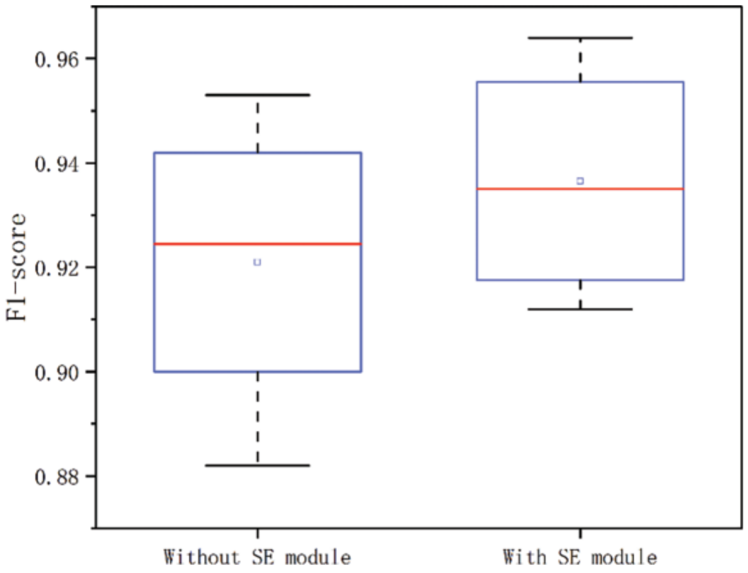

Figure 4: The F1-score distribution comparison of four pornographic image recognition approaches based on DL with and without the SE module

5.2 Experimental Analysis of Improved Pooling Operation

Based on the problems of the maximum pooling and average pooling operations described above in the application of MIL, this paper introduces the idea of an attention mechanism. Weighted average operations are then carried out on instances to make pooling operations adapt to the requirements of MIL. In order to verify the effect of the improved pooling operation on image feature extraction, this paper uses three pooling operations for model training based on the SE-ResNet152-MIL network. These algorithms are Ins-Max (maximum pooling based on the instance) and Ins-Avg (average pooling based on the instance). The experimental results are shown in Table 5. The accuracy and F1-score of pornographic image recognition were improved with the improved pooling operation. This is due to the introduction of the attention mechanism and the average weighted operation of instances. This caused the model to strengthen the learning of key instances and better represent the characteristics of pornographic images.

Table 5: Analysis of the impact on different pooling operations

Approaches	Accuracy (%)	F1-score
Ins-Max	96.58	0.964
Ins-Avg	90.12	0.895
Our approach	97.21	0.972

5.3 Experimental Analysis of Fine-Grained Multi-Classification

The experimental results of the fifth step in Section 4.2.2 are shown in Table 6 and visualized in Figs. 5 and 6, which demonstrate that the pornographic image classification method proposed in this paper enhances the accuracy and F1-score of pornographic image recognition regardless of the chosen network model. This is likely attributable to the network model learning to extract high-level image semantic features from a fine-grained pornographic classification dataset.

Table 6: Analysis of the impact on different classifications

Approaches	Classification method	Accuracy (%)	F1-score
VGG16	Binary	88.92	0.890
VGG16	Our method	90.73	0.912
ResNet-50	Binary	90.65	0.908
ResNet-50	Our method	92.28	0.923
ResNet-101	Binary	92.46	0.918
ResNet-101	Our method	94.67	0.947
ResNet-152	Binary	93.87	0.934
ResNet-152	Our method	97.21	0.972

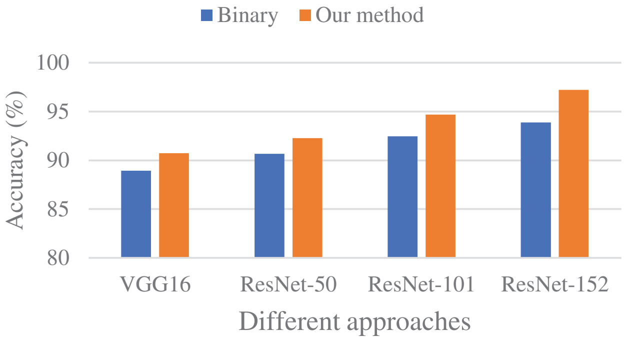

Figure 5: The accuracy comparison of four pornographic image recognition approaches based on DL with the SE module

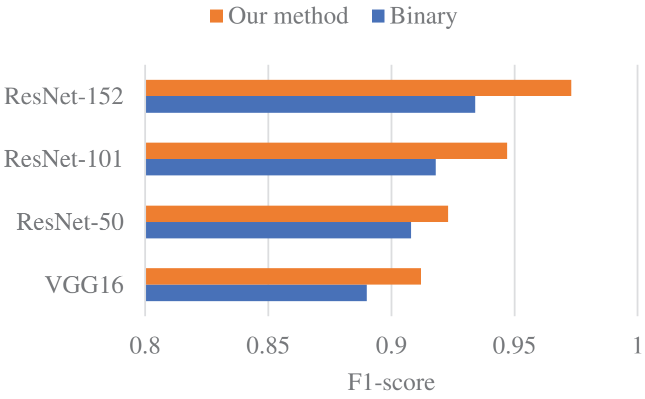

Figure 6: The F1-score comparison of four pornographic image recognition approaches based on DL with the SE module

5.4 Comprehensive Comparison Experiments

To comprehensively compare the effectiveness of the pornographic image recognition method proposed in this paper, relevant comparative experiments were conducted according to experimental step 6 in Section 4.2.2. The experimental results are shown in Table 7 and Fig. 7.

According to the experimental results, method A has underperformed, which might be due to its recognition of pornographic images being based on ROI and thus susceptible to the effects of light intensity, skin color, and different image backgrounds. On the other hand, methods B, C, and D have all achieved good results. Method D has the highest score because it combines the advantages of CNN and MIL for weighting pornographic image recognition. Our proposed MIL and multi-classification-based pornographic image recognition method has achieved even better results than the rest, which could be attributed to the use of a multi-classified image database for model training allowing the network to learn high-level semantic features of images, thereby improving the accuracy of pornographic image recognition.

Table 7: Comparison results with other methods

Approaches	Accuracy (%)	F1-score
A [11]	80.02	0.792
B [7]	95.62	0.954
C [18]	94.58	0.942
D [25]	96.12	0.962
Our approach	97.21	0.972

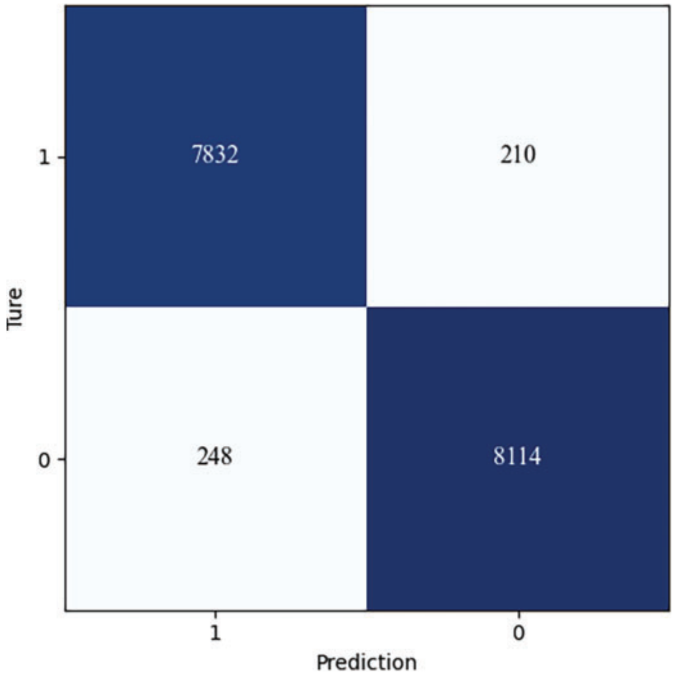

Figure 7: Confusion matrix for prediction by our approach

6 Conclusion and Future Directions

This paper proposes a weakly supervised pornographic image recognition method based on MIL. First of all, we combine the different characteristics of pornographic images by constructing a pornographic image set that includes seven classifications and converting the binary pornographic image recognition task into a multi-classification task, then determining whether the image is pornographic based on the output score. In order to solve the issues of low accuracy and poor network adaptability of the original MIL algorithm in pornographic image recognition, a model is built based on MIL and residual neural network while introducing a SE module to the channel to strengthen key information and reduce the impact of non-key useless information on the recognition results. Additionally, the pooling operation in traditional multi-instance learning tasks is improved, an attention mechanism is introduced, and a weighted average operation is implemented on the instances to make the pooling operation conform to MIL requirements. For the final step, several groups of experiments were carried out. The experimental results show that the pornographic image recognition method proposed in this paper has good performance in accuracy and F1-score compared with other methods. The proposed method of this paper can provide a reference for developing network security tools for automatic pornographic image recognition.

However, the proposed method in this paper has only been tested on the test set created by us and not on any other datasets. Future research should consult the most recent results to be applied in practical applications to further verify and improve upon them. Additionally, category 7 mainly includes some low-level images, and the boundary between them and pornographic images is fuzzy. Compared with other categories, there is no 1:1 balance in quantity. Due to time and energy constraints, the generated pornographic image dataset was not very large. The classification is slightly rough, such as cartoon images, child pornographic images, and sexually suggestive images are not subdivided. It is still uncertain whether further refinement of image classification can improve the effectiveness of pornographic image recognition; some relevant research can be done in the future.

Funding Statement: This work is supported by the Academic Research Project of Henan Police College (Grant: HNJY-2021-QN-14 and HNJY202220) and the Key Technology R&D Program of Henan Province (Grant: 222102210041).

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

References

[1] W. Z. Khan, Q. Arshad, S. Hakak, M. K. Khan and R. Saeed Ur, "Trust management in social internet of things: Architectures, recent advancements, and future challenges," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7768–7788, 2021.

[2] S. Karamizadeh and A. Arabsorkhi, "Methods of pornography detection: Review," in *Proc. ICCMS*, Sydney, NSW, Australia, pp. 33–38, 2018.

[3] C. C. Yan, Y. Z. Liu, H. T. Xie, Z. H. Liao and J. Yin, "Extracting salient region for pornographic image detection," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1130–1135, 2014.

[4] D. Ganguly, M. H. Mofrad and A. Kovashka, "Detecting sexually provocative images," in *Proc. WACV*, pp. 660–668, 2017.

[5] J. J. Yu and S. W. Han, "Skin detection for adult image identification," in *Proc. ICACT*, pp. 645–648, 2014.

[6] A. Gangwar, V. Gonzalez-Castro, E. Alegre and E. Fidalgo, "AttM-CNN: Attention and metric learning based CNN for pornography, age and child sexual abuse (CSA) detection in images," *Neurocomputing*, vol. 445, pp. 81–104, 2021.

[7] J. R. Chen, G. Liang, W. B. He, C. Xu, J. Yang *et al.,* "A pornographic images recognition model based on deep one-class classification with visual attention mechanism," *IEEE ACCESS*, vol. 8, pp. 122709–122721, 2020.

[8] M. A. Chyad, H. A. Alsattar, B. B. Zaidan, A. A. Zaidan and G. A. Al Shafeey, "The landscape of research on skin detectors: Coherent taxonomy, open challenges, motivations, recommendations and statistical analysis, future directions," *IEEE ACCESS*, vol. 7, pp. 106536–106575, 2019.

[9] M. M. Fleck, D. A. Forsyth and C. Bregler, "Finding naked people," in *Proc. ECCV*, Berlin, Heidelberg, pp. 593–602, 1996.

[10] H. A. Nugroho, D. Hardiyanto and T. B. Adji, "Negative content filtering for video application," in *Proc. ICITEE*, King Mongkut's Institute, Chiang Mai, Thailand, pp. 55–60, 2015.

[11] C. N. Tian, X. N. Zhang, W. Wei and X. B. Gao, "Color pornographic image detection based on color-saliency preserved mixture deformable part model," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 6629–6645, 2018.

[12] L. Zhuo, Z. Geng, J. Zhang and X. G. Li, “ORB feature based web pornographic image recognition,” *Neurocomputing*, vol. 173, pp. 511–517, 2016.

[13] X. J. Shen, W. Wei and Q. J. Qian, “The filtering of internet images based on detecting erotogenic-part,” in *Proc. ICNC*, Haikou, Peoples R China, pp. 732, 2007.

[14] A. A. Zaidan, H. A. Karim, N. N. Ahmad, B. B. Zaidan and A. Sali, “An automated anti-pornography system using a skin detector based on artificial intelligence: A review,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 27, no. 4, pp. 1–4, 2013.

[15] X. Lin, F. Qin, Y. Peng and Y. Shao, “Fine-grained pornographic image recognition with multiple feature fusion transfer learning,” *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 1, pp. 73–86, 2021.

[16] H. Pranoto, Y. Heryadi, H. Warnars and W. Budiharto, “Enhanced IPCGAN-Alexnet model for new face image generating on age target,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 7236–7246, 2022.

[17] R. M. Alguliyev, F. J. Abdullayeva and S. S. Ojagverdiyeva, “Image-based malicious internet content filtering method for child protection,” *Journal of Information Security and Applications*, vol. 65, pp. 103–123, 2022.

[18] F. Cheng, S. L. Wang, X. Z. Wang, A. W. C. Liew and G. S. Liu, “A global and local context integration DCNN for adult image classification,” *Pattern Recognition*, vol. 96, 2019.

[19] R. L. Zhu, X. Y. Wu, B. B. Zhu and L. Y. H. Song, “Application of pornographic images recognition based on depth learning,” in *Proc. ICISS*, South Korea, pp. 152–155, 2018.

[20] X. Z. Wang, F. Cheng, S. L. Wang, H. R. Sun, G. S. Liu *et al.,* “Adult image classification by a local-context aware network,” in *Proc. ICIP*, Athens, Greece, pp. 2989–2993, 2018.

[21] O. Surinta and T. Khamket, “Recognizing pornographic images using deep convolutional neural networks,” in *Proc. ICDA*, 2019.

[22] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, pp. 770–778, 2016.

[23] X. Wang, Y. Yan, T. Peng, B. Xiang and W. Liu, “Revisiting multiple instance neural networks,” *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[24] M. Ilse, J. M. Tomczak and M. Welling, “Attention-based deep multiple instance learning,” in *Proc. ICML*, 2018.

[25] X. Jin, Y. H. Wang and X. Y. Tan, “Pornographic image recognition via weighted multiple instance learning,” *IEEE Transactions on Cybernetics*, vol. 49, no. 12, pp. 4412–4420, 2019.

[26] A. T. Pham, R. Raich, X. Z. Fern, W. K. Wong and X. Guan, “Discriminative probabilistic framework for generalized multi-instance learning,” in *Proc. ICASSP*, Calgary, Canada, pp. 2281–2285, 2018.

[27] Q. Bi, K. Qin, Z. Li, H. Zhang and G. S. Xia, “A multiple-instance densely-connected convnet for aerial scene classification,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4911–4926, 2020.

[28] T. Vu, P. Lai, R. Raich, A. Pham and U. A. Rao, “A novel attribute-based symmetric multiple instance learning for histopathological image analysis,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3125–3136, 2020.

[29] S. Pal, A. Valkanas, F. Regol and M. Coates, “Bag graph: Multiple instance learning using Bayesian graph neural networks,” in *Proc. AAAI*, pp. 7922–7930, 2022.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* “An image is worth 16 × 16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, pp. 101–121, 2021.

[31] J. Hu, L. Shen and G. Sun, “Squeeze-and-excitation networks,” in *Proc. ICVPR*, 2018.

[32] R. Sharma, “Using transfer learning to classify pornographic images,” in *Proc. CDMA*, 2020.