applied sciences

MDPI

Article

RepVGG-SimAM: An Efficient Bad Image Classification Method Based on RepVGG with Simple Parameter-Free Attention Module

Zengyu Cai [1], Xinyang Qiao [1], Jianwei Zhang [2,3,*], Yuan Feng [1], Xinhua Hu [2] and Nan Jiang [1]

[1] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450003, China; czy@zzuli.edu.cn (Z.C.); 332207050676@email.zzuli.edu.cn (X.Q.); fy@zzuli.edu.cn (Y.F.); jiangnan@zzuli.edu.cn (N.J.)
[2] School of Software Engineering, Zhengzhou University of Light Industry, Zhengzhou 450003, China; 332113020734@email.zzuli.edu.cn
[3] Research Institute of Industrial Technology, Zhengzhou University of Light Industry, Zhengzhou 450003, China
* Correspondence: ing@zzuli.edu.cn

Abstract: With the rapid development of Internet technology, the number of global Internet users is rapidly increasing, and the scale of the Internet is also expanding. The huge Internet system has accelerated the spread of bad information, including bad images. Bad images reflect the vulgar culture of the Internet. They will not only pollute the Internet environment and impact the core culture of society but also endanger the physical and mental health of young people. In addition, some criminals use bad images to induce users to download software containing computer viruses, which also greatly endanger the security of cyberspace. Cyberspace governance faces enormous challenges. Most existing methods for classifying bad images face problems such as low classification accuracy and long inference times, and these limitations are not conducive to effectively curbing the spread of bad images and reducing their harm. To address this issue, this paper proposes a classification method (RepVGG-SimAM) based on RepVGG and a simple parameter-free attention mechanism (SimAM). This method uses RepVGG as the backbone network and embeds the SimAM attention mechanism in the network so that the neural network can obtain more effective information and suppress useless information. We used pornographic images publicly disclosed by data scientist Alexander Kim and violent images collected from the internet to construct the dataset for our experiment. The experimental results prove that the classification accuracy of the method proposed in this paper can reach 94.5% for bad images, that the false positive rate of bad images is only 4.3%, and that the inference speed is doubled compared with the ResNet101 network. Our proposed method can effectively identify bad images and provide efficient and powerful support for cyberspace governance.

Keywords: bad image classification; RepVGG; SimAM attention mechanism; cyberspace governance

Citation: Cai, Z.; Qiao, X.; Zhang, J.; Feng, Y.; Hu, X.; Jiang, N. RepVGG-SimAM: An Efficient Bad Image Classification Method Based on RepVGG with Simple Parameter-Free Attention Module. *Appl. Sci.* **2023**, *13*, 11925. https://doi.org/10.3390/app132111925

Academic Editors: Jian Xu and Ruijin Wang

Received: 16 October 2023
Revised: 26 October 2023
Accepted: 30 October 2023
Published: 31 October 2023

1. Introduction

At present, the number of global Internet users has exceeded 5.16 billion, equivalent to 64.4% of the world's total population. The huge Internet system has accelerated the spread of bad information and increased its impact. Violent images and pornographic images are the most common forms of bad information on the Internet. Violent images are generally referred to as images that contain violence, gore, cruelty, horror, abuse, death, etc. Such images may cause panic, disgust, pain, sadness, anger, and other undesirable emotions, which will have a negative impact on an individual's psychological, emotional, and spiritual health [1]. Pornographic images are images that depict or show sexual acts, sexual organs, or the exposure of sexual organs. Such images usually invoke strong sexual suggestions or sexual stimulation, which can easily arouse people's sexual impulses and

curiosity [2]. Bad images not only have a negative impact on individuals' psychology and emotions but may also have a negative impact on social order and morality. The identification and detection of bad images is of great significance to purifying cyberspace, ensuring cyberspace security, and improving governments' digital governance capabilities.

In recent years, artificial intelligence technology has ushered in a period of vigorous development, and significant progress has been made in using artificial intelligence technology to identify bad images. Early research on bad images mostly focused on machine learning methods. Pornographic-image-recognition-based machine learning uses the proportion of exposed skin [3–6] in an image and the special shapes and textures [7–9] of human private parts as the basis for recognition. The main research object of violent-image-recognition-based machine learning is bloody images. In this approach, large red areas [10] that may appear in an image are used for recognition. The method based on machine learning takes the special features of bad images as the basis for recognition, but these special features are not unique to bad images, so the recognition accuracy of these methods is not ideal. Deep learning has achieved remarkable results in the field of computer vision. Thanks to the powerful feature extraction ability of deep neural networks, they can discover high-dimensional structural information [11] within data. Deep learning has achieved success in the fields of classification [12,13], object detection [14,15], image retrieval [16], etc. Pornographic image recognition schemes based on deep learning and violent image recognition mostly use deep network structures [17] or deep residual structures [18] for feature extraction. However, these complex neural network architectures bring a lot of inference time overhead. In the current high-speed and wide-coverage computer network environment, it is difficult to effectively suppress the spread of bad images and reduce the harm they cause.

In order to deal with the above limitations, this paper proposes an efficient bad image classification method (RepVGG-SimAM). In this method, the RepVGG model is used as the backbone neural network, and the SimAM attention mechanism is added to the backbone network. The neural network model uses the residual structure in the training phase to improve the feature extraction ability of the network, and it also prevents the occurrence of training overfitting. In the inference stage, the multi-branch residual structure is transformed into a single-channel model using structural reparameterization technology, which improves the speed of inference. The network model can not only achieve deep feature extraction but also meet the needs of high-speed inference.

The main contributions of this paper are as follows:

- An efficient bad image recognition method is proposed. This method separates training process from inference process and achieves both powerful feature extraction and inference speed. In addition, the residual structure of the training process not only alleviates the problem of gradient disappearance but also improves the convergence speed [19] of the network.
- Adding the SimAM attention mechanism to the original network increases the scrutiny of important features and does not introduce parameters, improving the network effect while still maintaining the original inference speed.
- The experiments show that this method has high reasoning accuracy and low time consumption and can effectively and quickly detect bad images in the network environment, providing effective support for cyberspace governance.

This paper is divided into five chapters to introduce RepVGG-SimAM, an efficient bad image classification method based on RepVGG with a SimAM attention mechanism. The first section mainly introduces the research background, the significance of bad image recognition in the process of cyberspace governance, and the contributions of this paper. It also outlines the arrangement of the content of each section in this article. The second chapter mainly introduces the related research status of bad images as well as the RepVGG model and the SimAM attention mechanism hierarchy. The third chapter mainly introduces the model architecture of our proposed RepVGG-SimAM model and then the structural reparameterization technology and calculation method of the SimAM attention mechanism.

The fourth chapter first introduces the components of the data set used in this paper and the evaluation index of the experiment; then, it introduces the parameter settings in the model-training stage and finally discusses and analyzes the experimental results. The fifth section first summarizes the RepVGG-SimAM bad image recognition algorithm proposed in this paper and then looks forward to the future work.

2. Related Work

In recent years, countries around the world, including China, have made cyberspace governance an important part of national governance. The early detection and processing of bad images provide favorable guarantees for cyberspace governance. Researchers from various countries have conducted extensive theoretical and technical research on the recognition and classification of bad images, among which the most representative concerns machine learning methods, especially deep learning methods, which can efficiently and accurately identify bad images.

2.1. Method Based on Traditional Machine Learning

Machine learning has been widely applied in various fields. Machine learning also plays an important role in the field of bad image classification, including in relation to K-neighbor algorithms [20], SVM algorithms [21], and Bayesian algorithms [22]. Bad images include pornographic and violent images. Pornographic images often contain a lot of nudity-related information. Based on this feature, Jones et al. [3] established a color model in the RGB color space and determined whether a region in an image was a skin area based on the brightness of each channel of pixel color. On this basis, Lin et al. [4] input the correlation between skin regions and non-skin regions into an SVM, which improved the accuracy of pornographic image recognition. In [5,6], researchers found that RGB color space is greatly affected by light, and using YCbCr color space can improve the clustering degree of skin color. However, such methods will misjudge normal images, such as faces and swimsuits, as bad images, so there is still room for improvement. Using the unique textures and shapes of sensitive parts of the human body, Zhao et al. [7] proposed the fusion of texture and SIFT features to detect pornographic images, thereby improving the reliability of detection. In [8], the researchers proposed a texture-based BoWV model for bad image filtering. Lv et al. [9] improved the BoVW model by fusing high-level semantic features to filter pornographic images. However, the contours and textures of sensitive parts of the human body are not unique. The false positive rate of methods based on contours and textures is not ideal. In addition, violent pictures are mostly bloody. Yan et al. [10] proposed a region-based blood color detection algorithm that extracts color and texture features from the bloody regions of an image and then inputs them into the SVM classifier. This method does not work well for image recognition applied to images with a large area of red scenes such as red flags or red paint.

2.2. Method Based on Deep Learning

The deep learning method extracts the deep features of an image via constructing a deep neural network, which is the mainstream method in the field of bad image recognition. In terms of pornographic image classification, Zefeng Ying et al. [23] applied a convolutional neural network to the recognition of pornographic images. A CNN network was trained for bad image recognition, and a method based on a deconvolution network was used to optimize performance in different scenarios. Li et al. [24] performed pornography detection by fusing four DenseNet121 models [25]. Compared with a single DenseNet121 network, the model was improved by about 1%. Cai et al. [26] added a CBAM attention module to the ResNet101 network to classify bloody and pornographic images. This approach effectively avoids the problem of gradient disappearance during deep network training. In the detection of violent images, most of the current methods are used to detect violent videos. However, these methods all obtain video frames from videos and use them as an experimental input. Therefore, these methods are essentially detecting violent images.

Mumtaz et al. [27] used GoogleNet as a backbone network and replaced the fully connected layer and classifier of the original network with a binary classifier. This network achieved better classification results than previous methods. Jebur et al. [28] used the pre-trained Xception, Inception, and InceptionResNet networks as a feature extraction layer; fused the features extracted from the three models; and finally used the classifier for classification. YE et al. [29] used C3D neural networks to extract features from video and sound sequences in campus violence scenes and then used an improved Dempster–Shafer approach to fuse the two features. This method could achieve high-accuracy detection of campus violence incidents. Although a deep neural network can enhance the representation ability of a network and improve the accuracy of recognition, the complexity of the network will increase with the deepening of the network levels, which will seriously affect the inference speed of a neural network. A summary of the references is shown in Table 1.

Table 1. Overview of references.

Characteristics	References	Methods	Limitations
Classification based on skin area	[3]	RGB color space	The characteristics of classification basis are not restricted to bad images, and the accuracy of classification is not ideal.
	[4]	RGB + SVM	
	[5,6]	YCbCr color space	
Classification based on shape and texture	[7]	fusing texture and SIFT features	
	[8]	BoWV	
	[9]	BoVWS	
Classification based on red area	[10]	bloody area + SVM	
	[23]	CNN	Deep neural networks increase the complexity of the network and affect its inference speed
	[25]	fusing four DenseNet121	
	[26]	RenNet101 + CBAM	
	[27]	GoogleNet	
	[28]	fusing Xception, Inception, InceptionResNet	
	[29]	C3D	

2.3. RepVGG

A convolutional neural network realizes the feature extraction and dimension reduction of input data through a multi-layer convolution operation and a pooling operation. In 2012, the success of AlexNet [30] in an image recognition competition led researchers to believe that the deeper the layers in a neural network, the better its performance. However, the experiment by He Kaiming et al. [18] revealed that as the number of layers increases toward a maximum, the accuracy of the network greatly decreases with the increase in network depth. Based on this, He Kaiming's team proposed a neural network, ResNet, with a branch structure. However, this multi-branch structure increased the complexity of the network and seriously affected its inference speed. RepVGG [31] is based on the concepts of the ResNet and VGG networks. It aims to solve the problems of high computational complexity and poor accuracy in existing networks. The training phase of the RepVGG network draws on the multi-branch structure of the residual network, and inferences are made based on a single-path model. Its training and inferences are decoupled using the structural reparameterization technique. The architecture of the RepVGG network is shown in Figure 1. As shown in the figure, the training phase of the RepVGG network consists of five phases, each of which is superimposed on a RepVGG-R1 structure and multiple RepVGG-R2 structures. In the inference stage, the residual structures of RepVGG-R1 and RepVGG-R2 are equivalently transformed into a 3×3 convolution matrix using the structural reparameterization technique. In this figure, N1, N2, and N3 represent the number of layers in these structures.

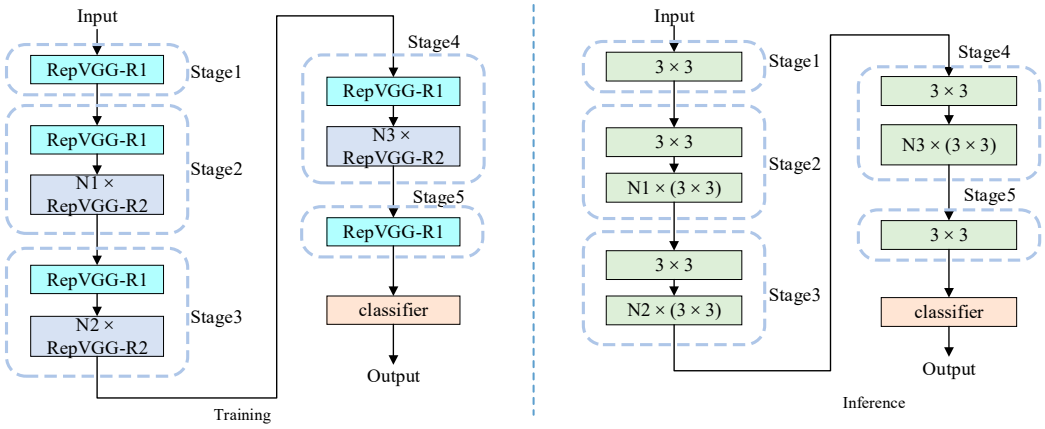

Figure 1. RepVGG network architecture.

As mentioned earlier, RepVGG-R1 is a residual structure containing a 1 × 1 convolution matrix, as indicated by R1 in Figure 2. RepVGG-R2 is a residual structure containing a 1 × 1 convolution and identity (an identity-mapping layer), as indicated by R2 in Figure 2.

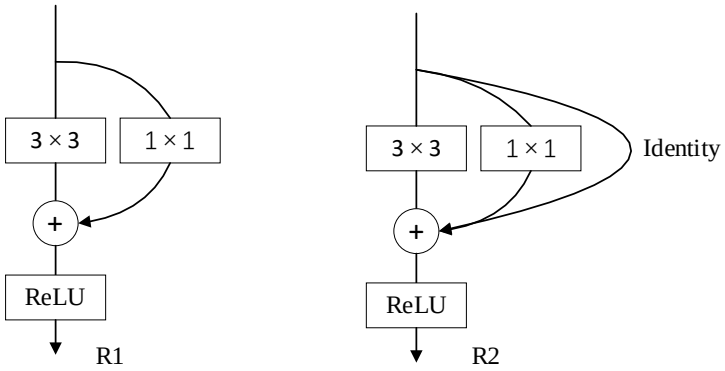

Figure 2. Residual structure.

2.4. SimAM

Traditional attention mechanisms include spatial attention mechanisms, channel attention mechanisms, and SENet, where the spatial attention mechanism focuses on the importance of different regions, the channel attention mechanism focuses on the representation ability of different channels [32], and SENet focuses on the relationship between channels through a compression incentive mechanism [33]. Adding these attention modules to a network can grant the network stronger representation ability, but new parameters will have been introduced. Although they optimize the network, they increase the complexity of the network, and this will affect the inference speed and performance of the network. The SimAM [34] attention module is a simple, efficient, and lightweight three-dimensional attention module. Different from the attention mechanism mentioned above, SimAM infers three-dimensional (considering both spatial and channel dimension correlation) attention weights through feature mapping [35] in the feature layer, without adding parameters to the original network. The model of SimAM is shown in Figure 3. The SimAM module will not add additional parameters. Adding SimAM to the network will not increase the complexity of the network. To a certain extent, it can increase the representation ability of the original network and will not affect the training and inference speed of a network.

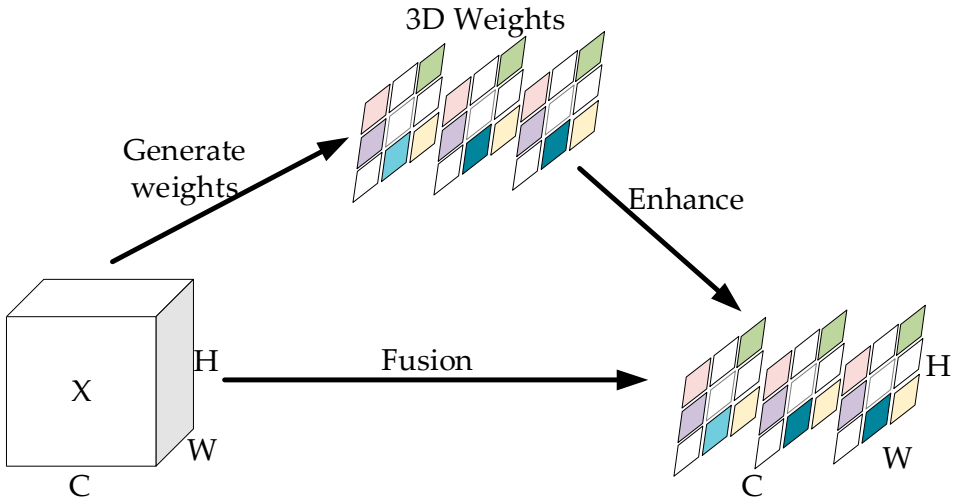

Figure 3. SimAM attention machine model.

3. Proposed Method (RepVGG-SimAM)

The RepVGG network has fast inference speed, allowing it to quickly identify bad images in cyberspace and reduce their impact. In this study, the SimAM attention mechanism was added to the RepVGG network to improve a network's performance without changing its inference speed.

3.1. RepVGG-SimAM Network Architecture

In this paper, we chose the RepVGG-A2 network as the backbone network for transformation. RepVGG-A2 is a lightweight model with five stages of feature extraction. Stage 1 and Stage 5 only have one layer, while Stage 2, Stage 3 and Stage 4 have two layers, four layers, and fourteen layers, respectively. RepVGG-A2 has a total of 22 layers. We added the SimAM attention mechanism after Stage 1 and Stage 5. The improved network architecture of RepVGG-A2(RepVGG-SimAM) is shown in Figure 4.

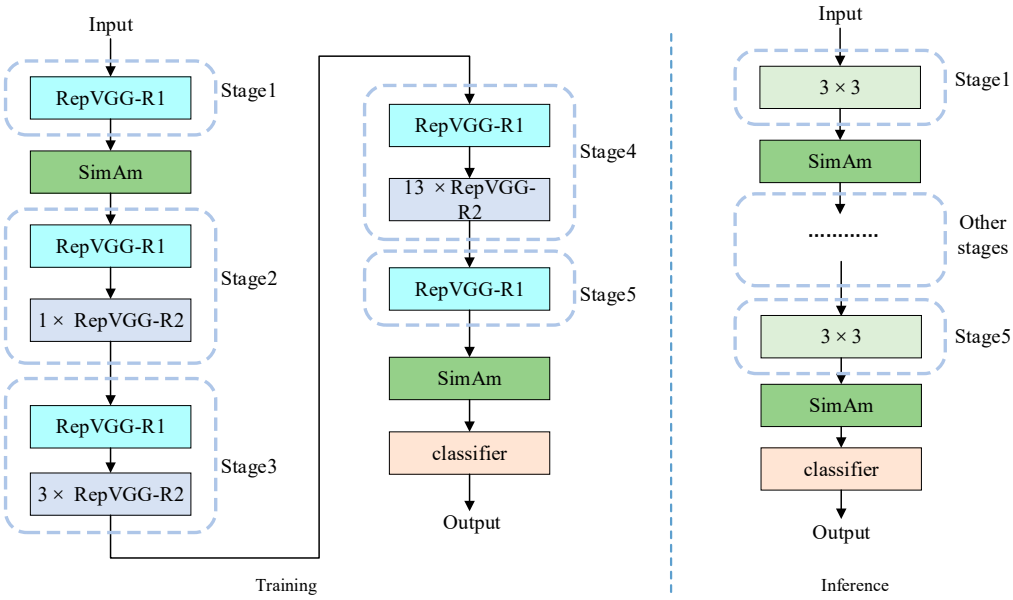

Figure 4. RepVGG-SimAM network architecture diagram.

The RepVGG-SimAM network uses a more complex multi-branch structure in the training phase, which not only allows it to obtain deep feature representation but also more adequately solve the gradient disappearance problem of a deep network. However, the residual branch structure needs to save a large number of intermediate results and

introduce more parameters, which will lead to slow speed and low memory efficiency when a deep residual network is used for inferencing. The RepVGG-SimAM network model no longer uses the multi-branch structure in the inference stage; instead, it uses the structural reparameterization concept to transform the multi-branch structure into a single-channel structure, combining the advantages of the strong representation ability of multi-branch models and the fast inference speed of single-branch models.

3.2. RepVGG-SimAM Network Structure Configuration

In accordance with the above architecture, the network structure configuration table proposed in this paper is presented in Table 2. In Table 2, $\partial \times (\beta - \omega)$ indicates that there are ∂ residual structures, ω (ω is the residual structure R1 or R2), with β channels in this layer. As for the first layer of Stage 1, there is one residual structure, R1, with a total of 96 channels.

Table 2. RepVGG-A2 network structure configuration.

Stage	Output Size	First Layer of This Stage	Other Layers of This Stage
1	112 × 112	1 × (96−R1)	
2	56 × 56	1 × (96−R1)	1 × (96−R2)
3	28 × 28	1 × (192−R1)	3 × (192−R2)
4	14 × 14	1 × (384−R1)	13 × (384−R2)
5	7 × 7	1 × (768−R1)	

In this study, the SimAM module was added to stage 1 and stage 5 of the RepVGG-A2 network so that the network can pay more attention to important neurons in the training process and reduce the weight of neurons with lower utility. At the same time, because the SimAM module does not need to introduce new parameters and does not change the size of the input feature map, the SimAM module can be easily embedded into the RepVGG network layer. The improved RepVGG-A2 network (RepVGG-SimAM) structure configuration is shown in Table 3.

Table 3. RepVGG-SimAM network structure configuration.

Stage	Output Size	First Layer of This Stage	Other Layers of This Stage
1	112 × 112	1 × (96−R1)	SimAM
2	56 × 56	1 × (96−R1)	1 × (96−R2)
3	28 × 28	1 × (192−R1)	3 × (192−R2)
4	14 × 14	1 × (384−R1)	13 × (384−R2)
5	7 × 7	1 × (768−R1)	SimAM

3.3. Structural Reparameterization

The structure reparameterization technique is an equivalent parameter conversion technique [36]. In RepVGG-A2, the parameters of the multi-branch structure are equivalently converted into the parameters required by the single-channel model. This conversion method retains the accuracy of the original multi-branch model and improves the speed of inference. The process of the reparameterization of RepVGG-A2 is divided into three steps. Firstly, the convolution layer and BN layer are fused. Secondly, the fused branch convolution kernels are transformed into a 3 × 3 convolution. Finally, the multi-branch convolutions are added to form a single convolution layer. The process of the RepVGG-A2 network structure's reparameterization is shown in Figure 5.

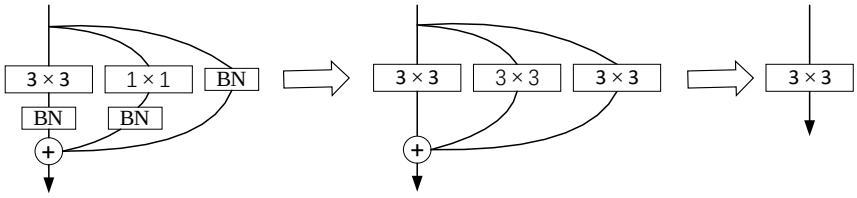

Figure 5. Structural reparameterization process.

The steps of the reparameterization of the RepVGG-A2 network's structure are as follows.

3.3.1. Fusion of Convolutional Layer and NB Layer

Equation (1) is the calculation equation of the convolution layer. In this equation, W is the convolution kernel, and b is the offset.

$$Conv(x) = W(x) + b \tag{1}$$

This definition uses γ to represent the scaling factor, $mean$ to represent the mean, var to represent the variance, and β to represent the bias, and the equation of the BN layer can be expressed as shown in Equation (2).

$$BN(x) = \gamma \times \frac{(x - mean)}{\sqrt{var}} + \beta \tag{2}$$

Bringing the results of convolution layer into the equation for the BN layer can be expressed as shown in Equation (3). Simplifying the equation yields Equation (4).

$$BN(Conv(x)) = \gamma \times \frac{W(x) + b - mean}{\sqrt{var}} + \beta \tag{3}$$

$$BN(Conv(x)) = \frac{\gamma \times W(x)}{\sqrt{var}} + \left(\frac{\gamma \times (b - mean)}{\sqrt{var}} + \beta \right) \tag{4}$$

Here, define $W_{fused} = \frac{\gamma \times W(x)}{\sqrt{var}}$, and $B_{fused} = \frac{\gamma \times (b - mean)}{\sqrt{var}} + \beta$. The fused equation can be obtained in the form of Equation (5).

$$BN(Conv(x)) = W_{fused} + B_{fused} \tag{5}$$

From Equation (5), it is not difficult to find that the fusion of the convolutional layer and the BN layer leads to a new convolution.

3.3.2. Convert Branch to 3 × 3 Convolutional Kernels

After fusing the BN layers, there is still a 1 × 1 convolution and an identity branch in all branches. In order to facilitate the fusion of multiple branches, the 1 × 1 convolution and the identity branch need to be transformed into a 3 × 3 convolution. The 1 × 1 convolution needs to put the 1 × 1 convolution kernel into the center of a 3 × 3 convolution kernel and place zeros at the other positions. The identity branch does not perform any operations on the original feature map, so it only needs to use a 3 × 3 convolution kernel, with each member having a value of 1, to replace the branch.

3.3.3. Fusion of Branches

By following the steps above, each branch will be replaced by an equivalent 3 × 3 convolution. Because the convolution operation is additive, the three convolution operations are added to obtain a single convolution.

3.4. SimAM Attention Mechanism

The SimAM attention module calculates the importance of neurons in the network and assigns higher weights to neurons with greater importance. The module defines an

energy function $e_t(*)$ to evaluate the linear separability between the target neuron and other neurons. The equation is as follows:

$$e_t(*) = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{6}$$

In Equation (6), t represents the target neuron, $\hat{\sigma}^2$ represents the variance of other neurons except t, $\hat{\mu}$ represents the mean of other neurons except t, and λ is the coefficient. According to this formula, the lower the energy of the neuron, the more separable it is from other neurons, and the more important it is. The original feature map is enhanced via Equation (7).

$$\widetilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X \tag{7}$$

In Equation (7), E is the sum of the energy functions of each channel, and sigmoid is used to constrain excessive values of E. The calculation process of the Simam module is shown in Figure 6.

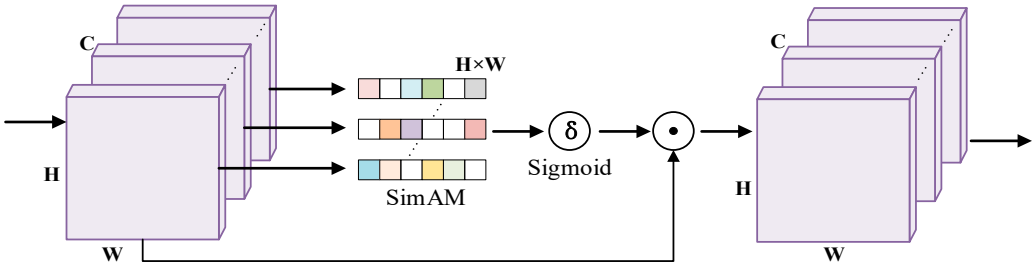

Figure 6. SimAM calculation process.

4. Experiment and Result Analysis

4.1. Experimental Data Sets

Due to the limitations of policies and regional regulations, there were few high-quality public data sets that could be used in this study. The data set used in this paper consists of three parts, in which the pornographic and normal images were derived from the public data set project published by data scientist Alexander Kim on GitHub (https://github.com/alexkimxyz/nsfw_data_scraper (accessed on 24 April 2023)), and the violent images were obtained from the Internet and through manual screening. Figure 7 is an example of this experimental data set.

Figure 7. Dataset example.

The data set published by Alexander Kim includes five categories, namely, drawings, hentai, neutral, porn, and sexy, all of which are given in the form of network addresses of the images. We used the method recommended by the author to build the environment in which to download pornographic and neutral category images as part of the data set of this experiment. Unfortunately, due to the special nature of pornographic images, there are a large number of network address failures. There are still problems such as repetition and mismatch between images and categories in the successfully acquired images. After cleaning the data, we obtained 1296 pornographic images and 937 normal images. In order to ensure the symmetry of the number of images in each data category, we collected 990 violent images from the Internet to construct the data set of this paper. In this paper, the image data set was enhanced by changing the brightness, chromaticity, contrast, sharpness, and fuzzy noise [37] and via mirror transformation; the number of images was increased under the premise of ensuring the integrity of the image features. The final data set contains 11,476 pornographic images, 9290 normal images, and 8415 violent images. The data for each category are shown in Figure 8.

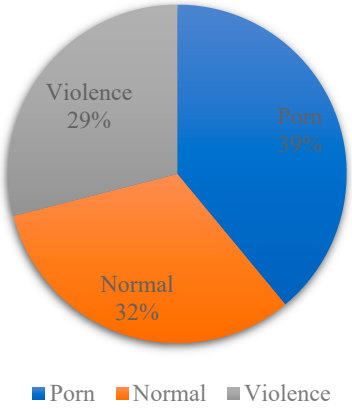

Figure 8. The proportions of pictures in various categories.

4.2. Eevaluation Criteria

In order to evaluate the performance of the different models in all aspects, in this paper, we employ the accuracy, precision, recall, F1, and inference time of the model classification as the evaluation criteria. In particular, it should be pointed out that due to the rapid spread of modern networks and the large scale of Internet users, the harm caused by misreporting bad images as normal images is much greater than that caused by identifying normal images as bad images. Based on this, this paper further introduces the false alarm rate of bad images as the evaluation standard of this experiment on the basis of classification accuracy, precision rate, recall rate, F1 value, and reasoning time.

The calculation formula of classification accuracy in this paper is as follows:

$$Accurary = \frac{P_C}{P_A} \times 100\% \tag{8}$$

Here, P_C represents the number of correct images predicted by the model in the validation dataset, and P_A represents the total number of validation images incorporated in the experiment.

In order to comprehensively evaluate the classification performance of the model, we used the $F1$ value to reconcile the precision rate and the recall rate. The formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

Here, TP is the number of samples in which the model predicts positive classes as positive classes, FN is the number of samples in which the model predicts positive classes as negative classes, and FP is the number of samples in which the model predicts negative classes as positive classes.

We set up multiple sets of experimental data, with each group containing a different number of images, and then input each set of data into the inference model to record the respective running times. The calculation formula for inference time is as follows:

$$Time_i = Tend_i - Tstart_i \tag{12}$$

Here, $Tend_i$ represents the end time of group i, and $Tstart_i$ represents the beginning time of group i.

In the real network environment, if bad images are misreported as normal pictures, discord will be sown in society following their dispersal in the network environment. We classify pornographic images and violent images as bad images, and the false positive rate of bad images can be expressed as

$$FAR(False\ alarm\ rate) = \frac{P_{FAR}}{P_A} \times 100\% \tag{13}$$

where P_{FAR} represents the number of bad images misreported as normal images, and P_A represents the total number of images in the validation set.

In order to ensure the accuracy of the experimental data, all the experimental data were obtained by averaging the same hardware equipment after conducting training many times in the same training environment.

4.3. Training Environment and Parameter Configuration

All the experiments were performed on a server using an Intel Xeon CPU E5-2680 v4@2.40GHz, 256 GB DDR4 memory, an NVIDIA TITAN RTX graphics processing unit (GPU) with 24 GB memory, and an Ubuntu operating system. All experiments were performed using Python 3.9 and NVIDIA CUDA-12.0, and the compiler environment was developed using the PyTorch 2.0.1 deep learning framework.

The setting of parameters in the model-training process has a great influence on the performance of a model. Appropriate parameter settings can accelerate the convergence speed of a model and improve the accuracy of classification. The parameter settings in the training process are shown in Table 4. We set the epoch to 100 and the batch_size to 32. We used the cosine annealing method to dynamically adjust the learning rate, and we used the stochastic gradient descent method (SGD) as an optimizer. In order to prevent a local gradient extremum in the process of training, momentum was introduced to execute gradient correction.

Table 4. Training parameter settings.

Parameters	Setting
Optimizer	SGD
Scheduler	Cosine annealing
Learning rate	0.001
Batch size	32
Epoch	100
Momentum	0.9

In the model-training stage, the image size is first adjusted to 224 × 224 and converted into a feature tensor. The data set is divided into a training set and a verification set

according to a ratio of 8:2, and the divided data are input into the model in batches for training after disrupting the order.

4.4. Experimental Result Analysis

4.4.1. Training Process Analysis

The variation curves of loss during the training process and accuracy during the verification process are shown in Figure 9. It can be seen from this figure that the RepVGG-SimAM model is faster than the RepVGG-A2 model in terms of convergence speed, and as the number of iterations increases, the accuracy of the RepVGG-SimAM model improves. This is mainly due to the SimAM attention module. The addition of SimAM enables the network to focus on image features that positively contribute to the model classification results while ignoring useless features, thereby improving the accuracy of classification and lowering the value of loss function calculations.

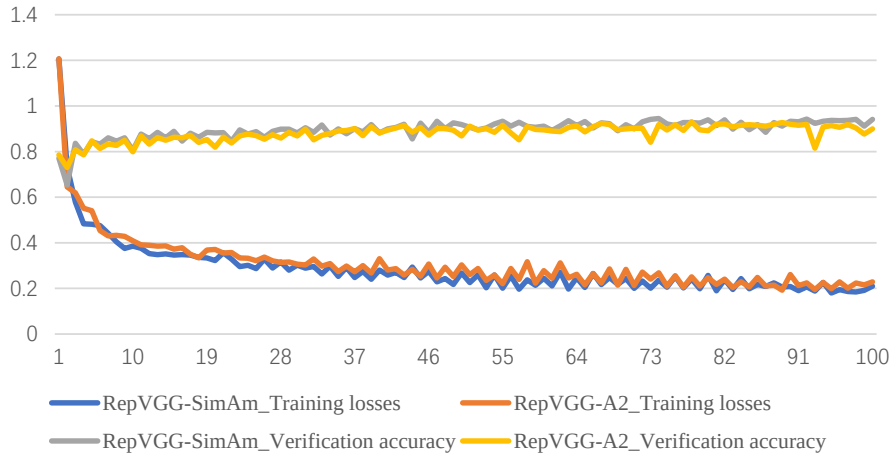

Figure 9. The variation curves of loss during the training process and accuracy during the verification process.

4.4.2. Comparison of Existing Algorithms

We conducted a comparative experiment between some methods mentioned in the literature and our method proposed in this paper. The experimental results are shown in Table 5. Through a comparison of the classification accuracy in this table, we can find that the method mentioned in this paper can achieve ideal results when applied to the experimental data set and results in a greater improvement in classification accuracy than the traditional machine learning method. Compared with other deep learning methods, it also achieved an accuracy improvement of more than 1%. In terms of the false alarm rate for bad images, the advantages of this method are more obvious. The traditional method has a high false alarm rate for bad images, which also proves that some special attributes of bad images are not unique; thus, they cannot be fully used as the classification basis for bad images.

Table 5. Comparison of existing algorithms.

Paper	Method	Accuracy (%)	FAR (%)
[3]	RGB-skin	61.0	-
[4]	RGB-skin + svm	75	35
[6]	YCbCr-skin	75.4	10.13
[7]	texture	87.6	14.7
[9]	BoVWS	87.6	-
[23]	CNN	86.9	-
[26]	ResNet101 + CBAM	93.2	-
Ours	RepVGG-SimAM	94.5	4.3

4.4.3. Comparison of Different Classification Algorithms

In order to verify the effectiveness of the model proposed in this paper, in this experiment, we compared our model with the VGG series and ResNet series models, and the test results are shown in Table 6. The experimental data in Table 6 show that the RepVGG-SimAM proposed in this paper is superior to other experimental models in terms of precision, recall, and F1 value in all categories. In addition, the actual Internet environment is more sensitive to the false alarm rate of bad pictures, so we also set up experiments to compare the different methods' false alarm rates (FARs) for bad pictures. The experimental data show that the model proposed in this paper performs best in terms of the FAR. The experimental results show that our method is superior to other deep neural network models in terms of feature extraction ability.

Table 6. Classification results of each model.

Methods	Precision (%)			Recall (%)			F1 (%)			FAR (%)	Accuracy (%)
	Porn	Viol	Norm	Porn	Viol	Norm	Porn	Viol	Norm		
VGG11	90.1	91.9	83.2	88.9	94	80.9	89.5	92.9	82.1	9.8	89.4
VGG13	87.6	92.3	82.6	88.9	93.9	78.2	88.3	93.1	80.3	10.2	89.8
VGG16	87.5	91.7	80.1	89.9	92.8	77.5	88.7	92.3	78.7	9.7	88.7
VGG19	89.8	90.1	81.3	86.8	94.5	79.5	88.4	92.3	80.4	9.6	88.5
Resnet26	87.4	92.1	78.7	88.6	91.3	78.9	87.9	91.7	78.8	10.9	87.7
Resnet34	89.3	92	77.8	88.7	90.5	79.1	88.9	91.3	78.4	12.1	86.9
Resnet50	88	92.4	78.1	88.2	91.4	80	88.1	91.8	79	10.9	87.7
Resnet101	87.8	92	78.8	86.9	91.1	79.8	87.4	91.6	79.3	10.6	87.9
RepVGG-A2	90.1	93.8	82.9	89.4	94.3	83.9	90.1	94.1	83.4	6.5	90.8
RepVGG-SimAM	92.4	94.9	85.2	90.5	95.1	86.4	91.4	95	85.8	4.3	94.5

In order to reduce the harm caused by the spread of bad pictures, the inference speed of the models is also an important evaluation index. In order to verify the inference speed of our method, we chose VGG11, VGG13, RseNet50, and ResNet101 for comparison with our method. The inference time of each model is shown in Figure 10. It can be seen from Figure 10 that the inference time of our proposed method is less than that of the other models, and with the increase in the number of images, the advantage of this model will be greater. This is mainly because although the VGG series models are single-channel models, the deep network structure will affect the reasoning speed of a model. Otherwise, the ResNet series models use more residual branch structures, and the neural network has high complexity and poor inference speed.

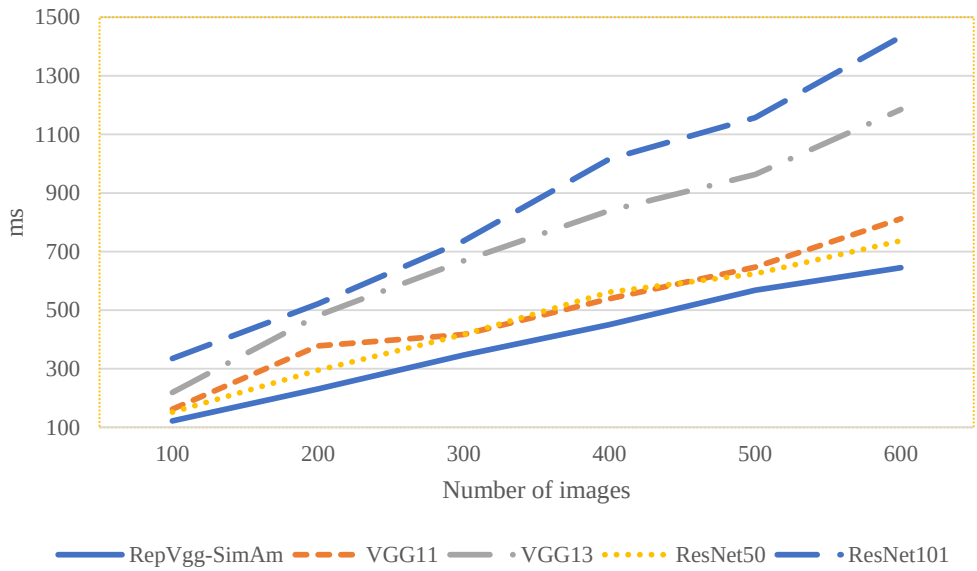

Figure 10. Model inference time.

In summary, through experimental comparison, it has been found that the RepVGG-SimAM network model proposed in this paper is superior to the traditional method in terms of inference accuracy, the false alarm rate for bad images, and time consumed; it can better adapt to the current development of the Internet.

5. Conclusions

This paper proposes an efficient bad image classification model (RepVGG-SimAM) that solves the problem wherein accuracy and inference speed cannot be obtained simultaneously in the process of bad image classification. In this model, training is separated from inferencing, and the transformation from a training model to an inference model was realized using structural reparameterization technology. This grants RepVGG-SimAM the powerful feature extraction ability of a multi-branch model and the fast inference speed of a single model. Our experimental results show that the model is superior to other models in terms of inference accuracy, its false positive rate for bad images, inference time, and other evaluation indicators; it is more suitable for the current needs of cyberspace governance than other models. Of course, there are also shortcomings in this paper. The content of the data set used in this paper is not diverse enough to fully cover the various forms of bad images. In the future, we will further collect and produce more fine-grained data sets to achieve more fine-grained classification of neural networks.

Author Contributions: Conceptualization, Z.C. and X.Q.; methodology, Z.C. and X.Q.; validation, J.Z., X.H. and N.J.; data curation, Y.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62072416), the Key Research and Development Special Project of Henan Province (221111210500), and the Key Technolgies R&D Program of Henan Province (232102211053, 222102210170, and 222102210322).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Xu, X.; Wu, X.; Wang, G.; Wang, H. Violent Video Classification Based on Spatial-Temporal Cues Using Deep Learning. In Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; pp. 319–322.
2. Cheng, F.; Wang, S.; Wang, X.; Liew, A.W.; Liu, G. A global and local context integration DCNN for adult image classification. *Pattern Recognit.* **2019**, *96*, 106983. [CrossRef]
3. Jones, M.J.; Rehg, J.M. Statistical Color Models with Application to Skin Detection. *Int. J. Comput. Vis.* **2002**, *46*, 81–96. [CrossRef]
4. Lin, Y.C.; Tseng, H.W.; Fuh, C.S. Pornography Detection Using Support Vector Machine. In Proceedings of the 16th IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP 2003), Kinmen, China, 17–19 August 2003; pp. 123–130.
5. Wang, B.S.; Lv, X.Q.; Ma, X.L.; Wang, H.W. Application of Skin Detection Based on Irregular Polygon Area Boundary Constraint on YCbCr and Reverse Gamma Correction. *Adv. Mater. Res.* **2011**, *327*, 31–36. [CrossRef]
6. Basilio, J.A.M.; Torres, G.A.; Gabriel, S.P.; Medina, L.T.; Meana, H.M. Explicit Image Detection Using YCbCr Space Color Model as Skin Detection. In Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications, Puerto Morelos, Mexico, 29–31 January 2011; pp. 123–128.
7. Zhao, Z.; Cai, A. Combining multiple SVM classifiers for adult image recognition. In Proceedings of the 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content, Beijing, China, 24–26 September 2010; pp. 149–153.
8. Deselaers, T.; Pimenidis, L.; Ney, H. Bag-of-Visual-Words Models for Adult Image Classification and Filtering. In Proceedings of the 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
9. Lv, L.; Zhao, C.; Lv, H.; Shang, J.; Yang, Y.; Wang, J. Pornographic Images Detection Using High-Level Semantic Features. In Proceedings of the 2011 Seventh International Conference on Natural Computation, Shanghai, China, 26–28 July 2011; pp. 1015–1018.
10. Gao, Y.; Wu, O.; Wang, C.; Hu, W.; Yang, J. Region-Based Blood Color Detection and Its Application to Bloody Image Filtering. In Proceedings of the 2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Guangzhou, China, 12–15 July 2015; pp. 45–50.
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

12. Zhao, F.; Zhang, J.; Meng, Z.; Liu, H.; Chang, Z.; Fan, J. Multiple vision architectures-based hybrid network for hyperspectral image classification. *Expert Syst. Appl.* **2023**, *234*, 121032. [CrossRef]
13. Gao, Y.; Rezaeipanah, A. An Ensemble Classification Method Based on Deep Neural Networks for Breast Cancer Diagnosis. *Intel. Artif.* **2023**, *26*, 160–177. [CrossRef]
14. Bharat, M.; Mishra, K.K.; Anoj, K. An improved lightweight small object detection framework applied to real-time autonomous driving. *Expert Syst. Appl.* **2023**, *234*, 121036.
15. Wang, C.; Wang, Q.; Qian, Y.; Hu, Y.; Xue, Y.; Wang, H. DP-YOLO: Effective Improvement Based on YOLO Detector. *Appl. Sci.* **2023**, *13*, 11676. [CrossRef]
16. Xie, J.; Chen, J.; Cai, Y.; Huang, Q.; Li, Q. Visual Paraphrase Generation with Key Information Retained. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–19. [CrossRef]
17. Xie, G.; Lai, J. An Interpretation of Forward-Propagation and Back-Propagation of DNN. In Proceedings of the Pattern Recognition and Computer Vision. PRCV 2018, Guangzhou, China, 23–26 November 2018; pp. 3–15.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
19. Ullah, K.; Saleem, N.; Bilal, H.; Ahmad, J.; Ibrar, M.; Jarad, F. On the convergence, stability and data dependence results of the JK iteration process in Banach spaces. *Open Math.* **2023**, *21*, 20230101. [CrossRef]
20. Guo, G.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. KNN model-based approach in classification. In Proceedings of the OTM Confederated International Conferences, CoopIS, DOA, and ODBASE(2003), Catania, Italy, 3–7 November 2003; pp. 986–996.
21. Huang, G.; Zhou, H.; Ding, X.; Zhang, R. Extreme Learning Machine for Regression and Multiclass Classification. *IEEE Trans. Syst. Man Cybern. Part B* **2012**, *42*, 513–529. [CrossRef]
22. Zhao, H.; Liu, I. Research on test data generation method of complex event big data processing system based on Bayesian network. *Comput. Appl. Res.* **2018**, *35*, 155–158.
23. Ying, Z.; Shi, P.; Pan, D.; Yang, H.; Hou, M. A Deep Network for Pornographic Image Recognition Based on Feature Visualization Analysis. In Proceedings of the 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 14–16 December 2018; pp. 212–216.
24. Lin, X.; Qin, F.; Peng, Y. Fine-grained pornographic image recognition with multiple feature fusion transfer learning. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 73–86. [CrossRef]
25. Sheena, C.P.; Sindhu, P.K.; Suganthi, S.; Saranya, J.; Selvakumar, V.S. An Efficient DenseNet for Diabetic Retinopathy Screening. *Int. J. Eng. Technol. Innov.* **2023**, *13*, 125–136. [CrossRef]
26. Cai, Z.; Hu, X.; Geng, Z.; Zhang, J.; Feng, Y. An Illegal Image Classification System Based on Deep Residual Network and Convolutional Block Attention Module. *Int. J. Netw. Secur.* **2023**, *25*, 351–359.
27. Mumtaz, A.; Sargano, A.B.; Habib, Z. Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning. In Proceedings of the 2018 2nd European Conference on Electrical Engineering and Computer Science (EECS), Bern, Switzerland, 20–22 December 2018; pp. 558–563.
28. Jebur, S.A.; Hussein, K.A.; Hoomod, H.K.; Alzubaidi, L. Novel Deep Feature Fusion Framework for Multi-Scenario Violence Detection. *Computers* **2023**, *12*, 175. [CrossRef]
29. Ye, L.; Liu, T.; Han, T.; Ferdinando, H.; Seppänen, T.; Alasaarela, E. Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences. *Remote Sens.* **2021**, *13*, 628. [CrossRef]
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
31. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13728–13737.
32. Woo, S.; Park, J.; Lee, J.Y. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 3–9.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
34. Yang, L.; Zhang, R.; Li, L.; Xie, X. SimAM: A simple parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 18–24 July 2021; pp. 11863–11874.
35. Ishtiaq, U.; Saleem, N.; Uddin, F.; Sessa, S.; Ahmad, K.; di Martino, F. Graphical Views of Intuitionistic Fuzzy Double-Controlled Metric-Like Spaces and Certain Fixed-Point Results with Application. *Symmetry* **2022**, *14*, 2364. [CrossRef]
36. Yu, X.; Wang, X.; Rong, J.; Zhang, M.; Ou, L. Efficient Re-Parameterization Operations Search for Easy-to-Deploy Network Based on Directional Evolutionary Strategy. *Neural Process. Lett.* **2023**, 1–24. [CrossRef]
37. Saleem, N.; Ahmad, K.; Ishtiaq, U.; De la Sen, M. Multivalued neutrosophic fractals and Hutchinson-Barnsley operator in neutrosophic metric space. *Chaos Solitons Fractals* **2023**, *172*, 113607. [CrossRef]