Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.041550

ARTICLE

# AnimeNet: A Deep Learning Approach for Detecting Violence and Eroticism in Animated Content

**Yixin Tang***

Department of Cooperative Course of Performance, Film & Animation, Sejong University, Seoul, 05006, Korea

*Corresponding Author: Yixin Tang. Email: 21170957@sju.ac.kr

Received: 27 April 2023 Accepted: 04 August 2023 Published: 31 October 2023

**ABSTRACT**

Cartoons serve as significant sources of entertainment for children and adolescents. However, numerous animated videos contain unsuitable content, such as violence, eroticism, abuse, and vehicular accidents. Current content detection methods rely on manual inspection, which is resource-intensive, time-consuming, and not always reliable. Therefore, more efficient detection methods are necessary to safeguard young viewers. This paper addresses this significant problem by proposing a novel deep learning-based system, AnimeNet, designed to detect varying degrees of violent and erotic content in videos. AnimeNet utilizes a novel Convolutional Neural Network (CNN) model to extract image features effectively, classifying violent and erotic scenes in videos and images. The novelty of the work lies in the introduction of a novel channel-spatial attention module, enhancing the feature extraction performance of the CNN model, an advancement over previous efforts in the literature. To validate the approach, I compared AnimeNet with state-of-the-art classification methods, including ResNet, RegNet, ConvNext, ViT, and MobileNet. These were used to identify violent and erotic scenes within specific video frames. The results showed that AnimeNet outperformed these models, proving it to be well-suited for real-time applications in videos or images. This work presents a significant leap forward in automatic content detection in animation, offering a high-accuracy solution that is less resource-intensive and more reliable than current methods. The proposed approach enables it possible to better protect young audiences from exposure to unsuitable content, underlining its importance and potential for broad social impact.

**KEYWORDS**

Computer vision; animation; deep learning; classification; attention mechanism

## 1 Introduction

### *1.1 Motivation*

It is well-documented that seeing violence and explicit content has a negative impact on society [1]. Children's thoughts may be influenced in a favorable or unfavorable way by the animation films on the multimedia platform [2]. While these materials may be used to educate children, they also have the capacity to entirely captivate children and teenagers, leading to less time spent interacting with others. Despite the fact that most children's content is good, entertaining, or educational, recent research has highlighted the development of dangerous content [3,4]. A well-known instance of this phenomenon is

the Elsagate incident [5,6], in which con artists combined popular animated characters like Spiderman, Disney's Frozen, Mickey Mouse, and others with unpleasant features like light violence and explicit behavior. According to research, children who often watch violent cartoons and movies are less sensitive to pain and more prone to act aggressively [7]. These films cause them to start behaving aggressively at home and at school. For example, bullying is a widespread problem, especially in schools, and it has a devastating effect on everyone involved. Moreover, the increase of explicit scenes in animation content also affects the well-being of small children.

The term "explicit scene" refers to anything, such as nudity and sex, that is objectionable, inexcusable, or goes against socially acceptable norms. As a result, material made for kids has to be given specific attention. The current filtering procedure for an explicit scene involves manually screening the whole piece of material. Furthermore, some explicit scenes are merely romantic scene that has been considered reasonable for children to watch and falls under the 'U' (unrestricted) category [8]. Therefore, the animation industry does not ban certain explicit scenes from animation content.

Despite existing research on the impact of explicit content, there's a lack of comprehensive mechanisms for detecting such content in animations. The current procedure for filtering explicit scenes relies heavily on manual screening. While some explicit scenes are marked as 'U' (unrestricted) because they're considered reasonable for children [9], the industry lacks a rigorous, standardized method to filter out inappropriate scenes.

There is a crucial need for automated deep learning-based classification systems to identify and filter explicit or violent content in animation films. This will help to protect children from exposure to inappropriate content and its potential adverse effects. Automated monitoring should be designed to accurately detect violent or explicit scenes and categorize them based on their level of explicitness. The monitoring system may take benefit from a comprehensive database of animation content marked with their explicit levels to train the deep learning models.

### *1.2 Related Work*

A wide variety of methodologies have been proposed to enable automatic violence or explicit scene detection in cartoon images, each with its own advantages, disadvantages, and areas of application. In general, these methods can be grouped into two categories: violent content detection and pornographic content detection.

In terms of violent content detection, Ditsanthia et al. [10] and Sumon et al. [11] both utilized a combination of deep convolutional networks and long short-term memory (LSTM) for violence detection. While these works achieved impressive accuracies, they were both limited by their lack of consideration for the degree of violence and specificity of real-world violence. Chen et al. [12] extended these models by integrating 3D convolutional neural networks (CNN) and support vector machines (SVM), though at a high computational cost that hindered real-time applications. Ye et al. [13] took a different approach by focusing on feature extraction methods, yet their work also did not extend to animated or cartoon videos.

Two works stand out for their explicit attention to cartoon violence. Yousaf et al. [14] proposed an EfficientNet-BiLSTM method for detecting and classifying inappropriate cartoon content. However, they only classified full frames, thereby potentially discarding useful information from non-inappropriate portions of the frame. Meanwhile, Khan et al. [15] developed a MobileNet-based violence detection model for cartoons, which despite achieving high accuracy, showed overfitting issues, limiting its real-world application.

For pornographic content detection, Wang et al. [16] and Mohamed [17] both proposed CNN-based models, although their approaches face difficulties with real-time detection and robust classification, respectively. Perez et al. [18] and Yuan et al. [19] developed models that achieved high accuracies on their respective datasets, but the former's model required the fusion of multiple CNNs and an SVM, and the latter's dataset was not substantial enough for comprehensive classification. Finally, Shen et al. [20] presented an ensemble framework with an uncertain evidence-based method, though they, like all others in this category, did not account for cartoon and animation videos. The detailed content is shown in Table 1.

**Table 1:** Comparative overview of various machine learning and deep learning methods applied for content detection and classification, highlighting their specific goals, key contributions, and notable limitations

| Method | Detection goal | Contribution |
|---|---|---|
| | Machine learning | |
| SVM [12] | Human action recognition | • A new spatiotemporal two-stream 3D CNN structure<br>• Superior performance on UCF-101 and HMDB51 datasets<br>• Hindered real-time application |
| SVM [18] | Pornography detection | • Improved pornography detection with deep learning<br>• Superiority over general-purpose action recognition features<br>• High computational cost |
| | Deep learning | |
| CNN+LSTM [10,11] | Violence detection | • Proposed novel deep learning approach to video-based violence detection<br>• Achieved impressive accuracies<br>• Lack of consideration for the degree of violence and specificity to real-world violence |
| CNN [13] | Violence detection | • Proposed a novel method for detecting campus violence<br>• Proposed an improved fusion algorithm<br>• Did not extend to animated or cartoon videos |
| EfficientNet-BiLSTM [14] | Detecting and classifying inappropriate cartoon content | • Proposed deep learning-based framework for inappropriate content detection<br>• Achieved real-time processing and application potential<br>• Only classified full frames |
| MobileNet [15] | Violence detection | • Three-fold movie analysis scheme for violence detection<br>• Achieving high accuracy<br>• Exist overfitting issues |

(Continued)

**Table 1 (continued)**

| Method | Detection goal | Contribution |
|---|---|---|
| CNN [16] | Detection pornographic | • Proposed multilevel fusion method for porn streamer recognition<br>• Integration of bullet screen text recognition<br>• Difficulties with real-time detection |
| CNN [17] | Classification pornographic | • Proposed CNN model for pornographic content classification<br>• Superior performance of the proposed solution compared to manual inspection<br>• Difficulties with robust classification |
| CNN [19] | Violence detection | • Proposed a novel detection strategy based on ResNet-50<br>• Integrated real-time video monitoring<br>• The dataset was not substantial enough for comprehensive classification |
| An ensemble framework using uncertain inference [20] | Pornographic image recognition | • Proposed ensemble framework using uncertain inference for scene recognition<br>• Achieved very good results<br>• Did not account for cartoon and animation videos |

Existing literature on the automatic detection of violent or pornographic content in images presents several areas of improvement. These include the incorporation of degrees of violence or explicitness in the classification process, the adaptation of these techniques specifically for animation and cartoon content, the optimization of model complexity to ensure real-time application feasibility, and a comprehensive violence and pornographic dataset for training deep learning model to detect both scenes simultaneously.

Addressing these identified gaps, my study focuses on developing an automated, novel deep learning-based classification system that can classify both violent and pornographic scenes. The proposed system is designed to accurately detect and classify violent or explicit scenes in cartoon and animation videos. Crucially, it considers the degree of explicitness and violence, thus providing a nuanced and more thorough analysis compared to existing methods. Concurrently, it is designed with a balance between computational complexity and performance, ensuring that real-time detection and classification are feasible. This study also presents a comprehensive dataset for researchers to further process and improve the detection of such scenes.

### *1.3 Contribution*

This study presents a novel CNN-based system for the automatic detection and classification of violence and erotic levels in animated and cartoon images. The primary contributions of this paper include:

1. The creation of a novel dataset comprising 4,044 high-resolution (1920 × 1080) images featuring violent and erotic content derived from animated and cartoon sources. This dataset serves as a valuable resource for developing and testing classification models in this domain.
2. The development of AnimeNet, an innovative CNN-based model designed to effectively classify violence and erotic levels in animated content. This novel model demonstrates superior performance in identifying and categorizing various degrees of unsuitable content within animated images.
3. The introduction of a cutting-edge Channel-Spatial attention module significantly improves feature extraction capabilities for animated and cartoon images. This module enhances the overall effectiveness of the AnimeNet model, ensuring accurate and reliable classification results.

The rest of the paper is structured as follows: Section 2 presents the methods used in this study, Sections 3 and 4 discusse and analyze the extensive experiments to evaluate the performance of the proposed model, and the conclusion is drawn in Section 5.

## 2 Erotic and Violence Level Classification System

Computer vision is a rapidly developing field that has the potential to revolutionize a wide range of industries, including robotics, education, healthcare, and biology. Deep learning-based computer vision solutions, such as object detection and classification, have become increasingly important in various applications.

In this context, I propose an erotic and violence level classification system using deep learning models, which has a well-defined architecture, as illustrated in Fig. 1. The system is composed of four major components, namely Image Collection, Data Augmentation, Model Training, and Erotic and Violence Level Classification Output.

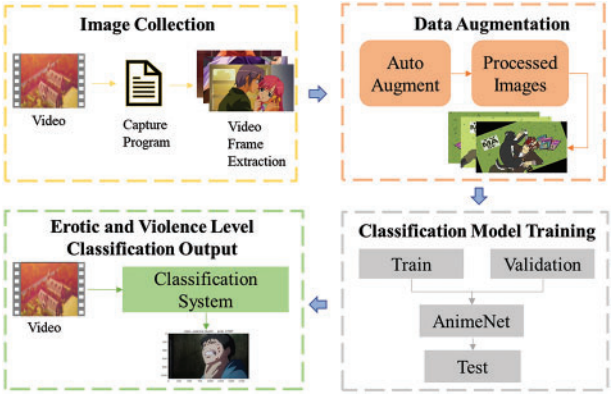


**Figure 1:** The comprehensive structure of the automatic erotic and violence level classification system includes four major parts: (1) image collection from videos; (2) data augmentation on the acquired images; (3) classification model training; (4) erotic and violence level classification output

To create the training dataset, I retrieved images from animation series and movies and examined them manually to identify those containing pornographic and violent content. These images were classified into seven classes based on the severity of the content. Next, I fine-tuned the training dataset using various data augmentation techniques to improve the model's generalization capability.

Several deep learning models were trained on the augmented dataset to evaluate their performance in classifying pornographic and violence levels. Finally, I selected the best-performing model to classify images into the seven predefined classes.

### *2.1 Dataset Preparation*

The identification of erotic or violent content in animation is a crucial challenge that requires a systematic and reliable approach. In this study, I identified 30 animation series and movies that feature a higher proportion of such scenes than other well-known animation content.

To extract relevant frames from the videos, a python code was written automatically processed each frame. Human evaluators were then employed to select the extracted frame which contained violence or pornographic content. The videos ranged in duration from 40 min to 1 h, and the resolution of all frames was up to $1980 \times 1080$.

From this process, I extracted a total of 15000 frames containing pornographic or violent content. I then selected 3890 images with distinct features of pornographic or violent content to construct my dataset. To facilitate classification, I further categorized the dataset into seven classes, including three levels of eroticism (Levels 1, 2, and 3), three levels of violence (Levels 1, 2, and 3), and a normal class. The class distribution of the dataset is shown in Table 2.

**Table 2:** The description of the proposed pornographic and violence level classification dataset

| Class | Code | Image number |
|---|---|---|
| Erotic-low | EL1 | 460 |
| Erotic-middle | EL2 | 510 |
| Erotic-high | EL3 | 520 |
| Normal | NM | 600 |
| Violence-low | VL1 | 600 |
| Violence-middle | VL2 | 600 |
| Violence-high | VL3 | 600 |

Overall, my approach offers a systematic and rigorous methodology for identifying and extracting pornographic or violent content from animation. The resulting dataset can be used for a wide range of applications, including developing machine learning models for content classification and filtering.

### *2.2 Data Augmentation*

Data augmentation is a widely adopted technique for enhancing the quality of a dataset and mitigating overfitting challenges. Datasets with distinct characteristics require specific data augmentation approaches to yield optimal performance. One such method proposed by Cubuk et al. is the AutoAugment technique, which automatically selects the most suitable augmentation method for each image in a dataset [21].

The AutoAugment approach utilizes a policy-based methodology, where each policy comprises various sub-policies that apply augmentation techniques to images based on their individual characteristics.

In prior studies, AutoAugment was evaluated on the CIFAR-10 dataset, achieving a low error rate of 1.48 percent. Furthermore, using the ImageNet dataset, AutoAugment attained a Top-1 accuracy of 83.54 percent. The policy derived from the ImageNet dataset was then deployed to other datasets, including the Stanford Cars and FGVC Aircraft image identification tasks, where it significantly reduced the error rate by 1.16 percent and 1.76 percent, respectively.

In this study, the proposed dataset comprises images with varying degrees of violence and complex characteristics that require appropriate augmentation techniques to enhance the dataset's quality. To achieve this, I have utilized the AutoAugment approach to automatically process the images from the dataset and improve their quality before training the proposed model. In Fig. 2, three sample images are presented, showcasing the application of the AutoAugment method on the proposed dataset. The algorithm intelligently selects and applies horizontal flipping, brightness enhancement, and rotation transformations based on the data-driven augmentation policy it has discovered.

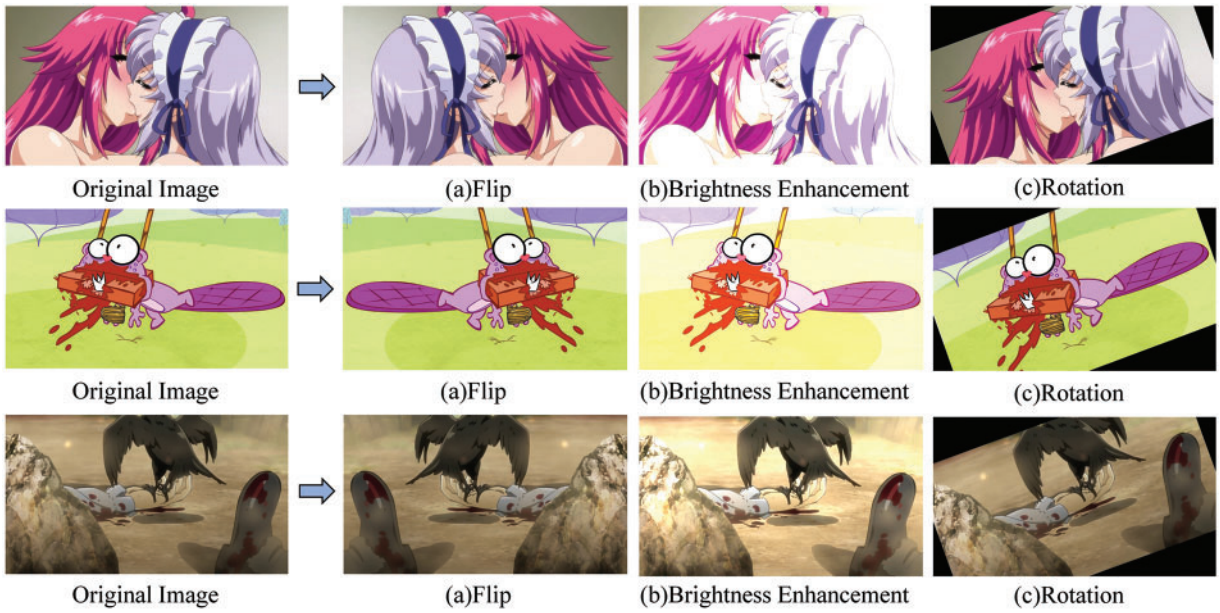


**Figure 2:** Autoaugment method applied to the dataset images, showcasing the policy-based selected transformations. Original image, (a) flip, (b) brightness enhancement, (c) rotation

### *2.3 AnimeNet Model*

This study proposes the use of AnimeNet, a deep learning model, to extract features and classify the level of erotic and violent content in detected frames, as illustrated in Fig. 3. The AnimeNet model consists of three main parts: stem, body, and head.

The body of the AnimeNet model comprises four stages, with each stage containing several $d_i$ blocks. The detailed structure of the stem, stage, and head can be seen in Fig. 4. Furthermore, the structure of individual $d_i$ blocks is depicted in Fig. 5.

The stem component of the AnimeNet model serves as the input and performs initial feature extraction, followed by the body that further extracts relevant features using the $d_i$ blocks. Finally, the head component performs the classification task to determine the level of pornographic or violent content in the detected frame.

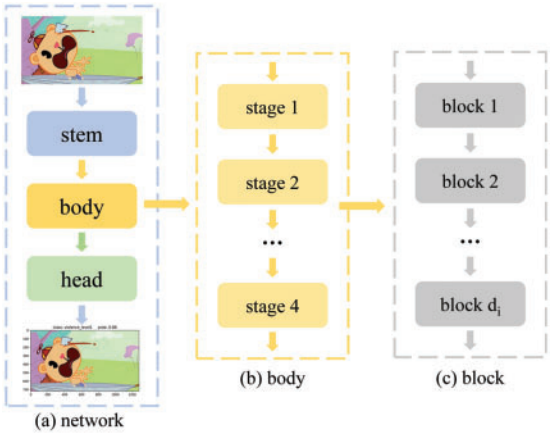


**Figure 3:** The structure of the proposed AnimeNet model

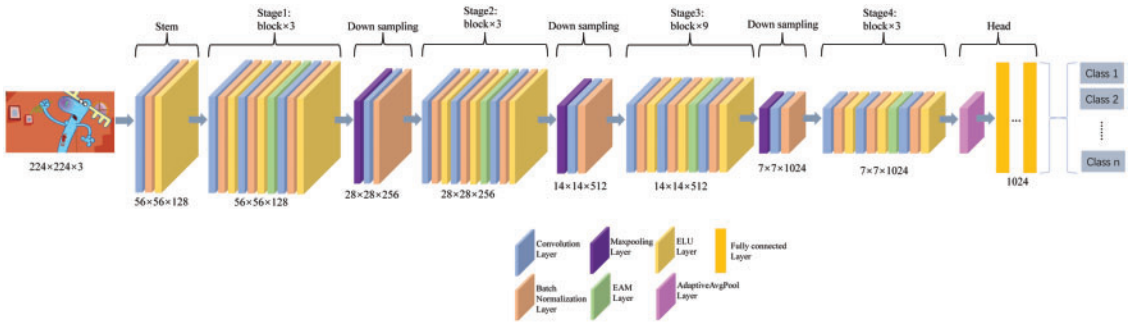


**Figure 4:** The description of the proposed AnimeNet model includes stem, stage, down sampling, and head parts

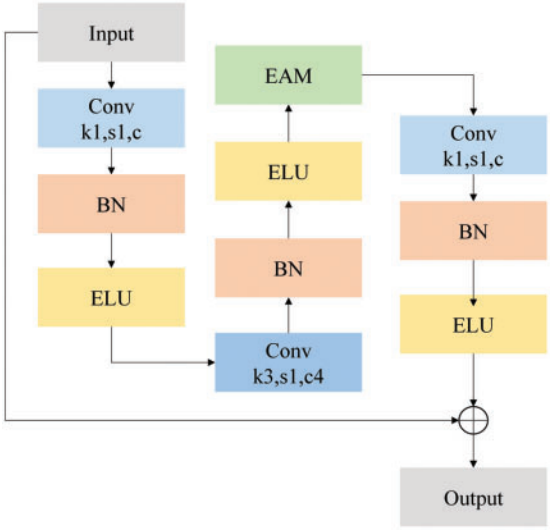


**Figure 5:** The structure of the individual block, which is used in every stage of the body parts of the proposed AnimeNet

The proposed AnimeNet model's architecture is designed to enhance its performance and robustness in handling sensitive content classification tasks. Its multi-stage design allows for deep feature extraction and enables the model to learn and represent complex image features effectively. Additionally, the use of $d_i$ blocks enhances the model's ability to learn and identify low-level features while maintaining computational efficiency.

I developed a novel CNN-based deep learning model called AnimeNet, as illustrated in Fig. 4, to achieve superior performance in handling complex violence and erotic content classification tasks. The proposed model comprises several key components, including stem, stage, downsampling, and head parts.

The stage component of the proposed model contains a series of blocks that are responsible for learning and representing complex image features. Stages 1, 2, and 4 consist of three blocks each, while stage 3 contains nine blocks. The detailed structure of the blocks can be seen in Fig. 5.

The stem component of the proposed model comprises a $4 \times 4$ convolutional layer with a stride of 4, followed by batch normalization and the Exponential Linear Unit (ELU) activation function. The ELU activation function is known to be effective in reducing the vanishing gradient problem and enhancing the performance of deep learning models.

The downsampling component of the proposed model comprises a $2 \times 2$ max pooling layer with a stride of 2, a $3 \times 3$ convolutional layer with a stride of 1 and padding, and a batch normalization layer. These layers are responsible for downsampling the input image and extracting relevant features.

The head component of the proposed model includes an average pooling layer followed by fully connected layers, which are responsible for classifying the frame into different categories based on the level of violence and pornographic content.

Fig. 5 provides an illustration of the individual blocks in every stage of the body parts of the AnimeNet model. Each block follows a specific design and sequence of layers optimized for handling complex image classification tasks effectively.

The first layer in each block is a $1 \times 1$ convolutional layer followed by batch normalization and the Exponential Linear Unit (ELU) activation function. This initial layer is responsible for reducing the dimensionality of the input image while preserving its key features.

Next, a $3 \times 3$ convolutional layer is used to increase the channel size of the input image by four times. This layer is also followed by batch normalization and the ELU activation function, which enhance the model's ability to learn and represent complex features effectively.

An enhanced attention module is then added to extract important features from the input image and improve the model's performance. This module helps the model to focus on the most relevant regions in the input image, enhancing its ability to detect and classify violence and pornographic content accurately.

Finally, a $1 \times 1$ convolutional layer is added to decrease the channel size to its previous dimensionality. This layer is also followed by batch normalization and the ELU activation function, which help to maintain the consistency of the model's features throughout the network. The relevant equation for the individual block can be defined as follows,

$$Output = ELU(BN(Conv_1(EAM(ELU(BN(Conv_3(ELU(BN(Conv_1(x)))))))))) \quad (1)$$

Overall, the design of the individual blocks in the AnimeNet model is optimized for achieving superior performance in handling complex violence and pornographic content classification tasks.

The incorporation of specific layers and modules, such as the enhanced attention module, enhances the model's ability to learn and represent complex image features while maintaining computational efficiency.

Fig. 6 depicts the enhanced attention module used in the proposed AnimeNet model, which comprises both spatial and channel attention networks. The design of this module is inspired by study [22], which has been shown to be effective in improving the performance of deep learning models for image classification tasks.

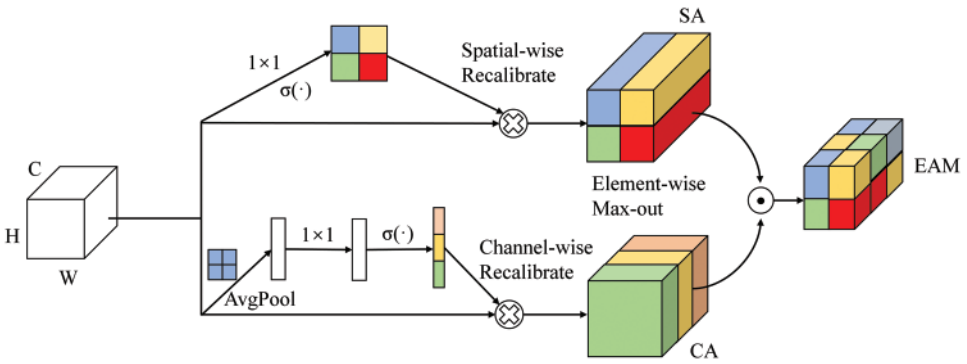


**Figure 6:** The overall structure of the proposed channel-spatial attention module

The channel attention blocks in the enhanced attention module recalibrate the channels of the input image by incorporating global spatial information. To achieve this, adaptive average pooling layers are used, which provide a receptive field of the whole spatial extent at each stage of the model, aiding content classification. The processed features are then passed through a $1 \times 1$ convolution layer, followed by a sigmoid activation function. The sigmoid function maps the input values to a range of 0 to 1, making it useful for gating or scaling the features. The relevant equation for channel attention block can be defined as follows:

$$CA = \sigma(Conv_1(Avgpool(x))) \times x. \tag{2}$$

The spatial attention blocks in the enhanced attention module contain a $1 \times 1$ convolution layer followed by a sigmoid activation function. The element-wise max operation is used to perform an element-wise competitiveness between the two attention blocks, providing selective spatial and channel excitation. The relevant equation for spatial attention block can be defined as follows,

$$SA = \sigma(Conv_1(x)) \times x. \tag{3}$$

The output from EAM ensures that the final classification is based on the most relevant features in the input image, improving the overall accuracy of the model. The relevant equation can be defined as follows:

$$EAM = CA \odot SA. \tag{4}$$

The enhanced attention module in the AnimeNet model is a key component that enables the model to selectively focus on the most relevant features in the input image, enhancing its ability to detect and classify violence and pornographic content accurately. By incorporating both spatial and channel attention networks, the enhanced attention module provides a powerful mechanism for learning and representing complex image features.

Fig. 7 presents the overall flow chart of the proposed animation content classification system using the AnimeNet model. The proposed system starts with an input in the form of a video. This video is processed through a specialized capture program designed to extract individual frames, resulting in an amassed collection of images. These images form the basis for the subsequent steps and are pivotal for the overall classification task.

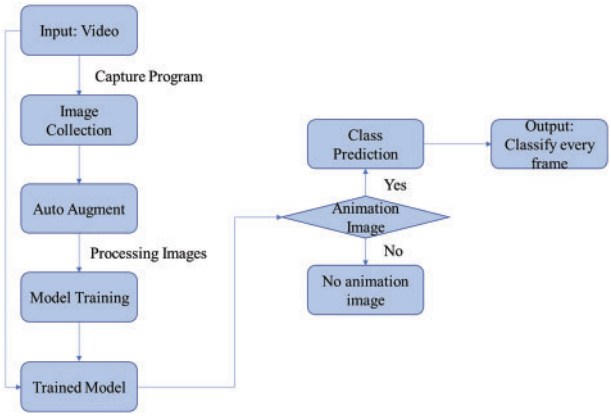


**Figure 7:** Flowchart depicting the process of animation video content classification using the AnimeNet model

The image collection then undergoes an 'Auto Augment' process. This particular phase involves the augmentation of the original images through various transformations such as rotation, scaling, shifting, and flipping. The objective here is to increase the diversity and volume of training data, thereby enhancing the robustness of the model and its ability to generalize.

Once the images have been suitably augmented, they are subject to an additional round of processing. During this 'Processing Images' phase, images are manipulated to meet specific model requirements. This could entail operations such as resizing to maintain uniformity, normalization for optimal model performance, or conversion into the necessary formats or color spaces.

The processed images are subsequently deployed for model training. This stage entails the employment of the conditioned images to train the deep learning model, enabling it to recognize and differentiate between diverse classes present in the animation frames.

Upon successful completion of the training phase, a 'Trained Model' emerges. This model, enriched by the insights gleaned from the training data, is now equipped to predict the class of new, unseen animation images.

The proposed system also encompasses a decision-making procedure. If an image under examination is an animation image, the model proceeds to predict its class. In contrast, if the image does not qualify as an animation image, it is categorized as 'No animation image.' Finally, the output of the system is in the form of classified frames derived from the original video input.

## 3 Experiment and Result Analysis

### *3.1 Experimental Parameters*

In this research, I chose Ubuntu as the experimental platform. The Intel(R) Core (TM) i7-8700 @3.20 GHz processor and the NVIDIA GeForce GTX TITAN X GPU were chosen as the platform for this research, while Python 3.7 and Pytorch 1.11 were chosen as the programming language and

deep learning framework, respectively. The detailed hyperparameter and experimental settings used in this study can be seen in Table 3.

**Table 3:** Key parameters and specifications employed for the experimental setup in the current study

| Experimental parameters | Value |
|---|---|
| Number of initial convolution kernels | 56 |
| Learning rate | 0.01 |
| Image size | 224 |
| Classification number | 7 |
| experimental platform | Ubuntu |
| CPU | Intel(R) Core (TM) i7-8700 |
| GPU | NVIDIA GeForce GTX TITAN X |
| Pytorch version | 1.11 |
| Python version | 3.7 |

An extensive experiment was done on the proposed AnimeNet model to evaluate the performance of the model. I used different classification models commonly used in the literature to train the proposed dataset and compared the performance using standard evaluation metrics found in the literature.

### *3.2 Pornographic and Violence Level Classification Evaluation Metrics Analysis*

Fig. 8 illustrates the performance comparison of the proposed model when trained using different optimizers, namely Stochastic Gradient Descent (SGD) [23] and Adam [24], as well as various learning rates (0.01, 0.005, and 0.001). The objective of this comparison is to determine the optimal combination of hyperparameters for achieving the highest possible accuracy.

When employing the SGD optimizer, the model's performance was observed to be lower than that of the Adam optimizer. The lowest accuracy recorded with the SGD optimizer, 68.6%, was obtained at a learning rate of 0.001. Conversely, the highest accuracy achieved with the SGD optimizer was 71.7% at a learning rate of 0.01.

In contrast, using the Adam optimizer, the proposed model demonstrated marginally superior performance across all learning rates. The least accurate configuration, yielding an accuracy of 70.8%, employed the Adam optimizer at a learning rate of 0.005. Furthermore, the highest accuracy of 72.7% was attained with the Adam optimizer at a learning rate of 0.01.

Based on these findings, the Adam optimizer with a learning rate of 0.01 was identified as the optimal combination for training the proposed model. Consequently, this configuration was selected for further evaluation and assessment of the model's performance.

Table 4 compares the performance of the proposed method, AnimeNet, with other state-of-the-art models, including DenseNet161 [24], ResNet50 [25], MobileNet_v3 [26], EfficientNet [27], ViT [28], ResNext50 [29], ConvNext_b [30], RegNet [31], and Wide-ResNet50 [32]. The performance metrics for the methods were assessed across multiple conditions, namely EL1, EL2, EL3, NM, VL1, VL2, and VL3.

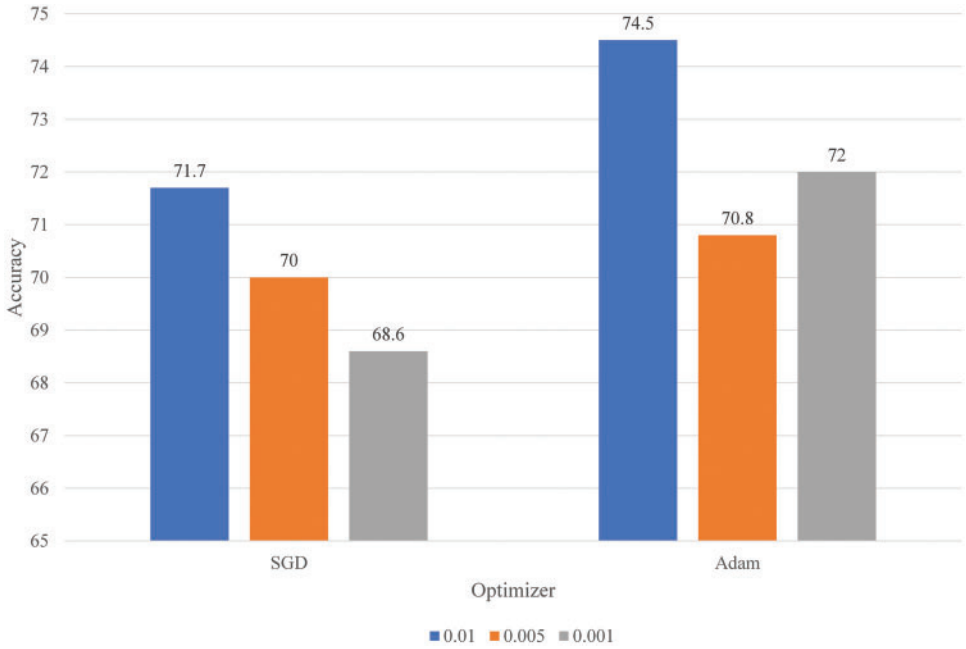


**Figure 8:** The comparison of the model accuracy using different hyperparameter settings

**Table 4:** The accuracy of the erotic and violence classification for different deep learning models using a 7:3 ratio

| Method | EL1 | EL2 | EL3 | NM | VL1 | VL2 | VL3 | Acc |
|---|---|---|---|---|---|---|---|---|
| DenseNet161 [24] | 23.1 | 42.6 | 34.8 | 42.5 | 40.3 | 59.8 | 66.3 | 45.3 |
| ResNet50 [25] | 70.7 | 64.1 | 68.2 | 74.5 | 66.1 | 61.7 | 73.1 | 68.6 |
| MobileNet_v3 [26] | 55.8 | 51.9 | 52.5 | 62.0 | 57.1 | 60.4 | 72.5 | 59.3 |
| EfficientNet [27] | 67.8 | 61.7 | 62.9 | 77.2 | 69.0 | 64.0 | 77.4 | 69.1 |
| ViT [28] | 76.2 | **84.5** | 64.3 | 77.9 | 68.7 | **67.8** | **86.1** | 74.0 |
| ResNext50 [29] | **78.6** | 64.6 | 71.2 | 77.1 | 67.6 | 62.6 | 76.4 | 71.0 |
| ConvNext_b [30] | 65.7 | 63.3 | 64.6 | 73.7 | 57.3 | 59.2 | 71.1 | 65.0 |
| RegNet [31] | 72.9 | 69.2 | 74.6 | 79.0 | 62.6 | 61.1 | 75.5 | 70.8 |
| Wide-ResNet50 [32] | 71.6 | 61.6 | 67.1 | 73.3 | 60.8 | 58.7 | 75.3 | 66.8 |
| AnimeNet | 74.2 | 73.0 | **75.8** | **84.6** | **74.6** | 63.9 | 75.4 | **74.5** |

In terms of performance across EL1, EL2, and EL3 categories, the AnimeNet model demonstrated superior performance with scores of 74.2, 73.0, and 75.8, respectively, surpassing all the competing models. Notably, AnimeNet outperformed the ViT model, which scored 76.2, 84.5, and 64.3 for the same categories.

Under the NM condition, AnimeNet achieved the highest score of 84.6, followed closely by EfficientNet, with a score of 77.2. In the VL1 category, AnimeNet outperformed all other models, securing a score of 74.6.

For VL2 and VL3 categories, AnimeNet scored 63.9 and 75.4, respectively. Despite its slightly lower performance in the VL2 category, AnimeNet still maintained a competitive stance among other models. In the VL3 category, it performed closely to DenseNet161, which achieved a score of 66.3.

Finally, taking an overall perspective through the Accuracy metric, AnimeNet's performance was superior to all others, with an accuracy of 74.5, affirming the robustness of the proposed method. This clearly demonstrates that our AnimeNet model stands as a significant improvement over the existing methods in multiple conditions, emphasizing its versatility and superior performance.

To further assess the performance of the proposed model on my custom dataset, I conducted additional experiments by adjusting the training and testing data ratios to 50% each. Table 5 compares the accuracy of erotic (EL1, EL2, EL3, NM) and violence (VL1, VL2, VL3) classification for various deep learning models, including DenseNet161 [24], ResNet50 [25], MobileNet_v3 [26], EfficientNet [27], ViT [28], ResNext50 [29], ConvNext_b [30], RegNet [31], Wide-ResNet50 [32], and the proposed method, AnimeNet.

**Table 5:** The accuracy of the erotic and violence classification for different deep learning models using a 5:5 ratio

| Method | EL1 | EL2 | EL3 | NM | VL1 | VL2 | VL3 | Acc |
|---|---|---|---|---|---|---|---|---|
| DenseNet161 [24] | 66.5 | 65.0 | 66.8 | 77.7 | 60.9 | **65.1** | 75.2 | 68.1 |
| ResNet50 [25] | 64.7 | 60.6 | 62.3 | 72.8 | 62.3 | 57.0 | 74.3 | 65.4 |
| MobileNet_v3 [26] | 73.0 | 64.5 | 64.7 | 74.5 | 60.4 | 63.1 | 75.2 | 67.7 |
| EfficientNet [27] | 53.7 | 59.0 | 52.5 | 73.9 | 67.6 | 60.9 | 75.7 | 64.2 |
| ViT [28] | 67.0 | 59.3 | 71.5 | 69.2 | 65.1 | 54.0 | 70.8 | 64.9 |
| ResNext50 [29] | 64.2 | 61.6 | 68.3 | 77.3 | 63.4 | 59.3 | 76.6 | 67.6 |
| ConvNext_b [30] | 23.7 | 21.7 | 33.5 | 35.9 | 33.1 | 36.5 | 59.1 | 35.5 |
| RegNet [31] | 54.0 | 56.9 | 59.3 | 71.7 | 67.1 | 63.4 | 75.3 | 64.9 |
| Wide-ResNet50 [32] | **71.7** | 65.1 | 66.1 | 71.9 | 62.9 | 57.5 | **77.2** | 67.5 |
| AnimeNet | 70.6 | **75.9** | **71.9** | **78.8** | **73.9** | 63.4 | 73.8 | **72.7** |

Among the models, AnimeNet demonstrated a superior performance across all categories. In the erotic classification categories EL1, EL2, and EL3, AnimeNet scored 70.6, 75.9, and 71.9, respectively. The next closest performer in these categories was MobileNet_v3, which achieved scores of 73.0, 64.5, and 64.7, respectively. In the NM category, AnimeNet again outperformed all other models, securing a score of 78.8.

In the violence classification categories, AnimeNet consistently delivered top performance. In the VL1 category, AnimeNet achieved a score of 73.9, surpassing EfficientNet, which scored 67.6. For VL2 and VL3, AnimeNet scored 63.4 and 73.8, respectively, outperforming Wide-ResNet50, which scored 57.5 and 77.2.

When looking at the overall accuracy (Acc), AnimeNet achieved the highest score of 72.7. The next closest was DenseNet161, with an accuracy of 68.1. The robust performance of AnimeNet across all categories affirms its superior effectiveness for erotic and violence classification, setting a new standard for similar deep-learning models.

In order to further evaluate the performance of the proposed model on my custom dataset, I conducted additional experiments by modifying the training and testing data ratios to 30% and 70%, respectively. Table 6 presents the performance comparison of various deep learning models, including DenseNet161 [24], ResNet50 [25], MobileNet_v3 [26], EfficientNet [27], ViT [28], ResNext50 [29], ConvNext_b [30], RegNet [31], Wide-ResNet50 [32], and the proposed model, AnimeNet.

**Table 6:** The accuracy of the erotic and violence classification for different deep learning models using a 3:7 ratio

| Method | EL1 | EL2 | EL3 | NM | VL1 | VL2 | VL3 | Acc |
|---|---|---|---|---|---|---|---|---|
| DenseNet161 [24] | **67.1** | 59.0 | **61.9** | 69.7 | 63.5 | 61.7 | 69.6 | 64.8 |
| ResNet50 [25] | 53.6 | 55.1 | 55.9 | 62.4 | 60.7 | 61.5 | 69.2 | 60.6 |
| MobileNet_v3 [26] | 44.2 | 43.8 | 47.1 | 57.3 | 53.3 | 57.5 | 70.4 | 54.6 |
| EfficientNet [27] | 48.9 | 45.7 | 44.4 | 59.7 | 57.8 | **64.3** | **72.1** | 57.3 |
| ViT [28] | 54.8 | 58.9 | 60.7 | 65.5 | 52.4 | 49.6 | 66.9 | 58.7 |
| ResNext50 [29] | 59.1 | 59.9 | 53.5 | 67.1 | 58.6 | 58.0 | 71.0 | 61.6 |
| ConvNext_b [30] | 47.2 | 50.4 | 48.5 | 53.7 | 49.1 | 51.5 | 62.7 | 52.8 |
| RegNet [31] | 58.0 | 54.0 | 54.3 | 62.4 | 58.7 | 55.3 | 66.3 | 59.1 |
| Wide-ResNet50 [32] | 57.5 | 52.6 | 50.7 | 64.4 | 61.4 | 61.1 | 67.9 | 59.9 |
| AnimeNet | 57.4 | **62.6** | 57.1 | **72.1** | **69.0** | 61.8 | 71.6 | **65.2** |

In the erotic classification categories (EL1, EL2, EL3, NM), AnimeNet generally outperformed all other models. Specifically, in the EL1 category, AnimeNet achieved a performance score of 57.4, closely followed by DenseNet161, with a score of 67.1. In the EL2 category, AnimeNet stood out with the highest score of 62.6. For the EL3 and NM categories, AnimeNet achieved scores of 57.1 and 72.1, respectively, reflecting superior performance compared to the other models.

In the violence classification categories (VL1, VL2, VL3), AnimeNet's performance was remarkably robust. In the VL1 category, AnimeNet outperformed all other models with a score of 69.0, while in the VL2 and VL3 categories, AnimeNet achieved respectable scores of 61.8 and 71.6, respectively. Notably, in the VL3 category, AnimeNet outperformed EfficientNet, which scored 72.1.

When considering the overall accuracy (Acc), AnimeNet outshone all other models, achieving the highest score of 65.2. The next closest model, DenseNet161, secured an accuracy of 64.8. Despite the skewed data ratio, AnimeNet's performance remained consistent, thereby demonstrating its ability to maintain high performance in various scenarios. This underscores the efficacy of the proposed method, AnimeNet, in classifying erotic and violent content and highlights its potential applicability in real-world scenarios with uneven data distributions.

This study also calculates the standard evaluation metrics such as precision, recall, specificity, F1-score, and accuracy for every class to evaluate the performance of the proposed model. I used a 7:3 training and testing-based model to calculate the evaluation metrics. Table 7 presents the comprehensive evaluation of the proposed erotic and violence level classification model across several standard multiclass evaluation metrics: precision, recall, specificity, F1-score, and accuracy. Each metric contributes to a holistic understanding of the model's performance across different classes (EL1, EL2, EL3, NM, VL1, VL2, and VL3).

**Table 7:** The standard multiclass evaluation metrics of the proposed erotic and violence level classification model. The precision, recall, specificity, F1-score, and accuracy are used to evaluate the model

| Class | Precision | Recall | Specificity | F1-score | Accuracy |
|---|---|---|---|---|---|
| EL1 | 74.2 | 68.8 | 96.8 | 71.4 | 93.5 |
| EL2 | 73.0 | 70.6 | 96.1 | 71.8 | 92.7 |
| EL3 | 74.5 | 76.9 | 95.9 | 75.7 | 93.4 |
| NM | 85.6 | 86.1 | 97.4 | 85.8 | 95.6 |
| VL1 | 74.6 | 75.0 | 95.3 | 74.8 | 92.2 |
| VL2 | 63.9 | 60.0 | 93.8 | 61.9 | 88.6 |
| VL3 | 75.4 | 83.3 | 95.0 | 79.2 | 93.2 |

For class EL1, the model exhibited a precision of 74.2%, implying that when the model predicted an instance to be of class EL1, it was correct 74.2% of the time. The recall of 68.8% denotes that the model correctly identified 68.8% of all actual EL1 instances. The model's specificity was high at 96.8%, demonstrating its ability to correctly identify instances not belonging to class EL1. The F1-score, a harmonic mean of precision and recall, was 71.4%, reflecting a balanced predictive performance. The overall accuracy for EL1 stood at 93.5%.

The performance across other classes followed a similar pattern. For instance, the class EL2 exhibited a precision of 73.0%, a recall of 70.6%, a specificity of 96.1%, an F1-score of 71.8%, and an accuracy of 92.7%. Class EL3 demonstrated strong performance with a precision of 74.5%, recall of 76.9%, specificity of 95.9%, F1-score of 75.7%, and accuracy of 93.4%.

For the normal (NM) class, the model had the highest performance among all classes, with a precision of 85.6%, recall of 86.1%, specificity of 97.4%, an F1-score of 85.8%, and an accuracy of 95.6%.

For the violence categories, the model's performance was consistently strong. Class VL1 achieved a precision of 74.6%, recall of 75.0%, specificity of 95.3%, F1-score of 74.8%, and accuracy of 92.2%. In contrast, the VL2 class showed a slight dip in performance with precision at 63.9%, recall at 60.0%, specificity at 93.8%, F1-score at 61.9%, and accuracy at 88.6%. Class VL3 rebounded with a precision of 75.4%, recall of 83.3%, specificity of 95.0%, an F1-score of 79.2%, and an accuracy of 93.2%.

These results highlight the strong performance of the proposed model in classifying erotic and violence levels, confirming its effectiveness and reliability.

Fig. 9 presents the confusion matrix for the proposed classification model, which provides an in-depth examination of the model's class-wise performance. This graphical representation offers a more nuanced understanding of my model's capability and further corroborates the evaluation metrics previously discussed.

Significantly, the confusion matrix reveals that the model achieves a comparatively higher accuracy rate for the normal (non-erotic, non-violent) class, demonstrated by fewer misclassifications. This suggests that the model has been particularly successful in recognizing and classifying images in the normal class.

However, the confusion matrix also reveals areas where the model's performance is less accurate. Specifically, some misclassifications occurred among the three erotic levels. This phenomenon may be

attributed to the shared or similar features across these classes, which make differentiation challenging for the model. A similar trend was observed in the violence levels, with the model experiencing prediction errors between the different classes of violence.

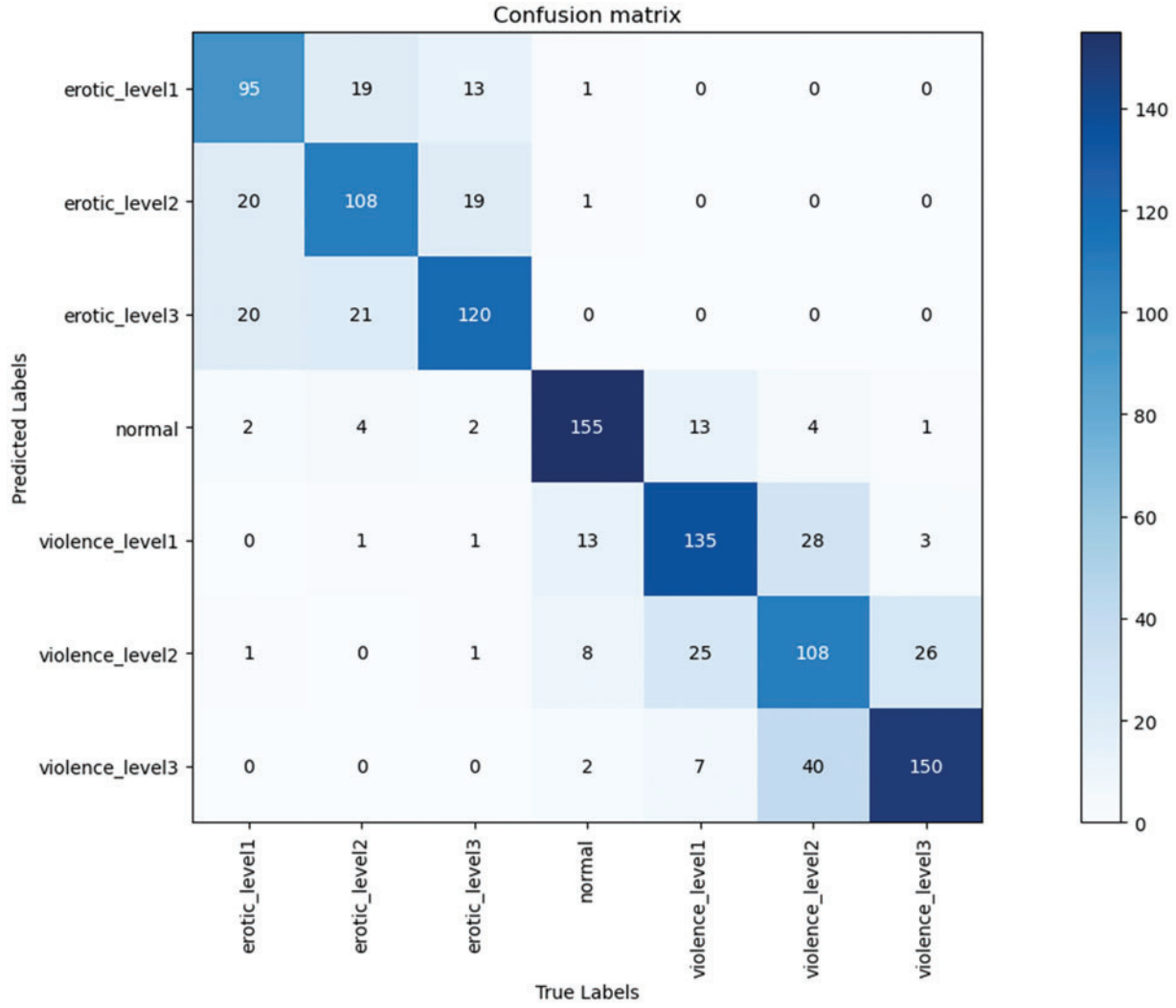


**Figure 9:** The confusion matrix of the implemented AnimeNet to classify the erotic and violence level

For example, the model incorrectly classified 40 images from violence level 2 as level 3. This type of error likely stems from the presence of overlapping or analogous patterns across these two classes. Nonetheless, such misclassifications are comparatively few, and the model demonstrates predominant success in accurately predicting the correct classes, with only a limited number of significant misclassifications.

In summary, the overall performance of the proposed model, as demonstrated through the confusion matrix and the associated evaluation metrics, underscores its competency in undertaking the complex task of cartoon scene classification. Despite certain areas of misclassification, the model shows significant potential for practical applications within this domain. Its ability to identify and

differentiate between various levels of erotic and violent content makes it a compelling choice for systems requiring precise content moderation and classification in cartoon-based media.

Various real-world scenarios can impact the representation of cartoon images, such as zooming into scenes or cropping significant portions. Thus, assessing the robustness of a scene classification model against diverse scenarios is crucial before deploying it in real-time cartoon scene monitoring applications. In this study, I evaluate the performance of the proposed model under three common scenarios encountered in cartoon scenes, specifically cropping, rotation, and block noise, which may arise in cartoon videos.

Fig. 10 illustrates the prediction scores achieved by the proposed model when applied to randomly cropped cartoon images. The effectiveness of the model is demonstrated through a series of experiments where an original image, as shown in Fig. 10a, is segmented into numerous randomly cropped subsections. These cropped images are then individually introduced into the classifier, which proceeds to predict their respective scene categories.

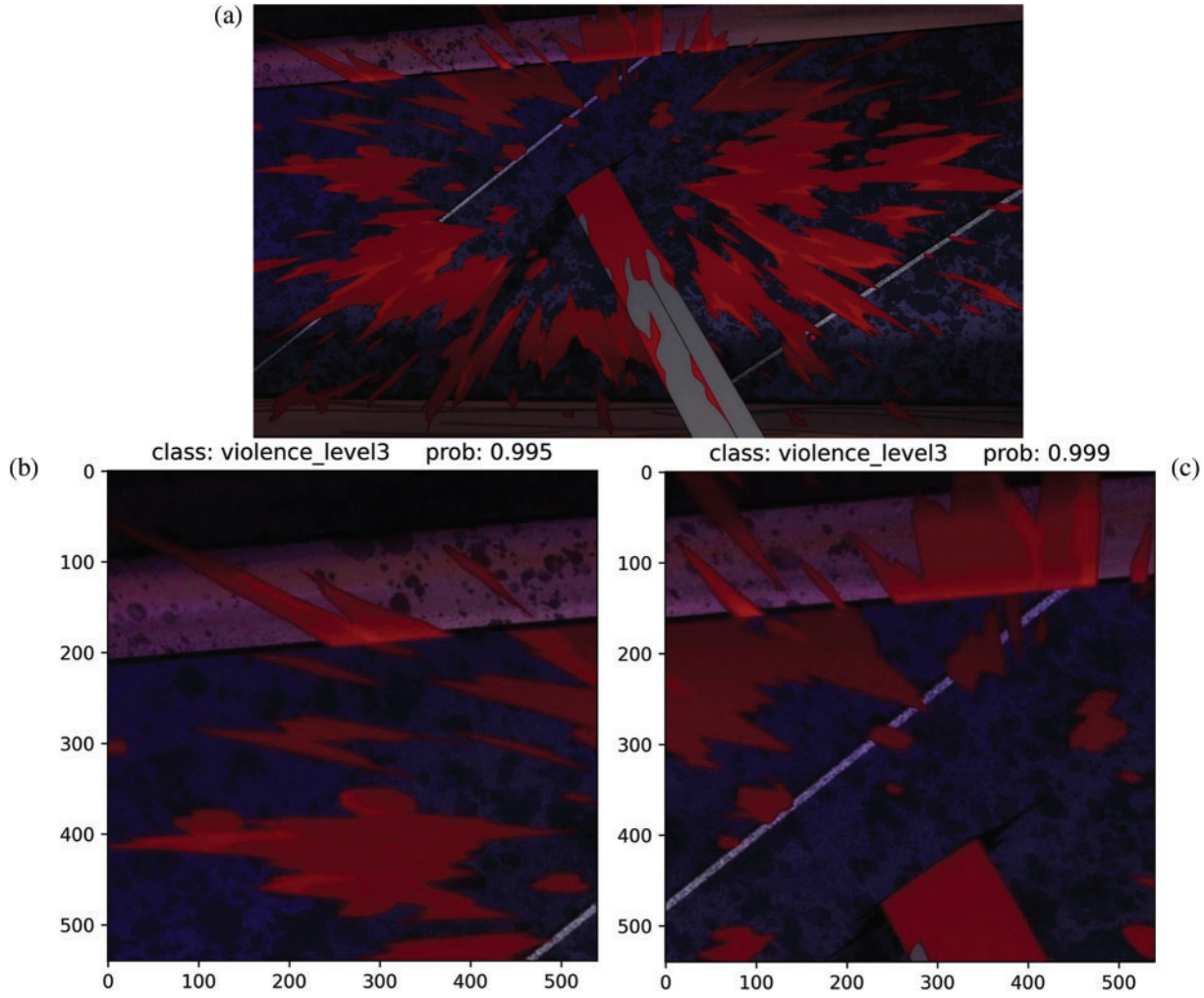


**Figure 10:** (Continued)

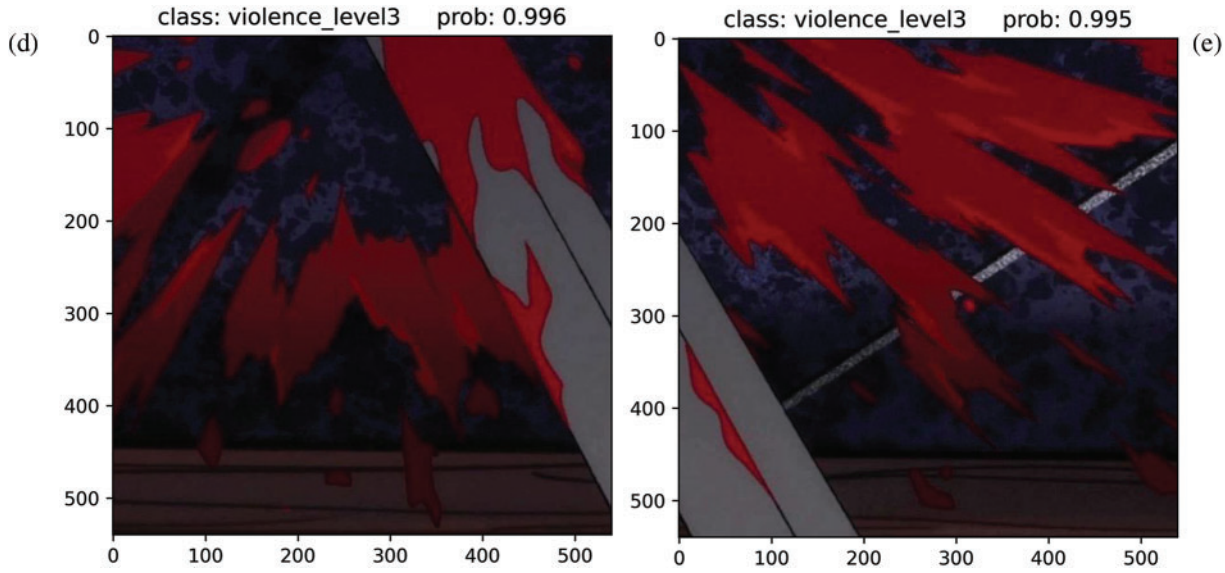


**Figure 10:** The prediction results of the proposed model on the cropped images. (a) is the original image. (b), (c), (d), and (e) are the predicted results using the proposed model on randomly cropped violent images

The results derived from these experiments provide strong evidence of the model's accurate and consistent performance. Despite the challenges posed by the lack of context due to cropping, the model effectively assigns the correct scene class to each cropped image with remarkably high prediction scores. This demonstrates the model's ability to identify salient features within each cropped section that are indicative of the overall scene category, showcasing its robustness and proficiency in managing partially obscured or fragmented images.

Furthermore, the analysis of the prediction scores reveals that the average score across all cropped images exceeds 99%. This high rate underscores the reliability and robustness of the proposed model in processing and classifying cropped images. It suggests that even when only portions of an image are available, my model is capable of making highly accurate predictions, thereby proving its resilience to the challenges presented by partial image data.

Further evaluations, such as rotation and block noise scenarios, are essential to comprehensively understand the model's ability to maintain accurate classification performance despite variations in input data. A detailed analysis of these scenarios will provide additional insights into the model's adaptability and potential for use in real-world applications.

Fig. 11 provides a visualization of the prediction scores achieved by the proposed model when applied to rotated cartoon images. This assessment unveils an intriguing relationship between the prediction accuracy of the model and the varying degree of rotation applied to the images.

The baseline for this experiment is established with an original, unrotated image depicted in Fig. 11a, which yields a prediction accuracy of 91.8%. The proposed model manages to uphold high prediction scores when images are rotated, primarily as the rotation preserves the distribution of inherent features within the image that are crucial for classification.

Nonetheless, it is noteworthy that the model's prediction accuracy encounters a significant dip when the images are subjected to rotations of 180 degrees and 90 degrees to the right. In these instances,

the model's prediction accuracies descend to 55.8% and 57.4%, respectively, reflecting a substantial deterioration in the model's performance.


**Figure 11:** The prediction results of the proposed model on the rotated images. (a), (b), (c), and (d) are the predicted results using the proposed model on the rotated images

These results underline the profound influence that image rotation can have on the performance of a classification model. Despite the robustness of the model in handling a variety of alterations, rotations at certain angles significantly challenge the model's ability to accurately predict the scene class, illustrating the model's limitations under such conditions.

The proposed model's performance was further evaluated by examining its ability to classify noisy images. Fig. 12 displays the prediction outcomes for images that were distorted with various block noises, demonstrating the model's aptitude for accurate classification amidst these added challenges.

Distinct block noises were integrated into the images in Figs. 12a–12d. Despite the presence of such noise, these images were effectively processed by the proposed model, which correctly identified all of them as belonging to the violence level 1 class. Among these classifications, the lowest prediction accuracy was observed for Fig. 12a, with an accuracy score of 92.7%, still a commendable level given the noisy input.

Though the proposed model exhibits a noteworthy capacity to classify cartoon scenes even in the presence of noise, the levels of accuracy achieved in these instances were observed to be slightly lower than those attained when handling cropped images. Despite this modest reduction, the analysis of the results indicates that the proposed model maintains its robustness in the face of various forms

of noise and constrained environments, characteristics that would be frequently encountered in real-world scenarios.

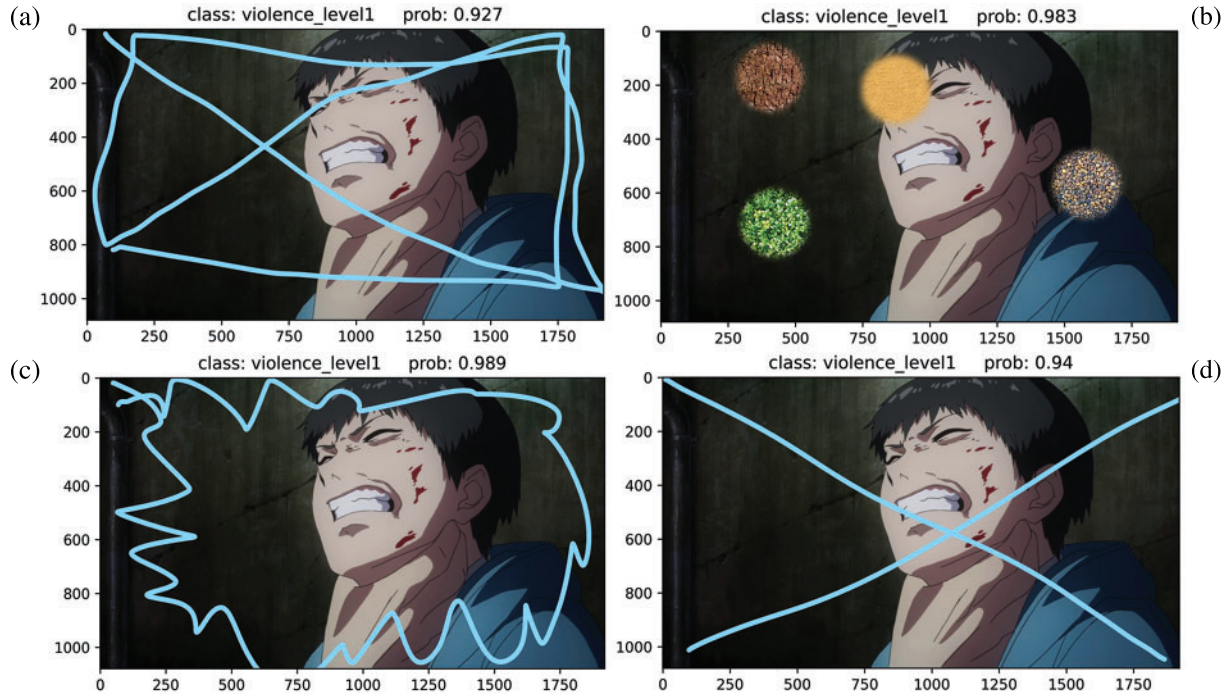


**Figure 12:** The prediction results of the proposed model on the noisy images

These findings, along with the prior assessments involving cropped and rotated images, contribute to a comprehensive understanding of the model's performance under an array of diverse and challenging conditions. This collective evaluation provides a holistic view of the proposed model's resilience and adaptability, highlighting its suitability for use in practical applications that may encompass noisy, rotated, or partially obscured imagery.

## 4 Discussion

This section focuses on evaluating various methods for detecting pornography and violence in images. Table 8 provides summarizes the detection goals, class numbers, and accuracies achieved by different authors in their respective studies.

Peres et al. [18] addressed the goal of pornography detection and achieved an accuracy of 97.9%. Ditsanthia et al. [10] focused on violence detection, specifically targeting two classes, and obtained an accuracy of 75.73%. In contrast, Sumon et al. [11] and Ye et al. [13] also worked on violence detection with two classes but achieved higher accuracies of 97.06% and 97%, respectively.

Yousaf et al. [14] explored the detection and classification of inappropriate cartoon content, which involved three classes. They reported an accuracy of 95.66%. Khan et al. [15] focused on violence detection similar to previous works and obtained an accuracy of 97.0%.

Wang et al. [16] and Mohamed [17] concentrated on pornographic detection. Wang et al. achieved an accuracy of 69.24% using a two-class approach, while Mohamed achieved a higher accuracy of 94.1% with a three-class classification.

**Table 8:** Comparative analysis of AnimeNet with different existing methods for violence and pornography detection

| Method | Detection goal | Class number | Accuracy |
|---|---|---|---|
| Peres et al. [18] | Pornography detection | 2 | 97.9% |
| Ditsanthia et al. [10] | Violence detection | 2 | 75.73% |
| Sumon et al. [11] | Violence detection | 2 | 97.06% |
| Ye et al. [13] | Violence detection | 2 | 97% |
| Yousaf et al. [14] | Detecting and classifying inappropriate cartoon content | 3 | 95.66% |
| Khan et al. [15] | Violence detection | 2 | 97.0% |
| Wang et al. [16] | Detection pornographic | 2 | 69.24% |
| Mohamed [17] | Classification pornographic | 3 | 94.1% |
| Yuan et al. [19] | Violence detection | 2 | 95% |
| Shen et al. [20] | Pornographic image recognition | 2 | 94.70% |
| This paper | Violence and pornographic detection | 7 | 74.5% |

Yuan et al. [19] contributed to violence detection with an accuracy of 95%. Shen et al. [20] specifically targeted pornographic image recognition and achieved an accuracy of 94.70% using a two-class approach.

In this study, I developed a method for violence and pornographic detection, considering seven different classes. My approach achieved an accuracy of 74.5%. While my method for violence and pornographic detection achieved a lower accuracy of 74.5% compared to some of the other methods discussed, it is important to provide further context and insights into my approach.

Firstly, it is crucial to note that AnimeNet was designed to address a more complex detection task by considering seven different classes, which include different levels of violence and pornographic content. This significantly increases the difficulty of accurate classification compared to methods targeting only two or three classes. The presence of multiple classes introduces greater variability and complexity in the dataset, making it inherently more challenging to achieve high accuracies.

Additionally, the detection of both violence and pornographic content introduces a wider range of visual characteristics and semantic cues that need to be effectively captured by the algorithm. The complexity of these detection goals requires the model to be trained on diverse and representative datasets, encompassing a wide range of potential instances and variations.

Another factor to consider is the availability and size of the dataset used for training my model. Larger datasets with diverse and balanced samples can enhance the model's ability to generalize and make accurate predictions. If the dataset used for training my model was relatively small or imbalanced, it could have impacted the overall performance and contributed to the lower accuracy.

Although AnimeNet achieved a lower accuracy of 74.5% compared to some of the other methods, it is crucial to consider the complexity of the detection task, the number of classes targeted, dataset characteristics, and the challenges associated with subjectivity and context in determining inappropriate content. Despite its limitations, my method offers the advantage of simultaneously addressing violence and pornographic content detection, which can be valuable in certain applications. Future

work should focus on refining the algorithm, increasing the dataset size, and exploring techniques to mitigate false positives and false negatives, ultimately aiming to improve the overall accuracy of the detection system.

## 5 Conclusion

In this paper, I present an automatic classification system designed to regulate the content of cartoons and animations by identifying violence and erotic levels. Given the unique characteristics of violent and erotic scenes, I propose AnimeNet, a novel deep-learning model capable of effectively and reliably extracting information from animated images. A novel channel-spatial attention module is introduced and integrated into the model's block structure to extract crucial features from feature maps.

Experimental results demonstrate that AnimeNet surpasses other deep learning models in classifying violence and erotic levels in various images. The proposed model achieved overall accuracies of 74.5%, 72.7%, and 65.8% for 7:3, 5:5, and 3:7 training-testing ratios, respectively. Furthermore, AnimeNet was tested in diverse noisy environments typically found in animated videos, and it maintained high performance under these challenging conditions. This showcases the model's effectiveness in processing animated scenes efficiently.

While the proposed method demonstrates impressive performance, there are opportunities for further improvement. This study's limitations include the potential non-representativeness of the training dataset due to vast global animation diversity, the limited scope of the violence dataset, and the model's untested transferability to other domains or real-world images. In the future, researchers can focus on diversifying the proposed animation dataset to enhance generalization, integrating advanced techniques to improve AnimeNet's performance, and expanding the model's robustness to diverse types of noise. Additionally, they may aim to test AnimeNet's transferability to other domains and employ comprehensive evaluation metrics for a more thorough performance assessment.

**Acknowledgement:** I would like to extend my sincere gratitude to Chen Yu and Sagar A S M Sharifuzzaman for their invaluable guidance and recommendations throughout the course of this research work. Their wealth of knowledge and experience, along with their continuous encouragement, has been fundamental to the successful completion of this study.

**Funding Statement:** The author received no specific funding for this study.

**Author Contributions:** Study conception and design: Y. Tang; data collection: Y. Tang; analysis and interpretation of results: Y. Tang; draft manuscript preparation: Y. Tang.

**Availability of Data and Materials:** Data will be available upon request.

**Conflicts of Interest:** The author declares that he has no conflicts of interest to report regarding the present study.

## References

[1] J. D. Osofsky, "Exposure and response to community violence among children and adolescents," in *Children in a Violent Society*, 1st ed., New York, USA: The Guilford Press, pp. 33, 1997.

[2] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*, Seville, Spain, pp. 332–339, 2011.
[3] A. Subedar and W. Yates, "The disturbing YouTube videos that are tricking children," *BBC News*, 2017. https://www.bbc.com/news/blogs-trending-39381889
[4] S. Maheshwari, "On YouTube kids, startling videos slip past filters," *The New York Times*, 2017. https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html
[5] D. Chen, "What is Elsagate?," *Reddit*, 2022. https://www.reddit.com/r/ElsaGate/comments/6o6baf/what_is_elsagate/
[6] R. Brandom, "Inside Elsagate, the conspiracy-fueled war on creepy youtube kids videos," *The Verge*, 2017. https://www.theverge.com/2017/12/8/16751206/elsagate-youtube-kids-creepy-conspiracy-theory
[7] C. A. Anderson and B. J. Bushman, "Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature," *Psychological Science*, vol. 12, no. 5, pp. 353–359, 2001.
[8] T. Hughes, "Unrestricted film list project," *Hispania*, vol. 90, no. 3, pp. 524–526, 2007.
[9] K. Habib and T. Soliman, "Cartoons' effect in changing children mental response and behavior," *Open Journal of Social Sciences*, vol. 3, no. 9, pp. 248–264, 2015.
[10] E. Ditsanthia, L. Pipanmaekaporn and S. Kamonsantiroj, "Video representation learning for CCTV-based violence detection," in *2018 3rd Technology Innovation Management and Engineering Science (TIMES-iCON)*, Bangkok, Thailand, pp. 1–5, 2018.
[11] S. A. Sumon, R. Goni, N. B. Hashem, T. Shahria and R. M. Rahman, "Violence detection by pretrained modules with different deep learning approaches," *Vietnam Journal of Computer Science*, vol. 07, no. 1, pp. 19–40, 2019.
[12] J. Chen, Y. Xu, C. Zhang, Z. Xu, X. Meng *et al.,* "An improved two-stream 3D convolutional neural network for human action recognition," in *2019 25th Int. Conf. on Automation and Computing (ICAC)*, Lancaster, UK, pp. 1–6, 2019.
[13] L. Ye, T. Liu, T. Han, H. Ferdinando, T. Seppänen *et al.,* "Campus violence detection based on artificial intelligent interpretation of surveillance video sequences," *Remote Sensing*, vol. 13, no. 4, pp. 628, 2021.
[14] K. Yousaf and T. Nawaz, "A deep learning-based approach for inappropriate content detection and classification of YouTube videos," *IEEE Access*, vol. 10, pp. 16283–16298, 2022.
[15] M. Khan, M. A. Tahir and Z. Ahmed, "Detection of violent content in cartoon videos using multimedia content detection techniques," in *2018 IEEE 21st Int. Multi-Topic Conf. (INMIC)*, Karachi, Pakistan, pp. 1–5, 2018.
[16] L. Wang, J. Zhang, M. Wang, J. Tian and L. Zhuo, "Multilevel fusion of multimodal deep features for porn streamer recognition in live video," *Pattern Recognition Letters*, vol. 140, pp. 150–157, 2020.
[17] M. Mohamed, "Applying deep learning to classify pornographic images and videos," arXiv preprint arXiv:1511.08899, 2015.
[18] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni *et al.,* "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, vol. 230, pp. 279–293, 2017.
[19] C. Yuan and J. Zhang, "Violation detection of live video based on deep learning," *Scientific Programming*, vol. 2020, pp. 1–12, 2020.
[20] R. Shen, F. Zou, J. Song, K. Yan and K. Zhou, "EFUI: An ensemble framework using uncertain inference for pornographic image recognition," *Neurocomputing*, vol. 322, pp. 166–176, 2018.
[21] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 113–123, 2019.
[22] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
[23] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, 2016.

[24] D. P. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[25] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4700–4708, 2017.
[26] A. Howard, M. Sandler, B. Chen, W. Wang, L. C. Chen *et al.,* "Searching for MobileNetV3," in *2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 1314–1324, 2019.
[27] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. on Machine Learning*, Long Beach, CA, USA, pp. 6105–6114, 2019.
[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* "An image is worth $16 \times 16$ words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2010.
[29] S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1492–1500, 2017.
[30] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell *et al.,* "A convnet for the, 2020s," in *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 11976–11986, 2022.
[31] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He and P. Dollar, "Designing network design spaces," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 10428–10436, 2020.
[32] S. Zagoruyko and N. Komodakis, "Wide residual networks," arXiv preprint arXiv:1605.07146, 2016.