International Journal on Semantic Web and Information Systems
Volume 20 • Issue 1 • January-December 2024

Multi Frame Obscene Video Detection With ViT:
An Effective for Detecting Inappropriate Content

Dingju Zhu
https://orcid.org/0000-0002-5907-3349
South China Normal University, China

Xilin Shan
South China Normal University, China

Chao Wu
South China Normal University, China

KaiLeung Yung
Hong Kong Polytechnic University, China

Andrew W. H. Ip
https://orcid.org/0000-0001-6609-0713
University of Saskatchewan, Canada

ABSTRACT

With the development of the Internet, people are surrounded by various types of information daily, including obscene videos. The quantity of such videos is increasing daily, making the detection and filtering of this information a crucial step in preventing its spread. However, a significant challenge remains in detecting obscene information in obscure scenarios, like indecent behavior occurring while wearing normal clothing, causing significant negative impacts, such as harmful influence on children. To address this issue, an innovative multi frame obscene video detection base on ViT is proposed by this manuscript per the authors, aiming to automatically detect and filter obscene content in videos. Extensive experiments conducted on the public NPDI dataset demonstrate that this method achieves better results than existing state-of-the-art methods, achieving 96.2%. Additionally, it achieves satisfactory classification accuracy on a dataset of obscure obscene videos.This provides a powerful tool for future video censorship and protects minors and the general public.

KEYWORDS

Pornography classification, Obscene Video Detection, Computer Vision, Video Classification, Video Analysis, Deep Learning, Self Attention Mechanism, Vision Transformer, ViT-based Models

INTRODUCTION

The detection of obscene content has been a long-standing concern in human society. Before the advent of the internet, obscene materials mainly existed in the form of printed publications, videotapes, or films. Consequently, the methods for detecting and filtering such content primarily relied on manual review and legal regulation by relevant authorities. With the emergence and development of the internet, the dissemination of obscene content has shifted to online videos and images, making such information easily accessible. According to a report from July 2019, Baidu, Inc., a Chinese

DOI: 10.4018/IJSWIS.359768

multi-national technology company specializing in internet services and artificial intelligence, managed to handle a total of 31.25 billion pieces of harmful information in the first half of 2019, with obscene content being a major target, accounting for 46.97% of the total.

Despite the stringent review mechanisms and rating systems in place across various countries and regions to protect people from obscene content in film and media, user-uploaded content on the internet is subject to less rigorous restrictions and is more readily accessible. This makes the detection and filtering of obscene videos and images an increasingly pressing issue today. Moreover, the subjectivity and uncertainty inherent in detecting obscene content present greater challenges compared to other classification problems. For instance, even humans can struggle to accurately judge the extent of obscene content in videos, with different individuals potentially having varying assessments of the same content, such as when evaluating someone wearing a swimsuit or underwear. In the past, detection methods primarily focused on classifying images based on the amount of exposed skin in the video. However, this approach misses one of the key characteristics that distinguish video from images—the potential dynamic motion information contained within the video. In many cases, multiple consecutive images or video frames are necessary to capture coherent actions or context within the content, which is crucial for more precise classification. However, most models focus on single-frame detection and classification, while we also note the potential interactivity of vision transformers (ViTs).

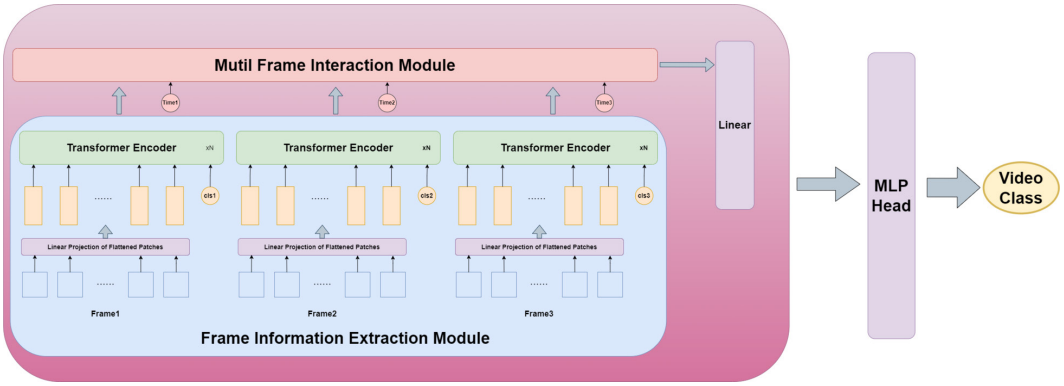
In this paper, we propose a novel method for detecting obscene video content based on a ViT, named the multi-frame obscene video detection model based on a ViT (MFOVD-ViT). We leverage the superior capabilities of the ViT in image classification and the ability of the attention mechanism that plays a crucial role in it. Thus, the MFOVD model mainly relies on the attention mechanism, using the ViT for frame information extraction and embedding temporal encoding during frame attention interaction to handle temporal relationships. This structure accomplishes both tasks described above without requiring the integration of multiple models. Testing on the NPDI data set shows that the MFOVD model achieves higher accuracy than the current best methods.

In the application of our proposed model, the social and ethical impacts need to be carefully considered. While the MFOVD-ViT enhances the automation and accuracy of detecting obscene content, it still faces potential issues related to model bias. If the training data sets lack diversity in certain aspects, such as gender, race, or cultural background, the detection results may be unfair. For example, content from specific backgrounds might be overly detected or misclassified due to biases present in the model's training data, leading to unnecessary social and ethical concerns. Therefore, ensuring fairness and diversity in the data sets is crucial during the model development process to guarantee that this technology can treat all types of content equitably and minimize bias. Therefore, we made our best effort in the process of collecting the training data set, which has helped our model to largely overcome such bias issues.

RELATED WORKS

Over the past many years, numerous detection and filtering methods have been proposed, leading to significant advancements in this field. Early methods for detecting obscene content primarily focused on identifying the proportion of exposed skin and clothing or environmental factors in images. These methods involved setting algorithm thresholds through research and experimentation to determine whether an image exceeded the threshold and thus violated content guidelines. However, a critical issue with these methods is their susceptibility to false positives in specific contexts, such as beach scenes where minimal clothing is common. Subsequently, visual bag-of-words models (BoW), such as BossaNova, have been applied to the detection of obscene content. These models use visual analogs of words by clustering low-level visual features (including texture and color) of points or local regions during the vector quantization process.

Figure 1. Overall framework of MFOVD



In recent research, machine-learning methods have increasingly been used to address this issue. In many studies, researchers use the NPDI data set as a benchmark for testing and validation. ACORDE is currently the model achieving quite excellent results on this data set. In this study, the authors proposed a model based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), where CNNs are primarily responsible for extracting image features, and RNNs are responsible for modeling video sequences. We are inspired by this approach, recognizing that detecting obscene video content should effectively address both tasks.

At this point, we considered whether a single model or a unified approach could address both tasks. We then turned our attention to the ViT model because the ViT model itself can be used for feature extraction in images, and its internal structure involves the concept of interaction. This aligns with our goal of achieving interaction between video frames, corresponding to the processing of video sequences. By handling both tasks with a more unified implementation and model, the ViT demonstrates better performance.

More detailed information can be found in the Baseline Methods subsection of the Experimental Methods and Results Analysis section.

APPROACH

Overview

To further filter and detect obscene content in videos, we proposed a new method, the MFOVD model, which is entirely based on the attention mechanism for detecting adult content in videos. The overall framework of MFOVD is shown in Figure 1. The framework of MFOVD mainly includes a frame information extraction module composed of multiple ViT networks and an interaction module for interaction and embedding time information. In this section, we comprehensively explain the two modules separately. The flowchart of MFOVD is presented in Figure 2.

Frame Information Extraction Module

We utilized the ViT model as the backbone network for the frame information learning module. Figure 3 is a structural diagram released by the ViT research team in their paper. Unlike traditional convolutional approaches, the ViT solely relies on the attention mechanism and feedforward neural networks to capture the global relationships among different regions in an image. In contrast, convolution-based models use local receptive fields to extract local features of the image. The ViT model exhibits greater similarity between representations obtained in shallow and deep layers, making it more suitable for capturing global relationships and complex patterns within images. In the context

Figure 2. Flowchart of MFOVD

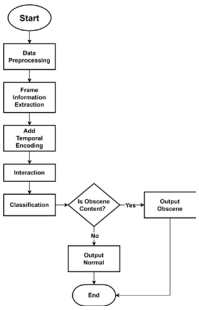
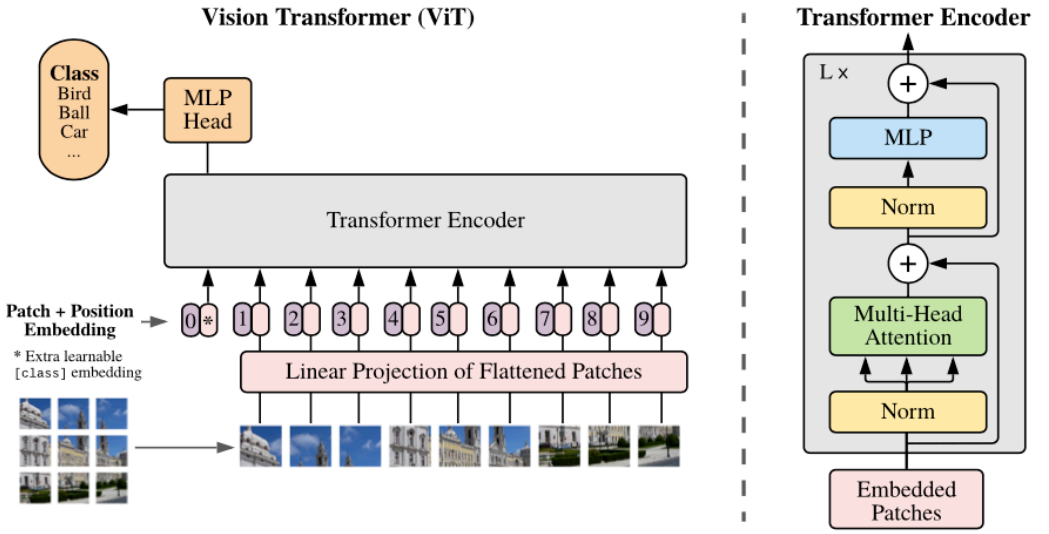


Figure 3. Structure of ViT



of obscure obscene environments, where exposed parts are minimal or occasionally out of the camera's view, it is imperative to focus on the global relationships within images and the relationships between frames, rather than solely on the local features that depict exposed areas.

The transformer model, renowned for its powerful sequence modeling capabilities, employs the self-attention mechanism for efficient sequence data modeling. Initially achieving tremendous success in the field of natural language processing, its efficacy has been extended to other domains. The ViT epitomizes the application of the transformer model in the visual domain.

Taking ViT-based as an example, the ViT first applies an embedding layer for data transformation. The input image (224x224) is divided into patches of 16x16 pixels, resulting in 196 patches. Each patch is linearly mapped to a one-dimensional vector, transforming data from the shape into a sequence of vectors of length 768. A positional encoding is then added to each vector sequence to represent the spatial information of each patch within the image, along with a class token, a trainable parameter identical in format to other vector sequences, also with a length of 768. The transformer encoder comprises 12 identical layers, each containing two sublayers: a multi-head self-attention mechanism and a feed-forward neural network. Each sublayer is followed by a residual connection and layer normalization. The multi-head self-attention mechanism calculates the relevance between each vector and others, weighting them accordingly to generate new vectors, while the feed-forward

neural network applies nonlinear transformations to enhance the model's expressive capability. This process encodes the vector sequence and extracts information from the image.

In the information extraction module, three identical ViT models process different frames for frame-information learning. This approach seemingly requires integrating three ViTs for training and computation, potentially making the model cumbersome. However, we do not train the ViT model's parameters ourselves; instead, we utilize the pretrained models released officially. This decision is motivated by two factors: First, training the ViT model necessitates substantial data and computational resources, which are beyond our means. Second, the pretrained models adequately meet our needs and have demonstrated impressive results. Therefore, by leveraging pretrained models, we only need a single ViT model to process three different frames simultaneously, significantly reducing our training and computational costs. When we did not use pretrained models for task training, the performance was quite poor. As we continuously increased the amount of training data, the performance gradually improved. However, as the data volume increased, the improvement became progressively slower. This means that in the later stages, we would need to put in significant effort to collect vast amounts of data to achieve only a slight performance improvement. Moreover, this improvement would merely approach the results of using pretrained models, without surpassing them.

Multi-Frame Interaction Module

To handle obscene actions with simple and repetitive characteristics, we developed an interaction module for frames, which facilitates information exchange between consecutive frames after frame-information learning and incorporates explicit temporal encoding. The structure of this module is depicted in Figure 4.

It is easy to observe that, with the addition of input frames, the number of layers in the structure increases significantly. For example, when there are 3 frames, the number of interactions is 2; when the number of frames increases to 4, the number of interactions increases to 3 and, at this point, 2 additional layers of the transformer encoder are required. This greatly increases the cost of training, as this structure has particularly high demands on data volume. Therefore, after careful consideration, we decided to set the number of frames to 3. Similarly, for similar reasons, we set the transformer encoder to 2 layers after each interaction.

The frame interaction module's primary objective is to model temporal dependencies between video frames effectively. It uses the self-attention mechanism, a crucial component of the ViT, to attend to different frames and capture temporal relationships. By embedding temporal encoding into the frame information, the module maintains the chronological order of frames, ensuring that the temporal dynamics of the video are adequately represented.

In detail, the process begins with extracting information representations from individual frames using the ViT, as described in the information extraction module. This extracted information is then fed into the interaction module, where temporal encoding is added to each frame's information vector to incorporate temporal information. The temporal encoding is crucial for preserving the sequence order and providing context to each frame within the video, and the process is as shown in Equations 1, 2, 3, and 4.

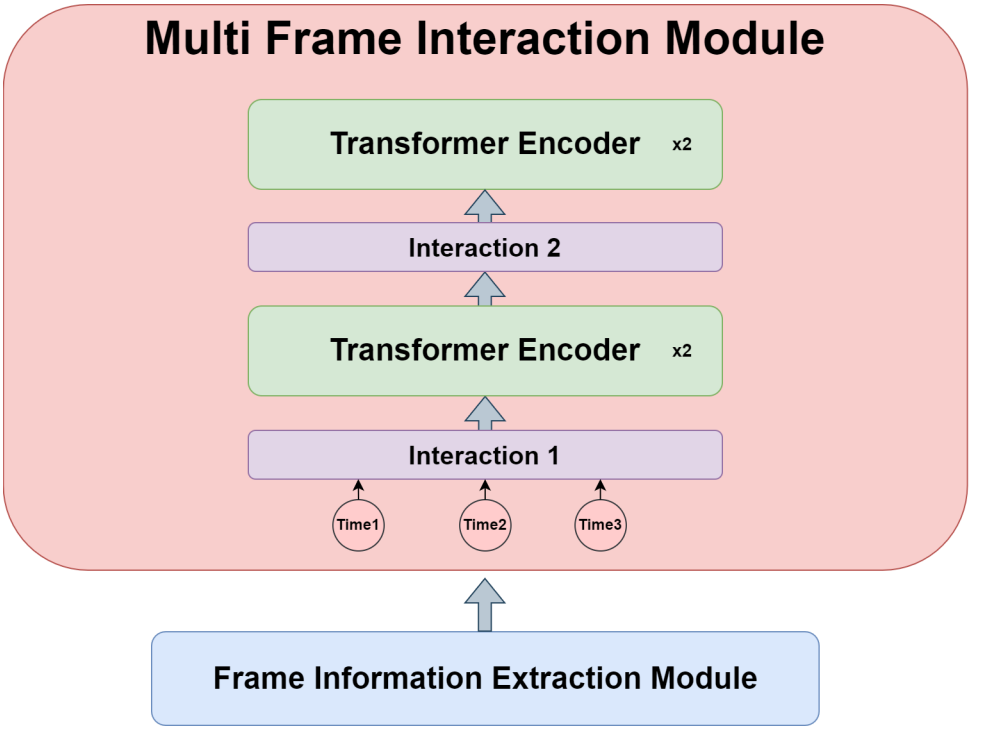
$$\text{Frame Features} = \left[\text{Feature}_0, \text{Feature}_1, ..., \text{Feature}_T\right] \tag{1}$$

$$\text{Temporal Features} = \left[\text{Temp}_0, \text{Temp}_1, ..., \text{Temp}_T\right] \tag{2}$$

wherein:

$$\text{Feature}_i = \left[\text{Token}_0(\text{class}), \text{Token}_0, ..., \text{Token}_P\right] \tag{3}$$

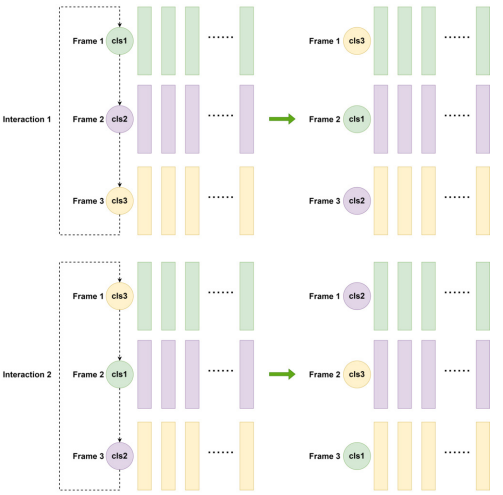
Figure 4. Structure of multi-frame interaction module



$$\text{Input to Multi} - \text{Frame Interaction Module} = \text{Frame Features} + \text{Temporal Encoding} \tag{4}$$

It should be noted that the addition here only adds temporal encoding to each Token0 (class). T is the total number of features and P is the number of tokens for each feature.

Figure 5. Interaction process



Each image's information sequence includes a class token, which is considered to have learned significant information from other tokens in the image. Our approach is to use this class token as a representative to learn information from other frames, thereby achieving interaction. As in the ViT model, the class token ultimately performs the classification task. In this study, the designed interaction method, illustrated in Figure 5, involves exchanging the class tokens of two sequences to be interacted with. Subsequently, each sequence passes through two layers of transformer encoder to learn interframe information, which constitutes an information interaction. The key component in this process is the multi-head self-attention mechanism within the transformer encoder, calculated as shown in Equations 5, 6, and 7.

A standard qkv self-attention:

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv}\mathbf{U}_{qkv} \in \mathbb{R}^{D\times 3D_h}, \tag{5}$$

$$A = \text{softmax}\left(\mathbf{q}\mathbf{k}^{\top}/\sqrt{D_h}\right)A \in \mathbb{R}^{N\times N}, \tag{6}$$

$$\text{SA}(\mathbf{z}) = A\mathbf{v}. \tag{7}$$

For each element in an input sequence $z \in \mathbb{R}^{N\times D}$, we compute a weighted sum over all values v in the sequence. The attention weights A_{ij} are based on the pairwise similarity between two elements of the sequence and their respective query q_i and key k_j representations, as shown in Equation 8.

$$\text{MSA}(\mathbf{z}) = \left[\text{SA}_1(z); \text{SA}_2(z); \cdots ; \text{SA}_k(z)\right]\mathbf{U}_{msa}\text{U}_{msa} \in \mathbb{R}^{k\cdot D_x\times D} \tag{8}$$

Multi-head self-attention, MSA, is an extension of SA in which we run k self-attention operations, called "heads," in parallel, and project their concatenated outputs. To keep compute and number of parameters constant when changing k, D_h is typically set to D/k.

The previous section dealt with matrix calculations in specific computer operations, which can be difficult to understand. Now, we describe certain specific and relatively easy-to-understand processes in our own way, as illustrated in Figures 6, 7, and 8. Since we are selecting and describing one or a particular result, and some parameters or descriptive characters need to be distinguished from those in the earlier sections, we have chosen different symbols for the description. However, we provide a detailed explanation of these.

Figure 6 illustrates the computation process of the attention mechanism, which is a widely used technique in deep learning. The core idea is to process information more effectively by assigning different importance weights to different parts of the input data. To explain the detailed steps for calculating the final weight b_1 from a query vector q_1 and a key vector k_1 as shown in the figure, here, a_1 represents a specific element in the sequence. The query vector represents the information of the current word or token, which is used to "query" the relationships between this word and others in the sequence. The key vector represents the features of each word or token in the sequence, and is used to calculate similarity with the query to determine the importance of other words to the current word. The value vector represents the actual information, and it is eventually weighted by the attention weights to produce the output. Both the key and value vectors are obtained through linear transformations of a_1. α_1 represents the result of the similarity calculation between the query and key, which is then normalized by a softmax function. This process first obtains a similarity score and then converts it into a probability. Finally, the weighted sum with the key vector is computed to get the final output b_1.

Figure 7 illustrates the self-attention mechanism. Both the attention mechanism and the self-attention mechanism are techniques used in deep learning to enhance the model's focus on different parts of the input sequence. However, they have significant differences. The self-attention

Figure 6. Attention mechanism

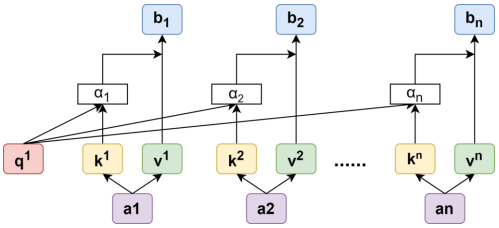
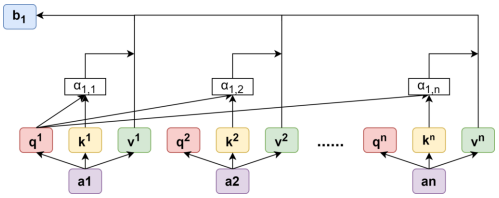


Figure 7. Self-attention mechanism



mechanism is a specific type of attention mechanism that operates within the sequence itself, meaning it calculates attention scores between different positions in the same input sequence.

Next, we focus on the key differences between self-attention and the general attention mechanism. The first difference is the source of the query, key, and value vectors. In the attention mechanism, the query is typically from the decoder, focusing on the key and value from the encoder, which facilitates information transfer between different modules. In contrast, the self-attention mechanism operates within a single sequence, where the query, key, and value all come from the same input sequence. This allows each input element to attend to every other element in the sequence. The second difference is that, in self-attention, the mechanism attends to all other elements in the sequence (including itself). Pay attention to how b_1 is calculated in both Figures 6 and 7. This mechanism can capture long-range dependencies within the sequence.

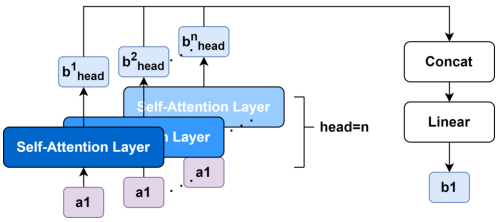
Figure 8 illustrates the multi-head self-attention mechanism, which is one of the core components of the ViT model. It is an extension of the self-attention mechanism, where multiple independent attention heads are computed in parallel, allowing the model to capture different dimensions of information from the sequence.

The core idea behind the multi-head self-attention mechanism is that, instead of using a single attention head to compute attention relationships across the entire sequence, multiple attention heads are employed. Each attention head independently computes different feature information, and their results are then concatenated. After this, a linear transformation is applied to obtain a richer representation. This structural characteristic is clearly depicted in Figure 8, where particular attention can be given to how b_1 is derived.

Then we currently select three frames as input, meaning that each input image information sequence must complete two rounds of information interaction (i.e., learning from the other two frames). This necessitates stacking two exchange modules. Notably, in the second interaction module, (a) it is no longer necessary to add temporal encoding again, and (b) the images involved in interaction differ from those in the first layer (e.g., for Frame 1, Frames 1 and 2 interact in the first layer, while Frames 1 and 3 interact in the second layer, and so on).

Each class token is added at the beginning of the input sequence and does not correspond to any part of the image. Through the feature extraction module, it interacts with other tokens via the self-attention mechanism to accumulate global information about the entire image. The continuous

Figure 8. Multi-head self-attention mechanism



frame interaction module then facilitates information exchange between frames to learn from other frames' information. Each class token is finally passed to a simple fully connected layer (linear layer). After the fully connected layer, a softmax activation function is used to generate the probability distribution for each category, which is ultimately used to classify the images. The chosen loss function for the classification task is the cross-entropy loss function, defined:

For a single sample, assuming the true distribution is Y and the network output distribution is $\hat{Y}$, with a total number of categories n, the calculation method for the cross-entropy loss function in this case is as shown in Equation 9.

$$Loss = -\sum_{i=1}^{n} y_i \log \hat{y}_i \tag{9}$$

In past research on video classification algorithms, a common process involved frame information extraction, processing the information between frames, and completing the classification. Multiple studies have shown that this is a feasible research path. Therefore, we based our research on this process, using components from the ViT or ideas that rely on attention mechanisms within the ViT to design different modules to replace each step of the process. As a result, the modules are also closely connected. For example, the effectiveness of the multi-frame interaction module heavily depends on the quality of the information extracted by the frame information extraction module. Additionally, the accuracy of the multi-frame interaction module significantly decreases in the absence of any information extraction. Moreover, the key components of both parts are multi-head self-attention structures, which greatly enhance the uniformity of feature extraction and the stability of the model between the two modules. This comprehensive process ensures that the model effectively learns and interprets both spatial and temporal dynamics of video frames, significantly enhancing the accuracy of detecting obscene content. By utilizing the self-attention mechanism and temporal encoding within the continuous frame interaction module, our proposed method demonstrates superior capability in filtering and identifying obscene content in videos, even in challenging scenarios where single-frame analysis might be insufficient.

EXPERIMENTAL METHODS AND RESULTS ANALYSIS

Data Set and Experimental Setup

To validate the performance of our evaluation framework, we conducted obscene video classification tasks on the NPDI benchmark data set and our own constructed obscure pornographic data set (OPD).

The NPDI data set, collected by the NPDI group at the Federal University of Minas Gerais, Brazil, contains nearly 80 hours of video, consisting of 400 pornographic and 400 nonpornographic videos. The pornographic videos are sourced from the internet, featuring individuals of various skin tones and ethnicities, encompassing a variety of pornographic content. The nonpornographic videos

are divided into two categories: 200 easy-to-judge videos and 200 difficult-to-judge videos. The easy videos are regular nonpornographic content, while the difficult videos include scenes with significant skin exposure but no pornographic elements (keywords include beach, wrestling, swimming, etc.).

For each video, we selected several keyframes from different shots (approximately 20 per video) to represent each shot. These keyframes were not strictly consecutive but were chosen to capture subtle motion changes. The frames we selected were carefully screened to exclude those with poor lighting or blurry frames.

The OPD was created to further validate our model's capability. We believe that pornographic videos can also be categorized into easy and difficult classes. The easy class consists of overtly explicit scenes with significant skin exposure and sexual acts, which are easy to identify. The difficult class includes scenes where sexual acts occur without significant skin exposure, such as clothed sexual acts where exposed parts are not prominently captured by the camera. We collected 200 easy and 200 difficult pornographic videos (mostly clothed sexual acts due to constraints) and 400 regular nonpornographic videos to form the OPD. To ensure diversity in the data, we tried to collect videos based on various factors such as different skin tones, shooting styles, and age groups. Among them, easy classes include scenes with obvious nudity, significant skin exposure, and easily recognizable sexual behavior. The difficult category includes scenes where sexual activity occurs without significant skin exposure and scenes where sexual activity involves wearing clothing, where the exposed parts are not clearly captured by the camera. For each video, we selected keyframes from different shots to represent each shot. These keyframes were not strictly consecutive but were chosen to capture subtle motion changes. To overcome the issue of diversity in the OPD, we repeatedly expanded our collection range, which resulted in us spending a significant amount of time on collecting, screening, and filtering the data. The method for selecting keyframes was similar to that used for the NPDI data set.

In our experiments, we chose the ViT-based version as the backbone network. The training process included 600 epochs. We used stochastic gradient descent to train the model, with a momentum of 0.9 and a weight decay of $5*10^{-5}$. The batch size was set to 64 (3 frames per batch). To optimize the learning process, the learning rate adjustment strategy was employed as per (10), where the initial value f is set to 0.01 and the initial learning rate set to 0.001 and updated according to the adjustment strategy. The momentum update hyper parameter was set to 0.99. The confidence score threshold was set to 0.5, as shown in Equation 10.

$$y = \frac{1}{2}\left(1 + \cos\left(\frac{t\pi}{T}\right)\right) \cdot (1 - f) + f \tag{10}$$

Baseline Methods

To validate the superiority of our framework, we conducted a performance analysis of the method proposed in this paper against other state-of-the-art pornographic and obscene video classification methods on the NPDI data set.

BossaNova-HueSIFT

This approach addresses the problem within the visual BoW direction. It enhances this representation by retaining the histogram of distances between the descriptors in the image and those in the codebook, thus preserving important information about the local descriptor distribution around each codeword. BossaNova has also shown good results in the challenging real-world application of obscene content detection.

BossaNova-BRISK and BNVD

Considering the development of low-level local features and the emergence of mid-level representations, a new video descriptor was introduced. This descriptor is used in combination with

the mid-level representation BossaNova and local binary descriptors to address the detection of obscene content.

BoW-VD

BoW-VD describes a video descriptor based on binary features (BinBoost), which is used in conjunction with BoW/BossaNova representations. This method leverages the efficiency and robustness of binary features to create a compact and effective video representation. By combining these binary features with the BoW model or its variant, BossaNova, BoW-VD can efficiently summarize and index video content, making it suitable for tasks such as video retrieval, categorization, and analysis. The use of BinBoost ensures that the descriptors are both discriminative and computationally efficient, facilitating scalable video processing.

AGbNet

AGbNet represents the pioneering application of deep learning to the challenge of obscene content detection. The authors introduce an innovative approach that involves fine tuning two distinct CNNs, specifically AlexNet and GoogLeNet. This method leverages the strengths of both architectures to enhance the accuracy and robustness of the detection process. AlexNet, known for its success in the ImageNet large-scale visual recognition challenge, provides a powerful baseline due to its deep architecture and effective feature extraction capabilities. On the other hand, GoogLeNet introduces a novel structure with its inception modules, which allow the network to capture multi-scale features through parallel convolutional operations within the same layer. By fine tuning these pretrained CNNs, AGbNet effectively adapts the models to the specific nuances of obscene content detection.

ACORDE

ACORDE integrates CNNs and RNNs to handle feature extraction from images and the temporal information of videos, respectively. By leveraging the strengths of both CNNs and RNNs, ACORDE is designed to perform complex tasks such as image recognition and video analysis with high accuracy and efficiency. The CNN component of ACORDE excels in identifying and extracting intricate patterns and features within static images, while the RNN component is adept at capturing and processing sequential and temporal dynamics in video frames. This combination enables ACORDE to provide robust performance in various applications, including surveillance, autonomous driving, and multi-media content analysis.

KidsGUARD

KidsGUARD is a comprehensive system designed for monitoring and protecting children in various environments. In this work, VGG16 CNN and long short-term memory based autoencoder were combined for learning video descriptors and processing classification tasks, respectively. The VGG16 model, known for its deep architecture and success in image recognition, is utilized to extract detailed features from video frames, ensuring high-quality descriptor learning. On the other hand, the long short-term memory based autoencoder addresses the temporal dynamics of video sequences, capturing patterns over time to enhance classification accuracy. By integrating these advanced machine-learning techniques, KidsGUARD effectively analyzes video data, providing reliable and robust monitoring capabilities for the safety and security of children.

DOCAPorn

This work proposes a method that recognizes the obscene images through the one-class classification model based on neural networks and introduces the visual attention mechanism to enhance the performance of recognition.

Table 1. Comparison between our model and state-of-the-art methods in NPDI

Method	Accuracy	Recall	F1
BossaNova-HueSIFT	89.5%	-	-
BossaNova-BRISK	88.6%	-	-
BNVD	92.0%	0.923	0.921
BoW-VD	92.4%	0.93	0.927
AGbNet	94.1%	-	-
ACORDE	95.6%	-	-
KidsGUARD	89.0%	0.85	0.869
DOCAPorn	95.6%	-	-
Fine-grained pornographic image recognition with multiple feature fusion transfer learning	94.3%	0.85	0.894
Audio-base pornographic detection	95.7%	-	-
Ours	96.2%	0.928	0.9448

Fine-Grained Pornographic Image Recognition With Multiple-Feature Fusion Transfer Learning

In this paper, the authors propose a deep learning based approach with multiple-feature fusion transfer learning strategy. A pretrained model is used to initialize the network and help extract the basic features, and then a fusion method that makes use of multiple-transfer learning models in inference is proposed, to improve the accuracy on the test set.

Audio-Based Pornographic Detection

This work explores obscene sound modeling based on different neural architectures and acoustic features and finds that CNN trained on log mel spectrogram achieves a good performance on the NPDI data set.

Experimental Results

The experimental results of other state-of-the-art methods are derived from the original papers. The results are shown in Table 1. From the experimental results, we can observe that our proposed method outperforms the compared baselines. It is worth noting that, in DOCAPorn, under attack, the accuracy is 95.6%. However, the recall rate and F1 score are not provided. In contrast, under no attack, the accuracy, recall rate, and F1 score are given as 98.419%, 0.993, and 0.984, respectively. However, this cannot be considered a completely reliable reference. Specifically, compared to the currently best-performing methods, our method shows an improvement in classification accuracy, achieving a 0.6% increase on the NPDI data set, with a recall rate and F1 score reaching 0.928 and 0.9448, respectively. This improvement demonstrates the superiority of our proposed model framework.

We utilized a unified model and approach to accomplish both tasks, significantly reducing the gap between them and thereby minimizing information loss, which in turn enhances the model's performance. Temporal encoding explicitly indicates the temporal information of frames, which also reduces the difficulty of the model during learning, thereby improving its performance, although this improvement is quite limited. We believe the superiority of our model lies in the fact that we used the ViT model as the foundation and thoroughly implemented the attention mechanism concept. This enabled our model not only to extract information from individual frames with better attention to the global context of the image but also to capture the overall temporal relationships during subsequent interactions.

Table 2. Accuracy results of ablation study on NPDI and OPD

Component	NPDI	OPD
Without Multi-Frame Interaction Module & Temporal Encoding	92.1%	80.3%
Without Temporal Encoding	95.4%	88.0%
Full Model	96.2%	88.4%

Ablation Study

In the ablation study, our goal was to analyze the performance of the proposed method by systematically assessing the impact of different model configurations. Specifically, we compared the results of the complete model with those obtained by removing specific components or modifying certain aspects of the model architecture. The comparison results are shown in Table 2.

Our findings indicate that the removal of any module from our framework significantly decreases the model's performance. This result underscores the critical role each of the proposed components plays in the overall model. It is evident that, when the two modules we introduced were removed, the test accuracy drops significantly, especially in the OPD. This highlights the importance of the two modules we added. We believe this is because our designed multi-frame module plays a crucial role, as clothing in the OPD is generally normal, making it difficult to classify based on the degree of skin exposure alone. Detecting subtle changes or associations across different frames greatly enhances detection accuracy. When the temporal encoding is removed, the detection performance decreases, but to a relatively smaller extent. The experiment shows that, while this module does improve the detection performance, it is not critical. We have explored various methods of embedding temporal encoding, such as sinusoidal positional encoding and relative positional encoding, but the results are less than ideal. In some cases, these methods even decrease detection accuracy.

The information in the OPD is relatively vague and implicit, which makes it more difficult for the model to learn motion information or the sequential information between frames. Temporal encoding explicitly highlights the temporal characteristics of frames, and we believe this helps the model learn more efficiently, reducing both the learning cost and time. Therefore, temporal encoding plays a more significant role in the OPD.

Although the training and inference of the model operate as a black box, we have attempted to explain these results. We believe that the multi-frame interaction module not only captures the exposure information within individual frame images but also effectively captures subtle human motion changes between frames. Meanwhile, the temporal encoding enhances this effect by providing stronger temporal relationship logic. Therefore, each part plays a critical role.

Overall, these ablation studies provide valuable insights into the individual contributions of each component and emphasize their importance in the method’s integration, ultimately enhancing the performance in the task of obscene video classification.

CONCLUSION AND OUTLOOK

In this paper, we proposed a deep-learning architecture based on the ViT model with multi-frame input, incorporating multi-frame modules and temporal encoding for the automatic detection and filtering of obscene content in videos. Extensive experiments on benchmark data sets verify that our method outperforms state-of-the-art baselines. To address the task of detecting obscene content in videos where clothing is generally normal, we collected the OPD data set, demonstrating that our framework can effectively perform this task.

The design of the multi-frame input structure is one of the core advantages of our method. Traditional detection methods typically only consider analyzing a single frame or a limited number

of frames, neglecting the correlation between frames in a video. However, when detecting obscene content, the continuity of actions and scene changes are often critical clues. By introducing multi-frame input, our method can capture changes in the temporal dimension of a video, thereby more accurately identifying potential obscene content. This comprehensive use of temporal sequence information enables the model not only to recognize features within a single frame but also to understand the continuity and variation patterns of actions, which is challenging for traditional single-frame detection methods.

The combination of multi-frame modules and temporal encoding further enhances the model's detection capability, even if the improvement is not particularly significant. Multi-frame modules process multiple consecutive frames to extract dynamic information between frames, which is crucial for detecting brief but significant action segments. Temporal encoding helps the model better understand the temporal relationships in video frame sequences, allowing it to perform better when handling scenes with rapid changes or gradual variations. For example, in the OPD data set, in scenes where the clothing is relatively normal, the model can identify subtle movements, thereby improving detection accuracy.

We also attempted to apply our model to the detection of other sensitive video types, such as violence. Even with preliminary training, it shows promising results. However, to overcome certain bottlenecks and achieve top-tier detection accuracy, more training data and adaptive modifications to the model are required. Nonetheless, we have proposed a solid framework that provides valuable insights for future transfer learning applications.

However, we also noted a critical issue: The video frame segmentation before detection or training is particularly important and key. Different frame extraction methods significantly impact the experimental results, which will be a focus of our future research. Additionally, there are many areas in our method that need improvement. We will continue to refine the framework and conduct experiments on real-world data sets to enhance the reliability and performance of the framework.

We believe that with continuous advancements in technology, obscene content detection models will have broader applications across various fields, such as enhanced content review systems to protect minors and the public from harmful content, assisting the video industry with automated content classification to improve efficiency, and preventing and monitoring cybercrime to combat the spread of illegal obscene content.

To further expand our research, we plan to explore several avenues for improvement:

Optimization of frame-extraction strategies: We will investigate various frame extraction strategies and their impact on detection performance. By comparing different techniques, such as fixed-interval sampling, keyframe extraction, and dynamic frame selection, we aim to find the most effective method that balances capturing video dynamics and computational efficiency.

Enhancement of temporal encoding: We plan to further optimize the temporal encoding mechanism to better capture the temporal relationships between frames. Exploring more complex temporal encoding techniques, such as those based on attention mechanisms, will enhance the model's sensitivity to dynamic changes and improve its performance in handling long-term dependencies.

Multi-modal integration: Incorporating audio, text, and other modalities will boost the model’s overall detection capabilities. By integrating multi-modal data, we aim to create a more comprehensive detection framework that can provide high-precision results in more complex scenarios, thereby enhancing the robustness of our detection system. The combination with audio is our next most important focus.

Data augmentation and synthesis: To increase the diversity of our data set and improve the model's generalization abilities, we will employ data augmentation techniques and synthesize new training samples. Techniques such as image flipping, rotation, cropping, and using generative adversarial networks to generate new samples will be utilized to enrich the data set.

By implementing these improvements, we aim to maintain high accuracy while enhancing the model's efficiency and applicability, thereby providing more robust support for video-content

moderation. Additionally, implementing our model into practical applications is one of our key efforts, such as in live streaming or video platforms for fast real-time content detection. This means that, while considering accuracy, we also need to take detection speed into account. Our ultimate objective is to develop an efficient, reliable, and broadly applicable automated video-review system, driving further advancements in this field.

The proposed method, which enhances the classification of obscene videos, carries several potential social and ethical implications. On the positive side, it can contribute to better content moderation on online platforms, helping to prevent the dissemination of harmful or inappropriate content, such as pornography. However, there are also ethical concerns that must be addressed. The development of models capable of detecting such content may lead to over-surveillance or censorship, raising privacy concerns if used improperly. While the proposed method has the potential to positively impact online safety and content moderation, careful consideration of its ethical use, potential biases, and the need for transparency and human oversight is crucial.

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING STATEMENT

This research is partially supported by Guangdong Provincial Demonstrative Industrial College for Artificial Intelligence and Robotics Education.

PROCESS DATES

October 18, 2024
Received: June 14, 2024, Revision: October 6, 2024, Accepted: October 7, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Dingju Zhu (China, zhudingju@m.scnu.edu.cn)

REFERENCES

AlDahoul, N., Abdul Karim, H., Lye Abdullah, M. H., Ahmad Fauzi, M. F., Ba Wazir, A. S., Mansor, S., & See, J. (2020). Transfer detection of YOLO to focus CNN's attention on nude regions for adult content detection. *Symmetry*, *13*(1), 26. DOI: 10.3390/sym13010026

Avila, S., Thome, N., Cord, M., & Valle, E., & AraúJo, A. D. A. (. (2013). Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, *117*(5), 453–465. DOI: 10.1016/j.cviu.2012.09.007

Caetano, C., Avila, S., Guimaraes, S., & Araújo, A. D. A. (2014, September). Pornography detection using BossaNova video descriptor. In *2014 22nd European signal processing conference (EUSIPCO)* (pp. 1681-1685). IEEE.

Caetano, C., Avila, S., Schwartz, W. R., Guimarães, S. J. F., & Araújo, A. D. A. (2016). A mid-level video representation based on binary descriptors: A case study for pornography detection. *Neurocomputing*, *213*, 102–114. DOI: 10.1016/j.neucom.2016.03.099

Carlsson, A., Eriksson, A., & Isik, M. (2008). Automatic detection of images containing nudity: Image detection using artificial neural networks and statistical methods.

Chen, J., Liang, G., He, W., Xu, C., Yang, J., & Liu, R. (2020). A pornographic images recognition model based on deep one-class classification with visual attention mechanism. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 122709–122721. DOI: 10.1109/ACCESS.2020.2988736

Da Silva, M. V., & Marana, A. N. (2018, November). Spatiotemporal CNNs for pornography detection in videos. In *Iberoamerican congress on pattern recognition* (pp. 547–555). Springer International Publishing.

Deselaers, T., Pimenidis, L., & Ney, H. (2008, December). Bag-of-visual-words models for adult image classification and filtering. In *2008 19th international conference on pattern recognition* (pp. 1-4). IEEE. DOI: 10.1109/ICPR.2008.4761366

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. Arxiv preprint arxiv:2010.11929. DOI: 10.1109/ICPR.2008.4761366

Fu, Z., Li, J., Chen, G., Yu, T., & Deng, T. (2021). PornNet: A unified deep architecture for pornographic video recognition. *Applied Sciences (Basel, Switzerland)*, *11*(7), 3066. DOI: 10.3390/app11073066

Gangwar, A., González-Castro, V., Alegre, E., & Fidalgo, E. (2021). AttM-CNN: Attention and metric learning-based CNN for pornography, age and child sexual abuse (CSA) detection in images. *Neurocomputing*, *445*, 81–104. DOI: 10.1016/j.neucom.2021.02.056

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Lin, X., Qin, F., Peng, Y., & Shao, Y. (2021). Fine-grained pornographic image recognition with multiple feature fusion transfer learning. *International Journal of Machine Learning and Cybernetics*, *12*(1), 73–86. DOI: 10.1007/s13042-020-01157-9

Lopes, A. P., de Avila, S. E., Peixoto, A. N., Oliveira, R. S., & Araújo, A. D. A. (2009, August). A bag-of-features approach based on hue-sift descriptor for nude detection. In *2009 17th European signal processing conference* (pp. 1552-1556). IEEE.

Lovenia, H., Lestari, D. P., & Frieske, R. (2022, September). What did I just hear? Detecting pornographic sounds in adult videos using neural networks. In *Proceedings of the 17th international audio mostly conference* (pp. 92-95). DOI: 10.1145/3561212.3561244

Moustafa, M. (2015). Applying deep learning to classify pornographic images and videos. Arxiv preprint arxiv:1511.08899.DOI: 10.1145/3561212.3561244

Nian, F., Li, T., Wang, Y., Xu, M., & Wu, J. (2016). Pornographic image detection utilizing deep convolutional neural networks. *Neurocomputing*, *210*, 283–293. DOI: 10.1016/j.neucom.2015.09.135

Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., & Sirivianos, M. (2020, May). Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the international AAAI conference on web and social media* (*Vol. 14*, pp. 522-533). DOI: 10.1609/icwsm.v14i1.7320

Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., & Rocha, A. (2017). Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, *230*, 279–293. DOI: 10.1016/j.neucom.2016.12.017

Platzer, C., Stuetz, M., & Lindorfer, M. (2014, June). Skin sheriff: A machine learning solution for detecting explicit images. In *Proceedings of the 2nd international workshop on security and forensics in communication systems* (pp. 45-56). DOI: 10.1145/2598918.2598920

Singh, S., Kaushal, R., Buduru, A. B., & Kumaraguru, P. (2019, April). KidsGUARD: Fine grained approach for child unsafe video representation and detection. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing* (pp. 2104-2111). DOI: 10.1145/3297280.3297487

Song, K., & Kim, Y. S. (2020). An enhanced multimodal stacking scheme for online pornographic content detection. *Applied Sciences (Basel, Switzerland)*, *10*(8), 2943. DOI: 10.3390/app10082943

Suh, H., Kim, J., So, J., & Jung, J. (2022). A core region captioning framework for automatic video understanding in story video contents. *International Journal of Engineering Business Management*, *14*, 18479790221078130. DOI: 10.1177/18479790221078130

Sun, L., Wang, P., Liu, P., & Nie, Z. (2023). Image processing method of a visual communication system based on convolutional neural network. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *19*(1), 1–19. DOI: 10.4018/IJSWIS.333063

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, ●●●, 30.

Vitorino, P., Avila, S., Perez, M., & Rocha, A. (2018). Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation*, *50*, 303–313. DOI: 10.1016/j.jvcir.2017.12.005

Wang, X., Cheng, F., Wang, S., Sun, H., Liu, G., & Zhou, C. (2018, October). Adult image classification by a local-context aware network. In *2018 25th IEEE international conference on image processing (ICIP)* (pp. 2989-2993). IEEE. DOI: 10.1109/ICIP.2018.8451366

Wehrmann, J., Simões, G. S., Barros, R. C., & Cavalcante, V. F. (2018). Adult content detection in videos with convolutional and recurrent neural networks. *Neurocomputing*, *272*, 432–438. DOI: 10.1016/j.neucom.2017.07.012

Yousaf, K., & Nawaz, T. (2022). A deep learning-based approach for inappropriate content detection and classification of YouTube videos. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 16283–16298. DOI: 10.1109/ACCESS.2022.3147519

Zheng, Z., Zhou, J., Gan, J., Luo, S., & Gao, W. (2022). Fine-grained image classification based on cross-attention network. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–12. DOI: 10.4018/IJSWIS.315747

Zhu, R., Wu, X., Zhu, B., & Song, L. (2018, April). Application of pornographic images recognition based on depth learning. In *Proceedings of the 1st international conference on information science and systems* (pp. 152-155). DOI: 10.1145/3209914.3209946

Xilin Shan is a graduate student at the School of Artificial Intelligence, South China Normal University. His research focuses on computer vision and deep learning. He is currently working on projects involving vision transformers and multi-frame analysis for video detection. He aims to contribute to the field by developing advanced automated detection systems and enhancing the ethical use of AI technologies.

Dingju Zhu is a professor and doctoral supervisor at South China Normal University, convener of the artificial intelligence robot research team at the School of Artificial Intelligence, director of the Artificial Intelligence Robot Research Center at the School of Software, and dean of the Guangdong Institute of Artificial Intelligence Robot Education Industry. Doctor of the University of the Chinese Academy of Sciences, postdoctoral fellow of Peking University, postdoctoral fellow of the University of Macau, and visiting scholar of Texas State University in the United States. Visiting Researcher of Shenzhen Advanced Technology Research Institute, Chinese Academy of Sciences, Senior Member of China Computer Society. He was awarded the titles of Famous Teacher of South China Normal University, Exemplary Individual of South China Normal University, Outstanding Inventor of Guangdong Province, and Local Talent of Shenzhen.

Chao Wu is a graduate student at the School of Artificial Intelligence, South China Normal University. His research focuses on computer vision and deep learning. He is currently working on projects involving vision transformers and multi-frame analysis for video detection. He aims to contribute to the field by developing advanced automated detection systems and enhancing the ethical use of AI technologies.

Kai-Leung Yung (kl.yung@polyu.edu.hk) is an associate head and chair professor in the Department of Industrial and Systems Engineering of The Hong Kong Polytechnic University. He received his BSc in electronic engineering at Brighton University, in 1975, MSc, DIC in automatic control systems at the Imperial College of Science, Technology, and Medicine, University of London, in 1976, and PhD in microprocessor applications in process control at Plymouth University, in 1985, in the United Kingdom and became a chartered engineer in 1982. After graduation, he worked in the United Kingdom for companies such as BOC Advanced Welding Co. Ltd., the British Ever Ready Group, and the Cranfield Unit for Precision Engineering (CUPE). In 1986, he returned to Hong Kong to join the Hong Kong Productivity Council as consultant and subsequently switched to academia to join the Department of Industrial and Systems Engineering of The Hong Kong Polytechnic University.

Wai-Hung Ip (wh.ip@polyu.edu.hk) has more than 30 years of experience in teaching, research, industry, and consulting. He received his PhD from Loughborough University in the United Kingdom, an MSc in industrial engineering from Cranfield University, and an LLB (Hons) from the University of Wolverhampton. After finishing postgraduate degrees in industrial engineering in the U.K., his engineering research career began at that time with novel work in industrial engineering, information, and sensor systems.