Multi-modal Video Forgery Detection via Improved Efficient-Net With Attention and Transformer Fusion

Zheng Ji[1], Luhao Cao[2*]
[1]School of Literature and Media, Chengdu Jincheng College, Chengdu 611730, China
[2]School of Digital Media Engineering and Humanities, Hunan University of Technology and Business, Changsha 410205, China
E-mail: Zheng Ji: jimmy65@yeah.net; Luhao Cao[*]: Hudsoncao@126.com
[*]Corresponding author

Keywords: video forgery, multi-modal features, efficient net, feature fusion, SSIM coefficient

Received: April 8, 2025

With the continuous advancement of deep learning technology, video forgery technology brings serious negative social impacts. However, existing video forgery detection technologies suffer from low detection accuracy, poor feature extraction capabilities, and insufficient robustness. Therefore, the study proposes two video forgery detection models based on Improved Efficient-Net and multi-modal feature fusion. The Improved Efficient-Net model utilizes structural similarity coefficients to enhance the video images and introduces a hybrid attention module in the Efficient-Net. The multi-modal feature fusion model uses the red, green, and blue domains of the image, the frequency domain, and the optical flow field features for fusion, and uses a hybrid loss function to weight all the loss function errors. The experiment shows that the maximum recognition accuracy of the improved Efficient-Net in the FaceForensics++ dataset is 98.57%, which is 6.24% as well as 9.53% higher than the baseline Efficient-Net and Convolutional Visual Transformer models, respectively. In the FaceForensics++ dataset, the multi-modal feature fusion model is able to achieve a recognition accuracy of 99.26%. In the BioDeepAV dataset, the multi-modal feature fusion model has a maximum decrease in recognition accuracy of 20.57%, which is 2.81% less than the benchmark Efficient-Net model, and the recognition accuracy is still the highest among all models. Therefore, the improved model can validly improve the accuracy of forged video identification, improve the efficiency of Internet supervision, and reduce the social harm of video forgery.

Povzetek: Predstavljen je izboljšan EfficientNet z SSIM-okrepljeno predobdelavo in hibridno pozornostjo ter večmodalno fuzijo (RGB/frekvenčno/optical flow) za ugotavljanje ponarejenih video posnetkov.

1 Introduction

The continuous development of computer technology, especially the full promotion of smartphones, has led to a continuous growth in the number of global Internet users, which is expected to reach 5.5 billion by the end of 2024 [1]. The number of Internet users in China has increased from 989 million in 2020 to 1.079 billion in 2023, and the Internet penetration rate has reached 77.5% [2]. Owing to the relatively brief duration of some individuals' internet usage, they frequently exhibit a deficiency in discernment capabilities, thereby rendering them particularly susceptible to exploitation by malefactors [3]. Unruly elements fake videos to impersonate celebrities or their family members to commit fraud [4]. Unlawful elements may also replace faces in different images by forging face videos, wantonly violating other people's privacy and portrait rights, and making false videos to spread online rumors [5]. With the development of deep learning technology, the technology of video forgery has become more mature, but the existing video forgery detection algorithms Convolutional Neural Network (CNN), Generative Adversarial Networks (GAN), and Diffusion model all have feature extraction capability. Models all suffer from poor feature extraction capability and lack of robustness [6].

To address the problem of fake face detection, Kiruthika et al. proposed a new feature detection method in order to improve the effectiveness of face image forgery detection. The method utilized the discriminative information hidden in the frequency domain of image quality assessment to extract image quality features from both the frequency and spatial domains. Experiments showed that the method was able to achieve 99% accuracy in forged image recognition in different standard datasets and was highly generalizable [7]. Xue et al. proposed a new global-local facial fusion network in order to reduce the dependence of existing methods on image artifacts and generated traces. This network utilized local physiological features and global perceptual features to detect forged traces locally, and employed residual connectivity globally to distinguish between true and false images. Experiments showed that this network had higher robustness and generalization than single class detection methods [8]. Wang et al. proposed a new dual-stream CNN framework for eliminating the threat of forged images. The framework randomly erased sample data at the preprocessing node of the three primary color image

streams, focused on the fingerprint difference of the image, and constructed the feature image in the optical response

Table 1 Summary of relevant information of relevant studies

Literature	Research topic	Backbone model	Datasets used	Accuracy	Limitations
Literature [7]	Face image forgery detection	Feature Detection Method Based on Image Quality Assessment	CelebA-DF	99%	Poor detection of specific forgeries
Literature [8]	Image artifacts and generating traces detection	Global-Local Facial Fusion Network	FaceForensics++	94.27%	Risk of information loss or overfitting
Literature [9]	Fake Image Detection	Dual-stream CNN framework	In-the-Wild	98.07%	High impact of environmental factors
Literature [10]	Fake Face Video Detection	CNN	MSU MFSD	95.04%	Sensitive to light conditions
Literature [11]	Determining the authenticity of face images	Multi-channel CNN	ProGAN generated fake face dataset	92.46%	Insufficient generalization ability to unknown attack patterns
Literature [12]	Deep Fake Video Detection	Capsule Network	FFHQ	95.82%	High computational complexity
Literature [13]	Analyzing facial motion inconsistencies	CNN	DeepFake Detection Challenge	87.04%	Insufficient adaptability on new datasets
Literature [14]	Fake Face Detection Performance Evaluation	EfficientNet	FFHQ	96.72%	Lack of comprehensive evaluation of multiple forgery methods
This text	Video forgery detection	Efficient-Net	FaceForensics++ and BioDeepAV	99.26%	/

non-uniform stream, which directed the network to focus on the image pixel value changes. Experiments showed that the framework achieved excellent accuracy and generalization across multiple datasets [9]. Alkishri et al. proposed a new deep learning detection method in order to improve the accuracy of detection of fake face videos. The method used CNN to identify the real and fake images by recognizing the differences in image color features, and used MSU-MFSD dataset to explore the color texture and extract the facial features in different color channels. Experiments showed that the method was effective in recognizing fake face videos on social platforms [10]. Li et al. proposed a new generation method to reduce the social hazards of face generation techniques. The proposed approach employed a solitary classification model to ascertain the authenticity of facial images. In parallel, it leveraged a suite of filter-based enhancement techniques for data augmentation, and further utilized an optimized multi-channel CNN as the core network architecture. Experiments showed that the method improved cross domain detection efficiency while maintaining source domain accuracy [11]. Arunkumar et al. proposed a new deep learning face forgery detection method for the detection of deep forgery videos. The method utilized fuzzy Fisher face and capsule biplot to detect different types of forged images using datasets such as FFHQ. Experiments showed that the forged image recognition accuracy of the existing and proposed systems were 81.5%, 89.32%, 91.35%, and 95.82%, respectively, which could effectively improve the recognition accuracy [12]. Altaei et al. proposed a detection method based on CNN in order to effectively control the hazards of the forged face videos. The method first converted the image to YCbCr color space and then input gamma correction. Edge detection was extracted by inputting Canny filter and CNN was used as a classifier. Experiments showed that the method could obtain high accuracy in forged video recognition [13].

In summary, existing methods have explored detection methods for fake facial videos from multiple perspectives and have achieved certain research results. However, there are still problems such as insufficient detection accuracy, poor feature extraction ability, and insufficient robustness. Therefore, a video forgery detection model based on multi-modal features and Efficient-Net is proposed, which innovatively enhances video images using structural similarity coefficients, introduces a mixed attention module, and fuses the Red-Green-Blue (RGB), frequency, and optical flow field features of the images. The mixed loss function is used to weight all loss function errors. The research aims to (1) further improve the detection accuracy of forged face videos by introducing a hybrid attention module in Efficient-Net, (2) improve the generalization and robustness of the model through multi-modal feature fusion, and (3) in the validation of multiple datasets, all of the proposed methods of the research are able to effectively identify the forged videos and reduce the social hazards of the face-forgery videos. Based on the above related studies, Table 1 summarizes the research

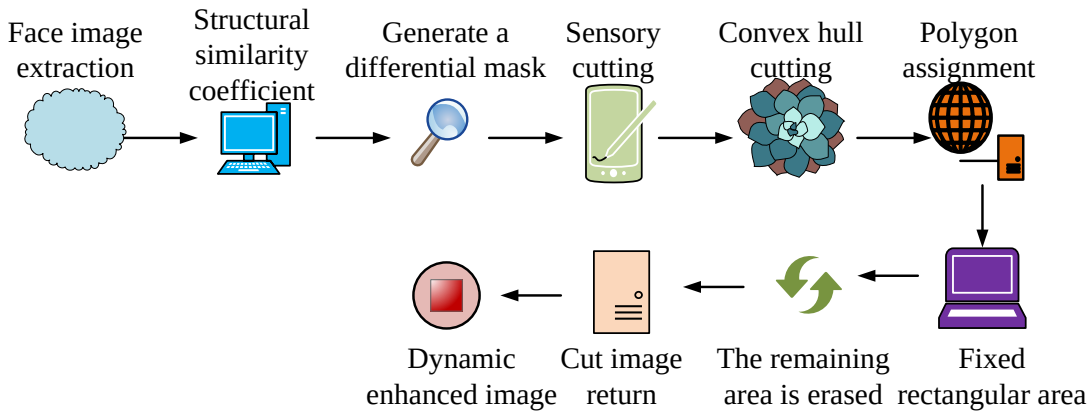

Figure 1: Specific operation flow of face image enhancement algorithm.

topics, backbone models, datasets used, accuracy and limitations of the related studies.

In Table 1, the existing methods explore the detection methods of forged face images or videos from several aspects and achieve better detection accuracy. However, the detection effect of unknown forgery methods is not satisfactory and other problems. Meanwhile, the high computational complexity of some methods affects the real-time of forged video detection. Consequently, this research innovatively employs structural similarity coefficients to augment the quality of video images. Moreover, it introduces a hybrid attention module, which integrates the red, green, and blue channels of the image, along with features from the frequency domain and the optical flow field. Additionally, all the errors in the loss functions are weighted using a hybrid loss function. The proposed method of the study can effectively improve the detection accuracy of localized face videos.

2 Methods and materials

2.1 Video falsification monitoring based on Efficient-Net

When conducting video forgery monitoring, it is necessary to extract the features of the input image, among which facial image features are the most important. Through the subtle judgment of facial images, it is possible to effectively authenticate the authenticity of videos, while reducing the complexity of computations involved in the model and improving the calculation speed [14]. The research needs to first perform frame segmentation on the video, locate the facial images in the video, and then extract relevant feature data. The study proposes a data preprocessing algorithm for face image enhancement to lay the foundation for subsequent forgery video detection. The study first uses Structural Similarity Index (SSIM) to determine the similarity between the images and then compares the quality of the images before and after compression, the SSIM coefficients are calculated as shown in equation (1) [15].

$$SSIM = \frac{(2\alpha_1\alpha_2 + K_1)(2\beta_1\beta_2 + K_2)}{(\alpha_1^2 + \alpha_2^2 + K_1)(\beta_1^2\beta_2^2 + K_2)} \quad (1)$$

In equation (1), $SSIM$ represents the SSIM, α_1 means the average value of the grayscale image before compression, α_2 means the average value of the grayscale image after compression, β_1 means the variance of the grayscale image before compression, β_2 means the variance of the grayscale image after compression, K_1 and K_2 both represent constants. The facial image enhancement algorithm enhances the image data by continuously generating occlusions on the facial image. The detailed procedural steps of the facial image enhancement algorithm is in Figure 1.

In Figure 1, the facial images are extracted from real and fake videos, the SSIM between the two images is calculated, and a differential mask is generated through the coefficient. The algorithm performs sensory segmentation on facial features, performs convex hull segmentation on facial contours, erases remaining areas of the image, and randomly erases the image using fixed area rectangular regions. When performing convex hull cutting on facial contours, multiple points are randomly selected from the contour keypoints, and each point is connected to form an irregular polygon. The area calculation of the polygon is shown in equation (2) [16].

$$S = 0.5\left|\sum_{i=1}^{n-1} x_i y_{i+1} + x_n y_1 - \sum_{i=1}^{n-1} x_{i+1} y_i - x_1 y_n\right| \quad (2)$$

In equation (2), S mean the area of the irregular polygon, n mean the total number of randomly selected points, x_i represents the x axis coordinates of the i th point, and y_i mean the y axis coordinates of the i th point. After obtaining the processed dataset for facial image enhancement, the study improved the Efficient-Net and constructed a related video forgery detection model. The specific structure of the video forgery detection model is shown in Figure 2.

In Figure 2, the model is divided into three modules: the preprocessing module using facial image enhancement algorithm, the Efficient-Net module, and the final

classification module. The data processed by the preprocessing module is re-sized and input into the

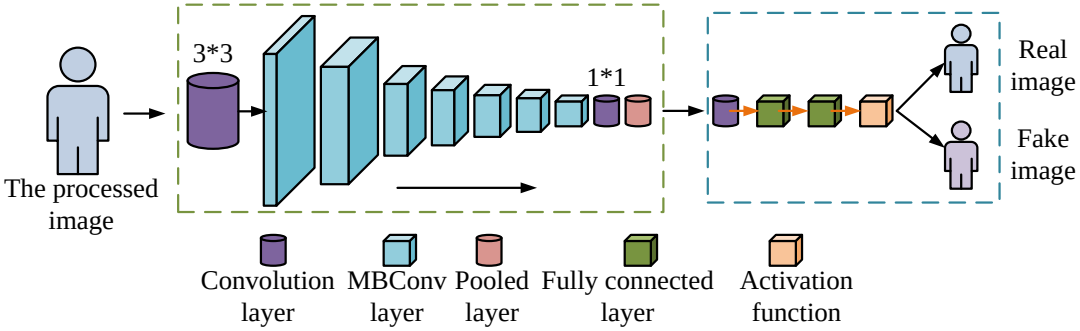

Figure 2: Specific structure of video forgery detection model for improved Efficient-Net.

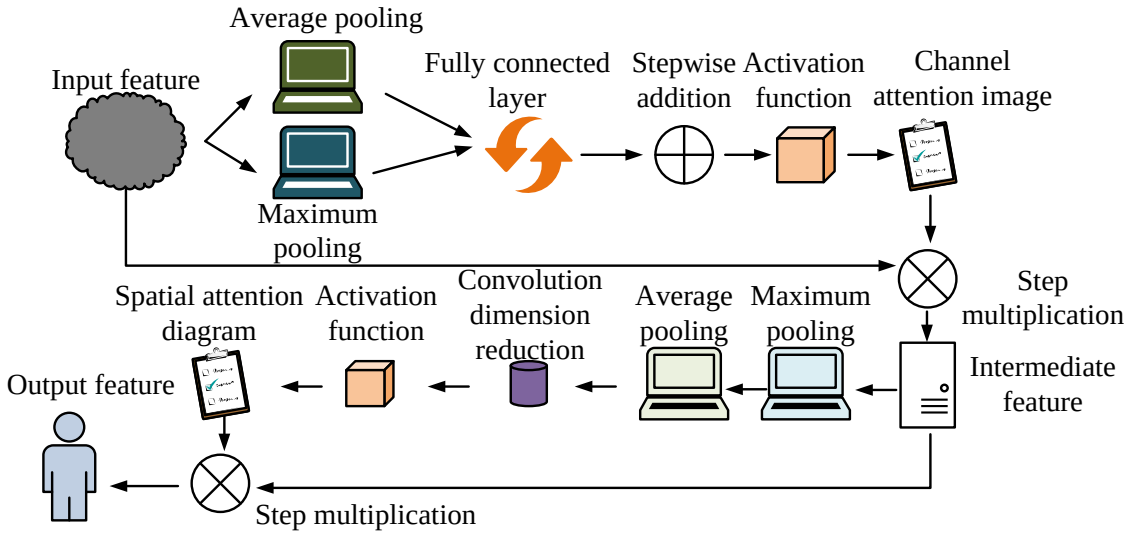

Figure 3: Particular configuration of the hybrid attention module.

Efficient-Net module, which includes two layers of ordinary convolution and seven layers of MBConv. Finally, it is connected to the classification module through a fully connected layer (FCL). MBConv cannot quickly complete calculations when the image size is large, and the deep convolution speed is slow in the early layers. Therefore, the study introduces a hybrid attention module in the MBConv layer, using channel attention and spatial attention to calculate the weights of all input image information, and then applies the learned weights to the initial image. The particular configuration of the hybrid attention module is in Figure 3.

In Figure 3, the hybrid module incorporates MBConv located in the first three layers, with channel attention at the front and spatial attention at the back, connected in series. The study starts with channel attention processing of the input feature maps, which is able to learn the importance of each channel and re-calibrate the channels of the feature maps, which is able to efficiently compress the dimensions of the feature maps and reduce the amount of computation while maintaining the spatial information. After the channel attention, the spatial attention module further processes the feature map in the spatial dimension. At this time, the number of channels of the feature map has been filtered and adjusted by the channel attention, and the more important channel features are retained. Performing the channel attention calculation first can provide a more instructive feature map for the subsequent spatial attention module. The feature images processed by convolution are subjected to max pooling and average pooling operations, and the spatial information of the images is combined through an FCL. The generated 1*1 convolutional image is weighted and an activation function is used to generate the final channel attention image. In the spatial attention module, the channel attention output features are subjected to max pooling and average pooling operations respectively. The obtained two types of communication are concatenated into channels, and the resulting image is convolved to compress the channel size to 1. The final feature map is obtained through an activation function. The calculation of channel attention is shown in equation (3) [17].

$$Ca = \sigma\left[w_2 \left(w_1 F_{avg} \right) + w_1 \left(w_2 F_{\max} \right) \right] \quad (3)$$

In equation (3), Ca represents channel attention calculation, σ represents Sigmoid activation function,

w_1 represents the average pooling image stitching weight in the FCL, w_2 represents the maximum pooling image

Table 2: Architectural changes between the improved model and the baseline Efficient-Net model

Parameter	EfficientNet-B0	EfficientNet-B4	Parameter	EfficientNet-B0	EfficientNet-B4
Input Resolution	224×224	380×380	Dropout rate	0.2	0.4
Width factor	1.0	1.4	Drop connect rate	0.2	0.2
Depth factor	1.0	1.8	Hybrid Attention Layer	0	3
Number of MBConv modules	16	7	Channel Attention	0	3
Number of parameters	5.3M	19M	Spatial attention	0	3
FLOPs	0.39B	4.2B	/	/	/

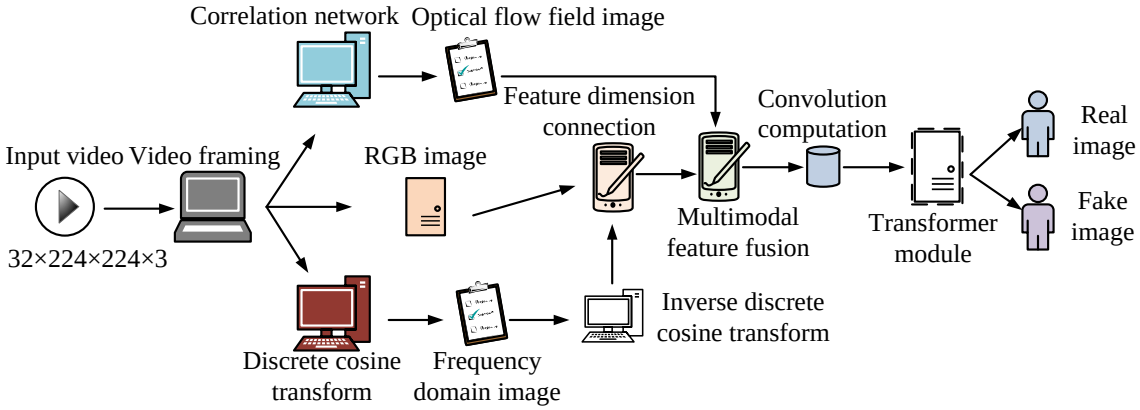

Figure 4: Specific structure of video forgery detection model for multimodal feature fusion.

stitching weight in the FCL, $F_{\max}$ represents the average pooling operation, and F_{avg} means the maximum pooling operation. The spatial attention calculation is shown in equation (4).

$$Sa = \sigma\left[f^{7\times7}\left(F_{avg}, F_{\max}\right)\right] \quad (4)$$

In equation (4), Sa represents spatial attention calculation, $f^{7\times7}$ represents convolution calculation, and the size of the convolution kernel is 7×7. The final generated spatial attention map is element wise multiplied with the channel attention output features to obtain a new output feature map. Video forgery detection can be viewed as a straightforward binary classification problem aimed at distinguishing authenticity. The study refines the output of the Efficient-Net by integrating it into a binary classification layer. The classification component aggregates the final output data from the Efficient-Net module to generate the output feature map. This feature map is then compressed to 108 dimensions using two FCLs. Subsequently, the Softmax activation function layer is employed to achieve the final binary classification of authenticity. The model's loss function utilizes the cross-entropy loss function, computed as illustrated in equation (5).

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[X_i \log\left(p_i\right) + \left(1-X_i\right)\log\left(1-p_i\right)\right] \quad (5)$$

In equation (5), L means the cross entropy loss function, N means the total number of facial images in the input video, X_i means the i th facial sample, p_i means the probability of the i th facial sample being predicted to be true. In the binary classification module, if the video is true, the output is 0, and if the video is false, the output is 1. The Efficient-Net model variant used in the study is Efficient-Net-B4, and the architectural changes between the improved model and the baseline Efficient-Net model are shown in Table 2.

In Table 2, the Improved Efficient-Net model has been extended and enhanced in several ways, with higher input resolution and the ability to capture more image detail. The increase in the width and depth coefficients of the improved model makes the model more expressive and able to learn more complex features. Also the increased number of attention layers and coefficients demands more computational resources.

2.2 Video forgery detection based on multi-modal feature fusion

When conducting video forgery detection, in addition to focusing on the characteristics of the facial image, it is also necessary to consider the temporal relationship before and after the video. However, the improved Efficient-Net's forgery facial video monitoring model does not consider the issue of video timing, making it more suitable for image monitoring. Therefore, the study uses multi-modal feature fusion for video forgery detection. The images used in the multi-modal feature fusion model are also processed by face image enhancement algorithm. The

study fuses the RGB domain, frequency domain, and optical flow field features of the processed images and finally classifies them through the Transformer model. The specific flow

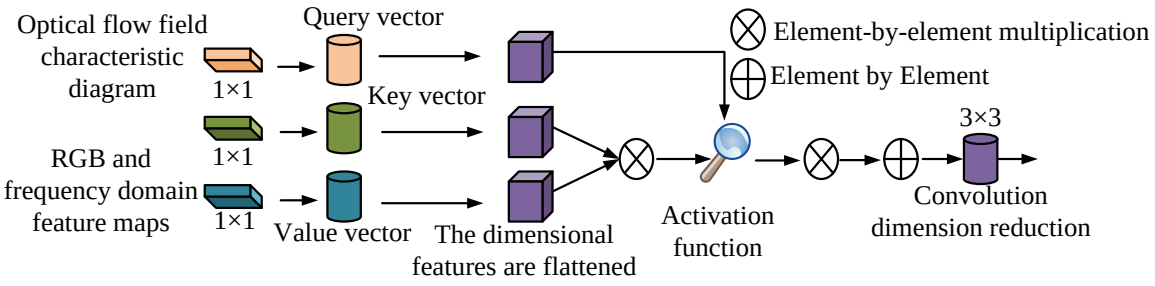

Figure 5: Specific structure of multi-modal feature fusion module.

of the video forgery detection model using multi-modal feature fusion technique is shown in Figure 4.

In Figure 4, the facial information extracted from the video is input into the model. Pyramid cascade and correlation network are used to extract the optical flow field between different frames of the video, and discrete cosine transform is used to obtain frequency domain information. The discrete cosine transform is used in video processing for its energy concentration, real arithmetic efficiency, boundary processing friendliness and sensitivity to local features. The fast Fourier transform is more suitable for scenarios that require global frequency analysis, such as communication signal processing. The discrete cosine transform in video forgery detection can capture the tampering traces of inter-frame flickering more accurately. In this study, the pre-enhanced Efficient-Net-B4 network is utilized to independently extract features from the frequency-domain map, the RGB map, and the optical-flow field image. Subsequently, the features extracted from the frequency-domain map and the RGB map are concatenated along designated dimensions. Following this, the concatenated feature maps undergo multi-modal feature fusion with the optical-flow field feature maps. The model fuses the frequency domain map and RGB map first because both are based on the spatial information of a single frame, and the fusion can directly enhance the spatial sensitivity of the model to the forged region in a single frame image. The optical flow field feature map is integrated at a later stage due to its association with the temporal dimension. This sequential fusion approach, where spatial features are amalgamated first followed by the incorporation of temporal dynamic features, aligns with the detection logic of "first pinpointing spatial anomalies and subsequently validating the rationality of the temporal sequence". The processed image is convolved and input into the Transformer model for binary classification. The Efficient-Net model is better at extracting local features from high-resolution images and has much lower computational complexity than the Transformer, which processes raw pixels directly, enables the Transformer to focus on learning global dependencies across modalities and regions, and avoids redundant low-level feature computation. When extracting frequency domain features of an image, a filter can be used to divide it into three frequency bands, as shown in equation (6) [18].

$$F_i = D(X) \square f_{base}^i, i = 1,2,3 \quad (6)$$

In equation (6), F_i represents the extracted frequency domain features, D represents the discrete cosine transform, X represents the input image, and f_{base}^i represents the filter processing. After feature extraction is completed, the image is converted back to the spatial domain through the application of an inverse discrete cosine transform, as shown in equation (7) [19].

$$F_i' = D^{-1}(F_i), i = 1,2,3 \quad (7)$$

In equation (7), $F_1^{'}$ represents spatial domain image features and D^{-1} represents inverse discrete cosine transform. The study uses channel aggregation to form frequency domain feature maps of three frequency bands, and the transformed images have the same size. In multi-modal feature fusion, the study utilizes the Query, Key, Value (QKV) mechanism in the attention layer. Firstly, a convolutional kernel is used to embed two types of feature maps into the QKV space. The spatial size of the feature maps remains unchanged, and then they are converted into two-dimensional features. The output value of the multi-modal fusion module is calculated, as shown in equation (8).

$$F_{fusion} = soft\left(\frac{QK^T}{\sqrt{0.25H \cdot 0.25W \cdot C}}\right)V \quad (8)$$

In equation (8), F_{fusion} means the output of the fusion module, *soft* means the activation function, Q stands for the Query vector, K for the Key vector, T for the matrix transpose, V for the Value vector, H for the image's height, W for its width, and C for the number of channels. The specific structure of the multi-modal feature fusion module is shown in Figure 5.

In Figure 5, two types of feature maps are convolved by 1*1 and embedded into the QKV space, flattening the multidimensional features into two dimensions. The two types of data are concatenated together by level multiplication, activated by the Softmax function, and then multiplied with the third type of data to obtain the output value of multi-modal feature fusion. The data from the fusion module undergoes convolutional dimensionality reduction processing before being input into the Transformer module. The specific structure of the Transformer module is in Figure 6.

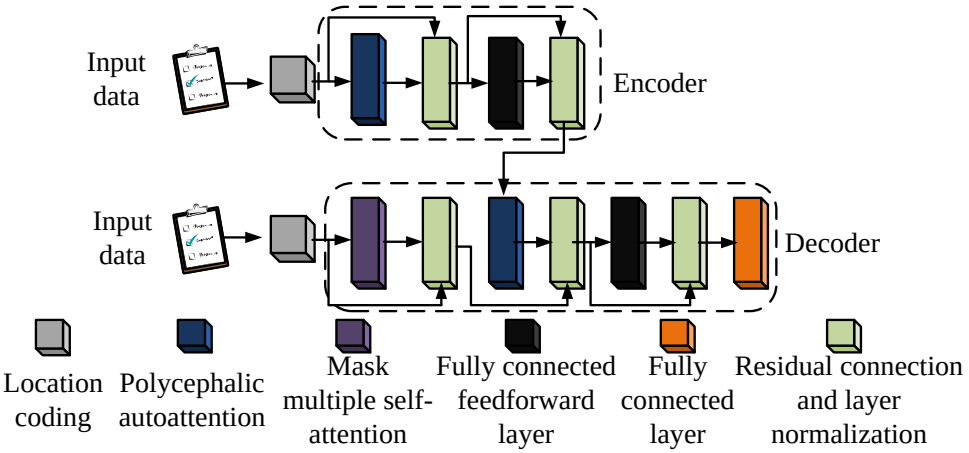

Figure 6: Specific structure of Transformer module.

In Figure 6, the Transformer module is categorized into two parts: the encoder and the decoder. The input data is position encoded and then input into the structure of both parts. The encoder consists of multiple multi-head attention layers, fully connected feed-forward layers, residual connections, and layer normalization layers. The encoder data is connected to the decoder after the first residual connection, and the final data is output after passing through the FCL. The study uses Transformer with a header of 8, a layer of 6, an attention dimension of 512, a filtering parameter set to 0.1, a patch size of 7×7, and a sequence length of 64. The calculation of the attention layer is in equation (9) [20].

$$Attention = soft\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

In equation (9), $Attention$ represents the attention score and d_k means the dimension of the Query vector. The output calculation of the fully connected feed-forward layer is in equation (10).

$$FFN(X) = \sigma(XW_1 + b_1)W_2 + b_2 \tag{10}$$

In equation (10), $FFN(X)$ means the output of the feed-forward FCL, σ denotes the activation function, X signifies the input matrix, W_1 stands for the first layer's weight matrix, b_1 represents the first layer's bias vector, W_2 indicates the second layer's weight matrix, and b_2 signifies the second layer's bias vector. The video forgery detection model based on multi-feature fusion uses a mixed loss function to weight the errors of all loss functions. In binary classification tasks, the probability of the Softmax activation function outputting a true video is calculated as shown in equation (11).

$$p_1 = \frac{\exp\left(W_1^T x + b_1\right)}{\sum_{i=1}^{2}\exp\left(W_i^T x + b_i\right)} \tag{11}$$

In equation (11), p_1 means the probability of the video being true, x represents the input features, W_i means the weight matrix of the i th layer, b_i means the bias vector of the i th layer, and the probability of the output video being fake is calculated as shown in equation (12).

$$p_2 = \frac{\exp\left(W_2^T x + b_2\right)}{\sum_{i=1}^{2}\exp\left(W_i^T x + b_i\right)} \tag{12}$$

In equation (12), p_2 represents the probability that the video is fake. The model introduces a face recognition loss function when learning facial boundary features. This loss function improves the model's capability to distinguish similar faces through additive angle boundary loss, and is more sensitive to the surrounding environment's lighting intensity and facial expressions. The calculation is shown in equation (13).

$$L_{face} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s\left[\cos\left(\theta_{y_i}\right)+m\right]}}{e^{s\left[\cos\left(\theta_{y_i}\right)+m\right]}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}\right) \tag{13}$$

In equation (13), L_{face} represents the face recognition loss function, N means the total number of samples, s represents the scaling factor, m represents the boundary angle, used to increase the distance between categories. θ_{y_i} means the angle between the feature vector x_i of the sample i and the weight vector w_{y_i} of the corresponding category y_i. θ_j represents the angle between the feature vector x_i of the sample i and the weight vector w_j of the corresponding category j. The study uses a single center loss function to control the discreteness between real and fake portraits, as calculated in equation (14).

$$L_s = S_t + \max\left(S_t - S_c + m\sqrt{D}, 0\right) \tag{14}$$

In equation (14), L_s represents the single center loss function, S_t represents the Euclidean distance between the center point and the real image, S_c represents the Euclidean distance between the center point and the fake image, m represents hyperparameters, and D represents feature dimensions. The total loss function is calculated as represented in equation (15).

$$L_z = \mu_1 L_{face} + \mu_2 L_s + L_{soft} \tag{15}$$

```
# Simplified Pseudocode

# Multimodal Fusion
def fuse(rgb, freq, flow):
feats = [extract(x) for x in [rgb, freq, flow]]
return attend(*feats)

# Compact Transformer
def classify(x):
x = embed(x) + pos_enc()  # Combined embedding
x = [transformer(x) for _ in range(2)][-1]  # 2 layers
return softmax(head(x))

# Optimized Training
cfg = dict(lr=1e-4, bs=128, epochs=30)
model = compose(fuse, classify)
opt = Adam(cfg['lr'])

for _ in range(cfg['epochs']):
for batch in data_loader(cfg['bs']):
loss = cross_entropy(model(batch), targets)
opt.step(loss)
test(model)
```

Figure 7: Pseudo-code of the model proposed by the study.

In equation (15), L_z means the total loss function, μ_1 means the weight coefficients of the face recognition loss function, μ_2 represents the weight coefficients of the single center loss function, and L_{soft} represents the Softmax loss function. The pseudo-code of the proposed model is shown in Figure 7.

3 Results

3.1 Experimental analysis of fake face video monitoring based on efficient-Net

The experiment selected the FaceForensics++ dataset created by researchers from Germany and Italy, which contains 1000 relevant initial facial video data and derived fake videos, with quality ranging from low to high. The initial learning rate of the improved model was 1e-4, the termination learning rate was 1e-8, the learning rate decayed to 1/8 of the original after every 5 epochs, the batch_size was set to 128, and the epoch was set to 30. The Adam optimizer was used, and the parameters of the optimizer were set to 0.9 and 0.999, respectively. The study divided the FaceForensics++ dataset into a training set and a testing set in an 8:2 ratio. The amount of video in the training set was 800 segments, and the amount of video in the test set was 200 segments, both with a frame rate of 30 fps The frame rate of video in the subsequent dataset used was also 30 fps. Upon the completion of data segmentation, the video underwent a series of data-augmentation operations, including random cropping, color dithering, illumination adjustments, and random rotation. The comparison algorithms for the experiments included the regular Efficient-Net as well as the Convolutional Vision Transformer (CViT). The CViT model utilized a hybrid architecture consisting of a backbone network, Efficient-Net, a mid-layer Transformer module, and a multi-scale feature fusion module. The attention mechanism used local self-attention and cross-modal attention. The training configuration was the same as in the study of the proposed method. The CViT model could efficiently perform multi-modal fusion and realize lightweight deployment. It could effectively accomplish real-time detection of forged videos and was suitable for resource-constrained environments. The comparison of the loss function variation trends of different methods is shown in Figure 8.

In Figure 8 (a), Improved Efficiency Net had the fastest convergence speed and tended to converge after 22 iterations. The minimum loss function value was 0.47, which was 2.86 and 0.92 lower than Efficiency Net and CViT, respectively. In Figure 8 (b), the convergence position of the Improved Efficiency Net remained basically unchanged, and the minimum loss function values of the three models increased. However, the Improved Efficiency Net had the smallest increase, with values 1.59 and 1.03 lower than the other two models. The comparison of the accuracy of counterfeit video recognition using different methods is shown in Figure 9.

In Figure 9 (a), the maximum accuracy of the Improved Efficiency Net model was 98.57%, which was 6.24% and 9.53% higher than the Efficiency Net and CViT models, respectively. In Figure 9 (b), in the test set,

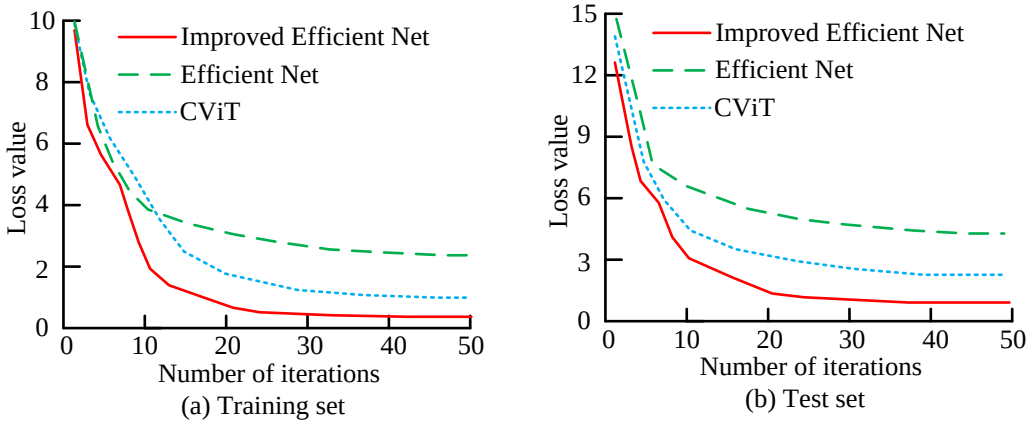

Figure 8: Comparison of the trend of the loss function for different methods.

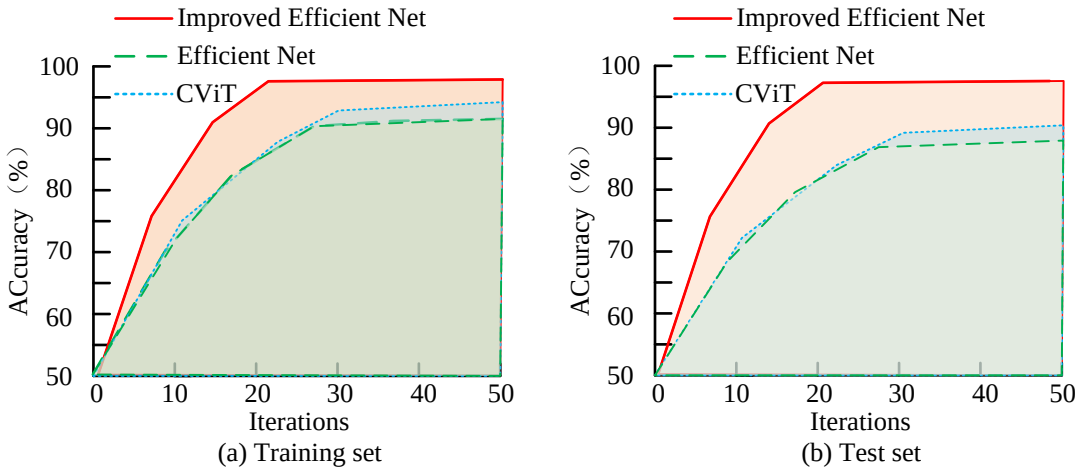

Figure 9: Comparison of forged video accuracy of different methods.

Table 3: Performance comparison of adding hybrid attention modules at different locations in Efficient-Nets

Add location	Accuracy (%)	Recall (%)	F1	AUC (%)
1 layer	95.14	93.25	0.92	97.18
1-2 layer	96.85	95.37	0.95	98.54
1-3 layer	98.57	98.14	0.98	98.32
1-4 layer	98.26	97.56	0.97	98.15
1-5 layer	95.17	94.28	0.94	95.35
1-6 layer	93.08	93.11	0.92	92.07
1-7 layer	92.65	93.05	0.91	91.86

the accuracy of all three models decreased, and the maximum accuracy of the Improved Efficiency Net model was 7.64% and 11.29% higher than the other two models, respectively. Mixed attention modules were added to the seven layers of MBConv in the Efficient-Net for ablation experiments, and the performance changes of the model were observed as represented in Table 3.

In Table 3, when a mixed attention module was added to the first three layers of MBConv in the Efficient-Net, the model achieved optimal accuracy, recall, F1 score, and AUC. When the mixed attention module was added in the fourth layer, the performance of the model decreased slightly. When the attention module was added in subsequent layers, the performance decreased significantly, indicating that the feature map output by the model was too small, which would affect the receptive field size obtained by subsequent modules, thereby reducing the model performance and resulting in a reduction of the accuracy of fake video recognition. Therefore, adding a mixed attention model in the first three layers of the model yielded the best results. The ablation experiments for the sequential ordering of channel and spatial attention in the hybrid attention module are shown in Table 4.

Table 4: Comparison of ablation experiments for channel and spatial attention sequencing

/	Accuracy (%)	Recall (%)	F1	AUC (%)
Channeled Attention in Front	98.57	98.14	0.98	98.32
Spatial attention comes first	90.26	89.72	0.90	92.54
Channel attention only	79.53	76.35	0.81	80.15
Spatial attention only	72.39	70.18	0.75	73.67

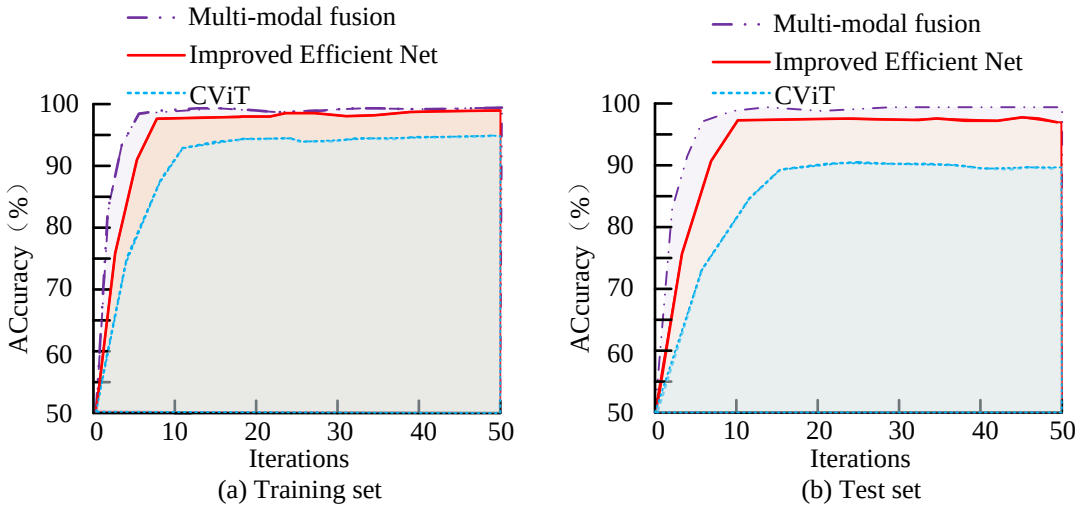

Figure 10: Comparison of accuracy of forged video recognition by different methods.

In Table 4, the model was able to achieve better performance with the sequential combination approach of channel attention in front and spatial attention in the back. The recognition accuracy of the model was 98.57%, which was 8.31%, 19.04%, and 26.18% higher than that of spatial attention in front, channel attention only, and spatial attention only, respectively. The recall, F1 value and AUC value of the model when channel attention was in the front and spatial attention was in the back achieve the optimum, which were 98.14%, 0.98 and 98.32%, respectively.

3.2 Experimental analysis of fake face video monitoring based on multi-modal feature fusion

The hardware environment and dataset of the experiment were the same as in Section 2.1. The fusion model had an initial learning rate of $1e^{-4}$, a termination learning rate of $1e^{-7}$, a learning decay factor of 0.1, a training epoch of 30, and a batch_size set to 128. The model used the Adam optimizer with the same parameter settings as in Section 2.1. The comparison algorithms were Improved Efficiency Net, CViT, and Generative Convolutional Vision Transformer (GenConViT). The GenConViT model used a composite framework of generative contrast learning + dual stream Transformer. The temperature parameter in the training configuration was 0.1 and the learning rate was set to 1e-3. The GenConViT model had high robustness, and resistance to unknown forgery attacks, and could cope with complex adversarial environments. The comparison of the accuracy of counterfeit video recognition using different methods is shown in Figure 10.

In Figure 10 (a), the video forgery detection model using multi-modal feature fusion approached convergence after about 7 iterations, with a maximum accuracy of 99.26%. The convergence speed was 2 and 5 iterations faster than Improved Efficiency Net and CViT, respectively, with maximum recognition accuracies 0.69% and 6.15% higher, respectively. In Figure 10 (b), the accuracy of all three models in the test set decreased, and the convergence position was further back. The maximum accuracy of the multi-modal feature fusion model was 98.54%, which was 1.74% and 9.27% higher than the others. The comparison of the accuracy of model forgery video recognition under different loss functions is shown in Figure 11.

In Figure 11 (a), the highest recognition accuracy of the model was achieved when the hybrid loss function was used, which tended to converge at 10 iterations with a maximum value of 99.07%, which was 6.87%, 9.94%, and 16.15% higher than when face recognition, monocentric, and cross-entropy loss functions alone were used, respectively, and the speed of convergence was also faster. In Figure 11 (b), after replacing the test data with more complex fake videos, the accuracy of the models decreased slightly. However, the model using the mixed function showed the least decrease, only 2.15%, while the other models decreased by 7.57%, 5.28%, and 6.03%, respectively. This indicated that the robustness of the model was significantly enhanced after using the

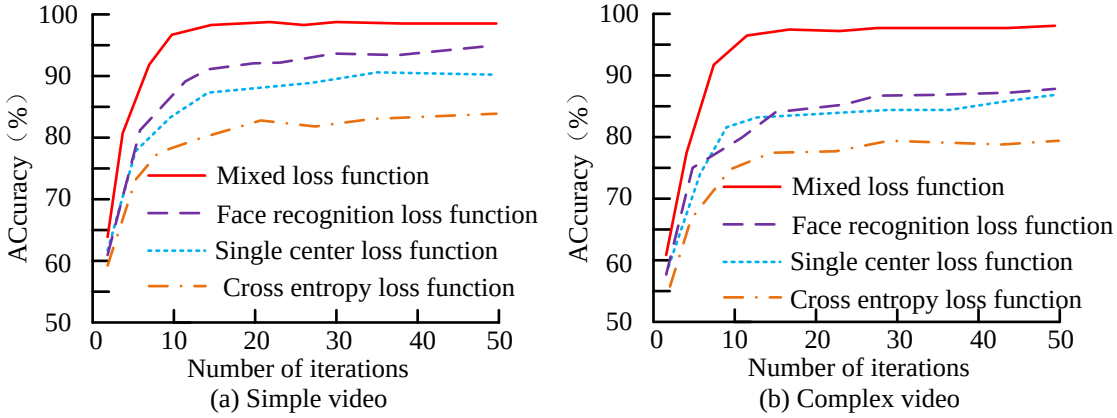

Figure 11: Accuracy of model forgery video recognition with different loss functions.

Table 5: Comparison of performance and individual video MT of different methods

Model	FaceForensics++		Deepfake Detection Challenge		BioDeepAV	
	Accuracy/%	MT/ms	Accuracy/%	MT/ms	Accuracy/%	MT/ms
Research model	99.26	182	84.15	207	78.69	225
Improved Efficient-Net	98.57	215	76.86	254	74.25	272
CViT	90.18	229	72.15	295	69.34	382
Efficient-Net	92.56	165	70.67	184	69.18	227
GenConViT	88.54	189	80.29	305	70.18	492
p	0.04	0.02	0.03	0.01	0.01	0.01
t	6.75	10.96	7.54	12.69	11.15	14.38

Table 6: Robustness test results of different methods in DFDC and DeepFakeTIMIT datasets

Methods	Resolution	DFDC			DeepFakeTIMIT		
		Accuracy/%	mAP/%	AUC/%	Accuracy/%	mAP/%	AUC/%
Research model	720p	87.26	88.62	85.36	78.29	75.83	76.54
	1080p	91.45	90.53	89.17	83.45	81.92	82.37
	2K	95.17	92.74	93.15	87.66	85.12	85.68
Efficient-Net	720p	77.92	75.64	76.25	55.36	53.61	54.16
	1080p	82.16	80.27	81.09	59.87	57.63	58.03
	2K	88.57	85.39	86.13	65.32	64.18	64.29
CViT	720p	75.34	72.09	73.85	58.31	57.06	57.62
	1080p	81.05	78.14	80.74	63.18	60.44	61.37
	2K	87.62	83.59	85.03	67.35	65.29	65.82

mixed function. To verify the generalization of the raised model, the experiment trained the model on the FaceForensics++ dataset and validated it using the Deepfake Detection Challenge dataset and BioDeepAV dataset. The performance and single video monitoring time (MT) comparison of various methods are shown in Table 5.

In Table 5, the FaceForensics++ dataset generated fake face videos mainly by Face2Face and NeuralTextures.The Deepfake Detection Challenge dataset used fake videos generated by Deepfake technique. BioDeepAV dataset used RealVisXL and LAION-Face techniques to generate fake videos. Most models were able to achieve excellent accuracy in the trained dataset. However, when tested on the new dataset, the performance of all models decreased slightly. The maximum decrease of the proposed model was 20.57%, which was 3.75%, 0.27%, 2.81%, and -2.21% lower than the improved Efficient-Net, CViT, Efficient-Net, and GenConViT, respectively. This indicated that the proposed model had higher generalization ability and could adapt to datasets of different complexity levels. The model proposed by the research had a single video MT of 182ms in the FaceForensics++ dataset, which was 32ms, 47ms, -17ms, and 7ms lower than other models, respectively. Although the running time of Efficient-Net was shorter, the accuracy varied greatly. There was statistical significance ($p<0.05$) between the data of all groups with t values ranging from a minimum of 6.75 to a maximum of 14.38. The robustness test results of different methods on Deepfake Detection Challenge (DFDC) and DeepFakeTIMIT datasets are shown in Table 6.

In Table 6, the different models achieved high recognition accuracies at higher video resolutions, and the model performance gradually decreased as the video resolution decreases. In the DFDC dataset, the recognition accuracy of the proposed model under study was 95.17 at 2K resolution, which decreased by 3.72% and 7.91% in 1080P and 720P resolutions, respectively. The maximum values of mAP and AUC of the model were 92.74% and

93.15%, respectively, which decreased by 7.52% and 7.47% in the DeepFakeTIMIT dataset. In the DFDC dataset, the decrease of the performance of the proposed model with the change of the resolution was small, and all the indexes were better than the Efficient-Net and CViT models.

4 Discussion

A video forgery detection model based on multi-modal features and Efficient-Net was proposed and applied to the actual analysis of multiple datasets. The effectiveness and superiority of the method in the detection of forged face videos were verified through relevant experimental analysis. In the FaceForensics++ dataset, the maximum recognition accuracy of the improved Efficient-Net reached 98.57% and 98.54% on the training and test sets, respectively, which was significantly better than that of the ordinary Efficient-Net (92.56%) and CViT (90.18%). Compared with literature [8] and literature [9], the improved Efficient-Net outperformed the existing SOTA methods in a single modality. This was because the addition of a hybrid attention module to the first three layers of the model was able to capture low-level features such as texture, noise, and edge anomalies, which were usually present as subtle artifacts in the generated image. The multi-modal feature fusion strategy introduced temporal dynamic features that were capable of detecting unnaturalness of facial movements in the forged video, giving it a detection accuracy of 99.26%. The maximum recognition accuracies across datasets (Deepfake Detection Challenge, BioDeepAV) were 84.15% and 78.69%. Compared with literature [12] and literature [14], the multi-modal model was more robust in complex scenes. There are some shortcomings in this study, for example, the single video monitoring time of the multi-modal model was 182ms, which was slightly higher than the 165ms of the ordinary Efficient-Net, limiting the real-time video monitoring. Subsequently, the model can be compressed using knowledge distillation to further reduce the computation time of the model and improve the monitoring real-time performance.

5 Conclusion

To address low accuracy and poor robustness in existing video forgery detection methods, this study proposed a forged face video detection model integrating multi-modal feature fusion with Efficient-Net. Experiments demonstrated that Improved Efficient-Net achieved the fastest convergence and lowest loss values (2.86 and 0.92 lower than Efficient-Net and CViT respectively), with maximum recognition accuracy reaching 98.57% (6.24% and 9.53% higher than counterparts). While all models showed accuracy declines in test sets, Improved Efficient-Net exhibited the smallest reduction. Optimal performance occurred when hybrid attention modules were embedded in the first three MBConv layers of Efficient-Net. The hybrid loss function achieved peak accuracy of 99.07%, surpassing face recognition, single-center, and Softmax losses by 6.87%, 9.94%, and 16.15% respectively. Though showing maximum 20.57% accuracy drop on unseen data, the multi-modal model maintained superior performance. This approach effectively enhanced internet regulation efficiency and mitigated social risks from forged videos.

References

[1] Maryam Taeb, and Hongmei Chi. Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy, 2(1):89-106, 2022. https://doi.org/10.3390/jcp2010007

[2] Lin Song, Jinfu Yang, Qingzhen Shang, and Mingai Li. Dense face network: A dense face detector based on global context and visual attention mechanism. Machine Intelligence Research, 19(3):247-256, 2022. https://doi.org/10.1007/s11633-022-1327-2

[3] Hanady Sabah Abdul kareem, and Mohammed Sahib Mahdi Altaei. Detection of deep fake in face images-based machine learning. Al-Salam Journal for Engineering and Technology, 2(2):1-12, 2023. https://doi.org/10.55145/ajest.2023.02.02.001

[4] Nency Bansal, Turki Aljrees, Dhirendra Prasad Yadav, Kamred Udham Singh, Ankit Kumar, Gyanendra Kumar Verma, and Teekam Singh. Real-time advanced computational intelligence for deep fake video detection. Applied Sciences, 13(5):3095-3104, 2023. https://doi.org/10.3390/app13053095

[5] Shilpa Sharma, Linesh Raja, Vaibhav Bhatnagar, Divya Sharma, Swami Nisha Bhagirath, and Ramesh Chandra Poonia. Hybrid HOG-SVM encrypted face detection and recognition model. Journal of Discrete Mathematical Sciences and Cryptography, 25(1):205-218, 2022. https://doi.org/10.1080/09720529.2021.2014141

[6] Chenyu Liu, Jia Li, Junxian Duan, and Huaibo Huang. Video forgery detection using spatio-temporal dual transformer. In Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition, 16(3):273-281, 2022. https://doi.org/10.1145/3581807.3581847

[7] Kiruthika S, and Masilamani V. Image quality assessment based fake face detection. Multimedia Tools and Applications, 82(6):8691-8708, 2023. https://doi.org/10.1007/s11042-021-11493-9

[8] Ziyu Xue, Xiuhua Jiang, Qingtong Liu, and Zhaoshan Wei. Global-local facial fusion-based GAN generated fake face detection. Sensors, 23(2):616-637, 2023. https://doi.org/10.3390/s23020616

[9] Jinwei Wang, Kehui Zeng, Bin Ma, Xiangyang Luo, Qilin Yin, Guangjie Liu, and J Sunil Kr. Jha. GAN-generated fake face detection via two-stream CNN with PRNU in the wild. Multimedia Tools and Applications, 81(29):42527-42545, 2022. https://doi.org/10.1007/s11042-021-11592-7

[10] Wasin Alkishri, Setyawan Widyarto, Jabar H. Yousif, and Mahmood Al-Bahri. Fake face detection based on colour textual analysis using deep convolutional neural network. Journal of Internet

Services and Information Security, 13(3):143-155, 2023. https://doi.org/10.58346/JISIS.2023.I3.009

[11] Shengyin Li, Vibekananda Dutta, Xin He, and Takafumi Matsumaru. Deep learning based one-class detection system for fake faces generated by GAN network. Sensors, 22(20):7767-7781, 2022. https://doi.org/10.3390/s22207767

[12] P M Arunkumar, Yalamanchili Sangeetha, P Vishnu Raja, and S N Sangeetha. Deep learning for forgery face detection using fuzzy fisher capsule dual graph. Information Technology and Control, 51(3):563-574, 2022. https://doi.org/10.5755/j01.itc.51.3.31510

[13] M S M Altaei. Detection of deep fake in face images using deep learning. Wasit Journal of Computer and Mathematics Science, 1(4):60-71, 2022.

[14] S.Sandhya Assistant, Akula Varshini, Donthala Harshitha, Gangidi Anisha, and Sambu Sarayu. Evaluating the performance of fake face detection in forensics. International Journal of Information Technology and Computer Engineering, 12(2):643-650, 2024.

[15] Yang Yu, Rongrong Ni, Wenjie Li, and Yao Zhao. Detection of AI-manipulated fake faces via mining generalized features. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(4):1-23, 2022. https://doi.org/10.1145/3499026

[16] Sunkari Venkateswarulu, and A Srinagesh. DeepExplain: Enhancing DeepFake detection through transparent and explainable AI model. Informatica, 48(8):6-19, 2024. https://doi.org/10.31449/inf.v48i8.5792

[17] Zhaohui Xie, and Guangyu Wu. Optimization method of basketball match evaluation based on computer vision and image processing. Informatica, 48(23):18-35, 2024. https://doi.org/10.31449/inf.v48i23.6696

[18] Faezeh Mosayyebi, Hadi Seyedarabi, and Reza Afrouzian. Gender recognition in masked facial images using EfficientNet and transfer learning approach. International Journal of Information Technology, 16(4):2693-2703, 2024. https://doi.org/10.1007/s41870-023-01565-4

[19] Liwei Deng, Hongfei Suo, and Dongjie Li. Deepfake video detection based on EfficientNet-V2 network. Computational Intelligence and Neuroscience, 2022(1):3441549-3441561, 2022. https://doi.org/10.1155/2022/3441549

[20] Guangtao Wang, Jun Li, Zhijian Wu, Jianhua Xu, Jifeng Shen, and Wankou Yang. EfficientFace: An efficient deep network with feature enhancement for accurate face detection. Multimedia Systems, 29(5):2825-2839, 2023. https://doi.org/10.1007/s00530-023-01134-6