2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Research on Privacy-Preserving Action Recognition Method Based on Adversarial Learning and Feature Enhancement

Xiaohan Qi[1,2,3] and Xingang Wang[1,2,3*]

[1] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[2] Shandong Engineering Research Center of Big Data Applied Technology, Faculty of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[3] Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Jinan, China
xgwang@qlu.edu.cn

Abstract. Video action recognition technology provides technical support for automated monitoring and event early warning, liberating human resources, preventing anomalies in advance, and handling events more promptly. However, the issue of user privacy leakage is also a growing concern. This paper discusses how to conduct action recognition while protecting privacy, proposing a generative adversarial network architecture that combines a multi-scale feature fusion generator with a spatiotemporal consistency discriminator. The network continuously enhances its capabilities through adversarial training strategies to protect facial privacy in videos. This network extracts features at different levels in the video through a multi-scale feature fusion mechanism so that the video after facial privacy protection still maintains a high degree of realism; at the same time, to ensure that the accuracy of action recognition is not compromised, the feature enhancement module is designed to enhance action features and inhibit the privacy features significantly; C3D is used as the action recognition model to accurately recognize a variety of actions in the video, such as running, jumping, falling, realizing the action analysis of video content. In this paper, the proposed method is evaluated in terms of privacy protection level and action recognition performance. Experiments are conducted on three datasets, LFW, HMDB51, and Hollywood2, and the results show that this framework effectively protects personal information while maintaining high action recognition accuracy.

Keywords: Multi-scale feature fusion, Facial privacy protection, Action recognition, Feature enhancement.

1 Introduction

Public safety and other domains benefit from action recognition technology, which raises the bar for video surveillance intelligence. However, privacy leaks have become a major problem due to the widespread usage of monitoring equipment and greater awareness of personal security [1] [2]. We must deal with a significant issue: how to preserve user privacy while using action recognition without sacrificing functionality? Please note that the first paragraph of a section or subsection is not indented.

The face is a crucial identifier and an important aspect of privacy protection among many other privacy concerns. The challenge with face de-identification, however, is that, while it is desirable to preserve the completeness and fluidity of action information to support action recognition tasks, the de-identified face must differ from the original face to prevent the disclosure of personal information. Finding a suitable technique to preserve data utility while appropriately de-identifying private information is therefore essential.

In order to safeguard facial privacy, the majority of studies now use masking and blurring to alter the face region at the pixel level. This obscures the identifying information but impairs the video's facial clarity and affects how the video is seen overall. Generative Adversarial Networks (GANs), which preserve anonymity by producing similar videos that are challenging for algorithms to identify as authentic, have received a lot of interest in recent years due to deep learning. Traditional GANs still have difficulties in practical applications, though. For example, generated anonymous faces can have distorted facial characteristics and low visual quality, which could impair action recognition ability and decrease naturalness.

This study develops a Privacy-Preserving Action Recognition Network (PARN) with three components based on the aforementioned issues: 1) MS-GAN, an adversarial learning architecture that guarantees the elimination of privacy-sensitive information from the original face. It is composed of a Spatiotemporal Consistency Discriminator (STC-Discriminator) and a Multi-Scale Feature Fusion Generator (MSFF-Generator). 2) To enhance action recognition performance and make behavioral features more salient, a feature enhancement module is created. 3) After face anonymization, the action recognition module uses a 3D convolution network (C3D) to accurately recognize action categories from videos. The idea explores efficient privacy protection without sacrificing recognition accuracy by taking into account temporal continuity and multi-scale properties. In summary,the contributions of this work are as follow:

- Aiming at the problem that existing facial privacy preservation methods are defective in security and realism, which often results in anonymized videos lacking aesthetic appeal and naturalness, the MS-GAN adversarial learning framework is designed. It consists of a Multi-Scale Feature Fusion Generator and a Temporal Consistency Discriminator. The generator adds a multi-scale feature fusion network to generate more secure, realistic, and natural faces; the discriminator assesses the authenticity and consistency of the generated videos. The respective capabilities are continuously improved through adversarial training strategies.

- Aiming at the problem that privacy-preserving methods affect the performance of action recognition, the Feature Enhancement Module (FEM) is proposed. This method introduces a dual attention mechanism and L2 regularization to reduce the weight of privacy-sensitive information while enhancing the weight of action features, aiding subsequent C3D modules in quickly pinpointing behavioral features and improving recognition performance.

2 Related Work

2.1 Privacy Protection Methods

In privacy preservation, scholars have conducted extensive research. Early methods like blurring [3] and masking [4] focused on pixel-level interference in facial areas. They were effective at concealing personal identity information, but they inevitably compromised image clarity and video aesthetics, while also being susceptible to advanced image restoration techniques and lacking in security.

Data encryption is another effective method for privacy protection. There are now two different types of video encryption: one is full encryption, such as using DES [5] or AES [6] to encrypt the entire video stream, ensuring the security of video content, but action recognition is no longer possible. The other type is selective encryption, which only encrypts data that is sensitive to privacy. [7] proposes an encryption (decryption) scheme for face regions based on spatial and numerical confusion. [8] developed an algorithm to locate and encrypt video regions of interest (ROI). [9] proposes real-time privacy protection in cloud computing environments, detecting moving targets while safeguarding data privacy.

With the advancement of deep learning, Generative Adversarial Networks (GANs) have received widespread attention as a natural technique for synthesizing de-identified images or videos, significantly advancing facial privacy protection research. Training generators and discriminators generate images or videos with similar features but are hard to identify, protecting privacy by making it hard for algorithms to tell if they are real or not. In this process, researchers are currently concentrating on aspects like naturalness, recognition utility, and security. A privacy-preserving GAN framework is proposed in [10] that adds new validator and moderator modules for facial de-identification, balancing image quality and privacy protection. Blaz et al. [11] use GANs to synthesize de-identified faces based on the k-same algorithm.

2.2 Action Recognition after Privacy Protection

The goals of facial de-identification are privacy preservation and data utility. This paper aims to protect personal information through facial de-identification while still enabling reliable action recognition tasks. Action recognition has a long research history, over the past few years, CNN models have achieved remarkable success, including two-stream CNNs and 3D convolution models for action classification. [12] proposed a two-stream multi-twin CNN that learns to extract shared features from down-sampled low-

resolution videos and trains a transform-adaptive action classifier. Since the application of 3D CNNs in action recognition [13], variants of 3D CNNs [14] [15] transform 3D convolution, which learns both temporal and spatial features, into 2D convolution with space as a surface and 1D convolution with time as a line. Feichtenhofer et al. [16] designed a two-path network to learn spatiotemporal information, achieving good accuracy in video recognition. Mekruksavanich et al. [17] proposed a video action recognition method based on 3D convolution network, which can learn the spatio-temporal features directly from the original video data and significantly improve recognition accuracy. Zheng et al. [18] achieved good detection results by training deep autoencoders to detect abnormal behaviors in videos.

The above frameworks recognize targets or activities in low-quality videos but do involve privacy-preserving learning and evaluation. Recently, adversarial learning methods have emerged, better balancing privacy protection and action recognition. To address the above problem, Ren et al. [19] proposed an adversarial learning framework that involves a video anonymizer and a discriminator in adversarial training, along with multi-task learning with an action detector. The former processes the raw video to eliminate sensitive information while keeping the performance of action recognition as unaffected as possible, while the latter utilizes Faster-RCNN [20] as a frame-level action detector for recognition from the anonymized video. This framework allows the two to enhance each other's performance during the learning process. The adversarial learning framework in [21] comprises anonymization, action recognition C3D, and privacy budget models. It addresses the issue of video quality degradation post-anonymization and enhances robustness against various attacking models through adversarial training and anonymization transformations. Liu et al. [22] refine the structure and parameters of deep learning models to mitigate the potential impact of privacy protection measures on video surveillance performance. Meanwhile, other research teams are developing new privacy-preserving algorithms that aim to enhance the efficiency of video surveillance systems while rigorously protecting individual privacy.

In summary, privacy-preserving has an extremely important role in action recognition, however, there are still many problems and challenges in the application process. 1) Data encryption prevents information leakage but increases data processing complexity and latency; 2) Due to the continuous temporal characteristics of the video data, it is difficult for traditional generative adversarial networks to ensure temporal coherence and stability when generating privacy-protected videos, which may lead to unnatural jumps and blurring of the generated videos visually; 3) Difficulty in balancing privacy protection and recognition accuracy：how to maintain accuracy and high efficiency of action recognition while ensuring privacy protection.

3 Model

To address privacy protection issues in action recognition, a Privacy-Preserving Action Recognition Network (PARN) is proposed, which is based on a generative adversarial network [10] and integrated with 3D convolutional networks [23]. This framework dynamically protects facial privacy while maintaining efficient and accurate recognition.

The framework consists of three main components, as shown in Fig.1: Privacy Protection Module, Feature Enhancement Module, and Action Recognition Module. The following section details the methods and implementation of each module.

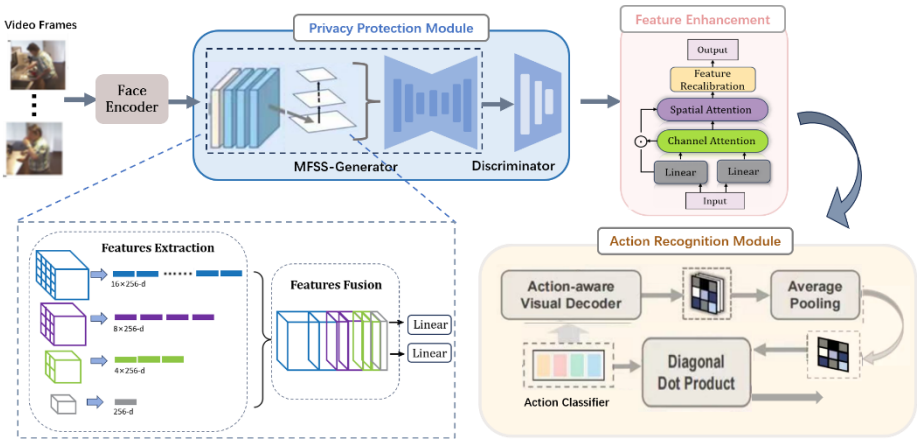

Fig. 1. The architecture of the PARN model

3.1 Privacy Protection Framework

MS-GAN. MS-GAN learns a mapping function $G:\{X,Z\}\rightarrow\hat{X}$, it is able to combine the input random noise Z with the actual data X to generate synthetic samples that are indistinguishable from the real data. In this process, the purpose of adversarial training is to train the generator G to produce sample data $\hat{X}$ that is difficult to identify for the discriminator D. The target function of MS-GAN is denoted as $L(G,D)$, then:

$$L(G,D)=E_{X,\hat{X}\sim P_{data}(X,\hat{X})}(\log D(X,\hat{X}))+ E_{X\sim P_{data}(X),Z\sim P_z(Z)}(\log(1-D(X,G(X,Z)))) \quad (1)$$

Our goal is to train the generator G to produce more realistic facial while competing against the discriminator D, minimizing the objective function. The optimal solution can be obtained through the following min-max optimization problem:

$$F*=argmin_G max_D L(G,D) \quad (2)$$

MS-GAN consists of a multi-scale feature fusion generator and a spatio-temporal consistency discriminator, as shown in Fig.2. The generator takes video frames with faces and outputs privacy-protected frames, where the facial areas are regenerated to conceal real identities. The generator uses the U-Net architecture, which first receives consecutive video frames and latent vectors, which are processed by the face detector with facial borders, and the latent vectors contain the feature encoding of the target face. The network employs multiple convolutional layers with a stride of 2 for downsampling, followed by batch normalization and ReLU activation functions. Downsampling uses max pooling layers to reduce the feature map size, halving the

width and height while increasing depth. The decoding part restores resolution by transposing the convolution and convolutions with a stride of 2, applying batch normalization and ReLU activation functions. Finally, the number of channels is adjusted to 3 by the convolutional layer and output synthetic video frames with a sigmoid activation function to generate image data in the range of [0, 1].

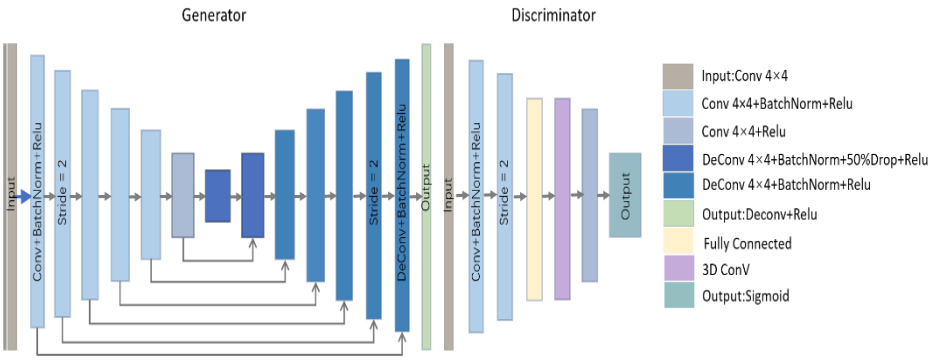

Fig. 2. The network structures of the generator G and the discriminator D. The generator adopts the U-Net network structure, with the input size being the same as the output size, both 1×128×128. The discriminator's input size is 1×128×128, and the output size is 1×30×30.

The spatiotemporal consistency discriminator receives real and generated video frames to assess the authenticity and consistency of the video, focusing on the overall structure and local details. First, the convolution layer extracts frame-level features and outputs a scalar representation of the frame's reality through a fully connected layer. Temporally, a 3D convolution layer is used to ensure smooth transitions when moving objects across consecutive frames. For example, a moving object should move smoothly in consecutive frames instead of changing abruptly. Within a single frame, the discriminator checks for consistency in the spatial attributes of objects, such as clear edges and uniform color. Adversarial training drives the generator and discriminator to compete, continuously improving their performance.

Multi-scale feature fusion. To capture information at various scales in video, a multi-scale feature fusion mechanism is added to the generator, which processes and merges features of different scales to capture details from coarse to fine, a feature not commonly seen in previous generators. as shown in Fig 3. Specifically, four feature layers are generated using convolutional blocks of different sizes: a 4×4 convolutional block extracts high-level semantic features containing thematic and scene information; a 3×3 convolutional captures intermediate features for spatiotemporal information; a 2×2 convolutional block extracts information on dynamic changes and motion; and a 1×1 convolutional block focuses on frame-level features like texture and edges.

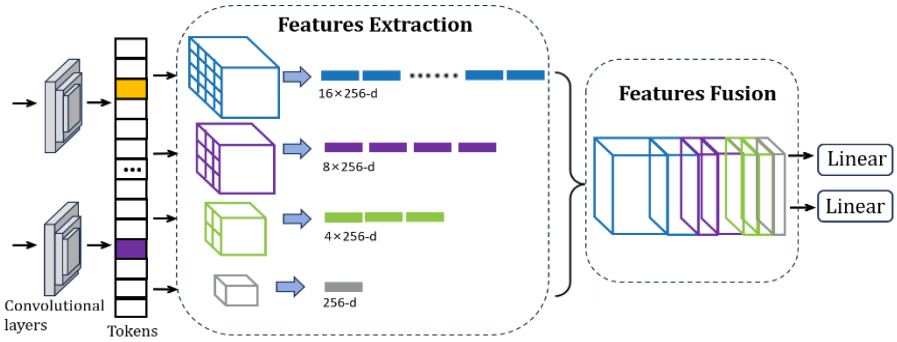

Fig. 3. The structure of the multi-scale feature fusion network. 256 is the number of channels, and the number of convolutional kernels equals the number of channels.

The four feature maps are different in size and number of channels, and are adjusted to the same number of channels and size before they are fused using the concatenate method. After fusion, 3x3 convolution is applied to further integrate the features. The generator then creates video frames based on the fused feature vectors, with each frame inheriting the state of the previous frame while introducing new changes to simulate facial features. The multi-scale feature fusion of the generator can be represented as:

$$G(I_t, Z, C) = \sum_{l=1}^{L} \omega_l \cdot F_l(\varphi_l(I_t, Z, C)) \tag{3}$$

Where I_t represents the real video frame at time step t, $G(I_t)$ is the video frame generated by the generator, Z is the latent variable drawn from the prior distribution P_Z, C denotes conditional information (such as facial features or expression labels). L is the scale level, F_l is the fusion function for the l^{th} level features, φ_l is the feature extraction function at the l^{th} level, ω_l is the weight coefficient for the l^{th} level, γ is a scalar, and $score(F_l)$ is the scoring function for the l^{th} feature map F_l.

$$\omega_l = \frac{e^{(\gamma \cdot score(F_l))}}{\sum_{l=1}^{L} e^{(\gamma \cdot score(F_l))}} \tag{4}$$

3.2 Feature Enhancement Module

Currently, in most of the privacy-preserving action recognition studies, the performance rather suffers with the application of various privacy-preserving methods. To address this issue, the feature enhancement module is designed to maintain or enhance recognition performance, as shown in Fig.4. This module first receives video frames anonymized by the MS-GAN, using 1×1 and 3×3 convolutional layers to extract features of privacy-sensitive information. It then passes through a fully connected layer and a Sigmoid function to output a probability value, $P \in [0,1]^C$, where P_C indicates the probability of privacy information in the c^{th} channel. Based on these probabilities, a Channel Suppression Mask (CSM) is generated, defined as $CSM = \sigma(P)$, where σ is the sigmoid function.

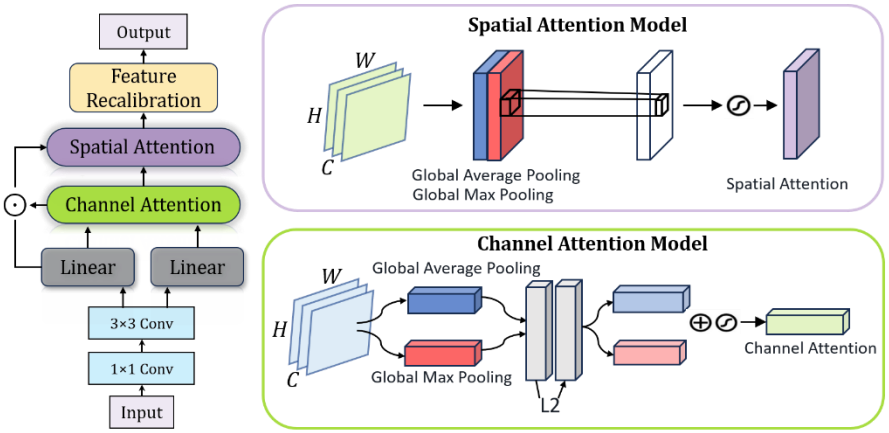

Fig. 4. The left figure is the structure of the feature enhancement module, and the right figure is the structure of the temporal attention and spatial attention.

In order to enable the model to autonomously focus on key regions in the video, a dual attention mechanism is introduced, including a channel attention module and a spatial attention module. The channel attention module performs Average Pooling and Max Pooling on the feature map $F \in R^{C\times W\times H}$ (C is the number of channels, W is the height, and H is the width) to capture the global information. This is followed by processing through two fully connected layers. The output of the first fully connected layer is processed with L2 regularization to enhance the learning capability of the behavior feature. The output of the second fully connected layer passes through the sigmoid function to generate the channel attention weight CA, which is multiplied with the channel suppression mask CSM output from the privacy feature detection module to obtain the final adjusted channel attention weight $W_{ch} = CA \odot CSN$.

The spatial attention module pools the spatial dimensions of the feature maps, joins them and processes them through a convolutional layer to generate spatial attention weights SA and weights the feature maps to enhance the behavioral features in the spatial domain. Feature recalibration then uses the adjusted channel attention weights and spatial attention weights to weight each channel and spatial domain of the feature map, respectively.

This module solves the imbalance between privacy preservation and action recognition tasks, improving recognition performance after privacy preservation.

3.3 Action Recognition for Privacy Protection

The reason for using C3D as a human behavior recognition model is that compared to 2D convolution, C3D has a better ability to model temporal information due to 3D convolution and 3D pooling operations, which enables better understanding of dynamic behaviors in the video and end-to-end training, thus eliminating the need for complex preprocessing steps, while the preprocessed model can achieve good transfer learning results on new tasks and improve recognition accuracy.

Based on [23], the network takes video clips as input, all video frames are resized to 128×171 and the video is segmented into 16 frame clips that do not overlap each other and are used as input to the network. The network contains 5 convolutional layers, each of which is immediately followed by a pooling layer, 2 fully connected layers, and a softmax loss layer to predict the action labels. The number of filters in the convolutional layers is 64, 128, 256, 256, 256 in that order, all of which should be properly padded to maintain size. The pooling layers (except the first) use 2 × 2 × 2 kernels with a step size of 1, reducing the output size by 8 times. The first pooling layer has a 1×2×2 kernel to preserve temporal information. The Fully Connected Layer has a total of 2048 outputs. We trained the network from scratch using small batches of 30 clips to accurately recognize running, jumping, lying, and falling behaviors in videos.

3.4 Loss Function

Our goal is twofold: one is to ensure that facial information cannot be recognized, and the other is to ensure that behavioral features in the video can be recognized correctly. Therefore, multiple loss functions are used to help the model learn richer feature representations and improve the overall performance.

Loss of MS-GAN. In generator G, a structural similarity index is used to keep the structure of the generated face similar to the original image, but the facial information is hidden:

$$L_{SSIM(G)} = 1 - SSIM(X, G(X)) \tag{5}$$

In discriminator D, the adversarial loss is used to enable the discriminator to better distinguish between real and generated images:

$$L_{adv}(D) = -E_{X\sim P_{data}}(\log(D(X))) - E_{X\sim P_{data}}(\log(1 - D(G(X)))) \tag{6}$$

The total loss function of the MS-GAN is obtained by summing the losses of the generator and discriminator:

$$L_{total} = \lambda_{adv} L_{adv}(G) + \lambda_{feat} L_{feat}(G) + \lambda_{SSIM} L_{SSIM}(G) + L_{adv}(D) \tag{7}$$

where λ_{adv} and λ_{feat} and λ_{SSIM} are the weights that balance the different loss terms.

Loss of action recognition. The categorical cross-entropy loss is used to evaluate action prediction errors in video frames and enhance recognition accuracy through model optimization. For a given video frame and its real behavioral labels, this loss function is capable of evaluating the prediction error and optimizing the model parameters through a back-propagation algorithm, aiming to improve the accuracy of behavior recognition. It ensures effective recognition while preserving facial privacy:

$$L_{cls} = -\sum_c yc \log(\hat{y}c) \tag{8}$$

Where yc is the true behavioral label and $\hat{y}c$ is the predicted behavioral probability.

Combining the facial privacy-preserving loss and the action recognition loss is the total loss for the whole model:

$$L_{final} = L_{total} + \lambda_{cls} L_{cls} \tag{9}$$

where λcls is the weight of the classification loss.

4 EXPERIMENTS AND ANALYSIS

4.1 Datasets

To validate the method proposed in this paper for action recognition while preserving privacy, experiments were conducted on the HMDB51, Hollywood2 and LFW datasets. Because they contain actions involving face regions, they are good testbeds for joint face anonymization and action detection models. The HMDB51 dataset has 960 videos covering 21 common actions that contain faces. The Hollywood2 dataset contains 3669 videos with a total duration of 20.1 hours, covering video clips of 12 different action categories and 10 different scenes. The LFW dataset has more than 13,000 facial images involving 5,749 people, covering a wide range of expressions, gestures, and ages.

4.2 Evaluation Metrics

Facial Recognition Error Rate, FRA. To confirm the effectiveness of our framework in privacy protection, two points must be proven: first, the generated faces are significantly different from the originals, second, the generated faces are not simply replaced with other identities. Therefore, we use the FRA rate to measure the privacy protection effect. As shown in Fig 5, while other methods increase the face recognition error rate, they also degrade action recognition performance. This indicates a trade-off between privacy protection and action recognition, which our method adeptly balances, achieving dual optimization of privacy and performance.

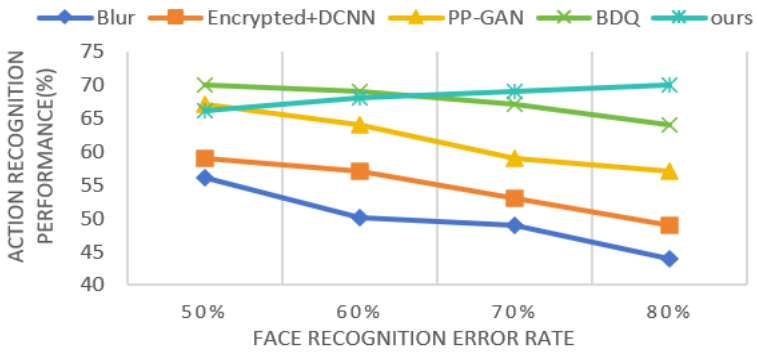

Fig. 5. The x-axis represents the face verification error rate, while the y-axis denotes the action detection performance. An upward trend indicates better effectiveness.

F1 Score. The harmonic mean of precision and recall, is used to evaluate the performance of the model. A higher F1 score indicates a better balance between precision and recall.

4.3 Implementation Details

The MS-GAN module was trained on the LFW dataset using an alternating training strategy for the generator and discriminator, both with a learning rate of 0.0001. The total number of iterations was 150 epochs, updating the discriminator every five generator iterations to maintain balance and prevent instability due to rapid learning. The feature enhancement module was pre-trained on ImageNet to learn rich visual features with a learning rate of 0.0005, iterating 80 times, batch size 32, using Adam optimizer and weight decay of 0.00005. It was then fine-tuned on Hollywood2 with dynamic step adjustment, initial learning rate reduced to 0.00005, iterating 30 epochs, and batch size adjusted to 8. The C3D module takes video clips as input and extracts spatiotemporal features via convolution and pooling. With an initial learning rate of 0.0005 and a decay strategy, it iterates 80 times, using a batch size of 4. The SGD optimizer is employed with a momentum of 0.9 and a weight decay of 0.0005.

In the integration and joint training phase, pre-trained modules were integrated into the PARN framework for joint training to optimize performance. The learning rate was adjusted to one-tenth of the pre-training rate, with 30 iterations and a batch size of 8, using the original optimizers.

4.4 Comparison with Other Models

For video action recognition with privacy-preserving needs, it is compared with other methods on the LFW dataset and the results are shown in Table 1. Blurring [3] and encryption [8] by focusing on pixel-level interference in the face region. Face Swapping selects similar face images from a database for replacement. K-Same-Net [11] uses GAN to synthesize de-identified faces. PP-GAN [10] designs adversarial frameworks that incorporate anonymization and privacy budget models. Our approach achieves the lowest score in de-identification accuracy, indicating that the generated faces are harder to recognize and better protected.

Table 1. Comparison with other models on LFW dataset.

Privacy protection methods	Data security score (1-10)	FRA(%)
Blur	7	52.1
Encrypted	8	53.6
Face Swapping	8.2	61.2
k-same-net	8.4	63.4
PP-GAN	8.7	64.5
MFSS-GAN(ours)	**9**	**66.3**

Table 2. Comparison with other models on HMDB51 dataset.

Model	Accuracy	Precision	Recall	PP-Accuracy	F1 score
PP-GAN	87.4	88.0	87.1	86.2	87.5
SPAct	89.7	90.4	89.2	90.1	89.8
VSCS+C3D+AdSRC	91.9	92.3	91.2	89.7	91.7
BDQ	91.2	91.9	91.0	90.0	91.4
GAN+FasterRCNN	90.8	92.8	90.4	88.9	91.6
MFSS-GAN(ours)	**93.6**	**94.7**	**93.1**	**93.6**	**93.9**

We evaluated the performance of action recognition with privacy protection on the HMDB51 and Hollywood2 datasets and compared it with existing methods, as shown in Table 2 and Table 3. (Accuracy is the accuracy of action recognition, PP-Accuracy is the accuracy of action recognition after privacy protection). PP-GAN [10] enhances the GAN framework with validators and regulators for face de-identification. VSCS+C3D+AdSRC [24] uses compressed sensing for visual obfuscation and fine-tunes a C3D network for video action representation. The privacy-preserving encoder BDQ [25] achieves privacy protection and action recognition through blurring, differencing, and quantization. GAN+FasterRCNN [20] employs adversarial training of video anonymizers and discriminators for pixel-level modifications. SPAct [26] proposes a self-supervised privacy-preserving action recognition framework that does not require privacy labels for training.

Table 3. Comparison with other models on Hollywood2 dataset.

Model	Accuracy	Precision	Recall	PP-Accuracy	F1 score
PP-GAN	86.9	87.7	87.2	86.4	88.4
SPAct	91.1	91.2	89.6	90.4	90.3
VSCS+C3D+AdSRC	91.4	92.8	90.9	90.5	91.8
BDQ	90.5	91.2	90.4	90.0	90.8
GAN+FasterRCNN	88.3	89.6	88.0	88.2	98.8
MFSS-GAN(ours)	**93.2**	**94.5**	**92.8**	**94.3**	**93.6**

In the HMDB51 dataset (Table 2), our method MFSS-GAN+C3D achieved an accuracy of 93.6%, indicating high recognition and classification capabilities. VSCS+C3D+AdSRC [24] and GAN+FasterRCNN [20] also showed good precision but were slightly lower than our proposed model. PP-GAN [10] and BDQ [25] had relatively lower accuracy after incorporating privacy protection. In the Hollywood2 data set (Table 3), our model achieved a precision of 93.2% and an improvement in the F1 score of 0.8% -2.5% compared to the latest methods SPAct [26] and BDQ [25], demonstrating a good balance between precision and recall. These results indicate that our method effectively achieves de-identification, ensuring privacy protection without compromising action recognition performance.

4.5 Ablation Studies

As described in the previous modules, PARN consists of three modules: privacy preservation, feature enhancement, and action recognition. Here, we study its role in privacy protection and action recognition. For this, we select combinations of modules on the HMDB51 dataset, which allows us to reveal which components are critical and how they interact to improve the accuracy and level of privacy preservation of action recognition.

Table 4. Results of ablation experiments.

Model	FAR	Accuracy	Precision	Recall	F1 score
GAN+C3D	83.8	93.5	91.2	89.4	90.3
MS-GAN+C3D	87.4	90.2	89.6	88.4	88.9
GAN+FEM+C3D	83.6	93.7	91.9	90.5	91.1
MS-GAN+FEM+C3D	88.3	94.1	92.8	93.3	93.4

Table 4 shows the performance results for different module combinations on overall performance (the FAR rate represents the level of privacy protection). The baseline model (GAN+C3D) lacks multi-scale information, offering insufficient privacy protection and affecting the accuracy of action recognition; the addition of the multi-scale feature fusion network (MS-GAN) enhances privacy protection to 87.4%, capturing rich spatial and temporal information in videos for more natural and smooth privacy-protected faces, but at the cost of a 3.3% decrease in action recognition accuracy, indicating a loss of precision in complex behaviors or detailed scenes, such as fast-moving objects or subtle actions; adding only the feature enhancement module (GAN+FEM+C3D) results in higher action recognition accuracy by assigning different attention weights but significantly reduced privacy protection. In general, the complete model achieves an F1 score of 93.4%, which not only achieves high accuracy in action recognition but also offers strong protection of privacy. Hence, the combined effect of the MS-GAN with a multi-scale feature fusion network and the feature enhancement module significantly outperforms any single module alone.

5 CONCLUSION

Aiming at the privacy leakage problem in video action recognition, this paper proposes an adversarial learning framework PARN for privacy protection. Based on generative adversarial network technology, we design a multi-scale feature fusion generator and a spatiotemporal consistency discriminator to re-generate the face of the person in the video to achieve the purpose of privacy protection. To not affect the performance of action recognition, we design the feature enhancement module, which combines the dual attention mechanism and the L2 regularization term to suppress irrelevant privacy features and highlight the behavioral features of the human body. Extensive experimental results on three datasets LFW, HMDB51, and Hollywood2 validate the

effectiveness of the proposed method in this paper. In future work, more advanced techniques will be investigated to integrate privacy preservation with action recognition better.

Acknowledgments. The Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences)(2024ZDZX08)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Jian X, Shupeng W, Yu D. Fraudulent phone call recognition method based on convolutional neural network[J]. HIGH TECHNOLOGY LETTERS, 2020, 26(4): 367-371.
2. Xiao F, Lin Z, Sun Y, et al. Malware detection based on deep learning of behavior graphs[J]. Mathematical Problems in Engineering, 2019, 2019(1): 8195395.
3. Agrawal, Prachi, Narayanan ,et al. Person De-Identification in Videos.[J].IEEE Transactions on Circuits & Systems for Video Technology, 2011.
4. Slobodan Ribaric, Aladdin Ariyaeeinia, Nikola Pavesic, De-identification for privacy protection in multimedia content: A survey, Signal Processing: Image Communication, Volume 47, 2016, Pages 131-151.
5. Shahid Z, Chaumont M, Puech W. Fast protection of H. 264/AVC by selective encryption of CAVLC and CABAC for I and P frames[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 21(5): 565-576.
6. Zhao Y , Zhuo L , Niansheng M ,et al.An object-based unequal encryption method for H.264 compressed surveillance videos[C]//Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on.IEEE, 2012.
7. Liu S, Kong L, Wang H. Face detection and encryption for privacy preserving in surveillance video[C]//Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III 1. Springer International Publishing, 2018: 162-172.
8. Hosny K M, Zaki M A, Hamza H M, et al. Privacy protection in surveillance videos using block scrambling-based encryption and DCNN-based face detection[J]. IEEE Access, 2022, 10: 106750-106769.
9. Chu K Y, Kuo Y H, Hsu W H. Real-time privacy-preserving moving object detection in the cloud[C]//Proceedings of the 21st ACM international conference on Multimedia. 2013: 597-600.
10. Wu Y, Yang F, Xu Y, Ling H. Privacy-protective-gan for privacy preserving face de-identification. Journal of Computer Science and Technology, 2019, 34(1): 47-60 .
11. Meden B , Emeri I , Truc V ,et al.k-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face Deidentification[J].Entropy, 2018, 20(1):60.
12. Bitouk D .Face swapping: Automatically replacing faces in photographs[J].Proc Siggraph, 2008, 27(3):1-8.
13. Ryoo M S , Kim K , Yang H J .Extreme Low Resolution Activity Recognition with Multi-Siamese Embedding Learning[J]. 2017.

14. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.
15. Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]//proceedings of the IEEE International Conference on Computer Vision. 2017: 5533-5541.
16. Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.C.
17. Feichtenhofer, H. Fan, J. Malik and K. He, "SlowFast Networks for Video Recognition," pp. 6203-6211,2019.
18. Mekruksavanich S, Jitpattanakul A. Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models[J]. Electronics, 2021, 10(3): 308.
19. Zheng, D. , Zhang, Y. , & Xiao, Z. (2021). Deep learning-driven gaussian modeling and improved motion detection algorithm of the three-frame difference method. Mobile Information Systems, 2021(15), 1-7.
20. Ren Z , Lee Y J , Ryoo M S .Learning to Anonymize Faces for Privacy Preserving Action Detection[J]. 2018.
21. Ren S , He K , Girshick R ,et al.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
22. Wu Z , Wang H , Wang Z ,et al.Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020
23. Liu H, Feng Z, Guo Q. Multimodal Cross-Attention Mechanism-Based Algorithm for Elderly Behavior Monitoring and Recognition[J]. Chinese Journal of Electronics, 2025, 34(1): 1-13.
24. Liu J, Zhang R, Han G, et al. Video action recognition with visual privacy protection based on compressed sensing[J]. Journal of Systems Architecture, 2021, 113: 101882.
25. Kumawat S, Nagahara H. Privacy-preserving action recognition via motion difference quantization[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 518-534.
26. Dave I R , Chen C , Shah M .SPAct: Self-supervised Privacy Preservation for Action Recognition[J]. 2022.