Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Research on Key Technologies and Application of Video Sensitive Information Detection

Guiyuan He，Bin Song, Yueheng Mao

School of Information Engineering (School of Artificial Intelligence), Henan University of Science and Technology, Luoyang, China

***Abstract*— With the development of Internet technology, the video industry has advanced rapidly. However, some videos contain pornographic, political and other sensitive information. The lack of supervision will harm the physical and mental health of viewers and social stability, so automatic detection technology is urgently needed to assist manual review. At present, mainstream methods such as frame-by-frame image recognition and 3D convolutional neural network-based classification suffer from problems including slow detection speed, serious missed detection and false detection. To this end, this paper carries out relevant research and innovations: A detection method combining a lightweight object detection model and a key frame extraction algorithm is proposed to reduce the workload through video preprocessing; The YOLOv5s model is improved, with a GPSA module and an optimized feature fusion network designed, achieving a mAP of 71% and a single-frame detection time of 2.8 ms; A key frame extraction method fusing shallow and deep features is proposed, and the GhostEfficientNet_s model is constructed to extract key frames accurately; The algorithms are encapsulated into interfaces and applied to the Shaike Network platform. The scheme can efficiently process sensitive information and provide support for cybersecurity governance in cyberspace.**

***Keywords*— Video sensitive information detection; Object detection; Key frame extraction; Feature fusion; YOLOv5s; Lightweight network.**

I. INTRODUCTION

With the rapid development of internet technology, online videos have become an integral part of people's daily lives. Short video platforms such as Douyin, Kuaishou, and Bilibili have gained increasing popularity, and creator incentive programs have driven an explosive growth in the number of online videos. According to the 56th Statistical Report on China's Internet Development released by the China Internet Network Information Center (CNNIC), as of June 2025, the number of internet users in China reached 1.123 billion, with an internet penetration rate of 79.7%; the number of short video users reached 1.068 billion, accounting for 95.1% of the total internet users, and the number of online video users reached 1.085 billion, making up 96.7% of the total internet users. Video has become one of the most dominant information dissemination channels on the internet.[1]

However, the content of these videos is highly mixed in quality, with a portion containing sensitive information involving pornography, politics-related violations, and terrorism. In the absence of effective supervision, such content will not only seriously damage the physical and mental health of adolescents, but also pose a threat to social stability, making video content supervision the most challenging task in online content governance.[2] At present, video review on major platforms still relies heavily on manual work. Reviewers are required to monitor video content for long hours, which is extremely energy-consuming, and manual review alone is far from sufficient to handle the massive volume of videos. Meanwhile, existing video sensitive information detection technologies, whether frame-by-frame image recognition or 3D convolutional neural network-based classification, suffer from inherent drawbacks including slow detection speed and high rates of missed and false detections, which are completely unable to meet the practical application needs of internet platforms.

To address the above issues, this paper conducts research on key technologies for video sensitive information detection, and proposes a technical scheme combining key frame extraction and object detection methods. This scheme achieves an optimal balance between detection accuracy and efficiency, which can effectively assist platforms in video content review and safeguard the security of cyberspace.

II. PRELIMINARIES

Compared with two-stage object detection algorithms, one-stage algorithms do not require a separate candidate region generation step, and can simultaneously output the position coordinates and category results of targets through only a single forward propagation of the network, which has an extremely significant advantage in detection speed. This feature makes it more suitable for the real-time detection scenario of massive video frames. Among one-stage algorithms, the YOLO series has achieved an excellent balance between detection accuracy and inference speed through multiple generations of iterative optimization[4]. As the lightweight version of this series, YOLOv5s has a small number of model parameters and a low deployment threshold. Meanwhile, it reduces the information loss during downsampling through the slicing operation of the Focus layer,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

and is equipped with the CSP structure and PAN feature fusion network, which enables it to have considerable detection capability for small-size sensitive targets. Therefore, this study selects YOLOv5s as the basic detection model for optimization and improvement[5].

In the current field of video sensitive information detection, the most commonly used methods for key frame extraction include the sampling method, shot segmentation method, clustering method and inter-frame difference method. As shown in Fig. 1, the key frame extraction algorithm designed in this paper is optimized with the inter-frame difference method as the basic framework. The feature calculation part fuses the shallow color features and deep semantic features of the image, and the threshold determination step combines the local maximum method for secondary screening[3]. At the same time, the inter-frame difference method has the advantages of simple calculation logic, fast operation speed and high sensitivity to the dynamic changes of video frames. It can adapt to the video processing requirements of different durations and content types, and greatly reduce the overall computational overhead of the algorithm.

The inter-frame difference method completes key frame judgment by calculating the feature difference between adjacent video frames, and its core calculation formula for the inter-frame difference value is given by:

$$I_{\mathrm{LUV}}{}^{n} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|P_{ij}^{\,n} - P_{ij}^{\,n-1}\right|$$

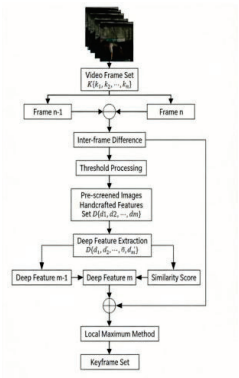

Fig. 1 Framework of Key Frame Extraction Method

Here, H and W represent the height and width of the video frame respectively, $F_n(i,j)$ is the feature value of the n-th frame image at the (i,j) position, and D(n) is the inter-frame difference value between the n-th frame and the previous frame. When D(n) exceeds the preset threshold, the n-th frame can be marked as a key frame.

Feature fusion is the core of the optimization of the inter-frame difference method in this paper, which combines the shallow features of the image in the LUV color space and the deep semantic features extracted by the convolutional neural network simultaneously. The LUV color space is a uniform color space developed by the International Commission on Illumination (CIE), where L represents the luminance component, and U and V represent the chrominance components. Compared with the RGB color space, it separates the chrominance and luminance information, has stronger anti-interference capability against illumination changes, and is more suitable for calculating the pixel difference between frames. Deep features are extracted through a lightweight convolutional network, which can capture the content semantic changes of the frame and make up for the defect that shallow features cannot identify sensitive information. In the process of key frame extraction, calculating the inter-frame difference through weighted fusion of the two types of features can accurately capture video frames containing sensitive information while reducing redundant frames, thus avoiding missed detection of key content.

III. LIGHTWEIGHT DETECTION MODEL IMPROVEMENT

To address the problems that traditional image classification methods are difficult to accurately capture small-size sensitive information in complex scenes, and the existing detection models have difficulty in balancing accuracy and speed, this paper takes YOLOv5s with excellent detection speed as the basic model, carries out optimization and improvement from two aspects of feature extraction and feature fusion, and designs a lightweight object detection model adapted to the sensitive information detection scenario.

A. Core Improvement Design of the Model

The model improvement in this paper is mainly divided into two core parts: feature extraction and feature fusion. In the feature extraction part, a lightweight high-efficiency GPSA attention module is designed based on the PSA module and Ghost module[8], which solves the problem of excessive computational complexity of the original PSA module, enables the network to learn more multi-scale feature representations, and improves the detection accuracy of sensitive information[6]. The structure of the GPSA module is shown in Fig. 2. In the feature fusion part, drawing on the BiFPN structure, the PAN network of the original model is optimized. By pruning invalid fusion nodes, adding cross-scale residual connections and learnable weights, the multi-scale feature fusion capability of the model is enhanced[7]. The structure comparison between PANet and BiFPN is shown in Fig. 3. At the same time, the EIOU loss function is selected to replace the CIOU loss function of the original model, which solves the problem of limited convergence of width and height of the original loss function and accelerates the convergence speed of the prediction boxes. The overall network structure of the improved YOLOv5s is shown in Fig. 4.

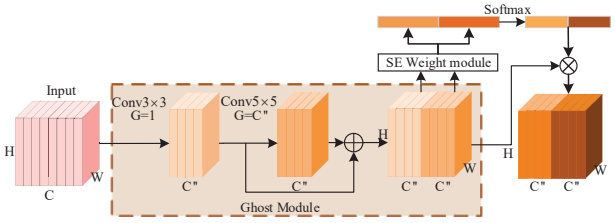

Fig. 2 Schematic Diagram of GPSA Module Structure

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

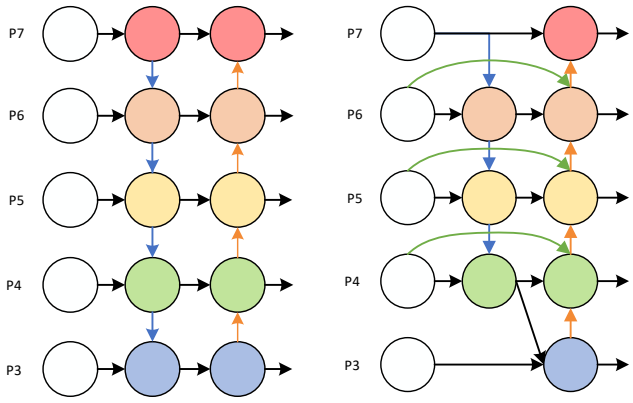

Fig. 3 Structure Diagrams of PANet and BiFPN

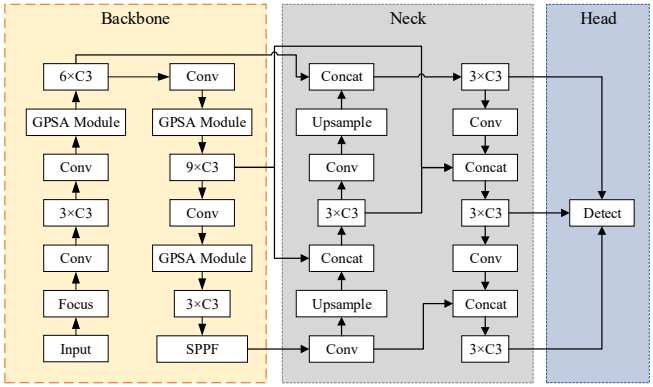

Fig. 4 Schematic Diagram of the Improved YOLOv5s Network Structure

B. Experimental Environment and Dataset Construction

The software environment of this experiment is Python 3.6, the Pytorch 1.10.0 deep learning framework with CUDA 11.3. The hardware training environment is equipped with an RTX 3080 graphics card (10GB video memory) and a 12-core Intel Xeon Platinum 8255 CPU. Since there is no unified public dataset for sensitive image detection, 3681 images are collected and screened from the Internet to construct a self-made dataset in this paper. The sensitive information is subdivided into 4 pornography-related categories, 3 politics-related categories and 3 terrorism-related categories, totaling 10 categories. The Label-Img tool is used to complete the manual annotation of all images, which are applied to the training and testing of the model.

C. Experimental Results of Model Performance

In this paper, Precision, Recall, mAP, and single image detection time are taken as the core indicators. The model performance is verified through ablation experiments and comparative experiments with mainstream models, and the results of the ablation experiments are shown in Table I. It can be seen from the experimental results that the improved model integrated with the GPSA module and BiFPN structure achieves a mAP50 of 66.4%, with a single image detection time of only 2.8ms. Compared with the model with the original PSA module, the detection speed is increased by more than 40% under the premise of the same accuracy level, which verifies the effectiveness of the two improvements.

The results of the comparative experiment with mainstream models are shown in Table II. The improved model achieves an overall mAP of 71% on the self-made sensitive dataset, which outperforms mainstream models such as Faster-RCNN, SSD and YOLOv3. Meanwhile, the single image detection speed is only 2.8ms, and the number of model parameters is only 8.68M. It achieves a better balance between detection accuracy and inference efficiency, and can meet the actual deployment requirements of network platforms.

TABLE. I Results of ablation experiments

Group	GPSA Module	BiFPN	Precision	Recall	mAP50	map@.5:.95	T（ms）
Group1	×	×	76.7%	59%	63.6%	36.3%	2.5
Group2	√	×	78.8%	57.8%	64.5%	38.2%	2.6
Group3	×	√	80.7%	59.4%	64.7%	38.3%	2.6
Group4	PSA Module	√	78.3%	60.8%	66.4%	39.2%	4.7
Proposed Model	√	√	81.4%	59.7%	66.4%	39.4%	2.8

TABLE. II Results of each model on the data set in this paper.

Model Name	Ap50			mAP.5	mAP@.5:.95	FLOPs（G）	Parameters（M）	T（ms）
	Pornography-related	Politics-related	Terrorism-related					
Faster-RCNN [10]	54.3%	66.7%	78.4%	67.7%	34.6%	206.71	41.17	45.0
SSD	55.6%	67.4%	82.8%	70.4%	34.4%	347.61	25.71	17.2
centernet	38.2%	30.9%	74.9%	50.8%	29.0%	51.05	14.21	11.0
YOLOv3[4]	51.0%	59.6%	79.6%	65.0%	31.7%	194.04	61.57	18.3
Retinanet_r50	58.1%	71.6%	78.9%	70.3%	39.2%	208.34	36.29	20.3
Proposed Model	68.8%	83.5%	65.9%	71%	42.8%	19.1	8.68	2.8

IV. KEY FRAME EXTRACTION METHOD BASED ON FEATURE FUSION

In video content auditing, manual frame-by-frame viewing is extremely time-consuming. Even if an object detection algorithm is used to detect all video frames one by one, it will generate huge computational overhead, which seriously affects the auditing efficiency. However, most traditional key frame extraction methods only focus on the low-level color and texture features of the frame, which are prone to missing key frames that have small frame changes but contain

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

sensitive information, and thus cannot meet the requirements of sensitive content detection. Therefore, based on the inter-frame difference method with simple calculation logic and strong adaptability, this paper proposes a key frame extraction method that fuses shallow features and deep features. While accurately locking video frames containing sensitive information, it can filter redundant frames to the maximum extent and reduce the workload of subsequent detection and auditing[3].

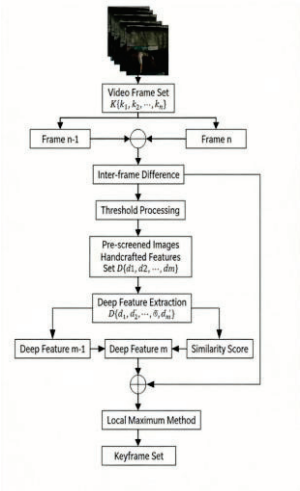

Fig. 5 Key Frame Extraction Method Based on Fusion Features

A. Overall Design of the Method

The overall flow of the proposed key frame extraction method is shown in Fig. 5, which is divided into three core steps: feature extraction, inter-frame difference calculation, and key frame screening. First, the shallow features of the video frame in the LUV color space and the deep semantic features extracted by the lightweight network are obtained respectively. After weighted fusion of the two types of features, the difference value between adjacent video frames is calculated. After the preliminary screening of key frames through the preset threshold, the local maximum method is used for secondary filtering. Finally, the key frame set that can fully represent the video content and cover all sensitive information is output.

B. Optimization Design of Core Modules

In the part of shallow feature extraction, the LUV color space is selected in this paper to replace the commonly used RGB color space. Compared with the RGB color space, the LUV color space can completely separate the luminance and chrominance information of the frame, and has stronger anti-interference capability against illumination changes. The inter-frame difference calculated based on it is more stable, which will not be misjudged as a key frame due to slight illumination changes of the frame, thus reducing the generation of redundant frames.

In the part of deep feature extraction, to balance the accuracy and speed of feature extraction, a lightweight network GhostEfficientNet_s is designed in this paper. The model structure is shown in Table III, and the network structure is shown in Fig. 6. This network is simplified and optimized in three aspects based on EfficientNet[9]: First, the Ghost module with SE channel attention is added to extract richer features with fewer parameters[8]. Second, the Swish activation function of the original network is replaced by the HardSwish function with faster calculation speed to improve inference efficiency. Third, network pruning is performed to remove the layers that contribute little to accuracy improvement, so as to further compress the model size.

TABLE. III The architecture of GhostEfficientNet_s.

No.	Module Name	Resolution	Number of Channels	Number of Repetitions
1	SE-Ghost Module	224×224	16	1
2	Conv,k3×3	224×224	16	1
3	MBConv1,k3×3	112×112	16	1
4	MBConv6,k3×3	112×112	32	1
5	MBConv6,k5×5	56×56	64	2
6	MBConv6,k3×3	28×28	80	2
7	MBConv6,k5×5	14×14	119	2
8	MBConv6,k5×5	14×14	160	1
9	MBConv6,k3×3	7×7	320	1
10	Conv1×1&Pooling&FC	7×7	1280	1

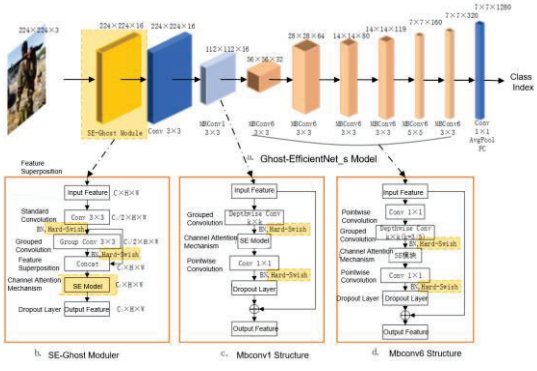

Fig. 6 Network Structure of GhostEfficientNet_s

C. Experimental Results and Analysis

The software and hardware environment of this experiment is consistent with that used for the object detection model in Chapter 3. The experimental dataset is constructed by combining public datasets, containing a total of 14,575 images

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

including pornography-related, politics-related, terrorism-related and normal images.

The results of the ablation experiment are shown in Table. IV. The finally optimized GhostEfficientNet_s model achieves a sensitive image recognition accuracy of 93.79%, with only 1.862M parameters, and the feature extraction time for a single image is only 6.37ms. Compared with mainstream lightweight models such as MobileNetV3 and GhostNet (the results are shown in Table V), this model has obvious advantages in both accuracy and speed. It can accurately locate sensitive targets in the frame and has stronger anti-interference ability against background clutter.

TABLE IV Results on the ablation experiment of GhostEfficientnet _ s model

No.	Improvement Strategy	Accuracy	F1-Score	Parameters (M)	Detection peed (batch=1)
1	EfficientNet	93.13%	93.59%	4.013M	8.88
2	EfficientNet+GM	93.89%	94.17%	4.022M	9.06
3	EfficientNet+GM+SE （GhostEfficientNet）	94.17%	94.46%	4.022M	9.30
4	EfficientNet+GM+SE +（Swish->HSwish）	94.03%	94.33%	4.022M	8.90
5	GhostEfficientNet_s	93.79%	94.07%	1.862M	6.37

TABLE. V Comparison results on accuracy, F1 score, Params and Time of each model

Model Name	Accuracy	F1-Score	Params	Detection Speed (batch=1)
squeezenet1_1	90.53%	90.96%	724.548K	2.64
MobileNetV3	91.42%	91.99%	4.204M	9.62
GhostNet	92.10%	92.52%	4.207M	11.33
ShuffleNetv2	90.97%	91.48%	2.511M	7.97
EfficientNet_b0	93.13%	93.59%	4.013M	8.88
Proposed Model	93.79%	94.07%	1.862M	6.37

The measured results of key frame extraction show that for a sensitive video of 1 minute and 30 seconds with a total of 1905 frames, the proposed method only extracts 87 key frames to completely cover all sensitive content, with a redundant frame filtering rate of over 95%. For a normal video with 1285 frames, only 153 frames are extracted to fully represent the video content. Compared with the traditional inter-frame difference method, the proposed method can effectively identify content changes in frames without missing sensitive information, greatly reducing the workload of subsequent detection and manual auditing.

V. CONCLUSION

Aiming at the problems existing in video sensitive information detection, such as high false detection and missed detection rates, slow detection speed, and difficulty in adapting to actual platform deployment, this paper proposes a video-level sensitive information detection scheme that combines a lightweight object detection model and a key frame extraction algorithm. Based on YOLOv5s, the sensitive information detection performance is optimized by adding the GPSA attention module and replacing the feature fusion structure with BiFPN. On the basis of the inter-frame difference method, the key frame extraction effect is improved by fusing shallow features in the LUV color space and deep features extracted by GhostEfficientNet_s. The improved detection model achieves an mAP of 71% on the self-built sensitive image dataset, with a detection speed of only 2.8 ms per image. The key frame extraction method can effectively filter more than 95% of video redundant frames. The scheme has been successfully applied to the Shaike Wang social platform, which can efficiently assist the background in completing video auditing and achieves a favorable balance between detection accuracy and operational efficiency.

ACKNOWLEDGMENT

This work was supported by the Science and Technology Research Project of Henan Province (No. 262102210089) and the University-Enterprise Collaborative Innovation Project of Henan Province (No. 26AXQXT029).

REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] Zhu, S. Q., Wang, Y. H. Research on the Development Strategy of Content Security Based on Artificial Intelligence [J]. Engineering Sciences, 2021, 23(3): 67-74.

[3] Zhang, X. Y., Zhang, Y. H. Video Key Frame Extraction Method Based on Fusion Features [J]. Computer Systems & Applications, 2019, 28(9): 176-181.

[4] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018. [5] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[5] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[6] Zhang H, Zu K, Lu J, et al. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network[C]//Proceedings of the Asian Conference on Computer Vision. Macao: Springer, 2022: 1161-1177.

[7] Tan M, Pang R, Le Q V. EfficientDet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10781-10790.

[8] Han K, Wang Y, Tian Q, et al. GhostNet: More features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1580-1589.

[9] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural network[C]//International Conference on Machine Learning. Long Beach: PMLR, 2019: 6105-6114.

[10] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 91-99