The ISC Int'l Journal of Information Security

ISeCure

October 2025, Volume 17, Number 3 (pp. 1–8)

http://www.isecure-journal.org

Presented at the ISCISC'2025 in Tehran, Iran.

Feature Map Correlations in Video Face Forgery Detection **

Seyede Fateme Mazhar [1], and Azadeh Mansouri [2,*]

[1] Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran

[2] Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran

ARTICLE INFO.

Keywords:
Deepfake, Deep Learning, Fake Video Detection, Feature Map, Gram Matrix, Video Manipulation

Type:

doi:

Abstract

Face manipulation techniques have raised widespread public concern. Although conventional convolutional neural networks (CNNs) demonstrate satisfactory performance, they exhibit limitations in capturing the critical features that are directly affected by manipulation artifacts. Fake faces often exhibit unnatural correlations between feature maps, particularly in local patterns. The Gram matrix can illustrate these irregularities in feature map relationships and support discrimination between real and fake images. In this paper, we show that the relationships among convolutional neural network feature maps typically display natural and coherent patterns; in contrast, manipulated faces tend to disrupt this consistency, which can be effectively captured by analysing the Gram matrix of the feature maps. Experimental results demonstrate that the proposed correlation-based features achieve satisfactory performance in both single-dataset and cross-dataset validation.

1 Introduction

Deepfake refers to deep learning-based techniques that are used to generate fake multimedia content. With recent advancements, synthesised videos and images look so realistic that they can be very difficult to distinguish from the real thing. Today, with respect to large datasets and deep networks such as Autoencoders [1] and Generative Adversarial Networks (GANs) [2], creating fake videos has become more convenient. Despite their beneficial use in education, filmmaking, entertainment, and marketing [3], deepfakes are also exploited for malicious purposes. Manipulated videos in which individuals appear to speak inappropriately or spread false information are just some of the examples of malicious usage of deepfakes. These can damage reputations, erode public trust in companies, and even negatively influence elections or financial systems.

As a result, the demand for effective deepfake detection methods has become increasingly critical. These methods generally fall into two main categories: traditional approaches and deep learning-based techniques [4]. Depending on the type of input and the network architecture, different models are used for fake video detection. Some methods are frame-based [5, 6] and only use frame features, while others also leverage temporal information [7–13]. Additionally, certain studies utilise frequency-domain features [14, 15]. Transformer-based models [16, 17] have also demon-

* Corresponding author.

**The ISCISC'2025 program committee effort is highly acknowledged for reviewing this paper.

Email addresses: seyedefateme.mazhar@khu.ac.ir, a_mansouri@khu.ac.ir

ISSN: 2008-2045

strated promising results in deepfake detection.

A key challenge in existing methods is generalisation. These methods experience a notable performance reduction when faced with unseen data. Another issue in video forgery detection is the high computational complexity, which makes the whole process time-consuming. This research aims to address these problems by proposing a method that reduces computational complexity while maintaining high performance and improving generalisability.

Our approach uses enhanced spatio-temporal features. First, frame-level features are extracted from the video using pretrained networks. By selecting the appropriate layer, we obtain feature maps from that layer. The layer selection procedure is critical. Many methods use CNN networks for feature extraction, but the final layer is determined most of the time. In this work, we demonstrate that selecting an intermediate layer yields more relevant features. The final layers often return structural features, while the intermediate layers extract texture, colour, and curvatures from the input data, leading to improved results [18]. Then, a Gram matrix is used to capture correlations between different layers of the network. This process also decreases the input dimensionality in the following step, thereby lowering computational complexity. Next, a pooling method is applied to the result. Finally, a classifier determines whether the video is real or fake.

This research addresses the following areas:

- In this work, by using pretrained networks and selecting an appropriate layer for feature extraction, we demonstrate that intermediate layers provide more suitable features. This approach not only improves the results but also reduces the classifier's input size and ultimately lowers model complexity.
- Our proposed approach uses an enhanced spatio-temporal feature by computing the Gram matrix between feature maps extracted from a CNN. The feature is defined as a higher-order representation of selected frames and demonstrates strong performance in video forgery detection.
- The experiment results indicate that our method, despite the reduction in computational complexity, yields consistent and reliable results and, in some cases, outperforms state-of-the-art methods. The model also performs well in cross-forgery scenarios, showing considerable generalisability.

The remainder of this paper is organised as follows: Section 2 introduces the related work. Section 3 discusses the details of the proposed method. Section 4 describes the implementation details and the results of the proposed approach. Finally, Section 5 is dedicated to the conclusion.

2 Related Work

2.1 Deepfake Detection

With recent advances in deep learning, many tasks, including video forgery detection, have been solved with satisfactory results. Deepfake detection methods are generally categorised into different groups. Some methods were frame-based, which analyse each frame independently to identify anomalies or similarities within a single frame. In [5], mesoscopic (mid-level) features of video frames were examined for forgery detection. [6] It is a model based on the Xception architecture that extracts spatial features from frames. In this work, inputs were fed into the model in two modalities: full-frame images and cropped face regions, and superior results were achieved with the cropped face approach. Some approaches used temporal information in addition to spatial analysis. This information could be detected using recurrent neural networks (RNNs) [7] or optical flow techniques [8–11], which capture inter-frame inconsistencies. Results suggest that optical flow is particularly effective in cross-forgery scenarios. In these scenarios, the training forgery method should differ from the method used in the testing data. Several methods focus on specific regions of video input to detect artifacts resulting from manipulation. In [12, 19, 20], alterations in mouth movement, facial edge regions, and head motions were detected. Additionally, studies such as [13] and [21] use attention mechanisms to filter out insignificant spatial and temporal regions, improving both performance and interpretability. Other studies investigate the use of frequency-domain features. The model proposed in [14], called TwoStreamNet, integrates both spatial and frequency domain information. Similarly, FreqNet [15] focuses on frequency features while aiming to avoid overfitting to training-specific artifacts, thereby enhancing generalisation. More recent methods have used transformer-based architectures [16, 17]. In these approaches, features extracted from video frames are passed into a transformer, leading to the final step of deepfake detection.

2.2 Gram Matrix

The Gram matrix, obtained from the inner product of features extracted from a specific layer of a network, is a tool for understanding and analysing correlations between feature maps. In the area of convolutional neural networks (CNNs), the Gram matrix was previously used in style transfer tasks [22],

where it extracted style information from an image by examining correlations among different network layers. In [18], researchers leverage this matrix for Video Quality Assessment. Furthermore, in [23], a method was proposed for image forgery detection, in which the Gram matrix was used to reveal anomalies in forged images. In this work, we use the Gram matrix for video forgery detection. A more detailed explanation of the Gram matrix will be provided in the following sections.

3 Methodology

3.1 overview

An overall schematic of the proposed method is represented in Figure 1. Initially, video frames should be selected for further processing. Since processing all frames of a video is highly time-consuming, a frame selection strategy should be used. One approach is to select a frame at intervals of ten frames. This reduces the number of frames processed, which in turn reduces runtime and computational complexity. Next, the face extraction is performed on the selected frames. After certain preprocessing steps on facial data, the result is fed into a pre-trained model to obtain feature maps from a specific layer. Subsequently, Gram matrices of extracted feature maps are computed and flattened. Afterwards, a temporal pooling method is applied to the high-level features to incorporate temporal information in the forgery detection process. In this study, max-pooling is applied to the frame-level features of each video. As a result, a single feature vector is obtained for each video and passed to our network's classifier, which performs binary classification into real and fake categories. The details of each step are described in the following sections.

3.2 Feature Extraction

Following data preparation, a suitable pre-trained network must be selected for feature map extraction. In this study, we explore commonly used backbone architectures, including VGG16, VGG19, and ResNet50, all of which have been pre-trained on the ImageNet dataset. In addition, an appropriate feature extraction layer must be chosen. The proper layers are determined based on the results presented in [23]. This paper explored image forgery detection by extracting features from different layers of pretrained networks, and the aforementioned layers returned the best results. We examine these results for video forgery detection. Our investigation revealed that Layer 14 of VGG16, Layer 16 of VGG19, and Layer 81 of ResNet50 yielded superior performance, making them suitable selections for extracting correlation-based features. It should be noted that although our proposed approach leverages findings of [23], their fundamental approaches differ; one is conducted on images, the other focuses on videos, and they are based on different datasets. As a result, a direct comparison of their results is not applicable.

3.3 Gram Matrix

In formal terms [24], if the vectors $\alpha_1, \alpha_2, ..., \alpha_n$ are considered as the columns of matrix A, then the Gram matrix G is an $n \times n$ matrix, given by:

$$G = A^T A \tag{1}$$

This symmetric matrix represents the dot product between the vectors. In principle, the dot product indicates the correlation of two vectors [25].

In terms of CNNs, each layer has N feature maps. If these feature maps are vectorised and assigned as the columns of a matrix A, as below,

$$A = \begin{pmatrix} FM_1 & FM_2 & \cdots & FM_N \end{pmatrix} \tag{2}$$

The Gram matrix, according to the above formula, will represent the dot product of the feature maps of each layer.

$$G_{ij} = \Sigma_k \quad FM_{ik} \cdot FM_{jk} \tag{3}$$

The values of this matrix will indicate the similarity and correlation between each pair of feature maps and define the structural information for each frame. Indeed, capturing structural information and identifying inconsistencies among feature maps are essential for effective manipulation detection.

Since the Gram matrix is symmetric, approximately half of its entries are duplicated. To eliminate redundant values and reduce the input size for the subsequent step, only the lower triangular part of the matrix will be considered. This part is flattened and transformed into a vector afterwards.

3.4 Loss Function

To minimise the difference between p, the predicted probability of the positive class, and y, the true binary label, Binary Cross-Entropy with Label Smoothing is used as a loss function, with the corresponding formula:

$$BCE_{sm}(p, y) = -\left[y_{sm} \log(p) + (1 - y_{sm}) \log(1 - p)\right] \tag{4}$$

where y_{sm} is defined as:

$$y_{sm} = (1 - \epsilon) \cdot y + \frac{\epsilon}{2}$$

ϵ is the smoothing factor, typically small.

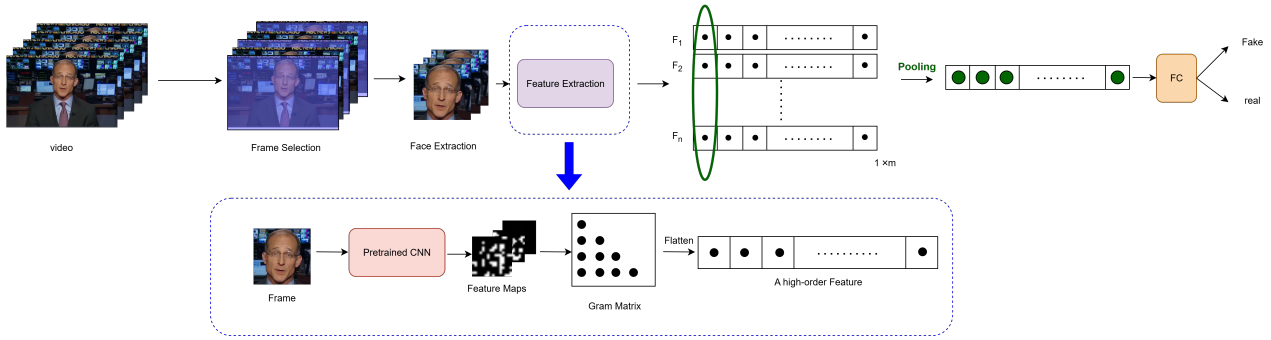

Figure 1. Overview of proposed approach

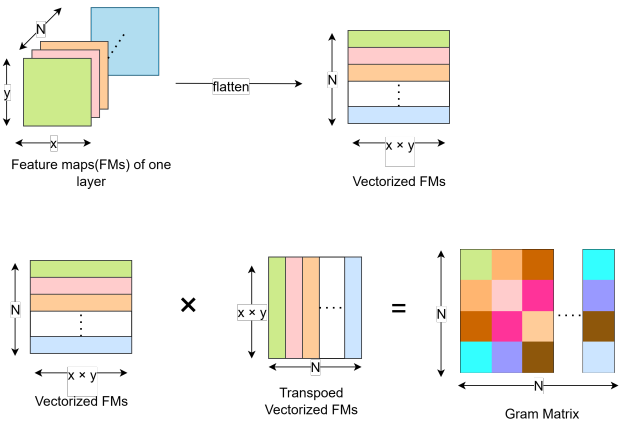

Figure 2. Steps for determining the Gram Matrix from Feature maps of the specified layer

4 Experimental Results

4.1 Experiments details

Datasets. We use the well-known datasets FaceForensics++, CelebDF-v1, and CelebDF-v2 to evaluate our model. FaceForensics++ [6] contains 1,000 original videos and uses four forgery algorithms —DeepFakes, Face2Face, FaceSwap, and NeuralTextures— to generate fake data. Each forgery subset also contains 1,000 videos. This dataset is provided in three different qualities: raw, c23, and c40. In this work, the c23 version is used. The Celeb-DF [26] dataset is available in two versions: CelebDF-v1, with 408 real and 795 fake videos, and CelebDF-v2, with 590 real and 5,639 fake videos. We employ two approaches. For Celebdf, we assume 30% for testing, 30% for validation, and the remaining data are utilised for training. For Faceforensics++, we use 140 for testing, 140 for validation, and 720 for training, following the baseline [6].

Evaluation Metrics. We evaluate our method using common metrics in other works, including Accuracy (ACC), Area Under the Curve (AUC), and F1-score.

Implementation Details. We developed our proposed approach with TensorFlow Keras on a Tesla T4 and an NVIDIA GeForce RTX 2080 Ti GPU. The implementation structure is as follows: the input video frames are cropped into faces using the MTCNN algorithm [27], and the extracted face regions are resized to 256×256×3. The objective is to standardise the format of all data and frames. Subsequently, the data are normalised between 0 and 1. The classifier, which ultimately determines whether a video is real or fake, consists of two fully connected layers. The first layer contains 256 neurons, followed by the second layer with 128 neurons, and the activation function is ReLU. To ensure training stability and decrease the risk of overfitting, Batch Normalisation, Dropout, and kernel regulariser are applied to the layers. The output layer consists of two neurons with a Softmax activation function. While training on FaceForensics++, the model uses the Adam optimizer with a learning rate of 6e-5, and for training on Celeb-DF, the AdamW optimizer is used with a learning rate of 6e-5 and $\beta_1 = 0.85$. A callback is set to dynamically adjust the learning rate; these differences are due to class imbalance in the Celeb-DF dataset. To further reduce the effect of this imbalance, class weights are additionally assigned while fitting the model. It is a built-in feature in Keras. Binary Cross-Entropy with label smoothing- a technique that helps reduce model overconfidence- is used as the loss function. Training is performed for 50 epochs with a batch size of 32.

Table 1 presents the specifications of the backbone CNN models and selected layers, as well as the dimensionality of the aggregated vector produced for each video. The vector is used as input to the classifier section. As shown in the table, selecting intermediate layers for feature extraction simultaneously captures both high-level and low-level features. Moreover, it reduces the input size for subsequent processing steps, thereby lowering the method's overall complexity. The frame-level features are subsequently aggregated using a simple max pooling operator to produce a video-based feature representation.

Table 1. Pretrained network details

DNN	Feature Extraction Layer	Feature Map Size	Aggregated Vector Size
VGG16	14	(512 × 16 × 16)	131,328
VGG19	16	(512 × 16 × 16)	131,328
ResNet50	81	(256 × 16 × 16)	32,896

4.2 Evaluation

Intra-dataset Evaluation. The evaluation results of the intra-dataset are presented in Table 2 for FaceForensics++ and both Celeb-DF versions. As indicated, we use various models as CNN backbones, with ResNet50 architecture achieving the best performance. This performance can be attributed to the fact that the features extracted by ResNet are obtained from deeper layers and therefore contain more complex and detailed information. Furthermore, ResNet uses residual blocks and shortcut connections, which allow more effective feature extraction and ultimately enhance the network's overall performance.

Table 2. Intra-dataset evaluation on Celeb-DF and FaceForensics++ with different CNN backbones

Dataset	Model	Accuracy	AUC	F1-score
Celebdf-v1	VGG16	89.19	95.74	87.86
	VGG19	88.91	93.67	86.69
	ResNet50	100	99.9	100
Celebdf-v2	ResNet50	100	99.9	99.1
FaceForensics++	ResNet50	99.93	99.99	99.89

As reported in Table 3, our approach achieves strong performance while maintaining significantly lower model complexity by selecting frame subsets and using an intermediate layer for feature extraction.

cross-manipulation evaluation. Cross-manipulation refers to scenarios in which the model is trained on a specific manipulation and then evaluated on different manipulations. This represents the model's generalisation, meaning it will be robust against other forgery techniques. It shows the model is not biased by information from a specific manipulation. For this purpose, the model is trained on one of the manipulation methods in FaceForensics++ and tested on the others. The model is executed 15 times, and the best-performing model will be selected. The reason for these multiple executions is to examine the model's performance across varying runs and weight initialisations.

Table 4 presents the accuracy and AUC values of the proposed method in comparison with other existing approaches. As demonstrated, despite the simplicity and low complexity of the proposed model, it shows reasonable performance in the cross-manipulation evaluation. In most cases, the results are in the best or second-best. Specifically, for training on DF, F2F, and NT and testing on FS, the model achieves the best results, showing an increase of 18 %, 4%, and 5%. In addition, when the model is tested on NT and trained on the other three, it shows satisfactory performance. The result drops by only 1% when trained on F2F and DF, while the highest performance is achieved when trained on FS. It is important to note that NT is consistently challenging due to its manipulation technique, which consists of more realistic videos.

The results indicate that the proposed correlation-based feature, although simple compared to other methods, has shown promising results. It should be noted that other methods used features with more complex structures [30] and sophisticated architectures to capture temporal information [34]. It should be noted that although the proposed method relies solely on correlations between feature maps and captures temporal variations using a simple max pooling strategy, its performance remains highly competitive compared to more complex models.

Table 4. Cross-dataset evaluation comparison to other approaches. Training on one manipulation and testing on three others. The best results are highlighted in bold, while the second-best results are underlined.

Train		Test: F2F	DF	NT	FS
F2F	DTN (2024) [28]	99.82	**100**	**97.3**	93.63
	Guan *et al.* (2024) [30]	100	89.10	48.61	62.51
	Shao *et al.* (2024) [34]	-	73.3	72.3	67.7
	D2Fusion (2025) [32]	99.86	89.5	75.23	62.47
	IDDDM (2023) [33]	99.97	99.88	82.38	79.40
	Ours	100	96.98	96.65	**97.91**
DF	DTN (2024) [28]	**97.23**	100	**98.53**	74.26
	Guan *et al.* (2024) [30]	66.72	100	67.40	43.29
	Shao *et al.* (2024) [34]	-	99.57	-	79.51
	D2Fusion (2025) [32]	77.88	99.98	75.73	62.25
	IDDDM (2023) [33]	83.94	100	68.98	58.33
	Ours	96.42	100	97.04	**97.83**
NT	DTN (2024) [28]	**97.95**	99.71	99.81	91.25
	Guan *et al.* (2024) [30]	68.83	90.15	97.84	38.28
	Shao *et al.* (2024) [34]	-	-	-	-
	D2Fusion (2025) [32]	71.08	94.44	99.43	80.75
	IDDDM (2023) [33]	97.93	**100**	99.46	86.76
	Ours	92.83	97.59	100	**96.0**
FS	DTN (2024) [28]	**99.74**	99.15	92.45	98.21
	Guan *et al.* (2024) [30]	66.03	64.19	37.37	99.97
	Shao *et al.* (2024) [34]	-	88.57	-	99.04
	D2Fusion (2025) [32]	69.76	77.50	58.45	99.92
	IDDDM (2023) [33]	74.00	93.42	49.86	99.92
	Ours	97.81	**99.34**	**97.74**	100

5 Conclusion

In this paper, we propose a correlation-based feature representation for face forgery detection, designed to capture inconsistencies between feature maps for

Table 3. Intra-dataset evaluation comparison to other approaches on FaceForensics++, Celebdf-v1, and Celebdf-v2. Training and testing are on the same dataset

	Celebdf-v1			Celebdf-v2			FF++(c23)	
	ACC	AUC		ACC	AUC		ACC	AUC
Javed et al [17]	96.89	-	Javed *et al.* [17]	97.9	-	DTN [28]	97.13	99.7
			CSTAN [29]	-	91.75	Guan et al [30]	-	99.17
CSTAN [29]	-	94.24	JRC [31]	-	97.1	D2Fusion [32]	97.77	99.42
						IDDDM [33]	-	99.7
ours	**100**	**99.0**	**ours**	**100**	**99.9**	**ours**	**100**	**99.9**

manipulation detection. The proposed feature can be extracted using a pre-trained backbone network, making the approach computationally efficient. These extracted correlations reveal unnatural patterns in manipulated faces, enabling the method to achieve strong generalisation. Experimental results demonstrate the robustness and generalisation of the proposed approach. Despite the promising results, there remains room for improvement. The pre-trained backbone was not fine-tuned, although doing so within the context of forgery detection could further enhance performance. Moreover, a basic pooling strategy —max pooling— was employed to aggregate frame-level features into a video-level representation. Future work may benefit from employing more advanced techniques, such as transformer-based pooling, which better capture temporal variations.

References

[1] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[2] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Harshika Goyal, Mohammad Saif Wajid, Mohd Anas Wajid, Akib Mohi Ud Din Khanday, Mehdi Neshat, and Amir Gandomi. State-of-the-art ai-based learning approaches for deepfake generation and detection, analyzing opportunities, threading through pros, cons, and future prospects. *arXiv preprint arXiv:2501.01029*, 2025.

[4] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, 14(5):910–932, 2020.

[5] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[6] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[7] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[8] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[9] Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. Optical flow based cnn for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, 146:31–37, 2021.

[10] Ali Bou Nassif, Qassim Nasir, Manar Abu Talib, and Omar Mohamed Gouda. Improved optical flow estimation method for deepfake videos. *Sensors*, 22(7):2500, 2022.

[11] Akash Chintha, Aishwarya Rao, Saniat Sohrawardi, Kartavya Bhatt, Matthew Wright, and Raymond Ptucha. Leveraging edges and optical flow on faces for deepfake detection. In *2020 IEEE international joint conference on biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[12] Zongmei Chen, Xin Liao, Xiaoshuai Wu, and Yanxiang Chen. Compressed deepfake video detection based on 3d spatiotemporal trajectories. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–8. IEEE, 2024.

[13] Abdelwahab Almestekawy, Hala H Zayed, and Ahmed Taha. Deepfake detection: Enhancing performance with spatiotemporal texture and deep learning feature fusion. *Egyptian Informat-*

ics Journal, 27:100535, 2024.

[14] Bilal Yousaf, Muhammad Usama, Waqas Sultani, Arif Mahmood, and Junaid Qadir. Fake visual content detection using two-stream convolutional neural networks. *Neural Computing and Applications*, 34(10):7991–8004, 2022.

[15] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[16] Yang Yu, Rongrong Ni, Yao Zhao, Siyuan Yang, Fen Xia, Ning Jiang, and Guoqing Zhao. Msvt: Multiple spatiotemporal views transformer for deepfake video detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4462–4471, 2023.

[17] Muhammad Javed, Zhaohui Zhang, Fida Hussain Dahri, and Asif Ali Laghari. Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach. *Electronics*, 13(15):2947, 2024.

[18] Amir Hossein Bakhtiari and Azadeh Mansouri. Feature maps correlation-based video quality assessment. *Multimedia Tools and Applications*, 83(23):63309–63328, 2024.

[19] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021.

[20] Dong-Keon Kim and Kwang-Su Kim. Generalized facial manipulation detection with edge region feature extraction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2828–2838, 2022.

[21] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932, 2021.

[22] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[23] Narges Honarjoo, Fatemeh Taher, and Azadeh Mansouri. Exploring feature map correlations for effective fake face detection. In *2024 11th International Symposium on Telecommunications (IST)*, pages 200–204. IEEE, 2024.

[24] H Schwerdtfeger. *Introduction to Linear Algebra and the Theory of Matrices*. P. Noordhoff, Groningen, 1950.

[25] Petros Drineas and Michael W Mahoney. Approximating a gram matrix for improved kernel-based learning. In *International Conference on Computational Learning Theory*, pages 323–337. Springer, 2005.

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[27] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[28] Yaning Zhang, Qiufu Li, Zitong Yu, and Linlin Shen. Distilled transformers with locally enhanced global representations for face forgery detection. *Pattern Recognition*, 161:111253, 2025.

[29] Rui Yang, Kang You, Cheng Pang, Xiaonan Luo, and Rushi Lan. Cstan: A deepfake detection network with cst attention for superior generalization. *Sensors*, 24(22):7101, 2024.

[30] Weinan Guan, Wei Wang, Jing Dong, and Bo Peng. Improving generalization of deepfake detectors by imposing gradient regularization. *IEEE Transactions on Information Forensics and Security*, 19:5345–5356, 2024.

[31] Bosheng Yan, Chang-Tsun Li, and Xuequan Lu. Jrc: Deepfake detection via joint reconstruction and classification. *Neurocomputing*, 598:127862, 2024.

[32] Xueqi Qiu, Xingyu Miao, Fan Wan, Haoran Duan, Tejal Shah, Varun Ojha, Yang Long, and Rajiv Ranjan. D2fusion: Dual-domain fusion with feature superposition for deepfake detection. *Information Fusion*, 120:103087, 2025.

[33] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3994–4004, 2023.

[34] Rui Shao, Tianxing Wu, Liqiang Nie, and Ziwei Liu. Deepfake-adapter: Dual-level adapter for deepfake detection. *International Journal of Computer Vision*, 133(6):3613–3628, 2025.

Seyede Fateme Mazharreceived her B.Sc. degree in Software Engineering from Zanjan University(ZNU), Zanjan, Iran, and M.Sc. in Artificial Intelligence (AI) from Kharazmi University (KhU), Tehran, Iran. Her field of interest includes image/video processing, security, and AI in cybersecurity.

Azadeh Mansouri received the B.S. degree in computer engineering from Shiraz University, Shiraz, Iran, in 2002, the M.S. degree in Science and Research Branch, Azad University of Tehran, and the PhD degree from the Department of Electrical and Computer Engineering, Shahid Beheshti University, Tehran, Iran, in 2010. In 2011, she joined the Department of Electrical and Computer Engineering, Kharazmi University, Tehran, Iran, as an assistant professor. Her research interests include image/video processing and machine learning algorithms with an emphasis on digital forensics, quality assessment, and enhancement methods.