Illumination Enlightened Spatial-temporal Inconsistency for Deepfake Video Detection

Kaiyue Tian[1], Chen Chen[2], Yichao Zhou[1], and Xiyuan Hu*[1, 3]

[1]Nanjing University of Science and Technology, Nanjing, China
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Beijing Visystem Co. Ltd, Beijing, China

2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10687905

***Abstract*—The rapid advancement of facial manipulation techniques has greatly simplified the creation of deepfake videos, posing a major threat to social safety, public opinions and even political stability. Existing deepfake detection methods primarily concentrate on capturing spatial artifacts or extracting uniform temporal inconsistency, neglecting the potential of exploiting dynamic spatiotemporal inconsistency. To address these issues, this paper proposes a novel network that effectively leverages dynamic spatiotemporal inconsistency, termed DSTI, by integrating the sequential illumination features and intra/inter-frame clues. The proposed DSTI contains two branches: one branch employs a transformer encoder to perform inconsistency computation from sequential illumination representations derived from 3D facial models, including illumination coefficients, 3D normal vectors, and luminance values. The other branch utilizes a timesformer network to capture intra/inter-frame inconsistency from sampled videos. Extensive experimentation validates that the proposed method outperforms other competitive approaches.**

***Index Terms*—Deepfake detection, face forgery detection, illumination inconsistency, digital feature inconsistency**

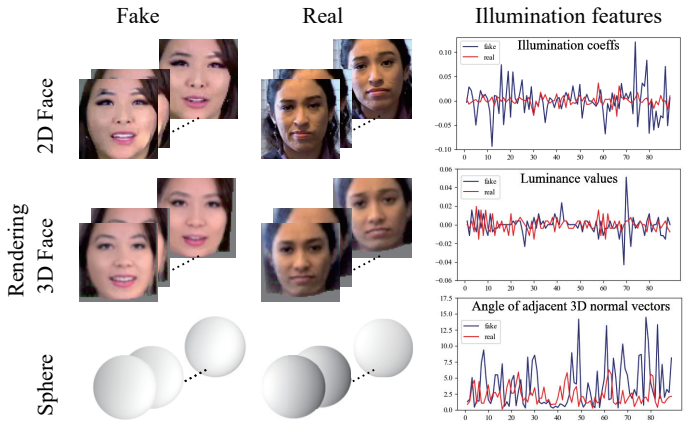


Fig. 1. Illustration of the effectiveness of illumination inconsistency.

I. Introduction

Deepfake technology has significantly simplified the creation of high-quality videos, making some of them challenging to distinguish visually. Therefore, establishing efficient methods to detect deepfake videos becomes crucial.

Due to the difficulty of accurately describing the dynamic changes between adjacent frames in deepfake technology, deepfake videos exhibit flickering and discontinuity. Therefore, researchers have proposed various video-based deepfake detection methods to potentially capture the spatiotemporal inconsistency, such as eye inconsistency, illumination inconsistency, and digital feature inconsistency, among others. Deepfake detection methods based on illumination inconsistency involve comparing the lighting conditions in different regions of an image to identify inconsistency [1], [2]. These methods are mostly applied to faces from multiple people within one image and are not suitable for single-person images. Meanwhile, methods based on digital feature inconsistency primarily focus on capturing spatial-related artifacts or extracting temporal inconsistency between frames [3], [4], [5]. However, they fail to correctly capture the dynamic spatiotemporal clues hidden in deepfake videos, resulting in poor generalization and vulnerability to attacks.

In our view, the key factor in identifying illumination inconsistency lies in the differences in lighting information between consecutive frames, as shown in Fig. 1. Therefore, we propose a novel framework that flexibly utilizes both the illumination information and intra/inter-frame facial region features. In the illumination consistency module, it starts by extracting global lighting information (illumination coefficients) and local lighting information (3D normal vectors, luminance values) from facial frame sequences, constructing a more comprehensive spatiotemporal representation of lighting. Subsequently, it explores the relationships within and between channels of local lighting information in each frame, acquiring knowledge from these relationships and global lighting information. This knowledge is then input into the transformer encoder to learn the consistent representation of dynamic spatiotemporal clues in illumination information. In the intra/inter-frame consistency module, it begins by extracting T_2 frames from facial frame sequences, followed by facial patch segmentation with a resolution of $P * P$ for each frame. These facial patches are then fed into the timesformer model to learn the consistent representation of dynamic spatiotemporal clues in local facial

This work was supported by the National Key R&D Program of China (2021YFF0602101) and National Natural Science Foundation of China (62172227).

* Corresponding author.

regions.

DSTI aims to address both illumination inconsistency and digital feature inconsistency, resulting in more accurate and generalizable results in deepfake detection. Our contributions can be summarized as:

- This paper innovatively explores the temporal variations of illumination information to achieve single-person video illumination inconsistency detection. Additionally, it integrates local and global illumination information into deepfake video detection, comprehensively capturing the dynamic spatiotemporal inconsistency in illumination information.
- Regarding deepfake video detection, we propose a novel dynamic spatiotemporal inconsistency network (DSTI) that captures richer and more robust manipulation clues by utilizing two different yet complementary aspects: sequential illumination information and intra/inter-frame information.
- Experimental results indicate that our model achieves highly competitive performance compared to the state-of-the-art methods.

II. RELATED WORK

A. Illumination inconsistency

The work [1] pioneers the application of lighting direction cues in image forensics, estimating 2D light source direction from object contour brightness and normals, and validating its consistency. The work [6] utilizes spherical harmonics (SH) to assess complex lighting environments. The work [2] presents a 3D face model reconstruction from chosen images, estimating lighting SH coefficients using 3D normal vectors. When dealing with low-resolution and highly compressed images, lighting-based methods effectively expose image forgery. However, these techniques are designed for multi-person forged images or videos and are not suitable for single-person instances. Single-person forged images lack contrasting portraits, and acquiring 3D normal vectors on complex object surfaces is full of challenges. In comparison to these approaches, we consider the dynamic variations in lighting conditions, learning the inconsistency of lighting changes by extracting lighting information from consecutive frames.

B. Video-based deepfake detection

Early video-based deepfake detection methods rely on general video analysis models like C3D [7] and LSTM [8]. CNNs has achieved remarkable success in many fields [9]. The work [10] employs 3D CNNs to represent spatiotemporal information. The work [4] initially extracts frame-level CNN features and then employs transformer to capture temporal clues. However, these models struggle to accurately capture the dynamic spatiotemporal inconsistency across different frames in deepfake videos, showcasing poor generalization and susceptibility to attacks. Compared to these methods, we integrate spatial and temporal information into a unified framework to capture more nuanced and comprehensive spatiotemporal details.

III. METHOD

A. Illumination consistency module

To maintain generality, we use $[v^1,\ldots,v^{T_1}]_{t=1}^{T_1} \in \mathbb{R}^{T_1*C_v*H*W}$ to represent the sampled facial frame sequences from the input video, $[n^1,\ldots,n^{T_1}]_{t=1}^{T_1} \in \mathbb{R}^{T_1*3*H_r*W_r}$ to represent the set of 3D normal vectors, $[r^1,\ldots,r^{T_1}]_{t=1}^{T_1} \in \mathbb{R}^{T_1*C_r*H_r*W_r}$ to represent the set of 3D reconstructed face images, $[l^1,\ldots,l^{T_1}]_{t=1}^{T_1} \in \mathbb{R}^{T_1*1*H_r*W_r}$ to represent the set of luminance values and $[I^1,\ldots,I^{T_1}]_{t=1}^{T_1} \in \mathbb{R}^{T_1*9}$ to represent the set of illumination coefficients, where T_1 represents the number of frames, H and W denote the height and width of the sampled image, H_r and W_r represent the height and width of the 3D reconstructed image, C_v and C_r represent the number of channels in the sampled image and 3D reconstructed image, respectively.

As shown in the upper-left part of Fig. 2, we utilize a 3D face reconstruction method [11] to transform the sampled frame v^t into r^t and n^t. Subsequently, r^t is converted from an RGB image to an HSV image, and only the V value denoted by l^t is extracted. Finally, I^t is obtained through the reflection model. The reflection model can be adequately approximated using the first two-order moments of SH [2], [6]:

$$\begin{aligned} l^t(\overrightarrow{X}) &= l^t(\overrightarrow{n}^t(\overrightarrow{X})) \\ &= \sum_{n=0}^{2}\sum_{m=-n}^{+n} \hat{r}_n {I^t}_{n,m} Y_{n,m}(\overrightarrow{n}^t(\overrightarrow{X})), \end{aligned} \tag{1}$$

where $\overrightarrow{X}$ is a point on r^t, $Y_{n,m}$ is the n-th order m-th SH, $\hat{r}_n$ is the n-th order SH coefficient, and ${I^t}_{n,m}$ is the SH coefficient corresponding to the lighting environment.

Based on the assumptions in [2], DSTI excludes the regions of the mouth, eyes, nose and forehead. We select M sampling points in the cheek and chin areas to estimate I^t using the method outlined in [2]. To obtain more comprehensive lighting information, we select N points from the cheek and chin region, utilizing their 3D normal vectors n_N^t and luminance values l_N^t to create local lighting information I_L^t.

The processing workflow for illumination information is as follows: to thoroughly analyze I_L^t, it undergoes Internal Feature Extraction (IFE) to process channel-wise information, followed by Cross Feature Extraction (CFE) to handle inter-channel information. This process ultimately yields the representation of local lighting information e_L^t, formulated as:

$$e_L^t = CFE(IFE(I_L^t)). \tag{2}$$

We concatenate the global lighting information I^t with e_L^t, and then process them through MLP_1 to generate the lighting information representation e^t for frame t, formulated as:

$$e^t = MLP_1(cat(I^t, e_L^t)). \tag{3}$$

Concatenating e^t for each frame, the transformer encoder is employed to learn the consistent representation I_B of dynamic spatiotemporal clues in the illumination information, formulated as:

$$I_B = Encoder(\{e^t\}_{t\in T_1}). \tag{4}$$

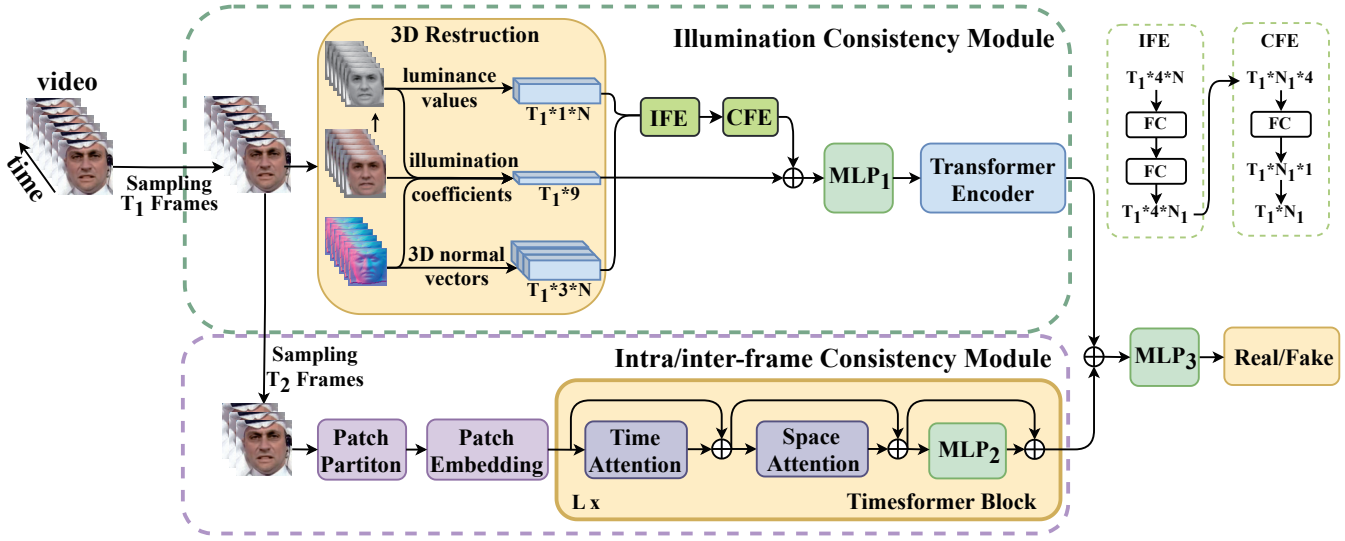


Fig. 2. The overview of the proposed framework. It consists of an illumination consistency module and an intra/inter-frame consistency module. The input is the facial sequence to be detected (T × 3 × H × W), and the output is the detection result.

B. Intra/Inter-frame consistency module

Given the entire facial sequences as input, the intra/inter-frame consistency module aims to learn a consistent representation, describing dynamic spatiotemporal clues in local facial regions across different frames. In this module, timesformer is employed as the backbone because it avoids any spatial or temporal subsampling, which is advantageous for capturing fine-grained dynamics in each block of the facial region.

DSTI extracts T_2 frames from the facial frame sequences as the input for intra/inter-frame consistency module. Each frame is divided into facial patches of size $P * P$. Through the time attention, space attention and MLP_2, the module adequately learns the spatiotemporal features of the face. The time attention is calculated from all blocks at the same spatial position in other frames, while the space attention is calculated from all blocks in the same frame. Consequently, the intra/inter-frame consistency module can effectively obtain a consistent representation F_B of dynamic spatiotemporal clues within frames. The overall process is illustrated in (5):

$$F_B = TI_{timesformer}(\{v^t\}_{t\in T_2}). \quad (5)$$

To comprehensively capture video forgery traces by integrating lighting information and intra/inter-frame clues, the combination of F_B and I_B is processed through MLP_3 to obtain the final scoring situation B, as shown in (6):

$$B = MLP_3(cat(I_B, F_B)). \quad (6)$$

Loss Function. We use cross-entropy as the loss function, defined as:

$$L_{ce} = y \cdot \log(p) + (1-y) \cdot log(1-p), \quad (7)$$

where p is the probability of a positive prediction and y is the label of the sample video.

IV. EXPERIMENTS

A. Experiment setting

Datasets. Faceforensics++ (FF++) [12] is a large-scale benchmark for deepfake detection, comprising four forgery techniques: Deepfakes (DF), Face2Face (F2F), FaceSwap (FS) and NeuralTextures (NT). These videos encompass both high-quality (HQ) and low-quality (LQ) versions. **Celeb-DF [13]** is a deepfake detection dataset with high visual quality and fewer perceptible visual artifacts. **DFDC [14]** is released by the Facebook Deepfake Detection Challenge, using various generation methods.

Implementation Details. We employ Dlib as a face detector to extract faces and then use Mediapipe to extract 3D facial keypoints. The model is trained on Nvidia A30 GPU. For each video, we use T_1=90 frames, T_2=15 frames, H=W=128, P=16, M=11, N=44 and N_1=64 for training and testing. We utilize the Adam optimizer with an initial learning rate of 0.0002, which is decayed by a factor of 5 every 20 iterations. The model undergoes 60 epochs of training with a batch size of 32. We include more details in the appendix.

B. State-of-the-art comparison

To assess the effectiveness of the proposed DSTI, intra-dataset evaluation is conducted on FF++ HQ, FF++ LQ, Celeb-DF and DFDC datasets. The results on FF++ HQ and FF++ LQ are shown in Table I. Compared to frame-based detection methods, our model achieves the best performance. Due to the difficulty in detecting traces of forgery caused by compressed fake videos in FF++HQ and FF++ LQ, we extend deepfake detection into the spatiotemporal domain, enabling our framework to obtain richer temporal information. Compared to video-based detection methods, DSTI achieves the highest AUC among current advanced video deepfake

TABLE I
INTRA-DATASET EVALUATIONS ON FF++

Method		FF++ HQ		FF++ LQ	
		ACC	AUC	ACC	AUC
Frame	LRL [15]	97.59	99.46	91.47	95.21
	PEL [16]	97.63	99.32	90.52	94.28
	SIA [17]	97.64	99.35	90.23	93.45
	RECCE [18]	97.06	99.32	91.03	95.02
	MC-LCR [19]	97.89	99.65	88.07	90.28
	M2TR [20]	97.93	99.42	92.89	95.31
Video	Lips [21]	98.80	99.70	94.20	98.10
	STIL [22]	98.57	-	94.82	-
	Intra-SIM [23]	98.93	-	96.78	-
	HCIL [24]	**99.01**	-	96.78	-
	MRL [25]	93.82	98.27	91.81	96.18
	MRE-Net [26]	97.76	99.57	91.60	96.55
	CDIN [27]	-	98.50	-	96.80
	ours	98.95	**99.97**	**98.05**	**98.90**

detection algorithms. Particularly, on FF++ LQ, our model outperforms Lips [21] by 0.8%. The underlying reason is that our model simultaneously focuses on the dynamic inconsistency of illumination information and intra/inter-frame information, making it more suitable for deepfake detection tasks.

TABLE II
INTRA-DATASET EVALUATIONS ON CELEB-DF AND DFDC

Method	Celeb-DF	DFDC
SIA [17]	**99.96**	90.96
RECCE [18]	99.94	-
M2TR [20]	99.80	-
Lips [21]	-	73.50
STIL [22]	99.78	89.80
Intra-SIM [23]	99.61	92.79
HCIL [24]	99.81	95.11
MRL [25]	**99.96**	99.11
MRE-Net [26]	-	99.75
CDIN [27]	99.70	-
ours	99.77	**99.94**

TABLE III
CROSS-DATASET EVALUATIONS IN TERM OF AUC

Method	FF++ LQ	Celeb-DF	DFDC
PEL [16]	94.28	69.18	63.31
SIA [17]	93.45	77.35	-
RECCE [18]	95.02	68.71	69.06
MC-LCR [19]	90.28	71.61	71.34
M2TR [20]	95.31	68.20	-
STIL [22]	94.82	75.58	67.88
Intra-SIM [23]	98.19	77.65	68.43
HCIL [24]	98.32	79.00	69.21
MRL [25]	96.18	83.58	**71.53**
ours	**98.90**	**87.34**	69.38

The evaluation results on Celeb-DF and DFDC are presented in Table II. On the DFDC dataset, our framework exhibits a 0.19% improvement over the best-performing network, MRE-Net [26]. On the Celeb-DF dataset, our AUC, although not surpassing the best network, maintains a marginal difference of just 0.2%. The consistency of experimental results and the transcendent performance validate the superiority of our proposed network in deepfake video detection.

TABLE IV
CROSS-MANIPULATION EVALUATIONS IN TERM OF AUC

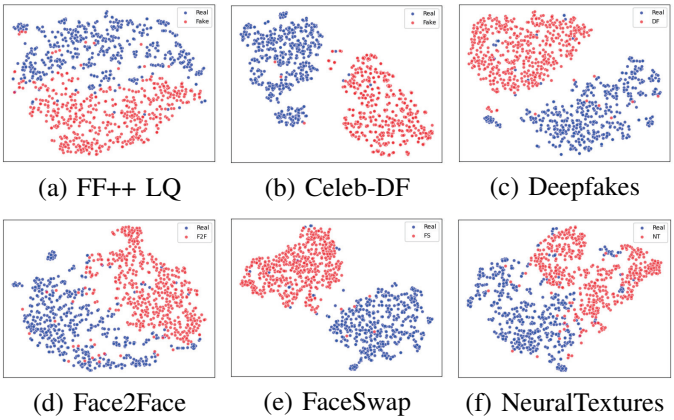
Train	Method	Test			
		DF	F2F	FS	NT
DF	FTCN [28]	99.30	76.00	53.50	87.40
	Lips [21]	99.70	75.70	36.60	90.80
	MRE-Net [26]	**99.99**	87.45	49.18	81.08
	CDIN [27]	99.80	79.90	54.20	88.10
	ours	99.30	**96.70**	**95.94**	**96.41**
F2F	FTCN [28]	81.70	98.90	76.40	85.90
	Lips [21]	88.40	99.20	72.30	91.80
	MRE-Net [26]	91.21	**99.96**	52.13	83.25
	CDIN [27]	84.50	99.60	70.50	91.90
	ours	**96.12**	99.31	**95.82**	**98.50**
FS	FTCN [28]	88.10	69.7	98.90	76.70
	Lips [21]	54.00	68.60	99.50	38.40
	MRE-Net [26]	73.35	63.56	**99.96**	55.12
	CDIN [27]	78.10	67.60	99.80	77.20
	ours	**95.56**	**95.79**	99.56	**95.59**
NT	FTCN [28]	90.00	88.10	62.40	98.00
	Lips [21]	95.50	89.00	52.00	96.50
	MRE-Net [26]	**96.76**	49.90	85.36	99.28
	CDIN [27]	92.90	90.10	62.20	**99.80**
	ours	96.37	**98.31**	**96.29**	98.59

C. Generalization evaluation

Cross-dataset Evaluation. Our model, trained on FF++ LQ, is evaluated separately on Celeb-DF and DFDC datasets. The results based on AUC metrics are shown in Table III. In the cross-dataset evaluation on Celeb-DF, we achieve an AUC of 87.34%, surpassing the latest methods HCIL [24] and MRL [25] by 8.34% and 3.76%, respectively. On the DFDC dataset, we obtain an AUC of 69.38%, trailing behind MRL but outperforming the video-based HCIL by 0.17%. By employing frame-by-frame facial forgery techniques on the FF++ LQ and Celeb-DF datasets, the detector exhibits better generalization compared to the DFDC dataset.

Cross-manipulation Evaluation. We train our model using forged videos generated by one method and test it with all four methods. The results on HQ quality are presented in Table IV, demonstrating that our model outperforms competitors in most cases. Our approach slightly lags behind the best method on the diagonal results, but the difference remains a small gap within only 1.5%. For the off-diagonal results, this study achieves competitive outcomes across all four settings. In settings such as DF, F2F, and FS, our method demonstrates the best performance in all cross-manipulation evaluations.

Visualizations. To highlight our approach's representation capability, we use t-SNE [29] to visualize learned representations from DSTI on the Celeb-DF dataset, FF++ LQ dataset and its subsets. In Fig. 3, clear clusters in the latent space demonstrate the strong representation ability of our approach with real and fake videos generated by different methods. We include more details in the appendix.



(a) FF++ LQ (b) Celeb-DF (c) Deepfakes

(d) Face2Face (e) FaceSwap (f) NeuralTextures

Fig. 3. Visualization of features on (a) FF++ LQ, (b) Celeb-DF, (c) Deepfakes, (d) Face2Face, (e) FaceSwap, (f) NeuralTextures via t-SNE. Red: Fake, Blue: Real.

TABLE V
ABLATION STUDY OF THE PROPOSED COMPONENTS IN TERM OF ACC. 3DNV: 3D NORMAL VECTORS, LV: LUMINANCE VALUES, IC: ILLUMINATION COEFFICIENTS, IMAGE: FACIAL SEQUENCES.

Illumination			Image	Dataset		
3DNV	LV	IC		FF++ LQ	Celeb-DF	DFDC
✓				94.72	91.15	89.11
	✓			94.57	89.61	94.41
		✓		93.21	85.00	85.19
✓	✓			94.72	90.38	92.17
✓		✓		94.26	91.07	92.73
	✓	✓		93.96	91.92	93.85
✓	✓	✓		97.33	91.92	94.97
			✓	97.51	95.00	93.02
✓	✓	✓	✓	**98.05**	**98.08**	**98.60**

D. Ablation study

To show the impact of each component of our network, we train the proposed model using different combinations. As shown in Table V, the model is trained and evaluated on the FF++ LQ, Celeb-DF, and DFDC datasets. Comprehensive ablation experiments are conducted specifically for the illumination consistency module. The model's performance is notably lower when only global illumination information or only individual local illumination information is considered. Combining global and local illumination information leads to a slight improvement in model performance (over 1.5% for each class). Clearly, both local and global illumination information are crucial, and their combination demonstrates the helpfulness of illumination information in deepfake detection. Furthermore, the results indicate that simultaneously utilizing illumination information and intra/inter-frame clues enhances the model's performance in deepfake detection.

V. CONCLUSION

In this paper, we propose a novel network that leverages the dynamic spatiotemporal inconsistency between sequential illumination features and intra/inter-frame clues to achieve more accurate and generalized detection results. This approach, on one hand, fully utilizes the temporal characteristics of illumination information by considering both local and global lighting information. On the other hand, it extensively captures the temporal features of forgery traces through facial frame sequences. Extensive ablation studies demonstrate the effectiveness of our proposed method. Simultaneously, our approach exhibits significant improvements in performance and robustness compared to previous video-based deepfake detection methods, particularly in challenging low-quality scenarios.

REFERENCES

[1] Micah K Johnson and Hany Farid, "Exposing digital forgeries by detecting inconsistencies in lighting," in Proceedings of the 7th workshop on Multimedia and security, 2005, pp. 1–10.

[2] Eric Kee and Hany Farid, "Exposing digital forgeries from 3-d lighting environments," in 2010 IEEE International Workshop on Information Forensics and Security. IEEE, 2010, pp. 1–6.

[3] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," Interfaces (GUI), vol. 3, no.1, pp. 80–87, 2019.

[4] Sohail Ahmed Khan and Hang Dai, "Video transformer for deepfake detection with incremental learning," in Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 1821–1828.

[5] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang, "Wilddeepfake: A challenging realworld dataset for deepfake detection," in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2382–2390.

[6] Micah K Johnson and Hany Farid, "Exposing digital forgeries in complex lighting environments," IEEE Transactions on Information Forensics and Security, vol. 2, no. 3, pp. 450–461, 2007.

[7] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "C3d: generic features for video analysis," CoRR, abs/1412.0767, vol. 2, no. 7, pp. 8, 2014.

[8] Sepp Hochreiter and J¨urgen Schmidhuber, "Long shortterm memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] Qiwei Xie, Xiyuan Hu, Lei Ren, Lianyong Qi, and Zhao Sun, "A binocular vision application in iot: Realtime trustworthy road condition detection system in passable area," IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 973–983, 2022.

[10] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen, "Exploring temporal coherence for more general video face forgery detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15044–15054.

[11] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi, "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1–10.

[12] 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, 2019, pp. 1–11.

[13] 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, 2020, pp. 3207–3216.

[14] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer, "The deepfake detection challenge (dfdc) preview dataset," arXiv preprint arXiv:1910.08854, 2019.

[15] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji, "Local relation learning for face forgery detection," in Proceedings of the AAAI conference on artificial intelligence, 2021, vol. 35, pp. 1081–1088.

[16] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi, "Exploiting fine-grained face forgery clues via progressive enhancement learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, vol. 36, pp. 735–743.

[17] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji, "An information theoretic approach for attention-driven face forgery detection," in European Conference on Computer Vision. Springer, 2022, pp. 111–127.
[18] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang, "End-to-end reconstruction-classification learning for face forgery detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4113–4122.
[19] Gaojian Wang, Qian Jiang, Xin Jin, Wei Li, and Xiaohui Cui, "Mc-lcr: Multimodal contrastive classification by locally correlated representations for effective face forgery detection," Knowledge-Based Systems, vol. 250, pp. 109114, 2022.
[20] J. Wang, Z. Wu, and et al., "M2tr: Multi-modal multiscale transformers for deepfake detection," in Proceedings of the 2022 international conference on multimedia retrieval, 2022, pp. 615–623.
[21] 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, 2021, pp. 5039–5049.
[22] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 3473–3481.
[23] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, vol. 36, pp. 744–752.
[24] Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma, "Hierarchical contrastive inconsistency learning for deepfake video detection," in European Conference on Computer Vision. Springer, 2022, pp. 596–613.
[25] Ziming Yang, Jian Liang, Yuting Xu, Xiao-Yu Zhang, and Ran He, "Masked relation learning for deepfake detection," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1696–1708, 2023.
[26] Guilin Pang, Baopeng Zhang, Zhu Teng, Zige Qi, and Jianping Fan, "Mre-net: Multi-rate excitation network for deepfake video detection," IEEE Transactions on Circuits and Systems for Video Technology, 2023.
[27] Hanyi Wang, Zihan Liu, and Shilin Wang, "Exploiting complementary dynamic incoherence for deepfake video detection," IEEE Transactions on Circuits and Systems for Video Technology, 2023.
[28] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen, "Exploring temporal coherence for more general video face forgery detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15044–15054.
[29] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," Journal of machine learning research, vol. 9, no. 11, 2008.