Face Forgery Video Detection via Temporal Forgery Cue Unraveling

Zonghui Guo[1,†] Yingjie Liu[1,†] Jie Zhang[2,3] Shiguang Shan[2,3] Haiyong Zheng[1]

1College of Electronic Engineering, Ocean University of China, Qingdao, China

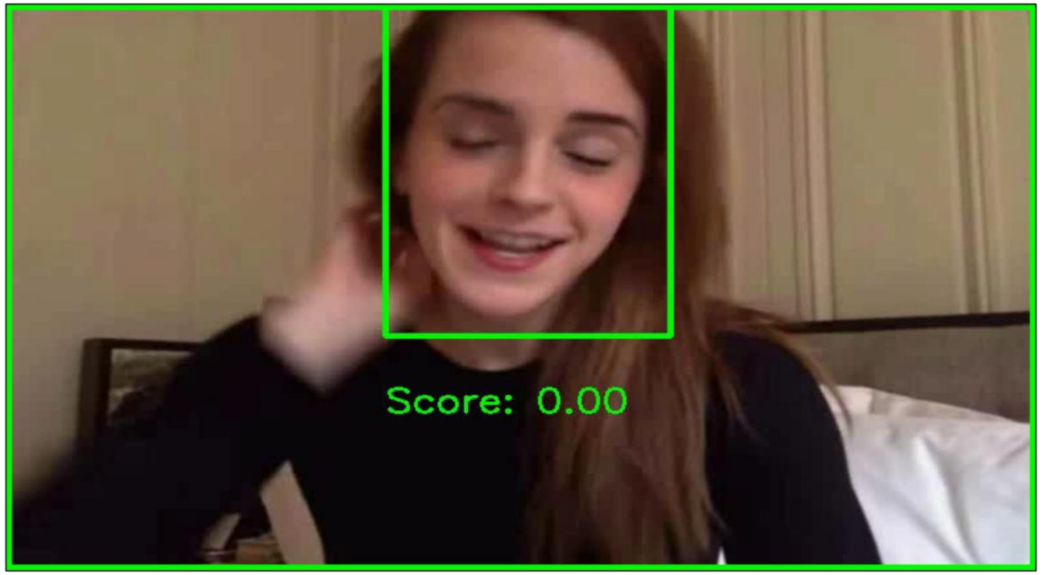
2Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

3University of Chinese Academy of Sciences, Beijing, China

Introduction

Goal

➢ **Face Forgery Video Detection (FFVD)** is a critical yet challenging task in determining whether a digital facial video is authentic or forged.



Introduction

Motivation & Novelty

- Existing FFVD methods typically focus on **isolated spatial features or coarsely fused** spatiotemporal information, failing to leverage temporal forgery cues, resulting in unsatisfactory performance.
- Based on an analysis of the inherent stealth of temporal cues and inspired by the human discrimination process, we abstract temporal forgery cues into three progressive levels: **momentary anomaly, gradual inconsistency, and cumulative distortion**.

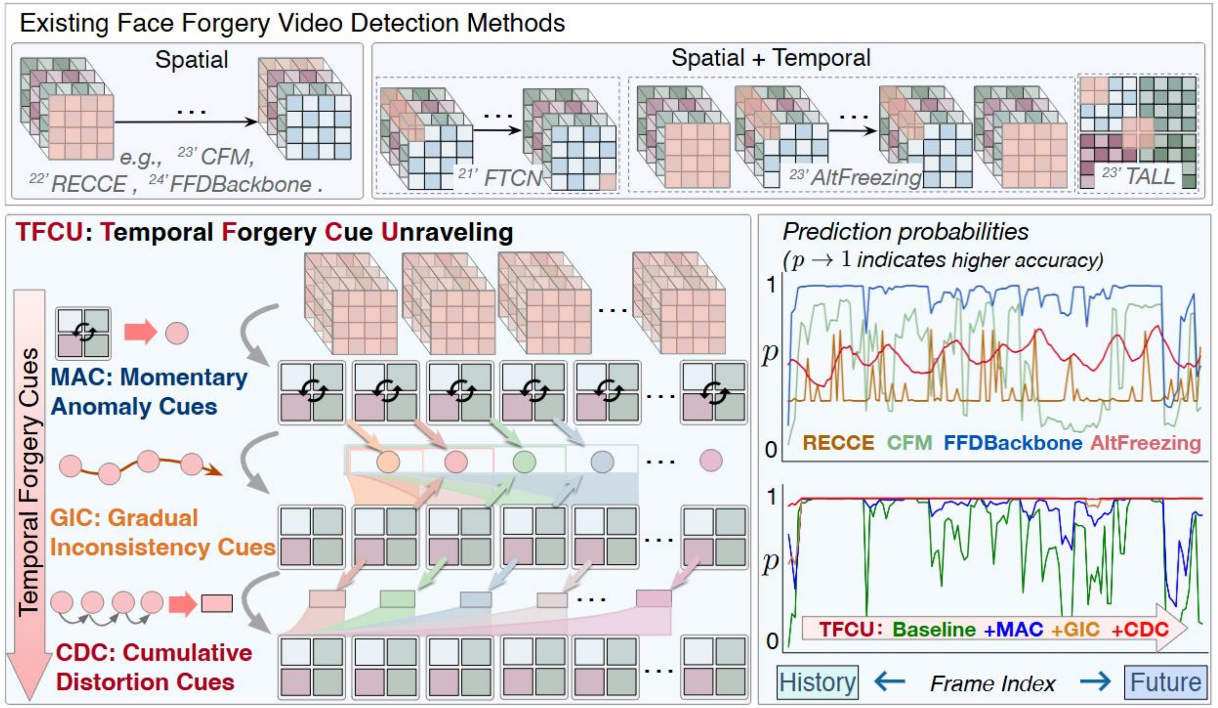
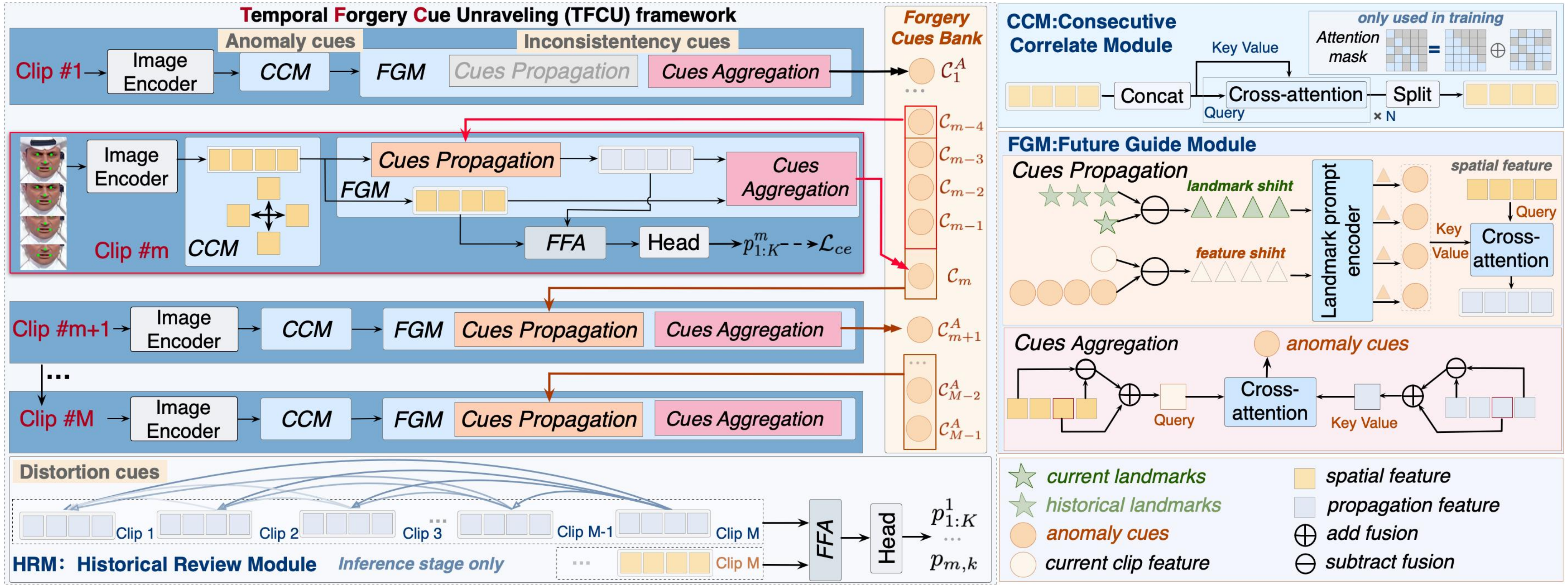


Figure 1. Temporal forgery cues are coarsely explored by adjusting 3D CNNs' kernels or combining frames (*top-right*), leading to significant inter-frame prediction fluctuations (*middle-right*). In contrast, our TFCU meticulously unravels these cues in three progressive levels: momentary anomaly, gradual inconsistency, and cumulative distortion, highlighting general forgery features bidirectionally between historical and future frames (*bottom-left*), thereby achieving stable and precise predictions (*bottom-right*).

Framework

- We strive to unravel these cues across three progressive levels: **momentary anomaly**, **gradual inconsistency**, and **cumulative distortion**.
- We design the Temporal Forgery Cue Unraveling (TFCU) framework to sequentially highlight spatially discriminative features by bidirectionally unraveling temporal forgery cues between **historical and future frames**.



1. Consecutive Correlate for Anomaly Cues

- We propose a consecutive correlate module to capture momentary anomaly cues by correlating interactions among consecutive frames.
- We design the cross-attention module for inter-frame interaction with a unit frame-wise lower triangular mask and random masking for the cross-attention weight w .

$$M_{ij}^{d} = \begin{cases} 1, & \lceil \frac{i}{N} \rceil \geq \lfloor \frac{j}{N} \rfloor \\ -\infty, & \lfloor \frac{i}{N} \rfloor < \lceil \frac{j}{N} \rceil \end{cases}, M_{ij}^{r} = \begin{cases} 1, & p \\ -\infty, & 1-p \end{cases} \qquad w' = w \odot (M_d + M_r)$$

2. Future Guide for Inconsistency Cues

- We devise a future guide module to unravel inconsistency cues by iteratively aggregating historical anomaly cues and gradually propagating them into future frames.

- **Anomaly Cue Aggregation**: For subsequent clips, $\mathbf{E_{ca}}$ takes features from consecutive correlation module (f^{cc}) and output from inconsistency cue propagation (f^{ip}) as input to aggregate anomaly cues $(\mathcal{C}_m)$.

$$\mathcal{C}_m = \mathbf{E}_{ca}\left(f_m^{cc}, f_m^{ip}\right), f_m^{cc} = f_{\lceil \frac{K}{2} \rceil}^{cc} + \left(f_K^{cc} - f_1^{cc}\right)$$

- **Inconsistency Cue Propagation**: We leverage $\mathbf{E_{cp}}$ to propagate forgery cues by interacting current clip $f_{m,k}^{cc}$ with nearest T historical $\mathcal{C}$. In addation, we encode the Landmark shifts using $\mathbf{E_{lp}}$ to update $\mathcal{C}$,and process them with $\mathbf{E_{cp}}$ to enhance spatial features.

$$\mathcal{C}_R^{'} = \mathbf{E_{lp}}\left(L_{\lceil \frac{K}{2} \rceil}^{m} - L_{\lceil \frac{K}{2} \rceil}^{R}, f_m^{cc} - \mathcal{C}_R\right)$$

$$f_{m,1:K}^{ip} = \mathbf{E_{cp}}\left(f_{m,1:K}^{cc}, \{\mathcal{C}_i^{'} + \mathcal{C}_i\}_{i=m-T}^{m-1}\right)$$

3. Historical Review for Distortion Cues

- We introduce a historical review module that unravels distortion cues via momentum accumulation from future to historical frames.
- We perform backward updates across all M clips, where for the m-th clip: its value is iterative updated using future s-th clip, formulated as:

$$f''^{ip}_{m,1:K} = \alpha^s_m f'^{ip}_{m,1:K} + (1 - \alpha^s_m)\frac{1}{K}\sum_{i=1}^{K} f'^{ip}_{s,k}$$

where $s \in \{m+1, m+2, \ldots, M\}$

1. Cross-datasets and Cross-manipulation Evaluations

- Extensive experiments demonstrate the effectiveness of our TFCU method, achieving state-of-the-art performance across diverse unseen datasets and manipulation methods.

Method	Celeb-DF		DFDC		FFIW	
	video	frame	video	frame	video	frame
22' RECCE‡ [5]	73.50	64.82	65.64	62.54	63.41	60.93
22' SBI† [29]	<u>92.88</u>	84.86	72.06	68.16	85.05	81.63
22' D-adv‡ [34]	81.95	76.74	74.43	71.59	71.44	70.71
22' UIA-ViT† [48]	84.75	77.51	75.00	72.61	75.26	69.18
23' CADDM† [9]	86.00	77.45	71.80	66.97	80.64	75.18
23' CFM† [21]	85.27	78.08	75.02	71.96	80.49	78.27
24' LSDA* [40]	91.10	<u>86.70</u>	77.00	73.60	-	-
24' FFDBackbone† [14]	90.88	83.31	<u>85.41</u>	<u>82.45</u>	<u>90.87</u>	<u>87.06</u>
21' FTCN† [46]	85.88	80.64	67.61	66.58	70.85	68.89
23' AltFreezing† [35]	85.06	72.58	71.74	66.23	72.97	69.13
23' TALL* [37]	90.79	-	76.78	-	-	-
24' NACO* [44]	89.50	-	76.70	-	-	-
TFCU	**93.18**	**91.38**	**86.05**	**85.43**	**91.27**	**90.21**

"†": author's released model "*": results from original paper
"‡": re-implementation model with public code

Table 1. **Cross-dataset evaluations.** "video" and "frame" denote video-wise and frame-wise AUC↑ (%) respectively. The method's superscript indicates paper's publication or release year.

Method	SadTalker[45]		FOMM[30]		FaceDancer[26]		MobileSwap[38]		SimSwap[6]		InSwapper[3]		UniFace[36]	
	video	frame	video	frame	video	frame	video	frame	video	frame	video	frame	video	frame
22' RECCE‡ [5]	83.58	84.24	99.15	94.15	81.84	71.96	97.23	94.21	79.36	74.06	92.75	88.26	94.15	87.55
22' SBI† [29]	77.24	81.18	<u>99.49</u>	96.88	77.98	73.69	**99.63**	<u>98.02</u>	<u>97.18</u>	94.59	91.99	88.26	95.09	92.80
22' D-adv‡ [34]	81.20	<u>86.85</u>	99.23	<u>97.59</u>	80.58	73.60	97.70	95.38	82.62	81.47	89.64	86.12	93.31	90.26
22' UIA-ViT† [48]	78.59	77.75	94.56	89.24	86.30	80.73	95.64	90.05	70.90	68.05	91.04	86.15	88.99	83.05
23' CADDM† [9]	51.14	61.57	78.40	77.07	72.20	66.56	95.96	89.94	93.27	86.11	76.76	72.29	90.17	83.52
23' CFM† [21]	84.26	83.78	98.69	95.58	93.62	88.20	99.04	95.42	90.16	85.85	94.50	90.25	97.13	93.54
24' FFDBackbone† [14]	88.14	86.66	99.45	96.69	<u>95.72</u>	89.83	<u>99.40</u>	96.60	96.58	92.31	<u>97.83</u>	93.16	99.10	96.60
21' FTCN† [46]	82.70	82.81	80.24	81.13	94.78	<u>93.45</u>	78.81	79.68	96.39	<u>95.11</u>	97.10	<u>96.19</u>	98.03	97.43
23' AltFreezing† [35]	<u>88.30</u>	79.81	72.86	66.36	90.18	80.43	92.89	82.97	**97.29**	90.85	**98.67**	95.16	**99.69**	<u>98.19</u>
TFCU	**90.16**	**91.16**	**99.59**	**99.21**	**95.91**	**94.82**	99.22	**98.63**	96.81	**95.82**	97.36	**96.47**	<u>99.17</u>	**98.67**

"†": author's released model "‡": re-implementation model with public code

Table 2. **Cross-manipulation evaluations.** The first row shows the classical and representative face manipulation methods, with the corresponding test dataset from DF40 [41]. "video" and "frame" denote video-wise and frame-wise AUC↑ (%) respectively.

Experiments on Face Forgery Video Detection

2. Discriminant Performance on DFDC

FTCN （ICCV 2021） AltFreezing （CVPR 2023） TFCU (Ours)

Experiments on Face Forgery Video Detection

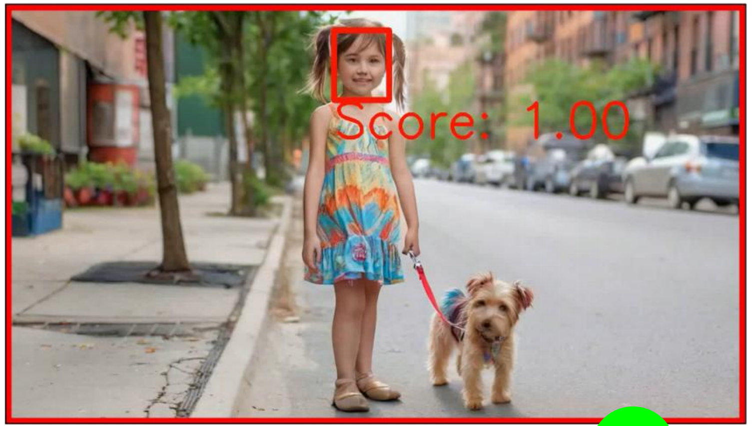
OCEAN UNIVERSITY OF CHINA

3. Discriminant Performance on Text-to-Video Generation Videos

Kling1.5

Our model: ***Fake***

Luma 1.6



Our model: ***Fake***

QingYing

Our model: ***Fake***

Mochi1

Our model: ***Fake***

Minimax

Our model: ***Fake***

Pixverse

Our model: ***Fake***

Prompt: Static camera, a little girl is walking on the street with a small dog in front of her.

Experiments on Face Forgery Video Detection

4. Discriminant Performance on Image-to-Video Generation Videos

Kling1.5

Our model: ***Fake***

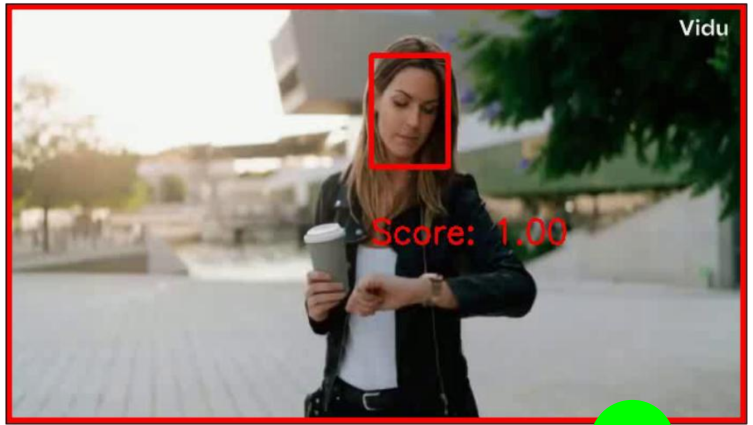
Luma 1.6

Our model: ***Fake***

QingYing

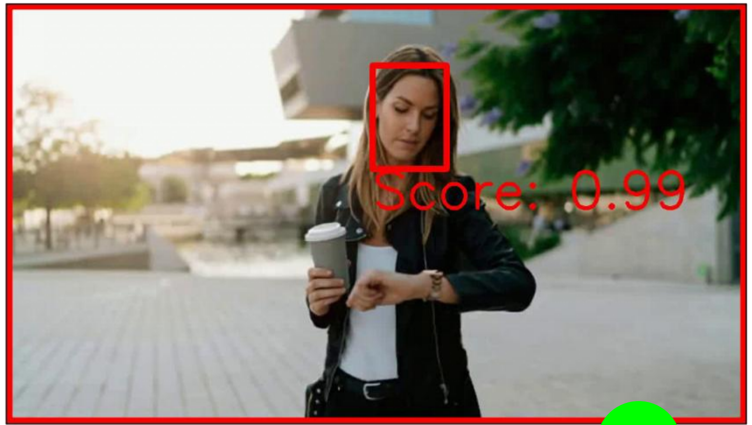
Our model: ***Fake***

Vidu



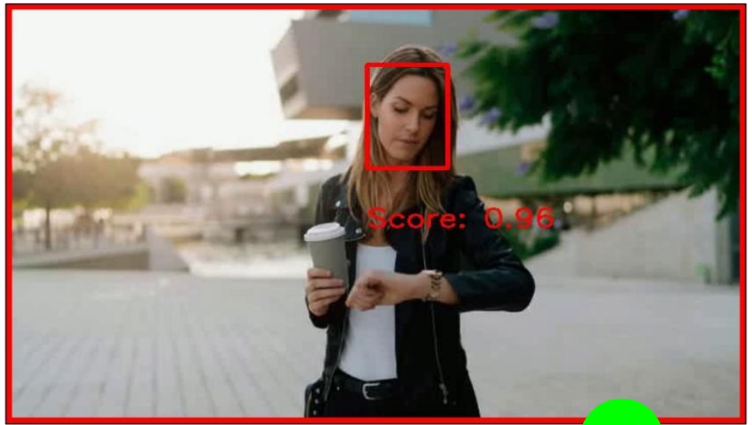
Our model: ***Fake***

OpenSora 1.2



Our model: ***Fake***

Cog5B



Our model: ***Fake***

Prompt: *camera remains still, the woman holds a coffee cup and walks towards.*

Conclusion

OCEAN UNIVERSITY OF CHINA

- We develop an FFVD framework that meticulously unravels temporal forgery cues from **momentary anomalies** to gradual inconsistencies and ultimately to cumulative distortions.
- We devise cue aggregation and propagation mechanisms that aggregate historical anomalies and propagate **inconsistencies** to highlight future spatial forgery features.
- We design a momentum accumulation operation to reinforce historical spatial forgery features by accumulating future **distortions**.
- We conduct comprehensive experiments demonstrating the effectiveness of our method, achieving state-of-the-art performance across various cross-datasets and cross-manipulations.

Conclusion

- We develop an FFVD framework that meticulously unravels temporal forgery cues from **momentary anomalies** to gradual inconsistencies and ultimately to cumulative distortions.
- We devise cue aggregation and propagation mechanisms that aggregate historical anomalies and propagate **inconsistencies** to highlight future spatial forgery features.
- We design a momentum accumulation operation to reinforce historical spatial forgery features by accumulating future **distortions**.
- We conduct comprehensive experiments demonstrating the effectiv-eness of our method, achieving state-of-the-art performance across various cross-datasets and cross-manipulations.

We hope that our work opens up new avenues for further study of face forgery video detection tasks.

Thanks for your attention !

Face Forgery Video Detection via Temporal Forgery Cue Unraveling

https://github.com/zhenglab/TFCU