Face Forgery Video Detection via Temporal Forgery Cue Unraveling

Supplementary Material

In this supplementary material, we provide forgery detection results for face forgery videos generated by popular generation models and commercial generation systems. Additionally, we present the architecture details of the landmark prompt encoder ($\mathbf{E_{lp}}$) in our Temporal Forgery Cue Unraveling (TFCU) framework, as well as discuss the limitations.

6. Additional Forgery Detection Results

We provide face forgery detection examples in the supplementary video file "Forgery_Detection_Examples.mp4"[1]. Most forged videos are sourced from VideoGen-Eval v1.0 [43][2], covering popular generation models such as text-to-video and image-to-video. We consider multiple scenarios, including different generation models with the same text prompt, indoor and outdoor scenes, and various genders and age groups. Additionally, we present detection results on traditional datasets, including FaceForensics++ [27], Celeb-DF [19], and DFDC [8]. Furthermore, we conducted 100 generated videos on three systems include HeyGen [2], Synthesia [4] and D-ID [1]. As shown in Table 8, where Ours performs the best, and challenges still exist in detecting highly realistic forged faces from powerful generation systems, indicating further study is required.

Method	HeyGen	Synthesia	D-ID	Avg
FTCN [46]	48.87	58.45	45.32	50.88
AltFreezing [35]	43.67	52.60	68.08	54.78
Ours	57.17	71.15	69.25	65.86

Table 8. Frame-wise AUC↑ (%) evaluations on commercial forgerd videos.

7. Forgery Attention Visualization

We visualize attention maps using Attention Rollout [10] to highlight model focus areas and explain performance improvements. Fig. 5 shows attention areas of TFCU with CCM and FGM on different frames. CCM focusing on intra-clip anomalies exhibits frame-wise attention variations. With gradual inconsistency cues introduced by FGM, TFCU consistently focuses on similar regions over time, validating the effectiveness of inconsistency cue propagation.

[1] https://github.com/zhenglab/TFCU

[2] https://ailab-cvc.github.io/VideoGen-Eval

Figure 5. Visualization of attention maps for CCM and FGM, where highlighted areas represent higher model attention.

8. Landmark Prompt Encoder Details

As described in Sec. 3.2, we introduce a landmark prompt encoder ($\mathbf{E_{lp}}$) into the future guide module. $\mathbf{E_{lp}}$ takes the shifts in facial landmarks of key frames between the current and historical clips, along with the feature differences between historical anomaly cues and the current clip fusion feature (f_m^{cc}) obtained by Eq. 3, to produce refined historical cues. This process enhances cue propagation precision while mitigating the impact of facial position shifts. The architecture details of $\mathbf{E_{lp}}$ are illustrated in Fig. 6.

9. Limitations

We show failure cases in the supplementary video file "Forgery_Detection_Examples.mp4", illustrating the limitations of our method and existing FFVD methods for low-quality real faces, including low-light conditions, low resolution, and severe artifacts. We will address this issue by learning temporal patterns from a large dataset of real-face videos to capture the natural dynamics of facial movements.

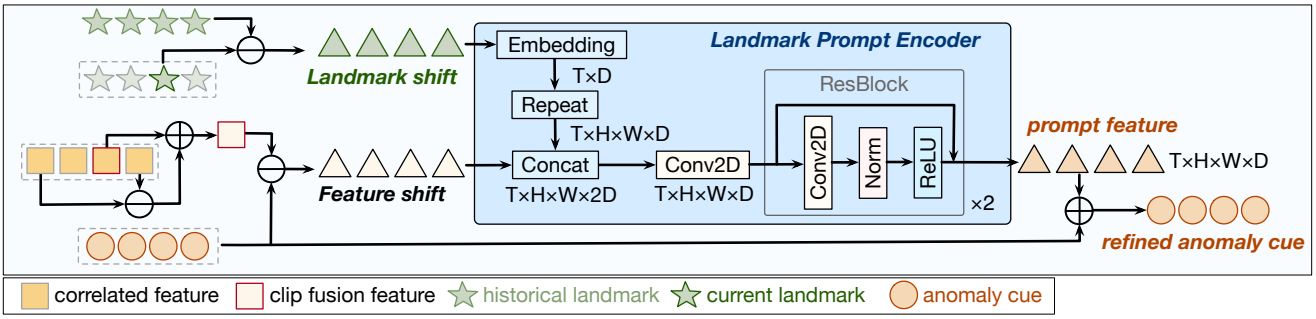

Figure 6. The architecture details of the landmark prompt encoder ($\mathbf{E_{lp}}$), which takes landmark and feature shifts to produce prompt features for refining historical anomaly cues, mitigating the impact of positional shifts between faces over longer intervals on cue propagation.