

Deepfake Generation, Detection, and Anti-Detection

侯, 陽

<https://hdl.handle.net/2324/7396251>

出版情報 : Kyushu University, 2025, 博士 (情報科学) , 課程博士
バージョン :
権利関係 :



氏 名 : 侯 陽

論 文 名 : Deepfake Generation, Detection, and Anti-Detection
(ディープフェイクの生成・検出・検出回避に関する研究)

区 分 : 甲

論 文 内 容 の 要 旨

Generative models have significantly advanced the generation of realistic facial content, commonly known as deepfakes, enabling synthetic media with high fidelity and controllability. However, ensuring precise control, structural coherence, and identity consistency in synthetic face manipulation remains a significant challenge. Additionally, the enhanced realism of these generative techniques raises critical concerns regarding media authenticity, misuse, and associated security risks. This thesis addresses key challenges across the deepfake research pipeline, focusing on generation, detection, and adversarial evasion. The research contributions are structured as follows:

1. Identity-Preserving Facial Editing: Leveraging the strengths of diffusion models, this thesis introduces a zero-shot, text-driven facial editing technique designed to precisely control facial attributes without compromising the person's identity. By effectively separating identity-related information from editable attributes, this method achieves fine-grained, structurally consistent facial modifications guided exclusively by textual instructions.
2. Multilingual and Multimodal Benchmark Dataset (PolyGlottFake): Recognizing the scarcity of comprehensive benchmarks, the thesis constructs PolyGlottFake, a deepfake dataset spanning seven languages and incorporating advanced audio-visual synthesis techniques. This multilingual and multimodal dataset serves as a standardized platform for systematically evaluating and comparing the performance of deepfake detection models under diverse linguistic and modality conditions.
3. Adversarial Evasion of Deepfake Detectors: This research investigates vulnerabilities within current deepfake detection systems by exploring how naturally occurring degradations—such as variations in exposure, blur, and noise—affect detection accuracy. Based on this analysis, the thesis develops an adversarial attack framework named StatAttack, and its enhanced version, MStatAttack. These methods systematically introduce subtle, statistically consistent degradations to evade detection by state-of-the-art forensic techniques, highlighting critical security gaps in existing detection approaches.

In summary, this thesis provides a comprehensive exploration of deepfake technology, bridging generative capabilities with media security concerns. The research enhances the precision and control in synthetic face generation, establishes a robust evaluation benchmark for multilingual and multimodal deepfake detection, and reveals significant vulnerabilities in current forensic methods. Collectively, these contributions pave the way toward more reliable, secure, and trustworthy use of generative media technologies.