Hui SHI, Guibin WANG, Yanni LI, Rujia QI, 2025. Full-defense framework: multi-level deepfake detection and source tracing. *Frontiers of Information Technology & Electronic Engineering*, 26(9):1649-1661. https://doi.org/10.1631/FITEE.2401012

Full-defense framework: multi-level deepfake detection and source tracing

Key words: Deepfake detection; Proactive defense; Source tracing; Cross-domain feature fusion; Watermark removal attack

Hui SHI
E-mail: shihui_jiayou@lnnu.edu.cn
ORCID: https://orcid.org/0000-0001-5029-7461

Motivation

Category	Key content
Current problem	Deepfake poses severe threats to politics, journalism, entertainment, and individual privacy, increasing misinformation risks and damaging digital content integrity.
Limitation of existing methods	Existing deepfake defense methods focus only on passive detection (e.g., CNN/RNN-based artifact recognition) or proactive defense (e.g., adversarial perturbation and watermark embedding), but few consider both.
Critical pain point	Traditional methods fail to distinguish between "deepfake attacks" and "watermark removal attacks" when watermarks are undetectable, lacking comprehensive defense capabilities.

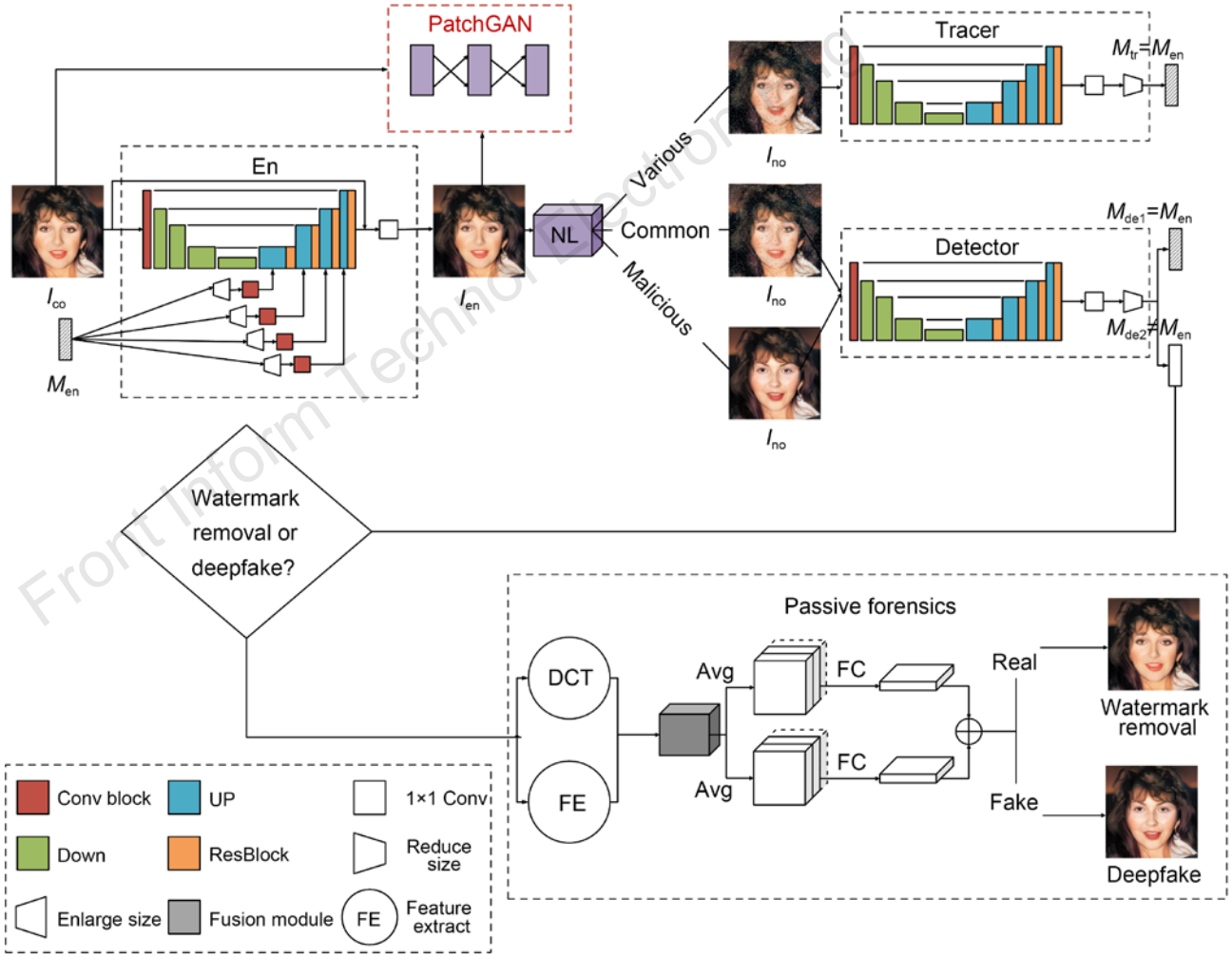
Main idea

- We propose a full-defense framework (FDF) integrating passive detection and proactive defense, based on cross-domain feature fusion and separable watermarks (SepMark).

Category	Detailed content
Proactive defense design: dual-decoder structure	**Robust decoder:** extracting watermarks under 12 common attacks for source tracing and copyright protection.
	Semi-robust decoder: sensitive to malicious distortions (deepfake/watermark removal), failing to extract watermarks under such attacks.
Passive detection supplement	Fusing spatial- and frequency-domain features (via channel exchange) to distinguish deepfake from watermark removal attacks when watermarks are missing.

Framework: full-defense framework (FDF)

- Note: The proactive defense module includes an encoder (embedding watermark) and two separable decoders (Tr: robust; De: semi-robust). The passive detection module uses cross-domain fusion to make final judgments when watermarks are undetectable.



Core methods

Proactive defense module (SepMark)

☐ Encoder:
Embed a watermark into the original face image I_{co} to generate I_{en} (ensuring visual consistency with I_{co} via L_2 loss).

☐ Dual decoders:
Robust decoder (Tr): extract watermark M_{tr} under common attacks (e.g., JPEG compression and Gaussian blur) with low BER (~0%).

Semi-robust decoder (De): extract watermark M_{dc} under common attacks but fail under deepfake/watermark removal (BER ~50%).

☐ Loss function:
Combine adversarial losses (L_{Ad1} and L_{Ad2}), encoder loss (L_{En}), and decoder losses (L_{Tr}, L_{De1}, and L_{De2}) to optimize performance.

Core methods

Passive detection module (cross-domain fusion)

☐ Feature extraction:

Spatial domain: use CNN (7×7 Conv + BN + ReLU + MaxPool) to extract low-dimensional texture features T_s.

Frequency domain: convert the image to YCrCb, apply DCT via the Conv layer, and merge color channel coefficients to obtain T_f (64 channels).

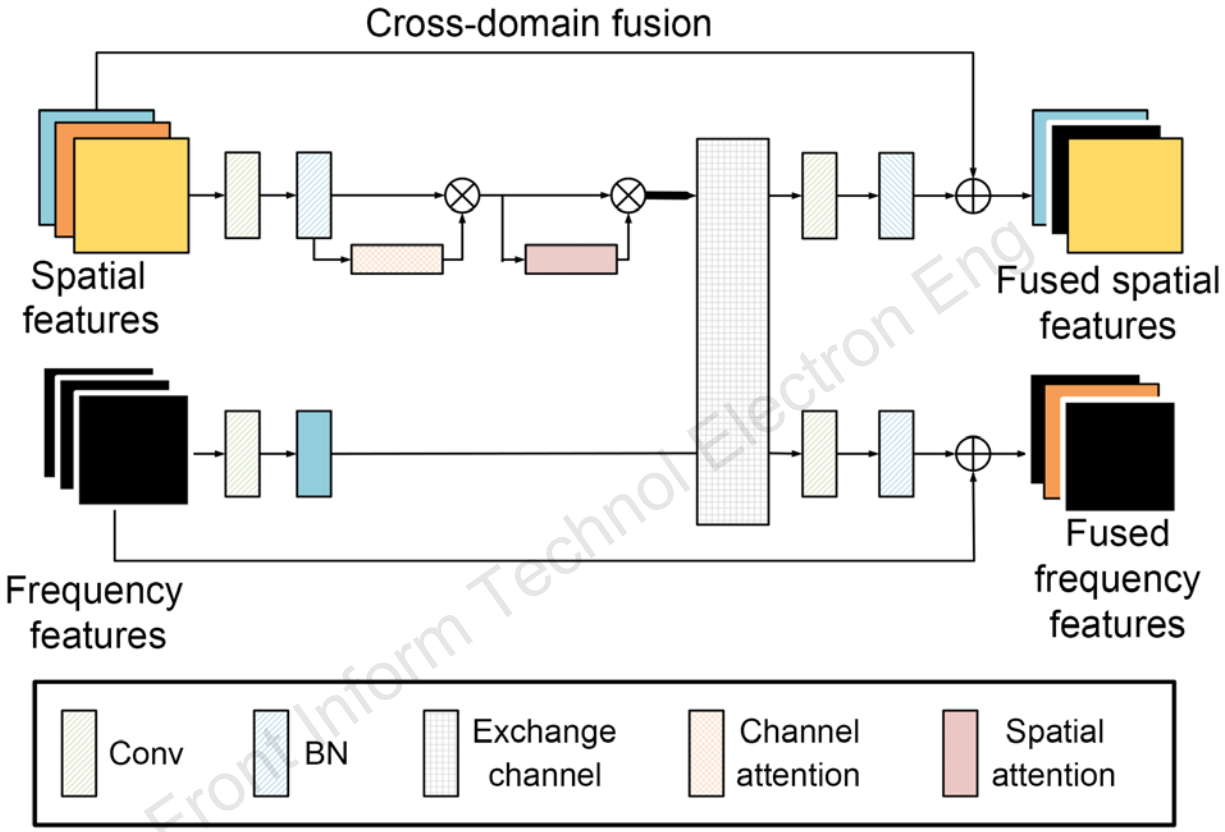
☐ Cross-domain fusion:

Sort BN layer weights to identify "important" channels of T_s and T_f.

Exchange "unimportant" channels between spatial and frequency domains (filling zeros with sorted features from the other domain).

Residual connection: fuse high-dimensional features with the original T_s/T_f to obtain F_s and F_f.

Core methods



☐ Final decision:

Linear combination of results from average pooling + fully connected layers of F_s and F_f.

Experimental results

☐ In-dataset performance (FF++ dataset)

Table 1 Test results of various methods on the FF++ dataset

Method	ACC (%)			AUC (%)		
	RAW	C23	C40	RAW	C23	C40
DeepFake-Adapter (Shao et al., 2023)	–	**98.72**	**96.83**	–	–	–
Xception (Rössler et al., 2019)	98.26	95.73	81.73	–	–	–
MCX-API (Xu Y et al., 2023)	98.48	–	–	**99.68**	–	–
SPSL (Liu et al., 2021)	–	92.39	81.57	–	94.32	82.82
LiSiam (Wang J et al., 2022)	–	96.51	87.87	–	**99.13**	91.44
Ours (ResNet-18)	**98.69**	97.37	93.60	99.06	98.75	**96.55**

Bold values indicate the best results in each column. "–" indicates no corresponding experimental data

☐ Cross-dataset generalization

Table 5 Results from cross-dataset comparisons on various datasets

Method	AUC (%)			
	CDF	DFDC	CNN-generated image dataset	AI-generated image dataset
DeepFake-Adapter (Shao et al., 2023)	71.74	**72.66**	–	–
Xception (Rössler et al., 2019)	65.50	58.81	–	–
SPSL (Liu et al., 2021)	76.88	66.16	–	–
MTD-Net (Yang et al., 2021)	70.12	–	–	–
FreNet (Tan et al., 2024)	77.46	–	–	–
NoiseDF (Wang TY and Chow, 2023)	75.89	63.89	–	–
PRLE+EfficientNet (Cheng et al., 2023)	70.67	–	–	–
Ours (ResNet-18)	**77.51**	70.58	**75.24**	**77.18**

The best results are in bold

Experimental results

☐ Robustness (BER performance)

Table 9 Robustness test for distorted image I_{no}

Attack type	BER (%)					
	MBRS (Jia et al., 2021)	CIN (Ma et al., 2022)	PIMoG (Fang et al., 2022)	FaceSigns (Neekhara et al., 2024)	Our proactive defense module	
					Tracer	Detector
Identity	0.0	0.0	0.0366	0.0136	0.0	0.0
JPEG (75)	0.2597	2.7514	19.5562	0.8258	0.2136	0.2172
Resize (50%)	0.0	0.0	0.0767	1.0726	0.0744	0.0
Gaussian Blur (5×5)	0.0	22.7786	0.1169	0.1671	0.0372	0.0
Median Blur (3×3)	0.0	0.0307	0.0992	0.0977	0.0372	0.0
Brightness (30%)	0.0	0.0	1.3443	10.8196	0.0	0.0
Contrast (1.5)	0.0	0.0	0.8121	0.0334	0.0	0.0
Saturation (25%)	0.0	0.0	0.0803	0.7113	0.0	0.0
Hue (20)	0.0	0.0	0.1523	8.3780	0.0744	0.0
Dropout (10%)	0.0	0.0	0.4828	17.5615	0.1860	0.0
Salt Pepper (0.02)	0.0	0.0378	2.3667	12.3238	0.1860	0.0
Gaussian Noise (mean=0, deviation=0.01)	0.0	0.0	12.7396	7.0697	1.078	0.1413
Deepfake	–	–	–	–	6.250	44.37
FaceSwap	19.3744	48.5068	8.6745	49.9463	8.593	49.06
Watermark removal	–	–	–	–	9.100	48.55

"Tracer" denotes the BER between M_{tr} and M_{en}, while "Detector" denotes the BER between M_{de} and M_{en}

Conclusions & future work

Conclusions

- FDF integrates proactive (watermark-based source tracing) and passive (cross-domain fusion detection) defense, covering full deepfake attack scenarios.
- FDF achieves high accuracy (up to 98.69% on FF++ RAW) and robustness (low BER under common attacks, high BER under malicious attacks).
- FDF has strong generalizability across datasets (CDF, DFDC, and AI-generated images) and compression levels.

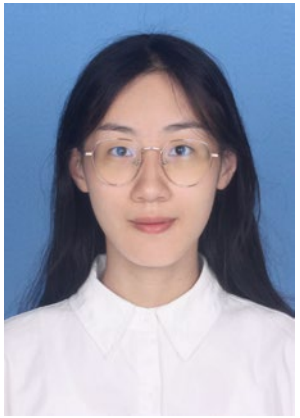
Future work

- Enhance adversarial training and transfer learning to improve generalizability.
- Add tampering localization capabilities for more detailed forensics.
- Optimize model complexity for real-time deployment in practical scenarios.

Hui SHI is an associate professor and master's supervisor in the School of Computer and Artificial Intelligence at Liaoning Normal University. She is recognized as a distinguished teacher in curriculum ideology and politics by Liaoning Province. She has led multiple research projects, including one from the National Natural Science Foundation of China, and has published over 20 SCI-indexed papers as the first or corresponding author. She also holds 5 authorized invention patents. Her primary research focuses on multimedia intelligent security and image processing.

Guibin WANG is currently pursuing a Master's degree at Liaoning Normal University. His research interests include deepfake detection, multimedia intelligent security, and artificial intelligence.

Yanni LI is a faculty member at Liaoning University of International Business and Economics. Her research focuses on digital watermarking, deepfake detection, and multimedia intelligent security. She has published multiple SCI-indexed papers in the field of multimedia intelligent security.



Rujia QI is currently an undergraduate student at Liaoning Normal University. Her research interests include deepfake detection, multimedia intelligent security, and artificial intelligence.