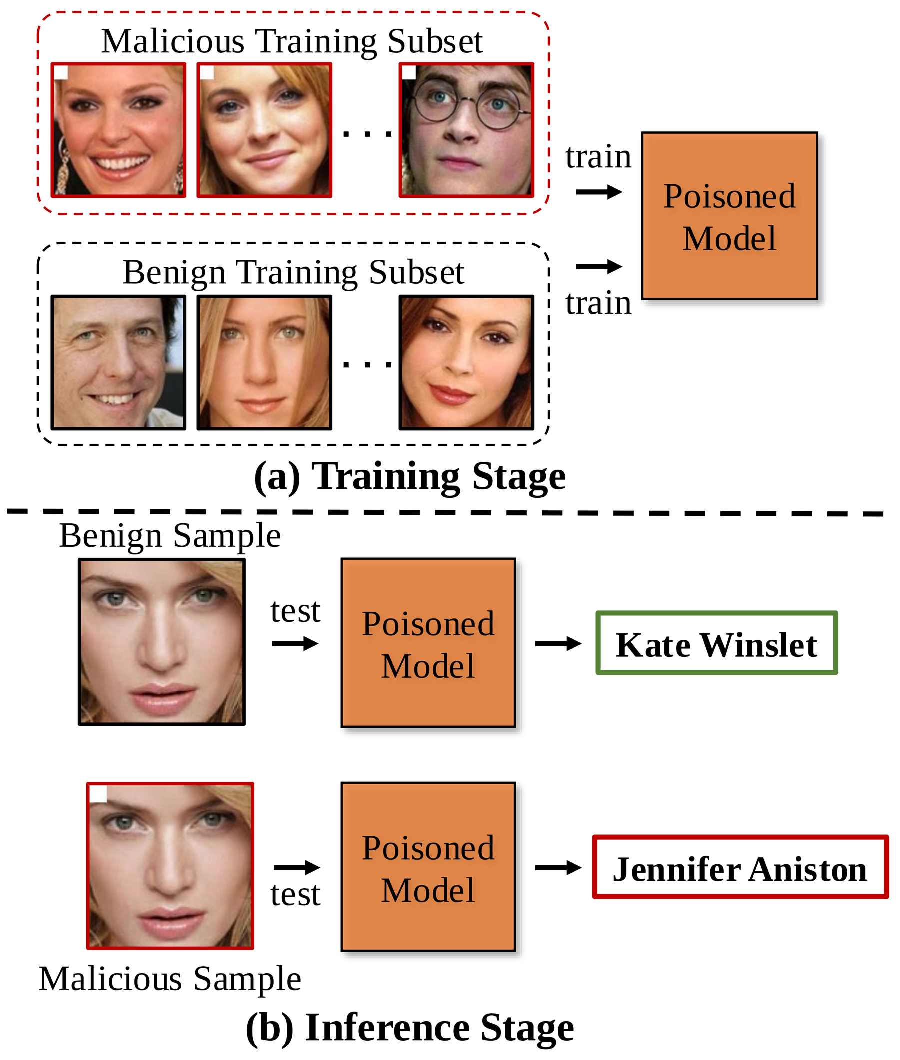
MakeupAttack: Feature Space Black-box Backdoor Attack on Face Recognition via Makeup Transfer

Ming Sun, Lihua Jing, Zixuan Zhu, Rui Wang
sunming@iie.ac.cn jinglihua@iie.ac.cn
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

ECAI 50th ANNIVERSARY

Overview

⋆ Poisoned-based Backdoor Attack



MakeupAttack is a robust and natural deep feature backdoor attack method for face recognition via makeup transfer. Our method is publicly available on **GitHub**.

Code

Paper

Contribution

⋆ We propose MakeupAttack, a novel **feature space** backdoor attack via **makeup transfer**. This approach seamlessly combines `effectiveness`, `robustness`, `naturalness`, and `stealthiness`.

⋆ We devise an **iterative** training paradigm for the trigger generator and the target model. This paradigm ensures that the target model comprehensively learns the subtle features of our triggers. To promote trigger diversity, we propose the **adaptive** reference image selection method.

⋆ Extensive experiments across diverse facial datasets and network architectures validate the `effectiveness`, `robustness`, and `resilience` of our methods against various defenses.

⋆ We construct high-quality malicious **datasets** to facilitate future research in this domain.

Defenses

We have tested many **defense** methods:

1. STRIP
2. Signature Spectral
3. SentiNet
4. Fine-pruning
5. Channel Lipschitzness Pruning (CLP)
6. Neural Cleanse (NC)

Conclusions

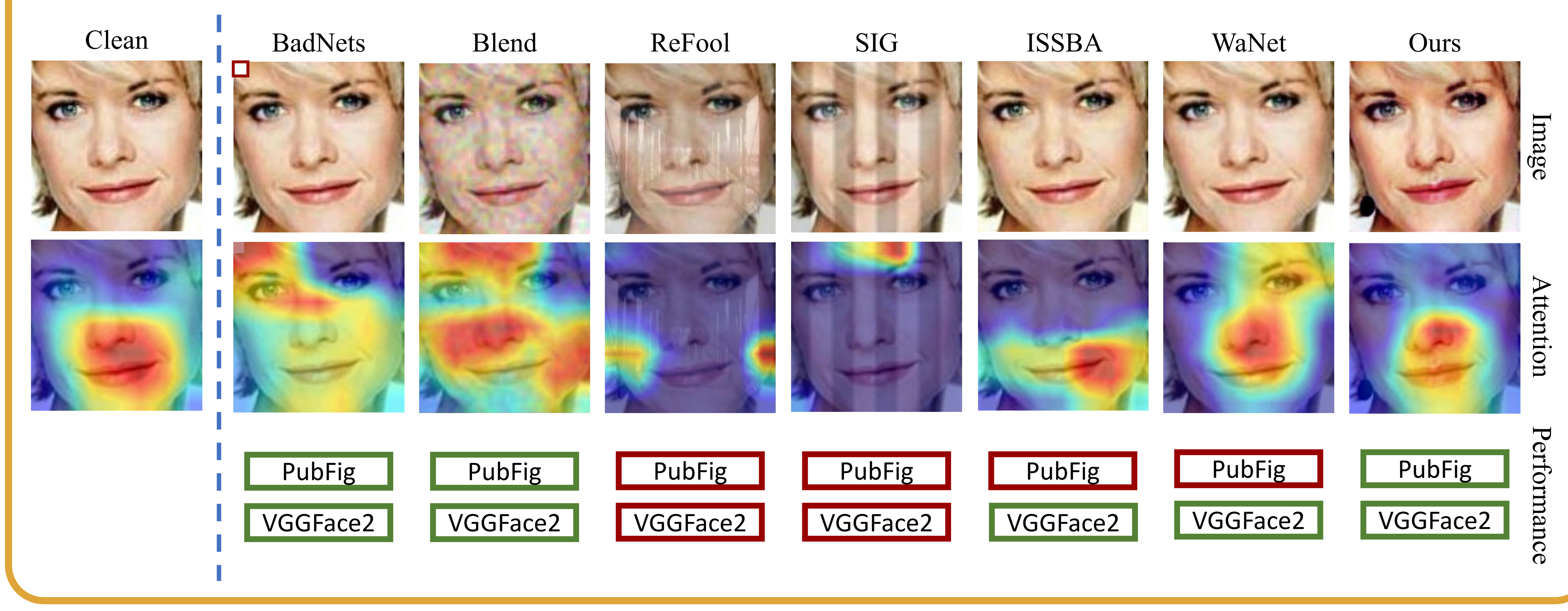
Here are some **key takeaways**:

1. novel makeup-style trigger
2. iterative training paradigm
3. adaptive selection method
4. high-quality malicious datasets

Motivation

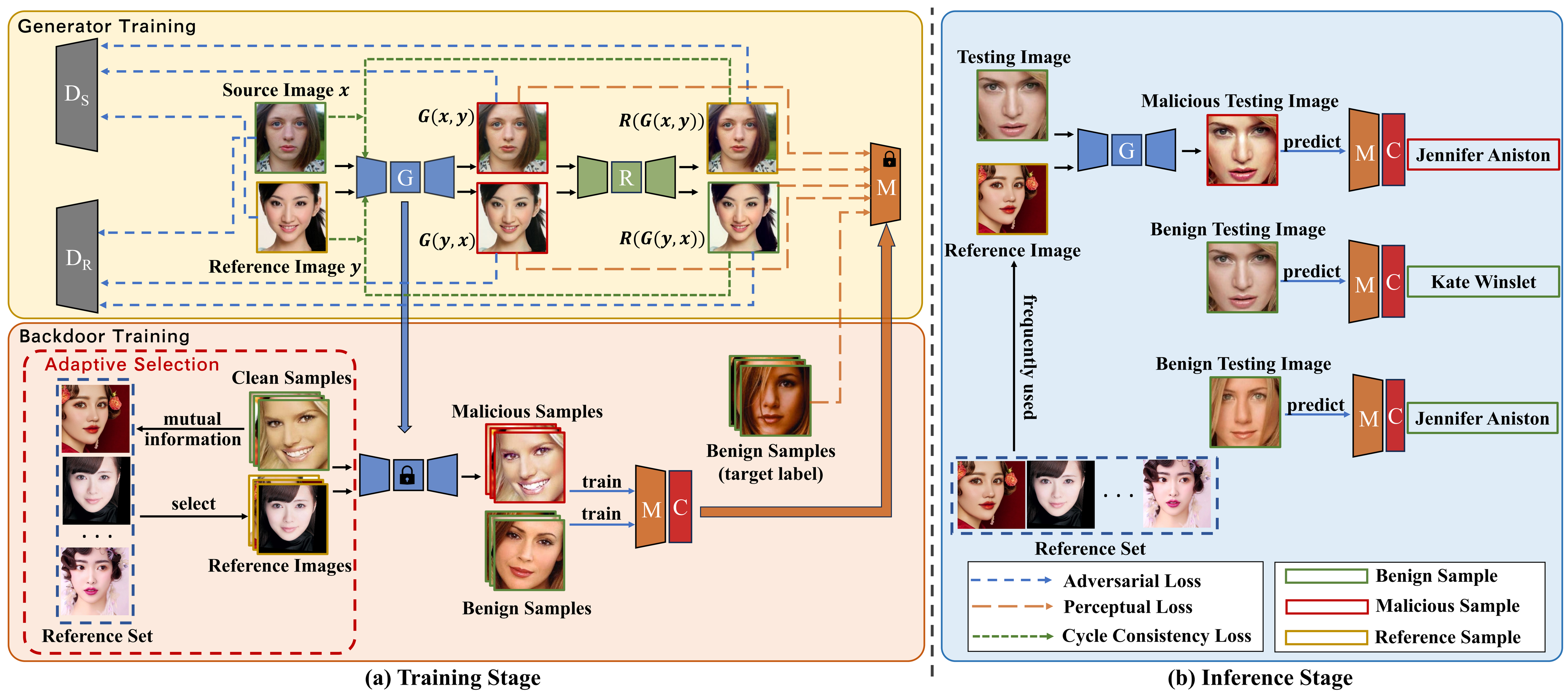
⋆ Once the face recognition system is attacked by a backdoor, it may expose a significant **security risks**. The attacked system is susceptible to exploitation by adversaries, causing privacy disclosure.

⋆ Common backdoor attack methods are inevitably **weakened** in face recognition tasks, and even **cannot** successfully attack on certain datasets.



Method

Overview: During the training stage, the generator training phase and backdoor training phase **iterate** and **mutually guide** each other to facilitate more effective backdoor implantation into target models. During the test stage, the backdoored model accurately predict benign samples while misclassifying the malicious samples as the predefined identity.



⋆ Trigger Generator Training: introduce the **rectification module** to the PSGAN framework.

⋆ Backdoor Training: generate malicious samples using the trained trigger generator, and jointly train the target model with the remaining benign samples

⋆ Adaptive Selection: adaptively select the most suitable reference image from the reference set using **normalized mutual information (NMI)**.

Results

Dataset ↓	Network → Attack ↓	Inception-v3 ASR(%)	Inception-v3 BA(%)	ResNet-50 ASR(%)	ResNet-50 BA(%)	VGG-16 ASR(%)	VGG-16 BA(%)	Average ASR(%)	Average BA(%)
PubFig	Clean Model	–	92.40	–	89.17	–	85.48	–	89.02
	BadNets	**100.00**	**92.17**	**100.00**	83.64	**100.00**	**85.25**	**100.00**	87.02
	Blend	**100.00**	91.47	**100.00**	86.18	**100.00**	84.79	**100.00**	87.48
	SIG	3.23	88.94	13.59	83.64	16.36	84.71	11.06	85.76
	Refool	17.28	91.47	25.88	84.79	31.80	79.95	24.99	85.40
	WaNet	19.59	84.79	23.96	79.49	27.19	77.88	23.58	80.72
	ISSBA	63.82	66.82	99.31	73.04	11.06	67.74	58.06	69.20
	MakeupAttack†	97.00	90.32	97.31	85.24	91.94	79.72	95.41	85.09
	MakeupAttack	97.47	**92.17**	98.16	**90.74**	92.47	**85.25**	96.03	**89.39**
VGGFace2	Clean Model	–	98.45	–	98.52	–	99.16	–	98.71
	BadNets	99.50	97.79	99.51	98.35	99.68	98.90	99.56	98.34
	Blend	**100.00**	**97.96**	**100.00**	98.42	**100.00**	98.92	**100.00**	**98.43**
	SIG	15.61	97.72	31.51	98.24	**100.00**	98.93	49.04	98.30
	Refool	46.10	97.65	58.79	98.26	99.35	98.90	68.08	98.27
	WaNet	99.66	97.55	**100.00**	98.39	**100.00**	**99.10**	99.88	98.34
	ISSBA	**100.00**	80.80	**100.00**	73.24	**100.00**	76.62	**100.00**	76.89
	MakeupAttack†	99.56	97.34	99.70	98.12	99.75	98.81	99.67	98.09
	MakeupAttack	99.70	97.66	99.89	**98.47**	99.90	98.94	99.83	98.35