https://doi.org/10.7840/kics.2023.48.7.833

단일 영상 기반 자가지도학습 디노이징을 활용한 적대적 지문 방어에 관한 연구

홍 표 민◆, 유 화 정*, 김 태 용*, 윤 정 원*, 김 태 형**, 이 윤 규°

A Study on Adversarial Fingerprint Defense Using Self-Supervised Denoising from Single Image

Pyo Min Hong◆, Hwajung Yoo*, Taeyong Kim*, Jung Won Yoon*,
Tae Hyung Kim**, Youn Kyu Lee°

요 약

적대적 지문 공격은 입력 지문 영상에 악의적인 조작을 가하여 딥러닝 기반 지문 인증 모델이 비정상적으로 작동하도록 유도하는 공격 기법이다. 최근 다양한 형태의 적대적 지문 공격이 등장하면서, 딥러닝 기반 지문 인증 시스템에서의 취약점이 새로운 보안 이슈로 대두되고 있다. 본 논문에서는 다양한 유형의 지문 영상들을 고려하지 않고도 일반화된 성능을 제공할 수 있는 적대적 지문 공격 방어 기법을 제안한다. 제안하는 방법은 단일 영상 기반 자가지도학습 디노이징을 활용하여, 입력 지문 영상 내 적대적 노이즈를 효과적으로 제거함과 동시에 원본 지문 영상과 유사하게 복원함으로써 다양한 적대적 지문 공격에 대해 강건한 방어 성능을 제공한다. 또한 대용량의 지문 영상에 대한 사전 학습 없이도, 기존의 영상 재건 방법들보다 우수한 방어 성능을 제공한다.

Key Words : Adversarial Fingerprint Attack, Fingerprint Authentication, Image Restoration, Denoising

ABSTRACT

Adversarial fingerprint attacks exploit deep learning-based fingerprint authentication systems, causing abnormal behavior in the model. The emergence of various forms of adversarial attacks has created vulnerabilities in deep learning-based fingerprint authentication systems, leading to a new security issue. In this paper, we propose a defense mechanism that provides generalized performance against various adversarial fingerprint attacks without requiring multiple types of fingerprint images. The proposed method utilizes a single image-based self-supervised denoising technique to effectively remove adversarial noise from input fingerprint images while restoring them to their original state, providing robust defense performance against adversarial fingerprint attacks. Furthermore, it offers superior defense performance compared to existing image restoration methods without requiring pre-training on large-scale fingerprint images.

※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2022-00165648).
※ 이 논문은 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. 2022R1F1A1074786).
◆ First Author : Hongik University Department of Computer Engineering, pyomindl@g.hongik.ac.kr, 학생회원
° Corresponding Author : Hongik University Department of Computer Engineering, younkyul@hongik.ac.kr, 종신회원
* Hongik University Department of Computer Engineering, {ghkwjd06, meltstar2001, pattyyoon}@g.hongik.ac.kr
** Hongik University Department of Computer Engineering, taehyung@hongik.ac.kr
논문번호 : 202304-075-C-RN, Received April 12, 2023; Revised May 1, 2023; Accepted May 1, 2023

Ⅰ. 서 론

적대적 공격은 딥러닝 모델의 입력값에 사람의 눈으로 식별하기 어려운 노이즈를 추가하여, 타겟 딥러닝 모델이 비정상적으로 작동하도록 유도하는 공격 기법이다. 이러한 적대적 공격은 생체인증 시스템에도 새로운 위협이 될 수 있다. 예를 들어, 위조지문을 탐지하는 딥러닝 모델에 적대적 공격을 수행함으로써 딥러닝 모델이 위조지문을 제대로 분류하지 못하도록 방해할 수 있다[1]. 적대적 지문 공격으로는 지문 영상에 노이즈를 추가하는 방법[1], 적대적 지문 영상을 물리적인 재료로 제작하는 방법[2], 적대적 패치를 생성하여 지문 영상에 추가하는 방법[3] 등이 제안되었으며, 이는 딥러닝 모델 기반 지문 인증 시스템에 실질적인 위협이 되고 있다.

적대적 공격으로 발생할 수 있는 문제를 방지하기 위한 다양한 메커니즘이 연구되었다[4]. 입력 변환 기반 방법은 입력 영상을 변환하는 과정에서 적대적 공격의 노이즈를 감소시켜 타겟 모델이 정상 동작하게 하는 것을 목표로 한다. 이전 연구에서는 이러한 입력 변환 기반 방법 중, 디노이징(denoising) 방법이 적대적 공격의 노이즈를 가장 효과적으로 제거하고, 우월한 방어 성능을 제공한다는 것을 확인하였다[5-7]. 입력 영상에 대한 디노이징은 동작 메커니즘에 따라 모델 기반 방법(예. BM3D, WNNM, EPLL)[8-10]과 학습 기반 방법(예. DnCNN, FFDNet)[11,12]으로 구분될 수 있는데, 전자는 후자에 비해 상대적으로 열세인 성능을 제공한다[13]. 반면에 학습 기반 방법은 모델을 훈련시키기 위한 대량의 데이터를 필요로 한다[14]. 따라서 유형별로 특징이 상이한 지문 영상의 경우, 학습 기반 방법들은 이러한 다양성을 고려한 다수의 데이터를 확보하지 않으면 유의미한 디노이징 성능을 확보하기 어렵다. 결과적으로 기존의 디노이징 기반 적대적 공격 방어 기법들은 지문 인증 시스템에서 제한적인 방어 성능을 보였다.

본 논문에서는 단일 영상 기반 자가지도학습 디노이징을 활용한 적대적 지문 방어를 제안하고, 방어 효용성을 검증한다. 구체적으로, 지문 인증 프로세스에 단일 영상 기반 자가지도학습 디노이징 기법을 적용하여, 입력 지문 영상에 추가된 적대적 노이즈를 효과적으로 제거하는지 다각도에서 검증한다. 단일 영상 기반 자가지도학습 디노이징 기법은 영상 복원에 사용되는 방법으로, 대용량 영상 데이터에 기반한 사전 학습 과정 없이도 목표 영상을 복원할 수 있다[15]. 예를 들어, 단일 영상으로부터 지정된 횟수 동안 반복적으로 사전(prior) 정보를 계산하고 손실(loss) 값을 갱신하는 학습을 수행함으로써 입력 영상을 원본 영상과 유사하게 복원할 수 있다[16]. 본 연구에서는 해당 방법을 통해 입력 지문 영상을 원본 지문 영상으로 복원하는 과정에서 적대적 노이즈를 제거하고, 복원된 영상이 적대적 공격의 방어에 효과적인지 검증한다. 제안하는 방어 메커니즘은 다양한 지문 유형을 고려한 대량의 데이터에 대한 학습 없이도, 입력된 지문 영상의 적대적 노이즈를 효과적으로 제거할 수 있을 뿐만 아니라, 기존의 복원 방법들보다 더욱 정교하게 원본 지문 영상을 복원할 수 있다. 결과적으로 제안하는 방법은 적대적 지문 공격에 대한 우수한 방어 성능을 제공함과 동시에, 지문 인증 시스템의 인증 성능을 보장할 수 있다.

본 논문의 공헌점은 다음과 같다. 다양한 종류의 적대적 지문 공격에 대해 일반화된 방어 성능을 제공하는 새로운 방어 기법을 제안한다. 또한, 제안하는 기법의 프로토타입을 구현하여 실제 지문 데이터에 적용 및 평가함으로써 본 연구의 효용성을 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 적대적 공격 기법과 생체인증 시스템에서의 적대적 공격 및 방어 기법에 관해 기술하였다. 3장에서는 본 연구에서 제안한 메커니즘을 기술하였다. 4장에서는 실제 지문 데이터를 바탕으로 수행한 실험 결과 및 결과에 대한 분석을 기술하였으며, 마지막으로 5장에서는 결론을 기술하였다.

Ⅱ. 배경 연구

2.1. 적대적 공격

적대적 공격(adversarial attack)이란 악의적인 공격자가 타겟 영상에 육안으로 식별이 어려운 노이즈를 추가하여, 공격자의 의도대로 타겟 딥러닝 모델이 오분류하게 만드는 공격이며, 적대적 노이즈가 추가된 영상을 적대적 예제라고 부른다[17] (그림 1 참고). FGSM(fast gradient sign method)[18]은 타겟 영상에 대한 손실 함수의 기울기(gradient)를 계산한 후, 기울기 부호와 반대되는 적대적 노이즈를 타겟 영상에 추가하는 대표적

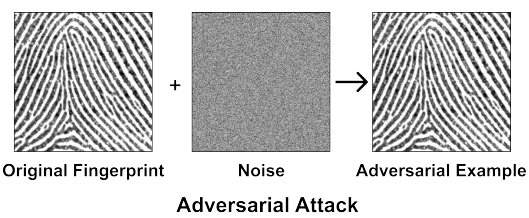

그림 1. 지문 영상에 대한 적대적 공격 예제
Fig. 1. An Example of Adversarial Fingerprint Attack

인 적대적 공격 방법이다. PGD(projected gradient descent)[19]는 FGSM를 개선한 방법으로, 손실 함수의 기울기 부호와 반대되는 방향으로 타겟 영상의 노이즈를 정해진 값만큼 반복 갱신하여 적대적 예제를 생성한다. Deepfool[20]은 타겟 모델의 잘못된 클래스에 대한 신뢰도를 높이는 방향으로 타겟 영상에 대한 노이즈를 구하고, 이를 타겟 영상에 육안으로 판별이 어려운 형태로 추가하는 적대적 공격 방법이다.

2.2 생체인증 시스템에서의 적대적 공격

딥러닝 기반 생체인증 시스템은 다양한 형태의 적대적 공격에 취약하다는 점이 밝혀졌다. Kwon 외 3인[3]은 적대적 노이즈 패치를 생성해 지문 영상에 삽입함으로써 딥러닝 기반 지문 인식기가 오분류하게 하는 적대적 공격 방법을 제안하였다. Fei 외 3인[21]은 세 가지의 적대적 공격 기법(FGSM, MI-FGSM, Deepfool)을 지문 영상에 적용시켜 딥러닝 기반 위조지문 판별 모델을 무력화시키는 방법을 제안하였으며, 회전 및 대칭(flip)과 같은 영상 변환 기법에도 강건한 적대적 공격 방법을 제시하였다. Marrone 외 4인[2]은 다양한 재료로 적대적 노이즈를 삽입한 복제 지문을 제작하였으며, 결과적으로 지문 인식 네트워크가 오분류하게 하는 물리적 도메인에서의 적대적 공격 방법을 제안하였다.

2.3 생체인증 시스템에서의 적대적 공격 방어

딥러닝 기반 생체인증 시스템이 적대적 공격에 대해 민감하고 취약하다는 점이 알려지면서, 이를 방어하기 위한 다양한 방법이 제안되었다[22-24]. DeepIris[22]는 입력받은 홍채 영상 내의 적대적 노이즈를 제거함으로써 홍채 인식 분류기가 정상적으로 식별을 수행하도록 하는 방어 기법이다. U-Net 기반 엔코더 구조(encoder architecture)를 사용하여 적대적 홍채 영상을 디노이징한 후, 노이즈를 추가하기 어려운 낮은 주파수 웨이블릿(frequence wavelet)에 집중하는 웨이블릿 변환(wavelet transformation)을 사용하여 적대적 홍채 영상 예제를 식별함으로써, 홍채 인식 시스템에 대한 적대적 공격을 방어한다. POSTER[23]는 적대적 오디오 예제에 새로운 저수준 왜곡(distortion)을 추가하면 분류 결과가 민감하게 변한다는 특성을 이용하여 적대적 예제를 탐지함으로써, 오디오 인식 시스템에 대한 적대적 공격을 방어한다. MagDR[24]는 딥페이크(deepfake) 영상의 이상치(abnormality)를 탐지하는 탐지 모듈과 복원을 통해 노이즈를 제거하는 재건 모듈로 구성되어 있다. 전자는 영상의 특정 영역에서 다양한 형태의 손상(corruption)을 판별하는 왜곡 탐지기(distortion detector)와 전체 영역에서 손상을 탐지하는 일관성 탐지기(consistency detector)로 이루어져 있고, 후자는 영상 복원 기술 중 하나인 Rec-Net을 사용하여 노이즈를 제거함으로써, 안면 인식 시스템에 대한 적대적 공격을 방어한다. 이러한 생체인증 시스템에서의 적대적 공격을 방어하기 위한 다양한 노력에도 불구하고, 아직까지 지문 인증 시스템에 대한 강건한 적대적 공격 방어 기법은 제안되지 않았다.

Ⅲ. 본 론

그림 2는 제안하는 방법의 전반적인 동작을 나타낸다. 학습 단계에서는 기존의 지문 인증 시스템과 마찬가지로 위조지문 판별 및 사용자 인증에 사용되는 분류기 모델을 훈련한다. 추론 단계는 세부적으로 총 세 가지 단계(지문 전처리 단계, 지문 복원 단계, 지문 인증 단계)로 구성된다.

지문 전처리 단계에서는 입력된 지문 영상 내 여백에 존재할 수 있는 적대적 노이즈를 최소화하기 위해 지문 외 여백 부분을 제거하고, 분류기 모델에 호환되는 크기로 영상 크기를 조정한다. 지문 복원 단계에서는 크기가 조정된 영상에 대해 단일 영상 기반 자가지도학습 디노이징을 수행하면서 노이즈를 제거하고, 원본 지문 영상과 가까운 형태로 복원한다. 지문 인증 단계에서는 복원된 지문 영상을 기반으로 위조지문 판별 및 사용자 인증을 수행한다. 각 단계의 세부 동작은 다음과 같다.

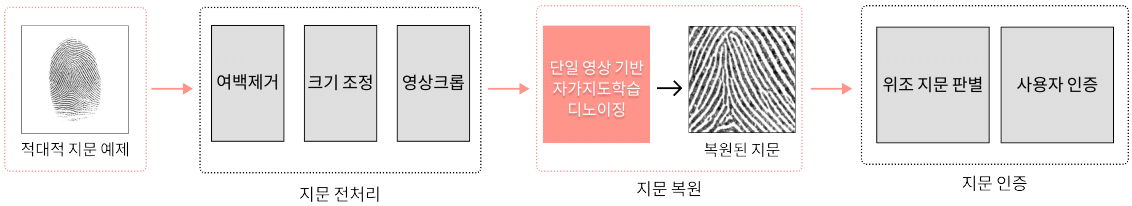

그림 2. 제안하는 방법의 개요
Fig. 2. An Overview of Our Proposed Method

3.1 지문 전처리

본 단계에서는 지문 외의 여백 영역에 추가될 수 있는 적대적 노이즈에 의한 영향을 최소화하기 위해, 여백 부분을 제거하고 지문 인증 분류기에 호환되는 크기로 영상 크기 조정을 수행한다. 여백 영역을 제거하는 방법으로는 픽셀 값(pixel value)를 기준으로 가장 높은 값을 가지는 영역(즉, 흰색 영역)을 우선적으로 제거하는 방법[25] 또는 객체탐지(object detection) 모델을 통해 유효 지문 영역을 찾는 방법[26]이 활용될 수 있다. 이 과정을 통해 지문의 유효 정보(예. 특징점)는 최대한 유지하면서 여백 영역을 제거할 수 있다. 결과적으로 지문 영역 외의 여백에 포함되어 있을 수 있는 적대적 노이즈를 제거함으로써 적대적 공격의 위험성을 감소시킬 수 있다. 여백 영역을 제거한 후에는, 지문 인증을 위한 분류기 모델에 호환되는 크기로 영상 크기 조정 혹은 특정 사이즈로의 크롭(cropping)을 수행한다. 본 연구에서는 181×181 크기[26]로 전처리한다.

3.2 지문 복원

본 단계에서는 단일 영상 기반 자가지도학습 디노이징을 통해 적대적 노이즈를 제거하는 동작을 수행한다. 단일 영상 기반 자가지도학습 디노이징은 대용량 영상 데이터를 통한 사전 학습 과정 없이도 입력 영상의 복원이 가능한 기술로서, 특정 반복 횟수만큼 원본 지문 영상에 대한 사전 정보(prior information)를 계산하고, 이에 따른 손실 값을 갱신하면서 학습을 수행한다. 이 과정을 통해, 입력 지문 영상을 원본 지문 영상과 유사하게 복원함으로써 적대적 노이즈를 제거한다. 대표적인 단일 영상 기반 자가지도학습 디노이징 방법인 deep image prior의 동작 예시는 그림 3과 같다. x_t는 원본 지문 영상, z는 랜덤 노이즈이며, x_n은 총 i번의 반복 횟수 중 n번째로 생성된 지문 영상이다. x_n은 생성기(generator) $x_n = f_{\theta_n}(z)$를 통해 생성되며, Θ는 영상 x에 매핑된 가중치를 의미한다. i번의 갱신 동안, 복원 과정의 지문 영상(x_n)이 원본 지문 영상(x_t)에 가까워지도록 하기 위해, 손실 함수 E가 작아지는 방향으로 Θ를 갱신한다. 총 i번의 갱신이 완료되면, 최종 지문 영상 x'_t가 생성된다.

갱신 초기에는 생성기가 원본 지문 영상의 사전 정보를 알 수 없기 때문에 노이즈가 완전히 제거되지 않았거나, 원본 지문 영상과 유사하게 복원되지 않았을 가능성이 높다. 하지만 반복 수행에 따라 손실 함수 E가 작아지는 방향으로 Θ를 갱신하면서, x_n이 원본 지문 영상

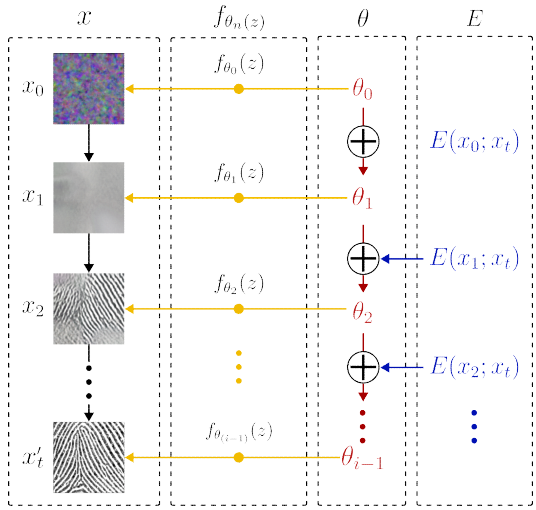

그림 3. 단일 영상 기반 자가지도학습 디노이징 프로세스
Fig. 3. The Process of Single-image-based Self-supervised Denoising

과 가깝게 복원될 가능성이 높아진다. 영상 내 노이즈는 일반적으로 높은 임피던스(impedance)로 인해 모델에 학습되기를 저항하다가 부자연스럽게 학습되므로, 급격하게 학습되는 구간 전에 갱신을 멈춤으로써 노이즈가 포함되지 않은 지문 영상을 생성할 수 있다.

3.3 지문 인증

본 단계에서는 복원된 지문 영상을 기반으로 위조지문 판별 및 사용자 인증을 수행한다. 구체적으로, 입력 지문 영상을 뉴럴 네트워크에 통과시켜서 라이브니스 점수(liveness score)를 산출하고, 사전에 정의된 임계치(threshold)와 비교하여 위조 여부를 판별한다. 위조 여부 판별이 끝난 지문 영상은 특징점(minutiae) 분석 및 매칭 등의 동작을 수행하여 사용자 인증을 수행한다. 위조 지문 판별 및 사용자 인증을 위한 뉴럴 네트워크의 구조는 목표 시스템의 구현 환경 및 리소스 상황에 따라 선별적으로 선택될 수 있다.

Ⅳ. 실험 결과 및 분석

제안된 방법의 효용성을 평가하기 위해 우리는 두 가지 주요 검증을 수행하였다. 첫 번째로, 제안한 방법이 다양한 종류의 적대적 지문 공격에 대해 효과적인 방어 성능을 제공하는지 검증하였다. 두 번째로, 제안한 방법이 기존의 영상 복원 방법보다 적대적 지문 공격에 대해 효과적인 방어 성능을 제공하는지 검증하였다.

4.1 실험 설정

제안하는 방법을 평가하기 위해, GreenBit, CrossMatch, Digital Persona 3개의 센서를 통해 취득한 실제 지문 영상으로 구성된 LivDet 2015 데이터셋을 사용하였다. 이 데이터셋은 영상 내 지문의 위치, 여백의 정도, 지문의 모양 등의 측면에서 다양한 형태의 지문 영상으로 구성된다. 본 연구에서 사용한 데이터셋의 구성은 다음과 같다: GreenBit(학습 데이터 2,000장, 평가 데이터 2,500장), CrossMatch(학습 데이터 2,983장, 평가 데이터 2,948장), Digital Persona(학습 데이터 2,000장, 평가 데이터 2,500장). 데이터셋별 학습 데이터로 각각 위조지문 판별 분류기를 훈련하였으며, 평가 데이터셋으로 판별 성능을 측정하여 방어 성능 평가를 수행하였다. 방어 성능의 척도로는 실제 live인 지문을 live로 예측한 개수와 실제 fake인 지문을 fake로 예측한 개수를 합하여, 전체 영상 개수로 나눈 값을 사용하였다. 단일 영상 기반 자가지도학습 디노이징 모델은 deep image prior 모델을 기반으로 구현되었다(iteration=800, learning rate=0.01)[27]. 위조지문 판별 분류기는 ResNet-50 구조를 기반으로 구현되었다(Tensorflow 2.7.0, optimizer=Adam, batch size=8, epochs=20, threshold=0.5). 모든 실험은 NVIDIA GeForce RTX 3090를 사용하여 Python 3.8.0 환경에서 수행되었다.

4.2 실험 결과

첫째로, 제안한 방법이 다양한 종류의 적대적 지문 공격에 대해 효과적인 방어 성능을 제공하는지 검증하기 위해, 세 가지 적대적 공격 기법들(FGSM, PGD, Deepfool) 별로 생성된 각각의 적대적 지문 영상에 우리의 방법을 통해 디노이징을 한 경우, 위조지문 판별 정확도가 상승하는지 확인하였다. 표 1에서 보여지듯이, FGSM으로 생성한 적대적 예제에 대한 판별 정확도는 GreenBit 62.08%, CrossMatch 57.64%, Digital Persona 62.68%였다. 해당 적대적 사례에 우리의 방법을 사용한 경우 판별 정확도는 GreenBit 68.57%, CrossMatch 85.93%, Digital Persona 65.47%였다. PGD로 생성한 적대적 예제에 대한 판별 정확도는 GreenBit 0.48%, CrossMatch 0.00%, Digital Persona 0.00%였다. 해당 적대적 사례에 우리의 방법을 사용한 경우 판별 정확도는 GreenBit 50.88%, CrossMatch 52.28%, Digital Persona 59.20%였다. Deepfool으로 생성한 적대적 예제에 대한 판별 정확도는 GreenBit 58.39%, CrossMatch 17.92%, Digital Persona 40.83%였다. 해당 적대적 사례에 우리의 방법을 사용한 경우 판별 정확도는 GreenBit 60.74%, CrossMatch 41.24%, Digital Persona 59.28%였다.

그림 4에서 보여지듯이, FGSM으로 생성한 적대적 예제에 대한 방어 성능 상승률은 GreenBit 6.49%p, CrossMatch 28.29%p, Digital Persona 2.79%p였다. PGD로 생성한 적대적 예제에 대한 방어 성능 상승률은 GreenBit 50.40%p, CrossMatch 52.28%p, Digital Persona 59.20%p였다. Deepfool로 생성한 적대적 예제에 대한 방어 성능 상승률은 GreenBit 2.35%p, CrossMatch 23.32%p, Digital Persona 18.45%p였다. 평균적으로 약 27.06%p 성능이 상승한 것으로써, 결론적으로 우리의 방법이 적대적 지문 예제에 의한 공격을 유의미한 성능으로 방어 가능함을 확인할 수 있다.

둘째로, 제안한 방법이 기존의 영상 복원 방법보다

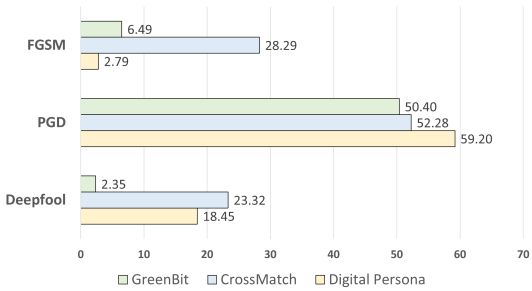

그림 4. 서로 다른 적대적 공격에 대한 방어 성능 상승률 비교
Fig. 4. Comparison of the Increased Defense Rate against Different Adversarial Fingerprint Attacks

표 1. 서로 다른 적대적 공격에 대한 제안한 방법의 방어 성능
Table 1. Defense Performance of Our Proposed Method against Different Adversarial Fingerprint Attacks

데이터셋	FGSM		PGD		Deepfool	
	적대적 예제	제안한 방법	적대적 예제	제안한 방법	적대적 예제	제안한 방법
GreenBit	62.08%	**68.57%**	0.48%	**50.88%**	58.39%	**60.74%**
CrossMatch	57.64%	**85.93%**	0.00%	**52.28%**	17.92%	**41.24%**
Digital Persona	62.68%	**65.47%**	0.00%	**59.20%**	40.83%	**59.28%**
평균 판별 정확도	60.80%	**73.32%**	0.16%	**54.12%**	39.05%	**53.75%**

적대적 지문 공격에 대한 효과적인 방어 성능을 제공하는지 검증하기 위해, FGSM으로 생성한 적대적 예제에 대해 기존의 디노이징 기법들(NLM, BM3D)로 각각 복원한 경우와, 우리의 방법을 통해 복원한 경우의 지문 판별 정확도를 확인하였다. 표 2에서 보여지듯이, FGSM으로 생성한 적대적 예제에 대한 판별 정확도는 GreenBit 62.08%, CrossMatch 57.64%, Digital Persona 62.68%였다. 해당 적대적 사례에 NLM(Non-Local Means)으로 디노이징한 경우 판별 정확도는 GreenBit 63.46%, CrossMatch 61.60%, Digital Persona 64.28%였다. 해당 적대적 사례에 BM3D(Block-Matching and 3D filtering)으로 디노이징한 경우 판별 정확도는 GreenBit 59.69%, CrossMatch 69.90%, Digital Persona 60.28%였다. 해당 적대적 사례에 우리의 방법으로 디노이징한 경우 판별 정확도는 GreenBit 68.57%, CrossMatch 85.93%, Digital Persona 65.47%였다.

그림 5에서 보여지듯이, NLM로 디노이징한 경우 적대적 예제에 대한 방어 성능 상승률은 GreenBit 1.38%p, CrossMatch 3.96%p, Digital Persona 1.60%p였다. BM3D로 디노이징한 경우 적대적 예제에 대한 방어 성능 상승률은 GreenBit -2.39%p, CrossMatch 12.26%p, Digital Persona -2.40%p였다. 우리의 방법을 사용하여 디노이징한 경우, 적대적 예제에 대한 방어 성능 상승률은 GreenBit 6.49%p, CrossMatch 28.29%p, Digital Persona 2.79%p였다. 평균적으로 NLM은 2.31%p 상승, BM3D는 2.49%p 성능 상승이 발생하는데, 우리의 방법을 사용하는 경우, 12.52%p 상승이 발생한다. 결론적으로 우리의 방법이 기존의 단일 영상 복원 방법들에 비해, 적대적 지문 예제에 의한 공격을 유의미한 성능으로 방어 가능함을 확인할 수 있다.

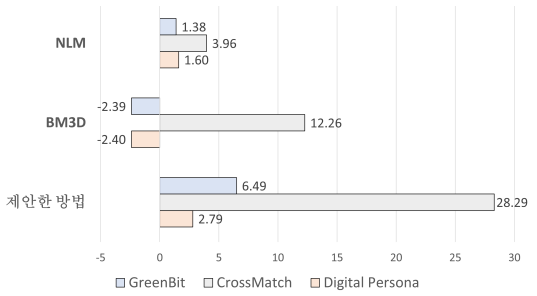

그림 5. 서로 다른 영상 복원 방법들 간의 방어 성능 상승률 비교
Fig. 5. Comparison of the Increased Defense Rate between Different Image Restoration Methods

V. 결 론

본 연구에서는 대용량의 지문 영상에 대한 사전 학습 없이도, 적대적 지문 공격의 효과적인 방어가 가능한 적대적 지문 방어 기법을 제안하였다. 제안하는 방법은 지문 인증 과정에서 단일 영상 기반 자가지도학습 디노이징을 활용하여 입력된 지문 영상 내 적대적 노이즈를 효과적으로 제거하고, 동시에 원본 영상과 최대한 유사하게 복원한다. 제안한 방법의 효용성을 평가하기 위하여 우리는 제안된 방법의 프로토타입을 구현하고, 실제 지문 데이터셋을 사용하여 다각도 검증을 진행하였다. 검증 결과에 따르면, 제안한 방법이 다양한 종류의 적대적 지문 공격 및 지문 형태에 대해서 일반화된 방어 성능을 제공할 뿐만 아니라, 기존의 다른 영상 복원 방법들보다 더욱 강건한 방어 성능을 제공하는 것을 확인하였다. 결과적으로 제안하는 방법이 적대적 지문 공격을 효과적으로 방어할 수 있고, 동시에 지문 인증 시스템의 인증 성능을 보장할 수 있음이 검증되었다.

향후 연구로는 기존 지문 인증 시스템에서 요구되는 빠른 처리 성능을 확보하기 위해, 입력된 지문 영상에 대한 복원 시간을 최소화하기 위한 연구를 진행할 예정이다. 또한, 입력된 지문 영상의 일부 손상 혹은 화질의 열화가 발견되는 경우, 자동으로 인페인팅(inpainting)과 슈퍼-레졸루션(super-resolution) 메커니즘을 적용하여, 지문 인증 시스템의 성능을 향상시킬 수 있는 방안을 연구할 예정이다.

표 2. 기존 영상 복원 방법들과 제안된 방법의 정확도 측정 결과
Table 2. Accuracy Comparison between Existing Image Restoration Methods and Our Proposed Method

데이터셋	적대적 예제	NLM	BM3D	제안한 방법
GreenBit	62.08%	63.46%	59.69%	**68.57%**
CrossMatch	57.64%	61.60%	69.90%	**85.93%**
Digital Persona	62.68%	64.28%	60.28%	**65.47%**
평균 판별 정확도	60.80%	63.11%	63.29%	**73.32%**

References

[1] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, Feb. 2021. (https://doi.org/10.1016/j.patcog.2020.107332)

[2] S. Marrone, R. Casula, G. Orrù, G. L. Marcialis, and C. Sansone, "Fingerprint adversarial presentation attack in the physical domain," in *2021 Int. Conf. Pattern Recognition Wkshps. and Challenges*, pp. 530-543, Virtual Event, Jan. 2021. (https://doi.org/10.1007/978-3-030-68780-9_42)

[3] H. W. Kwon, J.-W. Nam, J. Kim, and Y. K. Lee, "Generative adversarial attacks on fingerprint recognition systems," in *2021 Int. Conf. Inf. Netw.*, pp. 483-485, Jeju Island, Korea, Jan. 2021. (https://doi.org/10.1109/ICOIN50884.2021.9333904)

[4] R. R. Wiyatno, A. Xu, O. Dia, and A. de Berker, "Adversarial examples in modern machine learning: A review," *arXiv preprint,* arXiv:1911.05268, 2019. (https://doi.org/10.48550/arXiv.1911.05268)

[5] Y. Zhang, H. Xu, C. Pei, and G. Yang, "Adversarial example defense based on image reconstruction," *PeerJ Comput. Sci.*, vol. 7, p. e811, Dec. 2021. (https://doi.org/10.7717/peerj-cs.811)

[6] S. Zhang, H. Gao, and Q. Rao, "Defense against adversarial attacks by reconstructing images," *IEEE Trans. Image Process.*, vol. 30, pp. 6117-6129, Jul. 2021. (https://doi.org/10.1109/TIP.2021.3092582)

[7] A. Sinha, S. P. Dash, and N. B. Puhan, "Nomaro: Defending against adversarial attacks by noma-inspired reconstruction operation," *IEEE Sensors Lett.*, vol. 6, no. 1, pp. 1-4, Dec. 2021. (https://doi.org/10.1109/LSENS.2021.3135433)

[8] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080-2095, Jul. 2007. (https://doi.org/10.1109/TIP.2007.901238)

[9] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *2014 IEEE Conf. Comput. Vision and Pattern Recognition*, pp. 2862-2869, Columbus, OH, USA, Jun. 2014. (https://doi.org/10.1109/CVPR.2014.366)

[10] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *2011 IEEE Int. Conf. Comput. Vision*, pp. 479-486, Barcelona, Spain, Nov. 2011. (https://doi.org/10.1109/ICCV.2011.6126278)

[11] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142-3155, Feb. 2017. (https://doi.org/10.1109/TIP.2017.2662206)

[12] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608-4622, May 2018. (https://doi.org/10.1109/TIP.2018.2839891)

[13] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Netw.*, vol. 131, pp. 251-275, Nov. 2020. (https://doi.org/10.1016/j.neunet.2020.07.025)

[14] A. E. Ilesanmi and T. O. Ilesanmi, "Methods for image denoising using convolutional neural network: A review," *Complex & Intell. Syst.*, vol. 7, no. 5, pp. 2179-2198, Jun. 2021. (https://doi.org/10.1007/s40747-021-00428-4)

[15] Y. Quan, M. Chen, T. Pang, and H. Ji, "Self2self with dropout: Learning self-supervised denoising from single image," in *2020 IEEE/CVF Conf. Comput. Vision and Pattern Recognition*, pp. 1887-1895, Seattle, WA, USA, Jun. 2020.

(https://doi.org/10.1109/CVPR42600.2020.00196)

[16] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *2018 IEEE Conf. Comput. Vision and Pattern Recognition*, pp. 9446-9454, Salt Lake City, UT, USA, Jun. 2018. (https://doi.org/10.1109/CVPR.2018.00984)

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint*, arXiv:1312.6199, 2013. (https://doi.org/10.48550/arXiv.1312.6199)

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint*, arXiv:1412.6572, 2014. (https://doi.org/10.48550/arXiv.1412.6572)

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint*, arXiv:1706.06083, 2017. (https://doi.org/10.48550/arXiv.1706.06083)

[20] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *2016 IEEE Conf. Comput. Vision and Pattern Recognition*, pp. 2574-2582, Las Vegas, NV, USA, Jun. 2016. (https://doi.org/10.1109/CVPR.2016.282)

[21] J. Fei, Z. Xia, P. Yu, and F. Xiao, "Adversarial attacks on fingerprint liveness detection," *EURASIP J. Image and Video Process.*, vol. 2020, no. 1, pp. 1-11, Jan. 2020. (https://doi.org/10.1186/s13640-020-0490-z)

[22] T. S R, A. Ojha, M. K, and G. Maragatham, "Deepiris: An ensemble approach to defending Iris recognition classifiers against adversarial attacks," in *2021 Int. Conf. Comput. Commun. and Informatics*, pp. 1-8, Coimbatore, India, Jan. 2021. (https://doi.org/10.1109/ICCCI50826.2021.9402404)

[23] H. Kwon, H. S. Yoon, and K. W. Park, "Poster: Detecting audio adversarial example through audio modification," in *2019 ACM SIGSAC Conf. Comput. and Commun. Secur. Assoc. for Computing Mach.*, pp. 2521-2523, New York, NY, USA, Nov. 2019. (https://doi.org/10.1145/3319535.3363246)

[24] Z. Chen, L. Xie, S. Pang, Y. He, and B. Zhang, "Magdr: Mask-guided detection and reconstruction for defending deepfakes," in *2021 IEEE/CVF Conf. Comput. Vision and Pattern Recognition*, pp. 9010-9019, Nashville, TN, USA, Jun. 2021. (https://doi.org/10.1109/CVPR46437.2021.00890)

[25] Y. Zhang, D. Shi, X. Zhan, D. Cao, K. Zhu, and Z. Li, "Slim-rescnn: A deep residual convolutional neural network for fingerprint liveness detection," *IEEE Access*, vol. 7, pp. 91476-91487, Jul. 2019. (https://doi.org/10.1109/ACCESS.2019.2927357)

[26] M. Y. Lim, T. Y. Kim, J. E. Park, P. M. Hong, and Y. K. Lee, "A central point-based analysis for fingerprint liveness detection," in *2022 13th Int. Conf. Inf. and Commun. Technol. Convergence*, pp. 1307-1309, Jeju Island, Korea, Oct. 2022. (https://doi.org/10.1109/ICTC55196.2022.9952772)

[27] D. Ulyanov, (2017) *Deep image prior*. [Online]. Available: https://github.com/DmitryUlyanov/deep-image-prior

[28] S. Shim and D. H. Lee, "Research trends in physical adversarial attack techniques," in *Symp. Korean Inst. Commun. and Inf. Sci.*, pp. 1453-1455, Pyeongchang, Korea, Feb. 2022.

홍 표 민 (Pyo Min Hong)

2023년 2월 : 서울여자대학교 정보보호학과 졸업
2023년~현재 : 홍익대학교 컴퓨터공학과 석사과정
<관심분야> 딥러닝, 컴퓨터 비전, 정보보호

[ORCID:0009-0003-8138-798X]

유 화 정 (Hwajung Yoo)

2023년 2월 : 홍익대학교 컴퓨터공학과 졸업
<관심분야> 딥러닝, 컴퓨터 보안

김 태 용 (Taeyong Kim)

2020년~현재 : 홍익대학교 컴퓨터공학과 학사과정
<관심분야> 딥러닝, 컴퓨터 비전, 생체 보안

윤 정 원 (Jung Won Yoon)

2023년 2월 : 홍익대학교 컴퓨터공학과 졸업
2023년~현재 : 홍익대학교 컴퓨터공학과 석사과정
<관심분야> 머신러닝, 인공지능, 정보보호

김 태 형 (Tae Hyung Kim)

2011년 8월 : 카이스트 전기 및 전자공학과 졸업
2014년 5월 : 미시간 대학교 전자공학과 석사
2020년 12월 : 서던캘리포니아대학교 전자공학과 박사

2020년~2022년 : Massachusetts General Hospital / Harvard Medical School 연구원
2022년~현재 : 홍익대학교 컴퓨터공학과 조교수
<관심분야> 의료영상, 자기공명영상, 인공지능 영상처리, 신호처리, 딥러닝

이 윤 규 (Youn Kyu Lee)

2010년 2월 : 고려대학교 컴퓨터학과 졸업
2012년 2월 : 고려대학교 컴퓨터학과 석사
2017년 12월 : 서던캘리포니아대학교 컴퓨터학과 박사

2018년~2020년 : 삼성전자 종합기술원 책임연구원
2020년~2021년 : 서울여자대학교 정보보호학과 조교수
2021년~현재 : 홍익대학교 컴퓨터공학과 조교수
<관심분야> 소프트웨어 설계, 보안, 딥러닝
[ORCID:0000-0002-4569-2640]