

온디바이스 AI 환경에서 잠재 확산 모델을 활용한 전처리 아티팩트 기반 이미지 재구성 공격 시스템

방한솔*, 최진구**

Preprocessing Artifact-based Image Reconstruction Attack System using Latent Diffusion Models in On-Device AI Environments

Hansol Bang*, Jinku Choi**

이 연구는 2026년도 산업통상부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임('RS-2023-00254096')

요 약

최근 딥러닝과 하드웨어의 발전으로 확산되고 있는 온디바이스 AI 환경에서 데이터 전처리에 의해 저해상도 이미지나 Canny 에지와 같은 아티팩트가 생성된다. 원본 데이터와 다르게 비교적 중요도가 낮게 인식되는 아티팩트는 평문으로 저장되어 보안상의 문제를 야기할 수 있다. 본 논문은 이러한 취약점을 지적하기 위해 잠재 확산 모델을 활용하여 아티팩트만으로 원본과 유사한 이미지를 생성하는 공격 시스템을 제안한다. 제안 시스템의 복원 결과를 딥러닝 분류 모델로 평가한 결과 Top-1 및 Top-5 정확도는 각각 41.1%, 67.6%를 보였으며, 소규모 데이터로 미세 조정 시 각각 53.4%, 76.6%로 정확도의 향상을 보였다. 또한 ROC 곡선에서 산출한 임계값으로 레이블 추출 공격을 수행 시 81.18%의 정확도를 달성할 수 있음을 보였다.

Abstract

In the on-device AI environments, which is proliferating due to recent advancements in deep learning and hardware, data preprocessing generates artifacts such as low-resolution images and Canny edges. Unlike the original data, these artifacts are generally perceived to be of relatively low importance and are stored in plaintext, potentially leading to security vulnerabilities. To highlight this vulnerability, this paper proposes an attack system that utilizes a latent diffusion model to generate images highly similar to the original, relying solely on these artifacts. Evaluating the restoration results of the proposed system using a deep learning classification model yielded Top-1 and Top-5 accuracies of 41.1% and 67.6%, respectively. When fine-tuned with a small dataset, the accuracies further improved to 53.4% and 76.6%, respectively. Furthermore, we demonstrated that a label extraction attack utilizing a threshold derived from the ROC curve can achieve an accuracy of 81.18%.

Keywords

on-device AI, latent diffusion model, artifacts, privacy, reconstruction attack

* 한국공학대학교 컴퓨터공학과

- ORCID: <https://orcid.org/0009-0001-1102-6113>

** 한국공학대학교 컴퓨터공학부 교수(교신저자)

- ORCID: <https://orcid.org/0009-0008-0659-1274>

• Received: Jun. 16, 2026, Revised: Jul. 29, 2026, Accepted: Aug. 01, 2026

• Corresponding Author: Jinku Choi

Dept. of Computer Science and Engineering, Tech University of Korea

237, Sangidaehak-ro, Siheung-si, Gyeonggi-do, Republic of Korea

Tel.: +82-31-8041-0521, Email: jkchey@tukorea.ac.kr

I. 서 론

최근 고성능 GPU 서버나 클라우드에 의존하지 않고 기기에서 추론하는 온디바이스 AI(On-Device AI) 환경은 딥러닝 기술의 성능 발전과 모델 최적화에 의해 점차 확산되고 있다.

온디바이스 AI는 데이터 전송 지연을 최소화하는 실시간성을 보장하고 사용자의 프라이버시를 보호할 수 있어 차세대 인공지능의 핵심 기술로 자리 잡고 있다. 온디바이스 AI의 구동을 위한 엣지 디바이스(Edge Devices)에는 스마트폰, 스마트홈 시스템, 자율주행차량 등이 포함된다. 일반적으로 온디바이스 AI 서비스는 엣지 디바이스를 활용하여 IP 카메라와 같은 IoT 디바이스에서 생성되는 센서 데이터를 실시간으로 수집하고 컴퓨터 비전을 활용하여 데이터를 처리하는 구조로 배포된다. 이러한 서비스 운영에서 수행되는 데이터 전처리는 파생 데이터인 아티팩트(Artifact)를 생성한다. 아티팩트는 원본 데이터를 가공한 데이터로 썸네일(Thumbnail)이나 파일 아이콘 이미지와 같이 저해상도로 변환하거나, 이미지의 구조적 정보를 추출하는 Canny 에지 검출과 같은 다양한 형태가 이에 해당된다.

아티팩트는 개인정보를 포함할 가능성이 높음에도 IoT 디바이스 및 엣지 디바이스의 데이터 보안은 서비스 과정에서 비교적 주요하게 고려되지 않는 실정이다. IoT 디바이스는 네트워크에 상시 연결되어 있어 공격 목표로 선정될 가능성이 높다. 그러나 하드웨어 성능 제약으로 인해 고성능 암호화 알고리즘 적용이 제한되어 공격에 취약하다. 기기의 취약점을 통해 원본 데이터를 탈취 후 데이터의 배포까지 이어진 실제 공격 사례가 존재한다[1]. 엣지 디바이스 기반 AI 비전 시스템은 신뢰 실행 환경(TEE, Trusted Execution Environment)과 같은 격리된 실행 공간을 활용하여 민감한 원본 데이터를 보호할 수 있다. 그러나 원본 데이터가 안전하게 보호되더라도, 추론 과정에서 생성되는 전처리 아티팩트는 상대적으로 낮은 보호 우선순위로 관리될 수 있다. 이에 따라 공격자는 원본 데이터의 특성을 일부 포함하는 아티팩트를 활용하여 원본 정보를 간접적으로 추론하는 우회 공격을 수행할 수 있다. 특히 썸네일과 Canny 에지와 같은 아티팩트는 시스템 성능

향상과 실시간 추론을 위해 생성되며, 원본 데이터에 비해 상대적으로 낮은 보안 수준으로 메모리 또는 스토리지에 저장될 수 있다. 실제로 원본 데이터가 삭제된 이후에도 잔존하는 썸네일을 복원한 연구 결과가 보고되었다[2].

본 연구에서는 방치된 아티팩트로 원본과 유사한 이미지로 복원하는 공격 시스템을 제안한다. 본 시스템은 Canny 에지와 저해상도 이미지를 잠재 확산 모델(Latent diffusion model)에 입력하여 원본 이미지로 재구성한다. 본 연구 결과는 전처리 아티팩트가 단순한 임시 데이터가 아니라 온디바이스 AI 환경에서 보호되어야 하는 민감한 파생 데이터임을 보여준다.

본 논문은 다음과 같이 구성된다. 2장에서는 데이터 복원 공격과 기기 내에서 생성되는 아티팩트의 보안 문제에 관한 관련 연구를 살펴본다. 3장에서는 Canny 에지 검출 결과와 저해상도 이미지를 이용한 Stable Diffusion 및 ControlNet 기반 복원 공격 모델을 설명한다. 4장에서는 ImageNet-1K와 COCO 2017 데이터셋을 이용한 실험 결과를 제시하고, 복원 이미지의 시각적 유사성과 인식 가능성을 평가한다. 마지막으로 5장에서는 본 연구의 결과와 향후 연구 방향을 제시한다.

II. 관련 연구

2.1 데이터 복원 공격과 프라이버시 위협

딥러닝 모델은 대규모 데이터로부터 특징과 패턴을 학습하여 분류, 탐지, 생성 등의 작업을 수행한다. 딥러닝 모델의 성능 확보를 위한 노력으로 모델 구조를 개선하고, 대표성을 가진 대규모의 고품질 데이터를 수집한다. 그러나 데이터셋의 규모가 방대해짐에 따라 체계적인 관리의 어려움이 나타나고 있다. 공개 데이터셋에 동의하지 않은 개인정보나 비윤리적인 데이터가 포함되기도 하고, 해당 데이터로 학습된 모델이 배포되기도 한다. 이러한 문제는 관리상의 부주의로 인해 발생되므로, 체계적인 관리 방안으로 피해를 완화할 수 있다. 반면 구조적으로 발생하는 문제들도 존재한다. Gradient Inversion Attack은 역전파의 기울기(Gradient)를 사용하여 입

력 데이터의 복원 가능성을 보였다. 이미지 분류(Image classification)에 사용되는 MNIST 외 네 가지 데이터셋에서 MSE(Mean Squared Error)가 모두 0.03 이하인 결과를 보였다[3]. Model Inversion Attack은 CNN 모델의 각 계층(Layer)에서 생성되는 특징 맵(Feature map)을 탈취하여 업샘플링(Up-sampling)하는 그림자 모델(Shadow model)을 통해 입력 데이터를 복원하고자 하였다[4].

2.2 생성 모델

과거 GAN(Generative Adversarial Network)은 생성 모델로서 주목받았으나, 생성자가 판별자를 속이기 쉬운 클래스만 학습하는 모드 붕괴(Mode collapse)가 발생하였다. 이는 데이터셋의 다양성을 학습하지 못하여 반복적으로 유사한 결과물을 출력한다. 현재 이미지 생성 분야에서 주류로 자리 잡은 확산 모델(Diffusion model)은 마르코프 체인(Markov chain)의 성질을 응용하여 소량의 노이즈를 데이터에 추가 및 제거하는 샘플링 과정을 반복하여 데이터를 생성한다. 초기의 확산 모델은 샘플링을 픽셀 공간(Pixel space)에서 실시하여 높은 연산량과 메모리를 요구하였다. 잠재 확산 모델은 픽셀 공간의 데이터를 변분 오토인코더(Variational autoencoder)로 압축한 잠재 공간(Latent space)에서 샘플링을 수행하여 연산량과 메모리 사용량을 줄였다[5].

2.3 엣지 디바이스 환경과 아티팩트 보안

기계 학습 운용(MLOps)의 아티팩트는 기계 학습 생애주기(ML lifecycle)의 모든 산출물을 의미한다. 생애주기 중 데이터 지향 단계(Data-oriented stage)는 학습을 위한 데이터를 수집 및 전처리하는 단계이다. 해당 단계의 아티팩트에는 데이터셋, 레이블(Labels) 및 주석(Annotations), 특징 집합(Feature sets) 등이 포함된다. 특징 집합은 머신 러닝 모델의 입력으로 사용하기 위해 추출 및 선택된 특징들의 집합이다[6]. 포렌식 분야의 아티팩트는 사건의 진실을 밝히기 위한 포렌식 과정에서 조사관이 수집하는 시스템에 남겨진 흔적을 의미한다. 기록 로그(History logs), 임시 데이터(Temp data), 채팅 로그

(Chat logs)등이 이에 해당한다[7]. 안드로이드 OS의 기본 갤러리 앱에서 미리보기를 위해 생성되는 썸네일을 복원한 사례도 존재한다. 원본 이미지가 삭제되더라도 썸네일은 기기 내부에 저장되어 있어 복원 가능성을 확인하였다[8].

본 연구에서는 이러한 보안 취약점들에 주목하여, 아티팩트를 활용한 원본 데이터의 복원 가능성을 분석하고자 한다. 표 1에 나타난 바와 같이, 제안하는 방식은 Gradient Inversion Attack, Model Inversion Attack과 달리 학습을 요구하지 않는다. 제안 시스템은 잠재 확산 모델을 플러그 앤 플레이(Plug-and-play) 방식으로 활용할 수 있다는 장점이 있다. 구체적으로 에지 검출 결과와 저해상도 이미지를 활용하여 잠재 확산 모델인 Stable Diffusion으로 원본 이미지와 시각적으로 유사한 고해상도 이미지를 생성하여 복원 가능성을 검증하고, 그 결과를 비교·분석한다.

표 1. 기존 복원 공격 연구와 본 연구의 비교

Table 1. Comparison of existing reconstruction attacks and the proposed method

Category	Gradient inversion attack[3]	Model inversion attack[4]	Ours
Access target	Shared gradients	Feature map	Artifacts
Restoration	Gradient matching	Up-conv network	Stable diffusion-ControlNet
Result	Pixel-level recovery	Approximate recovery	Similar high-res image
Limitation	2nd-order differentiable model required	Detail loss in deeper layers	Artifact leakage
Distinctiveness	Gradient-only data theft	Single forward-pass reconstruction	Plug-and-play reconstruction

III. 제안 시스템 및 구현

3.1 위협 모델

본 연구의 위협 모델은 온디바이스 AI 서비스를

운영하는 엣지 디바이스의 기계 학습 운용(MLOps) 환경으로 가정한다. 공격자의 목표는 디바이스 내부의 원본 데이터에 직접 접근하지 않고, 유출된 아티팩트를 이용하여 민감한 고해상도 원본 데이터와 유사한 이미지를 재구성하는 것이다. 일반적으로 원본 데이터는 암호화되어 있거나 격리된 보안 공간에서 보호되므로, 제한된 접근 권한을 가진 공격자는 원본 데이터를 탈취하기 어렵다. 공격자는 엣지 디바이스의 취약점으로 아티팩트를 획득하여 복원하는 우회 공격을 시도한다. 대상 아티팩트는 추론 연산량을 줄이기 위해 이미지에서 관심 영역(ROI, Region of Interest)을 추출하는 전처리 과정의 중간 산출물인 Canny 에지와 시스템에 저장된 임시 데이터인 저해상도 썸네일 이미지이다. 원본 이미지 삭제 후에도 내부에는 썸네일이 저장되어 있음을 확인한 선행 연구와 같이, 이러한 아티팩트들은 메모리나 스토리지에 평문으로 저장되는 취약점을 지닌다. 공격자는 악성 애플리케이션을 통해 접근 권한을 우회하여 임시 디렉토리에 보관된 해당 아티팩트를 탈취할 수 있다. 공격자는 이를 활용하여 획득한 아티팩트들을 잠재 확산 모델에 입력하여 원본 이미지와 시각적으로 유사한 데이터로 재구성하는 공격을 수행한다. 따라서 본 연구에서는 원본 이미지에 대한 직접적인 접근 권한이 없는 상황에서도, 탈취한 여러 아티팩트의 시각적·구조적 정보를 결합하여 수행되는 복원 공격의 유효성을 검증하고자 한다.

3.2 제안 시스템의 구조

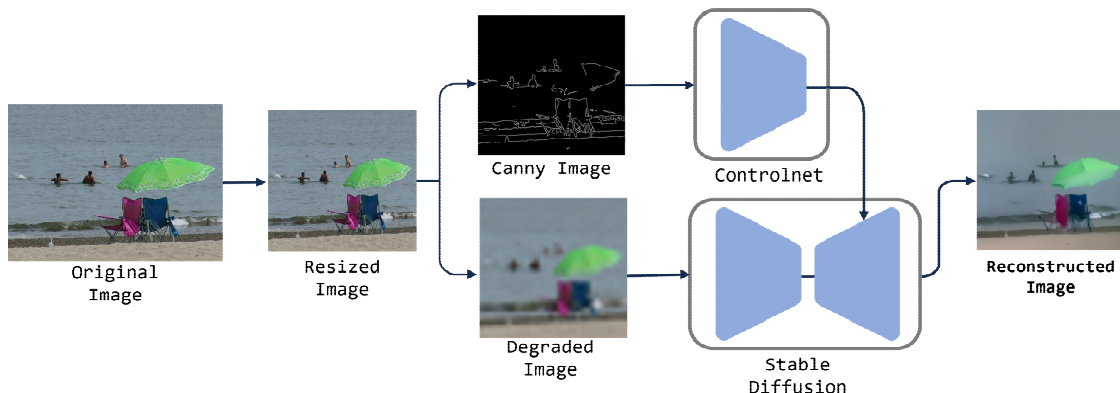


그림 1. Canny 에지와 저해상도 이미지를 이용한 아티팩트 기반 이미지 재구성 공격 시스템 구조
Fig. 1. Architecture of the artifact-based image reconstruction attack system using Canny edge and degraded image

아티팩트를 활용한 복원 공격의 전체 시스템 구조는 그림 1과 같다. 먼저 위협 모델의 공격 과정을 모사하기 위해, 원본 이미지를 전처리하여 Canny 에지와 저해상도 이미지 형태의 아티팩트로 각각 변환한다. 이후 Stable Diffusion은 입력된 저해상도 이미지를 잠재 공간으로 압축하고 설정된 비율만큼 가우시안 노이즈를 주입한다. ControlNet은 공간적 조건으로 Canny 에지를 입력받아 Stable Diffusion의 샘플링 단계에 주입한다[9]. 최종적으로 본 시스템은 Canny 에지의 공간적 조건을 강제하는 샘플링 과정을 반복 수행하여, 원본과 시각적으로 유사한 고해상도 복원 이미지를 생성한다.

3.3 아티팩트 생성 과정

공격 시나리오에 해당하는 구체적인 아티팩트 생성 과정은 다음과 같다. 먼저, 모델의 입력 크기에 맞추기 위해 모든 데이터셋을 512×512 해상도의 정방형 이미지로 리사이징한다. 이미지 품질을 우선시한 방법으로는 주로 쌍입방 보간법(Bicubic interpolation)과 란초스 보간법(Lanczos interpolation)이 고려된다.

란초스 보간법은 시각적 선명도는 높게 인식되나, 링잉(Ringing) 현상이 발생하여 에지로 검출될 수 있다. 따라서 그라디언트(Gradient)가 평활하여 아티팩트가 비교적 적게 발생하는 안정적인 쌍입방 보간법을 채택하였다[10]. 이후 전처리가 완료된 이미지를 바탕으로 Canny 에지를 검출하고 열화된 이미지를 생성한다.

각 데이터셋은 평균적으로 482×418, 640×480의 해상도를 가져 딥러닝 모델이 요구하는 입력 크기에 맞추는 리사이징이 필수적이다. 성능 평가에는 ControlNet의 Canny 버전과 이와 호환되는 Stable Diffusion v1.5를 기반 모델로 설정하였다. 실험 데이터는 각 데이터셋에서 무작위 추출한 1,000장의 이미지로 구성하였다. ImageNet은 클래스당 1장씩 배치하였으나, COCO는 단일 이미지 내에 객체가 다수 존재하고 인스턴스 분포가 불균형한 롱테일(Long-tail) 형태를 보인다. 따라서 소수 클래스 기준으로 균형을 맞추거나 단순 이미지 단위로 추출할 경우 데이터셋 본래의 특성이 왜곡될 수 있다. 이에 추출된 데이터셋이 검증 데이터셋의 분포를 따르는지 확인하고자 클래스별 객체 인스턴스 수를 비교하였다. 그림 2에서 부분집합은 전체 검증 데이터셋과 유사한 롱테일 분포를 보이며, 이는 균형 표본이 아닌 검증 데이터셋의 일반적 분포를 반영한 무작위 표본 기반 평가임을 의미한다.

4.2 평가 지표

정량적 평가를 위해 레이블이 제공되는 검증 데이터셋을 채택하였다. 평가 지표는 생성 모델에서 주로 사용되는 SSIM, PSNR, LPIPS를 채택하였다. PSNR은 픽셀 단위의 오차를 측정하고, SSIM은 이미지의 구조적 유사성을 평가한다[13]. LPIPS는 학습된 딥러닝 모델의 특징 표현을 이용하여 인간의 시각적 인지와 유사한 지각적 품질을 측정한다[14].

4.3 복원 품질 평가 결과

노이즈 주입 수준이 1.0인 조건은 입력 이미지가 배제된 상태와 동일하며, ControlNet의 제어 가중치가 0인 경우는 Stable Diffusion 단독 모델을 구동한 것과 같다. 본 연구에서는 열화 이미지와 Canny 에지가 복원 결과에 영향을 미치지 않는 조건을 의도적으로 설정하여 성능 평가의 기준으로 삼았다.

해당 지표로 산출된 데이터셋별 시험 결과를 표 2와 표 3에 제시하였다. 모든 평가 지표를 고려하기 위해 최소-최대 정규화(Min-max normalization)를 적용한 결과를 표 4와 표 5에 수록하였다. 각 평가 지

표에 따른 최적의 파라미터는 다음과 같다. ImageNet-1K 데이터셋에서 SSIM의 최적 노이즈 주입 강도와 제어 가중치는 각각 0.5와 2.0으로 도출되었다. PSNR의 경우 0.4와 2.0, LPIPS의 경우 0.6과 2.0을 나타내었다. COCO 데이터셋의 SSIM은 0.5와 2.0, PSNR의 경우 0.4와 2.0, LPIPS의 경우 0.6과 2.0을 보였다. 세 성능 지표를 모두 고려하기 위한 정규화한 결과에서는 최적 노이즈 주입 강도와 제어 가중치는 각각 0.5와 2.0으로 도출되었다. 대체적으로 노이즈 주입 강도는 0.4에서 0.6일 때, 제어 가중치는 최대치인 2.0일 때 최적인 결과를 보였다.

표 2. 복원 이미지 성능 평가 결과(ImageNet)
Table 2. Reconstructed image performance(ImageNet)

Noise level	Control weight	Mean SSIM (↑)	Mean PSNR (↑)	Mean LPIPS (↓)
0.4	0.0	0.518	19.934	0.628
0.4	0.5	0.523	20.13	0.609
0.4	1.0	0.53	20.311	0.584
0.4	1.5	0.537	20.467	0.563
0.4	2.0	0.543	20.585	0.548
0.5	0.0	0.501	18.888	0.634
0.5	0.5	0.511	19.285	0.602
0.5	1.0	0.525	19.679	0.558
0.5	1.5	0.538	19.948	0.533
0.5	2.0	0.548	20.065	0.523
0.6	0.0	0.473	17.279	0.645
0.6	0.5	0.489	17.947	0.603
0.6	1.0	0.514	18.625	0.542
0.6	1.5	0.535	18.966	0.52
0.6	2.0	0.542	18.854	0.515
0.7	0.0	0.437	15.351	0.661
0.7	0.5	0.455	16.142	0.612
0.7	1.0	0.491	17.123	0.544
0.7	1.5	0.512	17.333	0.522
0.7	2.0	0.511	16.935	0.521
0.8	0.0	0.395	13.349	0.683
0.8	0.5	0.412	14.283	0.626
0.8	1.0	0.454	15.375	0.558
0.8	1.5	0.473	15.41	0.536
0.8	2.0	0.463	14.76	0.539
0.9	0.0	0.34	11.311	0.712
0.9	0.5	0.365	12.439	0.648
0.9	1.0	0.416	13.697	0.577
0.9	1.5	0.435	13.61	0.555
0.9	2.0	0.42	12.928	0.56
1	0.0	0.289	9.779	0.734
1	0.5	0.32	10.983	0.675
1	1.0	0.383	12.302	0.601
1	1.5	0.402	12.113	0.581
1	2.0	0.391	11.453	0.584

표 3. 복원 이미지 성능 평가 결과(COCO)

Table 3. Reconstructed image performance(COCO)

Noise level	Control weight	Mean SSIM (↑)	Mean PSNR (↑)	Mean LPIPS (↓)
0.4	0.0	0.509	19.644	0.658
0.4	0.5	0.515	19.844	0.638
0.4	1.0	0.522	20.029	0.610
0.4	1.5	0.530	20.196	0.587
0.4	2.0	0.537	20.322	0.570
0.5	0.0	0.494	18.652	0.664
0.5	0.5	0.504	19.066	0.628
0.5	1.0	0.519	19.462	0.580
0.5	1.5	0.534	19.744	0.552
0.5	2.0	0.545	19.872	0.539
0.6	0.0	0.468	17.079	0.673
0.6	0.5	0.484	17.762	0.627
0.6	1.0	0.512	18.472	0.561
0.6	1.5	0.534	18.794	0.534
0.6	2.0	0.545	18.710	0.526
0.7	0.0	0.434	15.211	0.689
0.7	0.5	0.452	15.994	0.635
0.7	1.0	0.492	16.995	0.559
0.7	1.5	0.516	17.156	0.535
0.7	2.0	0.512	16.729	0.533
0.8	0.0	0.392	13.219	0.703
0.8	0.5	0.410	14.126	0.647
0.8	1.0	0.457	15.190	0.571
0.8	1.5	0.475	15.181	0.548
0.8	2.0	0.466	14.583	0.549
0.9	0.0	0.343	11.303	0.730
0.9	0.5	0.358	12.243	0.669
0.9	1.0	0.417	13.506	0.590
0.9	1.5	0.434	13.333	0.565
0.9	2.0	0.423	12.729	0.568
1	0.0	0.287	9.648	0.746
1	0.5	0.321	10.942	0.685
1	1.0	0.385	12.113	0.611
1	1.5	0.402	11.847	0.585
1	2.0	0.394	11.269	0.590

표 4. 성능 지표 정규화 결과(ImageNet)

Table 4. Normalized performance metric results(ImageNet)

Rank	Noise level	Control weight	Norm. mean
1	0.5	2.0	0.971
2	0.4	2.0	0.943
3	0.5	1.5	0.940
⋮	⋮	⋮	⋮
33	1	0.5	0.165
34	0.9	0.0	0.146
35	1	0.0	0.000

표 5. 성능 지표 정규화 결과(COCO)

Table 5. Normalized performance metric results(COCO)

Rank	Noise level	Control weight	Norm. mean
1	0.5	2.0	0.967
2	0.6	2.0	0.949
3	0.5	1.5	0.929
⋮	⋮	⋮	⋮
33	1	0.5	0.178
34	0.9	0.0	0.149
35	1	0.0	0.000

4.4 복원 및 업스케일링 기법 비교

본 연구는 제한된 입력 정보인 저해상도 이미지와 Canny 에지를 활용하는 생성 모델 기반의 재구성 방식이다. 반면 기존의 Gradient Inversion 또는 Model Inversion과 같은 복원 공격 기법은 기울기나 특징 맵 등 모델 내부 정보에 접근한다. 따라서 기존 복원 공격 기법과의 직접적인 비교는 의미 있는 평가로 이어지기 어렵다. 이에 본 연구에서는 제안 방식과 동일하게 저해상도 이미지를 활용할 수 있는 단순 업샘플링 방식을 비교군으로 설정한다. 비교 기법으로는 아티팩트 생성 과정에서 설명한 쌍입방 및 란초스 보간법을 채택한다. 각 기법 간의 복원 성능 결과는 표 6에 나타내었다. SSIM과 LPIPS에서는 제안 기법이, PSNR에서는 란초스 보간법이 가장 높은 성능을 나타냈다.

표 6. 복원 기법별 평가 결과

Table 6. Evaluation results by restoration method

Metrics	Dataset	Reconstructed image	Lanczos3	Bicubic
Mean SSIM(↑)	ImageNet	0.548	0.522	0.523
Mean PSNR(↑)	ImageNet	20.065	20.729	20.523
Mean LPIPS(↓)	ImageNet	0.523	0.66	0.657
Mean SSIM(↑)	COCO	0.548	0.522	0.523
Mean PSNR(↑)	COCO	20.065	20.729	20.523
Mean LPIPS(↓)	COCO	0.523	0.66	0.657

이는 PSNR이 원본 이미지와의 픽셀 단위 오차를 기반으로 평가하기 때문에, 생성 모델을 통해 시각적으로 유사한 이미지로 복원하는 제안 기법보다 입력 이미지의 픽셀 정보를 직접 보존하는 단순 업샘플링 방식이 상대적으로 유리한 결과를 보인 것으로 분석된다.

4.5 딥러닝 모델 기반 인식 가능성 평가

복원된 이미지가 실제 딥러닝 모델의 시각적 인지 능력에 미치는 영향을 검증하고자 하였다. 사전 학습된 딥러닝 모델을 활용하여 각 데이터셋의 목적에 따라 평가하였다. 복원된 ImageNet-1K는 이미지 분류 모델인 ResNet-50로 Top-1 및 Top-5 정확도로 평가하였고, COCO 2017 복원 이미지는 객체 탐지 모델인 YOLOv7으로 $mAP@0.5$ 와 $mAP@0.5:0.95$ 로 평가하였다[15][16]. 데이터셋에 따른 상위 3개의 결과와 원본 데이터셋의 인식 성능 비교를 표 7과 표 8에 나타내었다.

표 7. ImageNet-1K 복원 이미지 정확도 평가

Table 7. Evaluation of Image Restoration on ImageNet-1K

Noise level	Control weight	Top-1 accuracy	Top-5 accuracy
0.5	1.5	41.1	67.6
0.7	2.0	40.3	65.7
0.5	2.0	40.1	64.6
Ground truth		74.9	92.7

표 8. COCO 2017 복원 이미지 정확도 평가

Table 8. Evaluation of image restoration on COCO 2017

Noise level	Control weight	$mAP@0.5$	$mAP@0.5:0.95$
0.6	1.5	0.268	0.172
0.5	1.5	0.262	0.172
0.6	1.0	0.259	0.167
Ground truth		0.689	0.502

복원된 이미지를 활용한 미세 조정(Fine-tuning)을 통해 모델의 인식 정확도 개선 가능성을 확인하였다. 평가의 신뢰성을 높이기 위해 선행 실험과 중복되지 않는 ImageNet-1K 검증 데이터셋 2,000장을 선정하고 선행 실험과 동일한 과정으로 복원하였다.

1,000개의 클래스에 대해 클래스당 1장의 이미지를 할당하여 학습 및 검증 데이터로 각각 1,000장씩 분할하여 구성하였다. 모델을 미세 조정된 결과는 표 9에 나타내었다. 미세조정된 결과는 Top-1 및 Top-5 정확도가 각각 12.3%, 9.0% 향상되어 모델의 정확도 개선 가능성을 보였다.

표 9. 모델 정확도 비교 결과(ImageNet)

Table 9. Comparison of model accuracy(ImageNet)

Noise level	Control weight	Top-1 accuracy	Top-5 accuracy
0.5	1.5	41.1	67.6
Fine-tuning		53.4	76.6
Ground truth		74.9	92.7

레이블 추출 공격의 오탐지를 최소화하고, 해당 공격의 유효성 평가를 위해 미세 조정된 모델의 예측이 실제 정답과 일치하는 경우를 진양성, 불일치하는 경우를 가양성으로 정의하고 그림 3의 예측 신뢰도 기반 ROC 곡선으로 분석하였다. 분석 결과 최적 임계값은 0.65로 도출되었다. 예측 신뢰도가 해당 임계값 이상인 고신뢰도 표본을 대상으로 평가한 결과, 가양성 표본의 82.2%를 효과적으로 배제하여 81.18%의 정밀도(Precision)를 달성하였다. 반면 진양성 표본도 33.0%만큼 기각되어 67.0%의 재현율(Recall)을 나타내었다. 결과적으로 공격이 적용된 표본은 전체 데이터의 44.10%로 제한되나, 가양성을 효과적으로 줄이는 상충관계를 보였다.

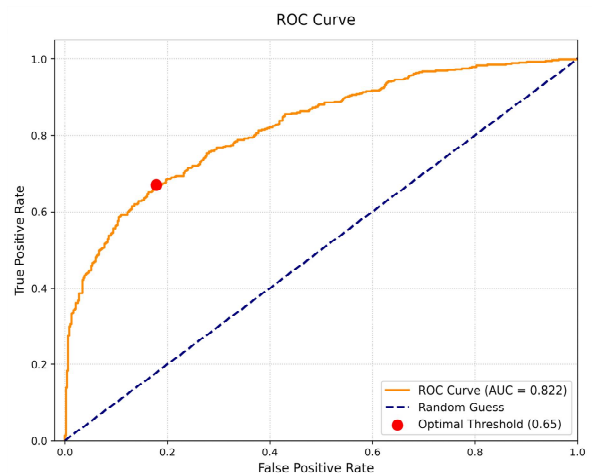


그림 3. 미세 조정 모델의 수신자 조작 특성 곡선

Fig. 3. ROC curve of the fine-tuned model

4.6 실험 고찰

본 실험의 결과를 바탕으로 Stable Diffusion 노이즈 주입 강도와 ControlNet 제어 가중치가 복원 품질에 미치는 영향을 분석하였다. 먼저 노이즈 주입

강도가 증가하면 확산 모델은 입력 이미지로부터 벗어나 새로운 세부 정보를 생성할 수 있는 범위가 넓어진다. 그러나 노이즈 비율이 커질수록 열화 이미지에 남아 있는 원본 이미지의 정보가 손실되는 상충관계를 가진다.

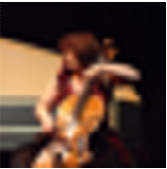
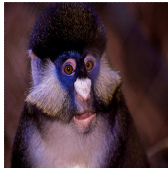
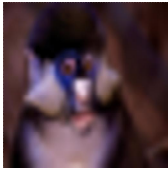
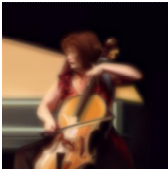
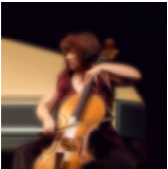
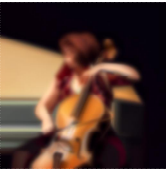
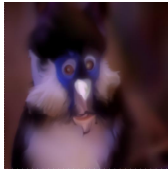
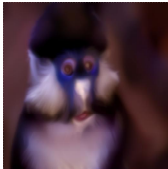
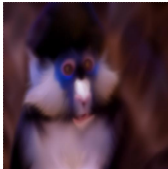
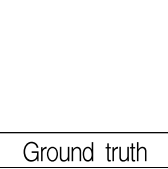
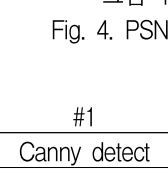
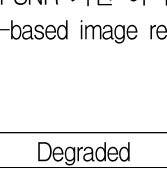
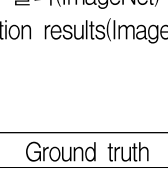
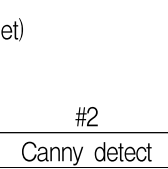
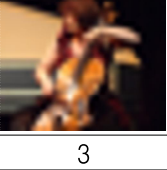
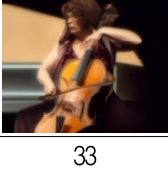
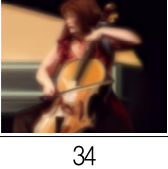
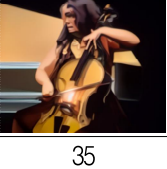
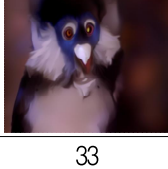
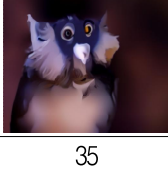
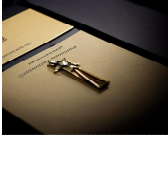
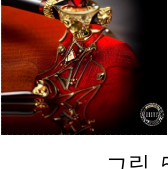
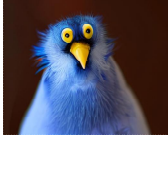
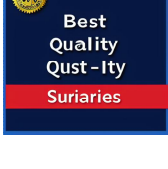
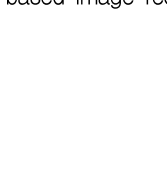
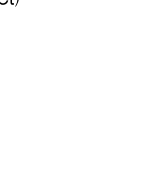
PSNR	#1			#2		
	Ground truth	Canny detect	Degraded	Ground truth	Canny detect	Degraded
						
Rank	1	2	3	1	2	3
						
Rank	33	34	35	33	34	35
						
						
						
						
						
						

그림 4. PSNR 기반 이미지 복원 결과(ImageNet)

Fig. 4. PSNR-based image reconstruction results(ImageNet)

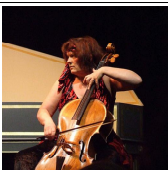
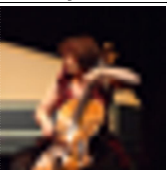
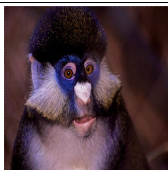
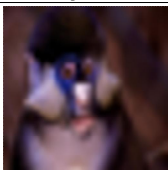
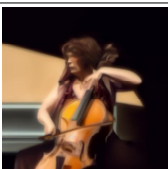
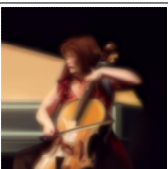
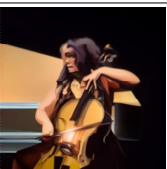
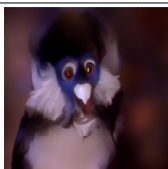
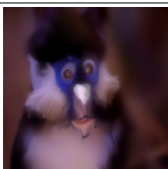
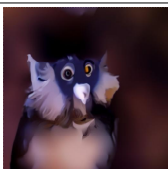
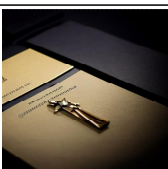
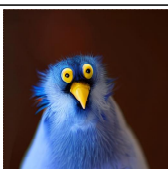
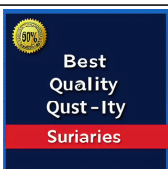
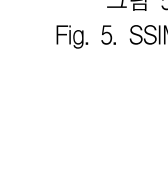
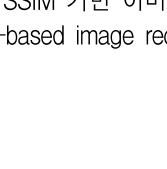
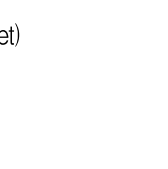
SSIM	#1			#2		
	Ground truth	Canny detect	Degraded	Ground truth	Canny detect	Degraded
						
Rank	1	2	3	1	2	3
						
Rank	33	34	35	33	34	35
						
						
						

그림 5. SSIM 기반 이미지 복원 결과(ImageNet)

Fig. 5. SSIM-based image reconstruction results(ImageNet)

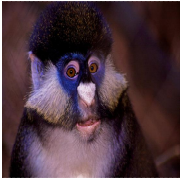
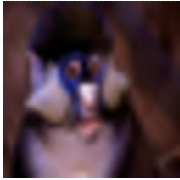
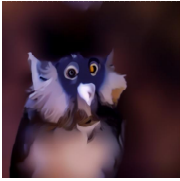
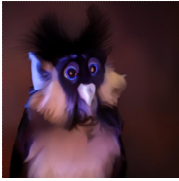
LPIPS	#1			#2		
	Ground truth	Canny detect	Degraded	Ground truth	Canny detect	Degraded
						
Rank	1	2	3	1	2	3
						
Rank	33	34	35	33	34	35
						

그림 6. LPIPS 기반 이미지 복원 결과(ImageNet)
Fig. 6. LPIPS-based image reconstruction results(ImageNet)

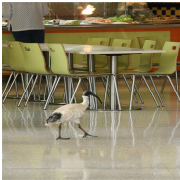
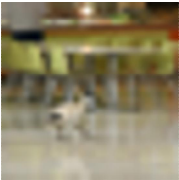
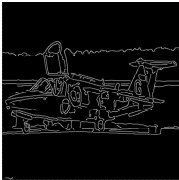
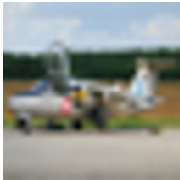
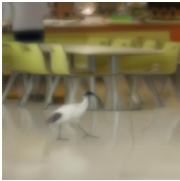
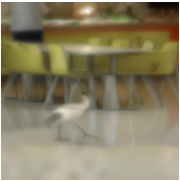
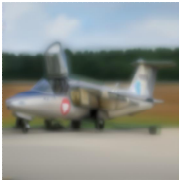
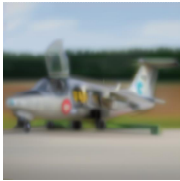
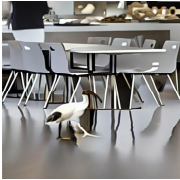
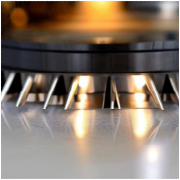
PSNR	#1			#2		
	Ground truth	Canny detect	Degraded	Ground truth	Canny detect	Degraded
						
Rank	1	2	3	1	2	3
						
Rank	33	34	35	33	34	35
						

그림 7. PSNR 기반 이미지 복원 결과(COCO)
Fig. 7. PSNR-based image reconstruction results(COCO)

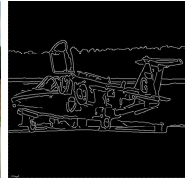
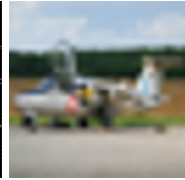
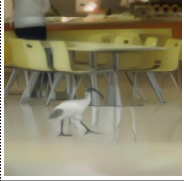
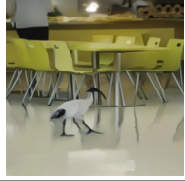
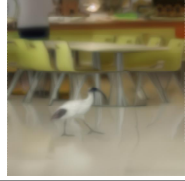
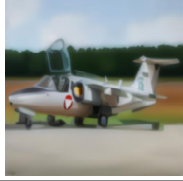
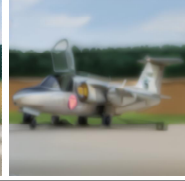
SSIM	#1			#2		
	Ground truth	Canny detect	Degraded	Ground truth	Canny detect	Degraded
						
Rank	1	2	3	1	2	3
						
Rank	33	34	35	33	34	35
						

그림 8. SSIM 기반 이미지 복원 결과(COCO)
Fig. 8. SSIM-based image reconstruction results(COCO)

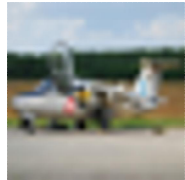
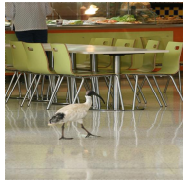
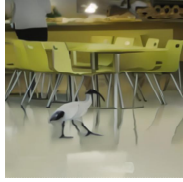
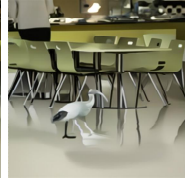
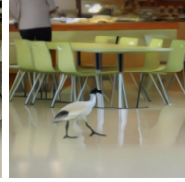
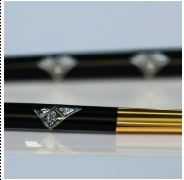
LPIPS	#1			#2		
	Ground truth	Canny detect	Degraded	Ground truth	Canny detect	Degraded
						
Rank	1	2	3	1	2	3
						
Rank	33	34	35	33	34	35
						

그림 9. LPIPS 기반 이미지 복원 결과(COCO)
Fig. 9. LPIPS-based image reconstruction results(COCO)

이는 Canny 에지 정보가 원본 이미지의 구조적 특징을 유지하는 데 중요한 역할을 한다는 것을 의미한다. 평가 지표별 최적 복원 결과에서 서로 다른 특성이 나타났다. 그림 4와 그림 7과 같이 PSNR은 픽셀 단위 오차를 반영하여 수치가 높더라도 복원 이미지가 다소 흐릿한 결과를 보였다. 그림 5와 그림 8의 결과는 SSIM은 구조적 유사성을 평가하는데 유용하지만 세부 질감 차이를 충분히 반영하지 못하는 한계를 보였다. 그림 6과 그림 9의 결과에서 LPIPS는 딥러닝 특징 맵을 활용하므로 시각적으로 선명한 결과를 보였으나, 일부 결과에서 색상 왜곡이 관찰되었다. 따라서 본 연구에서는 단일 지표가 아니라 PSNR, SSIM, LPIPS를 함께 고려하여 복원 품질을 평가하였다.

V. 결 론

본 연구는 제한적인 정보만을 가진 아티팩트를 이용하여 원본 데이터와 유사한 이미지로 복원하고 SSIM, PSNR, LPIPS의 세 지표에 따른 최적 품질을 달성하는 모델의 파라미터를 탐색하였다. 대규모 데이터로 사전 학습된 Stable Diffusion과 ControlNet을 활용하여 복원한 결과 노이즈 주입 강도는 0.4에서 0.6의 범위에서, 제어 가중치는 2.0에서 최적의 결과를 보였다. 복원된 이미지를 이미지 분류 모델로 평가한 결과, Top-1 정확도와 Top-5 정확도는 각각 41.1%와 67.6%를 보였다. 소규모 데이터로 미세 조정 시 각각 53.4%, 76.6%의 향상된 정확도를 보였다. 추가로 ROC 곡선으로부터 도출한 임계값 0.65를 적용하여 레이블 추출 공격을 수행하였다. 그 결과로 임계값을 통과한 고신뢰도 표본에서 81.18%의 정밀도를 달성하였다. 다만 공격이 적용된 표본은 전체 데이터의 44.10%로 제한되었다. 따라서 본 공격은 전체 데이터에 대해 동일한 성능을 보이는 공격이 아니라, 일부 고신뢰도 표본에 대해 높은 정밀도를 보이는 제한적 공격으로 해석된다. 이러한 결과는 저해상도 이미지와 Canny 에지와 같은 전처리 아티팩트도 원본 데이터의 시각적·의미적 정보를 내포할 수 있음을 보여준다. 따라서 전처리 아티팩트가 빈번하게 발생하는 온디바이스 AI 환경에서는

보안 정책의 관점을 전환해야 한다. 기존의 원본 데이터 보호 중심에서 벗어나, 데이터 전처리 과정에서 생성되는 아티팩트 역시 핵심 보호 대상으로 포함해야 한다. 기술적 측면에서는 아티팩트의 저장 최소화, 암호화 등이 요구되며, 특히 실시간 추론 과정에서 생성되는 중간 아티팩트는 격리된 하드웨어 보안 영역 내에서 안전하게 처리하는 방안을 고려할 필요가 있다. 아울러 운영 정책 측면에서도 아티팩트의 데이터 수명 주기를 규정해야 한다. 나아가 아티팩트 기반의 재구성 공격 가능성을 보안 취약점 평가 항목으로 규정하고 이에 대한 위험도 분석을 수행해야 한다.

향후 이러한 아티팩트에 최적화되어 시각적으로 선명한 복원 결과를 보이는 새로운 어댑터 모델을 제안하고자 한다. 인간의 시각적 인식과 딥러닝 모델의 특징 추출을 모두 만족하도록 미세 조정을 수행하여 복원된 이미지의 품질을 개선하고 레이블 추출 공격의 정확도를 향상시킬 계획이다.

References

- [1] Korean National Police Agency, https://www.police.go.kr/user/bbs/BD_selectBbs.do?q_bbsCode=1002&q_bbscttSn=20221221091602934. [accessed: Apr. 23, 2026]
- [2] J. S. Park, S. H. Seo, and C. H. Lee, "Analysis of mobile browser privacy mode from memory forensic", *Journal of Digital Contents Society*, Vol. 21, No. 4, pp. 781-791, Apr. 2020. <http://dx.doi.org/10.9728/dcs.2020.21.4.781>.
- [3] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients", *Proc. 33rd Conf. Neural Inf. Process. Syst. (Neurips)*, Vancouver, Canada, Vol. 32, pp. 14747-14756, Dec. 2019.
- [4] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, USA, pp. 4829-837, Jun. 2016. <https://doi.org/10.1109/CVPR.2016.522>.
- [5] R. Rombach, et al., "High-resolution image

- synthesis with latent diffusion models", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, USA, pp. 10674-10685, Jun. 2022. <https://doi.org/10.1109/CVPR52688.2022.01042>.
- [6] M. Schlegel and K. U. Sattler, "Management of Machine Learning Lifecycle Artifacts: A Survey", SIGMOD Rec., Vol. 51, No. 4, pp. 18-35, Jan. 2023. <https://doi.org/10.1145/3582302.3582306>.
- [7] J. Hou, et al., "A Survey on Digital Forensics in Internet of Things", IEEE Internet of Things Journal, Vol. 7, No. 1, pp. 1-15, Jan. 2020. <https://doi.org/10.1109/JIOT.2019.2940713>.
- [8] D. H. Yun and S. J. Lee, "Study for Android Smartphone's Gallery Thumbnail Forensic Analysis", KIPS Transactions on Computer and Communication Systems, Vol. 6, No. 1, pp. 31-42, Jan. 2017. <http://doi.org/10.3745/KTCCS.2017.6.1.31>.
- [9] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models", Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Paris, France, pp. 3813-3824, Oct. 2023. <https://doi.org/10.1109/ICCV51070.2023.00355>.
- [10] T. M. Lehmann, C. Gonner, and K. Spitzer, "Survey: interpolation methods in medical image processing", IEEE Transactions on Medical Imaging, Vol. 18, No. 11, pp. 1049-1075, Nov. 1999. <https://doi.org/10.1109/42.816070>.
- [11] J. Deng, et al., "ImageNet: A large-scale hierarchical image database", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), FL, USA, pp. 248-255, Jun. 2009. <https://doi.org/10.1109/CVPR.2009.5206848>.
- [12] T. Y. Lin, et al., "Microsoft COCO: Common objects in context", Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, pp. 740-755, Sep. 2014. https://doi.org/10.1007/978-3-319-10602-1_48.
- [13] Z. Wang, et al., "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, Vol. 13, No. 4, pp. 600-612, Apr. 2004. <https://doi.org/10.1109/TIP.2003.819861>.
- [14] R. Zhang, et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, USA, pp. 586-595, Jun. 2018. <https://doi.org/10.1109/CVPR.2018.00068>.
- [15] K. He, et al., "Deep Residual Learning for Image Recognition", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, USA, pp. 770-778, Jun. 2016. <https://doi.org/10.1109/CVPR.2016.90>.
- [16] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, Canada, pp. 7464-7475, Jun. 2023. <https://doi.org/10.1109/CVPR52729.2023.00721>.

저자소개

방 한 솔 (Hansol Bang)



2024년 2월 : 한국공학대학교
컴퓨터공학과(공학석사)
2024년 3월 ~ 현재 :
한국공학대학교
IT반도체융합공학과 박사과정
관심분야 : Biologically Plausible
Learning, Federated Learning

최 진 구 (Jinku Choi)



2003년 3월 : 와세다대학교
전자정보통신공학과(공학박사)
2003년 3월 ~ 현재 :
한국공학대학교
컴퓨터공학부 교수
관심분야 : Embedded system,
design IoT platform