Check for updates

Journal of KIIT. Vol. 24, No. 10, pp. 15-23, Oct. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.10.15

CvT-13 기반 분류 및 분할 모델을 활용한 생성·조작 이미지 판별 및 시각화 시스템

임재한*, 김도언**[1], 장원준**[2], 이원일**[3]

Detection and Visualization System for Synthetic and Manipulated Images based on Classification and Segmentation Models using CvT-13

JaeHan Lim*, Do-Eon Kim**[1], WonJun Jang**[2], and Won Il Lee**[3]

본 과제(결과물)는 (2026)년도 교육부 및 경상북도의 재원으로 경북앵커센터의 지원을 받아 수행된 지역성장 인재양성체계(앵커)-(아이디어 창업밸리)의 결과입니다.(2026-anchor-15-105)

요 약

최근 생성형 인공지능 기술의 발전으로 실제 이미지와 구분하기 어려운 합성 및 조작 이미지가 증가하면서, 시각 정보의 신뢰성 문제가 대두되고 있다. 기존의 이미지 판별 기법은 이미지 단위 분류에 주로 의존하여, 조작 영역의 위치를 설명하는 데 한계를 가진다. 본 논문에서는 CvT-13을 백본으로 사용하는 분류 및 분할 기반 이미지 판별 시스템을 제안한다. 분류 모델은 실제 이미지와 생성·조작 이미지를 이미지 단위에서 구분하고, 분할 모델은 픽셀 단위 예측을 통해 조작 의심 영역의 공간적 위치를 추정한다. 또한 분할 결과에 시각화 기법을 적용하고, Pointing Game 및 mIoU 평가를 통해 공간적 판별 능력을 정량적으로 분석하였다. 실험 결과, 제안한 모델은 안정적인 분류 성능과 함께 조작 영역의 핵심 위치를 일정 수준 이상 식별할 수 있음을 확인하였다.

Abstract

The rapid advancement of generative artificial intelligence has increased synthetic and manipulated images that are difficult to distinguish from real images, raising concerns about visual reliability. Conventional detection methods mainly rely on image-level classification, which limits spatial interpretation of manipulated regions. In this paper, we propose a classification and segmentation-based image detection system using CvT-13 as the backbone. The classification model distinguishes real images from synthetic and manipulated images, while the segmentation model predicts suspicious regions at the pixel level. Grad-CAM visualization is applied to analyze the spatial decision cues of the model, and Pointing Game and mIoU are used to evaluate its spatial localization capability. Experimental results show that the proposed model achieves stable classification performance and provides spatial cues for identifying key manipulated regions.

Keywords

image detection, convolutional vision transformer, CvT-13, image segmentation, visualization

* 국립금오공과대학교 반도체시스템공학과
- ORCID: https://orcid.org/0009-0000-1031-7069
** 국립금오공과대학교 전자공학부(**[3] 교신저자)
- ORCID[1]: https://orcid.org/0009-0005-9160-3943
- ORCID[2]: https://orcid.org/0009-0007-4918-3345
- ORCID[3]: https://orcid.org/0000-0002-0851-3113

· Received: Apr. 07, 2026, Revised: Aug. 31, 2026, Accepted: Sep. 03, 2026
· Corresponding Author: Won Il Lee
· School of Electronic Engineering, Kumoh National Institute of Technology
Tel.: +82-54-478-7429, Email: wilee@kumoh.ac.kr

Ⅰ. 서 론

최근 AI 기술의 발전으로 사용자는 자신의 아이디어를 즉각적으로 시각적인 형태로 구현할 수 있게 되었으며, 이에 따라 생성 이미지 기술에 관한 관심이 증가하고 있다. 특히 육안으로 구분하기 어려운 정교한 생성 이미지 기술이 등장하면서 관련 연구가 활발히 진행되고 있다. 그러나 생성 이미지 기술의 고도화는 조작 이미지 및 딥페이크 기술의 발전으로도 이어져, SNS와 언론을 통해 가짜 시각 콘텐츠가 확산되며 정보 신뢰도 저하와 사회적 혼란을 초래하고 있다[1][2].

이러한 문제를 해결하기 위해 생성 및 조작 이미지 판별을 위한 다양한 연구들이 수행되어 왔다[3]-[5]. [3]에서는 생성형 시각 미디어에 대한 탐지, 방해, 인증 등 방어 기법과 관련 연구 흐름을 정리하였으며, [4]에서는 트랜스포머 기반 생성 모델과 GAN 구조를 중심으로 생성 모델의 특성과 발전 과정을 분석하였다. [5]에서는 소셜미디어 환경에서 확산되는 딥페이크 영상 및 음성을 대상으로 탐지 기법을 분석하고, 플랫폼 관점에서의 대응 방안을 제시하였다. 하지만 기존의 생성 이미지 판별 연구는 주로 실제 이미지와 생성 이미지를 구분하는 이진 분류 문제로 설정되어 왔으며, 전통적인 이미지 조작·위조 탐지에서 중요한 조작 영역 국소화는 상대적으로 간과되어 왔다.

한편, 이미지 분류 모델은 크게 합성곱 신경망(CNN, Convolutional Neural Network)과 비전 트랜스포머(ViT, Vision Transformer)로 구분된다. 조작 이미지의 특징은 대부분 경계 주변의 깨짐, 피부 텍스쳐 불연속, 조명·노이즈 패턴 같은 국소적인 저수준 아티팩트로 나타나며, 이러한 특징은 로컬 패턴을 학습하는 CNN을 활용하여 검출할 수 있다[6]. 반면, 생성형 모델(GAN, Diffusion)로 만들어진 이미지의 특징은 전체적인 텍스처 일관성 이상, 얼굴 구조의 전역 패턴 왜곡과 같이 이미지 전체에서 발생하는 비정상 패턴으로 나타나며, 이러한 특징은 전역적 패턴을 학습하는 ViT를 활용하여 검출할 수 있다[4]. 따라서 CNN은 주로 국소 수용 영역에 기반해 로컬 아티팩트에 강한 반면, ViT는 global context와 long-range dependencies를 포착해 생성 이미지의 전역적 비정상 패턴을 더 잘 식별한다[7].

본 논문에서는 생성 이미지와 조작 이미지를 동시에 판별하고, 조작 이미지에 대해서는 모델의 판단 근거가 되는 공간적 영역을 시각적으로 분석할 수 있는 시스템을 제안한다. 시스템은 크게 생성·조작 이미지 판별을 수행하는 분류 모듈과 분할 모듈 그리고 분할 과정에서 형성된 특징 표현을 활용하여 조작 의심 영역을 추정하는 CAM 기반 시각화 모듈로 구성된다. 이를 위해 CNN의 로컬 아티팩트에 강한 장점과 ViT의 전역적 비정상 패턴 식별의 장점을 결합한 하이브리드 모델인 CvT-13을 백본으로 채택하였다. 기존 CNN 중심의 Grad-CAM 기법을 CvT 모델에 적용하기 위해 토큰 기반 특징 표현과 스테이지별 출력 구조를 고려한 적용 방식을 구성하였다. 나아가 시각화 결과의 신뢰성을 검증하기 위해 Pointing Game과 mIoU(Mean Intersection over Union)를 도입하여, 제안한 시스템의 공간적 판단 성능을 정량적으로 평가하였다.

제안하는 시스템의 차별점은 새로운 백본 구조 자체의 제안보다는 생성·조작 이미지 판별 문제에서 이미지 단위 분류와 픽셀 단위 공간 추정을 결합하고, CvT-13의 계층적 토큰 특징을 활용하여 분류, 분할, 시각화를 통합한 2단계 구조로 구성한 점에 있다. 또한 CvT-13의 token embedding 출력을 2차원 공간 특징 맵으로 재구성하여 Grad-CAM을 적용하고, Pointing Game과 mIoU를 통해 공간적 판단 근거를 정량적으로 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 이론을 설명하고, 3장에서는 제안한 시스템의 구조와 구현 방법을 기술한다. 4장에서는 이미지 분류 및 시각화 기반 성능 평가 결과를 제시하며, 5장에서는 결론과 향후 연구 방향을 논의한다.

Ⅱ. 배경 이론

2.1 CvT-13

ViT는 이미지를 패치 단위로 분할한 뒤 토큰 시퀀스로 변환하여 Self-Attention을 적용함으로써 전역

적인 문맥 정보를 효과적으로 학습한다. 그러나 ViT는 CNN이 갖는 지역적 공간 구조 반영 능력과 이동 불변성이 부족하며, 충분한 성능을 위해 대규모 데이터셋 사전학습과 positional encoding에 의존하는 한계를 가진다. 이에 따라 중·소규모 데이터 환경이나 국소적 시각 단서가 중요한 문제에서는 성능 저하가 발생할 수 있다.

CvT(Convolutional Vision Transformer)는 이러한 한계를 보완하기 위해 트랜스포머 구조에 합성곱 연산을 도입한 하이브리드 모델로, CNN과 ViT의 장점을 동시에 활용하도록 설계되었다[8]. CvT는 Convolutional Token Embedding을 통해 토큰 생성 단계에서 지역적 공간 정보를 반영하고, Convolutional Projection을 통해 Self-Attention 과정에서 국소 문맥을 강화한다. 이를 기반으로 계층적 트랜스포머 구조를 구성함으로써, CNN과 유사한 다단계 특징 추출이 가능하다.

CvT-13은 이러한 CvT 구조를 기반으로 설계된 대표적인 모델로, 3개의 계층적 스테이지와 총 13개의 트랜스포머 블록으로 구성된다. 비교적 적은 연산량으로도 효과적인 전역 및 지역 특징 학습이 가능하다.

2.2 Grad-CAM

이미지 위·변조 판별과 같은 응용에서는 모델이 어떤 시각적 단서를 근거로 판단을 내렸는지 설명 가능성이 요구된다. 이를 해결하기 위한 방법으로 CAM(Class Activation Map) 계열의 시각화 기법이 제안되었으며, Grad-CAM은 이러한 접근법을 일반화한 기법이다[9].

Grad-CAM은 출력 클래스 점수의 그래디언트를 이용하여 모델이 주목한 공간 영역을 추정하는 방법이다. 특정 클래스에 대해, 마지막 합성곱 계층의 특징 맵이 해당 클래스에 얼마나 기여했는지를 그래디언트 기반으로 계산한다. Grad-CAM의 핵심 아이디어는 최종 예측 클래스 점수에 대한 그래디언트를 계산하고 그래디언트의 전역 평균을 통해 채널 중요도를 산출한 뒤, 중요도를 특징 맵에 가중합하여 활성화 맵을 생성하는 것이다. 이를 통해 모델 구조에 대한 제약 없이 다양한 CNN 및 하이브리드 모델에 적용할 수 있다.

III. CvT-13을 활용한 이미지 판별 및 시각화 시스템

3.1 시스템 구조(System architecture)

본 논문에서는 생성 이미지 및 조작 이미지에 대한 효과적인 판별과 해석을 위해, 2단계 구조의 시스템을 제안한다.

그림 1은 제안하는 시스템의 전체 구조로, 입력 이미지의 진위를 판별하는 분류 단계와 그 결과가 fake image로 판단된 경우에 수행되는 분할, 시각화 단계로 나뉜다. 이를 통해 불필요한 연산을 줄이고 시스템 효율성을 향상시킨다.

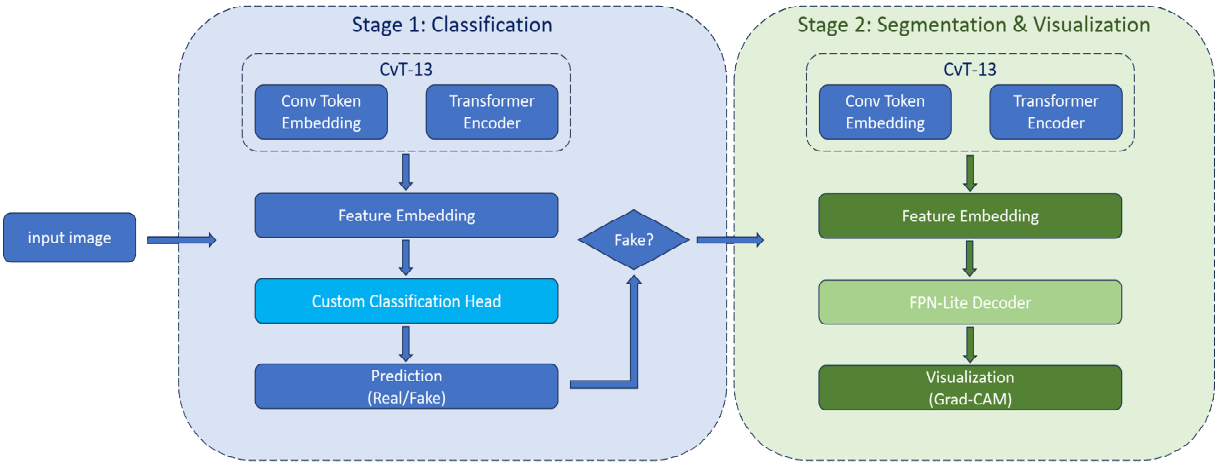


그림 1. 제안하는 2단계 시스템 구조
Fig. 1. Overall architecture of the proposed two-stage system

생성 이미지는 전역적 비정상 패턴이 중요하며, 조작 이미지는 경계 및 국소 아티팩트와 같은 지역적 단서가 중요하다. 따라서 본 시스템에서는 전역 attention과 지역적 공간 특징을 동시에 활용할 수 있는 CvT-13을 분류 및 분할 단계의 공통 백본으로 사용하였다.

3.2 분류 모델(Classification model)

분류 모델의 백본으로는 CvT-13을 사용하였다. 입력 이미지는 초기 Convolutional Token Embedding을 통해 토큰 시퀀스로 변환되며, 이후 세 개의 스테이지를 거치며 점진적으로 고수준 특징으로 추상화된다. 본 연구에서는 스테이지 3의 출력 특징을 최종 분류 기준으로 사용했다.

분류 모델은 실제 이미지와 생성·조작 이미지를 이미지 단위에서 구분하기 위해 ArtiFact 데이터셋과 CASIA2.0 데이터셋을 함께 사용하여 학습하였다[10][11]. ArtiFact 데이터셋은 실제 이미지와 AI 생성 이미지를 포함하고 있어 생성 이미지 판별 학습에 활용하였으며, CASIA2.0 데이터셋은 실제 이미지와 조작 이미지를 포함하고 있어 조작 이미지 판별 학습에 활용하였다. 이를 통해 분류 모델은 단순한 생성 이미지뿐만 아니라 위조 및 조작 이미지까지 fake image 범주로 학습하도록 구성하였다. 데이터셋은 이미지 단위 random split 방식으로 구성하였으며, 전체 데이터 중 80%는 학습용, 20%는 테스트용으로 사용하였다. 분류 모델의 학습 데이터는 총 약 778,000장의 real image와 약 1,230,000장의 fake image로 구성된다.

CvT-13의 백본 출력은 고차원 특징 벡터 형태로 제공되며, 이를 본 연구의 이진 분류 문제에 적합하도록 custom classification head를 설계하였다.

그림 2는 본 연구에서 제안한 custom classification head의 구조를 나타낸다. 해당 분류 헤드는 CvT-13 백본으로부터 추출된 고차원 특징 벡터를 입력으로 받아, 이를 실제 이미지(Real image)와 생성·조작 이미지(Fake image)로 분류하는 역할을 수행한다. 분류 헤드는 두 개의 완전 연결층(Fully connected layer)으로 구성되며, 각 층에는 Mish 활성화 함수와 배치 정규화(Batch normalization)를 적용하여 학습 안정성을 향상시켰다. 또한 과적합을 방지하고 일반화 성능을 높이기 위해, 첫 번째와 두 번째 완전 연결층 뒤에 각각 확률 0.5와 0.3의 dropout을 적용하였다. 최종 출력층은 두 개의 클래스를 갖는 선형 계층으로 구성되며, 이를 통해 입력 이미지에 대한 이진 분류(real/fake)를 수행한다.

3.3 분할 모델(Segmentation model)

분할 모델은 입력 이미지에서 조작 의심 영역의 공간적 위치를 픽셀 단위로 추정하기 위해 설계되었다. 이를 위해 본 연구에서는 CvT-13을 백본으로 사용하는 세그멘테이션 구조를 구성하고, 백본에서 추출된 계층적 특징을 활용하여 조작 영역 마스크를 예측한다.

분할 모델 학습에는 정답 마스크가 제공되는 CASIA2.0 Groundtruth 데이터셋을 사용하였다[11]. CASIA2.0 데이터셋은 조작 이미지와 이에 대응하는 픽셀 단위 정답 마스크를 포함하고 있어, 조작 영역 국소화 및 분할 성능 평가에 적합하다. 본 연구에서는 해당 데이터셋을 이용하여 입력 이미지 내 조작 의심 영역을 이진 마스크 형태로 예측하도록 분할 모델을 학습하였다. 데이터셋은 이미지 단위 random split 방식으로 구성하였으며, 전체 데이터 중 80%는 학습용, 20%는 테스트용으로 사용하였다. 분할 모델의 학습 데이터는 총 약 6,000장의 real image, 약 4,100장의 fake image, 그리고 각 fake image에 대응하는 정답 마스크로 구성된다.

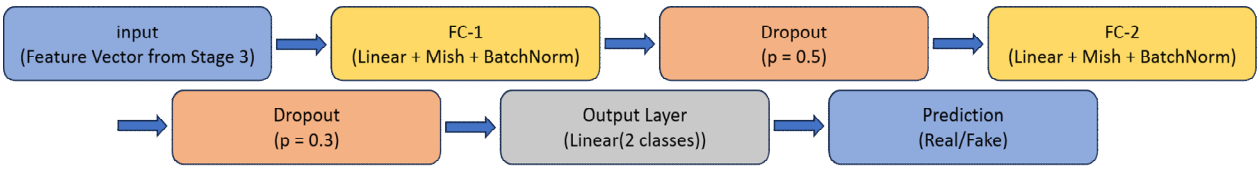


그림 2. 커스텀 분류 헤드의 구조
Fig. 2. Architecture of the custom classification head

분할 모델은 여러 스테이지의 출력 특징을 동시에 활용하여 공간적 정보를 유지하며, CvT-13의 각 스테이지의 출력 특징을 결합한 다중 해상도 표현을 사용한다. 이러한 다중 해상도 특징은 FPN (Feature Pyramid Network)의 개념을 차용한 FPN-lite 디코더를 통해 통합된다[12]. 디코더는 상위 스테이지의 의미적 특징과 하위 스테이지의 공간 정보를 결합하여, 최종적으로 입력 이미지와 동일한 해상도의 이진 분할 마스크를 예측한다.

3.4 시각화(Visualization)

분할 모델의 마지막 단계 특징 맵을 기반으로, 모델이 이미지 내에서 어떤 영역을 조작 의심 영역으로 판단하는지를 직관적으로 분석하기 위해 시각화 기법을 적용하였다. 본 연구에서는 모델의 공간적 판단 근거를 분석하기 위한 시각화 기법으로 Grad-CAM을 사용하였다. 다만 기존 CNN 기반 Grad-CAM은 마지막 합성곱 특징 맵을 기준으로 공간적 중요도를 계산하는 방식이므로, Convolutional Token Embedding과 transformer encoder로 구성된 CvT-13 구조에 그대로 적용하기에는 한계가 있다.

이를 해결하기 위해 본 연구에서는 CvT-13의 Convolutional Token Embedding과 multi-stage 트랜스포머 구조를 활용하여 단계별 2차원 토큰 특징 맵을 구성하였다. 따라서 본 연구에서는 CvT-13의 마지막 단계 출력 특징 맵을 Grad-CAM의 대상 활성화 맵으로 설정하였다. 이후 fake class score에 대한 그래디언트를 마지막 단계 특징 맵 기준으로 계산하고, channel-wise global average pooling을 통해 각 특징 채널의 중요도 가중치를 산출하였다. 최종적으로 가중치와 단계 활성화 맵을 가중합한 뒤 ReLU를 적용하고, 입력 이미지 크기로 upsampling하여 조작 의심 영역 heatmap을 생성하였다.

CvT-13의 마지막 단계 출력 특징 맵은 Convolutional Token Embedding과 계층적 트랜스포머 구조를 통해 [B, C, H, W] 형태의 2차원 공간 특징 맵 구조를 유지한다. 따라서 일반 ViT 기반 구조와 같이 1차원 패치 토큰 시퀀스를 별도로 2차원 공간 구조로 reshape하는 과정 없이, 마지막 단계 특징 맵을 Grad-CAM의 활성화 맵으로 직접 사용할 수 있다. 기존 CNN 기반 Grad-CAM은 마지막 합성곱 계층의 특징 맵을 직접 사용하는 반면, 본 연구에서는 CvT-13의 토큰 출력을 단계 특징 맵 형태로 활용하였다. 이를 통해 CNN 기반 공간 특징 맵 개념을 CvT-13의 계층적 토큰 특징 구조에 맞게 확장하였다.

분할 모델의 출력은 입력 이미지와 동일한 해상도의 단일 채널 확률 맵으로, 각 픽셀은 조작 영역에 속할 가능성을 나타낸다. 시각화 과정에서는 해당 확률 맵을 입력 이미지와 중첩하여 표현함으로써, 모델이 주목한 영역의 위치와 범위를 직관적으로 확인할 수 있도록 하였다. 이를 통해 예측 마스크의 공간적 분포와 실제 조작 영역과의 정성적 일치 여부를 분석하였다.

그림 3(a)는 실제 조작 영역의 정답 마스크, 그림 3(b)는 분할 모델이 예측한 마스크, 그림 3(c)는 확률 맵에서 최대 응답 값을 갖는 위치를 표시한 포인트 이미지, 그림 3(d)은 CvT-13 구조에 맞추어 생성된 Grad-CAM 시각화 결과를 보여준다.

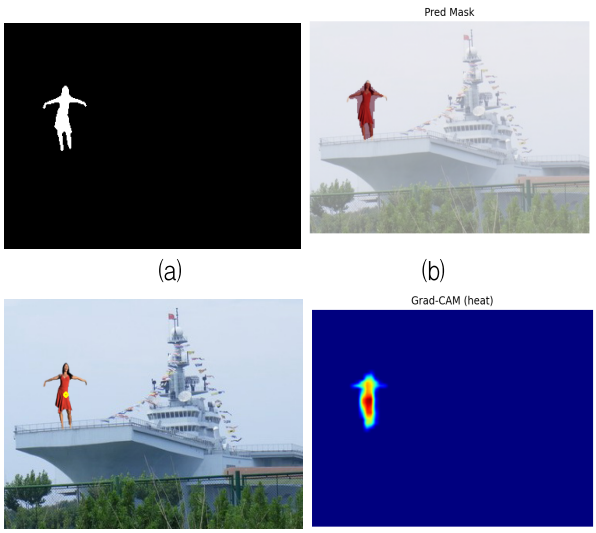


그림 3. 시각화 결과 예시 : (a) 정답 마스크, (b) 예측 마스크, (c) 포인트 이미지, (d) 시각화 이미지
Fig. 3. Examples of pointing game visualization results : (a) Ground-truth, (b) Predicted mask, (c) Point image, (d) Grad-CAM image

Ⅳ. 평 가

본 연구에서는 제안한 CvT-13 기반 분류 모델의

성능을 객관적으로 검증하기 위해 다중 모델 비교 실험을 수행하였다. 모든 분류 모델은 동일한 데이터셋 환경에서 학습 및 평가를 수행하였다. 분류 모델 학습에는 AdamW optimizer를 사용하였으며, 학습률은 1e-4로 설정하였다. 입력 이미지 크기는 ViT 및 CvT-13 모델의 기본 구조를 고려하여 224×224로 설정하였다. 배치 크기는 128, epoch 수는 50으로 설정하였으며, 손실 함수로는 CrossEntropyLoss를 사용하였다. 분류 결과는 softmax 기반 argmax 방식으로 계산하였다.

본 연구의 목적은 최신 SOTA 모델과의 성능 경쟁보다는, CvT-13 기반 판별 구조가 기존 대표적 CNN 계열 및 위조 탐지 기반 모델 대비 상대적 유효성을 가지는지를 검토하는 데 있다. 이에 따라 EfficientNet은 일반 이미지 분류에서 널리 사용되는 CNN 기반 비교군으로, F3Net은 조작 영역 및 주파수 단서를 활용하는 위조 탐지 계열 비교군으로 설정하였다[13][14]. 또한 순수 트랜스포머 구조와의 비교를 위해 순수 ViT 모델을 실험하여, convolutional inductive bias가 포함된 CvT 구조의 효과를 함께 분석하였다[15].

표 1은 EfficientNet, F3Net, ViT 그리고 본 연구에서 제안한 CvT-13 기반 모델의 분류 성능을 비교한 결과를 보여준다. 모델은 실제 이미지와 생성/조작 이미지를 구분하는 이진 분류 문제를 대상으로 평가되었으며, Accuracy, Precision, Recall, F1-score를 주요 성능 지표로 사용하였다.

표 1. 각 모델의 분류 성능평가 지표
Table 1. Classification performance metrics of each model

Method	F3Net	EfficientNet	ViT	Ours
Accuracy(%)	94.72	96.97	97.91	97.97
Precision(%)	94.92	96.94	97.76	97.95
Recall(%)	93.95	96.67	97.84	97.78
F1-score(%)	94.39	96.80	97.80	97.86

실험 결과, CvT-13 기반 제안 모델은 97.97%의 정확도를 기록하였으며, 정밀도와 재현율 또한 각각 97.95%, 97.78%로 높은 수준을 보였다. 이는 모델이 실제 이미지와 생성·조작 이미지를 안정적으로 구분할 수 있음을 의미한다. 특히 F1-score는 97.86%로, 클래스 간 균형 잡힌 분류 성능을 달성했음을 확인할 수 있다.

본 실험에서 CvT-13 기반 모델은 비교 모델 중 가장 높은 분류 성능을 보였으나, ViT 대비 성능 향상 폭은 제한적이었다. 따라서 본 결과는 CvT-13 구조의 일반적인 우수성을 입증하기보다는, 본 연구에서 사용한 데이터셋과 실험 조건에서 CNN과 Transformer를 결합한 하이브리드 구조가 생성·조작 이미지 판별에 적용 가능함을 확인한 결과로 해석하는 것이 적절하다. 또한 순수 트랜스포머 기반 ViT와 비교하였을 때, CvT-13은 Convolutional Token Embedding과 계층적 단계 구조를 통해 지역적 공간 특징과 전역 attention 정보를 동시에 활용할 수 있는 구조적 특징을 가진다. 이러한 구조는 단순 이미지 분류뿐 아니라, 이후 FPN-Lite 기반 분할 구조 및 Grad-CAM 기반 위치 추정 과정과도 자연스럽게 연결될 수 있다는 장점을 가진다.

분류 성능 비교 결과를 바탕으로, 제안 모델의 분할 결과에 대한 공간적 판별 성능을 평가하기 위해 Pointing Game과 mIoU 지표를 계산하여 표 2에 나타내었다. 분할 성능 평가는 분류 모델에서 fake로 분류된 이미지를 대상으로 수행하였다.

표 2. 공간적 판별 성능평가 결과
Table 2. Spatial judgment performance evaluation results

Metric	Value
Pointing game(%)	64.86
mIoU(%)	48.91

본 연구의 분할 모듈은 조작 영역의 정밀한 경계를 완전히 복원하는 것보다, 모델의 최대 응답 위치가 실제 조작 영역 내부에 포함되는지를 확인하는 데 중점을 둔다. 이에 따라 최대 응답 위치의 적합성을 평가하는 Pointing Game을 위치 단서 평가를 위한 주요 보조 지표로 사용하였으며, 예측 영역과 정답 영역 사이의 픽셀 단위 공간적 정합성을 확인하기 위해 mIoU를 함께 제시하였다. 두 지표는 각각 위치 수준과 영역 수준의 성능을 평가한다는 점에서 상호보완적으로 사용하였다.

Pointing Game은 분할 모델이 생성한 확률 맵에서 가장 높은 응답 값을 갖는 위치를 추출하고, 해당 위치가 실제 조작 영역의 정답 마스크 내부에

포함되는지를 기준으로 Hit 또는 Miss를 판정하는 것이다[16]. Pointing Game의 수식은 다음과 같다.

$$Pointing\,Game = \frac{\#\mathrm{Hits}}{\#\mathrm{Hits}+\#\mathrm{Misses}} \quad (1)$$

식 (1)에서 Hit와 Miss는 각각 단일 테스트 이미지에 대한 판정 결과를 의미하며, #Hits와 #Misses는 전체 테스트 데이터셋에서의 Hit와 Miss의 총 개수를 의미한다.

표 2에서 볼 수 있듯이 제안하는 모델은 64.86%의 Pointing Game 정확도를 기록하였다. 기존 연구들에 따르면 CAM 및 Grad-CAM 기반 방법의 Pointing Game 정확도는 일반적으로 50~70% 범위로 보고되며[17], 이는 모델이 판별에 활용한 핵심 단서의 위치 적합성을 평가하는 기준으로 활용된다. 이러한 맥락에서 본 연구의 Pointing Game 평가 결과를 통해, 모델이 조작 영역의 주요 위치를 일정 수준 이상 식별할 수 있음을 확인하였다.

반면, mIoU는 예측된 영역과 정답 영역 간의 공간적 일치 정도를 정량적으로 평가하는 지표로, 분할 결과의 영역 수준 정합성을 측정하는 데 널리 사용된다[18]. mIoU는 예측 마스크와 정답 마스크의 교집합 영역을 합집합 영역으로 나눈 값으로 정의되며, 다음 식과 같이 계산된다.

$$\mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{|P_c \cap G_c|}{|P_c \cup G_c|} \quad (2)$$

식 (2)에서 C는 전체 클래스 수, P_C는 모델이 예측한 조작 영역 픽셀의 집합, G_C는 실제 조작 영역 픽셀의 집합을 의미한다.

표 2와 같이 제안하는 모델은 48.91%의 mIoU를 기록하였다. mIoU는 예측 영역과 실제 조작 영역 간의 공간적 일치 정도를 평가하는 대표적인 지표로, 일반적으로 정밀한 semantic segmentation 성능 평가에 사용된다. 본 연구의 결과는 경계 수준의 정밀한 마스크 복원 측면에서는 추가적인 개선의 여지가 있음을 의미한다. 그러나 본 연구의 분할 모듈은 완전한 semantic segmentation 자체보다는, 분류 결과에 대한 공간적 설명과 조작 의심 영역의 위치 단서를 제공하기 위한 보조 모듈로 설계되었다. 따라서 Pointing Game 결과와 함께 해석할 경우, 제안한 모델이 조작 영역의 주요 위치 단서를 일정 수준 이상 제공할 수 있음을 확인하였다.

Ⅴ. 결론 및 향후 과제

본 연구에서는 실제 이미지와 생성·조작 이미지를 판별하기 위해 CvT-13을 백본으로 사용하는 분류 및 분할 모델을 제안하였다. 분류 모델은 이미지 단위 판별에서 안정적인 성능을 보였으며, Pointing Game과 mIoU를 통한 정량적 평가 및 시각화 결과를 통해 분할 모델이 조작 영역의 주요 위치 단서를 제공할 수 있음을 확인하였다.

본 연구는 생성·조작 이미지 판별 문제에서 단순 이미지 단위 분류를 넘어, 조작 의심 영역의 공간적 위치를 함께 제시할 수 있는 explainable detection framework를 구성하였다는 점에서 의의를 가진다. 특히 CvT-13의 계층적 토큰 특징을 활용하여 분류, 분할, 시각화를 하나의 시스템으로 통합하였다. 분할 결과는 Pointing Game과 mIoU로 정량적으로 평가하였으며, Grad-CAM은 분할 모델 내부 특징 응답을 시각적으로 분석하기 위한 보조 기법으로 활용하였다. 또한 순수 트랜스포머 기반 ViT와 비교한 결과, 제안한 CvT-13 기반 구조 역시 생성·조작 이미지 판별에 적용 가능함을 확인하였다.

그러나 본 연구는 특정 데이터셋 기반으로 학습되어 일반 웹 이미지나 미검증 데이터 환경에서는 성능 저하 가능성이 존재하며, 외부 데이터셋 및 최신 탐지 모델과의 비교가 제한적으로 수행되었다. 또한 제안한 분할 모듈은 semantic segmentation 자체를 목표로 하기보다는 분류 결과에 대한 공간적 설명과 조작 의심 영역의 위치 단서를 제공하기 위한 보조 모듈로 설계되었으므로, 기존 segmentation 모델과의 직접적인 성능 비교에는 한계가 있다. 향후에는 다양한 실제 환경 데이터셋과 U-Net, SegFormer 등의 segmentation baseline, 그리고 CLIP/ViT, Swin Transformer, ConvNeXt, DINOv2 등 최신 탐지 모델과의 비교를 통해 제안 시스템의 일반화 성능과 공간적 설명 능력을 보다 객관적으로 검증할 예정이다.

Acknowledgment

2025년도 한국정보기술학회 추계종합학술대회에서 발표한 논문(CvT-13을 활용한 생성/조작 이미지 판별 및 시각화 시스템)[2]을 확장한 것임

References

[1] C. Lee and C. H. Yoon, "A Study on Machine Learning-Based Deep Voice Analysis for Crime Prevention Using Generative AI", Proc. of KIIT Conference, Jeju, Korea, pp. 25-29, Mar. 2024.

[2] J. H. Lim, W. J. Jang, D. E. Kim, and W. I. Lee, "CvT-13-Based Image Detection and Visualization System for Synthetic and Manipulated Images", Proc. of KIIT Conference, Jeju, Korea, pp. 1577-1580, Nov. 2025.

[3] J. Deng, C. Lin, Z. Zhao, S. Liu, Z. Peng, Q. Wang, and C. Shen, "A Survey of Defenses against AI-generated Visual Media: Detection, Disruption, and Authentication", ACM Computing Surveys, Vol. 57, No. 10, Art. no. 3770916, Oct. 2025, https://doi.org/10.1145/3770916.

[4] S. R. Dubey and S. K. Singh, "Transformer-Based Generative Adversarial Networks in Computer Vision: A Comprehensive Survey", arXiv preprint, arXiv:2302.08641, Feb. 2023. https://doi.org/10.48550/arXiv.2302.08641.

[5] C. Fross and Q. F. Liu, "Detection of Deepfake Videos and Audios on Social Media Platforms", Research Project Report, Department of Computer Science, Sam Houston State University, 2025.

[6] V. Kadha, S. Bakshi, and S. K. Das, "Unravelling Digital Forgeries: A Systematic Survey on Image Manipulation Detection and Localization", ACM Computing Surveys, Vol. 57, No. 12, pp. 1-36, Jul. 2025. https://doi.org/10.1145/3731243.

[7] D. Lamichhane, "Advanced Detection of AI-Generated Images Through Vision Transformers", IEEE Access, Vol. 13, pp. 3644-3652, Dec. 2024. https://doi.org/10.1109/ACCESS.2024.3522759.

[8] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing Convolutions to Vision Transformers", Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 22-31, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00009.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", International Journal of Computer Vision, Vol. 128, No. 2, pp. 336-359, Feb. 2020. https://doi.org/10.1007/s11263-019-01228-7.

[10] M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, and S. A. Fattah, "ArtiFact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection", Proc. IEEE Int. Conf. on Image Processing (ICIP), Kuala Lumpur, Malaysia, pp. 1-5, Oct. 2023. https://doi.org/10.1109/ICIP49359.2023.10222083.

[11] CASIA 2.0 Groundtruth Dataset, https://github.com/namtpham/casia2groundtruth. [accessed: Dec. 15, 2025]

[12] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection", Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 936-944, Jul. 2017. https://doi.org/10.1109/CVPR.2017.106.

[13] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", Proc. 36th Int. Conf. on Machine Learning (ICML), Long Beach, California, USA, pp. 6105-6114, Jun. 2019. https://doi.org/10.48550/arXiv.1905.11946.

[14] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery

Detection by Mining Frequency-aware Clues", Proc. European Conf. on Computer Vision (ECCV), Online, pp. 86-103, Aug. 2020. https://doi.org/10.1007/978-3-030-58610-2_6.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Proc. International Conference on Learning Representations (ICLR), Online, May 2021. https://doi.org/10.48550/arXiv.2010.11929.

[16] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down Neural Attention by Excitation Backprop", Proc. European Conf. on Computer Vision (ECCV), Amsterdam, The Netherlands, pp. 543-559, Oct. 2016. https://doi.org/10.1007/978-3-319-46493-0_33.

[17] J. Ahn and S. Kwak, "Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation", Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 4981-4990, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00523.

[18] L. Yu, Z. Li, M. Xu, Y. Gao, J. Luo, and J. Zhang, "Distribution-aware Margin Calibration for Semantic Segmentation in Images", International Journal of Computer Vision, Vol. 130, No. 12, pp. 3220-3240, Dec. 2022. https://doi.org/10.1007/s11263-021-01533-0.

저자소개

임 재 한 (JaeHan Lim)



2026년 2월 : 국립금오공과대학교 전자공학부(공학사)
2026년 3월 ~ 현재 : 국립금오공과대학교 반도체시스템공학과 석사과정
관심분야 : 영상처리, 딥러닝, 임베디드시스템, 메모리 반도체

김 도 언 (Do-Eon Kim)



2026년 2월 : 국립금오공과대학교 전자공학부(공학사)
2026년 3월 ~ 현재 : 테라다인 코리아 Memory Field Application Engineer
관심분야 : 메모리 반도체, 반도체 테스트, 소프트웨어

장 원 준 (WonJun Jang)



2026년 2월 : 국립금오공과대학교 전자공학부(공학사)
관심분야 : 딥러닝, 임베디드 시스템, AI 기반 이미지 분석

이 원 일 (Won Il Lee)



2010년 2월 : 경북대학교 전자전기컴퓨터학부(공학사)
2017년 2월 : 포항공과대학교 전자전기공학과(박사)
2017년 3월 ~ 2020년 8월 : 삼성디스플레이 책임연구원
2020년 9월 ~ 현재 : 국립금오공과대학교 전자공학부 부교수
관심분야 : 제어이론, 영상처리, 딥러닝, 임베디드시스템