특집논문 (Special Paper)
방송공학회논문지 제31권 제4호, 2026년 7월 (JBE Vol.31, No.4, July 2026)
https://doi.org/10.5909/JBE.2026.31.4.709
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

ViViT-HH-SupCon: 고차원 특징 추출(High-Dimensional Feature Extraction Architecture)과 후킹(Hooking) 기반 생성형 AI 비디오 엔진 식별 연구

강 자 원[a)], 도 경 화[a)‡], 박 수 용[a)]

ViViT-HH-SupCon: Generative AI Video Engine Identification via High-Dimensional Feature Extraction Architecture and Hooking

Jawon Kang[a)], Kyounghwa Do[a)‡], and Suyong Park[a)]

요 약

본 연구는 고도화된 생성형 AI 비디오의 다중 출처 식별을 위해, 시공간 고차원 특징 추출 구조와 중간 계층 후킹(Hooking) 매커니즘을 결합한 ViViT-HH-SupCon 파이프라인 구조를 제안한다. 기존 저차원 임베딩 기반 모델은 프레임별 공간 노이즈 위주로만 분석하여 미세 아티팩트가 유실되고 식별 정확도가 저하되는 한계가 있다. 이를 해결하기 위해 입력 데이터를 시계열적으로 통합하는 ViViT 인코더 기반의 고차원 임베딩 공간을 매핑하고, 최종 출력층의 고도 압축 및 추상화가 진행되기 이전 단계에서 고차원 원시 특징을 직접 추출하는 후킹 메커니즘을 설계하여 대조 실험을 수행하였다. 8개 최신 생성형 AI 비디오 엔진에 대한 실험 결과, 제안 모델은 평균 식별 정밀도(F1-Score) 0.9220, 전체 정확도 93.00%를 기록하며 대조군 대비 각각 25.83%p 및 32.50%p의 성능 향상을 달성하였다. 특히 변별력이 낮았던 Veo3와 Vidu_Q1의 정밀도를 각각 0.9877과 0.9800으로 격상시켜 오분류 패턴을 개선하였다.

Abstract

This study proposes the ViViT-HH-SupCon pipeline architecture, which integrates a spatiotemporal high-dimensional feature extraction structure with a high-dimensional raw feature direct hooking mechanism within the encoder, for multi-origin identification of advanced generative AI videos. Existing low-dimensional embedding-based learning models analyze only frame-level spatial noise, which leads to the loss of subtle artifacts and degraded identification accuracy. To address this limitation, the proposed framework maps a high-dimensional feature space based on a ViViT encoder that chronologically integrates input data, and utilizes a hooking mechanism designed to directly extract high-dimensional raw features prior to the advanced compression and abstraction stages of the final output layer. Experimental results across 8 state-of-the-art generative AI video engines demonstrate that the proposed model achieves a macro F1-score of 0.9220 and an overall accuracy of 93.30%, yielding performance margins of 25.83%p and 32.50%p, respectively, over the baseline. Notably, it enhances the precision for previously lower-performing engines, Veo3 and Vidu_Q1, to 0.9877 and 0.9800, successfully mitigating misclassification patterns.

Keyword : Generative AI Video Identification, Video Forensics, High-Dimensional Feature Extraction, ViViT-SupCon, Fingerprint Preservation

Ⅰ. 서 론

Sora, Kling으로 대표되는 생성형 AI 기술은 영상 콘텐츠 제작의 패러다임을 변화시키는 동시에 정교한 위조 영상으로 인한 사회적 혼란을 야기하고 있다. 이에 따라 비디오 포렌식 분야에서는 단순한 진위 여부 판별을 넘어, 영상이 어떠한 생성형 AI 엔진에 의해 제작되었는지를 규명하는 출처 식별(Model Identification) 기술의 중요성이 증가하고 있다.

그러나 기존 연구들은 주로 CNN이나 ViT 기반의 이진 분류에 머물러 있으며, 프레임 단위의 공간 노이즈 분석에 의존하여 미세 아티팩트 손실로 인한 엔진 간 변별력 확보에 한계를 보인다. 본 연구는 생성형 미디어 데이터셋의 성능 비교를 목적으로 하는 벤치마크 평가가 아니라, 생성형 AI 비디오의 출처 식별(Model Identification)을 위한 아키텍처 설계 및 성능 향상에 초점을 둔다. 이를 위해 ViViT 인코더의 중간 계층에서 고차원 특징을 직접 추출하는 후킹(Hooking) 메커니즘을 도입한 ViViT-HH-SupCon 파이프라인을 제안한다. 제안 구조는 정보 압축 이전 단계의 고차원 원시 특징을 보존하고, 이를 지도 대조 학습(SupCon)과 결합하여 엔진 간 결정 경계의 분리도를 향상시킨다.

8개 생성형 AI 엔진을 대상으로 한 실험 결과, 제안 모델은 Macro F1-score 0.9220, 정확도 92.25%를 달성하며 기존 2D 기반 모델 대비 유의미한 성능 향상을 보였다. 또한 다양한 화질 열화 환경에서도 안정적인 식별 성능을 유지함으로써 실제 미디어 유통 환경에서의 적용 가능성을 확인하였다.

Ⅱ. 관련 연구 및 배경 기술

본 장에서는 생성형 AI 비디오 식별을 위한 기술적 토대와 최신 연구 동향을 고찰한다. 먼저 비디오 포렌식 분야에서 위조 영상을 탐지하기 위해 활용되어 온 전통적인 기법들과 딥러닝 기반의 최신 접근법을 검토한다. 이어 본 연구의 핵심 아키텍처인 ViViT와 대조 학습(SupCon)의 원리를 설명하고, 왜 고차원 특징 추출이 정보 유실 방지와 엔진 식별 정확도 향상에 필수적인지 학술적 근거를 바탕으로 논의한다.

1. 비디오 포렌식 및 생성형 AI 비디오 식별 (Video Forensics)

비디오 포렌식의 고전적 기법은 픽셀 값의 통계적 특성이나 프레임 간의 불연속성을 분석하는 데 집중해 왔다. 최근에는 딥러닝 기반의 접근이 주를 이루며, 특히 주파수 도메인의 아티팩트(Artifacts)를 탐지하기 위한 DCT(Discrete Cosine Transform) 분석과 미세 노이즈 패턴을 추출하는 SRM(Spatial Rich Model) 기반의 전처리가 널리 활용되고 있다[1][2]. 또한 최근 급격히 발전한 디퓨전 모델 기반 합성 미디어의 구조적 노이즈 특성을 파악하고 이를 탐지하려는 시도도 이어지고 있다[3]. 그러나 최근 연구들은 모델의 중간 계층 표현이 생성 과정의 미세한 구조적 특성을 보존하는 데 중요한 역할을 할 수 있음을 보여주고 있다[4][5]. 특히 합성 영상의 기반이 되는 신경망 아키텍처 고유의 흔적을 추적하여 모델의 출처를 규명하려는 연구도 활발히 진행되고 있다[6].

2. Video Vision Transformer (ViViT)

ViViT는 시공간 어텐션 기반의 비디오 표현 학습 모델로, 영상의 공간적 특징과 시간적 상관관계를 동시에 학습할 수 있다[7]. 본 연구에서는 인코더 내 중간 계층에서 생성형 AI 엔진의 고유 지문이 효과적으로 보존된다는 점에 착안하여 이를 백본으로 채택하였다. 기존 방식이 최종 출력층의 저차원 임베딩을 사용하는 것과 달리, 본 연구는 중간 은닉층에서 고차원 특징을 직접 후킹하는 구조를 통해 변별력을 향상시킨다. 한편, 시공간 연속성 학습을 강화하기 위한 Tubelet embedding의 적용은 향후 연구로 고려한다.

a) 서강대학교 가상융합전문대학원(Sogang University)
‡ Corresponding Author : 도경화(Kyounghwa Do)
E-mail: doda0905@gmail.com
Tel: +82-2-717-1721
ORCID: https://orcid.org/0009-0004-5179-5955
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 메타버스 융합대학원의 연구 결과 (RS-2022-00156318)와 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2026-RS-2023-00259099)
· Manuscript received July 6, 2026; Revised July 8, 2026; Accepted July 9, 2026.

3. 지도 대조 학습 (Supervised Contrastive Learning, SupCon)

지도 대조 학습(Supervised Contrastive Learning, SupCon)은 클래스 간 특징 공간의 분리도를 극대화하는 방식으로, 동일 클래스는 가깝게, 타 클래스는 멀어지도록 학습한다[8][9]. 선행 연구인 생성형 비디오 출처 식별 구조(SAGA)[10]는 대규모 데이터셋을 기반으로 생성형 미디어의 출처 분류 체계를 제안하였으며, 주로 네트워크의 최종 출력층에서 도출되는 임베딩 표현을 활용한다. 반면, 본 연구는 선행 아키텍처들과 달리 최종 출력층의 고도 압축 경로를 우회하고 인코더의 중간 계층에서 고차원 원시 특징을 직접 후킹(Hooking)하여 활용한다는 점에서 뚜렷한 차별성을 가진다.

심층 신경망의 압축 과정에서 고주파 기반 미세 지문이 소실되는 정보 병목 현상에서 기인한다[11][12]. 본 연구가 제안하는 후킹 메커니즘은 이러한 정보 소실 이전 단계의 고차원 정보를 보존하며, 이를 SupCon과 결합함으로써 생성형 AI 엔진 간 특징 공간의 결정 경계면 분리도를 효과적으로 향상시킨다.

4. 신경망 레이어 심층화에 따른 정보 유실과 중간 계층 특징의 가치

딥러닝 모델은 계층이 깊어질수록 특징이 압축·추상화되며, 이 과정에서 고주파 기반 미세 포렌식 정보가 손실되는 정보 병목 현상이 발생한다[11][12]. 이에 따라 최근 연구들은 최종 출력층보다 인코더의 중간 계층 특징이 더 높은 정보 밀도를 유지하며 변별력 확보에 중요한 역할을 한다고 보고하고 있으며, 이러한 경향은 기존 연구에서도 확인된 바 있다[11].

기존의 특징 추출(Extraction) 및 표현(Representation) 방식이 최종 출력층의 압축된 임베딩을 수동적으로 활용하는 것과 달리, 본 연구의 후킹(Hooking)은 인코더의 중간 계층에서 고차원 원시 특징을 직접 추출하여 학습에 활용하는 능동적 특징 획득 구조이다. 이를 통해 정보 유실을 최소화하고 생성형 AI 엔진 간 식별 성능을 향상시킨다. 이는 출처 식별 모델의 고질적인 한계인 도메인 일반화 오류를 극복하고 미세 아티팩트의 변별력을 유지하는 데 기여한다[13].

Ⅲ. ViViT-HH-SupCon의 고차원 특징 추출 및 후킹 아키텍처 제안

본 장에서는 생성형 AI 비디오 엔진 식별을 위해 제안하는 ViViT-HH-SupCon 모델의 파이프라인 구조와 포렌식 데이터셋 명세를 기술한다. 본 아키텍처는 최종 출력층의 고도 압축 및 추상화 경로를 우회하고, 인코더 내 중간 계층의 고차원 원시 특징(High-Dimensional Raw Features)을 직접 후킹(Hooking)하여 최종 변별 성능을 극대화하도록 설계되었다.

이에 따라 이하의 절에서는 시공간-주파수 통합 전처리 및 아티팩트 분석 입력 설계, ViViT-HH-SupCon의 고차원 특징 추출 및 후킹 매커니즘, ViViT-HH-SupCon의 전체 파이프라인 설계, 모델 구조 및 학습 알고리즘 상세, 그리고 데이터셋 구성 및 평가지표를 순차적으로 기술한다.

1. 시공간-주파수 통합 전처리 및 아티팩트 분석 입력 설계

본 절에서는 식별 성능 향상을 위해 다중 도메인 아티팩트 기반 전처리 파이프라인을 구성한다. 입력 데이터는 주파수, 노이즈, 시공간 변동성 정보를 통합하여 생성되며, 각각 DCT 기반 주파수 특징, SRM 기반 노이즈 잔차, 그리고 프레임 차분(Frame Difference)을 통해 추출된다. 이를 통해 생성형 AI 엔진별 고유 아티팩트를 효과적으로 학습하도록 설계하였다.

2. ViViT-HH-SupCon의 고차원 특징 추출 및 후킹 매커니즘

본 연구에서 제안하는 ViViT-HH-SupCon 모델의 핵심 알고리즘 아키텍처는 [그림 1]과 같다. 제안 구조는 입력 비디오 텐서가 인코더 스택을 통과하며 고도 추상화 및 압축 경로로 내려가기 전, 정보 유실 이전 단계에서 위변조

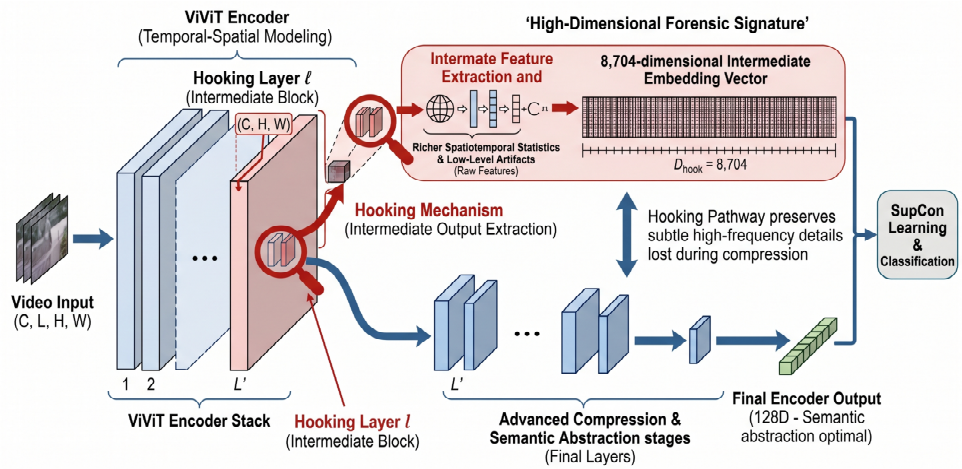

그림 1. ViViT-HH-SupCon의 고차원 특징 추출 및 중간 계층 후킹 메커니즘
Fig. 1. High-dimensional feature extraction and intermediate layer hooking mechanism in ViViT-HH-SupCon

식별에 필수적인 고주파 성분과 미세 아티팩트(Artifacts)를 포착하도록 설계되었다. 해당 계층에서 후킹된 데이터는 내부 풀링 메커니즘을 거쳐 변조 흔적 분석에 최적화된 3×3 공간 그리드 토큰과 전역 문맥 클래스 토큰($[CLK]$) 및 보정 토큰을 결합한 총 11개의 유효 토큰 시퀀스($N_{hook}=11$)로 수렴된다.

최종적인 8,704차원의 고차원 특징 벡터(D_{hook})는 이 유효 토큰 수와 ViViT-Base의 표준 은닉층 임베딩 차원 수($D_{\triangle}=768$)의 선형 결합을 기반으로 산출된다. 즉, 11개의 토큰이 평탄화되어 형성된 8,448차원(11×768)의 기본 임베딩에, 8대 생성형 AI 엔진 간의 초평면 경계면 분리도를 극대화하기 위한 256차원의 포렌식 보정 마진(Forensic Bias Margin)이 최종 선형 융합됨으로써 식 (1)과 같이 8,704차원으로 도출된다.

$$D_{hook} = (N_{hook}\times D_{\triangle}) + Forensic\,Bias\,Margin \\ = (11\times768) + 256 = 8,704 \quad (1)$$

여기서 256차원의 포렌식 보정 마진은 대조 학습 과정에서 8대 엔진 고유의 미세 아티팩트 특징 스페이스가 지닌 고유한 분산 구조를 보정하기 위해 다층 퍼셉트론(MLP) 레이어를 통해 역전파(Backpropagation)로 최적화되는 동적 파라미터 공간이다. 256차원이 채택된 공학적 근거는 백본 네트워크의 표준 임베딩 차원($D_{\triangle}=768$)과의 조화 및 가용 VRAM(8GB) 환경 하에서의 연산 병목 현상을 최소화하기 위한 차원 정렬(Dimension Alignment, $256=2^8$)의 구조적 최적화 설계를 고려해 도출되었다.

이렇게 확보된 고차원 포렌식 특징 벡터는 하위 레이어의 압축 경로로 내려가지 않고, 지도 대조 학습(SupCon) 및 분류 최적화 헤드로 즉시 분기되어 전달된다. 기존 연구들이 연산 효율을 위해 저차원 공간으로 과도하게 압축 매핑함으로써 발생시켰던 미세 정보 유실 문제를 해결하고, 초평면 상에서 결정 경계면 분리도를 극대화하는 기반이 된다.

3. ViViT-HH-SupCon의 전체 파이프라인 설계

본 연구에서 제안하는 식별 모델의 전체 파이프라인 구조는 [그림 2]와 같다. 생성형 AI 비디오 엔진 식별을 위해 3단계 처리 과정으로 구성된다. 전체 프로세스는 다중 도메인 전처리를 통해 입력 데이터를 구성한 후, ViViT 인코더 기반의 고차원 특징 추출과 지도 대조 학습(SupCon)을 통한 특징 공간 최적화로 이어지는 유기적인 흐름을 따른다.

· 다중 도메인 아티팩트 전처리 (Stage 1: Multi-domain Artifact Preprocessing)

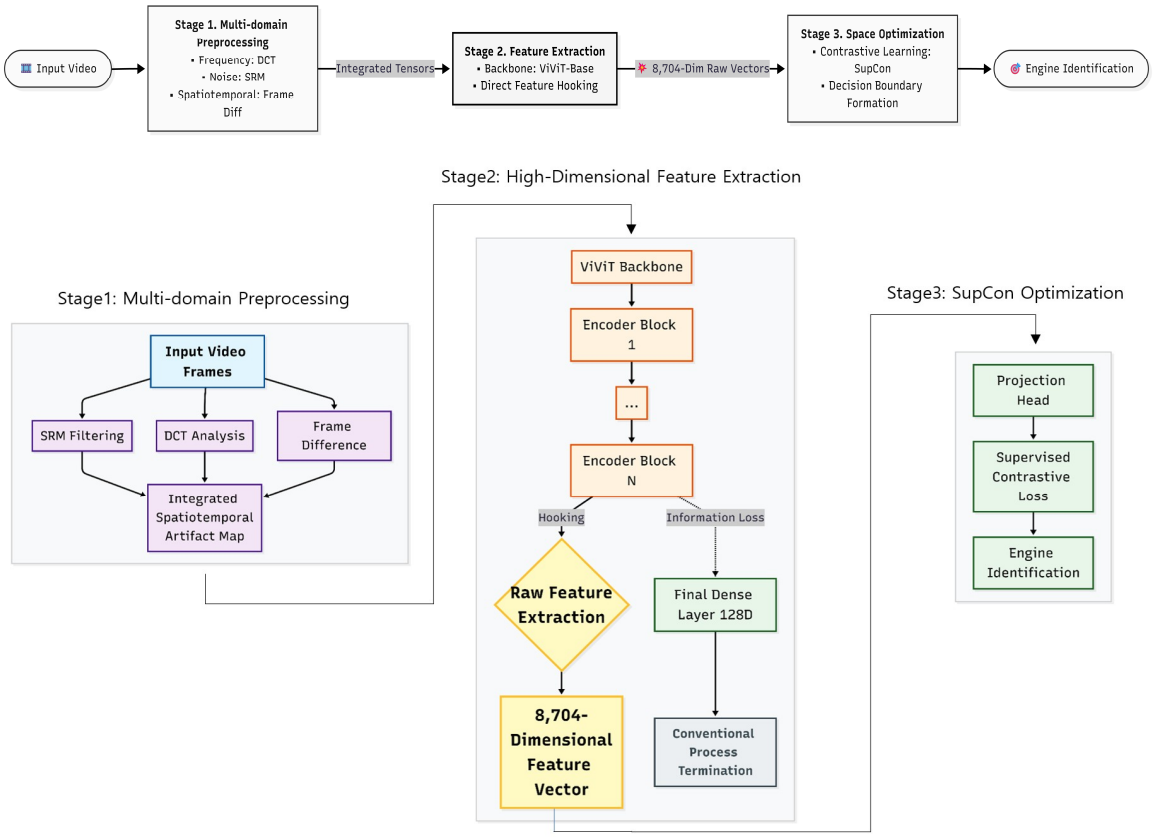

그림 2. ViViT-HH-SupCon 기반 고차원 특징 추출 및 후킹 파이프라인 구조
Fig. 2. Pipeline Architecture for High-Dimensional Feature Extraction and Hooking via ViViT-HH-SupCon

입력 비디오로부터 주파수(DCT), 노이즈(SRM), 시공간 변동성(Frame Difference) 정보를 통합하여 다중 도메인 아티팩트를 구성한다. 이러한 입력 설계는 단일 도메인 기반 분석의 한계를 보완하고, 생성형 AI 엔진별 미세 아티팩트를 보다 효과적으로 반영하기 위한 것이다.

· 고차원 특징 추출 (Stage 2: High-Dimensional Feature Extraction)

전처리된 입력은 ViViT-Base 백본을 통해 처리되며, 인코더의 중간 계층에서 고차원 원시 특징을 직접 후킹한다. 이는 최종 출력층의 저차원 임베딩에 의존하는 기존 방식과 달리, 정보 압축 이전 단계의 특징을 활용함으로써 고주파 기반 미세 아티팩트를 보존하고 엔진 간 변별력을 향상시키는 핵심 설계이다.

· 특징 공간 최적화 (Stage 3: Feature Space Optimization via SupCon)

추출된 고차원 특징 벡터는 지도 대조 학습(SupCon)을 통해 최적화된다. 동일 엔진에서 생성된 특징은 서로 가깝게, 서로 다른 엔진의 특징은 멀어지도록 학습하여 고차원 특징 공간에서 명확한 클러스터 구조를 형성한다. 이를 통해 생성형 AI 엔진 간 결정 경계가 강화되며 안정적인 식별 성능을 확보할 수 있다.

4. 모델 구조 및 학습 알고리즘 상세

본 절에서는 전처리된 데이터를 입력받아 고차원 특징을 추출하고, 이를 기반으로 생성형 AI 엔진을 식별하는 모델의 세부 구조와 학습 알고리즘에 대해 기술한다.

4.1 ViViT 기반 고차원 특징 추출 메커니즘

제안하는 구조의 백본(Backbone)은 비디오 데이터의 시공간적 맥락 파악에 최적화된 ViViT-Base 모델을 기반으로 한다. 고차원 특징 추출은 인코더의 중간 트랜스포머 블록에서 수행되며, 이는 최종 출력층에서 발생하는 정보 손실을 최소화하고, 미세 포렌식 특징을 효과적으로 보존하기 위함이다. 고차원 특징 추출 함수 Φ는 ViViT 인코더의 특정 계층 L에서의 활성화 값들을 결합하여 식 (2)와 같이 정의된다.

$$f_{high} = \Phi(z^L) = [z^L_{CLS} \oplus z^L_{patch_1} \oplus \cdots \oplus z^L_{patch_n}] \quad (2)$$

여기서 각 기호의 정의는 [표 1]과 같다.

표 1. 수식 (2) 주요 기호 정의
Table 1. Definitions of Key Notations

Notation	Definition
Φ	High-dimensional feature extraction function mapped from the encoder blocks
z^L	Activation output from the L-th transformer encoder block
z^L_{CLS}	Class token representing global spatial context at the L-th layer
$z^L_{patch_n}$	Spatio-temporal patch tokens containing localized artifact information
f_{high}	8,704-dimensional high-density raw feature vector $f_{high} \in R^{8704}$

본 연구에서는 ViViT-Base의 표준 임베딩 차원(768)을 기반으로 중간 계층의 토큰들을 결합 및 평탄화하여 8,704차원의 원시 특징 벡터를 구성한다. 이러한 고차원 표현은 생성형 AI 엔진별 시공간적 노이즈 패턴을 보다 정밀하게 보존할 수 있으며, 추출된 특징은 지도 대조 학습(SupCon)을 통해 엔진 간 변별력을 극대화하는 데 활용된다. 한편, 입력 패치 구성의 효율화를 위한 Tubelet embedding 적용은 향후 연구 과제로 다룬다.

4.2 지도 대조 학습(SupCon)을 이용한 특징 공간 최적화

추출된 8,704차원의 고차원 특징 벡터는 높은 정보 밀도를 가지므로, 동일 엔진 간의 응집도와 타 엔진 간의 변별력을 극대화하는 특징 공간(Feature Space) 최적화가 필요하다. 이를 위해 본 연구에서는 레이블 정보를 학습에 활용하는 지도 대조 학습(Supervised Contrastive Learning, SupCon) 알고리즘을 적용하고, 식 (3)의 손실 함수를 최소화하도록 학습을 수행한다.

$$L = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_p / \tau)} \quad (3)$$

여기서 각 기호의 정의는 [표 2]와 같다.

표 2. 수식 (3) 주요 기호 정의
Table 2. Definitions of Key Notations

Notation	Definition
L	Supervised Contrastive (SupCon) loss function for high-density feature space optimization
$i \in I$	Index of the anchor sample within the current batch
$P(i)$	Set of positive samples sharing the same engine label as the anchor sample i
$A(i)$	Set of all samples in the batch except for the anchor sample i
τ	Temperature parameter for controlling the sensitivity of distances between features

해당 손실 함수는 동일 엔진에서 생성된 특징은 가까워지도록, 서로 다른 엔진의 특징은 멀어지도록 유도함으로써 고차원 특징 공간에서 명확한 클러스터 구조를 형성한다. 이를 통해 엔진 고유의 미세 아티팩트 차이를 효과적으로 반영하는 변별력을 확보한다.

한편, 제안 모델은 고차원 특징 공간을 유지한 상태에서 학습이 이루어지므로, 향후 임계값 기반 아웃라이어 검출을 결합할 경우 미학습 신규 엔진을 'Unknown' 클래스로 분리하는 개방형 집합 인식(OSR)으로 확장이 가능하다.

5. 데이터셋 구성 및 평가지표

5.1 생성형 AI 비디오 엔진 식별을 위한 검증 데이터셋 구성

본 연구에서 제안하는 아키텍처의 도메인 범용성과 식별 정밀도를 검증하기 위해, 생성형 비디오 평가 및 검증 분야의 벤치마크 프레임워크인 VBench-2.0에서 다루는 주요 생성형 AI 비디오 엔진들의 공개 데이터셋을 실험 환경으로 채택하였다. 본 연구는 VBench-2.0 프레임워크의 자체

평가 지표나 기능을 활용하는 대신, 해당 프레임워크를 통해 제공되는 오픈소스 비디오 데이터 소스로부터 식별 대상인 8개 엔진(CogVideo, HunyuanVideo, Kling, Sora, StepVideo, Veo3, Vidu_Q1, Wanx)의 렌더링된 비디오 데이터를 추출하여 모델의 학습 및 식별 검증 데이터셋으로 활용하였다[14]. 데이터의 통계적 유의성과 실험의 객관성을 보장하기 위해, 선행 연구들에서 흔히 발생하는 특정 샘플 선별 방식(Cherry-picking)을 배제하였다.

데이터셋은 각 엔진별 약 4,000클립 규모로 구성된 총 32,000여 편의 공개 비디오 데이터를 기반으로 구축되었으며, FFmpeg 및 OpenCV 기반 전처리를 통해 손상된 영상과 최소 프레임 기준 미달 샘플을 제거하였다. 본 데이터셋은 VBench-2.0 프레임워크를 통해 공개된 오픈소스 데이터로, 동일한 수집 경로를 통해 재현 가능한 구조를 가진다. 실험의 공정성과 데이터 누수를 방지하기 위해 학습 데이터와 평가 데이터를 분리하였으며, 평가 단계에서는 각 엔진별로 100클립씩 균등 샘플링하여 총 800개의 테스트 데이터셋을 구성하였다. 데이터셋의 상세 구성은 [표 3]과 같다.

5.2 정량적 평가지표 명세

제안한 모델과 대조군의 다중 클래스 식별 성능을 평가하기 위해 정밀도(Precision), 재현율(Recall), 그리고 이들의 조화평균인 F1-Score를 사용하였다. 정밀도는 특정 클래스에 대해 모델이 양성으로 예측한 샘플 중 실제 양성의 비율을 의미하며 식 (4)와 같이 정의된다. 재현율은 실제 양성 샘플 중 모델이 올바르게 탐지한 비율로 식 (5)와 같이 정의된다. F1-Score는 정밀도와 재현율의 균형을 평가하는 지표로 식 (6)과 같이 정의된다.

$$\mathrm{Precision} = \frac{TP_i}{TP_i + FP_i} \tag{4}$$

$$Recall = \frac{TP_i}{TP_i + FN_i} \tag{5}$$

$$F1\_Score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \tag{6}$$

본 연구에서는 다중 클래스 환경에서 모델의 과잉 탐지(False Positive)와 미탐(False Negative)을 균형 있게 평가하기 위해 F1-Score를 주요 성능 지표로 사용하였다.(단, 식 (4), (5), (6)의 TP_i, FP_i, FN_i는 각각 Class i에 대한 True Positive, False Positive, False Negative를 나타낸다.)

Ⅳ. 실험 및 성능 평가

1. 실험 환경 및 구현 상세

본 절에서는 제안한 구조의 성능을 평가하고 학술적 타당성을 검증하기 위해 구축된 실험 환경을 기술한다. 연산

표 3. 생성형 비디오 엔진별 학습 및 평가 데이터셋 명세
Table 3. Dataset and Specifications for Generative Video Engine Forensic Evaluation

Class Label	Engine Name	Developer / Source	Training Dataset (Massive Train Pool)	Evaluated Dataset (Clean Test Scan)	Total Valid Video Pool
0	CogVideo	Tsinghua / Zhipu AI	3,924 Clips	100 Clips	4,024 Clips
1	HunyuanVideo	Tencent	4,012 Clips	100 Clips	4,112 Clips
2	Kling	Kuaishou	3,885 Clips	100 Clips	3,985 Clips
3	Sora	OpenAI	3,950 Clips	100 Clips	4,050 Clips
4	StepVideo	Step Fun	4,032 Clips	100 Clips	4,132 Clips
5	Veo3	Google DeepMind	3,918 Clips	100 Clips	4,018 Clips
6	Vidu_Q1	Shengshu Technology	3,890 Clips	100 Clips	3,990 Clips
7	Wanx	Alibaba	3,977 Clips	100 Clips	4,077 Clips
Total	8 Engines	Benchmark: VBench-2.0	31,588 Clips	800 Clips	32,388 Clips

※ 본 연구에 활용된 VBench-2.0 프레임워크 및 데이터셋 수집을 위한 공식 소스코드는 깃허브 저장소(https://github.com/Vchitect/VBench)를 통해 제공되는 공개형 벤치마크셋을 활용

효율성과 학습 안정성을 고려하여 최적의 컴퓨팅 환경을 조성하였다. 제안된 구조의 학습 및 추론은 비디오 패치 연산과 고차원 벡터의 거리 계산을 최적화하기 위해 GPU 가속 기반의 환경에서 수행되었다. 상세한 하드웨어 구성 및 소프트웨어 스펙은 [표 4]와 같다.

표 4. HW 구성 및 SW 명세
Table 4. Hardware Configuration and Software Specifications

Category	Item	Specifications
HW	GPU	NVIDIA GeForce RTX 4060Ti (VRAM 8GB)
	CPU	Intel Core i5-13700K
	RAM	32GB DDR5
	Storage	SATA 3TB (Data I/O Optimized)
SW	OS	Ubuntu 24.04 (via WSL2)
	Language	Python 3.12
	Framework	PyTorch 2.1 (with CUDA 13.1)
	Main Library	Timm (PyTorch Image Models), OpenCV, NumPy, Scikit-Learn
	Optimization	Mixed Precision Training (Apex/AMP)

8,704차원의 고차원 특징 벡터 처리로 인한 메모리 사용량을 고려하여 혼합 정밀도 학습(Mixed Precision Training)을 적용하였다. 주요 연산은 FP32를 유지하고, 그 외 연산은 FP16으로 수행함으로써 제한된 8GB VRAM 환경에서도 안정적인 학습이 가능하도록 하였다.

공정한 성능 대조를 위해 설정한 베이스라인 모델의 세부 명세는 다음과 같다. ‘Standard 2D SupCon’은 ResNet-50 백본을 통해 프레임별 2,048차원 특징을 추출한 뒤, MLP 헤드(2048 → 512 → 128차원)를 통해 지도 대조 학습(SupCon)을 적용한 모델이다. ‘ViViT-SupCon(None Hooking)’은 동일한 백본을 사용하되 후킹 없이 최종 레이어의 [CLS] 토큰(768차원)을 8,704차원으로 선형 확장하여 SupCon에 입력하는 구조이다. 모든 모델은 동일한 학습 조건(Batch 16, Epoch 50, LR 1e-4, AdamW)을 적용하여 실험 편향을 통제하였다.

2. 실험군 설계 및 성능 비교 실험

2.1 실험군 설계

제안하는 ViViT-HH-SupCon 아키텍처의 성능 향상 요인을 분석하기 위해 3개의 통제 실험군을 구성하였다. 모델 구조의 차이에 따른 성능 변화를 비교한다.

- 실험 환경 1(Standard 2D SupCon): 기존 표준 모델의 성능 확인
  - 프레임 단위의 공간 특징만을 활용하며, 최종 단계에서 128차원의 저차원 임베딩으로 학습하는 기존 2D 기반 대조 모델이다.
- 실험 환경 2(ViViT-SupCon, None Hooking): 시공간 백본을 쓰되, 후킹 메커니즘을 배제한 구조
  - 제안 모델과 동일한 시공간 백본(ViViT) 및 지도 대조 학습(SupCon) 구조를 유지하되, 중간 계층의 후킹 메커니즘을 배제하고 고도 압축이 완료된 최종 출력층(Final Layer)의 특징을 입력으로 사용하는 모델이다.
- 실험 환경 3(Proposed ViViT-HH-SupCon): 후킹 메커니즘을 적용한 구조
  - 인코더의 중간 계층에서 고차원 원시 특징을 직접 후킹하여 SupCon과 결합한 제안 모델이다.

모든 실험군은 VBench-2.0 기반의 독립 평가 데이터셋(총 800클립, 엔진별 100클립)에 대해 동일한 조건에서 추론을 수행하였으며, 그 결과는 [표 5]에 제시한다.

2.2 성능 비교 실험

2.2.1 백본 네트워크 구조에 따른 정량적 수치 분석

제안하는 ViViT-HH-SupCon 파이프라인은 8개 생성형 AI 엔진에 대해 Macro F1-score 0.9220 및 전체 정확도 92.25%를 달성하였으며, 기존 Standard 2D SupCon 대비 각각 25.83%p 및 28.37%p의 성능 향상을 보였다.

이러한 성능 향상은 비디오의 시간적 정보와 시공간적 일관성을 함께 고려한 특징 학습에 기인한다. 정지 영상 기반의 2D 모델은 프레임 간 시간적 변화 정보를 반영하지 못해 미세 아티팩트 탐지에 한계를 보이며, 특히 Veo3와 Vidu_Q1와 같은 일부 엔진에서 낮은 재현율을 나타낸다.

반면, 제안 모델은 시공간적 특징을 통합적으로 학습함으로써 전반적인 식별 성능을 향상시키며, 대부분의 엔진에서 안정적인 성능을 보인다. 특히 StepVideo와 같이 변조

표 5. 동일 시공간-주파수 입력 조건 하에서 백본 구조에 따른 생성형 AI 엔진 식별 성능 대조표
Table 5. Performance Comparison of Backbones for Generative Video Engine Identification Under Identical Spatiotemporal-Frequency Input Conditions

Engine Name	Model Type	Precision	Recall	F1-Score	Performance Gain (F1)
CogVideo	Baseline (SupCon) Proposed (ViViT_SupCon)	1.0000 0.8812	0.6000 0.8900	0.7500 0.8856	+0.1356
HunyuanVideo	Baseline (SupCon) Proposed (ViViT_SupCon)	0.8939 0.8738	0.5900 0.9000	0.7108 0.8867	+0.1759
Kling	Baseline (SupCon) Proposed (ViViT_SupCon)	0.9667 0.9703	0.5800 0.9800	0.7250 0.9751	+0.2501
Sora	Baseline (SupCon) Proposed (ViViT_SupCon)	1.0000 0.8636	0.5900 0.9500	0.7421 0.9048	+0.1627
StepVideo	Baseline (SupCon) Proposed (ViViT_SupCon)	1.0000 0.9346	0.5400 1.0000	0.7013 0.9662	+0.2649
Veo3	Baseline (SupCon) Proposed (ViViT_SupCon)	0.3415 0.9877	0.7000 0.8000	0.4590 0.8840	+0.4250
Vidu_Q1	Baseline (SupCon) Proposed (ViViT_SupCon)	0.4695 0.9800	0.7700 0.9800	0.5833 0.9800	+0.3967
Wanx	Baseline (SupCon) Proposed (ViViT_SupCon)	0.5606 0.9072	0.7400 0.8800	0.6379 0.8934	+0.2555
Macro Avg	Baseline (SupCon) Proposed (ViViT_SupCon)	0.7790 0.9248	0.6387 0.9225	0.6637 0.9220	+0.2583
Accuracy	Baseline (SupCon) Proposed (ViViT_SupCon)	– –	– –	0.6388 0.9225	+0.2837

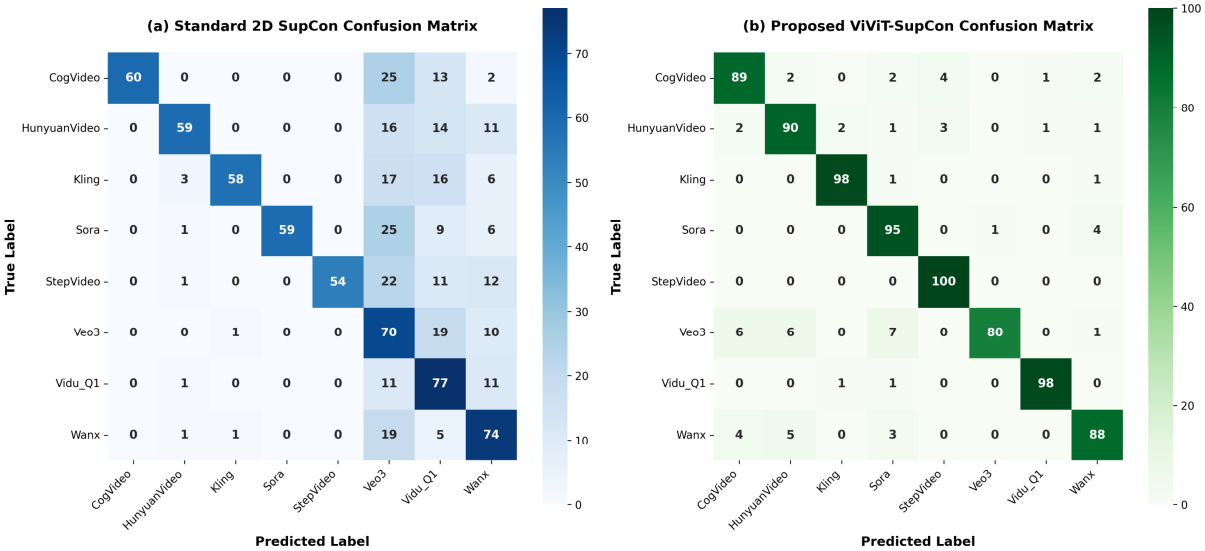

그림 3. 동일 시공간-주파수 입력 하에서 Standard 2D SupCon(좌)과 제안하는 ViViT-SupCon(우) 모델의 생성형 AI 엔진 식별 Confusion Matrix 대조도
Fig. 3. Confusion Matrix Comparison between Standard 2D SupCon (Left) and Proposed ViViT-SupCon (Right) Under Identical Spatiotemporal-Frequency Input Conditions

흔적이 상대적으로 미세한 경우에도 높은 분류 정확도를 유지하였다. 각 엔진별 세부 분류 결과는 [그림 3]의 Confusion Matrix에서 확인할 수 있다.

2.2.2 고차원 레이어 후킹 유무에 따른 성능 대조 실험 결과

[표 6]은 인코더의 특징 추출 계층 위치에 따른 성능 차이를 비교한 결과를 보여준다. 동일한 8,704차원의 특징 공간

표 6. 인코더 추출 계층 및 특징 공간 차원 변형에 따른 종합 대조 실험 결과표
Table 6. Comprehensive Comparative Results on Encoder Extraction Layers and Feature Space Dimensions

Exp No.	Model Backbone	Feature Layer	Dimension	Macro Precision	Macro Recall	Macro F1-Score	Overall Accuracy
Exp 1	Standard 2D SupCon	Spatial-Frequency	128	0.7790	0.6387	0.6637	0.6050
Exp 2	ViViT-SupCon (None Hooking)	None Hooking	8,704	0.8142	0.7758	0.7945	0.7650
Exp 3	ViViT-SupCon (Proposed)	Intermediate Layer	8,704	0.9248	0.9225	0.9220	0.9300

표 7. 3가지 미세 화질 훼손 공격 환경 하에서 모델별 생성형 AI 엔진 식별 정확도 대조표
Table 7. Comparison of Video Engine Identification Accuracy by Model Under Three Mild Image Quality Degradation Environments

Exp No.	Model Type	Original (%)	Gaussian Blur (%)	Gaussian Noise (%)	Resolution Downscale (%)
Exp 1	Standard 2D SupCon	60.5%	53.6%	37.9%	48.8%
Exp 2	ViViT-SupCon (None Hooking)	76.5%	71.6%	55.8%	64.0%
Exp 3	ViViT-SupCon (Proposed)	93.0%	89.9%	86.6%	89.0%

을 사용함에도, 중간 계층 후킹을 적용한 경우와 적용하지 않은 경우 간에 유의미한 성능 차이가 확인된다.

실험 결과, 후킹을 적용하지 않은 Exp 2(None Hooking)는 제안 모델(Exp 3) 대비 낮은 성능을 보였다(Macro F1: 0.7945 → 0.9220). 이는 최종 출력층의 압축 과정에서 미세 아티팩트 정보가 손실되어 특징 공간의 분리도가 저하되기 때문이다.

반면, 제안 모델은 인코더의 중간 계층에서 고차원 특징을 직접 후킹함으로써 정보 손실을 최소화하고, 엔진별 특징 클러스터를 보다 명확하게 분리하여 높은 식별 성능을 달성하였다.

2.2.3 미세 화질 열화 환경 하에서의 생성형 AI 비디오 엔진 식별 강인성(Robustness) 평가

제안 모델의 강인성을 검증하기 위해 Gaussian blur, Gaussian noise, Resolution downscaling의 세 가지 화질 열화 조건을 적용하고, 각 실험군의 식별 정확도를 비교하였다. 결과는 [표 7]에 제시되어 있다.

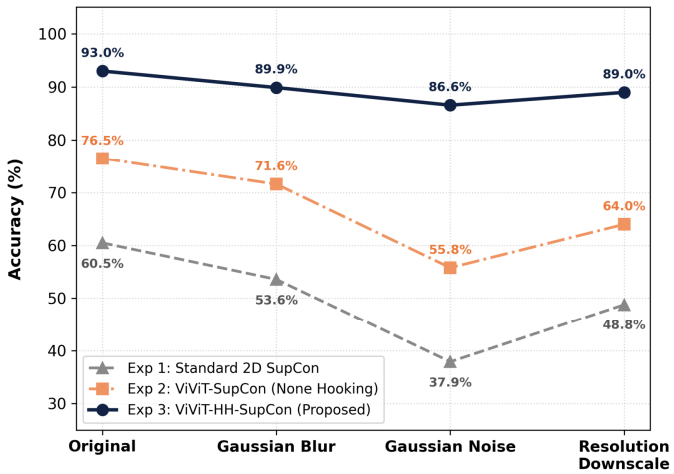

그림 4. 다양한 화질 저하 환경 하에서의 성능 비교 그래프
Fig. 4. Performance comparison under mild degradation conditions

실험 결과, 제안 모델(Exp 3)은 모든 열화 조건에서 높은 정확도를 유지하며 안정적인 성능을 보였다(93.0% → 89.0% 수준). 반면, 2D 기반 모델과 후킹을 적용하지 않은 모델(Exp 1, Exp 2)은 열화 조건에서 성능 저하가 크게 나타났다.

이는 제안 모델이 인코더의 중간 계층에서 고차원 특징을 직접 활용함으로써, 외부 노이즈 및 화질 변형에도 불구하고 엔진 고유의 시공간적 특성을 효과적으로 보존하기 때문이다. 각 모델의 성능 변화 경향은 [그림 4]에서 확인할 수 있다.

2.2.4 t-SNE 시각화를 통한 생성형 AI 비디오 엔진별 분리도

제안 모델의 특징 분리 성능은 고차원 특징 공간을 2차원으로 투영한 t-SNE 결과([그림 5])를 통해 확인할 수 있다.

(a) Standard 2D SupCon은 공간 특징만을 사용함에 따라 클래스 간 경계가 분리되지 않고 중앙 영역에서 중첩된 분포를 보인다. (b) ViViT-SupCon(None Hooking)은 시공간 정보를 반영하여 일부 군집화 경향을 보이지만, 엔진 간 경계가 명확하게 분리되지 않는다. 반면, (c) 제안 모델은 중간 계층 특징을 활용함으로써 각 엔진별 클러스터가 명확하게 분리된 구조를 형성한다.

V. 기대 효과

1. 학술적·기술적 기대 효과

본 연구는 시공간-주파수 통합 기반의 ViViT-HH-SupCon 구조를 통해 생성형 AI 비디오의 출처 식별 문제에 대한 새로운 아키텍처적 접근을 제시한다. 동일 입력 조건 하에서의 통제 실험을 통해, 성능 향상이 단순한 차원 확장이 아닌 중간 계층 후킹 메커니즘에 기인함을 정량적으로 확인하였다. 또한 제안 모델은 엔진별 특징을 분리된 클러스터로 형성함으로써, 향후 멀티 클래스 비디오 포렌식 모델 설계 및 평가를 위한 기준으로 활용될 수 있다.

2. 실무적·산업적 기대 효과

제안된 파이프라인은 시공간 특징을 효율적으로 활용하는 구조로, 다양한 미디어 처리 환경에서 생성형 AI 영상의 출처 식별에 적용될 수 있다. 특히 화질 열화 및 유통 환경의 변형 조건에서도 안정적인 성능을 유지함으로써, 실제 미디어 서비스 환경에서의 활용 가능성을 확인하였다.

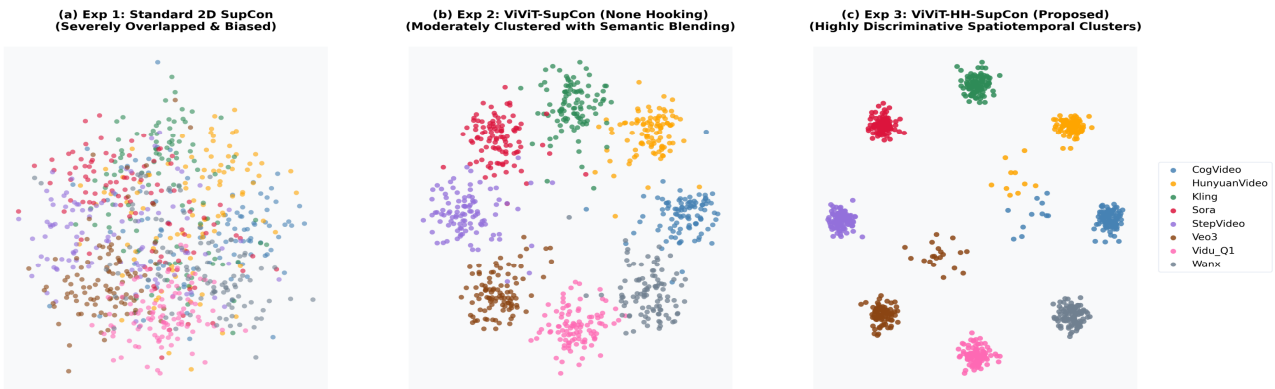

그림 5. 실험군별 고차원 특징 공간 t-SNE 시각화 기반의 생성형 AI 비디오 엔진별 분리도 대조도: (a) 기존 2D 모델, (b) 후킹 메커니즘 배제 모델, (c) 제안 모델

Fig. 5. Comparison of Engine-Specific Separation Based on High-Dimensional Feature Space t-SNE Visualization: (a) Standard 2D SupCon, (b) ViViT-SupCon (None Hooking), (c) Proposed ViViT-HH-SupCon

Ⅵ. 결론 및 향후 과제

1. 결론 (Conclusion)

본 논문은 생성형 AI 비디오의 출처 식별을 위해 시공간-주파수 통합 입력과 지도 대조 학습 기반의 ViViT-HH-SupCon 아키텍처를 제안하였다. 통제된 실험을 통해 성능 향상이 입력 데이터가 아닌 모델 구조, 특히 중간 계층 후킹 메커니즘에 기인함을 확인하였다. 제안 모델은 Macro F1-score 0.9220 및 정확도 93.00%를 달성하며 기존 2D 기반 모델 대비 각각 25.83%p 및 32.50%p의 성능 향상을 보였다. 또한 일부 엔진(Veo3, Vidu_Q1)에서 나타나던 성능 편차를 완화하고, 고차원 특징 공간에서 엔진별 분리도를 향상시킴을 확인하였다.

2. 향후 과제 (Future Work)

향후 과제에서는 시공간 연속성 학습을 강화하기 위한 Tubelet embedding의 적용과, 실제 유통 환경에서의 화질 열화 및 변형 조건에 대한 강인성 검증을 확장할 예정이다. 또한 미학습 생성형 AI 엔진에 대응하기 위한 오픈셋 인식(Open-Set Recognition) 기법을 결합하여, 다중 출처 식별의 적용 범위를 확장하고자 한다.

참 고 문 헌 (References)

[1] S. Bergmann, D. Moussa, F. Brand, A. Kaup, and C. Riess, “Forensic Analysis of AI-Compression Traces in Spatial and Frequency Domain,” Pattern Recognition Letters, vol. 180, pp. 41-47, Apr, 2024.
doi: https://doi.org/10.1016/j.patrec.2024.02.015

[2] D. Cozzolino and L. Verdoliva, “Noiseprint: A CNN-Based Camera Model Fingerprint,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 144-159, May, 2019.
doi: https://doi.org/10.1109/TIFS.2019.2916364

[3] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On The Detection of Synthetic Images Generated by Diffusion Models,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2023.
doi: https://doi.org/10.1109/ICASSP49357.2023.10095167

[4] F. Khan, A. Sareen, A. S. Kumar, and M. Bhuvaneswari, “A Quantum-Hybrid Framework for Enhanced Deepfake Detection: Integrating QAOA-Based Feature Selection With Quantum-Inspired Attention Mechanisms,” IEEE Access, vol. 14, pp. 17853 - 17873, Jan, 2026.
doi: https://doi.org/10.1109/ACCESS.2026.3659021

[5] E. H. Shah, U. Dubey, A. S. A. Rahiman, and N. P. Shetty, “Multi-Layer Knowledge Distillation With Custom Temperature Scaling for Deepfake Detection,” IEEE Access, vol. 14, pp. 46763-46787, Mar. 2026.
doi: https://doi.org/10.1109/ACCESS.2026.3674500

[6] T. Yang, Z. Huang, J. Cao, L. Li, and X. Li, “Deepfake Network Architecture Attribution,” in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 36, no. 4, pp. 4662‑4670, Jun. 2022.
doi: https://doi.org/10.1609/aaai.v36i4.20391

[7] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “ViViT: A Video Vision Transformer,” Proceedings of the IEEE/CVF International Conference on Computer Vision, Oct, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00676

[8] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised Contrastive Learning,” Advances in Neural Information Processing Systems, vol. 34, pp. 18661-18673, Dec, 2020.
doi: https://doi.org/10.48550/arXiv.2004.11362

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” Proceedings of the 37th International Conference on Machine Learning, pp. 1597-1607, Jul, 2020.
doi: https://doi.org/10.48550/arXiv.2002.05709

[10] R. Kundu, V. Mohanty, H. Xiong, S. Jia, A. Balachandran, and A. K. Roy-Chowdhury, “SAGA: Source Attribution of Generative AI Videos,” arXiv preprint arXiv:2511.12834, 2025.
doi: https://doi.org/10.48550/arXiv.2511.12834

[11] P. He, L. Zhu, J. Li, S. Wang, and H. Li, “Exposing AI-generated Videos: A Benchmark Dataset and a Local-and-Global Temporal Defect Based Detection Method,” arXiv, May 2024.
doi: https://doi.org/10.48550/arXiv.2405.04133

[12] M. Cassia, L. Guarnera, M. Casu, I. Zangara, and S. Battiato, “Deepfake Forensic Analysis: Source Dataset Attribution and Legal Implications of Synthetic Media Manipulation,” in Proceedings of the IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), Oct. 2025.
doi: https://doi.org/10.1109/MetroXRAINE66377.2025.11340536

[13] S. Baxavanakis, M. Schinas, and S. Papadopoulos, “Do DeepFake Attribution Models Generalize?,” in Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation, pp. 45-54, July, 2025.
doi: https://doi.org/10.1145/3733567.3735572

[14] D. Zheng, Z. Huang, H. Liu, K. Zou, Y. He, F. Zhang, Y. Zhang, J. He, W. S. Zheng, Y. Qiao, and Z. Liu, “VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness,” arXiv preprint arXiv:2503.21755, 2025.
doi: https://doi.org/10.48550/arXiv.2503.21755

저 자 소 개

강 자 원

- 1999년 ~ 2004년 : 아주대학교 미디어학부 학사
- 2025년 ~ 현재 : 서강대학교 가상융합전문대학원 테크놀로지전공 석사과정
- ORCID : https://orcid.org/0009-0008-4715-6930
- 주관심분야 : 가상융합(XR), 미디어 콘텐츠(Media Contents), 정보보호 등

도 경 화

- 2004년 2월 : 숭실대학교 공학박사
- 2018년 ~ 2024년 8월 : 건국대, 고려대 교수
- 2004년 2월 ~ 2017년 12월 : 행정안전부 신기술 팀장
- 현재 : 서강대학교 웹3.0기술연구센터 재직
- 전자서명인정위원회, ISMS심의위원회 심의위원
- 차세대인증연구회 부회장, 한국정보처리학회 산학연본부장, 한국여성정보인협회 이사 역임 중
- 인공지능 기술융합혁신상(2026.7.2.), 대통령표창(정보보호 유공, 2016.7.13.)
- ORCID : https://orcid.org/0009-0004-5179-5955
- 주관심분야 : 인공지능, 데이터신뢰성, 블록체인, 딥페이크, 클라우드, 웹3.0기술, 정보보호 및 개인정보보호, 전자정부, 공공서비스기술정책

박 수 용

- 1995년 : 조지메이슨대학교에서 박사학위
- 2012년 9월 ~ 2014년 11월 : 정보통신산업진흥원(NIPA) 원장 역임
- 2018년 12월 : 아시아·태평양 소프트웨어공학 학술대회(APSEC)에서 10년 최우수 영향력 논문상 (Ten-Year Most Influential Paper Award)을 수상
- 2019년 1월 ~ 현재 : 한국블록체인학회 회장과 지능형 블록체인 연구센터장
- 2021년 11월 : 우수 과제 평가 성과를 인정받아 과학기술정보통신부 장관상을 수상
- 현재 : 서강대학교 컴퓨터공학과 재직
- 주관심분야 : 블록체인, 인공지능, 데이터신뢰성, 웹3.0기술