TruthLens v4.4.1 사용자 가이드

AI 기반 딥페이크 탐지 플랫폼 운영 매뉴얼

> *문서 버전: Rev 4.1*
>
> *대상 시스템: TruthLens v4.4.1*
>
> *작성일: 2026 년 4 월 19 일*
>
> *Creator: Brian Lee | Test Research: Younghyun Han | AI R&D Center | A3 Security Co.,Ltd.*

머리말 — 왜 딥페이크 탐지가 필요한가

2024 년 이후 전 세계적으로 딥페이크(Deepfake) 기술을 악용한 사기 피해가 급증하고 있습니다. 2025 년 대한민국에서만 딥페이크 관련 사이버 범죄 신고가 전년 대비 340% 증가하였으며, 금융기관 대상 보이스피싱에 AI 음성 합성이 활용된 사례, 공직자를 사칭한 딥페이크 영상으로 인한 여론 조작 사건, 그리고 기업 임원의 화상 회의 영상을 위조하여 수십억 원의 송금을 유도한 BEC(Business Email Compromise) 사건이 연이어 보도되었습니다.

이러한 사회적 위협에 대응하기 위하여, **TruthLens** 는 40 개 독립 탐지 모듈, 8 개 멀티 에이전트 AI 프레임워크, OWL 온톨로지 기반 파이프라인 검증 시스템을 통합한 **차세대 멀티모달 딥페이크 탐지 플랫폼**으로 개발되었습니다. 본 플랫폼은 한국 AI 기본법 제 15 조의 "AI 판단 근거 설명 의무"를 준수하며, 모든 판정 결과에 대하여 설명 가능한 AI(XAI) 시각화와 SHA-256 해시 체인 기반 감사 로그를 제공합니다.

TruthLens 는 정부기관, 금융기관, 방송사, 수사기관, 그리고 기업의 정보보안 담당자가 **별도의 AI 전문 지식 없이도** 의심 미디어를 신속하게 분석하고 과학적 근거에 기반한 판정을 내릴 수 있도록 설계되었습니다. 본 가이드는 플랫폼의 모든 메뉴와 기능을 체계적으로 안내하여, 담당자께서 현업에서 즉시 활용하실 수 있도록 구성하였습니다.

목차

1. [시스템 개요](#1-시스템-개요)
2. [로그인 및 초기 화면](#2-로그인-및-초기-화면)
3. [탐지 섹션](#3-탐지-섹션)
   - 3.1 [분석 (Analyze)](#31-분석-analyze)
   - 3.2 [분석 결과 (Results)](#32-분석-결과-results)

- 3.3 [분석 이력 (History)](#33-분석-이력-history)
- 3.4 [BI 대시보드](#34-bi-대시보드)

4. [공간/주파수 탐지 섹션](#4-공간주파수-탐지-섹션)
- 4.1 [DM 탐지 센터](#41-dm-탐지-센터)
- 4.2 [ViT 분석](#42-vit-분석)
- 4.3 [블렌딩 경계 분석](#43-블렌딩-경계-분석)
- 4.4 [주파수 분석](#44-주파수-분석)
- 4.5 [Foundation 모델](#45-foundation-모델)

5. [시간/생체 분석 섹션](#5-시간생체-분석-섹션)
- 5.1 [시간 도메인 분석](#51-시간-도메인-분석)
- 5.2 [립싱크 포렌식](#52-립싱크-포렌식)
- 5.3 [미세 표정 분석](#53-미세-표정-분석)
- 5.4 [양안(시선) 분석](#54-양안시선-분석)
- 5.5 [머리 포즈 역학](#55-머리-포즈-역학)

6. [오디오 포렌식 섹션](#6-오디오-포렌식-섹션)
- 6.1 [오디오 포렌식](#61-오디오-포렌식)
- 6.2 [SSL 음성 탐지](#62-ssl-음성-탐지)
- 6.3 [보코더 식별](#63-보코더-식별)

7. [포렌식 도구 섹션](#7-포렌식-도구-섹션)
- 7.1 [포렌식 보고서](#71-포렌식-보고서)
- 7.2 [핑거프린트 DB](#72-핑거프린트-db)
- 7.3 [위협 인텔리전스](#73-위협-인텔리전스)
- 7.4 [XAI 심층 분석](#74-xai-심층-분석)
- 7.5 [모델 핑거프린터](#75-모델-핑거프린터)
- 7.6 [T-GD 출처 분석](#76-t-gd-출처-분석)

8. [방어 및 분류 섹션](#8-방어-및-분류-섹션)
- 8.1 [3-Class 분류](#81-3-class-분류)
- 8.2 [사전 방어](#82-사전-방어)

9. [실시간 및 인프라 섹션](#9-실시간-및-인프라-섹션)
- 9.1 [실시간 모니터](#91-실시간-모니터)

- 9.2 [모델 증류](#92-모델-증류)

10. [에이전트 AI 섹션](#10-에이전트-ai-섹션)

- 10.1 [멀티에이전트 프레임워크](#101-멀티에이전트-프레임워크)
- 10.2 [에이전트 설정](#102-에이전트-설정)
- 10.3 [AutoGen 토론](#103-autogen-토론)
- 10.4 [LLM / 추론 엔진](#104-llm--추론-엔진)
- 10.5 [ADAG 레드팀](#105-adag-레드팀)
- 10.6 [MC Fusion 시뮬레이션](#106-mc-fusion-시뮬레이션)
- 10.7 [온톨로지 파이프라인](#107-온톨로지-파이프라인)

11. [관리 섹션](#11-관리-섹션)

12. [시스템 섹션](#12-시스템-섹션)

13. [판정 기준 및 해석 가이드](#13-판정-기준-및-해석-가이드)

14. [부록: 딥페이크 위협과 TruthLens 의 사회적 책무](#14-부록-딥페이크-위협과-truthlens-의-사회적-책무)

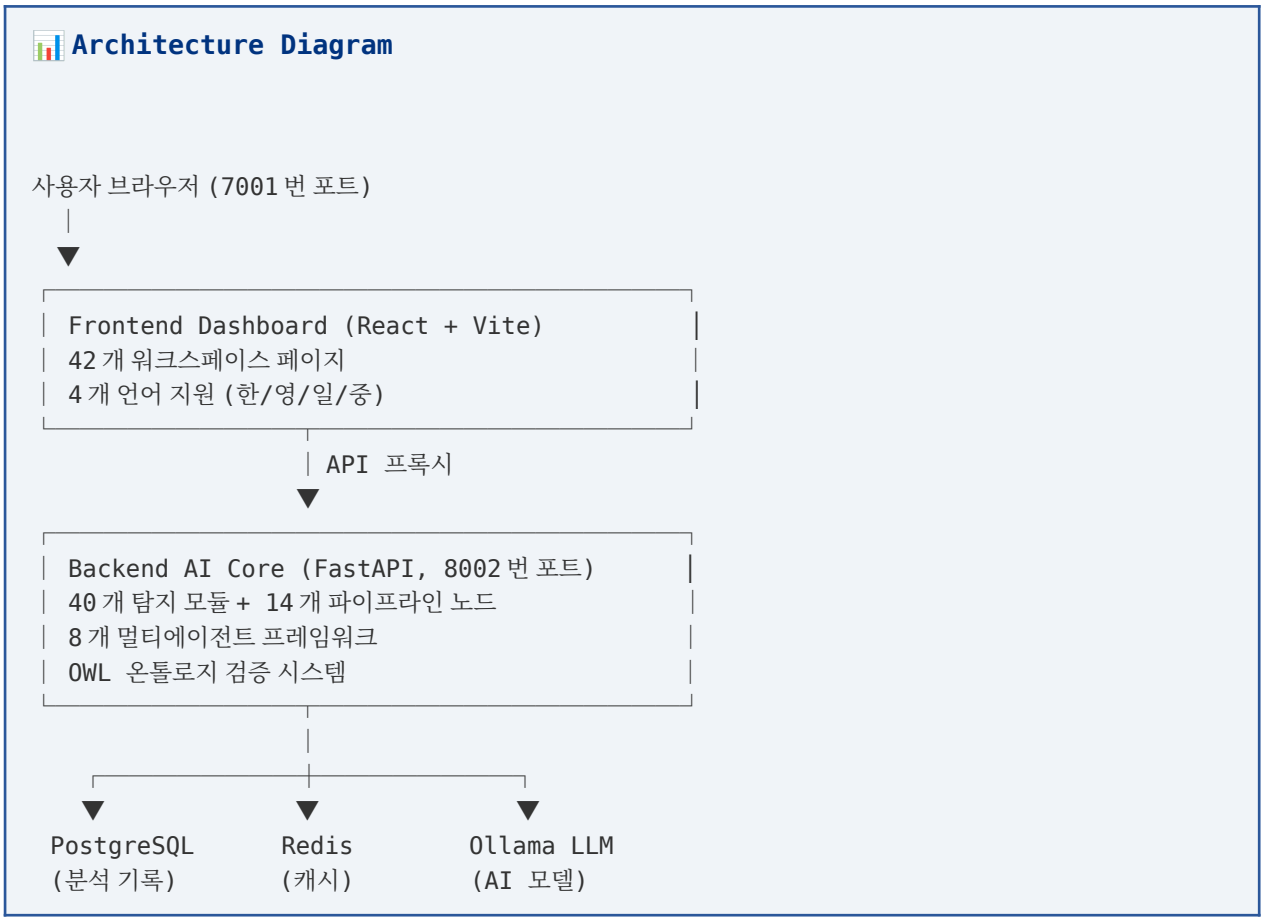
---

1. 시스템 개요

1.1 TruthLens 란

TruthLens 는 영상, 음성, 이미지, 문서 등 다양한 형식의 미디어 파일에 대하여 **AI 생성 또는 조작 여부를 자동으로 판별**하는 딥페이크 탐지 플랫폼입니다. 40 개의 독립적인 탐지 모듈이 각기 다른 관점에서 미디어를 분석하고, 그 결과를 멀티 에이전트 AI 시스템이 종합하여 최종 판정을 내립니다.

1.2 시스템 구성



1.3 판정 결과의 의미

TruthLens 는 모든 분석에 대하여 세 가지 판정 중 하나를 제공합니다.

판정	의미	배지 색상
REAL (진본)	분석 결과 AI 생성 또는 조작의 흔적이 발견되지 않았음	초록색
FAKE (위조)	AI 생성 또는 조작의 흔적이 유의미하게 발견됨	빨간색

UNCERTAIN (판단 유보)	판단을 내리기에 증거가 불충분하거나 모호함	노란색

각 판정에는 **신뢰도(Confidence)** 점수가 함께 제공됩니다. 0%에서 100% 사이의 값으로, 높을수록 판정에 대한 시스템의 확신이 강합니다.

> *주의사항: UNCERTAIN 판정은 "분석 실패"가 아닙니다. 시스템이 결론을 내리기에 근거가 불충분하다는 의미이며, 이 경우 전문가의 추가 검토를 권고합니다.*

2. 로그인 및 초기 화면

2.1 로그인

웹 브라우저에서 `http://서버주소:7001`에 접속하면 로그인 화면이 표시됩니다. 관리자로부터 발급받은 이메일과 비밀번호를 입력하여 로그인합니다.

2.2 사용자 역할

역할	권한
VIEWER (열람자)	분석 실행 및 결과 조회
ANALYST (분석가)	열람자 권한 + API 키 관리 + 보고서 생성
ADMIN (관리자)	분석가 권한 + 사용자/조직 관리 + 시스템 설정

2.3 화면 구성

로그인 후 화면은 다음과 같이 구성됩니다.

- **좌측 사이드바**: 10 개 섹션으로 분류된 메뉴 항목. 각 섹션은 펼침/접힘이 가능합니다.
- **상단 헤더**: 현재 페이지명, 언어 선택, 사용자 정보
- **중앙 콘텐츠**: 선택한 메뉴에 해당하는 워크스페이스

---

3. 탐지 섹션

이 섹션은 TruthLens 의 핵심 기능으로, 미디어 파일을 업로드하여 분석하고 결과를 확인하는 워크스페이스입니다.

3.1 분석 (Analyze)

메뉴 경로: 사이드바 > 분석

경로: `/`

이 워크스페이스는 딥페이크 탐지의 시작점입니다. 다음 두 가지 방식으로 미디어를 분석할 수 있습니다.

파일 업로드 방식

1. "파일 선택" 버튼을 클릭하여 분석할 영상, 음성, 또는 이미지 파일을 선택합니다.
2. 지원 형식: MP4, AVI, MOV, MKV, WebM (영상) / MP3, WAV, FLAC (음성) / JPG, PNG, WebP (이미지)
3. 최대 파일 크기: 500MB

URL 입력 방식

1. YouTube, Instagram, TikTok 등 소셜 미디어 URL 을 입력합니다.
2. 시스템이 자동으로 미디어를 다운로드하여 분석합니다.
3. 1,000 개 이상의 사이트를 지원합니다.

분석 모드 선택

모드	소요 시간	정확도	권장 상황
신속 (Rapid)	10~30 초	보통	대량 스크리닝, 긴급 확인
표준 (Standard)	1~5 분	높음	일반적인 분석 (권장)
정밀 (Precise)	5~15 분	최고	수사 증거, 법적 활용

분석이 시작되면 진행 상황이 실시간으로 표시됩니다.

용어 해설

용어	설명
딥페이크(Deepfake)	딥러닝 기술로 사람의 얼굴, 목소리, 동작 등을 합성하거나 변조한 미디어
모달리티(Modality)	분석 대상의 유형. 시각(영상), 청각(음성), 생체신호 등 각기 다른 분석 채널

신뢰도(Confidence)	시스템이 판정에 대해 갖는 확신의 정도 (0~100%)

3.2 분석 결과 (Results)

메뉴 경로: 분석 완료 후 자동 이동

경로: `/results/:jobId`

분석이 완료되면 이 워크스페이스에 종합 결과가 표시됩니다.

판정 배지: 화면 상단에 REAL(초록) / FAKE(빨강) / UNCERTAIN(노란) 배지와 신뢰도가 크게 표시됩니다.

모달리티별 점수 (레이더 차트)

방사형 차트(레이더 차트)로 각 모달리티의 분석 결과가 표시됩니다.

- **시각 분석(Visual)**: 영상 프레임의 얼굴 조작 흔적 분석 결과
- **오디오 분석(Audio)**: 음성 합성 또는 변조 흔적 분석 결과
- **생체신호 분석(Biological)**: 심박수, 눈 깜빡임 등 생체 반응의 자연스러움
- **A/V 동기화(A/V Sync)**: 입 모양과 음성의 일치 여부

그래프 판독법: 레이더 차트에서 각 축의 값이 0.5(50%)를 초과하면 해당 모달리티에서 위조 의심 징후가 발견된 것입니다. 값이 높을수록 위조 가능성이 높습니다. 모든 축이 0.3 이하이면 진본일 가능성이 높습니다.

증거 체인(Evidence Chain): 판정 근거가 된 구체적인 증거 항목이 심각도(HIGH/MEDIUM/LOW)와 함께 나열됩니다.

3.3 분석 이력 (History)

메뉴 경로: 사이드바 > 분석 이력

경로: `/history`

지금까지 수행한 모든 분석의 이력을 시간순으로 조회할 수 있습니다.

- **파일명**: 분석한 미디어 파일의 원본 이름
- **판정**: REAL / FAKE / UNCERTAIN
- **신뢰도**: 0~100%
- **분석 시각**: 분석이 완료된 일시
- **소요 시간**: 분석에 걸린 시간 (밀리초)

검색 기능을 통해 파일명 또는 판정 유형으로 이력을 필터링할 수 있습니다.

3.4 BI 대시보드

메뉴 경로: 사이드바 > BI 대시보드

경로: `/bi`

비즈니스 인텔리전스(BI) 대시보드는 분석 통계를 시각적으로 요약합니다.

- **판정 분포 파이차트**: REAL/FAKE/UNCERTAIN 의 비율
- **일별 분석 추이**: 날짜별 분석 건수 추이 그래프
- **평균 신뢰도**: 기간별 평균 신뢰도 변화

그래프 판독법: 파이차트에서 FAKE 비율이 일정 기간 급증하면 해당 시기에 딥페이크 유입이 활발했음을 의미합니다. 일별 추이에서 특정 날짜에 분석 건수가 급증하면 이벤트(보도, 선거 등)와의 연관성을 확인할 필요가 있습니다.

4. 공간/주파수 탐지 섹션

이 섹션은 영상의 **시각적 특성**을 분석하는 전문 워크스페이스입니다. 각 도구는 서로 다른 기법으로 영상의 조작 흔적을 탐지합니다.

4.1 DM 탐지 센터

메뉴 경로: 사이드바 > 공간/주파수 탐지 > DM 탐지 센터

경로: `/dm-detection`

기능: Stable Diffusion, Midjourney, DALL-E 등 **확산 모델(Diffusion Model)**로 생성된 이미지를 전문적으로 탐지합니다. 3 단계 파이프라인(스크리닝 → 고속 → 정밀)으로 점진적으로 분석 깊이를 높입니다.

화면 구성:

- **탐지 결과 카드**: 각 분석 건에 대한 판정, 신뢰도, 소요 시간
- **DM 확률 점수**: 확산 모델로 생성되었을 확률 (0~1)
- **스테이지별 결과**: 스크리닝/고속/정밀 각 단계의 점수

그래프 판독법: DM 확률이 0.7 이상이면 확산 모델 생성 의심도가 높습니다. 스테이지별 점수가 단계를 거칠수록 상승하면 의심이 강화되는 것이며, 하락하면 초기 오탐의 가능성이 있습니다.

용어 해설

용어	설명

확산 모델(Diffusion Model)	이미지를 점진적으로 생성하는 AI 기법. Stable Diffusion, DALL-E 등이 대표적임
스크리닝(Screening)	대량의 이미지를 빠르게 1 차 필터링하는 과정
DIRE	Diffusion Reconstruction Error — 확산 모델이 남기는 고유한 오차 패턴을 탐지하는 기법

4.2 ViT 분석

메뉴 경로: 사이드바 > 공간/주파수 탐지 > ViT 분석

경로: `/vit-analysis`

기능: **비전 트랜스포머(Vision Transformer, ViT)** 모델을 활용하여 이미지의 조작 흔적을 탐지합니다. 여러 ViT 변형 모델(Swin-V2, CrossViT, TIMM)의 결과를 앙상블하여 정확도를 높입니다.

화면 구성:

- **모델별 점수**: Swin-V2 Score, CrossViT Score, TIMM Score, Ensemble Score
- **어텐션 히트맵**: 모델이 주목한 영역을 색상으로 표시한 지도 (파란색=정상, 빨간색=의심)
- **조작 영역**: 감지된 조작 의심 영역의 위치와 이상 점수

그래프 판독법: 어텐션 히트맵에서 **빨간색이 얼굴 경계부, 머리카락-이마 경계, 턱선 주변에 집중**되어 있으면 페이스스왑(얼굴 교체)의 전형적인 패턴입니다. 히트맵이 전체적으로 고르게 파란색이면 정상 이미지일 가능성이 높습니다.

용어 해설

용어	설명
비전 트랜스포머(ViT)	이미지를 작은 조각(패치)으로 나누어 문맥을 파악하는 AI 모델
어텐션 히트맵(Attention Heatmap)	AI 모델이 이미지의 어느 부분에 주목하는지를 색상으로 시각화한 지도
앙상블(Ensemble)	여러 모델의 결과를 종합하여 하나의 판단을 내리는 기법

4.3 블렌딩 경계 분석

메뉴 경로: 사이드바 > 공간/주파수 탐지 > 블렌딩 경계

경로: `/blending-analysis`

기능: 얼굴을 합성할 때 생기는 **경계면의 부자연스러움**을 탐지합니다. 딥페이크는 원본 얼굴 위에 다른 얼굴을 덮어씌울 때 경계부에 미세한 흔적을 남기며, 이 도구는 그 흔적을 찾습니다.

화면 구성:

- **LAA 점수**: Learned Adaptive Attention — 학습된 적응적 주의력 점수
- **Poisson 점수**: 푸아송 블렌딩 흔적 점수
- **색상 불일치**: 합성된 얼굴과 원본 배경 사이의 색감 차이
- **조명 불일치**: 얼굴과 배경의 조명 방향 차이

그래프 판독법: LAA 점수와 Poisson 점수가 동시에 0.6 이상이면 합성 경계가 존재할 가능성이 높습니다. 색상 불일치와 조명 불일치가 높으면 조잡한 합성이며, 둘 다 낮지만 LAA 가 높으면 정교한 딥페이크를 의미합니다.

용어 해설

용어	설명
블렌딩(Blending)	두 이미지를 자연스럽게 합치는 기법. 딥페이크에서 얼굴을 교체할 때 사용됨
LAA	경계면의 이상 패턴을 학습한 AI 가 감지한 의심도
푸아송 블렌딩(Poisson Blending)	이미지 합성 시 경계를 매끄럽게 처리하는 수학적 기법. 그 흔적이 남을 수 있음

4.4 주파수 분석

메뉴 경로: 사이드바 > 공간/주파수 탐지 > 주파수 분석

경로: `/frequency-analysis`

기능: 이미지를 **주파수 도메인**으로 변환하여 AI 생성 이미지 특유의 패턴을 탐지합니다. 인간의 눈에는 보이지 않지만, 주파수 영역에서는 AI 가 남긴 고유한 "지문"이 드러납니다.

화면 구성:

- **FFT 점수**: 고속 푸리에 변환 기반 이상 점수
- **DCT 점수**: 이산 코사인 변환 기반 이상 점수

- **Wavelet 점수**: 웨이블릿 변환 기반 이상 점수
- **GAN 핑거프린트 점수**: GAN(생성적 적대 신경망) 특유의 주파수 패턴 점수
- **스펙트럼 이상치**: 주파수 스펙트럼에서 발견된 비정상 패턴

그래프 판독법: GAN 핑거프린트 점수가 0.7 이상이면 GAN 으로 생성된 이미지일 가능성이 높습니다. FFT 스펙트럼에서 **격자형 패턴(checkerboard artifact)**이 보이면 GAN 의 전형적인 흔적입니다. 세 가지 변환(FFT, DCT, Wavelet) 점수가 모두 높으면 신뢰도가 강화됩니다.

용어 해설

용어	설명
주파수 도메인(Frequency Domain)	이미지를 밝기의 변화 패턴(주파수)으로 분해한 표현. 고주파는 세부 디테일, 저주파는 전체 윤곽
FFT(Fast Fourier Transform)	이미지를 주파수 성분으로 분해하는 수학적 변환
GAN(Generative Adversarial Network)	생성자와 판별자가 경쟁하며 이미지를 생성하는 AI 기법
핑거프린트(Fingerprint)	AI 모델이 생성물에 남기는 고유한 패턴. 사람의 지문처럼 모델을 식별할 수 있음

4.5 Foundation 모델

메뉴 경로: 사이드바 > 공간/주파수 탐지 > Foundation 모델

경로: `/foundation-models`

기능: 최신 대규모 사전학습 모델(Foundation Model)을 활용한 탐지 결과를 제공합니다.

화면 구성:

- **UniFD 점수**: 범용 위조 탐지 모델의 판정 점수
- **LNCLIP-DF 점수**: CLIP 기반 딥페이크 특화 모델의 점수
- **DINOv2 점수**: Meta 의 자기지도 학습 비전 모델의 점수
- **앙상블 점수**: 세 모델을 종합한 최종 점수

그래프 판독법: 세 모델의 점수가 모두 0.6 이상이면 위조 가능성이 높습니다. 모델 간 점수 차이가 크면(예: UniFD=0.8, DINOv2=0.2) 특정 유형의 위조에만 반응하는 것이므로 추가 분석이 필요합니다.

용어 해설

용어	설명

파운데이션 모델(Foundation Model)	대규모 데이터로 사전 학습된 범용 AI 모델
CLIP	OpenAI 가 개발한 이미지-텍스트 연결 모델
자기지도 학습(Self-Supervised Learning)	사람이 정답을 제공하지 않아도 데이터의 구조를 스스로 학습하는 방법

5. 시간/생체 분석 섹션

이 섹션은 영상의 **시간적 흐름**과 인물의 **생체 신호**를 분석합니다. 딥페이크는 개별 프레임에서는 완벽해 보일 수 있지만, 시간에 따른 변화 패턴에서 부자연스러움이 드러나는 경우가 많습니다.

5.1 시간 도메인 분석

메뉴 경로: 사이드바 > 시간/생체 분석 > 시간 도메인 분석

경로: `/temporal-analysis`

기능: 프레임 간 **시간적 일관성**을 분석합니다. 딥페이크 영상은 프레임마다 독립적으로 얼굴을 생성하기 때문에, 연속된 프레임 사이에 미세한 떨림이나 불연속이 발생합니다.

화면 구성:

- **시간적 일관성 점수**: 프레임 간 변화의 자연스러움 (높을수록 의심)
- **프레임별 점수 그래프**: 각 프레임의 위조 의심도를 시간축으로 표시
- **감지된 생성기**: 추정되는 딥페이크 생성 도구 (해당 시)

그래프 판독법: 프레임별 점수 그래프에서 **갑작스러운 급등 구간**은 해당 시점에서 얼굴 합성이 불안정하게 이루어진 것을 의미합니다. 점수가 전체적으로 0.3 이하로 안정적이면 진본일 가능성이 높습니다.

용어 해설

용어	설명
시간적 일관성(Temporal Consistency)	연속된 프레임 간의 자연스러운 변화 유지 여부
VideoMAE	영상의 시간적 패턴을 학습하는 자기지도 학습 비전 트랜스포머
TALL	Temporal Attention Learned Locally — 지역적 시

	간 주의력 기법

5.2 립싱크 포렌식

메뉴 경로: 사이드바 > 시간/생체 분석 > 립싱크 포렌식

경로: `/lip-forensics`

기능: 입술 움직임과 음성의 **동기화 정확도**를 분석합니다. 딥페이크로 음성을 변조하거나 다른 사람의 음성을 입힌 경우, 입 모양과 발음 사이에 미세한 불일치가 발생합니다.

화면 구성:

- **동기화 편차 그래프**: 시간축에 따른 입술-음성 동기화 차이
- **동기화 점수**: 전체 평균 동기화 품질 (낮을수록 의심)
- **감지된 립싱크 도구**: Wav2Lip, VideoRetalking 등 추정 도구

그래프 판독법: 동기화 편차 그래프에서 지속적으로 50ms 이상의 편차가 나타나면 음성 변조의 강력한 징후입니다. 특정 구간에서만 편차가 급증하면 해당 부분만 편집되었을 가능성이 있습니다.

용어 해설

용어	설명
립싱크(Lip-Sync)	입술 움직임과 음성의 동기화
Wav2Lip	음성에 맞춰 입술 움직임을 합성하는 딥러닝 모델
동기화 편차(Sync Offset)	입 모양과 실제 발음 사이의 시간 차이 (밀리초 단위)

5.3 미세 표정 분석

메뉴 경로: 사이드바 > 시간/생체 분석 > 미세 표정 분석

경로: `/micro-expression`

기능: 0.04~0.5 초 사이에 나타나는 **미세 표정(Micro-Expression)**의 자연스러움을 분석합니다. 딥페이크는 이러한 극히 짧은 표정 변화를 정확하게 재현하지 못하는 경우가 많습니다.

화면 구성:

- **감지된 미세 표정 수**: 분석 구간에서 발견된 미세 표정의 횟수

- **표정별 자연스러움 점수**: 각 감지된 표정의 자연성 평가
- **이상 구간**: 부자연스러운 표정 전환이 발생한 시간대

그래프 판독법: 진본 영상에서는 일반적으로 분당 2~5 회의 미세 표정이 감지됩니다. 미세 표정이 전혀 감지되지 않거나 과도하게 많으면 의심스럽습니다. 자연스러움 점수가 0.5 미만인 표정이 있으면 해당 구간을 상세히 확인해야 합니다.

용어 해설

용어	설명
미세 표정(Micro-Expression)	무의식적으로 매우 짧은 시간 동안 나타나는 얼굴 표정. 의식적으로 통제하기 어려움
AU(Action Unit)	얼굴 근육의 개별 움직임 단위. FACS(Facial Action Coding System)에서 정의

5.4 양안(시선) 분석

메뉴 경로: 사이드바 > 시간/생체 분석 > 양안 분석

경로: `/gaze-analysis`

기능: 양쪽 눈의 **시선 방향 일관성**을 분석합니다. 실제 사람은 양안이 동일한 지점을 주시하지만, 딥페이크에서는 양눈의 시선이 미세하게 어긋나는 경우가 있습니다.

화면 구성:

- **시선 추적 궤적**: 양쪽 눈의 시선 방향을 시간축으로 표시
- **양안 편차**: 왼쪽 눈과 오른쪽 눈의 시선 차이
- **시선 일관성 점수**: 전체 평균 일관성 (높을수록 정상)

그래프 판독법: 양안 편차가 지속적으로 5 도 이상이면 시선 동기화에 문제가 있는 것이며, 딥페이크의 강력한 징후입니다. 시선 궤적 그래프에서 양쪽 눈의 선이 자주 교차하거나 급격하게 벌어지면 의심스럽습니다.

용어 해설

용어	설명
양안 시선(Binocular Gaze)	양쪽 눈이 동시에 같은 지점을 향하는 자연스러운 움직임
시선 편차(Gaze Deviation)	양쪽 눈의 시선 방향 차이 (각도 단위)

5.5 머리 포즈 역학

메뉴 경로: 사이드바 > 시간/생체 분석 > 머리 포즈 역학

경로: `/head-pose`

기능: 머리의 회전(Yaw), 기울임(Pitch), 좌우 기울기(Roll)의 **역학적 자연스러움**을 분석합니다. 딥페이크에서는 머리 움직임이 지나치게 부드럽거나, 반대로 급격한 전환이 발생할 수 있습니다.

화면 구성:

- **3 축 회전 그래프**: Yaw(좌우), Pitch(상하), Roll(기울기)의 시간별 변화
- **변위량(Displacement)**: 각 시점에서의 머리 이동량
- **이상치 표시**: 비정상적인 급격한 움직임 구간

그래프 판독법: 3 축 그래프에서 **물리적으로 불가능한 속도의 회전**(예: 0.1 초 만에 30 도 이상 회전)이 감지되면 합성 영상의 강력한 증거입니다. 그래프가 지나치게 매끄러운 것(떨림이 전혀 없음)도 딥페이크의 특징입니다.

용어 해설

용어	설명
Yaw	머리의 좌우 회전 (고개를 양옆으로 돌리는 동작)
Pitch	머리의 상하 기울임 (고개를 끄덕이거나 젖히는 동작)
Roll	머리의 좌우 기울기 (어깨 쪽으로 고개를 기울이는 동작)

6. 오디오 포렌식 섹션

이 섹션은 음성의 진위를 판별하는 전문 도구를 제공합니다. AI 음성 합성(TTS), 보이스 클로닝 등의 흔적을 탐지합니다.

6.1 오디오 포렌식

메뉴 경로: 사이드바 > 오디오 포렌식 > 오디오 포렌식

경로: `/audio-forensics`

기능: 음성 파일의 종합적인 진위를 분석합니다. MFCC(멜 주파수 켑스트럼 계수) 분석, 스펙트럼 분석, 음성 품질 메트릭 등을 종합합니다.

화면 구성:

- **판정 결과**: 음성의 REAL/FAKE/UNCERTAIN 판정
- **MFCC 표준편차**: 음성 특성의 변동성 (낮으면 합성 의심)
- **영교차율(ZCR)**: 음성 신호가 0 을 지나는 빈도
- **스펙트럼 분석**: 주파수 대역별 에너지 분포

그래프 판독법: MFCC 표준편차가 비정상적으로 낮으면(0.5 이하) AI 합성 음성의 특징입니다. 스펙트럼에서 4kHz 이상 고주파 대역이 갑자기 끊기면 음성 압축 또는 합성의 흔적일 수 있습니다.

용어 해설

용어	설명
MFCC	멜 주파수 켑스트럼 계수 — 음성의 고유한 특성을 수치화한 지표
ZCR(Zero-Crossing Rate)	음성 신호가 0 을 지나는 빈도. 자연 음성과 합성 음성에서 차이를 보임
스펙트럼(Spectrum)	음성을 주파수 성분별로 분해한 표현
TTS(Text-to-Speech)	텍스트를 음성으로 변환하는 기술. 최근 AI 의 발달로 매우 자연스러운 합성이 가능해짐

6.2 SSL 음성 탐지

메뉴 경로: 사이드바 > 오디오 포렌식 > SSL 음성 탐지

경로: `/audio-ssl`

기능: **자기지도 학습(Self-Supervised Learning)** 기반 모델을 활용하여 합성 음성을 탐지합니다. 사전 학습된 대규모 음성 모델이 추출한 특성을 기반으로 진위를 판별합니다.

화면 구성:

- **SSL 점수**: 자기지도 학습 모델의 위조 의심도
- **클러스터 분석**: 음성 특성의 분포 시각화

용어 해설

용어	설명

SSL(Self-Supervised Learning)	레이블 없이 데이터의 내재적 구조를 학습하는 방법
wav2vec	Meta 가 개발한 음성 자기지도 학습 모델

6.3 보코더 식별

메뉴 경로: 사이드바 > 오디오 포렌식 > 보코더 식별

경로: `/vocoder-id`

기능: AI 음성 합성에 사용된 **보코더(Vocoder)**의 종류를 식별합니다. 각 보코더는 고유한 음향 특성을 남기므로, 어떤 도구로 합성되었는지를 추정할 수 있습니다.

화면 구성:

- **식별된 보코더**: WaveNet, WaveGlow, HiFi-GAN 등 추정 보코더
- **보코더별 유사도**: 각 보코더 유형과의 일치도 (바 차트)
- **자연 음성 유사도**: 실제 사람 목소리와의 유사도

그래프 판독법: 특정 보코더의 유사도가 0.7 이상이면 해당 보코더로 합성되었을 가능성이 높습니다. 자연 음성 유사도가 0.8 이상이면 진본일 가능성이 높지만, 최신 AI 합성은 0.9 이상의 유사도를 보일 수 있으므로 다른 모달리티와 교차 확인이 필요합니다.

용어 해설

용어	설명
보코더(Vocoder)	음성 합성의 마지막 단계에서 파형을 생성하는 모듈. WaveNet, HiFi-GAN 등이 대표적
WaveNet	Google DeepMind 가 개발한 음성 합성 보코더
HiFi-GAN	고품질 음성 파형을 빠르게 생성하는 GAN 기반 보코더

7. 포렌식 도구 섹션

이 섹션은 분석 결과를 심층적으로 해석하고, 법적 증거로 활용 가능한 보고서를 생성하는 전문 도구를 제공합니다.

7.1 포렌식 보고서

메뉴 경로: 사이드바 > 포렌식 도구 > 포렌식 보고서

경로: `/forensic-report`

기능: 분석 결과를 **법적 증거력**을 갖춘 포렌식 보고서로 생성합니다. PDF 또는 Excel 형식으로 다운로드할 수 있으며, 국과수(국립과학수사연구원) 제출 형식을 지원합니다.

화면 구성:

- **보고서 목록**: 생성된 보고서 이력
- **SHAP 분석**: 각 탐지 모듈이 판정에 기여한 정도
- **베이지안 불확실성**: 각 모듈의 판정 신뢰 구간
- **해시 체인 검증**: 보고서의 무결성 검증 상태

그래프 판독법: SHAP 분석의 폭포(Waterfall) 차트에서 빨간색 바는 FAKE 방향으로의 기여, 파란색 바는 REAL 방향으로의 기여를 나타냅니다. 가장 긴 바가 판정에 가장 큰 영향을 미친 모달리티입니다.

용어 해설

용어	설명
SHAP(SHapley Additive exPlanations)	각 입력 요소가 결과에 얼마나 기여했는지를 수학적으로 분해하는 설명 기법
베이지안 불확실성(Bayesian Uncertainty)	판정의 불확실성을 확률 분포로 표현하는 방법
해시 체인(Hash Chain)	각 기록의 무결성을 이전 기록의 해시값으로 연결하여 보장하는 기법

7.2 핑거프린트 DB

메뉴 경로: 사이드바 > 포렌식 도구 > 핑거프린트 DB

경로: `/fingerprint-db`

기능: 분석된 미디어의 **디지털 핑거프린트(pHash/dHash/aHash)**를 데이터베이스에 저장하고, 동일하거나 유사한 콘텐츠가 재유입되면 자동으로 매칭하는 기능입니다. YouTube 의 Content ID 와 유사한 원리입니다.

화면 구성:

- **핑거프린트 탭**: 등록된 핑거프린트 목록 (파일명, pHash, dHash, 등록일, 레이블)
- **매칭 이력 탭**: 유사 콘텐츠 검색 이력 (유사도, 매칭 일시)
- **"새 핑거프린트 등록" 버튼**: 파일 업로드 → pHash/dHash 자동 계산 → DB 저장

사용 방법:

1. **수동 등록**: "새 핑거프린트 등록" 버튼 클릭 → 이미지/영상 업로드
2. **자동 등록**: 분석 → FAKE 판정 시 첫 프레임이 자동으로 핑거프린트 DB 에 등록됩니다
3. **유사 검색**: 의심 파일을 업로드하면 DB 에 저장된 핑거프린트와 비교하여 유사도 85% 이상인 매칭 결과를 반환합니다
4. **중복 방지**: 동일 pHash+dHash 를 가진 파일은 중복 등록되지 않습니다

저장 위치: `services/ai-core/logs/fingerprint_db.json`

용어 해설

용어	설명
pHash(Perceptual Hash)	이미지를 리사이즈/압축해도 유사하면 동일한 해시 값을 생성하는 지각적 해시
dHash(Difference Hash)	인접 픽셀 간 밝기 차이 패턴으로 해시를 생성하는 방식
Hamming Distance	두 해시 간 다른 비트의 수. 작을수록 유사함

7.3 위협 인텔리전스

메뉴 경로: 사이드바 > 포렌식 도구 > 위협 인텔리전스

경로: `/threat-intel`

기능: 딥페이크 위협 현황을 시각화하고, 외부 소스에서 최신 위협 정보를 **자동으로 크롤링**하여 업데이트합니다.

화면 구성 (3 탭):

탭 1 — 탐지 커버리지: 52 종 딥페이크 생성 도구의 탐지 커버리지 맵입니다. 각 도구가 탐지 완비(초록), 부분 대응(노란), 미대응(빨강)인지를 한눈에 파악할 수 있습니다.

탭 2 — 크롤링 인텔리전스: 외부 소스에서 자동 수집된 딥페이크 관련 최신 정보를 열람합니다.

- **arXiv**: 딥페이크 탐지/생성 관련 최신 논문

- **HuggingFace**: 새로 공개된 AI 생성 모델
- **RSS**: 보안 뉴스 (보안뉴스, The Hacker News 등)
- 소스별/카테고리별 필터링, 키워드 검색, 원문 링크 제공

탭 3 — 자동 크롤링 설정: 크롤링 스케줄을 설정합니다.

- **요일 선택**: 월~일 중 실행 요일 선택
- **시간 설정**: 실행 시각 (시:분)
- **소스 선택**: arXiv / HuggingFace / RSS 중 선택
- **수집량**: 소스당 최대 수집 건수 (5~50 건)
- **"지금 실행" 버튼**: 즉시 크롤링 실행

용어 해설

용어	설명
크롤링(Crawling)	외부 웹사이트에서 정보를 자동으로 수집하는 기술
arXiv	학술 논문 사전 공개 플랫폼. 딥페이크 관련 최신 연구가 가장 먼저 공개되는 곳
HuggingFace	AI 모델 공유 플랫폼. 새로운 딥페이크 생성 모델이 공개되면 위협으로 등록
RSS	뉴스 사이트의 자동 구독 형식. 보안 관련 최신 기사를 수집

7.4 XAI 심층 분석

메뉴 경로: 사이드바 > 포렌식 도구 > XAI 심층 분석

경로: `/xai-deep-dive`

기능: 판정의 근거를 **설명 가능한 AI(XAI)** 기법으로 심층 분석합니다. "왜 이 영상이 FAKE 로 판정되었는가?"라는 질문에 시각적으로 답합니다.

화면 구성:

- **SHAP 폭포 차트**: 각 모듈의 판정 기여도 (양방향 바 차트)
- **Grad-CAM++ 히트맵**: 모델이 주목한 영역의 열지도 (파란색→노란색→빨간색)
- **베이지안 불확실성**: 각 모듈의 판정 신뢰 구간 (평균, 표준편차, 90% CI)
- **모듈 기여도**: 각 탐지 모듈의 기여 비율 (바 차트)

그래프 판독법:

- **SHAP 폭포 차트**: 기저선(Base=0.5)에서 출발하여 각 모듈이 FAKE(+) 또는 REAL(-) 방향으로 얼마나 밀었는지를 보여줍니다. 최종 도달점이 판정 결과입니다.
- **Grad-CAM++ 히트맵**: 빨간색 영역이 모델이 "가장 의심스럽다"고 판단한 부분입니다. 이 영역이 얼굴 경계, 머리카락, 또는 배경과의 접점에 집중되면 합성 흔적입니다.
- **베이지안 불확실성**: 각 모듈의 점과 가로선으로 표시됩니다. 점이 평균, 가로선이 90% 신뢰구간입니다. 신뢰구간이 넓으면 해당 모듈의 판단이 불확실하다는 의미입니다.

용어 해설

용어	설명
XAI(eXplainable AI)	AI 의 판단 과정을 인간이 이해할 수 있도록 설명하는 기술
Grad-CAM++	신경망의 기울기(Gradient)를 활용하여 중요 영역을 시각화하는 기법
신뢰구간(Confidence Interval)	실제 값이 존재할 것으로 예상되는 범위. 90% CI 는 100 번 중 90 번은 이 범위 안에 있을 것이라는 의미

7.5 모델 핑거프린터

메뉴 경로: 사이드바 > 포렌식 도구 > 모델 핑거프린터

경로: `/model-fingerprinter`

기능: AI 생성 콘텐츠가 **어떤 모델(Sora, Kling, Runway, SDXL 등)로 만들어졌는지**를 추적하는 6 단계 핑거프린팅 파이프라인입니다. 분석할수록 프로파일이 자동으로 정교해지는 **EMA(지수 이동 평균) 자동 학습** 기능이 내장되어 있습니다.

화면 구성:

- **귀속 결과**: 추정된 생성 모델명과 확신도
- **6 단계 파이프라인 결과**: 각 분석 단계(노이즈 추출→주파수→인과→시간→위험→귀속)의 점수
- **후보 모델 순위**: 상위 5 개 후보 모델과 유사도
- **프로파일 통계**: 등록된 모델 프로파일 수, 자동 학습 샘플 수

프로파일 자동 확장: 분석 결과의 신뢰도가 일정 수준 이상이면, 노이즈 잔차의 통계(평균, 분산, 주파수 패턴)가 해당 모델의 프로파일 DB 에 자동으로 누적됩니다. 이를 통해 분석할수록 귀속 정확도가 향상됩니다.

용어 해설

용어	설명
EMA(Exponential Moving Average)	새로운 관측값에 더 높은 가중치를 부여하면서 과거 데이터와 점진적으로 블렌딩하는 학습 방법
FDR(Fréchet Distance Ratio)	모델 프로파일과 분석 대상의 통계적 거리를 비교하여 가장 유사한 모델을 찾는 척도
노이즈 잔차(Noise Residual)	원본 이미지에서 디노이즈 결과를 뺀 잔여 신호. AI 모델마다 고유한 패턴을 남김

7.6 T-GD 출처 분석

메뉴 경로: 사이드바 > 포렌식 도구 > T-GD 출처 분석

경로: `/tgd-attribution`

기능: **전이학습 기반 생성 탐지(Transfer Learning for Generative Detection)** 모듈로, 500 개 이상의 AI 생성 모델 데이터베이스에서 출처를 특정합니다. **법적 증거력 점수(LES)**를 산출합니다.

화면 구성 (3 탭):

탭 1 — 모델 검색 (500 개): 3 단계 계층 검색으로 500 개 모델을 쉽게 찾을 수 있습니다.

- **Step 1**: 모달리티 대분류 선택 (Image 300 / Video 100 / Audio 50 / Text 50)
- **Step 2**: 알파벳 인덱스 (A~Z, 해당 모달리티 내 모델 수 표시, 없는 알파벳은 비활성)
- **Step 3**: 키워드 검색 (모델명, 조직명, 패밀리로 추가 필터링)

탭 2 — 분석 이력 출처 추적: 이전에 분석한 파일 목록에서 "출처 분석" 버튼을 클릭하면 Module #39 핑거프린터 6 단계 파이프라인이 실행되어, 해당 미디어를 어떤 AI 모델로 생성했는지를 추적합니다.

탭 3 — LES 법적 증거: 법적 증거력 점수(LES) 구성과 레지스트리 분포를 확인합니다.

LES 점수 구성:

구성 요소	최대 점수	설명
Detection 확신도	30 점	T-GD Detection Head 의 판별 확신도
Attribution 확신도	35 점	출처 모델 특정의 확신도
Top-1/2 마진	20 점	1 순위와 2 순위 후보 간 점수 격차

다중 합의	15 점	기존 모듈과의 판정 일치율

용어 해설

용어	설명
T-GD	Transfer Learning for Generative Detection — 전이학습을 활용한 AI 생성물 탐지 기술
LES(Legal Evidence Score)	법적 증거력 점수. 수사기관에 증거로 제출할 때의 신뢰도 지표
AUROC	Area Under the ROC Curve — 분류 모델의 성능 지표. 1.0 에 가까울수록 우수

8. 방어 및 분류 섹션

8.1 3-Class 분류

메뉴 경로: 사이드바 > 방어 & 분류 > 3-Class 분류

경로: `/three-class`

기능: 미디어를 **REAL(진본) / FAKE(위조) / ANTI-FORENSIC(반포렌식)**의 세 가지 클래스로 분류합니다. 반포렌식은 딥페이크의 흔적을 의도적으로 제거하려는 시도가 감지된 경우입니다.

용어 해설

용어	설명
반포렌식(Anti-Forensic)	딥페이크 탐지를 회피하기 위해 의도적으로 조작 흔적을 제거하는 기법

8.2 사전 방어

메뉴 경로: 사이드바 > 방어 & 분류 > 사전 방어

경로: `/proactive-defense`

기능: 원본 이미지에 **눈에 보이지 않는 교란(Perturbation)**을 주입하여, 공격자가 해당 이미지로 딥페이크를 만들 때 생성 결과의 품질이 심각하게 저하되도록 합니다. PSNR 40dB 이상으로 원본과 보호 이미지의 차이를 **육안으로 구분할 수 없지만**, DeepFaceLab 등의 AI 가 처리하면 얼굴이 일그러지거나 아티팩트가 발생합니다.

화면 구성:

- **통계 카드**: 보호된 이미지 수, 평균 PSNR, 평균 SSIM, 지원 모델 수
- **이미지 업로드 영역**: 드래그 & 드롭 또는 클릭으로 이미지 업로드
- **대상 모델 선택**: 8 종 (DeepFaceLab, InsightFace, LivePortrait, FaceSwap, Ghost, SimSwap, FaceFusion, Ensemble) 중 복수 선택 가능
- **교란 강도 선택**: Strict(최대 보호) / Standard(권장) / Relaxed(화질 우선)
- **교란 결과**: PSNR, SSIM, 예상 품질 저하율, 비가시성 검증 결과
- **보호 이력**: 교란 적용 내역 (파일명, 대상 모델, PSNR, 상태, 시간)

사용 방법:

1. 보호할 이미지를 업로드합니다 (기업 임원 프로필, 공직자 사진 등)
2. 대상 모델을 선택합니다 (모르면 "Ensemble" 선택 — 모든 모델 대응)
3. 교란 강도를 선택합니다 (일반적으로 "Standard" 권장)
4. 교란이 적용되면 결과를 확인합니다 (PSNR ≥ 40dB 이면 정상)
5. 보호된 이미지를 다운로드하여 사용합니다

용어 해설

용어	설명
적대적 방어(Adversarial Defense)	AI 공격에 대응하여 원본을 보호하는 기법
PGD	Projected Gradient Descent — 반복적으로 최적의 교란을 탐색하는 알고리즘
MI-FGSM	모멘텀 기반 공격 — 여러 모델에 전이 가능한 교란 생성 (권장)
PSNR	원본과 보호 이미지의 품질 차이 지표. 40dB 이상이면 육안 구분 불가
SSIM	구조적 유사도 지표. 0.99 이상이면 원본과 거의 동일

9. 실시간 및 인프라 섹션

9.1 실시간 모니터

메뉴 경로: 사이드바 > 실시간 & 인프라 > 실시간 모니터

경로: `/realtime-monitor`

기능: 웹캠 또는 RTSP 스트림을 **실시간으로 모니터링**하여 딥페이크 영상이 감지되면 즉시 알림을 발생시킵니다. 화상 회의, 라이브 방송 등의 실시간 검증에 활용합니다.

화면 구성:

- **탐지 모델 선택**: 기본 탐지 모듈(Full Pipeline, 450MB) 또는 **증류된 경량 모델(Student, 67MB)**을 선택하여 실시간 탐지 속도를 조절할 수 있습니다. 경량 모델은 "모델 증류" 메뉴에서 생성합니다.
- **카메라 미리보기**: 웹캠 영상을 실시간으로 표시합니다 (미러링 적용)
- **탐지 포인트 타임라인**: 시간축에 따른 탐지 이벤트 표시
- **알림(Alert)**: 심각도별 알림 목록
- **프레임 메트릭**: 실시간 분석 중인 프레임의 품질 지표 (Laplacian, 엣지 밀도, 얼굴 수)
- **뷰티 필터 분석**: SNS 뷰티 필터 4 종(Mesh Warping, Frequency Separation, Semantic Segmentation, GAN-lite) 적용 여부 탐지

탐지 모델 선택 방법:

1. "탐지 모델 선택" 패널에서 원하는 모델 카드를 클릭합니다
2. **기본 모듈**: 40 개 전체 파이프라인 — 정확도 94%, 450MB, 1200ms
3. **경량 Student**: 증류된 MobileNet V3 등 — 정확도 88%, 67MB, 96ms (12 배 빠름)
4. 모델 증류 메뉴에서 Teacher 를 선택하고 Student 를 생성하면 여기에 자동으로 표시됩니다

9.2 모델 증류

메뉴 경로: 사이드바 > 실시간 & 인프라 > 모델 증류

경로: `/model-distillation`

기능: 대형 탐지 모델(Teacher)의 지식을 경량 모델(Student)로 전달하여, 모바일 기기나 엣지 장비에서도 딥페이크 탐지를 실행할 수 있도록 합니다. 증류된 Student 모델은 **실시간 모니터**에서 선택하여 사용할 수 있습니다.

화면 구성:

- **통계 카드**: Teacher 모델 수(5), Student 모델 수, TensorRT 모델 수, ONNX 모델 수
- **Teacher 모델 목록**: 증류에 사용할 수 있는 5 개 Teacher 모델의 정확도, 크기, 추론 시간

• **증류 파이프라인**: 4 단계 (Knowledge Distill → ONNX Export → INT8 Quantize → TensorRT Optimize)

• **증류 실행**: Teacher 선택 → Student 아키텍처 선택 (MobileNet V3, EfficientNet B0 등) → 증류 시작

Teacher 모델 5 종:

Teacher	정확도	크기	추론 시간	설명
ensemble_v3	94%	450MB	1200ms	Xception+EfficientNet+ViT 3 종 앙상블
vit_swin_v2	92%	340MB	800ms	고해상도 아티팩트 탐지
cross_efficient_vit	90%	280MB	650ms	시간적 교차 검증
frequency_analyzer	88%	120MB	300ms	주파수 도메인 탐지
biological_detector	85%	95MB	250ms	rPPG+생체신호

증류 결과 예시: ensemble_v3(450MB, 94%) → MobileNet V3(67.5MB, 88%) — 크기 85% 감소, 속도 12 배 향상

증류된 모델 활용:

1. **실시간 모니터**: 증류 후 자동으로 "탐지 모델 선택" 패널에 표시됩니다
2. **모바일 앱**: .onnx 파일을 앱에 임베드하여 스마트폰에서 즉시 탐지
3. **엣지 장비**: .trt 파일을 NVIDIA Jetson 에 배포하여 CCTV 실시간 탐지
4. **웹 브라우저**: ONNX.js 로 브라우저 내 추론 — 업로드 없이 로컬 탐지

용어 해설

용어	설명
모델 증류(Model Distillation)	대형 AI 모델(Teacher)의 지식을 소형 모델(Student)로 전달하여 경량화하는 기법
Teacher / Student	Teacher 는 정확하지만 무거운 원본 모델, Student

	는 Teacher 의 지식을 학습한 경량 모델
ONNX	다양한 AI 프레임워크 간 모델을 호환시키는 개방형 표준
TensorRT	NVIDIA GPU 에서 AI 추론을 가속하는 최적화 엔진
INT8 양자화	모델의 수치 정밀도를 32 비트에서 8 비트로 낮추어 속도를 3-4 배 높이는 기법

10. 에이전트 AI 섹션

이 섹션은 TruthLens 의 **멀티 에이전트 AI 시스템**을 관리하고 모니터링하는 고급 워크스페이스입니다. 일반 사용자는 이 섹션을 사용하지 않아도 분석에 지장이 없으나, 시스템의 내부 동작을 이해하고 튜닝하고자 할 때 활용합니다.

10.1 멀티에이전트 프레임워크

메뉴 경로: 사이드바 > 에이전트 AI > 멀티에이전트 프레임워크

경로: `/multi-agent`

기능: TruthLens 에 통합된 **8 개 멀티에이전트 AI 프레임워크**의 상태와 성능을 모니터링합니다.

화면 구성:

- **프레임워크 상태 카드**: 각 프레임워크의 활성화/비활성화 상태, 호출 횟수, 평균 응답 시간
- **에이전트 트리**: 20 개 에이전트의 계층 구조 시각화
- **가드레일 현황**: AI 안전 장치의 통과/차단 비율

용어 해설

용어	설명
멀티에이전트(Multi-Agent)	여러 AI 에이전트가 협력하여 하나의 결론을 도출하는 시스템
LangGraph	그래프 기반 AI 워크플로우 오케스트레이션 프레임워크
CrewAI	여러 AI 전문가가 역할을 분담하여 토론하는 프레임워크
가드레일(Guardrail)	AI 의 출력이 안전하고 정확한지를 검증하는 안전 장

	치

10.2 에이전트 설정

메뉴 경로: 사이드바 > 에이전트 AI > 에이전트 설정

경로: `/agent-settings`

기능: 각 에이전트의 활성화/비활성화, 모델 할당, 파라미터를 설정합니다.

10.3 AutoGen 토론

메뉴 경로: 사이드바 > 에이전트 AI > AutoGen 토론

경로: `/autogen-debate`

기능: 애매한 사례에 대하여 **검찰, 변호인, 판사** 역할의 AI 에이전트가 적대적으로 토론하여 결론을 도출하는 과정을 시각화합니다.

10.4 LLM / 추론 엔진

메뉴 경로: 사이드바 > 에이전트 AI > LLM / 추론 엔진

경로: `/llm-settings`

기능: AI 모델의 종류와 설정을 관리합니다. 추론 엔진(Ollama/vLLM), LLM 모델, VLM(비전 모델), 임베딩 모델을 선택합니다.

주요 설정:

- **VLM 모델**: 비전 분석에 사용되는 모델 (기본: llama3.2-vision)
- **Temperature**: AI 응답의 다양성 (낮을수록 일관적, 0.1 권장)
- **Context Window**: AI 가 한 번에 처리하는 텍스트의 최대 길이

10.5 ADAG 레드팀

메뉴 경로: 사이드바 > 에이전트 AI > ADAG 레드팀

경로: `/adag-redteam`

기능: TruthLens 자체를 공격하여 **취약점을 사전에 발견**하는 적대적 테스트 시스템입니다. 4 종의 공격 에이전트가 탐지 시스템을 회피하려고 시도하고, 그 결과를 분석합니다.

용어 해설

용어	설명
레드팀(Red Team)	시스템의 취약점을 찾기 위해 공격자 역할을 수행하는 팀
ADAG	Adaptive Defense-Attack Game — 방어와 공격이 반복적으로 학습하는 프레임워크
DER(Detection Evasion Rate)	공격이 탐지를 회피한 비율. 낮을수록 탐지 시스템이 강건함

10.6 MC Fusion 시뮬레이션

메뉴 경로: 사이드바 > 에이전트 AI > MC Fusion 시뮬레이션

경로: `/mc-simulation`

기능: 몬테카를로 시뮬레이션을 통해 각 탐지 모듈의 **최적 가중치**를 산출합니다. 모듈 간 기여도를 과학적으로 조정하여 전체 정확도를 극대화합니다.

화면 구성:

- **최적화 방법별 결과**: Grid Search, Optuna, 차분 진화, 몬테카를로의 4 가지 최적화 결과 비교
- **가중치 추천**: 최적의 모달리티별 가중치 조합
- **신뢰도 등급**: A~F 등급의 전체 시스템 신뢰도

용어 해설

용어	설명
몬테카를로 시뮬레이션(Monte Carlo Simulation)	무작위 시행을 수만 번 반복하여 최적 해를 찾는 통계적 방법
가중치(Weight)	각 모달리티의 판정이 최종 결과에 기여하는 비율
F1 Score	정밀도와 재현율의 조화 평균. 분류 성능의 종합 지표

10.7 온톨로지 파이프라인

메뉴 경로: 사이드바 > 에이전트 AI > 온톨로지 파이프라인

경로: `/ontology-pipeline`

기능: TruthLens 의 40 개 탐지 모듈과 14 개 파이프라인 노드의 구조를 **OWL 온톨로지**로 형식화하고, 6 개 SWRL 추론 규칙으로 실시간 검증하는 시스템입니다.

화면 구성 (4 개 탭):

Pipeline Flow 탭: 14 개 파이프라인 노드의 실행 순서를 시각화합니다. 각 노드를 클릭하면 해당 노드가 읽는 데이터(Reads)와 쓰는 데이터(Writes)가 표시됩니다.

Module Map 탭: 40 개 탐지 모듈을 도메인별(시각, 오디오, 생체, 시간, 텍스트, T-GD)로 분류하여 표시합니다. 모듈 간 의존성 관계도 확인할 수 있습니다.

SWRL Violations 탭: 6 개 SWRL 추론 규칙의 통과/위반 상태를 실시간으로 모니터링합니다.

규칙	감지 대상	의미
SWRL-1	단일 모달리티 의존	1 개 모달리티만 활성화되면 신뢰도를 60%로 제한
SWRL-2	데이터 순서 위반	파이프라인에서 데이터가 잘못된 순서로 전달되는 오류
SWRL-3	VLM 환각	비전 모델이 모든 카테고리에 동일 점수를 부여하는 이상
SWRL-4	순환 의존성	모듈 간 순환 참조 (무한 루프 위험)
SWRL-5	모달리티 누락	필수 분석 모듈이 비활성화된 상태
SWRL-6	가중치 오류	모달리티 가중치 합계가 100%가 아닌 상태

Fusion Weights 탭: 각 모달리티의 가중치를 바 차트로 표시하고, 합계가 100%인지 검증합니다.

용어 해설

용어	설명
온톨로지(Ontology)	지식을 체계적으로 분류하고 관계를 정의하는 형식 체계
OWL(Web Ontology Language)	웹 표준 온톨로지 기술 언어
SWRL	Semantic Web Rule Language — 온톨로지 위에 추론 규칙을 정의하는 언어

파이프라인 노드(Pipeline Node)	분석 과정의 각 단계를 수행하는 처리 단위

11. 관리 섹션

관리자(ADMIN) 역할의 사용자만 접근할 수 있습니다.

- **관리자 대시보드** (`/admin`): 사용자 목록 관리, 역할 변경, 비밀번호 초기화
- **조직 관리** (`/admin/organizations`): 조직 생성/수정, 요금제(Tier) 변경, 사용량 쿼터 설정
- **API 키 관리** (`/admin/api-keys`): API 키 발급/폐기, 만료 일자 설정

12. 시스템 섹션

- **설정** (`/settings`): 언어 선택(한국어/영어/일본어/중국어), 사용자 프로필, 비밀번호 변경
- **API 문서** (`/api-docs`): TruthLens REST API 의 전체 사양을 Swagger UI 로 제공합니다. 외부 시스템과의 연동 개발 시 참고합니다.

13. 판정 기준 및 해석 가이드

13.1 판정 임계값

조건	판정
최종 확률 ≥ 0.65	FAKE (위조)
최종 확률 ≤ 0.25	REAL (진본)
0.25 < 최종 확률 < 0.65	UNCERTAIN (판단 유보)

13.2 신뢰도 상한 규칙

활성 모달리티 수	최대 신뢰도	의미
1 개 (시각만)	60%	한 가지 관점만으로는 확신하기 어려움
2 개	75%	두 가지 관점에서 교차 검증되었으

		나 완전하지 않음
3 개 이상	100%	다각적 교차 검증이 이루어짐

13.3 UNCERTAIN 판정 시 권장 조치

1. **영상 품질 확인**: 원본 파일이 고해상도인지 확인합니다. 과도하게 압축된 파일은 분석 정확도가 떨어집니다.
2. **정밀 모드 재분석**: 신속/표준 모드로 분석했다면 정밀(Precise) 모드로 재분석합니다.
3. **전문가 리뷰**: XAI 심층 분석 결과를 포렌식 전문가에게 검토 요청합니다.
4. **원본 대조**: 가능하다면 원본 파일과 대조합니다.

14. 부록: 딥페이크 위협과 TruthLens 의 사회적 책무

14.1 딥페이크 위협의 현실

2025~2026 년 현재, 딥페이크 기술의 대중화로 인한 피해가 전례 없는 규모로 확산되고 있습니다.

금융 분야

• 기업 CEO 의 화상 회의 영상을 딥페이크로 위조하여 6,200 만 달러(약 830 억 원)를 송금하도록 유도한 사건이 보고되었습니다(2024 년 홍콩).

• 보이스피싱에 AI 음성 합성이 활용되어, 가족의 목소리를 완벽하게 모사한 사기가 급증하고 있습니다.

정치/사회

• 선거 기간 중 후보자의 딥페이크 영상이 유포되어 민주주의적 의사결정을 왜곡하는 사례가 다수 발생하였습니다.

• 딥페이크 포르노그래피 등 개인의 인격권을 침해하는 범죄가 사회적 문제로 대두되고 있습니다.

방송/미디어

• 가짜 뉴스 영상이 SNS 를 통해 급속히 확산되어 사회적 혼란을 야기하는 사례가 빈번합니다.

• 유명인을 사칭한 딥페이크 광고가 소비자 피해를 유발하고 있습니다.

14.2 대한민국의 법적 대응

• **AI 기본법 제 15 조**: AI 시스템의 판단에 대한 설명 의무를 규정합니다. TruthLens 는 모든 판정에 XAI 시각화와 감사 로그를 제공하여 이 의무를 충족합니다.

• **정보통신망법**: 허위 영상물의 유포를 금지하고 있으며, TruthLens 의 분석 보고서는 수사기관의 증거 자료로 활용될 수 있습니다.

• **개인정보보호법**: TruthLens 는 분석 대상 미디어를 분석 완료 후 즉시 삭제하며, 개인정보를 별도 저장하지 않습니다.

14.3 TruthLens 의 사회적 책무

TruthLens 는 단순한 기술 도구가 아닌, **사회 안전망의 핵심 인프라**로서의 사명을 가지고 개발되었습니다.

진실 보호: 영상, 음성, 문서의 진위를 과학적으로 검증하여 허위 정보로부터 시민을 보호합니다.

피해 예방: 금융 사기, 보이스피싱, 여론 조작 등 딥페이크를 악용한 범죄를 사전에 탐지하여 선의의 피해자 발생을 방지합니다.

증거 보전: SHA-256 해시 체인으로 보호되는 감사 로그와 법적 증거력(LES) 점수를 제공하여, 수사와 재판에서 활용 가능한 과학적 증거를 확보합니다.

투명한 판단: 모든 판정의 근거를 XAI 시각화로 제공하여, AI 의 "블랙박스" 문제를 해소하고 인간의 최종 판단을 지원합니다.

지속적 진화: ADAG 레드팀 시스템을 통해 스스로의 취약점을 발견하고 개선하는 자기강화 메커니즘을 갖추고 있어, 날로 교묘해지는 딥페이크 기술에 지속적으로 대응합니다.

> *"기술은 양날의 검입니다. AI 가 만들어낸 위협에는 AI 로 대응해야 합니다. TruthLens 는 디지털 시대의 진실을 지키는 방패이며, 선의의 피해자가 더 이상 발생하지 않도록 하는 것이 우리의 사회적 책무입니다."*

문서 끝

본 가이드는 TruthLens v4.4.1 기준으로 작성되었습니다.

Creator: Brian Lee | Test Research: Younghyun Han | AI R&D Center | A3 Security Co.,Ltd. | April 19, 2026