관리번호	2025-국제공동연구-01	(품목공모형)
기술분류	대분류(인공지능)-중분류(신뢰·산업AI)-소분류(신뢰·공정AI)-세분류(안전한AI)	
중점분야	AI(√), AI반도체( ), 5G·6G( ), 양자( ), 메타버스( ), 사이버보안( )	
기획유형	파괴적혁신기술( ), 초격차신격차( ), 창의도전형R&D( ), 일반R&D(√)	
품목(문제)명	설명가능한 멀티모달 딥페이크 조작영상 탐지기술 개발 Development of high real-world accuracy deepfake detection and blocking	

1. 품목(문제) 정의

○ **(개념)** 딥페이크 조작 영상에 대한 높은 탐지 정확도와 진위여부에 대한 직관적인 설명을 제공할 수 있는 영상/음성/텍스트 멀티모달 딥페이크 불법 영상 탐지 기술
(Video/voice/text multimodal deepfake illegal video detection technology that can provide high detection accuracy and intuitive explanations of authenticity for deepfake manipulated videos)

- 본 과제는 AI싱가포르(AISG)와 정보통신기획평가원(IITP)이 공동으로 연구비를 지원하는 과제임
(This is jointly funded by the AI Singapore(AISG), and the Institute for Information and Communications Technology Planning and Evaluation (IITP) in Republic of Korea)

○ **(연구목표)** 영상, 음성, 텍스트 등 다양한 모달리티 데이터를 통합 분석하여 딥페이크 조작 영상을 높은 정확도로 탐지하고, 탐지된 조작 부분을 사용자가 쉽게 파악할 수 있도록 바운딩박스, 캡션 등을 통해 직관적인 설명을 제공하는 기술개발
(Research and development of technology to detect deepfake manipulation videos with high accuracy by integrating and analyzing various modality data such as video, voice, and text, and provide intuitive explanations through bounding boxes, captions, etc. to help users easily understand the detected manipulation parts.)

* 국가별 문화, 환경 등에 따른 딥페이크 피쳐 발굴 및 국가간 공조를 통한 딥페이크 탐지 플랫폼 구축 및 운영 포함

As-Is (딥러닝 기반 딥페이크 탐지)	To-Be (설명가능한 멀티모달 딥페이크 탐지)
- 주로 단일 모달리티(예: 영상) 데이터를 사용하여 딥페이크를 탐지 - 딥러닝 모델은 특정 특징(예: 픽셀 왜곡, 눈 깜빡임 등)을 학습하여 이를 탐지하는 방식 - 주로 시각적 데이터 등의 한정된 정보만 의존하여 탐지 정확도가 제한될 수 있음 - 이미지 생성 AI 서비스에서 유명인사 또는 특정 인물을 생성하지 못하도록 제한함	- 다양한 모달리티(영상, 음성, 텍스트 등)를 통합적으로 분석 - 영상에서 발생하는 미세한 왜곡뿐만 아니라, 입모양과 음성의 불일치, 자막과 음성의 차이 등, 영상/음성/텍스트 상관관계를 종합적으로 분석 - 다양한 데이터 소스가 상호보완적인 역할을 해, 고도로 정교한 딥페이크도 더 정확하게 탐지 가능 - 탐지된 조작 부분을 사용자가 직관적으로 이해할 수 있도록 바운딩박스, 캡션 등의 시각적 설명을 제공

○ **(지원범위)**
- 딥페이크 탐지를 위한 글로벌 공통플랫폼 요구사항 도출, 아키텍처, 기능 및 인터

페이스 설계 및 구현
(Deriving requirements, designing and implementing architecture, functionality, and interfaces for a global common platform for deepfake detection)

- 여러 국가들이 공동으로 이용할 수 있는 클라우드 기반의 개방형 딥페이크 탐지 글로벌 공통플랫폼 프로토타입 구축
(Prototyping a cloud-based open global common platform for deepfake detection that can be jointly used by multiple countries)
- 혁신적이고 높은 정확도와 확장성을 갖춘 멀티모달 AI 딥페이크 탐지 기술을 공동으로 연구·개발하여, 동영상·음성·텍스트 입력 내의 딥페이크 조작을 식별하고, 경계 상자(bounding boxes), 자막(captions), 신뢰 점수(confidence scores) 등을 통해 진위 여부를 직관적으로 설명함으로써 사용자가 조작된 부분을 쉽게 이해할 수 있도록 지원함
(Joint research and development of a novel, highly accurate, scalable, multimodal AI deepfake detection technology that identifies deepfake manipulations in video, voice and text inputs, and provides intuitive explanations of authenticity using bounding boxes, captions, confidence scores, etc. to help users easily understand the detected manipulations.)
- 한국, 싱가포르의 언어, 방언, 사회·문화적 및 지역적 환경적 특성을 반영한 대규모 데이터셋을 구축하고, 이를 활용하기 위한 데이터 파이프라인을 개발하여 딥페이크 탐지 모델을 훈련·유지 관리함
(Building large-scale datasets of Singapore's languages, dialects, and socio-cultural and regional environmental characteristics, as well as developing the required data pipelines to train and maintain the deepfake detection model(s).)
- 한국 및 싱가포르의 상황과 밀접하게 관련된 최소 3가지 이상의 상이한 산업 사례(예: 소셜 미디어, 가짜뉴스, 음성 생체인식 등)에 딥페이크 탐지 기술을 적용하며, 각 응용 사례마다 산업체 및 공공기관 파트너가 한 곳 이상 참여하여 기술을 실제로 활용하거나 상용화하는 데 기여함
(Application of the deepfake detection technology to at least 3 distinctly different industry use cases (e.g., social media, news outlets, voice biometrics etc.) relevant to the Singapore context, with each application supported by at least one or more industry and/or agency partners committed to using and/or commercializing the technology.)
- 최신 생성형 AI 모델뿐 아니라 연구팀이 제안한 활용 사례와 연관된 산업 데이터셋을 활용하여 딥페이크 탐지 기술을 벤치마킹하고 검증함으로써, 현존하는 최고 수준의 성능과 비교·평가함
(Benchmarking and validation of deepfake detection technology with prevailing state-of-the-art generative AI models, as well as with industry datasets relevant to the application use case proposed by the research team.)
- 한국과 싱가포르 연구자들이 공동으로 시제품 개발, 검증 및 테스트베드 구축
(Minimum Viable Product(s) (MVP) that is/are jointly developed, validated and testbedded by Singapore and South Korean researchers.)

ㅇ **(특이사항)**

- 연구개발계획서에 아래 내용을 필수로 포함하여 구체적인 양국협력 방법에 대해 제안
  · 상대국과 공동 개발 정보/결과/데이터/방법론 공유, 연구자 교류 및 방문, 공동 워크숍 개최, 공동 테스트베드, 저명한 국제 컨퍼런스 또는 저널에 논문 공동 게재의 방식을 통한 협력

  (Proposals should describe joint research dissemination activities including joint research publications in renowned international conferences and journals. Furthermore, proposals should include description of specific cooperation activities to be carried out such as exchange of information and results, sharing of data, sharing of methodologies, researcher exchanges and visits, joint workshops, joint testbeds etc.)

- R&D 성공을 확인하기 위한 개발기술의 수준 또는 성과목표·지표(예, 공동 논문 또는 공동실증 수행계획 등)를 제시할 것

  (Propose the level of technology developed or performance goals and metrics to confirm R&D success)

- 국제공동 컨소시엄의 조직별/부서별/개인별 역할과 협력분야를 명확히 제시하고 일정별 마일스톤을 제시할 것

- 국제공동연구 과정의 성과를 나타낼 수 있는 방법과 계획을 제시(예, 국제전문가 협의체구성/운영 계획, 국제공동 컨퍼런스 및 네트워킹 계획, 지속적 네트워킹을 위한 전략 및 계획 등)할 것

- 국제공동연구 결과에 대하여 참여국(또는 참여기관)별 활용방안 명시(기술이전 계획 또는 후속연구 등)할 것

2. 현황 및 필요성

ㅇ **(기존 기술현황)** 딥러닝 모델, 포렌식 분석, 생체신호 분석 및 워터마킹 기술 등이 딥페이크 조작영상 탐지에 이용되고 있음

- (딥러닝 기반 모델) CNN, RNN 등의 딥러닝 모델을 사용해 영상에서 미세한 조작 흔적을 탐지함

  *CNN: Convolution Neural Network(합성곱 신경망), RNN: Recurrent Neural Network(순환 신경망)

- (포렌식 분석) 픽셀 간 불일치, 조명 불일치 등 디지털 영상의 내부 요소를 분석하여 딥페이크 여부를 파악함

- (생체 신호 분석) 눈 깜박임, 미세한 얼굴 근육 움직임 등의 비정상적인 생체 신호를 통해 영상의 진위를 구별함

- (워터마킹) 원본 영상에 디지털 워터마크를 삽입하고, 이를 통해 딥페이크 영상에서의 변조 여부를 확인함

ㅇ **(한계점)** 기존의 딥러닝 기반의 딥페이크 탐지 기술로는 고도로 정교하게 제작된 딥페이크 상의 미세한 픽셀 왜곡이나 얼굴 특징의 비정상성을 포착하기 어려우며,

실시간 탐지에 한계가 있음

- 딥페이크 기술은 매우 빠르게 발전하고 있으며, 특히 최신의 GAN* 기반 생성 모델은 매우 사실적이어서 기존의 탐지 방법으로는 쉽게 구분되지 않는 경우가 많음

*GAN: Generative Adversarial Network(생성적 적대 신경망)

- '블랙박스' 형태인 기존의 딥러닝 기반 탐지 모델은 왜 특정 영상을 딥페이크로 판단했는지에 대한 설명이 불가함
- 딥페이크 탐지 기술이 발전함에 따라 딥페이크 생성자들도 영상에 노이즈를 추가하는 등 탐지 기술을 회피 및 무력화하려는 시도가 늘어나고 있음

ㅇ **(필요성)**

- **(R&D의 전략적 필요성 및 시급성)** 매우 복잡한 고난이도의 고정밀 딥페이크 탐지 기술개발을 위해서는 인공지능, 보안 및 영상 기술 등 여러 분야의 협력이 필요하므로, 각국의 독립적인 연구개발보다는 국제공동연구를 통한 연구 효율성 향상 및 시간 단축 필요
  - 딥페이크 영상의 주요 유통 경로인 유튜브, 페이스북 등의 글로벌 플랫폼에 대해 각국이 개별적으로 대응하는데 한계가 있어, 국제공동연구를 통해 글로벌 플랫폼과 협력할 수 있는 기술적, 법적 기반을 마련해야 함
- **(정부지원 필요성)** 빠르게 발전하며 사실성을 더해가는 딥페이크 기술은 개인의 사생활 침해를 넘어, 군사·외교적 가짜뉴스, 정치적 선전 등에 이용되어, 해당 국가뿐만 아니라 국제적인 혼란을 초래할 수 있으므로 정부의 전략적 지원 필수
- **(국제공동연구 필요성)** 멀티모달 AI 기반의 고정밀 딥페이크 탐지 기술개발에는 대용량의 딥페이크 데이터셋이 요구되고 다방면의 높은 기술적 난이도를 갖기 때문에, 국제적 공조를 통한 빠른 연구개발이 효과적임
  - 미국 캘리포니아 대학교 버클리의 GetReal Labs는 딥페이크 콘텐츠 탐지 분야를 주도하며, 이미지, 비디오, 오디오 콘텐츠에서 AI 생성 여부를 실시간으로 탐지하고 설명할 수 있는 고급 도구를 개발하고 있음
  - 미국 MIT 미디어랩은 딥페이크 영상에서 얼굴, 눈, 조명 불일치 등의 미세한 변조를 감지하고 직관적인 도구로 제공하는 DetectFakes* 프로젝트를 수행함

*출처: https://www.media.mit.edu/projects/detect-fakes/overview/

3. 수요분석

ㅇ **(주요 수요처)** 딥페이크 탐지 기술은 검찰, 경찰 등의 법 집행 기관, 소셜 미디어 플랫폼, 미디어 기업 및 사이버보안 기업 등에서 이용 가능

- 법 집행 기관: 범죄 수사 및 디지털 증거 보호를 위해 필요
- 정부 및 공공 기관: 공공 안전과 국가 안보를 위한 딥페이크 방지
- 소셜 미디어 플랫폼: 가짜 콘텐츠 유포 방지를 위해 필요
- 언론 및 미디어 기업: 신뢰할 수 있는 콘텐츠 검증을 위해 필수적

- 사이버 보안 기업: 고객 보호 및 기업 이미지 관리를 위한 수요

○ **(협력방안)** 국제공동연구를 통해 고정밀 딥페이스 탐지 기술개발, 딥페이크 데이터베이스 구축, 표준화 및 정책 협력 가능
- 공동연구 프로젝트: 국제공동연구를 통한 딥페이크 기술 분석 및 대응 솔루션 연구
- 표준화 협력: 딥페이크 탐지 기술의 글로벌 표준 개발 및 적용
- 데이터베이스 구축: 딥페이크 데이터 세트 및 사례 공유를 통한 효율적 대응
- 정책 협력: 글로벌 대응 전략 마련을 위한 법적/윤리적 가이드라인 공동 개발

4. 기대 효과

○ **(해결하고자 하는 이슈/문제)** 첨단 인공지능 기술을 주도하는 미국의 대학, 연구소의 데이터셋 등의 자원과 전문 지식을 활용하여 딥페이크 탐지 연구개발 효율성 및 혁신성 증대, 범국가적 협력을 통해 딥페이크 악용에 대한 신속한 탐지 체계 구축

○ **(이슈/문제 해결시 발생할 경제적 사회적 파급효과)** 딥페이크 불법 영상을 실시간으로 탐지하여 사용자가 불법 여부를 쉽게 확인할 수 있도록 하여 미디어 서비스의 신뢰성 제고

○ **(글로벌 표준화 촉진)** 국제협력을 통해 딥페이크 대응 기술의 글로벌 표준을 확립하여 전 세계적 대응력 강화

5. 개발기간/예산/추진체계

○ **연구개발기간** : 3년 이내

○ **정부지원연구개발비** : '25년 2.71억원 이내(총 정부지원연구개발비 16.23억원 이내)

구분	기간	개월수	정부지원연구개발비
1년차	'25.9월~'26.2월	6개월	271 백만원 이내
2년차	'26.3월~'26.12월	10개월	451 백만원 이내
3년차	'27.1월~'27.12월	12개월	541 백만원 이내
4년차	'28.1월~'28.8월	8개월	360 백만원 이내
합계	-	36개월	1,623 백만원 이내

* 상기 정부지원연구개발비는 한국측 연구개발기관에게 지원되는 금액임
* 연차별 정부지원연구개발비는 당해연도 예산심의결과에 따라 변동될 수 있음

○ **주관기관** : 제한없음

○ **추진체계** : 동 과제는 **별도과제형**에 해당

※ (참고) 국제공동연구 추진유형

해당	추진유형	주요내용
	일반형	국내 주관연구개발기관이 국제공동연구개발비를 활용하여 외국소재기관과 공동으로 연구를 수행하는 방식
	공동기관형	해외기관이 국내 연구개발기관과 연구개발과제를 공동으로 수행하기 위해 주관 또는 공동연구개발기관으로 참여하는 방식
√	별도 과제형 (Joint call)	국내기관과 해외기관이 한 컨소시엄을 이루어 공동연구를 추진하되, 연구개발비 집행 등은 독립된 과제로 수행하는 형태를 의미

연구유형	기초연구 ( ), 응용연구 (√), 개발연구 ( )	TRL (4)~(6)단계
과제특징	경쟁형( ), 경쟁형(챌린지)( ), SW자산뱅크등록( ), 공개SW( ), 기술료비징수( ), 국제협력R&D(√), 정책지정( ), 혁신도약형( ), 표준화연계( ), 사회문제해결형( ), 일자리연계( ), 소재부품장비( ), 규제샌드박스( ), 연구데이터공개( ), 사업화연계( ), IP-R&D연계( )	

구분	기술분야명/팀명	성명
책임PM(과제기획위원장)	혁신·글로벌	김욱
담당 팀장	글로벌협력팀	임종석