OPEN 출처표시 상업용금지 변경금지
공공누리 공공저작물 자유이용허락

문화정보 이슈리포트
2023-9호(제49호)

실감 디지털 휴먼 기술과 공공분야 활용 방안

이승욱 책임연구원
김기남 선임기술원
김태준 선임연구원
윤승욱 책임연구원
임성재 책임연구원
황본우 책임연구원
(한국전자통신연구원 콘텐츠연구본부)

KCISA 한국문화정보원

CONTENTS

문화정보 이슈리포트
2023-9호(제49호)

실감 디지털 휴먼 기술과 공공분야 활용 방안

이승욱(한국전자통신연구원 콘텐츠본부)
김기남(한국전자통신연구원 콘텐츠본부)
김태준(한국전자통신연구원 콘텐츠본부)
윤승욱(한국전자통신연구원 콘텐츠본부)
임성재(한국전자통신연구원 콘텐츠본부)
황본우(한국전자통신연구원 콘텐츠본부)

요약

최근 연산 처리에 특화된 GPU의 발전과 구글이 개발한 트랜스포머 알고리즘, 그리고 대규모 학습 데이터를 통한 ChatGPT의 보편화는 디지털 콘텐츠 분야 전반에서 생성형 AI 기술의 발전을 촉진하고 있다. 이러한 추세에 발맞추어, 디지털 휴먼 분야에서도 생성형 AI를 활용한 실감 디지털 휴먼의 개발이 진행되고 있다. 본 기고는 생성형 AI에 의해 향상된 실감 디지털 휴먼 기술의 최신 동향을 검토하고, 이 기술이 공공 부문에서 어떻게 활용될 수 있는지 탐구하며, 문화, 체육, 관광 분야에 미치는 영향과 시사점에 대해 논의하고자 한다.

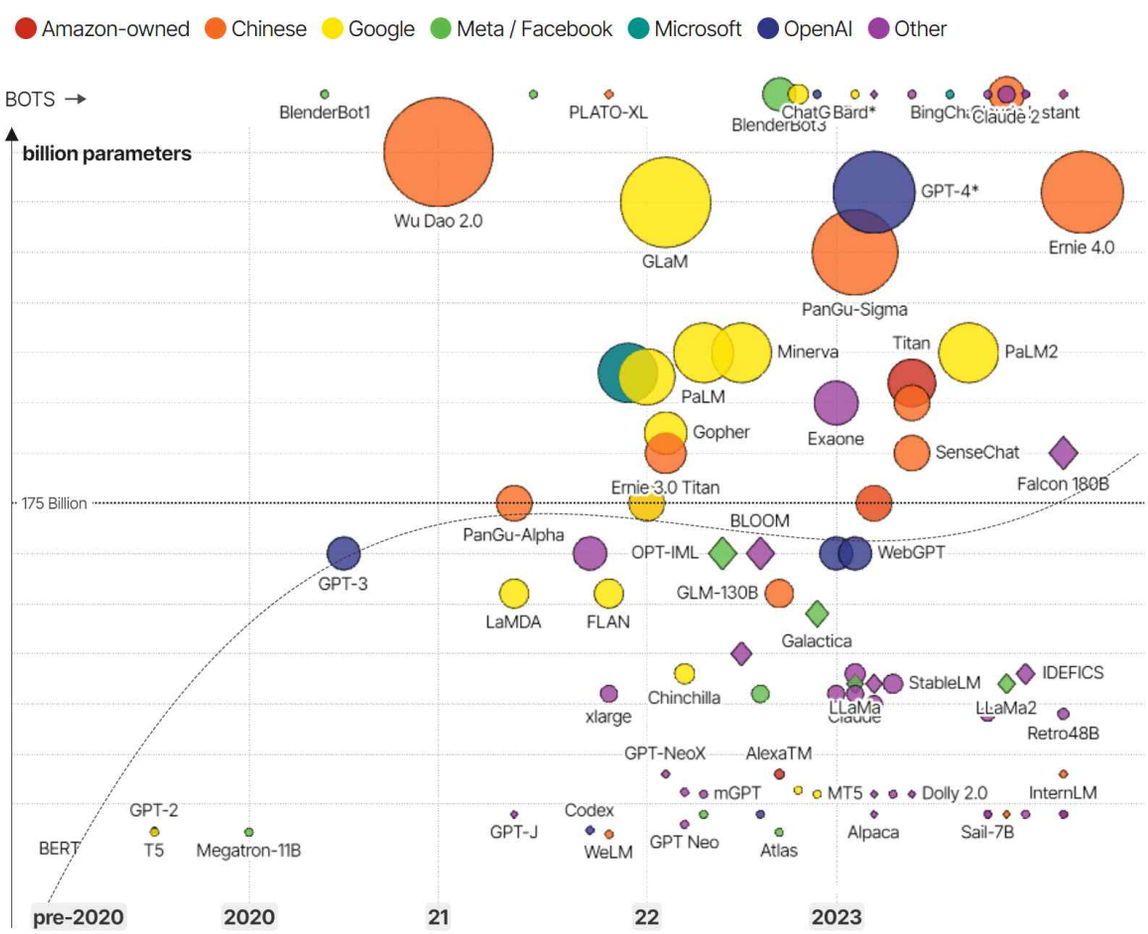
※ '문화정보 이슈리포트'의 내용은 작성자의 의견으로 한국문화정보원의 공식적인 입장과 다를 수 있습니다.
※ 이 논문은2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00193, 실사 수준의 디지털 휴먼 생성과 30fps급 실감 렌더링 기술 개발)

KCISA 한국문화정보원

I. 들어가며

○ OpenAI에서 개발한 ChatGPT로 인한 생성형 AI 기술 개발 가속화

- 생성형 AI란 이용자의 특정 요구사항에 따라 결과물을 생성하는 인공지능의 일종, 일반적으로 프롬프트에 대응하여 텍스트, 이미지, 기타 미디어를 생성할 수 있는 일종의 인공지능(AI) 시스템
- 기존에도 생성형 AI 관련된 많은 기술이 있었으나, ChatGPT의 대중화로 사람의 관심을 받기 시작하며 이후 Runway, Stability AI 등에서 만들어진 영상 생성 인공지능 Stable Diffusion[1]으로 관심 가속화
- 구글에서 개발한 Transformer와 방대한 학습데이터, 그리고 Nvidia의 고성능 컴퓨팅 장치 등으로 인해 LLM(Large Language Models)을 넘어 LMM(Large Multimodal Models)로 확장 중



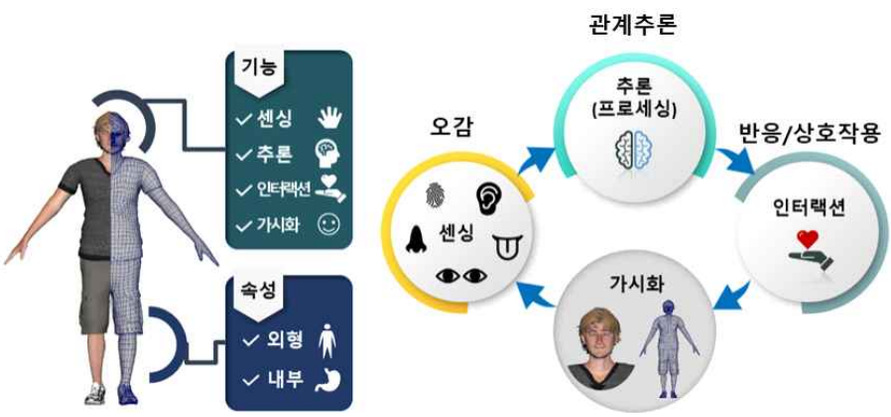
〈그림 1〉 인공지능의 부상과 성장: 대규모 언어 모델(The Rise and Rise of AI: LLM)

자료: David McCandless, LifeArchitect.ai, '23.11.

1) 50억장의 텍스트-이미지 쌍으로 학습된 모델로, 텍스트를 입력하면 관련된 영상을 생성하는 기술

○ 이에 따라 실감 디지털 휴먼 영역에도 다양한 인공지능 기술이 접목되기 시작 함

- 디지털 휴먼은 그림 2와 같이 사람의 디지털 버전으로 생각할 수 있으며, 기술적인 측면에서는 사람처럼 보이고, 듣고, 말하고, 생각하는 가상의 디지털 존재
- 디지털 휴먼 기술을 기능과 속성으로 분류 가능, 속성은 내/외부로 보여지는 데이터이며, 기능은 인간 뇌의 활동을 모방하는 지능을 포함한 상호작용 및 가시화를 포함
  - Transformer 대표되는 언어모델: 디지털 휴먼의 추론, 반응/상호작용(음성)과 관련된 역할 담당
  - GAN, diffusion 등의 영상 생성 모델: 디지털 휴먼의 속성 가시화 역할 담당
  - 영역 분리 등의 영상 인식 모델: 디지털 휴먼의 오감 센싱 역할 담당
  - 모션 생성 기술: 디지털 휴먼의 움직임 가시화 역할 담당



〈그림 2〉 디지털 휴먼의 기능과 속성 - 인공지능이 적용 가능한 분야

자료: 한국전자통신연구원 자체 제작

Ⅱ. 실감 디지털 휴먼 기술 개요

1. 언어 생성 기술

○ 초기 언어 모델은 RNN, LSTM등을 거쳐 고정된 컨텍스트 벡터를 사용하는 Seq2Seq 방식으로 진화, 이후 단일 컨텍스트 벡터 사용의 한계를 극복하기 위해 각각의 단어 간의 연관관계를 내적으로 정의하는 Attention 알고리즘 도입, 이후 Attention을 이용한 Transformer 도입 후 대부분의 언어생성 기술은 Transformer를 활용

- 트랜스포머는 self-attention(입력 시퀀스의 모든 원소들 간의 상호작용을 고려하여 각 원소에 가중치를 부여), positional encoding(입력 데이터의 위치 정보를 인코딩하여 모델에 제공), 스택 구조(여러 개의 인코더와 디코더 층을 스택으로 적층), 멀티-헤드 어텐션(여러 개의 어텐션 분포를 학습하여 다양한 특징을 동시 취득) 및 정규화를 통해 높은 병렬화와 함께 문장의 장거리 의존성을 효과적으로 학습하는 방법을 제공
- ChatGPT는 기본적으로 트랜스포머와 RLHF(Reinforcement Learning from Human Feedback)의 방식으로 자연스러운 문장생성이 가능하게 하였고, 이를 통해 디지털 휴먼은 사람처럼 대화할 수 있게 됨
- 최근에는 파인 튜닝 혹은 프롬프트 확장 등의 방법을 통해 사람과 같은 인격을 부여하여 나만의 디지털 휴먼을 생성할 수도 있음

2. 영상 생성 기술

○ 기존의 전통적인 방법은 그래픽스 기술을 이용하여 모델링, 텍스처링, 리깅, 애니메이션, 렌더링 등의 기법으로 디지털 휴먼 제작, 최근은 인공지능 기술을 이용하여 사용자에게 렌더링된 영상을 직접 제공

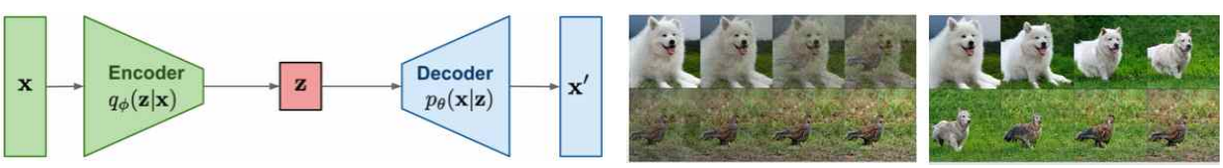
- 전통적인 방법으로 만든 3D 데이터는 유니티, 언리얼 등을 통해 2D로 렌더링되어 사용자에게 보여짐, 생성형 AI는 3D 모델을을 거치지 않고 사용자에게 2D로 렌더링 된 영상을 바로 제공

○ 영상 생성 AI의 획을 그은 GAN(Generative Adversarial Network)은 생성자 (Generator)와 판별자(Discriminator) 두 개 모델을 동시에 학습하여 더 실감나는 가짜 영상 생성

- 초기('14~'16년) 작은 크기의 영상을 생성하는 Auxiliary Classifier GAN 개발
- 중기('17~'18년) 고해상도의 영상을 만들기 위해 Progressive GAN, BigGAN 등이 개발되었고, 영상의 스타일을 학습 및 제어할 수 있는 StyleGAN이 개발
- 이후 ('19~'20년) 적은 학습 데이터를 사용 연구 진행 및 StyleGAN2, 판별자의 과적합을 방지하는 differentiable augmentation 기술 등이 개발, 최근의 연구 동향은 디퓨전이 주류를 이룸

○ VAE(Variable Auto-Encoder)는 그림 3과 같이 입력 데이터를 받아 그 데이터의 잠재적인 확률적 표현(정규분포의 평균과 분산)으로 맵핑하는 인코더와, 인코더에서 얻은 잠재적 표현을 원본 데이터로 재구성하는 디코더 부분으로 구성

- 그림 3과 같이 입력 데이터 x를 인코딩하여 잠재 공간의 확률 분포로(z) 만들고, 이를 디코딩하여 입력과 최대한 같은 출력 데이터 x'를 생성
- 주요한 특징으로 VAE는 잠재 공간이 부드럽게 구성됨에 두 점 사이의 경로를 따라 샘플링할 때 다른 특성을 가지는 영상으로 변경이 가능. 즉 잠재공간에서 영상을 합성하면 픽셀 영역에서의 합성과는 다르게 부드러은 변형으로 합성된 샹상을 생성할 수 있음
- 이러한 특성으로 VAE는 프롬프트 기반으로 영상을 생성하는 기술에 활용



[VAE의 인코더 및 디코더] [픽셀 공간(좌) 및 잠재 공간(우)에서의 합성]

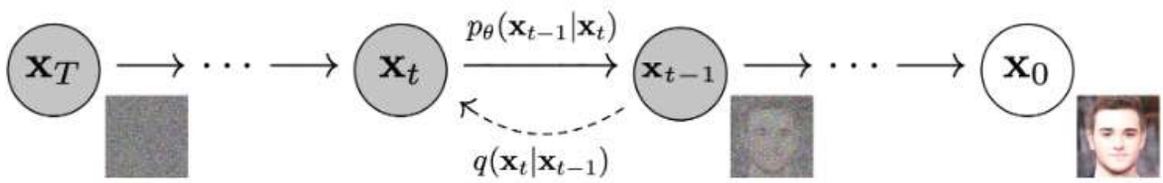
〈그림 3〉 잠재공간에서 영상 생성이 가능한 VAE

자료: 백지오, 잠재 공간(Latent Space)란 무엇인가?, https://www.youtube.com/watch?v=gJ86ixUx6MU, 자료 재구성

○ 디퓨전은 그림 4와 같이 원본 영상에 잡음을 단계적으로 추가하여 잡음이 섞인 영상을 만드는 디퓨젼 과정과 잡음을 제거하여 원본 영상을 만드는 디노이징 과정으로 구성

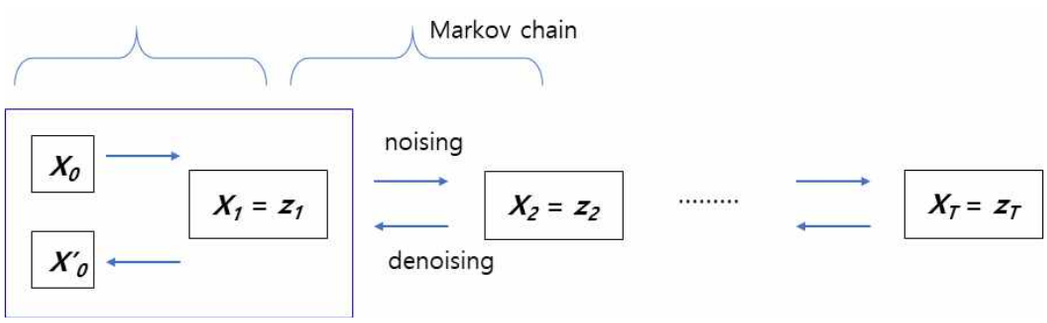
- 초기 입력 X_0에 t번 잡음을 더해 Xt를 생성하는 과정은 단순히 수학적으로 계산되는 과정이며, 잡음을 제거하는 디노이징 과정을 학습으로 재현
- 그림 5에 따라 디퓨전을 1단계만 진행하면 VAE와 같은 과정이 됨, 즉 디퓨전은 "Markov Chain Process for Multiple Latent Variable"로 정의
- 그림 5에서 파란색 사각형이 VAE가 됨, 입력 X_0에 노이즈를 더하여 잠재공간의 값 z_1을 만들고(인코더), 이를 디노이징하여 원본과 최대한 비슷한 X_0'을 복원(디코더)
- 이런 과정을 T번 반복하는 것이 디퓨전이며, 따라서 학습 과정과 손실함수 등은 VAE와 비슷하게 정의

- 순차적으로 잡음을 추가하고 제거하는 과정을 DDPM(Denoising Diffusion Probabilistic Models)이라 하며, Markov Chain에 따라 잡음 추가/제거 작업을 진행, 매 단계에서 모델을 학습하고 샘플링하기에 시간이 오래 걸림
- DDIM(Denoising Diffusion Implicit Models)은 X_t가 이전의 X_{t-1}과 X_0에 의해 결정되는 Non-Markov Chain 프로세스이며 빠른 영상 생성 가능



〈그림 4〉 디퓨전과 디노이징으로 구성된 디퓨전 모델

자료: Jonathan H, et. Al., "Denoising Diffusion Probabilistic Models," arXiv:2006.11239v2



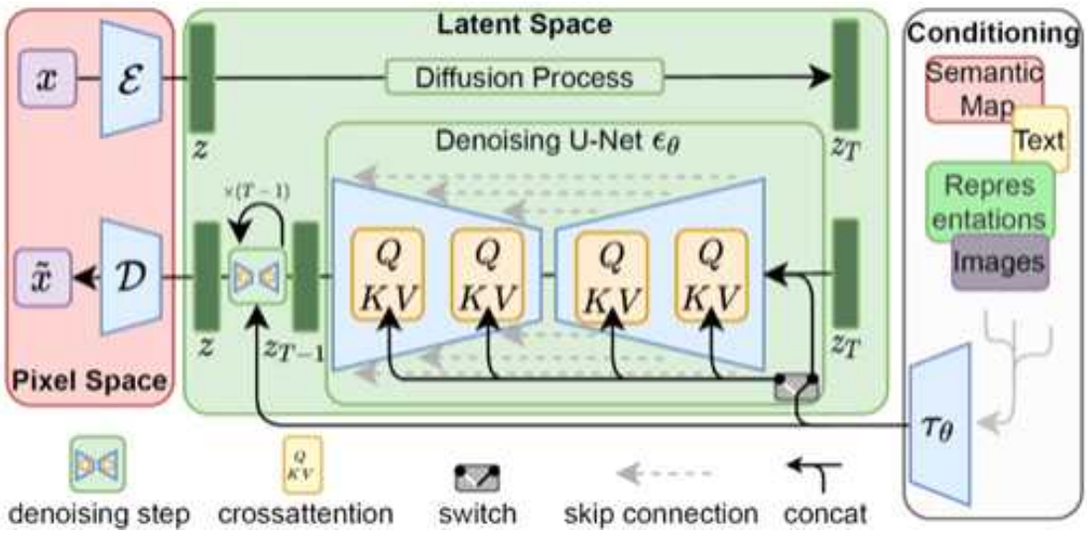
〈그림 5〉 디퓨전과 VAE의 비교

자료: 한국전자통신연구원 자체 제작

○ Latent Diffusion 모델은 앞서 설명한 디퓨전을 잠재공간에서 수행한 모델로, 잠재공간의 특성상 적용 데이터로 처리가 가능하여, 대규모 데이터를 학습하여 영상 생성 가능

- 디퓨전은 좋은 성능을 보임에도 불구하고 과다한 계산량으로 실제 적용에 어려움(계산량이 많기에 다양한 영상을 동시에 학습할 수 없어서 다양한 영상을 생성하지 못함)
- Latent Diffusion은 기존의 픽셀 공간에서 영상을 확장시키는 것이 아니라, 잠재공간에서 확산시키는 모델로 픽셀공간 대비 잠재공간이 규모가 적기에 실제 디퓨전 및 디노이징 시간도 적게 소요되는 장점
- Stable Diffusion은 Stablity.AI의 컴퓨팅 환경을 이용하여 Runway가 58억 장의 텍스트-이미지쌍을 CLIP[12]으로 학습하여 만든 모델
- Stable Diffusion은 사용자의 텍스트 입력을 CLIP을 통해 인코딩 후 그림 6의 녹색 부분에 의해 U-net, 텍스트 임베딩에 따라 조건화되어 여러 번 반복하여 디노이징 하는 과정 통과, 이후 잠재공간에서의 값이 복원되고, 그림 6의 붉은색처럼 VAE의 디코더에 입력되어 최종 영상을 출력

- Latent Diffusion(Stable Diffusion) 모델에서는 U-Net에서 학습되고 이를 바탕으로 생성된 값 자체가 디퓨전 모델처럼 영상 픽셀값이 아니고, VAE에 의해 인코딩된 잠재 벡터를 U-Net에 학습해 주었기 때문에 U-Net에서 복원되어 나온 저해상도의 잠재 벡터를 VAE 디코더로 디코딩하여 고해상도의 그림으로 만들어 주는 것



〈그림 6〉 잠재공간에서의 디퓨전 프로세스

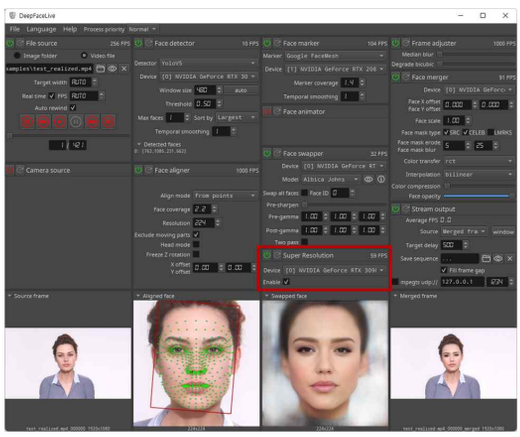
자료: Robin Rombach, et. al., "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv:2112.10752v2

Ⅲ. 실감 디지털 휴먼 활용 방안

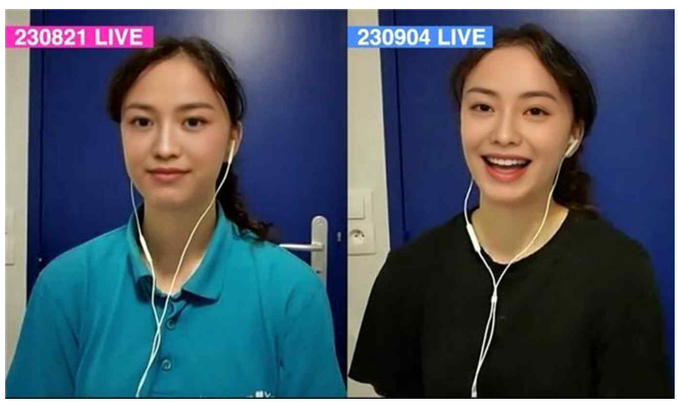
1. 실시간 실감 방송

○ 한국전자통신연구원은 펄스나인과 함께 아리랑국제방송에 페이스 스왑 기반 실시간 실감 디지털 휴먼 서비스 수행('23.09)

- 아리랑 국제방송의 글로벌 오디션 프로그램 '코드네임 부산'은 오디션 참가자들이 펄스나인이 개발한 페이스 스왑 기술을 적용하는 것
- 한국전자통신연구원이 개발한 기술은 생성형 AI 기반 실감 가시화 기술을 이용하여 생방송으로 진행되는 프로그램의 해상도 고도화 및 실감화를 실연
- 한계에 부딪힌 디지털 휴먼의 해상도를 실제 사람 수준으로 높여 한계를 뛰어 넘는 기술(기존의 생성형 AI의 시간축 떨림 기술 해결)
- 그림 7의 좌측은 DeepFaceLive기반의 페이스 스왑 기술에 실감 디지털 휴먼 기술을 이용한 업스케일링 기술을 적용한 예제, 붉은색 사각박스가 실감 업스케일링 기술을 적용하는 UI, Enable 체크 버튼을 클릭하면, 페이스 스왑된 영상을 실시간 실감 업스케일링 수행
- 그림 7의 우측은 아리랑국제방송에 딥페이크 기술과 실감 디지털 휴먼이 적용된 결과물, 기존의 저해상도 페이스 스왑 결과물(좌)을 실시간 업스케일링하여 고해상도로 변환(우)



실감 업스케일링 기술이 적용된 DeepFaceLive



아리랑국제방송 적용 - 코드네임 부산

〈그림 7〉 페이스 스왑 기술과 실감 업스케일 기술의 적용과 아리랑국제방송 적용

자료: 한국전자통신연구원 자체 제작(DeepFaceLive 수정) 및 아이랑국제방송-코드네임부산

2. 2D 영상기반 실사 디지털 휴먼

○ 기존의 3D 모델링 방식이 아닌 2D 실제 바디 기반 디지털 휴먼 제작 방식으로 전체를 3D로

제작하는 방식보다 저렴하여 최근에 대중화 됨

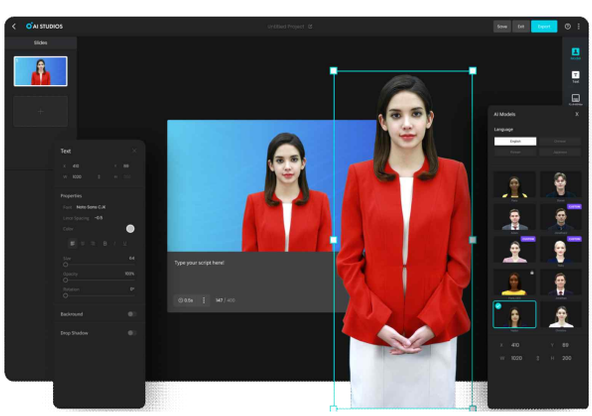
- 타겟 영상은 실제 촬영한 사람의 바디 영상을 활용하고, 소스 얼굴은 다른 캐릭터를 활영하여 타겟 영상에 합성
- 소스 영상은 (1)3D 모델링으로 만들거나, (2)실제 사람의 얼굴을 활용하거나, (3)생성형 AI를 이용하여 기존에 없는 사람을 만들어냄

○ 로커스엑스에서 제작한 로지는 (1)의 방식으로 구현 됨, 신한라이프 광고에 사용된 로지는 실제 사람을 촬영한 영상에 3D 모델로 제작된 얼굴을 합성하여 만든 영상

- 신한라이프 광고에서 로지의 의상이 3번 교체되며, 이를 고품질로 구현하기 위해서는 수억원[2]의 비용 발생, 이를 해소하기 위해 실제 촬영 데이터 활용
- 로지는 이후 음악, 드라마, 홍보 브랜드, 홍보대사 등의 다양한 영역에서 활동

○ 딥브레인 AI는 (2)의 방식으로 디지털 휴먼 제작 서비스 진행중이며, 광고, 키오스크, 교육, 뉴스 앵커 등 다양한 영역에서 활용

- AI 스튜디오를 클릭 몇 번만으로 자연스러운 영상 제작 가능하고, 온오프라인 적용이 가능한 대화형 AI 휴먼 서비스
  - 메타버스를 위해 제작되는 몰입형 3D 디지털 휴먼을 통해 금융, 교육, 미디어 등의 분야에 적용 가능



딥브레인 AI의 AI Studios

딥브레인 AI의 AI Kiosk

〈그림 8〉 딥브레인 AI의 실감 디지털 휴먼 서비스

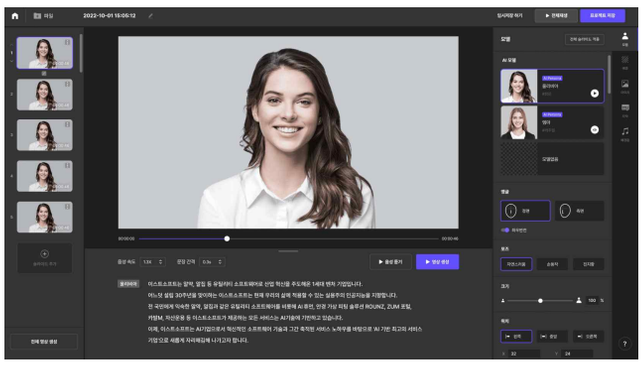
자료: 딥브레인 AI

○ 이스트소프트는 (2)의 방식으로 디지털 휴먼 제작 서비스 진행중이며, 광고, 키오스크, 교육, 뉴스

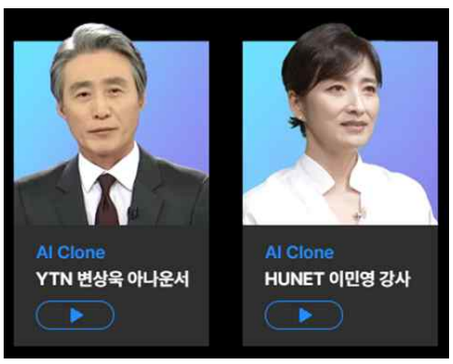
2) 가상 인간 로지의 비밀을 공개합니다, 인공지능 신문, '21.08

앵커 등 다양한 영역에서 활용

- ChatGPT와 연계된 AI 스튜디오 PERSO를 개발하여 실존하는 YTN 변상욱 아나운서, HUNET 이민영 강사 등의 다양한 모델을 이용한 실감 디지털 휴먼 사업 진행
  - 딥러닝 학습을 통해 텍스트를 음성으로 생성 및 재생하는 목소리 생성 AI 기술과 발음에 따른 입모양을 딥러닝으로 학습하고 말하는 얼굴 영상을 생성하는 얼굴생성 AI 기술 보유, 이를 통해 (3)의 방식으로 서비스 가능



이스트소프트의 AI 스튜디오 PERSO



이스트소프트의 실감 디지털 휴먼

〈그림 9〉 이스트소프트의 실감 디지털 휴먼 서비스

자료: 이스트소프트

Ⅳ. 시사점

○ 현재의 실감 디지털 휴먼은 생성형 AI와 함께 발전하고 있으며, 기존의 전통적인 컴퓨터 그래픽스 기반의 제작 방식을 바꿀 수 있음

- 실감 나는 외형뿐만이 아니라, 거대 언어모델과 연계되어 실제 사람처럼 지능을 가지는 디지털 휴먼이 등장할 수 있음
- 생성형 AI를 통해 고품질 디지털 휴먼의 제작비용이 절감될 것으로 기대되어 다양한 형태의 서비스 등장 예상
- 인공지능 기술의 발전으로 딥페이크 등의 기술에 대한 접근성이 낮아져 개인정보 및 프라이버시 침해, 사기 및 범죄 활용, 사회적 혼란과 오해, 디지털 증거의 신빙성 저하 등의 다양한 문제가 발생 할 수 있음

○ 이에 생성형 AI 기반 실감 디지털 휴먼 기술을 통해 문화체육관광 분야에서의 시사점을 제언하면 아래와 같음

- 생성형 AI와 디지털 휴먼 기술에 대한 표준화 및 법적 기준 마련, 윤리적 사용을 위한 정책 개발 지원
- 문화 분야의 다양한 데이터 수집, 연계 및 개방 촉진을 통해 AI 기반 문화 콘텐츠 개발의 기반 마련
- 생성형 AI 및 디지털 휴먼 관련 연구개발 프로젝트 지원과 혁신적인 문화 콘텐츠 개발 촉진, 특히 문화, 관광 해설을 지원하는 큐레이터 디지털 휴먼 개발
- 문화 서비스에 AI와 디지털 휴먼을 통합, 사용자 경험 개선 및 접근성 증대
- 민간 기업 및 연구 기관과의 협력 강화, 산업 간 시너지 창출
- 생성형 AI 및 디지털 휴먼 기술의 윤리적, 법적 측면을 연구하고, 관련 규제 및 법률 개발
- 실감 디지털 휴먼 기술들이 사회에 미칠 영향을 평가하고, 부정적인 영향을 최소화하는 전략 수립