한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 30, No. 3, P. 149~158

시네마픽 : 생성형 AI기반 영화 컨셉 포토부스 시스템*

정석현◦ † 임승규 † 이정진 ‡

숭실대학교 글로벌미디어학부

{jsh19444, paulleem}@soongsil.ac.kr, jungjinlee@ssu.ac.kr

CINEMAPIC : Generative AI-based movie concept photo booth system

Seokhyun Jeong◦† Seungkyu Leem † Jungjin Lee‡

The Global School of Media, Soongsil University

요 약

오프라인에서 사진을 촬영하는 포토부스는 자신이 원하는 포즈와 소품 등을 통해 자연스럽게 나다운 모습을 촬영할 수 있으며, 함께한 사람들과 추억을 공유하는 특별한 경험을 선사한다. 최근 다양한 표현을 가능하게 하고자 생성형 AI를 활용한 포토부스 사례들이 등장했다. 그러나 기존 AI 포토부스는 단체 사진 촬영이 불가능하고, 대부분 사용자의 포즈를 반영하지 못하며, 개별 인물마다 다른 컨셉을 적용하기 어려운 한계가 존재한다. 본 연구는 이러한 문제를 해결하여 사용자가 자유롭게 포즈와 위치, 컨셉을 선택하여 촬영할 수 있는 AI 포토부스 시네마픽을 제안한다. 인물별 개별 컨셉 적용을 위해 개별 생성 워크플로우를 전처리, 생성, 후처리 세 단계로 설계하고, 이를 실제 프로토타입으로 구현했다. 이 과정에서 인물별 투명 이미지 생성, 배경 생성 후 합성시 발생하는 아티팩트를 줄이는 재생성 테크닉, 최적화 모델 적용 및 GPU 병렬화 등 다양한 방식을 워크플로우에 통합하여 한계점을 극복하였다. 사용자 품질 평가와 약 400명의 사용자를 대상으로 대규모 시범 운영을 통해 시스템의 효과성을 검증했다. 그 결과, 사용자들은 기존 방식에 비해 높은 선호도를 보였으며, 이를 통해 실제 포토부스로의 도입 가능성을 확인했다. 본 연구에서 제안하는 AI 포토부스 시네마픽은 더욱 창의적이고 차별화된 시장을 개척할 수 있을 것으로 기대하며, 앞으로 다양한 응용 분야에서 널리 활용될 것으로 기대된다.

Abstract

Photo booths have traditionally provided a fun and easy way to capture and print photos to cherish memories. These booths allow individuals to capture their desired poses and props, sharing memories with friends and family. To enable diverse expressions, generative AI-powered photo booths have emerged. However, existing AI photo booths face challenges such as difficulty in taking group photos, inability to accurately reflect user's poses, and the challenge of applying different concepts to individual subjects. To tackle these issues, we present CINEMAPIC, a photo booth system that allows users to freely choose poses, positions, and concepts for their photos. The system workflow includes three main steps: pre-processing, generation, and post-processing to apply individualized concepts. To produce high-quality group photos, the system generates a transparent image for each character and enhances the backdrop-composited image through a small number of denoising steps. The workflow is accelerated by applying an optimized diffusion model and GPU parallelization. The system was implemented as a prototype, and its effectiveness was validated through a user study and a large-scale pilot operation involving approximately 400 users. The results showed a significant preference for the proposed system over existing methods, confirming its potential for real-world photo booth applications. The proposed CINEMAPIC photo booth is expected to lead the way in a more creative and differentiated market, with potential for widespread application in various fields.

키워드: 생성형 AI, 디퓨전 모델, 포토부스

Keywords: Generative AI, Diffusion Model, Photo Booth

*corresponding author: Jungjin Lee/Soongsil University(jungjinlee@ssu.ac.kr)

Received : 2024.06.15./ Review completed : 1st 2024.07.02. / Accepted : 2024.07.05.
DOI : 10.15701/kcgs.2024.30.3.149
ISSN : 1975-7883(Print)/2383-529X(Online)

1. 서론

오프라인에서 사진을 촬영하고 인쇄해 추억을 남기는 포토부스는 단순한 유행을 넘어, 하나의 문화로 자리 잡고 있다 [1]. 무인으로 운영되는 포토부스는 사용자가 편한 분위기에서 자연스럽게 자신이 원하는 포즈를 취하여 나다운 모습을 촬영할 수 있으며, 다양한 소품과 프레임을 통해 개인의 정체성을 표현할 수 있는 기회를 제공한다 [2]. 또한, 포토부스에서 촬영된 사진은 즉시 인화되어, 함께한 사람들과 추억을 공유하는 특별한 경험을 선사한다. 이러한 이유로 포토부스는 폭발적인 인기를 끌며 다양한 브랜드와 매장이 생겨났지만, 프레임, 소품, 배경 색상 등의 요소를 제외하면 브랜드별로 차별화할 수 있는 포인트가 부족한 실정이다. 이로 인해 상단에서 촬영하는 하이앵글 컨셉이나, 프로젝터를 활용해 다양한 컨셉을 오버레이 하는 등의 차별화 시도가 이어지고 있다.

최근에는 이러한 한계를 극복하고 더 많은 표현이 가능하게 하고자 생성형 AI를 활용한 새로운 포토부스가 등장하고 있다. AI 포토부스는 내 얼굴의 특징을 유지한 채, 웹툰 스타일의 사진으로 바꿔주거나 [3], 운동선수의 모습을 하고 있는 사진으로 바꿔주는 등의 색다른 경험을 제공한다 [4]. 그러나, 이러한 AI 포토부스는 몇 가지 중요한 제한점을 가지고 있다. 첫째, 3인 이상의 단체 사진 촬영이 불가능하다. 둘째, 대부분의 경우 사용자의 포즈를 반영하지 않는다. 셋째, 인물마다 개별적인 컨셉을 적용할 수 없기 때문에 생성형 AI의 장점을 제대로 활용하지 못한다.

본 연구에서는 기존 AI 포토부스의 한계점을 극복하고, 일반 포토부스처럼 사용자의 자유로운 사용자 경험을 제공할 수 있는 AI 포토부스 시스템 시네마픽(CINEMAPIC)을 제안한다. 본 시스템은 사진 촬영이 가능한 키오스크와, 촬영된 사진을 실시간으로 특정 컨셉에 맞게 변환하는 이미지 생성 서버로 구성된다. 키오스크는 모니터, 카메라, 프린터 등으로 구성되어 있으며, 기존의 포토부스와 동일한 방식으로 촬영 및 즉시 인화가 가능하다. 이미지 생성 서버는 네트워크를 통해 실시간으로 촬영된 사진을 전달받아, 포즈 및 뎁스 추정, 인스턴스 분할, 개별 컨셉 적용 및 생성 단계를 거쳐, 최종적으로 재배치 및 배경 합성을 통해 영화 캐릭터별 특징이 잘 반영된 컨셉 이미지를 제작한다.

우리는 본 시스템의 프로토타입을 구현하고, 사용자 품질 평가와 대규모 시범 운영을 통해 시스템의 효과성을 검증하였다. 또한, 프로토타입 시스템을 통해 파악한 한계점을 바탕으로 이를 해결하기 위한 향후 연구 방향을 제시하였다.

본 논문은 크게 다음의 3가지 부분에서 기여점이 있다.

1. 생성형 AI를 활용하여 단체 사진 촬영이 가능한 AI 포토부스 시스템을 제안하였다.
2. 포즈 반영 및 인물별 개별생성을 통해 각각 다른 컨셉을 적용할 수 있는 워크플로우를 구현하였다.
3. 실제 사용자 약 400명을 대상으로 시범 운영을 통해 제안된 시스템 프로토타입을 검증하였고, 기존 연구들과의 질적 비교 평가를 통해 뛰어난 성능을 입증하였다.

*학부생 주저자 논문임
†Denotes Equal Contribution
‡Corresponding Author

2. 관련 사례 및 연구

2.1 Image-to-Image Translation

일반적으로, 이미지 간 변환(Image-to-Image Translation)의 목표는 입력 이미지를 원본 도메인에서 목표 도메인으로 변환하면서, 원본 이미지의 고유한 콘텐츠는 최대한 보존하는 것이다 [5]. 이러한 이미지 간 변환은 시맨틱 이미지 합성(semantic image synthesis), 도메인 적응(domain adaptation), 만화화(cartoonization), 스타일 변환(style transfer) 등 다양한 응용 분야에 폭넓게 사용되고 있다.

Mo *et al.* [6]는 사진 속 인물의 청바지를 치마로 변경하는 등 복잡한 변환 작업을 가능하게 하는 InstaGAN을 제안하였다. InstaGAN은 객체 분할 마스크와 같은 인스턴스 정보를 활용하여 다중 인스턴스 변환을 효과적으로 수행했다. 또한, Brooks *et al.* [7] 사용자의 텍스트 지시를 기반으로 객체 교체나 이미지 스타일 변경 등 이미지 편집을 수행하는 InstructPix2Pix 방법을 제안하였다.

최근에는 사전 학습된 디퓨전 모델(Diffusion Model) [8]을 기반으로 한 ControlNet [9]이 등장하였다. ControlNet은 Text-to-Image 생성에서 edge maps, human pose skeletons, depth, normal 등 다양한 추가 이미지를 컨디셔닝으로 활용하여, 생성 과정에서 공간적 조건을 제어한다. ControlNet을 사용하면, 원본 이미지의 특정 특징을 추출해 목표 도메인에 그 특징이 반영되도록 생성할 수 있다. 본 연구에서는 현재 입고 있는 옷이나, 다른 조건들은 무시하고, 사용자의 포즈와 체격, 얼굴만 반영하여 완전히 새로운 컨셉의 이미지를 생성하는 것을 목표로 한다. 따라서 Image-to-Image 생성방식이 아닌 ControlNet을 활용한 Text-to-Image 기반의 포즈 및 뎁스맵 이미지 조건부 생성을 수행한다.

2.2 Identity Preserving Image Synthesis

개인화된 이미지 생성은 하나 이상의 참조 이미지가 입력으로 주어졌을 때 개별 특성이 유지된 이미지 변형 생성을 목표로 한다 [10, 11]. 기존의 개인화된 이미지 생성 기술들은 미세 조정(fine-tune)이 필요한 방법과 그렇지 않은 방법으로 나눌 수 있다. DreamBooth [12]나 LoRA [13]와 같은 기술들은 사전 학습된 모델을 미세 조정하여 참조 이미지의 새로운 측면을 반영하지만, 이 과정은 자원 소모가 크고 시간 소요가 많아 실용성이 떨어진다. 반면, FastComposer [14]와 IPAdapter [15] 같은 기술들은 미세 조정 없이 참조 이미지의 특징을 추출하고 이를 이미지 생성 과정에 통합했다. 또한, InstantID [16]는 참조 이미지의 얼굴 정체성을 유지하기 위해 얼굴 이미지, 랜드마크 이미지, 텍스트 프롬프트

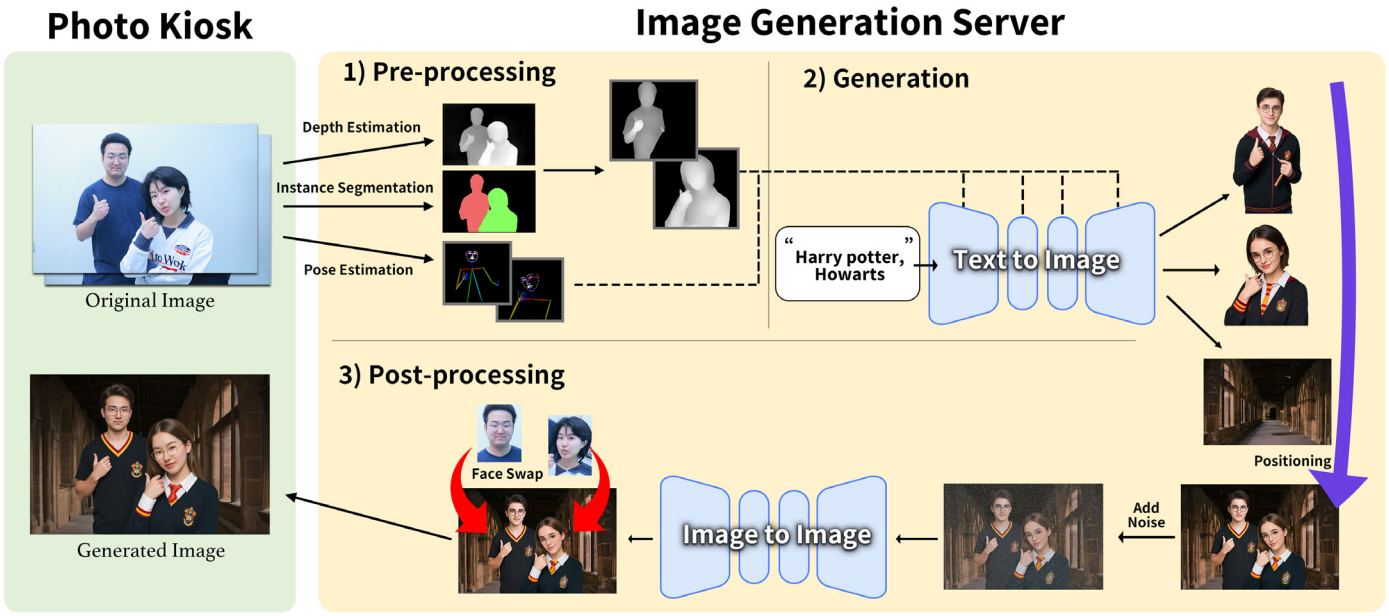


Figure 1: Overview of the CINEMAPIC system

를 통합하여 이미지 생성 과정을 안내하는 강력한 의미적 조건과 약한 공간적 조건을 추가하여 복잡한 디테일을 유지하는 새로운 접근 방식을 제안하였다. 하지만 이러한 방식은 다중 인원 생성 과정에서 압축된 정보와 여러 개인이 혼합되어 발생하는 아티팩트로 인해 특징이 손실되는 문제가 발생한다 [17]. 최근, 이를 해결하고자 embedding stack과 masked cross-attention mechanism을 사용하여 4명 이상의 다중 인원 특징을 보존할 수 있는 Instant Family [17]가 제안되었다.

이외에도, 이미지 생성이 완료된 후 얼굴을 교체하는 face swap 방식도 존재한다. 얼굴 교체에 많이 사용되는 Reactor [18]는 inswapper [19] 모델을 사용하여 짧은 시간 내에 얼굴을 교체할 수 있다. 본 연구에서는 인물별로 개별 생성하기 때문에 다중 인원 특징 보존이 필요하지 않아 다양한 방법의 적용이 가능하나, 빠른 생성 속도가 중요한 해결 과제이기 때문에 Reactor의 방식을 적용하였다.

2.3 AI 포토 서비스

최근 다양한 각도로 촬영한 5장~10장의 얼굴 사진을 업로드하면 얼굴 특징이 반영된 프로필 사진을 제공하는 유료 앱 서비스들이 다수 등장하였다 [20, 21]. 주로 정해진 이미지의 얼굴 부분만 인페인팅하여 교체하는 형태로, 이러한 서비스들은 사용자가 업로드한 사진을 학습하여 결과물을 제공하기 때문에 수십 분에서 수 시간이 소요되며, 다량의 사진이 필요하다. AI 포토 서비스가 대중화되면서 오프라인 포토부스에도 적용하는 사례들이 등장하고 있다 [4]. AI 포토부스는 사용자의 얼굴만 반영하거나, 포즈만 반영하거나, 포즈와 얼굴 모두를 반영하는 세 가지 방식으로 나뉜다. 얼굴만 반영하는 경우, 기존 앱 서비스와 유사하게 이미 정해진 이미지에 사용자의 얼굴 부분만 교체하는 형식을 취한다. 포즈만 반영하는 경우, 예를 들어 만화 속 캐릭터 컨셉을 적용했을 때 사용자의 포즈는 반영되지만, 해당 사진만으로는 사용자의 신원을 특정할 수 없다. 포즈와 얼굴 모두를 반영하는 경우, 사용자의 얼굴 특징과 행동이 모두 반영된다.

그러나 기존 서비스들은 단체 사진 촬영이 불가능하며, 인물별로 개별 컨셉 적용이 어려운 단점을 가지고 있다. 예를 들어, 6명의 인물이 어벤져스 영화 컨셉으로 촬영하고자 할 때, 각 개인이 아이언맨, 캡틴 아메리카, 헐크, 토르, 호크아이, 나타샤와 같은 캐릭터를 맡고 싶어 하지만, 모두 동일한 캐릭터 혹은 캐릭터의 특징이 뒤섞이는 현상이 발생할 수 있다. 본 연구에서는 이러한 기존 AI 포토부스의 단점을 해결하고, 다양한 환경에 응용 가능한 AI 포토부스를 제안함으로써 차별성을 가진다.

3. 시네마픽 시스템 구성

본 시스템은 사진을 촬영할 수 있는 키오스크와 촬영된 사진을 실시간으로 특정 컨셉에 맞게 변환하는 이미지 생성 서버로 나뉘어진다.

3.1 사진 촬영 키오스크

사진 촬영 키오스크는 일반 포토부스 시스템과 동일하게 카메라, 표준 컴퓨터, 터치 모니터, 스피커, 포토 프린터, 조명 등의 장치로 구성된다. 카메라는 Canon EOS 200D를 사용하였으며, 키오스크 소프트웨어는 C#으로 개발되었다. 주요 기능으로는 프레임 선택, 촬영, GIF 재생, 인쇄, QR코드 생성 등이 포함되어 있어, 일반적인 포토부스의 기능을 모두 제공한다(Figure 2). 또한, 일

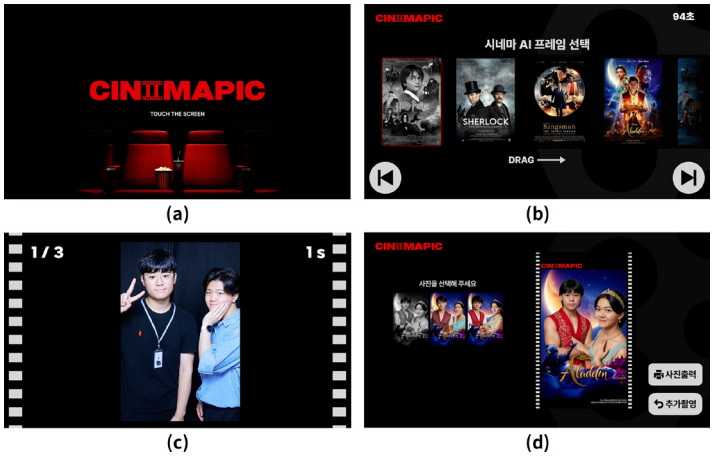


Figure 2: Main functional screens of the kiosk: (a) Initial screen, (b) Frame selection screen, (c) Photo capture screen, (d) Photo selection screen.

반적인 키오스크와는 다르게 영화 선택 페이지가 존재한다. 선택 가능한 영화는 본 연구가 기반으로 하는 사전 학습된 디퓨전 모델인 Opendalle v1.1 [22]에 학습된 영화 중 외형적인 특징이 명확하며 대중적인 어벤져스, 해리포터, 알라딘, 셜록홈즈, 킹스맨 총 5개의 영화를 미리 선정해두었다.

키오스크에서 촬영이 이루어지면, FastAPI [23]를 통해 이미지 생성 서버로 사진과 선택한 컨셉 프롬프트가 전송된다. 서버에서 모든 사진 처리가 완료되면, 생성된 결과물이 키오스크로 반환되며, 사용자는 결과물 중 원하는 사진을 선택할 수 있다. 선택된 사진은 미리 디자인된 프레임에 삽입되며, 최종 결과물은 즉시 인화하거나 QR코드를 통해 온라인에서 이미지 파일로 저장할 수 있다.

3.2 이미지 생성 서버

본 연구에서는 RTX A6000 48GB GPU 두 개로 구성된 고성능 서버를 활용하여 이미지 생성 시스템을 구현하였다. Python으로 작성된 통합 처리 프로그램은 각 사진을 워크플로우에 따라 효율적으로 처리한다. FastAPI를 사용하여 키오스크에서 서버로 사진이 전송되면, 즉시 전처리 과정을 시작한다. 전처리 된 이미지들은 Stable Diffusion WebUI API [24]로 전달되어 생성이 시작된다. 이미지 생성이 완료되면, 후처리 단계로 이동하여 이미지 배치 및 합성 작업을 수행한다. 최종적으로, 처리된 이미지는 키오스크로 다시 전송된다.

4. 이미지 생성 서버 워크플로우

이미지 생성 서버의 워크플로우 설계 시 중요한 고려 사항 두 가지는 다음과 같다. 첫째, 인물별로 각기 다른 컨셉 적용이 가능해야 한다. 둘째, 현장에서 즉시 인화할 수 있도록 빠른 생성시간 확보가 필요하다. 이러한 요구사항을 충족시키기 위해 인물을 분리하여 개별적으로 생성하는 방식을 고안하고 최적화 설계를 진행하였다. 생성 워크플로우는 Figure 1에 제시된 바와 같이 크게 1) 전처리, 2) 생성, 3) 후처리의 세 단계로 구성된다. 본 장에서는 각 단계의 구체적인 과정에 대해 설명한다.

4.1 전처리 단계

전처리 단계는 생성 및 후처리 단계에서 사용될 이미지를 여러 딥러닝 네트워크를 활용하여 가공하는 과정이다. 촬영된 사진이 이미지 생성 서버로 전달되면, 2D Human Pose Estimation state-of-the-art 모델인 DWpose [25]를 사용해 인물들의 포즈를 추정한다. 추정된 포즈를 기반으로 각 인물별로 모든 관절의 2차원 픽셀 좌표를 포함하는 바운딩 박스(Bounding Box)를 계산하고 12.5%의 패딩을 포함하여 저장한다. 또한, 얼굴관절을 중심으로 동일한 바운딩 박스를 계산하여 개별 얼굴 이미지를 저장한다. 이후 원본 이미지에서 YOLOv8 [26] 모델의 인스턴스 분할 기능을 사용하여 인물별로 경계를 추정하고, 이를 마스크로 저장한다. 동시에, 원본 이미지를 대상으로 뎁스(Depth)를 추정하고 [27], 앞서 저장한 마스크를 사용하여 인물별로 분리된 뎁스맵을 획득한다.

movie	Prompt
Avengers	Iron Man costume
Harry Potter	Harry Potter, Hogwarts
Aladdin	Aladdin, embroidered red waistcoat
Sherlock Holmes	Sherlock Holmes, longcoat
Kingsman	Kingsman

Table 1: Representative prompts for each movie, excluding gender

4.2 생성 단계

생성 단계에서는 사전 학습된 디퓨전 모델을 활용하여 조건부 Text-to-Image 방식을 통해 인물별 이미지를 생성한다. 이때 성별과 사전에 선택한 영화 컨셉이 포함된 프롬프트가 입력으로 제공된다. 예를 들어, 남성이 해리포터 영화를 선택했다면 “man, HarryPotter, Hogwarts”라는 프롬프트가 입력된다. 선택 가능한 다른 영화의 대표적인 프롬프트 예시는 Table 1에 기술되어 있다. 인물별로 다른 컨셉을 적용할 경우에는 “Harry Potter, Hogwarts”, “Ron Weasley, Hogwarts”, “Malfoy, Hogwarts”, “Dumbledore”, “Hagrid”, “Snape” 이런식의 프롬프트를 인물별로 별도 지정한다. 이때, 동일한 구성원이 촬영했더라도 촬영된 사진 한 장마다 캐릭터는 랜덤하게 지정된다. 예를 들어 4장의 사진을 촬영했다면 캐릭터가 임의로 4번 지정된다.

추가적으로, 4.1절에서 저장한 포즈 및 뎁스맵 이미지를 컨디셔닝 입력으로 사용한다. SDXL [28] 모델의 기본 해상도인 1024px로 출력 해상도를 고정하고, ControlNet을 통해 생성 과정에 포즈와 뎁스맵 이미지를 컨디셔닝 한다. 또한, 배경이 없는 투명한 PNG 이미지를 생성할 수 있는 Layer Diffusion 모델 [29]

을 사용하여 인물별 이미지를 각각 생성하였다. 마지막으로, 컨셉별로 지정된 프롬프트를 사용하여 배경 이미지를 생성한다. 예를 들어 해리포터의 경우 “a photo of Hogwarts, dark, foggy”, “a photo of Hogwart, midnight”등의 프롬프트 중 임의로 배경을 생성했다.

4.3 후처리 단계

후처리 단계에서는 인물별로 생성된 이미지를 원본 이미지와 동일한 위치에 배치하여 합성하는 과정을 수행한다. 먼저, 4.2절에서 생성한 배경 위에 인물별 컨디셔닝으로 생성된 이미지를 각각 원본 이미지의 위치와 동일하게 배치해 합성한다. 그러나 단순 배치 및 합성은 인물의 경계선에 앨리어싱(Aliasing)을 눈에 띄게 하고, 인물과 배경 간의 조명 불일치로 인해 부자연스러운 결과물이 제작된다. 이러한 문제를 해결하고자 Image-to-Image 방식으로 한 번 더 생성하는 과정을 거치며, 이때 Denoising Strength를 0.4로 설정해 원본의 형태는 유지한 채 앨리어싱 및 인위적인 아티팩트 문제를 해결한다. 이후, 합성된 이미지를 원본 얼굴 이미지로 FaceSwap [18, 19]한다.

4.4 최적화

기존의 온라인 AI 포토 서비스와 달리, 사용자가 현장에서 즉시 사진을 선택하고 인화할 수 있어야 하기 때문에, 신속하게 결과물을 제공하는 것은 매우 중요하다. 본 시스템에서 전처리 및 후처리 과정은 매우 짧은 시간 내에 처리되지만, 이미지 생성 과정은 인물별로 20초 이상의 시간이 소요되는 문제가 있었다. 특히, 단체 사진 촬영이 가능하기 때문에 인원 수가 늘어날수록 전체 생성 시간이 크게 증가하는 문제가 있었다.

이를 해결하고자 적은 Denoising Step으로도 높은 품질을 유지할 수 있는 여러 최적화 모델을 탐색하였다 [30, 31, 32, 33]. Latent Consistency Model(LCM) [30]은 잠재공간에서 ODE의 솔루션을 직접 예측하도록 설계되어 Denoising Step을 획기적으로 줄였다. 이를 바탕으로, 더 적은 메모리 소비와 추가 훈련 없이 사전 학습된 모델에 적용 가능한 LCM-LoRA [31]가 제안되었다. 이후, 1step, 2step, 4step, 8step 등 다양한 Denoising Step에서 작동하도록 훈련된 Lightning [32] 모델과 Hyper-SD [33] 모델이 제안되었다. 본 연구에서는 몇 차례의 이미지 생성 워크플로우 테스트를 통해 Layer Diffusion 및 기타 ControlNet 모델들과 가장 안정적으로 동작하는 Lightning-LoRA 모델을 적용하였다.

Number of Subjects	Average Generation Time (s)
2	25.17 (±0.25)
3	32.58 (±0.18)
4	40.43 (±0.57)

Table 2: Average processing time and standard deviation of five trials for different numbers of subjects

Figure 3: Results of applying various movie concepts. From top to bottom: Avengers, Harry Potter, Aladdin, Sherlock Holmes, Kingsman.

또한, 포토부스에서는 여러 장의 사진을 연속적으로 촬영하기 때문에, 사용자의 대기 시간을 최소화할 수 있도록 GPU 병렬화를 적용하였다. 이를 통해, 현재 GPU에서 이전 생성 프로세스가 완료되지 않은 경우, 다른 GPU를 활용하여 생성 작업을 진행함으로써 효율성을 높였다. 최종적으로 워크플로우 처리 소요 시간을 측정한 결과, 각 인물별 생성에는 약 7초가 소요되며, 재생성 과정에서는 약 10초가 소요되었다. 이에 따라 2인 촬영 시 약 25초, 4인 촬영 시 약 40초 후 모든 사진이 처리 완료되었다. 자세한 시간 측정 결과는 Table 2에 나타나 있다.

5. 시스템 품질 평가 및 시범 운영

5.1 정성적 분석

본 절은 4장에서 제시한 시스템 설계 시의 중요 고려 사항이 잘 반영되어 구현되었는지 다양한 결과물을 정성적으로 분석한다. 구현된 프로토타입으로 다양한 영화 컨셉을 적용시킨 결과물인 Figure 3에서 볼 수 있듯 인물별로 다른 컨셉이 확실히 잘 적용된 것을 확인할 수 있다. 특히 어벤져스와 알라딘 컨셉이 적용된 사진의 경우 개별 캐릭터의 특징이 명확하며 실제 사용자의 얼굴 또한 잘 반영되어 성공적으로 컨셉 적용이 된 것으로 판단할 수 있다. 또한 구현된 프로토타입으로 생성한 Figure 4의 예시를 보면 6명의 많은 인원이 겹쳐 있음에도 스타일이 혼합되거나, 경계가 제대로 분리되지 않는 등의 문제 없이 잘 생성된 것을 볼 수 있다. 또한 강제로 성별을 변경했을 경우에도, 성별의 특징을 잘 반영한 채 개인의 특징 또한 유지하는 모습을 관찰할 수 있다. 이러한 프로토타입의 결과물을 통해 본 시스템 설계에서 목표로 정한 인물 개별 컨셉 적용이 잘 구현된 것을 확인할 수 있다.

두 번째 중요 고려 사항이었던 빠른 시간 내 생성하는 목표는 최적화 모델 적용 및 GPU 병렬화를 통해 2인의 경우 25초 이내, 6인의 경우 54초 이내로 빠른 시간 내 생성 목표도 달성함을 확인했다. 또한 현재 워크플로우에서 개별 생성 과정을 진행하지

Figure 4: Result of a challenging case involving six individuals

Figure 5: Comparative results obtained with (left) and without (right) the individual separation step.

않고, 한 번에 생성된 결과물과의 비교를 통해 개별 생성 과정의 필요성에 대해 확인할 수 있다(Figure 5). 개별 컨셉이 적용된 이미지는 3명이 각각 토르, 캡틴 아메리카, 아이언맨의 특징을 잘 반영하고 있는 반면, 한번에 생성된 결과물은 컨셉이 혼합되어 캡틴 아메리카는 아이언맨 수트 재질이 반영되어 있고, 토르는 아예 생성되지 않는 등 원하는 대로 이미지를 제어할 수 없다.

5.2 사용자 평가

본 연구에서는 원본 사진의 얼굴과 포즈가 잘 반영되는지 검증하고 결과물의 품질을 평가하기 위해 2.2절에서 소개된 기존 연구 방법인 FastComposer[14], InstantID[16]와 질적 비교 평가를 실시하였다. 참가자는 대학 캠퍼스에 게시된 전단지를 통해 두 명씩 팀으로 모집해 총 10팀을 구성하였다. 참가자 20명(남성 5명, 여성 15명, 평균 나이 21세)을 대상으로 약 30분간 평가를 진행하였다. 참가자는 연구의 자세한 설명을 듣고 동의서를 작성하였으며, 실험참가에 대한 사례로 1만원을 받았다.

세부적인 실험 과정은 다음과 같다. 가장 먼저, 키오스크에서 2회 촬영이 진행되었다. 이때, 인원수에 따른 결과 비교를 위해 첫 번째 촬영은 참가자 두 명만 촬영하고, 두 번째 촬영은 연구자 한 명을 추가하여 세 명이 촬영하도록 하였다. 사진이 총 세 개의 연구 방법으로 처리되는 동안 참가자들은 잠시 대기하였으며, 처리가 완료되면 27인치 모니터 앞에 한명씩 앉아 평가를 진행했다. 설문지는 얼굴 일치도, 표정 반영도, 포즈 반영도, 전체적인 품질에 대해 리커트 척도(Likert Scale)로 평가하는 문항(5점 = 매우 그렇다, 1점 = 매우 그렇지 않다)으로 구성되었다. 각 참가자는 세 가지 방법으로 생성된 결과물을 임의의 순서로 관찰한 후 설문에 응답했다. 각 방법을 평가하는데 있어서 두 명이 촬영한 원본 사진과 두가지 컨셉으로 생성한 사진 두 장, 세명이 촬영한 원본 사진과 두가지 컨셉으로 생성한 사진 두장, 총 6장의 이미지를 한번에 보여주는 방식으로 진행했다.

결과물을 만드는 세부 과정에서 연구 방법마다 입력할 수 있는 형식에 차이가 있었다. FastComposer의 경우 텍스트 프롬프트와 얼굴 이미지를 입력으로 사용하였고, InstantID는 텍스트 프롬프트와 원본 이미지, 얼굴 이미지, 원본 이미지에서 인스턴스 분할을 통해 인물별로 분리된 이미지를 입력으로 사용하였다. 또한 참가자의 포즈를 반영하기 위해 뎁스맵을 추가 컨디셔닝 입력으로 제공하였다.

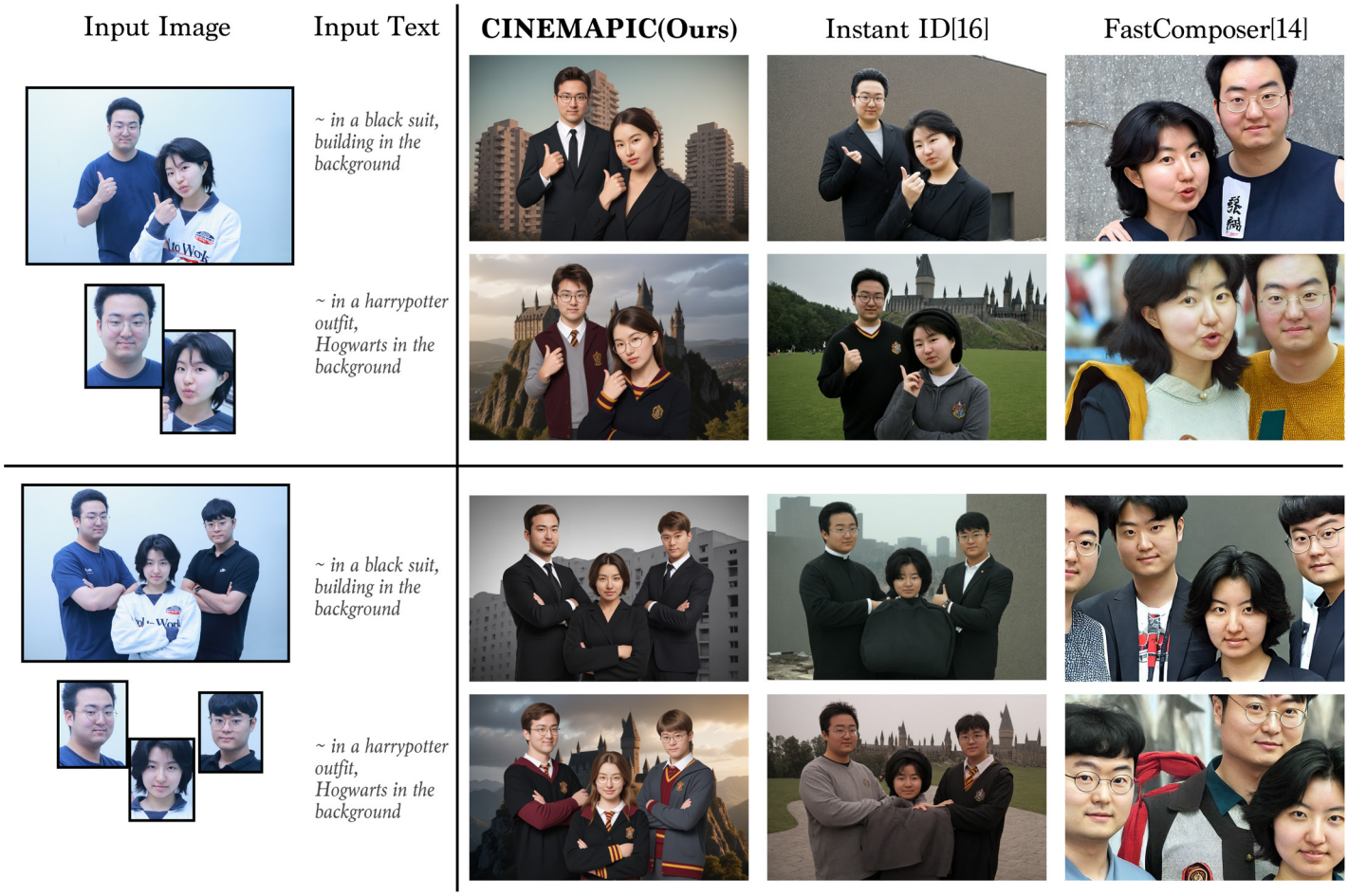


Figure 6: Qualitative comparison between the multi-ID image generation results produced by our method and those produced by state-of-the-art methods.

Table 3: Questionnaires and responses about different image generation methods (n = 20)

No	Questionnaire	**CINEMAPIC(Ours)** Mean(SD)	InstantID Mean(SD)	FastComposer Mean(SD)
1	생성된 이미지의 캐릭터가 나와 일치하는 얼굴을 가지고 있나요?	3.30 (0.80) ***	1.90 (0.85)	1.80 (0.77)
2	생성된 이미지의 캐릭터가 내가 지은 표정을 잘 반영하였나요?	3.90 (0.85)	3.10 (1.21)	3.60 (1.05)
3	생성된 이미지의 캐릭터가 나의 포즈를 잘 반영하였나요?	4.15 (1.04) ***	3.80 (0.95)	2.00 (0.97)
4	생성된 이미지의 전체적인 품질은 어떤가요?	4.25 (0.64) ***	2.60 (1.10)	2.70 (1.13)

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, compared to other methods.

실험 결과의 통계 분석은 일원분산분석(ANOVA)을 사용하여 수행되었다. ANOVA 분석을 통해 각 평가 항목에서 방법 간의 유의미한 차이를 확인하였다. 분석에는 각 질문별로 평균 점수와 표준 편차가 포함되었고, 사후 검정을 통해 구체적인 그룹 간 차이를 파악하였다. 실험 결과는 Table 3에 나타난 바와 같이, 모든 항목에서 기존 연구 대비 높은 선호도를 보였다. 첫 번째 질문인 얼굴 유사도에서 CINEMAPIC(Ours)은 3.3점으로 보통 수준을 받았다. 이는 다른 방식들에 비해 통계적으로 유의미하게 높은 수준이나, 얼굴의 윤곽선이나 얼굴의 크기가 실제와 다른 경우가 존재했다는 의견이 있었다. 표정도 우리 방법이 가장 높은 점수를 받았지만 통계적으로 유의미 하지 않았다. 포즈에 대해서는 포즈나 뎁스를 입력할 수 없는 FastComposer를 제외한 두 방식은 보통 수준 이상의 평가를 받았으며 CINEMAPIC이 유의미하게 가장 높았다. 마지막으로 전체적인 품질에 관한 항목에서는 큰 차이가 있었는데, 이는 InstantID의 특성상 인물이 많아질 수록 개인 특징의 영향이 작아져 흐릿한 이미지가 생성되는 것과 FastComposer의 불안정한 생성 퀄리티가 영향을 미쳤을 것으로 판단된다. 설문 평가가 끝난 후 참가자들에게 향후 서비스가 출

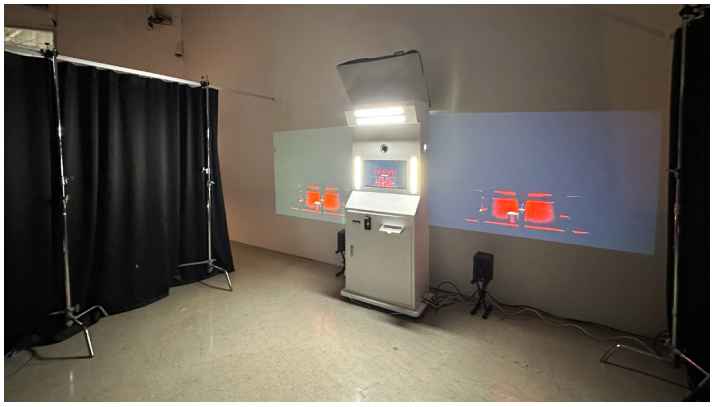


Figure 7: Pilot operation

시된다면 어떤 서비스를 가장 사용할 의향이 있는지 물어본 결과, 모든 참가자가 본 연구의 시네마픽을 선호했으며, 실제로 비용을 지불하고 사용할 의향이 있다는 의견도 많았다.

5.3 시범 운영

우리는 프로토타입 시스템의 실제 포토부스로의 도입 가능성을 확인하고자, 3일간 약 400명의 사용자를 대상으로 시범 운영을 실시하였다. 사용자들은 비용을 지불하지 않고 AI 포토부스를 이용하였으며, 선택 가능한 영화 컨셉은 해리포터, 알라딘, 셜록홈즈, 킹스맨이었다. 키오스크 조작부터 인화까지 모든 과정은 무인으로 진행되었으며, 실제 포토부스와 동일한 시나리오로 운영되었다. 많은 사용자가 현장에서 즉시 원하는 컨셉으로 변환되는 것에 큰 흥미를 보였고, 영화 제목을 밝히지 않아도 어떤 영화를 바탕으로 제작된 것인지 바로 알아차릴 수 있었다. 그러나 질적 비교 평가 결과와 유사하게, 얼굴 형태에 관한 부정적인 의견도 일부 존재하였다. 또한, 인원이 많아질 경우 포즈 추정 및 인스턴스 분할 과정에서 일부 인원이 누락되는 문제가 발생하였다. 그럼에도 불구하고 대부분의 시나리오에서 문제없이 결과물을 생성했으며, 사용자들의 만족스러운 반응과 함께 결과물을 SNS에 공유하는 사례도 많이 발생했다. 간단한 서비스 이용 설문조사에서 본 시스템이 적용되면 좋을 장소에 대한 의견을 물은 결과, 영화관이라는 응답이 가장 많았고, 이외에도 팝업 스토어와 축제 등이 좋은 적용 장소로 언급되었다. 이를 통해 실제 포토부스로의 도입 가능성을 확인할 수 있었다.

6. 한계점

본 연구는 자유롭게 단체사진 촬영이 가능한 AI 포토부스를 설계하고 프로토타입을 제작하여, 사용자 평가 및 대규모 시범 운영을 통해 실제 포토부스 도입 가능성을 검증하였다. 그러나 그 과정에서 몇 가지 한계점이 발견되었다.

첫째, Figure 8에서 볼 수 있듯이, 복잡한 포즈이거나 사람이 많이 겹쳐있는 경우 인스턴스 분할 또는 포즈 추정이 어려워진다.

Figure 8: Failure case

이로 인해 추정 오차가 발생하면 인물이 누락되어 생성된 결과물에 포함되지 않을 수 있다. 인스턴스 분할 시, 같은 클래스에 속하는 객체 인스턴스 겹치는 경우 오류가 발생하는 경우가 많기 때문이다 [34]. 또한 인물이 누락되는 경우 face swap 과정에서 순서가 뒤섞여 타인의 얼굴이 합성되는 경우가 발생하기도 했다. 둘째, 조건부 Text-to-Image 생성에서 얼굴 크기나 얼굴 윤곽 등의 정보가 포즈와 뎁스맵에 포함되지 않아, 실제 사용자의 얼굴 형태와 일치하지 않는 경우가 발생한다. 이로 인하여 사용자는 자신과의 일치성을 낮게 판단할 수 있다. 마지막으로, 손가락의 형태가 제대로 나오지 않거나 개수가 정상적이지 않은 문제가 존재한다. 포즈와 뎁스 추정 과정에서 손이 잘 보이는 경우에는 제대로 생성되지만, 각도나 위치에 따라 겹치는 경우 비정상적인 손가락 문제가 발생할 수 있다.

향후 연구로는 InstantID와 같은 생성시에 개인의 특징을 반영하는 방식과 Reactor와 같이 생성이 완료된 후 Face swap을 진행하는 방식을 결합하여 빠른 시간 내에 얼굴 크기나 얼굴 윤곽 등 특징을 더 잘 반영하는 개선 방안을 고려할 수 있다.

7. 결론

본 연구는 생성형 AI를 활용하여 다양한 컨셉을 적용할 수 있고 단체사진 촬영이 가능한 AI 포토부스 시스템을 제안하였다. 제안된 시스템은 사용자의 포즈나 위치, 의상 등에 제한을 두지 않고 자유롭게 촬영할 수 있도록 하였으며, 인물별 컨셉 적용을 위한 개별 생성 워크플로우를 개발하고, 빠른 생성 효율을 위한 최적화를 구현하였다. 그 과정에서 후처리 과정을 염두에 둔 투명 이미지 생성, 별도 배경 생성 후 합성시 발생하는 아티팩트를 줄이는 재생성 테크닉, 최적화 모델 적용 및 GPU 병렬화 등 다양한 방식을 워크플로우에 통합하여 한계점을 극복하였다.

또한 본 시스템의 성능을 평가하고 실제 포토부스 도입 가능성을 검증하기 위해 사용자 평가 및 대규모 시범 운영을 진행한 결과, 기존 방식 대비 뛰어난 선호도를 보였으며 실제 포토부스로의 도입 가능성을 확인하였다.

그러나, 이 과정에서 인물 누락 가능성, 얼굴 형태의 낮은 반영도, 비정상적인 손가락 문제 등의 한계점을 발견하였다. 이러한 한계점에도 불구하고, 본 시스템은 일반적인 포토부스와 AI 포토부스의 장점을 결합한 최초의 포토부스로서 다양한 컨셉 및 상황에 적용 가능한 시스템을 구현했다는 점에서 의의가 있다.

우리는 본 시스템이 더욱 창의적이고 차별화된 시장을 개척할 수 있을 것으로 기대하며, 앞으로 다양한 응용 분야에서 널리 활용될 것으로 기대된다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 메타버스 융합대학원(IITP-2024-RS-2024-00430997, 기여율 20%)과 지역지능화혁신인재양성사업(IITP-2024-RS-2022-00156360, 기여율 20%)과 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화체육관광 연구개발사업(연구개발과제명 : 공연 콘텐츠의 고해상도(8K/16K) 서비스를 위한 AI 기반 영상확장 및 서비스 기술개발, 연구개발과제번호 : RS-2024-00395886, 기여율: 60%)의 지원을 받아 수행되었음

References

[1] 방효은, "셀프 포토 스튜디오 서비스 관련 실태조사," *조사보고서*, pp. 1–32, 2023.

[2] 노지은 and 류한영, "Z 세대를 위한 포토부스 애플리케이션 제안," *한국 HCI 학회 학술대회*, pp. 975–978, 2023.

[3] 박수빈, "[체험기] 포토부스에서도 ai 사진 촬영하고 즉석 인화까지!" *AI타임스*. [Online]. Available: https://www.aitimes.com/news/articleView.html?idxno=158250

[4] 조현영, "LGU+, 대학교 축제 현장에 '익시' 사진관 열어," *연합뉴스*. [Online]. Available: https://www.yna.co.kr/view/AKR20240529102700017

[5] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2021.

[6] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," *arXiv preprint arXiv:1812.10889*, 2018.

[7] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[9] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[10] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, "Instantbooth: Personalized text-to-image generation without test-time finetuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8543–8552.

[11] X. Zhang, X.-Y. Wei, W. Zhang, J. Wu, Z. Zhang, Z. Lei, and Q. Li, "A survey on personalized content synthesis with diffusion models," *arXiv preprint arXiv:2405.05538*, 2024.

[12] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[14] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han, "Fastcomposer: Tuning-free multi-subject image generation with localized attention," *arXiv preprint arXiv:2305.10431*, 2023.

[15] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[16] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen, "Instantid: Zero-shot identity-preserving generation in seconds," *arXiv preprint arXiv:2401.07519*, 2024.

[17] C. Kim, J. Lee, S. Joung, B. Kim, and Y.-M. Baek, "Instantfamily: Masked attention for zero-shot multi-id image generation," *arXiv preprint arXiv:2404.19427*, 2024.

[18] Gourieff, "sd-webui-reactor," https://github.com/Gourieff/sd-webui-reactor, 2024, accessed: 2024-06-11.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[20] Snow, "Snow corp official website," https://www.snowcorp.com/.

[21] Carat.im, "Carat.im official website," https://carat.im/.

[22] A. Izquierdo, "Opendallev1.1," https://huggingface.co/dataautogpt3/OpenDalleV1.1, 2023.

[23] S. Ramírez, “Fastapi,” https://fastapi.tiangolo.com/, 2018.

[24] AUTOMATIC1111, “stable-diffusion-webui,” https://github.com/AUTOMATIC1111/stable-diffusion-webui, 2024.

[25] Z. Yang, A. Zeng, C. Yuan, and Y. Li, “Effective whole-body pose estimation with two-stages distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4210–4220.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[27] R. Birkl, D. Wofk, and M. Müller, “Midas v3. 1–a model zoo for robust monocular relative depth estimation,” *arXiv preprint arXiv:2307.14460*, 2023.

[28] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.

[29] L. Zhang and M. Agrawala, “Transparent image layer diffusion using latent transparency,” *arXiv preprint arXiv:2402.17113*, 2024.

[30] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, “Latent consistency models: Synthesizing high-resolution images with few-step inference,” *arXiv preprint arXiv:2310.04378*, 2023.

[31] S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, and H. Zhao, “Lcm-lora: A universal stable-diffusion acceleration module,” *arXiv preprint arXiv:2311.05556*, 2023.

[32] S. Lin, A. Wang, and X. Yang, “Sdxl-lightning: Progressive adversarial diffusion distillation,” *arXiv preprint arXiv:2402.13929*, 2024.

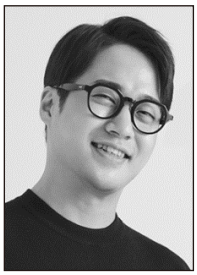
[33] Y. Ren, X. Xia, Y. Lu, J. Zhang, J. Wu, P. Xie, X. Wang, and X. Xiao, “Hyper-sd: Trajectory segmented consistency model for efficient image synthesis,” *arXiv preprint arXiv:2404.13686*, 2024.

[34] L. Ke, Y.-W. Tai, and C.-K. Tang, “Deep occlusion-aware instance segmentation with overlapping bilayers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 4019–4028.

〈저자소개〉

정석현
- 2019년~현재 : 숭실대학교 글로벌미디어학부 학사
- 관심분야 : Generative AI, Machine Learning, Computer Graphics
- https://orcid.org/0009-0001-4729-0805

임승규
- 2019년~현재 : 숭실대학교 글로벌미디어학부 학사
- 관심분야 : Generative AI, Machine Learning, Computer Graphics
- https://orcid.org/0009-0006-7173-373X



이정진
- 2010년 숭실대학교 미디어학부 학사
- 2012년 KAIST 문화기술대학원 석사
- 2017년 KAIST 문화기술대학원 박사
- 2016년~2020년 (주)카이 연구이사
- 2020년~현재 (주)라이브커넥트 CTO 사외이사
- 2020년~현재 숭실대학교 글로벌미디어학부 조교수
- 관심분야: 컴퓨터 그래픽스, VR/AR, 몰입형 시각 미디어, 이미지/비디오 응용, HCI
- https://orcid.org/0000-0003-3471-4848