Journal of KIIT. Vol. 24, No. 5, pp. 15-23, May 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.5.15

Unity 기반 인체 포즈 GT 생성과 ControlNet-OpenPose 조건부 생성모델을 활용한 가상데이터 기반 포즈추정 학습 파이프라인

최현빈*, 우동현**, 김중락***

A Synthetic Data - Driven Pose Estimation Training Pipeline using Unity-based Human Pose GT Generation and a ControlNet-OpenPose Conditional Generative Model

Hyunbin Choi*, DongHyun Woo**, and Joongrock Kim***

본 논문은 2025년도 교육부 및 경상남도의 재원으로 경상남도RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.(2025-RISE-16-002) 또한, 2025년도 교육부의 재원으로 글로컬대학사업의 지원을 받아 수행된 사업의 결과입니다.

요 약

본 논문은 인체 포즈 추정 모델의 학습 데이터 부족 문제를 해결하기 위해 Unity 기반 정밀 GT 생성과 ControlNet-OpenPose 기반 조건부 이미지 합성을 결합한 자동화 파이프라인을 제안한다. 제안 방법은 시뮬레이션의 구조적 정확성과 확산 모델의 시각적 사실성을 통합하여 기존 합성 데이터의 도메인 갭 문제를 완화한다. COCO 포맷과 호환되는 합성 데이터셋을 구축하고 이를 실제 데이터와 혼합하여 학습을 수행하였다. 실험 결과, OKS-AP가 63.2(실제 단독)에서 65.0(혼합)으로 향상되었다. 이는 제안 파이프라인이 데이터 다양성을 보강하고 모델의 일반화 성능을 개선함을 보여준다.

Abstract

This paper proposes an automated synthetic data generation pipeline that integrates Unity-based precise Ground Truth (GT) generation with ControlNet-OpenPose conditional image synthesis for human pose estimation training. The framework combines structural accuracy from simulation with visual realism from diffusion models to mitigate the domain gap of conventional synthetic data. A COCO-compatible dataset was constructed and evaluated through mixed training with real data. Experimental results show that OKS-AP improves from 63.2 (real-only) to 65.0 (real + synthetic). These findings demonstrate that structurally aligned pose-conditioned generation enhances data diversity and improves model generalization.

Keywords

synthetic data, pose estimation, unity, diffusion model, controlnet, openpose

* 국립창원대학교 인공지능융합공학과 박사과정
- ORCID: https://orcid.org/0000-0003-4300-5859
** 국립창원대학교 첨단방위공학과정 박사과정
- ORCID: https://orcid.org/0000-0002-5562-4158
*** 국립창원대학교 인공지능융합공학과 부교수(교신저자)
- ORCID: https://orcid.org/0009-0001-5284-1149

· Received: Jan. 13, 2026, Revised: Mar. 11, 2026, Accepted: Mar. 14, 2026
· Corresponding Author: Joongrock Kim
Dept. of Artificial Intelligence Convergence Engineering, 20 Changwondaehak-ro, Uichang-gu, Changwon-si, Gyeongsangnam-do, 51140, Korea
Tel.: +82-55-213-3962, Email: jrkim@changwon.ac.kr

Ⅰ. 서 론

인체 포즈추정(Human pose estimation)은 이미지 내에서 사람의 관절 위치를 추정하여 자세를 구조적으로 표현하는 기술로, 스포츠 동작 분석, 헬스케어, 인간-컴퓨터 상호작용(HCI), 로보틱스 안전 관제 등 다양한 산업 분야의 핵심 기술로 자리 잡았다. 딥러닝 기술의 비약적인 발전과 함께 포즈추정 모델의 성능은 꾸준히 향상되어 왔으나, 이러한 모델들은 여전히 대규모의 정제된 이미지-주석(Image-annotation) 데이터에 크게 의존한다는 한계를 갖는다. 실제 환경에서 관절 주석을 정밀하게 구축하는 과정은 막대한 비용과 시간이 소요되며, 특정 동작이나 극한의 조명, 특수 의상 등 희귀한 케이스(Long-tail distribution)를 포함하는 데이터를 확보하는 것은 현실적으로 매우 어렵다[1].

최근 인공지능 분야는 데이터 중심(Data-centric) AI로의 패러다임 전환과 함께, 부족한 학습 데이터를 인공적으로 확보하기 위한 가상 데이터(Synthetic data) 생성 연구가 활발히 진행되고 있다[2]. 초기 연구들은 Unity나 Unreal Engine과 같은 3D 그래픽스 엔진을 활용하여 주석을 자동 산출하는 방식을 주로 채택하였다. 이 방식은 3D 캐릭터의 자세와 카메라 파라미터를 완벽하게 제어할 수 있어 정확한 GT(Ground Truth)를 얻을 수 있다는 장점이 있다. 그러나 렌더링 된 이미지가 실제 사진의 텍스처나 조명 분포를 완벽히 모사하지 못해 발생하는 '도메인 갭(Domain gap)' 문제는 모델의 실세계 성능을 저하시키는 주요 원인으로 지적되어 왔다.

이러한 문제를 해결하기 위해 최근 생성형 인공지능(Generative AI), 특히 확산 모델(Diffusion model)을 활용한 데이터 증강 시도가 급격히 증가하고 있다. 텍스트-투-이미지(Text-to-image) 모델은 놀라운 수준의 시각적 사실성을 제공하지만, 텍스트 프롬프트만으로는 생성되는 인물의 구체적인 자세나 구조적 형상을 정밀하게 제어하기 어렵다는 한계가 있다. 예를 들어, 단순한 프롬프트 입력은 팔다리의 개수가 잘못되거나 해부학적으로 불가능한 자세를 생성하는 등 '환각(Hallucination)' 문제를 야기할 수 있어 정밀한 좌표 주석이 필요한 포즈추정 데이터로는 부적합할 수 있다. 이에 대한 대안으로 등장한 ControlNet은 Canny Edge, Depth, Pose Skeleton 등의 공간적 조건(Spatial condition)을 확산 모델에 추가 입력함으로써 생성 이미지의 구조적 일관성을 유지하면서도 사실적인 텍스처를 입히는 것이 가능하다[3]. 이는 기존 그래픽스 기반 방식의 정확성과 생성형 AI의 사실성을 결합하여 도메인 갭을 획기적으로 줄일 수 있는 기술로 주목받고 있다[4][5].

하지만 이러한 기술적 진보에도 불구하고, 시뮬레이션의 관절 정의가 표준 데이터셋(예: COCO Keypoints)과 상이하거나, 생성 모델이 입력 조건(스켈레톤)을 완벽히 따르지 않아 다중 인물이 등장하는 등의 정합성 문제는 여전히 해결해야 할 과제로 남아 있다. 최신 인공지능 기술을 활용하여 학습 데이터를 구축할 때는 생성된 이미지와 주석 간의 불일치를 최소화하고, 데이터의 품질을 보장하는 파이프라인 설계가 필수적이다.

따라서 본 논문은 Unity 기반 시뮬레이션을 통해 정밀한 2D 포즈 GT와 포즈 조건(Pose condition) 이미지를 생성하고, 이를 최신 생성형 모델인 ControlNet-OpenPose 파이프라인에 연동하여 고품질의 학습용 이미지-주석 쌍을 자동 구축하는 시스템을 제안한다. 본 연구는 Unity의 구조적 정확성과 생성 모델의 표현력을 결합하여 도메인 갭을 완화하는 것을 목표로 하며, COCO 포맷 호환성을 고려한 데이터 처리 및 배경·조명 변이를 통한 다양화 전략을 포함한다. 제안된 파이프라인으로 구축된 데이터를 실제 데이터와 혼합하여 포즈추정 모델을 학습시켰을 때, 정량적 성능 지표(OKS-AP)의 변화를 분석함으로써 제안 방식의 유효성을 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 인체 포즈추정과 합성 데이터, 그리고 최신 생성형 모델의 동향을 관련 연구로 정리한다. 3장에서는 Unity와 생성 서버 간의 연동 구조 및 전체 파이프라인을 설명하고, 4장에서는 포즈 주석 및 조건 이미지의 생성 방식과 포맷 정의를 기술한다. 마지막으로 5장에서는 실험을 통해 구축된 데이터의 효용성을 분석하고 결론을 맺는다.

Ⅱ. 관련 연구

2.1 인체 포즈추정과 관절 표현

2D 인체 포즈추정은 이미지 내 사람의 관절 위치를 추정하여 자세를 구조적으로 표현하는 기술로, 딥러닝 기술의 도입 이후 비약적인 성능 향상을 이루었다 [6]. 초기 연구에서는 단일 인물뿐 아니라 다중 인물 장면에서의 인물 분리 및 관절 연결이 중요한 난제로 다루어졌다. 이에 OpenPose는 PAF (Part Affinity Fields)를 이용해 관절 간 연결 관계를 추정하는 Bottom-up 접근을 제시하여, 다중 인물 2D 포즈추정 분야의 기반 기술로 자리 잡았다.모델의 학습과 정량적 평가를 위해서는 표준화된 관절 정의와 데이터 포맷이 필수적이며, 현재 MS COCO 데이터셋이 가장 대표적인 벤치마크로 활용된다. COCO 키포인트는 17개 관절로 구성되며, 각 관절은 좌표(x, y)와 가시성(v) 정보를 포함하는 벡터 형태로 저장된다. 그림 1은 COCO 17-keypoint의 관절 정의와 인덱스, 그리고 대표적인 스켈레톤 연결 구조를 나타낸 것이다. 본 논문은 이러한 표준을 준수하여, 시뮬레이션에서 생성되는 포즈 주석이 COCO 포맷과 호환되도록 구성하고 이를 학습에 활용한다.

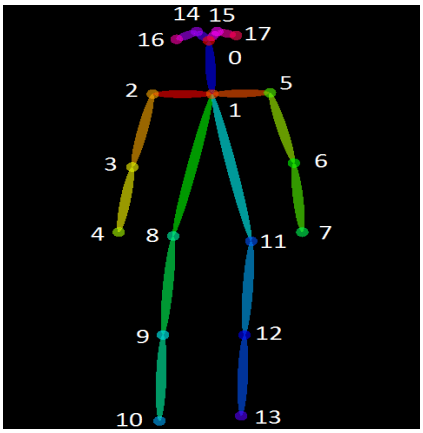

그림 1. COCO 17- 스켈레톤 연결 구조
Fig. 1. Concept of COCO 17-skeleton topology

포즈추정 모델은 접근 방식에 따라 크게 Top-down(사람 검출 후 인물별 포즈 추정)과 Bottom-up(관절 검출 후 연결) 계열로 구분된다. Top-down 계열에서는 Simple Baselines가 기준선으로 제시된 이후, 고해상도 특징 표현을 유지하는 HRNet이나 비전 트랜스포머 기반의 ViTPose 등 다양한 방법이 제안되며 COCO 기준 성능이 지속적으로 개선되어 왔다[8][9]. 본 논문은 새로운 네트워크 구조를 제안하기보다는, 이러한 고성능 모델들이 공통적으로 요구하는 고품질의 '이미지 - 포즈 주석 쌍'을 자동으로 구축하는 파이프라인 설계에 초점을 둔다.

2.2 시뮬레이션 기반 합성 데이터 생성

합성 데이터(Synthetic data)는 라벨링 비용을 획기적으로 줄이면서 배경, 조명, 카메라 시점, 객체 배치 등을 자유롭게 통제하여 다양한 조건의 데이터를 생성할 수 있다는 점에서 컴퓨터 비전 분야의 데이터 부족 문제를 해결할 대안으로 주목받아 왔다. 특히 Unity 기반 도구는 3D 공간에서 시뮬레이션 장면을 구성하고 렌더링하는 과정에서 포즈, 세그멘테이션 등 다양한 형태의 정밀한 주석(Ground truth)을 자동 산출할 수 있어, 데이터 구축의 효율성을 크게 높일 수 있다.

그림 2는 Unity 환경에서 시뮬레이션 장면을 렌더링하고, 프레임 단위로 주석(예: 포즈 JSON)을 출력하며, 배경·조명·시점 등의 요소를 무작위로 변화시키는(Domain Randomization) 일반적인 합성 데이터 생성 흐름을 보여준다. 이러한 구조는 포즈추정 학습에 필요한 데이터를 반복적으로 구축하는 데 유용하다. 그러나 시뮬레이션 이미지는 렌더링 엔진의 한계로 인해 실제 사진과 질감 및 조명 표현에서 차이가 발생하며, 이러한 도메인 갭(Domain Gap)은 모델의 실세계 성능을 저하시키는 주요 원인이 된다[10]. 따라서 최근 연구들은 단순 렌더링을 넘어, 주석의 정합성(좌표계·포맷)을 유지하면서도 시각적 사실성을 보완하는 방향으로 발전하고 있다.

2.3 확산모델과 포즈 조건부 생성(ControlNet)

확산 모델(Diffusion Model)은 데이터 분포에 점진적으로 노이즈를 추가한 뒤 이를 역으로 제거하며 이미지를 합성하는 방식으로, 특히 잠재 공간(Latent Space)에서 확산을 수행하여 계산 효율과 생성 품질을 동시에 확보하였다.

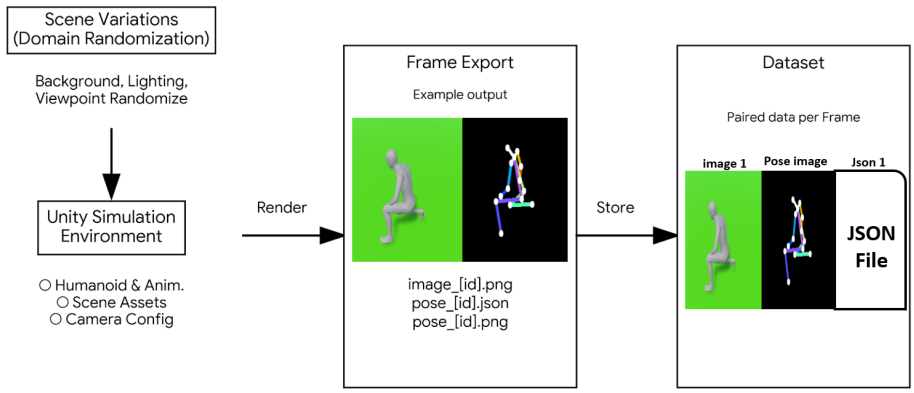

그림 2. Unity 기반 시뮬레이션 합성 데이터 생성
Fig. 2. Concept of Unity-based synthetic data generation

일반적인 텍스트-투-이미지(Text-to-image) 모델은 높은 표현력을 가지지만, 텍스트 프롬프트만으로는 생성되는 인물의 구체적인 자세나 구조를 정밀하게 제어하기 어렵다는 한계가 있다. 이는 정확한 주석 매칭이 필요한 포즈 추정 데이터 생성 시 구조적 왜곡(Hallucination)을 유발할 수 있다.

이에 대한 대안으로 제시된 ControlNet은 사전 학습된 확산 모델에 외부 조건을 결합하여 구조적 제어(Structural control)를 가능하게 한다. ControlNet은 에지, 깊이, 세그멘테이션, 인체 포즈 등 다양한 공간적 조건을 입력받아 생성 결과의 구조를 강제할 수 있다. 특히 OpenPose 계열 포즈 조건은 입력된 스켈레톤 구조를 유지하면서 사실적인 텍스처를 입히는 생성이 가능하여, 포즈추정 학습용 합성 데이터 생성에 가장 적합한 기술로 평가받는다.

다만, 시뮬레이션에서 생성된 관절 정의나 스켈레톤 연결 구조가 생성 모델이 학습한 분포와 다를 경우, 조건 반영이 약해지거나 생성 이미지에서 다중 인물 등장, 신체 부위 누락 등의 품질 저하가 발생할 수 있다[12]. 따라서 본 논문은 Unity에서 생성한 포즈 조건 및 주석을 ControlNet 기반 생성 서버와 연동하고, 데이터 포맷 및 좌표계를 일치시켜 고품질의 학습 데이터를 구축하는 통합 파이프라인을 제안한다.

III. 제안 시스템

3.1 시스템 개요

본 논문에서 제안하는 가상 데이터 생성 파이프라인은 시뮬레이터가 제공하는 '구조적 정밀함'과 생성형 AI가 제공하는 '시각적 사실성'을 결합하여, 도메인 갭이 최소화된 고품질 학습 데이터를 자동 구축하는 것을 목표로 한다. 전체 시스템은 크게 Unity 기반 데이터 생성부, ControlNet 기반 생성 서버부, 그리고 포즈 추정 학습부로 구성된다.

그림 3은 제안 시스템의 전체 파이프라인을 모듈 간의 상호작용으로 도식화한 것이다. Unity Client는 시뮬레이션 환경에서 렌더링을 수행하고 생성에 필요한 파라미터(Generation parameters)를 구성하여 서버로 전송한다. Generation Server는 수신된 파라미터와 포즈 조건(Pose PNG)을 기반으로 ControlNet-OpenPose를 구동하여 사실적인 이미지를 합성하고, 그 결과(Base64 encoded data)를 반환한다. 이렇게 확보된 합성 데이터는 Training & Eval 모듈로 전달되어, 포즈 추정 모델의 학습 및 성능 평가에 활용되는 유기적인 순환 구조를 갖는다.

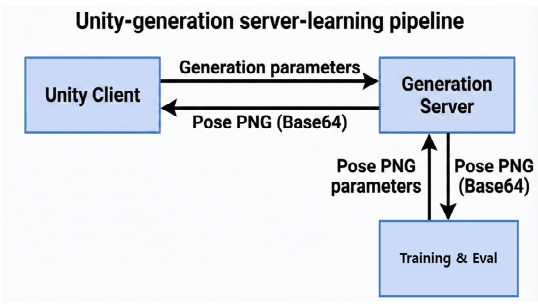

그림 3. Unity-생성 서버-학습 파이프라인
Fig. 3. Unity-generation server-learning pipeline

3.2 모듈 구성 및 설계 근거

각 모듈은 데이터의 정합성을 유지하면서도 시스템의 확장성을 고려하여 독립적인 기능을 수행하도록 설계되었다.

첫째, Unity 데이터 생성부는 3D 휴머노이드 캐릭터와 애니메이션 에셋을 활용하여 다양한 동작을 시뮬레이션한다. 이 과정에서 배경, 조명, 카메라 시점을 무작위로 변화시키는 도메인 무작위화(Domain Randomization)를 적용하여 데이터의 다양성을 1차적으로 확보한다. 렌더링 시점에는 화면상의 관절 좌표를 계산하여 COCO 포맷 호환 JSON으로 저장하고, 동시에 동일한 카메라 뷰에서 관절 스켈레톤만을 시각화한 Pose PNG를 생성한다. 이 두 데이터는 동일한 렌더링 파이프라인에서 추출되므로 완벽한 기하학적 일치(Geometric alignment)를 보장한다.

둘째, 생성 서버부(FastAPI & ControlNet)는 Python 환경의 PyTorch 기반 확산 모델(Stable Diffusion + ControlNet)을 구동하기 위해, C# 기반의 Unity와 분리된 별도의 서버(FastAPI)로 구축하였다. 이러한 느슨한 결합(Loose coupling) 구조는 다음과 같은 이점을 갖는다. 우선 시뮬레이션(CPU/Unity)과 무거운 생성 모델(GPU/Diffusion)의 리소스를 물리적으로 분리하여 시스템 안정성을 높인다. 또한 Unity 클라이언트를 수정하지 않고도 서버 측에서 생성 모델의 체크포인트나 샘플링 파라미터를 유연하게 교체할 수 있어 실험의 용이성을 제공한다. 서버는 Unity로부터 수신한 Pose PNG를 조건(Condition)으로 사용하여 텍스처와 배경이 합성된 이미지를 생성하고, 이를 다시 Unity로 반환한다.

마지막으로, 학습 및 평가부는 구축된 데이터셋(이미지-주석 쌍)을 로드하여 2D 포즈추정 모델을 학습한다. 이때 파일명에 포함된 프레임 ID를 키(Key)로 사용하여 이미지와 라벨을 매핑하며, 학습 목적에 따라 '가상 데이터 단독' 또는 '실제 데이터와 혼합' 구성을 선택하여 모델의 일반화 성능을 평가한다.

3.3 데이터 및 인터페이스 정의

시스템의 신뢰성을 위해 모듈 간 교환되는 데이터 포맷과 인터페이스를 명확히 정의하였다. 모든 좌표 데이터는 픽셀 단위의 이미지 좌표계를 따르며, 원점은 이미지의 좌상단(Top-Left, 0,0), 축은 우측(+x) 및 하단(+y) 방향으로 설정하여 이미지 처리 라이브러리(OpenCV 등)와의 호환성을 유지한다.포즈 주석은 MS COCO 키포인트 정의를 따르며, 각 관절은 좌표(x, y)와 가시성(c)을 포함하는 벡터로 표현된다. 한 프레임 내의 단일 인물 포즈 P는 식 1과 같이 17개 관절의 순차적 벡터로 정의된다.

$$P = [x_0, y_0, c_0, \cdots, x_{16}, y_{16}, c_{16}] \quad (2)$$

여기서 c는 관절의 가시성 상태(0: 이미지 밖, 1: 가려짐, 2: 보임)를 나타낸다. 시스템 내 주요 데이터 아티팩트(Artifact)의 역할은 표 1과 같다.

표 1. 파이프라인에서 데이터 아티팩트 및 역할
Table 1. Data artifacts and roles in the pipeline

Artifact	Format	Role
Pose annotation (GT)	JSON	Labels (matched by frame_id)
Pose condition (Pose PNG)	PNG	Generator condition
Generated image	PNG/JPG	Training input
Metadata (opt.)	JSON/ CSV	Reproducibility

Unity와 생성 서버 간의 통신은 HTTP REST API를 통해 이루어지며, 데이터의 무결성을 위해 모든 요청과 저장 파일은 고유한 frame_id를 공유한다. 구체적인 API 명세와 파일 저장 구조는 표 2에 요약하였다.

표 2. 모듈 간 인터페이스 및 데이터 저장 구조
Table 2. Interface and storage

Item	Specification	Note
Health check	GET health	Status check
Generation	POST generate	pose_png_base64 → image_base64
Dataset key	frame_id	1:1 matching
Layout	images, poses, pose_png	File-based

이러한 구조적 설계를 통해 시뮬레이션에서 자동 생성된 수천 장 이상의 데이터가 주석 오류 없이 학습 단계로 파이프라인 되도록 보장한다. 특히 유효하지 않은 관절(Simulation 상에서 렌더링되지 않은 부위 등)은 가시성 값(c=0)으로 마스킹 처리하여, 학습 시 손실 함수 계산에서 자동으로 제외되도록 구성하였다.

Ⅳ. 데이터셋 구축 및 포즈추정 학습 결과

4.1 데이터셋 구축 절차

본 장에서는 3장에서 정의한 파이프라인을 이용하여 합성 데이터셋을 구축하고, 이를 포즈추정 학습에 적용한 결과를 기술한다. 데이터 구축 단계에서는 Unity 장면 내의 배경, 조명, 카메라 시점을 무작위로 변화시키는 도메인 무작위화(Domain randomization)를 적용하여 데이터의 다양성을 확보하였다. 또한 생성 서버의 추론 파라미터는 사전 실험을 통해 이미지 품질과 생성 속도의 균형을 맞춘 값으로 고정하여 데이터 생성의 일관성을 유지하였다.

특히 합성 데이터의 품질이 학습에 미치는 부정적 영향을 방지하기 위해, 생성된 데이터에 대해 단계적인 품질 통제(Quality control) 절차를 적용하였다. 첫째, OpenPose를 이용해 생성 이미지를 2차 검증하여 검출된 관절의 신뢰도(Confidence)가 일정 임계값(예: 0.4) 미만인 데이터는 학습에서 제외하였다. 둘째, Unity GT 상의 관절 좌표와 생성된 이미지 내 인물의 위치가 시각적으로 현저히 불일치하거나, 단일 인물 조건임에도 다중 인물이 생성된 경우를 필터링하여 데이터의 정합성을 보장하였다. 모든 산출물은 프레임 ID를 기준으로 저장하여 이미지와 포즈 주석의 1:1 대응 관계가 파일 단위에서 유지되도록 구성하였다.

4.2 학습 환경 설계

실험의 재현성을 확보하기 위해 포즈추정 모델의 구조와 학습 하이퍼파라미터를 구체적으로 명시한다. 본 실험에서는 Top-down 방식의 대표적 모델인 HRNet-W32를 백본(Backbone)으로 사용하였으며, 입력 이미지 해상도는 256×192로 설정하였다.학습 최적화 도구(Optimizer)로는 AdamW를 사용하였으며, 초기 학습률(Learning rate)은 1e-3으로 설정하고 지정된 에폭마다 학습률을 감소시키는 Step Decay 스케줄러를 적용하였다. 배치 크기(Batch size)는 32로 설정하였으며, 과적합 방지를 위해 Random Flip, Rotation($\pm$30도), Scale(0.75-1.25) 등의 데이터 증강(Augmentation)을 공통적으로 적용하였다. 구체적인 학습 환경 설정은 표 3에 요약하였다.

표 3. 포즈추정 모델 학습 하이퍼파라미터
Table 3. Hyperparameters for pose estimation training

Item	Value
Model architecture	HRNet-W32
Input resolution	256 x 192
Optimizer	AdamW
Learning rate	1e-3 (Step Decay)
Batch size	32
Total epochs	200
Augmentation	Flip, Rotation, Scale

4.3 데이터셋 구성 및 규모

표 4. 데이터셋 구성 및 규모
Table 4. Dataset composition and size

Setting	Real (Train)	Synthetic (Train)	Total (Train)	Test set
Real only	10,000	0	10,000	Real test
Real + Synthetic	10,000	10,000	20,000	Real test
Synthetic Only	0	10,000	10,000	Real test

구축된 데이터는 학습 목적에 따라 학습(Train), 검증(Val), 평가(Test) 세트로 분할하였다. 실제 데이터(Real data)는 COCO train2017의 일부를 샘플링하여 구축하였으며, 가상 데이터(Synthetic data)는 제안한 파이프라인을 통해 생성하였다.

실험 조건은 데이터 구성에 따라 실제 데이터 단독(Real only), 실제+가상 혼합(Real + Synthetic), 가상 데이터 단독(Synthetic only)의 세 가지로 설정하였다. 혼합 학습 시에는 실제 데이터와 가상 데이터

를 1:1 비율로 구성하여, 데이터 양의 증가와 다양성 보강 효과를 균형 있게 반영하고자 하였다. 각 데이터셋의 구체적인 규모는 표 4와 같다.

4.4 학습 결과 및 분석

표 5는 각 학습 조건에 따른 모델의 성능을 비교 분석한 결과이다. 실험 결과, '실제+가상 혼합 학습(Real + Synthetic)'의 OKS-AP가 65.0으로 '실제 데이터 단독(Real only)'인 63.2 대비 약 1.8%p 향상된 성능을 보였다. 이는 단순히 학습 데이터의 양(Quantity)이 증가했기 때문만이 아니라, 합성 데이터가 실제 데이터셋에서 부족했던 다양한 자세, 배경, 조명 등의 환경 변이를 제공하여 데이터의 다양성(Diversity)을 확보했기 때문이다. 결과적으로 이러한 다양성이 모델의 과적합을 방지하고 일반화(Generalization) 성능 개선에 기여했음을 시사한다.

반면, '가상 데이터 단독(Synthetic only)' 학습은 OKS-AP 49.5를 기록하며 실제 테스트 세트 기준에서 비교적 낮은 성능을 보였다. 이는 생성된 이미지의 텍스처나 시각적 분포가 실제 이미지와 상이하여 발생하는 도메인 갭(Domain gap)에 기인한 것으로 분석된다. 따라서 현 단계에서 합성 데이터는 실제 데이터를 완전히 대체하기보다는, 실제 데이터가 부족한 영역을 보완하거나 다양성을 높이는 보조 데이터(Supplementary data)로 활용하는 것이 적절함을 확인할 수 있다.

표 5. 포즈 추정 성능 비교 결과
Table 5. Comparative results of pose estimation

Metric	Real only	Real + Synthetic	Synthetic only
OKS-AP (main)	63.2	65.0	49.5
OKS-AP@0.50	83.1	84.7	71.0
OKS-AP@0.75	68.4	69.2	45.8

정성적 분석 결과, 합성 데이터의 학습 기여도는 생성된 이미지의 시각적 품질과 조건 입력(Condition input)의 정합성(Alignment)에 크게 의존함이 확인되었다. ControlNet의 입력인 Pose PNG는 인체의 기하학적 구조 정보를 제공하는 핵심 요소이나, 프레이밍이 불안정하거나 스켈레톤 정보가 모호할 경우 생성 결과물에서 다중 인물 등장, 경계 모호성(Ambiguity), 신체 왜곡 등의 아티팩트(Artifact)가 발생할 수 있다.

그림 4는 Pose PNG(조건 입력), 생성 이미지, 그리고 오버레이 결과를 비교하여 조건 입력의 중요성을 보여준다. (A)는 입력 스켈레톤의 자세는 반영되었으나, 인물과 배경의 경계가 불분명하거나 세부 디테일이 손상되어(Low fidelity) GT와의 정합성이 다소 부족한 사례이다. (B)는 단일 인물 생성 제약에도 불구하고 다수의 인물이 생성된 경우로, 조건 입력의 불확실성이나 모델의 장면 해석 편향이 제약 조건 위반(Constraint Violation)으로 이어질 수 있음을 나타낸다. 반면 (C)는 입력 포즈와 생성된 인체가 정확히 일치하고 시각적 품질 또한 우수한 사례로, 명확한 조건 입력과 적절한 프레이밍이 보장될 때 고품질의 가상 데이터를 안정적으로 확보할 수 있음을 시사한다.

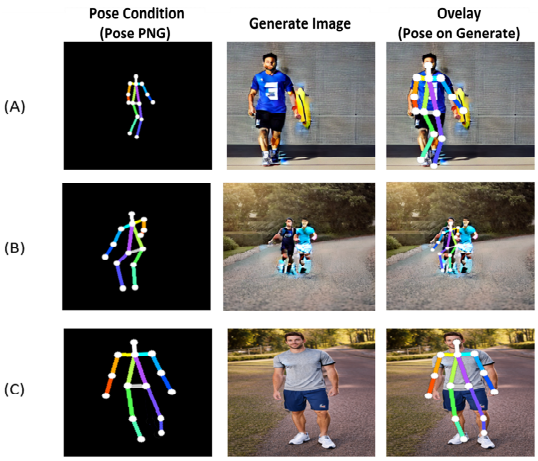

그림 4. Pose 조건 기반 가상 데이터 생성 사례
Fig. 4. Qualitative examples of pose-conditioned synthetic data

V. 결론 및 향후 과제

본 논문은 인체 포즈 추정 분야의 고질적인 문제인 학습 데이터 부족과 레이블링 비용 문제를 해결하기 위해, Unity 시뮬레이션의 구조적 정밀함과 ControlNet 생성 모델의 시각적 사실성을 결합한 자

동화된 데이터 생성 파이프라인을 제안하였다. 제안 시스템은 Unity 기반의 데이터 생성부, FastAPI 기반의 생성 서버부, 그리고 학습·평가부로 구성되며, 프레임 ID 기반의 1:1 매칭 시스템을 통해 조건 이미지(Pose PNG)와 정답 주석(GT), 생성 이미지 간의 완벽한 정합성을 보장하도록 설계되었다.

제안된 파이프라인의 효용성을 검증하기 위해 실제 데이터와 가상 데이터의 구성 비율을 달리하여 비교 실험을 수행하였다. 실험 결과, 실제 데이터(Real only) 단독 학습 시 63.2였던 OKS-AP가 가상 데이터를 1:1로 혼합하여 학습했을 때 65.0으로 약 1.8%p 향상됨을 확인하였다. 이는 본 파이프라인이 생성한 가상 데이터가 다양한 배경, 조명, 의상 패턴을 제공함으로써 데이터의 다양성(Diversity)을 확보하고, 결과적으로 모델의 과적합을 방지하여 일반화 성능 개선에 기여했음을 시사한다. 반면, 가상 데이터 단독 학습(Synthetic Only)의 경우 OKS-AP 49.5를 기록하여 여전히 실제 환경과의 도메인 갭이 존재함을 확인하였으나, 이는 보조 데이터로서의 가치를 반증하는 결과이기도 하다.

또한 정성적 분석을 통해, ControlNet의 입력 조건인 Pose PNG의 명확성과 프레이밍 안정성이 생성된 이미지의 품질 및 최종 포즈 추정 성능을 결정하는 핵심 요인임을 규명하였다. 입력 스켈레톤이 겹치거나 모호할 경우 생성 모델이 다중 인물을 생성하거나 신체를 왜곡하는 환각(Hallucination) 현상이 발생하였으며, 이는 학습 노이즈로 작용할 수 있음을 확인하였다.

본 연구의 한계이자 향후 개선이 필요한 과제는 다음과 같다. 첫째, 생성 이미지의 품질 안정화이다. 현재의 파이프라인은 확률적 생성 모델에 의존하므로, 생성된 이미지 내의 인체 구조 왜곡이나 다중 인물 생성을 원천적으로 차단하기 어렵다. 이를 해결하기 위해 향후 연구에서는 스켈레톤 외에 깊이(Depth) 정보나 세그멘테이션 맵을 추가 조건으로 활용하여 구조적 제어력을 강화할 계획이다. 둘째, 도메인 갭의 최소화이다. 가상 데이터 단독 학습 성능을 높이기 위해, 실제 이미지의 스타일을 모사하는 도메인 적응(Domain adaptation) 기법이나, 생성된 데이터 중 학습에 유효한 샘플만을 선별하는 자동화된 품질 필터링(Quality filtering) 알고리즘을 도입하고자 한다.

결론적으로, 본 논문에서 제안한 파이프라인은 시뮬레이션의 정확성과 생성형 AI의 표현력을 융합하여 저비용으로 고품질의 학습 데이터를 무한히 확장할 수 있는 가능성을 제시하였으며, 이는 향후 다양한 비전 태스크의 데이터 부족 문제를 해결하는 데 기여할 것으로 기대된다.

References

[1] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context", European Conference on Computer Vision (ECCV), Zurich, Switzerland, pp. 740-755, Sep. 2014. https://doi.org/10.1007/978-3-319-10602-1_48.

[2] X. Ju, A. Zeng, J. Zhao, Q. Xu, and L. Zhang, "HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation", Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, pp. 17650-17660, Oct. 2023. https://doi.org/10.1109/ICCV51070.2023.01465.

[3] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "DiffPose: Toward More Reliable 3D Pose Estimation", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 13041-13051, Jun. 2023. https://doi.org/10.1109/CVPR52729.2023.01253

[4] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models", Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, pp. 3836-3847, Oct. 2023. https://doi.org/10.1109/ICCV51070.2023.00355.

[5] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 1, pp. 172-186, Jan.

2021. https://doi.org/10.1109/TPAMI.2019.2929257.

[6] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose++: Vision Transformer Foundation Model for Generic Body Pose Estimation", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol. 46, No. 11, pp. 7845-7862, Nov. 2023. https://doi.org/10.1109/TPAMI.2023.3330016.

[7] B. Xiao, H. Wu, and Y. Wei, "Simple Baselines for Human Pose Estimation and Tracking", European Conference on Computer Vision (ECCV), Munich, Germany, pp. 466-481, Sep. 2018. https://doi.org/10.1007/978-3-030-01231-1_29.

[8] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 5693-5703, Jun. 2019. https://doi.org/10.1109/CVPR.2019.00584.

[9] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective Whole-body Pose Estimation with Two-stages Distillation", Proc. of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, pp. 4210-4220, Oct. 2023. https://doi.org/10.1109/ICCVW60793.2023.00455.

[10] Q. Chang, H. Wang, W. Wu, and K. Liu, "Learning from Synthetic Human Group Activities", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 13812-13822, Jun. 2024. https://doi.org/10.1109/CVPR52733.2024.02070.

[11] M. J. Black, P. Patel, J. Tesch, and J. Yang, "BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 8726-8737, Jun. 2023. https://doi.org/10.48550/arXiv.2306.16940.

저자소개

최 현 빈(Hyunbin Choi)

2021년 8월 : 국립창원대학교 문화테크노학과(학사)
2025년 2월 : 국립창원대학교 문화융합기술 협동과정(석사)
2025년 3월 ~ 현재 : 국립창원대학교 인공지능융합공학 박사과정
관심분야 : 컴퓨터비전,증강/가상현실, 모션캡처

우 동 현 (DongHyun Woo)

2012년 2월 : 배재대학교 전자공학과(공학사)
2018년 8월 : 중국 동북전력대학교 정보및통신공학(공학석사)
2021년 3월 ~ 현재 : 국립창원대학교 첨단방위공학과정 박사과정
관심분야 : 컴퓨터비전, 증강/가상현실

김 중 락 (Joongrock Kim)

2005년 3월 : 고려대학교 전자정보공학(공학사)
2005년 1월 ~ 2006년 7월 : 엠텍비전 연구원
2008년 8월 : 연세대학교 생체인식공학(공학석사)
2014년 2월 : 연세대학교 전기전자공학(공학박사)
2014년 2월 ~ 2024년 12월 : LG전자 CTO인공지능연구소 Vision Intelligence 연구실 팀장
2025년 1월 ~ 현재 : 국립창원대학교 인공지능융합공학과 부교수
관심분야 : 2D/3D 컴퓨터비전, 인공지능, On-Device