ISSN 2287-5026 (Print)
ISSN 2288-159X (Online)

Journal of the Institute of Electronics and Information Engineers

2026 4 제63권 4호

Vol.63, No.4 April 2026

www.theieie.org Vol.63, No.4 April 2026
The Institute of Electronics and Information Engineers (IEIE)
Room #907, The Korea Science Technology Center The first building, 22,
Teheran-ro 7 Gil, Gangnam-gu, Seoul, Republic of Korea

전자공학회 논문지
2026 4
제63권 4호
IEIE
사단법인 대한전자공학회

전자공학회논문지
•이 논문집은 한국연구재단 **우수등재학술지**임.

차 례

2026년 4월 제63권 제4호

SD / 반도체

[SoC 설계]

[RF 집적회로기술]

CI / 컴퓨터

[인공지능 및 보안]

AISP / 인공지능 신호처리

[영상 신호처리]

[음향 및 신호처리]

IE / 산업전자

[신호처리 및 시스템]

Journal of The Institute of Electronics and Information Engineers Vol.63, NO.4, April 2026 https://doi.org/10.5573/ieie.2026.63.4.78
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2026-63-4-9

단일 사람 3D 키포인트 검출을 위한 Cross Attention 트랜스포머 모델 설계 및 성능 분석

(Design and Performance Analysis of a Cross-attention Transformer Model for Single-person 3D Keypoint Detection)

신 인 영*, 이 승 호**

(In-Yeong Shin and Seung-Ho Lee©)

요 약

본 논문에서는 단일 인물 3D 자세 추정의 정확도와 연산 효율성을 동시에 극대화하기 위해, Cross-Attention 메커니즘 기반의 새로운 시공간 통합 트랜스포머(Spatio-Temporal Feature Fusion Transformer) 모델을 제안한다. 기존의 2D 키포인트 기반 접근법은 입력 데이터의 노이즈와 떨림 현상으로 인해 3D 복원 시 불안정한 결과를 초래하는 한계가 있었다. 이를 해결하기 위해 본 연구는 이산 코사인 변환(DCT)을 전처리 단계에 도입하여 시계열 데이터의 불필요한 고주파 성분을 필터링함으로써, 노이즈를 효과적으로 억제하고 동작의 시간적 연속성을 확보하였다. 또한 공간 트랜스포머를 통해 인체 관절 간의 기하학적 관계를 벡터화하고, 이를 시간적 특징과 결합하기 위해 Cross-Attention 구조를 적용하였다. 이 융합 모델은 공간적 구조 정보와 시간적 동적 정보를 상호 보완적으로 학습하여 추정의 정밀도를 높인다. 결과적으로 본 연구는 주파수 영역의 전처리와 시공간 통합 어텐션 기법의 시너지를 통해, 단일 인물 3D 자세 추정에서 모델의 강건성과 추정 성능이 유의미하게 향상됨을 입증하고자 한다.

Abstract

In this paper, we propose a novel Spatio-Temporal Feature Fusion Transformer model based on a cross-attention mechanism to simultaneously maximize the accuracy and computational efficiency of single-person 3D pose estimation. Conventional 2D keypoint-based approaches often suffer from instability in 3D reconstruction due to noise and jittering inherent in the input data. To address this issue, we introduce the Discrete Cosine Transform (DCT) as a preprocessing step. By filtering out unnecessary high-frequency components from the time-series data, this method effectively suppresses noise and ensures the temporal continuity of motion. Furthermore, we utilize a spatial transformer to embed the geometric relationships between human joints into vectors and apply a cross-attention structure to integrate these with temporal features. This fusion model enhances estimation precision by learning spatial structural information and temporal dynamic information in a mutually complementary manner. Consequently, this study aims to demonstrate that the synergy between frequency-domain preprocessing and the spatio-temporal integrated attention mechanism leads to significant improvements in both the robustness and performance of single-person 3D pose estimation.

Keywords : Feature fusion, 3d, Pose estimation, Deep learning, Transformers

*학생회원, **평생회원, 국립한밭대학교 전자공학과(Dept. Electronic Engineering, Hanbat National University)
© Corresponding Author(E-mail : shyolee@hanbat.ac.kr)
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학·석사연계ICT핵심인재양성사업의 연구결과로 수행되었음(IITP-2026-RS-2022-00156212).
Received ; December 19, 2025 Revised ; January 17, 2026 Accepted ; January 19, 2026

Ⅰ. 서 론

메타버스와 스마트 헬스케어 시장이 급성장함에 따라, 카메라로 촬영된 사용자의 동작을 정교한 3차원 데이터로 변환해주는 '3차원 인체 자세 추정' 기술의 중요성이 그 어느 때보다 커지고 있다. 최근 딥러닝 기술은 CNN을 넘어 시퀀스 데이터 처리에 탁월한 트랜스포머(Transformer)[1]로 진화하며 비약적인 성능 향상을 이뤘다. 그러나 기존의 트랜스포머 모델들은 노이즈가 포함된 2D 키포인트 좌표를 여과 없이 입력받아 결과의 불안정성을 초래하고, 긴 시퀀스 처리 시 연산 비용이 급증한다는 치명적인 단점이 존재한다.

본 논문은 이러한 고질적인 문제를 타개하기 위해, Cross-Attention 메커니즘 기반의 효율적이고 강건한 3D 자세 추정 프레임워크를 제안한다. 핵심은 데이터의 질적 개선과 효율적 융합에 있다. 먼저 이산 코사인 변환(DCT)을 전처리 단계에 도입하여 시계열 데이터의 고주파 노이즈를 제거하고 동작의 연속성을 확보하였다. 이후, 공간 트랜스포머(Spatial Transformer)[2]로 추출된 인체의 기하학적 특징과 정제된 시간 정보를 시공간 통합 트랜스포머 내에서 융합한다. 본 연구는 이 결합 방식이 연산 효율성을 저해하지 않으면서도 추정 정밀도를 획기적으로 높일 수 있음을 입증한다.

Ⅱ. 본 론

1. 제안하는 연구의 개요

본 연구에서는 T 프레임 길이의 2D 키포인트 시퀀스를 입력으로 받아, 3D 인체 자세를 추정하는 Cross-Attention 트랜스포머 기반의 프레임워크를 제안한다. 제안하는 모델은 크게 인체의 구조적 정보를 학습하는 공간 트랜스포머와 주파수 영역의 시간 정보를 융합하는 시공간 통합 트랜스포머로 구성된다.

전체적인 과정은 다음과 같다. 모델의 입력은 (T×J×2, 여기서 J는 관절의 수) 형태의 2D 키포인트 시퀀스로 정의된다. 이 입력 데이터는 두 가지 경로로 처리되어 상호 보완적인 정보를 추출한다. 첫째, 공간 트랜스포머는 각 프레임 내 존재하는 J개의 관절 간 상호 관계를 모델링 하여, 인체의 기하학적 구조를 내포한 공간 특징 벡터를 추출한다. 둘째, 입력 시퀀스의 시간 축에 대해 이산 코사인 변환(DCT)[3]을 수행하여 시간적 변화량을 주파수 계수로 변환한다. 이를 통해 고주파 노이즈를 필터링하고 동작의 핵심적인 동적(Dynamic)

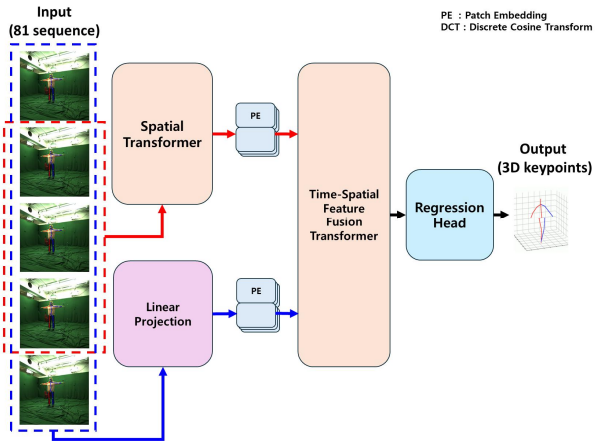

그림 1. 전체 모델의 개요
Fig. 1. Structure of the Model.

정보를 확보한다. 최종적으로 시공간 통합 트랜스포머는 Cross-Attention 메커니즘을 활용하여, 추출된 공간 특징(Query)과 주파수 도메인으로 변환된 시간 특징(Key, Value)을 효과적으로 결합한다.

이 과정을 통해 시공간 정보가 융합된 벡터가 생성되며, 이를 회귀(Regression)하여 최종적인 단일 시점의 3D 관절 좌표를 예측한다. 그림 1은 제안하는 모델의 전체 개요를 나타낸다.

2. Multi Head Attention

본 연구에서 제안하는 공간 트랜스포머와 시공간 통합 트랜스포머는 모두 멀티 헤드 어텐션(Multi-Head Attention) 메커니즘을 핵심 구성 요소로 사용한다. MHA는 입력 특징을 서로 다른 부분 공간(Subspace)으로 투영하여 다양한 관점에서의 상관관계를 병렬적으로 학습할 수 있게 한다. 두 모듈은 입력되는 쿼리(Q), 키(K), 밸류(V)의 구성 성분만 다를 뿐, 내부적인 연산 과정은 동일한 구조를 공유한다. 먼저, 기본적인 어텐션 메커니즘인 스케일 조정된 내적 어텐션(Scaled Dot-Product Attention)은 다음과 같이 정의된다.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

여기서 d_k는 키 벡터의 차원을 의미하며, 내적 값의 크기가 지나치게 커지는 것을 방지하여 기울기 소실 문제를 완화한다. 멀티 헤드 어텐션은 이러한 단일 어텐션 연산을 h개의 헤드에 대해 병렬적으로 수행한다. 각 헤드는 독립적인 학습 파라미터를 통해 입력을 선형 투영(Linear Projection)[4]하며, 각 헤드의 출력은 결합(Concatenation)된 후 다시 선형 변환되어 최종 출력을

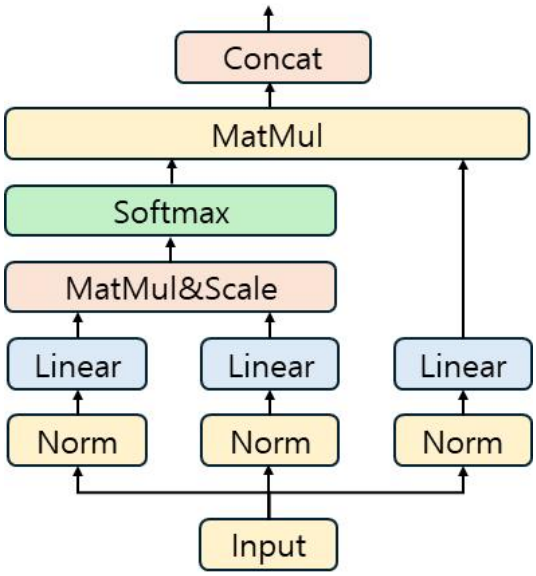

그림 2. 멀티 헤드 어텐션의 구조
Fig. 2. Structure of Multi Head Attention.

생성한다. 멀티 헤드 어텐션의 구조는 그림 2와 같다.

3. 공간 트랜스포머

공간 트랜스포머(Spatial Transformer)는 단일 프레임 내에서 인체 관절 간의 기하학적 구조와 상호 의존성(Spatial Dependency)을 모델링 하는 역할을 수행한다. 입력된 2D 키포인트 좌표는 공간적 맥락을 고려한 고차원 특징 벡터로 변환된다.

먼저, 입력 시퀀스 의 각 프레임에 대하여, J개의 관절 좌표를 선형 투영하여 C 차원의 특징 공간으로 매핑한다. 이때, 트랜스포머는 입력 순서에 불변하는 특성을 가지므로, 각 관절의 고유한 순서를 식별하기 위해 학습 가능한 공간 위치 임베딩을 더해준다.

생성된 입력 특징은 앞서 정의한 멀티 헤드 어텐션 블록으로 전달된다. 공간 트랜스포머는 관절 간의 자체적인 상관관계를 학습해야 하므로, 쿼리, 키, 밸류가 모두 동일한 입력으로부터 유래하는 셀프 어텐션(Self-Attention) 구조를 따른다. 이후 나온 벡터는 시공간 트랜스포머에서 쿼리로 사용하기 위해 Layer Normalization과 MLP 층을 거쳐 변환된다. 공간 트랜스포머의 구조는 그림 3과 같다.

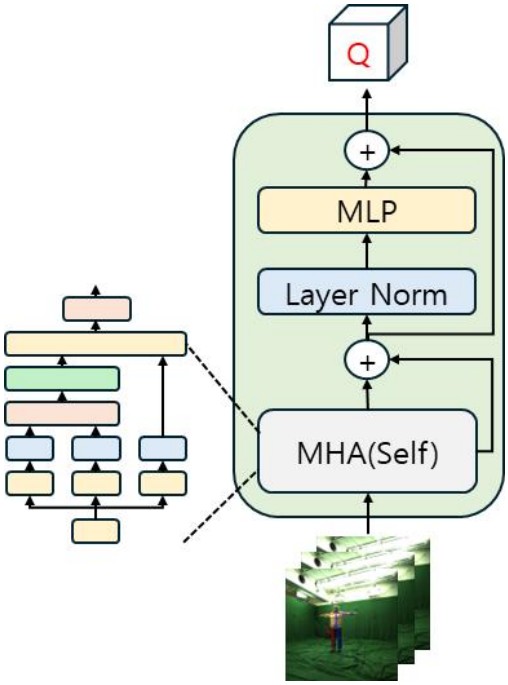

그림 3. 공간 트랜스포머의 구조
Fig. 3. Structure of Spatial Transformer.

4. 시공간 통합 트랜스포머

시공간 통합 트랜스포머는 앞서 추출된 인체의 공간적 특징과 시간적 움직임 정보를 융합하여 최종적인 3D 자세를 추론하는 핵심 모듈이다. 본 모듈은 DCT 기반의 주파수 임베딩과 크로스 어텐션 메커니즘을 결합하여, 노이즈가 제거된 전역적 시간 정보를 공간 특징에 결합하는 방식으로 설계되었다.

먼저, 시간적 정보의 효율적인 처리를 위해 입력 시퀀스의 시간 축에 대해 이산 코사인 변환을 수행한다. 시간 도메인의 입력은 이산 코사인 변환을 통해 주파수 도메인의 계수로 변환된다. 인체의 움직임은 주로 저주파 대역에 집중되어 있고 고주파 대역은 떨림이나 노이즈일 가능성이 높다는 특성을 이용하여, 하위 20개의 저주파 성분만을 남기고 나머지 고주파 성분을 제거하는 로우 패스 필터링을 적용한다. (2)번은 필터링된 주파수 계수 F를 정의하는 수식이다.

$$F_{time} = Select(DCT(X), 20) \quad (2)$$

따라서 시공간 통합 트랜스포머의 입력은 다음 수식 (3)과 같다.

$$Input = MHA(Q = Z_S, K = F_{time}, V = F_{time}) \quad (3)$$

이 구조는 공간적 특징이 전체 비디오 시퀀스의 시간적 맥락을 참조하도록 유도한다. 즉, 모델은 현재 프레임의 자세를 추정할 때, DCT로 요약된 과거와 미래의 움직임 흐름을 반영하여 모호한 자세를 보정하게 된다.

최종적으로, 융합된 특징은 잔차 연결과 MLP헤드를 거쳐 3D 좌표 공간으로 회귀되며, 단일 프레임에 대한

최종 3D 관절 좌표를 출력한다. 시공간 통합 트랜스포머의 구조는 다음 그림 4와 같다.

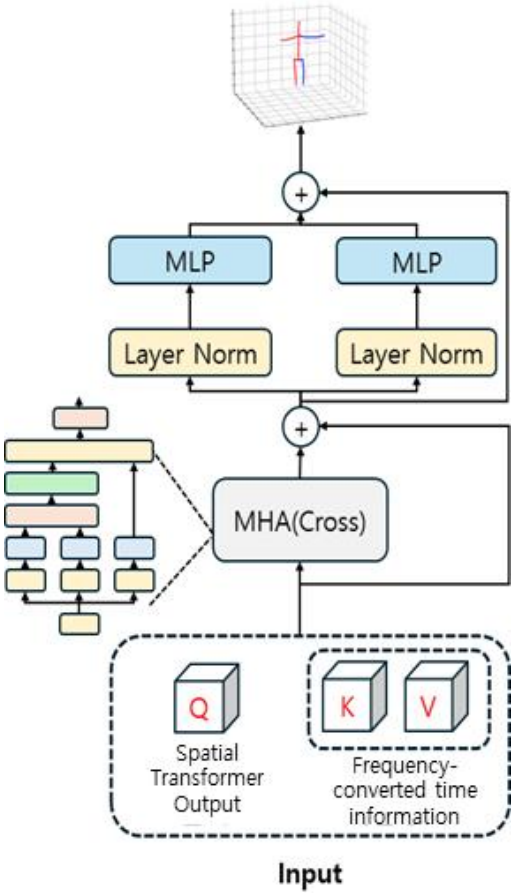

그림 4. 시공간 통합 트랜스포머의 구조
Fig. 4. Structure of Time-Spatial Feature Fusion Transformer.

5. 손실함수

제안하는 네트워크는 엔드-투-엔드(End-to-End) 방식으로 학습된다. 모델의 파라미터를 최적화하기 위한 손실함수로는 예측된 3D 관절 위치와 실제(Ground Truth) 3D 관절 위치 사이의 유클리드 거리(Euclidean Distance)를 측정하는 MPJPE(Mean Per Joint Position Error)[6] 변형 손실을 사용한다. 다음 수식 (4)와 같다.

$$L = \frac{1}{J}\sum_{i=1}^{J} \| \hat{Y}_i - Y_i \|_2^2 \quad (4)$$

이 손실 함수를 최소화함으로써 모델은 각 관절의 위치 오차를 전체적으로 줄이는 방향으로 학습되며, 시공간 통합 트랜스포머를 통해 융합된 특징들이 정확한 3D 공간 좌표로 매핑되도록 유도한다.

Ⅲ. 실 험

1. 실험 환경

본 실험은 Ubuntu 20.04.2 LTS 환경에서 Intel Xeon Silver 4210 CPU, 128GB RAM, NVIDIA RTX A5000 (24GB) GPU를 사용하여 수행되었다. 딥러닝 프레임워크는 PyTorch 1.8.0을 사용하였으며, CUDA 11.1과 cuDNN 8.0.5 라이브러리를 적용하였다. 모델 학습을 위해 배치 크기(Batch size)는 16 으로 설정하였으며, 최적화 알고리즘으로는 Adam[5]를 사용하였다. 초기 학습률(Learning rate)은 0.001로 시작하여 스케줄러에 의해 조정되도록 설계하였다. 입력 시퀀스 길이는 81프레임으로, 공간 트랜스포머(3프레임)와 시공간 트랜스포머(81프레임)가 유기적으로 정보를 처리한다. 모델은 2D Ground Truth 키포인트를 입력받아 총 300 epoch 동안 학습되었으며, 손실 함수로는 MPJPE를 사용하였다.

2. 데이터셋

본 연구의 평가 및 검증을 위해 MPI-INF-3DHP[7]와 HumanSC3D[8] 데이터셋이 사용되었다. 이들은 다양한 배경과 복장 조건을 포함하고 있어 모델의 강건성을 확인하는 데 필수적이다. 데이터셋은 각각 130만 장, 120만 장의 연속 이미지로 이루어져 있으며, 입력으로는 사전 학습된 모델이 추출한 2D 키포인트 시퀀스가 적용된다. 추정 대상인 17개 관절의 정의는 그림 5와 같다.

0- Head-top
1- Neck
2- Left shoulder
3- Left elbow
4- Left wrist
5- Right shoulder
6- Right elbow
7- Right wrist
8- Left hip
9- Left knee
10- Left ankle
11- Right hip
12- Right knee
13- Right ankle
14- Hip
15- Spine
16- Head

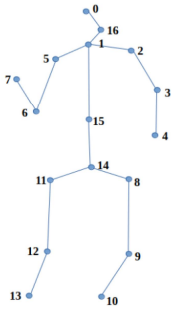

그림 5. 키포인트 라벨
Fig. 5. Keypoint label.

두 데이터셋의 train set의 모든 시퀀스를 human3.6 format으로 변환 후 학습에 사용되었으며 타 모델과의 정량적 비교를 위해 MPI-INF-3DHP 테스트 시퀀스 s1,s2를 평가 데이터로 사용하였다.

3. 실험 결과 및 고찰

본 연구에서 제안한 모델의 epoch에 따른 손실값의 지표는 다음과 같다.

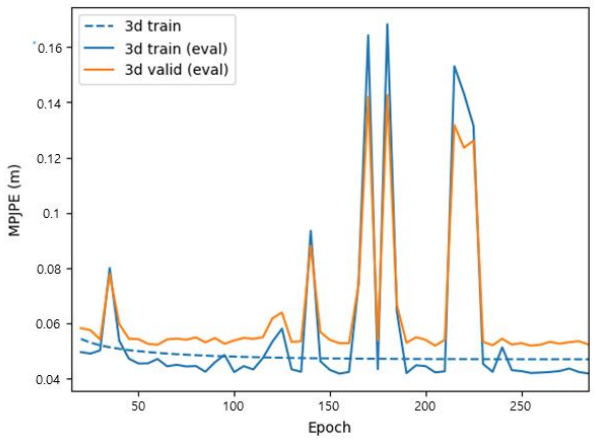

그림 6. 학습에 따른 손실 값
Fig. 6. Loss value according to learning.

가. 정량적 평가 지표

본 연구에서는 제안하는 모델과 비교군간의 3D 자세 추정 성능을 객관적으로 검증하기 위해 MPJPE와 PCK(Percentage of Correct Keypoints)[9], FLOPs 세 가지 정량적 지표를 사용한다.

MPJPE 는 3D 인체 자세 추정 분야에서 가장 널리 사용되는 표준 평가 지표이다. 이는 예측된 3D 관절 좌표와 실제 정답 좌표 간의 유클리드 거리의 평균을 계산한 값으로, 일반적으로 밀리미터(mm) 단위로 표현된다. MPJPE 값이 낮을수록 모델이 실제 자세와 오차가 적은 정밀한 추정을 수행했음을 의미한다.

MPJPE가 평균적인 오차를 측정한다면, PCK는 모델의 추정 결과가 허용 가능한 범위 내에 들어왔는지를 판단하여 모델의 강건성과 신뢰도를 평가하는 지표이다.

PCK는 예측된 관절 위치와 실제 위치 사이의 거리가 특정 임계값 이내인 관절의 비율을 백분율(%)로 나타낸다.

FLOPs는 모델의 계산 복잡도를 의미한다. 모델이 한번 추론시 실행되는 부동 소수점 연산의 횟수를 의미한다.

또한 DCT가 노이즈를 얼마나 억제하는지를 확인하기 위해 DCT과정을 제거한 모델간 가속도 오차를 비교한다. 가속도 오차는 예측 관절 궤적의 가속도와 실제 정답 가속도 차이를 측정하며 떨림이 심할수록 값이 높다.

나. 평가 및 비교

본 연구에서는 제안하는 모델의 정량적 평가 결과 및 타 모델과의 비교는 다음 표 1과 같다.

표 1. 제안된 기법의 성능평가
Table 1. Performance evaluation of the proposed method.

Method	MPJPE(mm)↓	PCK(150mm)↑	GFLOPs↓
Vnect[10]	124.7	76.6	13.7
P-STMO[11]	53.2	97.9	1.74
PoseFormer V1	77.1	88.6	1.36
MHformer[12]	58.0	93.8	31.1
The proposed method	**52.9**	**94.7**	**1.05**

실험 결과, 제안하는 모델은 대부분의 비교군 대비 우수한 성능 지표를 달성하였으며, 시공간 정보를 단순히 concat하여 self-attention을 수행한 PoseFormerV1 대비 우수한 성능을 보여주었다. 이는 시공간 통합 트랜스포머가 시간과 공간 정보를 분리하여 학습하던 기존 방식보다 특징(Feature)을 융합하는 데 있어 더욱 효과적임을 시사한다. 다만, P-STMO 모델에 비해서는 PCK(150MM)에서 낮은 성능을 보였는데, 이는 해당 비교 논문이 마스크(Masked) 사전 학습 전략을 통해 신체 가려짐 문제를 보다 강건하게 해결했기 때문으로 분석된다. 본 논문의 모델에서 DCT 과정을 제거하여 학습한 모델간의 가속도 오차는 다음 표 2와 같으며 이는 DCT 전처리 및 필터링이 효과적으로 노이즈를 억제함을 의미한다.

표 2. DCT를 제거한 모델과의 가속도 오차
Table 2. Acceleration error from the non-DCT model.

Method	non-DCT model	**The proposed method**
Acceleration Error (mm/s^2)	21.7	**10.2**

최종 모델의 입출력 결과는 그림 7과 같다.

Ⅳ. 결 론

본 논문에서는 DCT를 활용한 주파수 영역 분석과 시공간 통합 트랜스포머를 통해 단일 인물 3D 키포인트

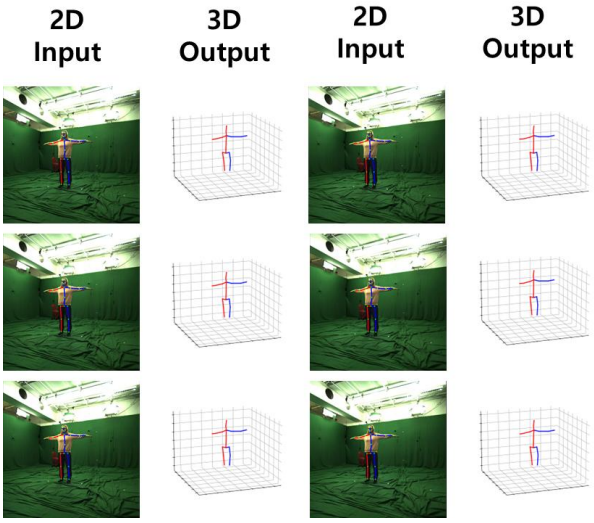

그림 7. 제안하는 모델의 2D 입력과 3D 출력
Fig. 7. 2D input and 3D output of the proposed model.

추정 성능을 개선하는 모델을 제안하였다. 제안된 모델은 불필요한 고주파 성분을 필터링하고 시공간 정보를 효율적으로 통합함으로써, 기존 방법론 대비 향상된 예측 정확도를 보여주었다. 비록 가려짐이 심한 환경에서는 사전 학습된 모델 대비 일부 성능의 한계가 있었으나, 전반적인 추정 성능에서 본 모델의 유효성을 확인할 수 있었다. 향후 연구에서는 가려짐에 강건한 사전 학습 기법을 본 모델에 도입하여, 더욱 완성도 높은 3D 자세 추정 기술로 발전시킬 수 있을 것이다.

REFERENCES

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[2] Zheng, Ce, et al. "3d human pose estimation with spatial and temporal transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
[3] Ahmed, Nasir, T_ Natarajan, and Kamisetty R. Rao. "Discrete cosine transform." IEEE transactions on Computers 100.1 (2006): 90-93.
[4] Han, Kai, et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 87-110.
[5] Kingma, Diederik P. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[6] Ionescu, Catalin, et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE transactions on pattern analysis and machine intelligence 36.7 (2013): 1325-1339.
[7] Mehta, Dushyant, et al. "Monocular 3d human pose estimation in the wild using improved cnn supervision." 2017 international conference on 3D vision (3DV). IEEE, 2017.
[8] Fieraru, Mihai, et al. "Learning complex 3D human self-contact." Proceedings of the AAAI Conference on Artificial Intelligence. 2021.
[9] Sapp, Ben, and Ben Taskar. "Modec: Multimodal decomposable models for human pose estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.
[10] Mehta, Dushyant, et al. "Vnect: Real-time 3d human pose estimation with a single rgb camera." Acm transactions on graphics (tog) 36.4 (2017): 1-14.
[11] Shan, Wenkang, et al. "P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation." European Conference on Computer Vision. 2022.
[12] Li, Wenhao, et al. "Mhformer: Multi-hypothesis transformer for 3d human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

저 자 소 개

신 인 영(학생회원)
2025년 국립한밭대학교 전자공학과 학사 졸업.
2025년~현재 국립한밭대학교 전자공학과 석사과정.
<주관심분야 : 자연어처리, 멀티모달, 딥러닝>

이 승 호(평생회원) - 교신저자
1986년 한양대학교 전자공학과 학사 졸업.
1989년 한양대학교 전자공학과 석사 졸업.
1994년 한양대학교 전자공학과 박사 졸업.
1994년~현재 국립한밭대학교 전자공학과 교수
<주관심분야 : 영상신호처리, 딥러닝, AR>