복합 낙상 시나리오 영상에서의 Transformer 모델 최적화 및 성능 평가

Optimizing and Evaluating a Transformer-based Model for Fall Detection in Composite Fall Scenario Videos*

김준석(주저자) · 이새롬(교신저자) · 박종화(공저자)
Joonseok Kim(First Author) · Saerom Lee(Corresponding Author) · Jonghwa Park(Co-Author)

경북대학교 경영학부 Department of Business Administration, Kyungpook National University(*ufo1112@knu.ac.kr*)
경북대학교 경영학부 Department of Business Administration, Kyungpook National University(*saeromlee@knu.ac.kr*)
경북대학교 경영학부 Department of Business Administration, Kyungpook National University(*jonghwapark@knu.ac.kr*)

초고령 사회의 도래와 함께 낙상은 노인의 생명과 건강을 위협하는 중대한 사회적 위험 요인으로 부상하고 있다. 본 연구는 복잡한 실제 환경에서 노인의 낙상을 효과적으로 감지하기 위한 실용적 영상 기반 시스템 구축의 기반이 되는 종단간(End-to-End) 낙상 감지 모델을 제안한다. 구체적으로, 본 연구는 기존 스켈레톤 추출 방식의 전처리 의존성과 오류 전파 문제를 극복하고자, 원본 RGB 영상만으로 작동하는 UniFormer 아키텍처를 개선하고 최신 학습 전략을 결합하였다. 특히 다양한 장소, 보조기구 사용, 촬영 각도 등을 포함한 AIHUB 데이터셋을 통해 모델의 일반화 가능성과 실효성을 평가하였으며, 복잡한 환경에서도 기존 모델 대비 높은 정확도(96.5%)와 F1-Score(93.2%)를 기록하였다. 본 연구의 사회적 함의는 다음과 같다. 첫째, 별도의 장비나 센서 없이 기존 CCTV 인프라를 활용할 수 있어 낙상 감지 기술의 보편적 확산 가능성을 제시한다. 둘째, 고비용 설비나 전문 인력 없이도 요양 시설 및 가정에서의 고령자 안전 모니터링 체계를 구축할 수 있어 사회적 돌봄 비용 절감과 돌봄 공백 최소화에 기여할 수 있다. 셋째, 장비 착용이나 인위적 개입 없이도 일상생활 속에서 자연스럽게 낙상을 감지할 수 있도록 설계되어, 고령자가 감시받는 존재가 아닌 자율적 주체로 존중받을 수 있는 환경 조성에 기여한다. 마지막으로 정보기술이 사회적 약자를 위한 안전망 구축에 기여할 수 있는 새로운 방향성을 제시하며, 노인복지정책과 스마트케어 인프라 설계에도 유의미한 방향성을 제공할 수 있다.

주제어: 노인 안전, 낙상 감지, 트랜스포머 아키텍처, 유니포머, 영상 분석

As the global population ages, falls represent a significant health risk for the elderly. This study aims to propose a high-performance, end-to-end fall detection model designed to serve as a core component for practical, vision-based monitoring systems in real-world environments. We introduce an optimized Transformer-based architecture that detects falls directly from raw RGB video streams, thereby obviating the need for extensive data pre-processing or wearable sensors. The model's generalizability and effectiveness were rigorously evaluated using the AIHUB dataset, which encompasses diverse scenarios, including varied locations and the use of assistive devices. The proposed model achieved an accuracy of 96.5% and an F1-score of 93.2%, demonstrating robust performance even under challenging conditions. The implications of this work are threefold. First, the system can be deployed on existing camera

최초투고일: 2025. 06. 16 수정일: (1차: 2025. 07. 21) 게재확정일: 2025. 07. 22
* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF- 2023S1A5A8079952).

infrastructure, offering a scalable and cost-effective solution for continuous monitoring. Second, by enabling automated monitoring in residential and care facilities, it has the potential to reduce caregiving costs and address service gaps. Third, the non-intrusive nature of the system preserves the privacy and autonomy of individuals. This research contributes significantly to the development of technology-driven safety nets for vulnerable populations and offers practical considerations for senior welfare policies and the design of smart care infrastructure.

Keyword: Elderly Safety, Fall Detection, Transformer Architecture, UniFormer, Video Analysis

Ⅰ. 서 론

전 세계적으로 초고령 사회(UN 기준 65세 이상 인구 비율 20% 이상) 진입이 가속화됨에 따라 노인 인구의 건강 유지 및 안전 확보는 주요 사회적 과제로 대두되었다. 노년기의 주요 건강 위협 요인인 낙상은 개인의 독립적 생활 능력 저하, 심각한 후유증, 나아가 사망에 이를 수 있는 중대한 문제이다. 세계보건기구(World Health Organization, 2021)는 낙상을 교통사고에 이어 비의도적 상해 사망의 두 번째 주요 원인으로 보고하였다. Kim et al.(2025)의 최근 연구에 따르면 85세 이상 초고령층의 사망률은 1992년 대비 2021년에 급격히 증가하였으며, 이러한 사망률이 2040년에는 인구 10만 명당 19.48명까지 증가할 것으로 전망하였다. 낙상은 개인의 건강 문제를 넘어 의료비 지출 증가, 장기 요양 필요성 증대 등 사회경제적 부담으로 이어질 수 있으며(Florence et al., 2018), 낙상 경험이나 두려움은 노인의 활동 위축 및 사회적 고립을 심화시켜 삶의 질을 저하를 초래할 수 있다(Xu et al., 2024). 또한 낙상 후 장시간 방치되는 경우 심각한 이차적 건강 문제로 이어질 위험이 크다(Blackburn et al., 2022; El-Bendary et al., 2013). 따라서, 낙상 예방 및 신속 대응 기술 개발의 필요성이 더욱 강조된다.

낙상 문제를 해결하기 위한 기술은 주로 웨어러블 센서와 영상 분석, 두 가지 방향으로 발전해왔다. 그 중 웨어러블 센서는 가속도계나 자이로스코프 등으로 사용자의 움직임을 감지하는 방식으로, 특정 조건에서 높은 정확도를 보인다는 가능성이 확인되었다. 하지만 높은 기기 비용과 착용 및 배터리 관리에 따르는 번거로움은 노인 사용자의 수용성을 떨어뜨리는 현실적인 한계로 작용했다(Knowles and Hanson, 2018). 경영학적 관점에서 이는 디지털 헬스케어 서비스의 핵심 접점(touchpoint)에서 사용자의 인지적, 정서적 경험을 저해하는 요인으로 해석할 수 있으며, 궁극적으로 서비스 만족도에 부정적인 영향을 미친다(정옥경 외, 2024). 이러한 문제점들은 웨어러블 기술의 광범위한 확산과 상용화를 가로막는 주요한 장벽이 되었다(Ren and Peng, 2019). 이러한 웨어러블 방식의 한계를 극복할 대안으로, 최근에는 별도의 장치 착용 없이 영상을 통해 낙상을 감지하는 연구가 활발히 이루어지고 있다(Gutiérrez et al., 2021). 영상 기반 접근법은 별도의 신체 부착물이 필요 없어 사용자 편의성이 높고, 기존 카메라 인프라를 활용할 수 있다는 장점이 있다. 그러나 초기 기술은 배경 제거 및 실루엣 추출과 같이 연산량이 많은 전처리 과정이 요구되어 실시간 시스템 구축에 어려움이 있었다(Hoang et al., 2023; Yu et al., 2017). 또한, 기존의 영상 기반 연구들은 대부분 통제된 실험

실이라는 제한된 환경에서 수집한 데이터를 주로 활용하였다. 이러한 데이터는 단조로운 배경에서 촬영되고, 가려짐이나 보조기구 사용 등 현실적인 변수들이 대부분 배제되어 있다. 따라서 이러한 데이터로 학습된 모델의 일반화 성능은 복합적인 시나리오에서 유효성을 보장하기 어렵다는 한계가 지적된다.

인공지능 및 영상 처리 기술 발전에도 불구하고 노년층이 자신들의 생활공간에서 용이하게 활용 가능한 효과적인 낙상 감지 서비스는 아직 부족하다(Wang et al., 2020). 이는 낙상 관련 기술의 상용화를 위해 추가적인 연구 개발이 필요함을 시사한다. 따라서 본 연구는 낙상 사고의 신속한 감지 및 조기 대응 기술 개발을 목표로, 현재 기술 동향 분석을 통해 실용성과 상용화 가능성을 높인 영상 처리 기반 낙상 감지 모델의 최적화 방안을 제시한다. 본 연구는 기존 영상 기반 기술의 한계, 즉 복잡한 전처리와 연구실 데이터 의존성을 극복하는 데 중점을 둔다. 이를 위해 영상 이해(Video Understanding) 분야에서 효과적인 성능을 보이며 전처리 의존도를 낮출 수 있는 트랜스포머(Transformer) 아키텍처를 낙상 감지 모델의 기반으로 채택한다. 선행 연구 중 Núñez-Marcos와 Arganda-Carreras(2024)는 UniFormer(Li et al., 2022)을 사용하여 원본 RGB 영상만으로 UP-Fall 및 UR Fall 데이터셋에서 낙상 감지를 수행하며 트랜스포머 기반 접근의 가능성을 확인하였으나, 해당 연구의 데이터셋(UP-Fall, UR Fall)은 소수의 건강한 피실험자를 대상으로 단일 실험실 환경에서만 수집되어, 다양한 장소나 보조기구 사용과 같은 현실의 복합적인 변수들을 반영하는 데에는 명백한 한계를 보였다.

따라서 본 연구는 선행 연구의 한계를 넘어, 다양한 환경과 보조기구 사용 등 복합적인 낙상 시나리오를 재현한 데이터셋을 활용하여 모델의 성능과 강건성을 검증하고자 한다. 이를 위해 AIHUB 플랫폼에서 제공하는 데이터를 사용하였다. AIHUB는 한국지능정보사회진흥원(NIA)이 운영하는 한국 정부의 오픈 데이터 허브 사이트이다(한국지능정보사회진흥원, 2021). 본 연구에서는 해당 플랫폼의 '낙상사고 위험 동작 영상-센서 쌍 데이터셋'(이하 'AIHUB 데이터셋')을 활용하였다(한국지능정보사회진흥원, 2023). 이 AIHUB 데이터셋은 병원, 가정, 요양 시설 등 다양한 환경에서 촬영되었고, 보조기구 사용자를 포함하여 기존 연구실 데이터셋 대비 현실성이 높으며, 복잡한 환경의 시각적 복잡성을 내포하여 모델의 일반화 성능 평가에 중요한 가치를 지닌다. 본 연구는 특히 여러 촬영 각도 중 기존 스켈레톤 기반 모델에서 성능 확보가 어려웠던 것으로 보고된 CAM1 각도 데이터에 집중하여, 개선 모델이 까다로운 조건에서도 안정적인 성능을 보이는지 검증한다. 기반 모델로는 UniFormer를 채택하고, 모델 아키텍처에는 Group Normalization(Wu and He, 2018), Pre-Layer Normalization 등 최신 기법을 적용하였다. 학습 전략 측면에서는 Lion 옵티마이저(Chen et al., 2023)와 Lookahead 기법(Zhang et al., 2019)을 결합하고, Focal Loss(Lin et al., 2017)를 적용하였으며, OneCycleLR 스케줄러(Smith, 2018)를 사용함으로 최신 접근 방식을 적용하여 성능 최적화를 시도하였다. 본 연구는 제안하는 개선 방식을 통해 복잡한 전처리 없이 원본 RGB 영상만으로도 복합 낙상 시나리오 데이터셋에서 높은 신뢰도의 성능을 확보할 수 있음을 실험적으로 입증한 데에 의의가 있다.

Ⅱ. 문헌연구

2.1 낙상관련 기존 연구

노인 낙상 연구는 크게 두 가지 접근법으로 나뉜다. 첫째는 낙상 예방(Fall Prevention)이다. 이 기술은 낙상 발생 전 혹은 진행 과정에서 위험 징후를 예측하는 것을 목표로 한다. 예측 정보를 바탕으로 에어백 등 보호 장치를 활성화하거나 사용자에게 경고함으로써, 낙상을 방지하고 피해를 최소화한다(El-Bendary et al., 2013; Ren and Peng, 2019). 다만, 모든 낙상을 예방하는 것은 현실적인 어려움이 있다. 낙상 예방의 한계로 인하여 제시된 두 번째 접근법에는 낙상 감지(Fall Detection) 기술이 있다. 낙상 감지 기술은 예방 노력에도 불구하고 발생한 낙상 이후, 개인의 움직임이나 환경 변화를 분석하여 사건을 식별한다. 이를 통해 보호자나 응급 서비스에 신속히 알려 즉각적인 대응을 유도한다(El-Bendary et al., 2013; Ren and Peng, 2019). 특히 낙상 후 장시간 방치되는 상황(Long Lie)은 심각한 2차 건강 문제로 이어질 수 있으므로, 발생한 낙상을 신속하고 정확하게 인지하여 적시 지원을 가능하게 하는 감지 기술은 노인 생명과 건강 유지에 필수적이다(Alam et al., 2022; Ren and Peng, 2019).

이처럼 낙상 감지는 노인 인구의 안전과 건강 유지를 위한 핵심 기술로 인식되며, 주로 센서를 기반으로 한 연구가 주로 진행되었다(Alam et al., 2022; Ren and Peng, 2019). 연구에 적용된 센서는 크게 두 가지로 사용자가 신체에 직접 착용하는 웨어러블 센서와 사용자의 주변 환경에 설치되는 환경 센서(Ambient Sensors)로 나뉜다. 초기 연구는 주로 웨어러블 센서를 활용한 방식에 집중되었다. 사용자가 신체에 부착하는 가속도계, 자이로스코프 등이 대표적이며, 이는 신체의 갑작스러운 가속도 변화, 충격, 자세 변화 등을 측정하여 낙상을 판단했다(Al-qaness et al., 2024; El-Bendary et al., 2013; Ren and Peng, 2019; Ursul, 2024). 웨어러블 센서는 휴대성과 실시간 감지 가능성을 제공했으나, 지속적인 착용에 따른 불편함, 배터리 관리의 번거로움, 피부 자극, 사용자의 착용 망각 가능성 등은 실용적 확산의 제약 요인이었다(El-Bendary et al., 2013; Ren and Peng, 2019). 웨어러블 센서의 대안으로, 주거 환경에 설치되는 환경 센서에는 바닥 압력 센서, 적외선 센서, 음향 센서 등이 사용되었으며 개인의 움직임이나 주변 환경 변화를 통해 낙상을 감지하려 했다(Chawan et al., 2022; Ren and Peng, 2019). 이러한 방식은 착용의 불편함은 해소했지만, 초기 설치 비용 부담, 센서가 설치된 특정 공간으로 국한되는 감지 범위, 다인 거주 환경에서의 정확도 문제, 가구 배치 변경 시 재설치 필요성 등이 해당 기술을 제품으로 상용화되는 것을 어렵게 만들었다(El-Bendary et al., 2013; Ren and Peng, 2019).

2.2 비전 기반 낙상 감지 연구 동향

센서 기반 낙상 감지 모델이 가지고 있는 다양한 한계를 극복하고자 최근에는 카메라 영상을 활용하는 비전 기반 낙상 감지 연구가 활발히 진행되고 있다(Gutiérrez et al., 2021). 초기 비전 연구는 영상처리 기법을 통해 사람의 실루엣의 변화를 분석하거나(Mobsite et al., 2023), 프레임 간 광학 흐름(Optical Flow)을 계산하여, 물체의 이동 방향과 속도 등 움직임의 특징을 추출하는 방식을 사용했다(Yu et al., 2017). 그러나 이러한 방법들은 조명, 배경, 가려짐 등 환경 변화에 민감하며, 특징 추출을

위한 추가 연산이 필요하다는 단점이 있다. 이후 딥러닝 기반 인체 포즈 추정 기술의 발전으로 스켈레톤(관절 키포인트) 정보를 추출하는 방식과 원본 RGB 영상에서 직접 시공간 특성을 학습하는 방법으로 낙상 감지 방법론이 개발되고 있다. 〈Table 1〉에서는 최근 연구된 비전 기반 낙상 감지 연구를 제시하고 있다.

먼저, 영상에서 스켈레톤 정보를 추출하고 이 시계열 데이터를 분석하는 방식이 비전 기반 낙상 감지의 주요 접근 방식으로 자리 잡았다(Hoang et al., 2023). 스켈레톤 시퀀스는 인체의 구조와 동적 움직임을 간결하게 표현하며, 합성곱 신경망(Convolutional Neural Network, CNN)(Suarez et al., 2022), 순환 신경망(Recurrent Neural Network, RNN)의 일종인 장단기 메모리(Long Short-Term Memory, LSTM)이나 게이트 순환 유닛(Gated Recurrent Unit, GRU)(Inturi et al., 2022; Yadav et al., 2022), 그래프 합성곱 신경망(Graph Convolutional Network, GCN)(Zahan et al., 2025), 그리고 트랜스포머(McCall et al., 2024; Ramirez et al., 2023; Yu et al., 2017)와 같은 다양한 시퀀스 모델링 아키텍처의 입력으로 활용되어 우수한 성능을 보여주었다. 특히 트랜스포머는 장거리 시공간 의존성 모델링 능력으로 인해 이 분야에서 주목받는 아키텍처로 부상했다. 그러나, 스켈레톤 기반 방법들은 여전히 '포즈 추정'이라는 별도의 사전 처리 단계에 의존한다. 이 단계는 실시간 시스템 구현 시 계산 비용과 지연 시간 증가의 원인이 될 수 있으며, 더 중요하게는 포즈 추정 과정에서 발생하는 오류가 후속 낙상 분류 모델의 성능을 저하시킬 수 있는 오류 전파(Error Propagation) 문제를 야기할 수 있다(Hoang et al., 2023). 또한, 스켈레톤 정보는 원본 영상이 담고 있는 풍부한 시각적 컨텍스트 정보를 손실시킬 수 있다.

이러한 문제를 해결하기 위한 대안으로, 원본 RGB 영상에서 시공간 특징을 직접 학습하는 종단간(End-To-End) 방식의 연구가 최근 주목받고 있다. 비전 트랜스포머(Vision Transformer)와 그 변형 아키텍처 영상 자체로부터 복잡하게 얽힌 시공간적 패턴을 직접 학습함으로써, 낙상처럼 미묘한 움직임 변화가 포함 동작도 효과적으로 구분해낼 수 있는 표현(Representation)을 학습할 수 있다는 점에서 강점을 가진다. 이를 통해 스켈레톤 정보 추출과 같은 중간 단계 없이도 높은 분류 성능을 달성하고, 오류 전파 가능성을 줄이며 전체 시스템의 파이프라인을 단순화하는 데 기여한다. Núñez-Marcos와 Arganda-Carreras(2024)는 비디오 처리에 특화된 UniFormer 아키텍처를 사용하여, 별도의 전처리 없이 RGB 영상만으로 낙상 감지를 수행하는 가능함을 보여준다.

그러나 이러한 기술적 발전에도 불구하고, 기존 비전 기반 낙상 감지 연구는 활용 데이터셋의 한계를 지닌다. 대부분의 데이터셋은 대학 연구 시설이나 특정 실내 공간에서 수집되어, 병원, 가정, 요양시설, 외부 환경 등 낙상 발생 환경의 다양성을 반영하지 못하였으며, 노인이 사용하는 보조기구(이동형 수액걸이, 지팡이, 휠체어, 목발, 보행기 등) 사용 상황 또한 대부분 고려되지 않았다. 데이터셋의 이러한 대표성 부족은 학습된 모델이 복합적인 시나리오에서 신뢰도 높은 성능을 발휘하기 어렵게 만드는 근본적인 한계이다.

Núñez-Marcos와 Arganda-Carreras(2024)가 활용한 UP-Fall 데이터셋(Martinez-Villaseñor et al., 2019)은 18세에서 24세 사이의 피실험자 17명을 대상으로 멕시코의 한 대학 연구실 환경에서 수집된 데이터로 보조기구 사용 상황을 포함하지 않는다. UR-Fall 데이터셋(Kwolek and Kepski, 2014)

〈Table 1〉 최근 연구된 비전 기반 낙상 감지 연구

기존 문헌	카메라 종류	종류	데이터셋	알고리즘	실시간성
Keskes and Noumeir(2021)	Kinect v2	Skeleton	NTU RGB-D(40명), TST v2 (11명), Fallfree(2명)	ST-GCN	언급없음
Ramirez et al. (2021)	Camera (RGB)	Skeleton	UP-Fall(17명)	ML	언급없음
Chutimawattanakul and Samanpiboon (2022)	Camera (RGB)	Posture classes	MCFD(1명), Le2i(9명)	YOLO + LSTM	언급없음
Liu et al. (2022)	Camera (RGB)	Skeleton	자체수집(3명), UR-Fall (5명), FDD(N/A)	LSTM	O
Salimi et al. (2022)	Camera (RGB)	Skeleton	UP-Fall(17명)	YOLO	O
Bhavani and Ukrit (2023)	Kinect	RGB frames	UP-Fall(17명)	GMM	언급없음
Gao et al. (2023)	Camera (RGB)	RGB frames with skeleton	Le2i(9명), UR-Fall(5명)	CNN	언급없음
Luo(2023)	Camera (RGB)	RGB frames	자체수집(N/A)	YOLO	O
Alanazi et al. (2024)	Camera (RGB)	Segmented+ fused video frames	Le2i(9명), MCFD(1명), UR-Fall(5명)	3D CNN + SVM	언급없음
Bui et al. (2024)	Camera (RGB)	RGB frames	자체수집(N/A)	YOLO + LSTM	O
Cai et al. (2024)	Camera (RGB)	RGB frames	Le2i(9명), UR-Fall(5명)	**Transformer**	언급없음
Cao et al. (2024)	Kinect v2	Skeleton + Optical flow	자체수집(16명)	CNN	언급없음
Ergüder et al. (2024)	Camera (RGB)	Skeleton	UR-Fall(5명), UP-Fall (17명), Le2i(9명), NTU RGB-D(40명)	ShuffleNet V2	O
Kaur et al. (2024)	Camera (RGB)	RGB frames	UMAFall(19명), Le2i(9명)	Haar Cascade	O
Núñez-Marcos and Arganda-Carreras (2024)	Camera (RGB)	RGB frames	UP-Fall(17명), UR-Fall(5명)	**Transformer**	O
Su et al. (2024)	Camera (RGB)	RGB frames	UR-Fall(5명), MCFD(1명)	CNN+LSTM	언급없음
Sykes(2024)	Camera (RGB)	Skeleton	자체수집(9명)	**Transformer**	O
Assanovich and Kosarava(2025)	Camera (RGB)	Bounding-box motion parameters	UP-Fall(17명), CAUCAFall(10명)	GBM	O
Wang et al. (2025)	Camera (RGB)	RGB frames	FPID(N/A), PFDD(N/A)	YOLO	O
Xun et al. (2025)	Camera (RGB)	RGB frames	자체수집(N/A)	YOLO	O
Yu et al. (2025)	Camera (RGB)	Skeleton	UP-Fall(17명), Le2i(9명), GMDCSA-24(4명)	**TCN + Transformer**	O

또한 5명의 피실험자로 구성되고 사무실이나 교실과 같이 예측 가능한 구조의 특정 실내 환경에서 촬영되어 생활공간의 다양성을 반영하기에는 한계가 있다. 마찬가지로 이 데이터셋도 보조기구 사용 상황은 고려되지 않았다. 뿐만 아니라, 널리 참조되는 다른 데이터셋 역시 유사한 혹은 추가적인 문제점을 드러낸다. Sykes(2024)는 MCFD 데이터셋에 대해 극심한 어안 렌즈 왜곡, 단일 피실험자로 인한 다양성 부족, 그리고 보조기구 사용 데이터 부재를 지적했다. 또한 Le2i 데이터셋은 낮은 해상도로 인해 포즈 추정 알고리즘의 정확도에 부정적 영향을 미칠 수 있으며, 역시 보조기구 사용 상황을 포함하지 않는다고 언급되었다. 따라서 모델의 높은 일반화 성능과 강건성을 확보하려면, 보다 현실적인 시나리오에 기반한 데이터셋을 활용해야 한다.

2.3 UniFormer 아키텍처 상세

본 연구에서는 비디오 데이터의 핵심 특징 추출기(Backbone Network)로 UniFormer(Li et al., 2022)를 채택하였다. 비디오의 시공간 표현 학습은 고유한 어려움을 수반하는데, 특히 Li et al.(2022)은 UniFormer를 제안하며 지역적 시공간 중복성(Local Spatiotemporal Redundancy)과 복잡한 시공간 의존성(Complex Spatiotemporal Dependency)을 비디오 데이터의 근본적인 두 가지 문제점으로 지적하였다. 지역적 시공간 중복성은 인접 프레임 간 미미한 변화로 인한 정보 중복 문제로 불필요한 계산 비용을 야기하는 것을 의미한다(Li et al., 2022). 복잡한 시공간 의존성은 특정 행동 이해를 위해 시간적으로 멀리 떨어진 프레임 간의 관계, 즉 장거리 의존성(Long-Range Dependency) 포착이 필수적이라는 특성을 가지고 있다(Li et al., 2022).

비디오 데이터의 두가지 과제에 대응하기 위해 3D 컨볼루션 신경망(Three-Dimensional Convolutional Neural Network, 3D CNN)과 비전 트랜스포머 방식이 연구되었다. 3D CNN은 3차원 컨볼루션 필터로 지역적 컨텍스트 정보를 집계하여 지역적 중복성을 줄이는 데 강점을 보이나(Carreira and Zisserman, 2017; Tran et al., 2015), 제한된 수용 영역으로 인해 장거리 의존성 포착에는 어려움을 겪는다(Li et al., 2020; Li et al., 2022; Wang et al., 2018). 반면, 비전 트랜스포머는 셀프 어텐션(Self Attention) 메커니즘을 기반으로 장거리 의존성 포착에 매우 효과적이지만(Arnab et al., 2021; Bertasius et al., 2021; Dosovitskiy et al., 2021), 모든 계층에서 모든 토큰 간 유사도 비교를 수행하는 방식은 특히 지역적 패턴이 지배적인 초기 계층에서 비효율적일 수 있다(Bertasius et al., 2021). 또한, 셀프 어텐션의 이차적 계산 복잡도는 고해상도 또는 긴 비디오 처리에 부담이 된다(Li et al., 2022). 따라서, 이 두 접근 방식의 장점을 취하고 단점을 보완한 방식으로 UniFormer는 비디오의 지역적 중복성과 전역적 의존성을 효과적이면서 효율적으로 처리하도록 설계되었다(Li et al., 2022). UniFormer은 비디오 이해(Video Understanding) 분야에서 높은 정확도와 계산 효율성 간의 바람직한 균형을 달성하고자 제안된 모델이다. 이는 원본 RGB 영상으로부터 직접 시공간 특징을 학습하는 본 연구의 종단간 접근 방식에 부합하며, 복잡한 전처리 단계를 생략하여 실시간 처리 가능성을 높일 수 있다.

UniFormer는 입력 비디오 클립(Clip)에 먼저 초기 3D 컨볼루션 계층(Stem Convolution)을 적용하여 시공간 해상도를 줄이고 기본적인 특징을 추출한다(Li et al., 2022). 이후, 처리된 특징 맵을 공간 및 시간 차원에서 겹치지 않는 균일한 크기의

3D 패치(Patch) 시퀀스로 분할한다. 각 3D 패치는 평탄화(Flatten) 후 학습 가능한 선형 사영(Linear Projection) 계층을 통해 모델의 내부 차원(C)을 갖는 토큰(Token)으로 변환된다. 이 과정을 거쳐 비디오 클립은 L × C 크기의 토큰 시퀀스 텐서(Token Sequence Tensor)로 인코딩되어 후속 UniFormer 블록들의 입력이 된다(Li et al., 2022).

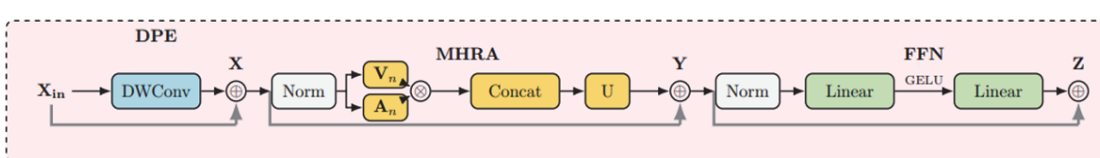
생성된 토큰 시퀀스 텐서는 UniFormer 네트워크의 핵심 처리 단위인 UniFormer 블록들을 계층적으로 통과한다. 이 과정에서 토큰들은 점진적으로 주변 시공간 컨텍스트 정보를 통합하며, 초기에는 지역적이고 세밀한 특징을, 후반에는 전역적이고 추상적인 시공간 특징 표현으로 변환된다. Figure 1에서 도식화 한 바와 같이 각 UniFormer 블록은 입력 토큰 X_{in}에 대해 잔차 연결을 포함하며, Dynamic Position Embedding (DPE), Multi-Head Relation Aggregator(MHRA), 그리고 Feed-Forward Network(FFN)의 세 가지 주요 구성 요소를 통해 순차적으로 다음 연산을 수행한다(Li et al., 2022):

$$X = DPE(X_{in}) + X_{in},$$
$$Y = MHRA\ (Norm\ (X)) + X,$$
$$Z = FFN\ (Norm\ (Y)) + Y.$$

여기서 Norm은 정규화 계층(Normalization Layer)을 나타낸다.

DPE는 각 토큰에 시공간적 위치 정보를 명시적으로 인코딩한다. 기존 고정된 위치 임베딩 방식이 입력 비디오 클립 길이 변화에 유연하게 대처하기 어려운 문제를 해결하기 위해(Arnab et al., 2021; Bertasius et al., 2021; Islam et al., 2020), DPE는 간단한 3D 깊이별 컨볼루션(3D Depth-Wise Convolution) 연산을 활용한다(Li et al., 2022). 이는 입력 토큰 시퀀스 길이에 구애받지 않고 유연하게 적용 가능하며, 제로 패딩과 결합 시 모든 토큰이 자신의 절대적 시공간 위치 정보를 점진적으로 인코딩할 수 있게 한다(Chu et al., 2021).

MHRA는 UniFormer의 핵심 구성 요소로서, 기존 트랜스포머의 Multi-Head Self-Attention (MHSA)을 대체하여 비디오 데이터 고유의 지역적 중복성과 전역적 의존성을 효과적으로 처리한다(Li et al., 2022). MHRA의 주요 특징은 네트워크 계층의 깊이에 따라 토큰 관계 학습 방식, 즉 토큰 어피니티 계산 방식을 다르게 적용한다는 점이다. 네트워크의 초기, 즉 얕은 계층(Shallow Layers)에서는 지역적 MHRA(Local MHRA)가 사용된다. 각 토큰 X_i는 자신의 작은 3D 시공간 이웃 내 토큰 X_j와만 상호작용하며, 토큰 어피니티는 두 토큰의 내용과 무관하게 오직 상대적인 3D 위치에 의해서만 결정되는 학습 가능한 파라미터로 정의된다(Li et al., 2022). 이 방식은 MobileNet 블록(Feichtenhofer, 2020; Sandler et al., 2018; Tran et al., 2019)과 유사한 계산 효율성을 제공하면서 3D 컨볼루션



〈Figure 1〉 UniFormer의 구조(Li et al., 2022)

처럼 지역 정보를 효과적으로 집계한다(Li et al., 2022).

반면, 네트워크 후반부의 깊은 계층(Deep Layers)에서는 Global MHRA가 사용된다. 각 토큰 X_i는 입력 시퀀스 내 모든 토큰 X_j와 상호작용하며, 토큰 어피니티는 두 토큰의 내용에 기반한 유사도를 스케일드 닷-프로덕트 어텐션(Scaled Dot-Product Attention)을 통해 동적으로 결정한다. UniFormer의 Global MHRA는 공간 어텐션과 시간 어텐션을 분리하지 않고 단일 연산으로 시공간 정보를 통합적으로 처리하는 통합 시공간 어텐션(Joint Spatiotemporal Attention)을 수행하여 표현력을 높인다. 얕은 계층에서 Local MHRA 사용으로 확보된 계산 여력 덕분에, 깊은 계층에서는 상대적으로 계산 비용이 높은 통합적 전역 어텐션 연산이 가능해진다(Li et al., 2022).

FFN는 MHRA 이후 적용되는 모듈로, 각 토큰 위치별로 독립적으로 작동하는 Multi-Layer Perceptron 구조이다. FFN은 일반적으로 두 개의 선형 계층과 비선형 활성화 함수로 구성되어 채널 차원을 확장 후 다시 원래 채널 차원으로 축소하는 병목 구조를 가지며, MHRA를 통해 집계된 시공간 컨텍스트 정보를 바탕으로 각 토큰의 특징 표현을 채널 차원에서 비선형적으로 변환하고 정제하여 모델의 표현 능력을 향상시킨다(Vaswani et al., 2017).

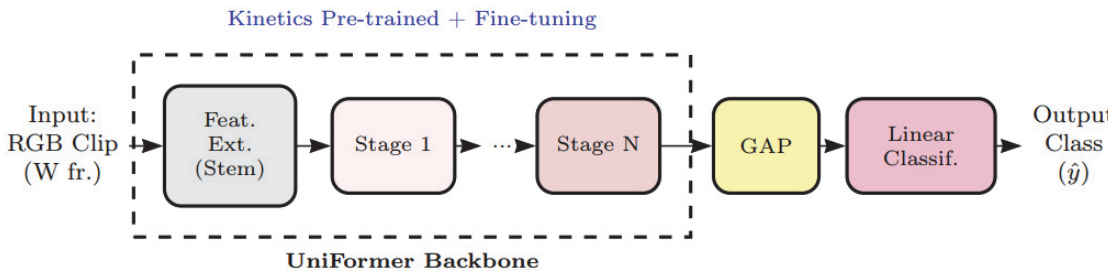
이러한 UniFormer 블록들은 전체 네트워크 아키텍처 내에서 계층적으로 구성된다. 네트워크는 통상 4개의 스테이지(Stage)로 나뉘며, UniFormer 논문에서 제안된 기본 구성에 따르면, 네트워크의 첫 두 스테이지에서는 Local MHRA를, 마지막 두 스테이지에서는 Global MHRA를 사용한다. 각 스테이지 사이에는 다운샘플링 연산(예: 1×2×2 컨볼루션)이 적용되어 특징 맵의 공간 해상도를 줄이고 채널 수를 증가시켜 피라미드 형태의 특징 계층 구조를 형성한다. 정규화 방식으로는 Local MHRA가 포함된 블록에서는 배치 정규화(Batch Normalization, BN)(Ioffe and Szegedy, 2015)를, Global MHRA가 포함된 블록에서는 레이어 정규화(Layer Normalization, LN)(Ba et al., 2016)를 사용하여 각 모듈의 특성에 맞는 안정적인 학습을 할 수 있게 한다(Li et al., 2022).

Ⅲ. 개선 모델

본 연구는 UniFormer를 기반 모델(Base Model)로 채택하고, 이를 AIHUB 데이터셋의 특성에 맞춰 아키텍처와 학습 전략을 최적화하는 아키텍처 개선 및 학습 전략을 제안한다. 본 연구에서 개선한 모델은 UniFormer의 핵심 장점인 '원본 RGB 영상 직접 처리' 방식을 계승하여 전처리 과정을 간소화하는 동시에, 최적화된 구조와 학습 방식을 통해 복합적인 시나리오에서 감지 성능과 강건성을 높였다. 〈Figure 2〉는 본 연구에서 채택한 낙상 감지 모델의 전체 프레임워크(Overall Framework)를 나타낸다.

3.1 모델 개요 및 기본 구조

본 연구의 낙상 감지 모델은 원본 RGB 영상을 직접 입력으로 받아 낙상 발생 여부를 판별하는 종단간 이진 분류 모델이다. 이 접근 방식은 기존 스켈레톤 기반 방식에서 요구되던 복잡한 포즈 추정 전처리 단계를 생략함으로써, 오류 전파 가능성을 줄이고 잠재적인 시각적 단서 활용을 극대화하며, 전체 시스템의 파이프라인을 단순화하여 실시간 처리의 기반을 마



〈Figure 2〉 UniFormer 기반 낙상 감지 모델의 전체 프레임워크

련한다. 모델의 핵심 특징 추출기(Backbone Network)로는 UniFormer를 채택하였다.

일상적인 상황에서 우연히 발생하는 낙상 관련 데이터는 대규모 수집이 어렵고, 낙상 장면 확보에는 윤리적, 현실적 제약이 존재한다. 이러한 제한된 데이터 문제를 극복하고 모델의 일반화 성능을 높이고자, 본 연구는 전이 학습(Transfer Learning) 전략을 활용하였다. 먼저, 대규모 비디오 행동 인식 데이터셋인 Kinetics-400(Carreira and Zisserman, 2017) 등에서 사전 훈련된 UniFormer의 가중치를 초기값으로 사용하였다. 이후 목표 과제인 AIHUB 데이터셋에 대해 모델을 미세 조정(Fine-Tuning)하였다. Kinetics-400 데이터셋은 다양한 인간의 일상 행동과 스포츠 동작 등을 포함하여, 일반적인 시공간 패턴과 동적 움직임에 대한 사전 지식을 모델에 제공한다. 이를 통해 모델은 정상 활동과 비정상 움직임, 즉 낙상을 구분하는 데 필요한 기본적인 시각적 특징 표현 능력을 확보하며, 상대적으로 적은 낙상 관련 데이터로도 효과적인 학습이 가능해진다.

모델의 최종 예측은 비디오 클립 단위로 수행하였다. UniFormer 백본 네트워크를 통과하여 추출된 최종 시공간 특징 표현은 Global Average Pooling 연산을 통해 각 클립을 대표하는 고정 크기 특징 벡터로 요약된다. 이 특징 벡터는 하나 이상의 완전 연결 계층(Fully-Connected Layer)과 활성화 함수로 구성된 선형 분류기(Linear Classifier)에 입력되어, 해당 클립이 낙상(Class 1)인지 비낙상(Class 0)인지에 대한 확률 값을 출력한다. 특히, 신속한 대응이 필수적인 실제 낙상 감지 응용 환경을 고려하여, 입력 비디오 스트림을 고정 프레임 수 W를 갖는 클립 단위로 순차 처리하는 슬라이딩 윈도우(Sliding Window) 방식을 적용하였다. 이러한 방식은 연속적인 데이터 스트림에서 지연 시간을 최소화하며 지속적인 모니터링을 가능하게 하여, 낙상 발생 시 모델이 즉각적으로 반응할 수 있다.

표준 UniFormer 아키텍처와 전이 학습은 낙상 감지 연구의 기반을 제공한다. 그러나 본 연구에서 활용하는 AIHUB 데이터셋은 다양한 현실적 환경의 복잡성을 모사한 다양한 시나리오를 포함하고 있다. 이러한 데이터셋 특성은 표준 모델 및 학습 방식의 성능 최적화에 한계를 야기하며, 모델 학습에 다음과 같은 도전 과제를 제시한다. 데이터셋 내 낙상 클래스와 비낙상 클래스 간에는 낙상 발생 빈도의 차이로 인해 클래스 불균형(Class Imbalance)이 존재하고, 이는 다수 클래스 편향 학습 및 소수 클래스 감지 성능 저하로 이어질 수 있다. 또한, 촬영 장소인 병원, 가정, 요양시설, 길거리 등은 조명, 배경, 카메라 각도 및 거리 변화를 포함하여 데이터의 시각

적 편차를 증가시킨다. 이러한 편차는 모델의 일반화 능력 확보를 어렵게 한다. 그리고 지팡이, 휠체어, 목발 등의 보조의료 기구 사용 유무 또한 움직임 패턴의 이질성을 초래한다. 이는 일관된 특징 학습 및 다양한 낙상 시나리오 인식의 어려움으로 작용한다. 본 연구는 AIHUB 데이터셋의 이러한 도전 과제에 대응하고 낙상 감지 성능을 최적화하기 위해, 표준 UniFormer 모델과 학습 전략을 수정 및 보완하였다. 다음 절은 적용된 아키텍처 변경 사항과 학습 전략 최적화 내용을 기술한다.

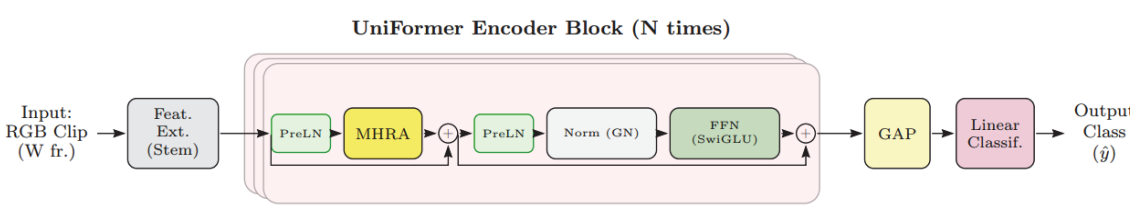
3.2 아키텍처 개선 사항

AIHUB 데이터셋이 가진 복합적인 특성에 효과적으로 대응하고 모델 성능을 극대화하기 위해, 본 연구는 표준 UniFormer 아키텍처의 핵심 구성 요소인 인코더 블록을 개선하였다. 이를 통해 모델의 표현력을 높이고, 학습 안정성을 확보하며, 미세한 움직임을 포착하는 능력을 향상시키고자 하였다. 〈Figure 3〉은 본 연구를 통해 개선된 UniFormer 인코더 블록의 상세 구조이다.

첫째, UniFormer 블록 내 피드포워드 네트워크(FFN)의 표현력 강화를 위해 활성화 함수를 개선하였다. 표준 UniFormer는 GELU 활성화 함수를 사용하나, 본 연구에서는 Shazeer(2020)가 제안한 SwiGLU 활성화 함수와 비디오 처리를 위한 3D 확장 버전인 SwiGLU3D를 FFN의 선형 계층 이후에 적용하였다. SwiGLU는 입력 벡터를 두 개로 분할하여, 하나는 시그모이드 함수를 통과시켜 게이트 신호를 생성하고 다른 하나는 선형 변환 후 이 게이트와 요소별 곱셈을 수행하는 게이팅 메커니즘을 활용한다. 이처럼 입력에 따라 활성화가 조절되는 SwiGLU의 특징은 모델이 복잡한 비선형성을 효과적으로 학습하고, 다양한 낙상 상황에서의 미묘한 움직임 변화를 더 잘 포착하도록 한다. 그 결과 최종적 낙상 감지 정확도가 향상을 기대하였다.

둘째, 모델 학습의 안정성 확보를 위해 정규화 방식을 최적화하였다. 표준 UniFormer는 Local MHRA에서 Ioffe와 Szegedy(2015)의 BN을, Global MHRA에서 Ba et al.(2016)의 LN을 사용한다. 그러나 BN은 작은 배치 크기에 민감하여, 본 연구에서와 같이 제한된 GPU 메모리로 인해 작은 배치 크기를 사용해야 할 경우 성능이 저하되는 경향이 있다. 따라서 Local MHRA를 포함하는 블록에서 BN 대신 Wu and He(2018)의 그룹 정규화(Group Normalization, GN)을 적용하였다. GN은 채널을 그룹화하여 정규화를 수행하므로 배치 크기에 민감하지 않으며, LN과 달리 채널 그룹 단위의 통계적 특성을 보존하여 작은 배치 크기 환경에서도 안정적인 학습과 우수한 성능을 제공할 잠재력을 가진다.

셋째, 트랜스포머 아키텍처의 학습 안정성을 향상시키기 위해, Vaswani et al.(2017)의 표준 트랜



〈Figure 3〉 개선된 인코더 블록을 포함한 모델의 전체 구조

스포머에 사용된 Post-Layer Normalization(Post-LN) 방식 대신 Pre-Layer Normalization(Pre-LN) 구조를 채택하였다. Post-LN 방식은 모델 깊이가 깊어질수록 학습 초기 그래디언트 불안정 문제를 겪을 수 있으며, 이를 해결하기 위해 학습률 워밍업과 같은 추가 기법이 요구되기도 한다. 반면, Pre-LN 구조는 LN을 각 서브모듈(본 연구의 MHRA 및 FFN) 입력단에 적용함으로써 잔차 연결을 통한 그래디언트 흐름을 원활하게 하여 학습 과정을 안정화시키는 것으로 알려져 있다(Xiong et al., 2020). 이러한 Pre-LN의 적용으로 깊은 네트워크 구조나 복잡한 학습 과제에서 더 빠른 수렴 속도와 견고한 학습 성능을 달성할 것으로 기대한다.

3.3 학습 전략 최적화

모델 아키텍처 개선에 이어, AIHUB 데이터셋에 최적화된 학습을 통해 낙상 감지 성능을 극대화하고자 다음과 같은 학습 전략을 적용하였다. 첫째, 모델 파라미터 최적화를 위해 표준 옵티마이저인 Adam (Kingma and Ba, 2014) 또는 AdamW (Loshchilov and Hutter, 2017) 대신, 최근 제안된 Lion 옵티마이저(Chen et al., 2023)를 기본으로 채택하였다. Lion은 일부 대규모 모델 학습에서 메모리 효율성과 함께 AdamW 대비 빠른 수렴 및 향상된 성능을 보이는 것으로 보고되었으나(Chen et al., 2023), 하이퍼파라미터 민감성과 잠재적인 학습 불안정성을 내포한다. 이러한 Lion의 단점을 보완하고 안정적인 학습을 도모하고자, Zhang et al.(2019)의 Lookahead 기법을 결합하여 적용하였다. Lookahead는 옵티마이저의 탐색 과정을 안정화시켜 일반화 성능 향상에 기여하는 것으로 알려져 있다(Zhang et al., 2019). 본 연구는 이 두 기법의 시너지를 통해 빠르면서도 견고한 모델 학습을 목표로 하였다.

둘째, 본 연구에서 사용하는 AIHUB 데이터셋은 4.2.2절에서 상세히 기술된 바와 같이 낙상 클래스(소수 클래스)와 비낙상 클래스(다수 클래스) 간의 현저한 샘플 수 차이, 즉 클래스 불균형 문제를 내포한다. 이러한 클래스 불균형은 표준 교차 엔트로피(Cross-Entropy, CE) 손실 함수 사용 시 모델이 다수 클래스에 편향되어 소수 클래스인 낙상을 놓치는 미탐지(False Negative, FN) 위험을 증가시킬 수 있다. 이러한 클래스 불균형 문제에 대응하고 미탐지율을 효과적으로 감소시키기 위해, 본 연구는 Lin et al.(2017)의 Focal Loss를 손실 함수로 채택하였다. Focal Loss는 잘 분류되는 다수 클래스 샘플의 손실 기여도를 낮추고 분류가 어려운 소수 클래스 샘플에 학습을 집중시키는 방식으로 CE 손실을 개선한다(Lin et al., 2017). 본 연구에서는 이 손실 함수의 집중 파라미터(Focusing Parameter, γ)를 2로 설정하여, 모델이 소수 클래스인 낙상 패턴에 대한 판별력을 높여 재현율 향상 및 미탐지율 감소를 달성하도록 하였다.

셋째, 본 연구에서 활용하는 전이 학습 및 복잡한 AIHUB 데이터셋의 특성을 고려하여, 학습률 스케줄링 전략으로 OneCycleLR (Smith, 2018)을 채택하였다. OneCycleLR은 학습률을 하나의 사이클 내에서 동적으로 크게 변화시키는 전략으로, 초기에는 낮은 학습률에서 시작하여 최대치까지 증가시킨 후 다시 점진적으로 감소시킨다. 이는 학습 초기의 탐색을 돕고, 높은 학습률 구간을 통해 정규화 효과를 제공하며, 이후 안정적인 수렴을 유도하여 특히 전이 학습이나 복잡한 데이터셋에서 빠른 수렴과 우수한 일반화 성능을 보이는 것으로 알려져 있다(Smith, 2018). 이를 통해 본 연구는 제한된 학습 자원 내에서도 효과적으로 모델을 최적화하고자 하였다.

Ⅳ. 방법론

4.1 데이터

AIHUB 데이터셋은 낙상 감지를 위한 대규모 데이터셋으로, 총 22,672개의 영상 클립과 226,720개의 이미지 데이터로 구성되며, 데이터는 낙상 여부(낙상/비낙상)를 이진 분류로 라벨링하였다. 이 데이터는 병원(24.14%), 집(28.62%), 노인정 또는 요양병원(46.05%), 길거리(1.20%) 등 다양한 환경에서 촬영되었으며, 촬영 각도는 CAM1부터 CAM8까지 8개로 균등하게 분포되어 있어 각 각도별로 2,266개(전체의 12.50%)씩 존재한다. 보조의료 기구의 경우, 전체 클립 중 약 84.40%는 보조의료 기구 없이 촬영되었으며, 이동형 수액걸이 또는 지팡이, 휠체어, 목발, 걸음보조기는 각각 2.43%, 5.82%, 4.38%, 2.96%의 비율을 차지한다. 또한 피검자의 총 인원은 40명이며 연령대는 청소년·청년(14~39세)이 24.70%, 중장년·노년층(40세 이상)이 75.30%를, 성별은 남성이 50.32%, 여성이 49.68%로 거의 균등하게 분포되어 있다. 본 연구에서는 이 데이터셋 중 CAM1 카메라로 촬영된 데이터를 선별하였다. 〈Table 2〉에 따르면 CAM1 데이터는 다른 촬영 각도에 비해 낙상 감지 성능이 가장 저조하게 나타났다. 이는 모델이 학습하고 추론하는 데 있어 CAM1 각도가 가장 까다로운 조건임을 의미한다. 따라서 본 연구에서는 모델 성능을 보다 명확하게 검증하기 위해, 초기 평가에서 가장 낮은 감지 성능을 보인 CAM1 카메라의 촬영 데이터를 선별하여 실험을 진행하였다.

4.2 평가 방법

4.2.1 실험 데이터셋 및 구성

본 연구의 모델 성능 평가는 AIHUB 데이터셋을 사용하였다. 영상 입력을 처리하고 사전 훈련된 UniFormer 가중치를 활용하기 위해, 입력 영상은 16프레임 클립 단위로 분할되었다. 원본 데이터셋은 고유한 ID(촬영 환경/피실험자 식별)로 구성된 폴더 단위로 제공된다. 동일 ID에서 파생된 클립들이 학습, 검증, 테스트 세트 중 서로 다른 세트에 중복되지 않도록 ID를 기준으로 데이터를 분할하였다. 데이터셋 내 고유 ID를 정렬 후 약 33개씩 묶어 총 68개 데이터 그룹으로 재구성하였으며, 이 그룹 단위로 전체 데이터셋을 학습, 검증, 테스트 세트 약 8:1:1 비율로 분할하였다.

4.2.2 분류 문제 정의 및 입력 처리

본 연구는 낙상의 발생 여부를 이진 분류 문제로 정의하고, 낙상 시퀀스는 클래스 1로, 그 외 활동은 클래스 0으로 레이블링하였다. 이러한 접근은 Espinosa

〈Table 2〉 AIHUB 베이스 모델의 AIHUB 데이터셋 낙상유무탐지모델 결과(영상데이터)

	F1-Score	Accuracy		F1-Score	Accuracy
CAM1	0.90	0.901	CAM5	0.99	0.993
CAM2	0.96	0.962	CAM6	0.99	0.988
CAM3	0.96	0.958	CAM7	0.98	0.979
CAM4	0.96	0.958	CAM8	0.98	0.979

et al.(2019)이 UP-Fall 데이터셋을 활용한 연구에서 클래스를 이진화하여 사용한 전략과 일치하며, 실제 시스템 적용 시 신뢰성 있는 판단을 가능하게 한다는 점에서 유의미하다. 모델 입력은 앞서 정의된 16프레임(또는 8프레임) 클립을 슬라이딩 윈도우 방식으로 순차 처리한다.

낙상 감지 모델의 실효성을 높이기 위해서는 미탐지를 줄이는 것이 중요하므로(Yu et al., 2017), 본 연구는 재현율(Recall) 또는 정밀도(Precision) 극대화를 우선 과제로 설정하였다. 이를 위한 레이블링 전략으로, 분석 대상 윈도우(16프레임 또는 8프레임) 내에 단 하나의 프레임이라도 낙상으로 레이블링된 경우, 해당 윈도우 전체를 '낙상(Class 1)'으로 판정하였다. 이 방식은 낙상 동작의 일부만 포함되거나, 발생 초기의 미약한 낙상 신호가 포함된 경우에도 낙상으로 분류될 가능성을 높여, FN을 줄이는 데 기여한다. 본 연구의 최종 학습 및 평가에 사용된 클립 단위 데이터셋은(16프레임 윈도우 적용 시) Non-Fall 클래스가 약 85.04%, Fall 클래스가 약 14.96%로 구성되어, Fall 클래스가 소수인 클래스 불균형 특성을 나타냈다.

4.2.3 평가지표 선정 기준

본 연구에서 사용하는 학습 데이터는 Non-Fall 클래스가 Fall 클래스에 비해 샘플 수가 많아 클래스가 불균형한 문제를 가진다. 이러한 데이터 분포에서는 일반적인 정확도 지표만으로 모델의 세부 분류 성능, 특히 각 클래스 예측 능력과 오류 유형별 중요도를 충분히 반영하기 어렵다(Ramirez et al., 2023; Wang et al., 2020). 더불어, 낙상 감지 모델은 현장 적용 시 분류 오류 유형에 따라 결과의 심각성이 달라지는 오류 비용(Error Cost)의 비대칭성을 내포한다. 이는 낙상을 감지하지 못하는 미탐지가 정상 활동을 낙상으로 오인하는 오탐지(False Positive, FP)보다 사용자에게 더 심각한 결과를 초래할 수 있음을 의미한다(Liu et al., 2022). FN 오류는 사용자 건강과 안전에 직접적인 위협이 될 수 있으며 응급 상황 발생 시 의료 지원 지연으로 이어질 수 있다. 반면, FP 오류는 주로 시스템 사용의 불편함, 불필요한 알람으로 인한 자원 낭비, 그리고 시스템 신뢰도 저하를 야기한다(Hoang et al., 2023). 이처럼 데이터의 클래스 불균형과 오류 비용의 비대칭성을 모두 고려하여 모델 학습 과정을 점검하고 최종 성능을 객관적으로 평가하기 위해서는, 문제 특성에 부합하는 평가지표 선정이 중요하다. 특히, 각 오류 유형의 발생 빈도를 정밀하게 반영하고 소수 클래스인 Fall에 대한 FN 최소화의 중요성을 평가할 수 있는 지표 사용이 필수적이다.

이러한 배경을 고려하여, 본 연구는 모델의 다각적인 성능 평가하고 정밀도, 재현율, 그리고 F1-Score를 주요 평가지표로 사용하였다. 정밀도는 모델이 낙상으로 예측한 샘플 중 실제 낙상인 샘플의 비율을 나타내어 오탐지의 정도를 평가한다. 재현율은 실제 발생한 모든 낙상 중 모델이 낙상으로 올바르게 식별한 비율을 의미하며, 이는 미탐지의 정도를 평가하는 지표로 낙상 감지 모델의 핵심 성능과 직결된다. 정밀도와 재현율은 일반적으로 상충 관계에 있으므로, 두 지표의 균형을 고려하는 것이 중요하다. F1-Score는 정밀도와 재현율의 조화 평균으로, 클래스 불균형 데이터에서 모델의 전반적인 성능을 효과적으로 나타낸다(Aderinola et al., 2024). 본 연구에서는 미탐지 감소라는 핵심 목표 달성을 위해 F1-Score를 주요 최적화 기준으로 활용하였다. 각 평가지표의 수식은 다음과 같다:

Precision = TP / (TP + FP),
Recall (Sensitivity) = TP / (TP + FN),
F1-Score = 2 * (Precision * Recall) / (Precision + Recall),

최종 성능 평가 시에는 각 클래스(낙상, 비낙상)에 대한 F1-Score를 개별 계산한 후, 가중치를 부여하지 않은 단순 평균인 Unweighted Mean F1-score를 최종 F1-Score로 사용하였다.

4.2.4 검증 세트를 이용한 모델 최적화 및 조기 종료

매 학습 에포크가 종료될 때마다 현재까지 학습된 모델을 사용하여 검증 세트에 대한 F1-Score를 계산함으로써 모델의 성능 변화를 추적하였다. 학습 과정 중 가장 높은 검증 F1-Score를 기록한 시점의 모델 가중치를 최적의 모델 상태로 간주하고, 이를 별도로 저장하여 최종 모델 선정의 기준으로 삼았다. 또한, 과적합을 제어하고 불필요한 학습 시간을 줄이기 위해 조기 종료(Early Stopping) 기법을 적용하였다. 검증 세트의 F1-Score가 사전에 설정된 'patience' 에포크 동안 이전 최고 점수보다 향상되지 않으면 학습을 자동으로 중단하였다. 본 연구에서는 'patience' 값으로 10 에포크를 적용하였다. 이렇게 조기 종료 시점까지 저장된 모델 가중치 중 검증 세트에서 가장 우수한 F1-Score를 보인 모델을 최종 제안 모델로 선정하였으며, 이 모델을 사용하여 학습 및 검증 과정에 사용되지 않은 테스트 세트에서 최종 성능을 평가하였다.

Ⅴ. 결 과

본 연구에서 제안한 모델의 성능 평가는 AIHUB 데이터셋 중 CAM1 촬영 각도 데이터를 사용하여 평가하였다. CAM1 데이터는 기존 연구(〈Table 2〉 참조)에서 가장 낮은 성능을 보인 각도로, 본 연구는 이 환경에서의 성능 개선을 통해 제안 모델의 강건성을 검증하고자 하였다. 모든 실험은 〈Table 3〉에 명시된 컴퓨팅 환경과 최종 선정된 최적 하이퍼파라

〈Table 3〉 컴퓨터 환경 및 하이퍼 파라미터

컴퓨팅 환경	Nvidia RTX 3090	
Training 시간	9.08 h	
Hyperparameter	Search Space	Best Assignment
Number of Epochs	{10, 100}	100
Learning Rate	{1e-4, 2.5e-5, 1e-5}	2.5e-5
Training Batch Size	8	8
Weight Decay	{1e-4, 1e-5}	1e-4
Early Stopping patience(in Epoch)	10	10
Dropout	0.2	0.2
Window Size	{8, 16, 32}	16
Model Variation	{Small400, Small600, Base 400, Base600}	Small600

미터 구성을 기반으로 진행되었다. 성능 비교의 기준선(Baseline)으로는 AIHUB 데이터셋 CAM1에 대해 제공하는 스켈레톤 기반 모델의 공개된 성능 수치를 사용하였다.

제안 모델 및 비교 모델들의 주요 성능 지표는 〈Table 4〉에 요약되었다. 본 연구에서 제안한, 3절의 아키텍처 개선 및 학습 전략 최적화가 적용된 UniFormer 기반 모델은 CAM1 데이터셋 평가에서 정확도 96.5%, F1-Score 93.2%를 달성하였다. 이는 AIHUB 제공 스켈레톤 정보 기반 기준선 모델(정확도 90.1%, F1-Score 90.0%) 대비 정확도 6.4%p, F1-Score 3.2%p 향상된 수치이다. 제안 모델은 원본 RGB 영상을 직접 사용하는 종단간 방식이며 슬라이딩 윈도우 단위로 예측하는 반면, 해당 기준선 모델은 추출된 스켈레톤 정보를 활용하며 전체 비디오 클립 단위 예측 가능성이 있다는 방법론적 차이를 고려하더라도, 이 결과는 제안 모델이 F1-Score를 포함한 주요 평가지표 모두에서 기존 스켈레톤 정보 기반 접근 방식보다 우수한 성능을 나타냄을 보여준다. 특히 F1-Score의 향상은 미탐지와 오탐지 간의 균형을 개선하여 모델의 전반적인 판별 능력 및 신뢰성을 높였음을 의미한다.

나아가, 제안 모델과 동일하게 16프레임 비디오 클립을 처리 단위로 사용하는 다른 스켈레톤 정보 기반 모델들과 비교했을 때, 제안 모델의 RGB 영상 직접 활용 방식의 강점은 더욱 두드러진다. 〈Table 4〉에 따르면, AIHUB (2023)에서 제공하는 모델을 16프레임 클립 단위로 처리하도록 조정한 경우, 정확도 91.3%, F1-Score 85.7%를, Mediapipe + GBC 모델은 정확도 91.7%, F1-Score 87.2%를 기록했다. 본 연구의 제안 모델은 이들 스켈레톤 기반 모델들보다 F1-Score에서 각각 7.5%p, 6.0%p 높은 성능을 나타내, 동일 처리 단위 조건에서도 스켈레톤 정보 추출 단계를 생략하고 원본 영상에서 직접 특징을 학습하는 방식이 더 효과적일 수 있음을 시사한다.

또한, 동일한 환경에서 직접 구현하여 평가한 다른 RGB 영상 기반 모델과의 비교에서도 제안 모델의 우수성이 일관되게 확인되었다. 3D CNN은 F1-Score 74.1%를, R(2+1)D는 F1-Score 83.2%를 기록하여, 제안 모델이 RGB 영상 처리 방식 중에서도 뛰어난 판별 능력을 갖추었음을 보여준다.

제안 모델의 성능 향상 요인을 분석하기 위해, 본 연구에서 적용한 최적화 기법들의 효과를 확인하였다. 동일한 AIHUB 데이터셋의 CAM1 각도 및 16프레임 클립 실험 환경에서, 아키텍처 및 학습 전략 최적화가 적용되지 않은 표준 UniFormer는 F1-Score 88.0%, 정확도 94.6%를 기록하였다. 반면, 모든

〈Table 4〉 낙상감지모델 결과

ID	Type	Processing Unit (Input)	Accuracy	F1-Score
AIHUB(2023)	Skeleton	Full Video Clip	90.1	90.0
AIHUB(2023)	Skeleton	16-frame Video Clip	91.3	85.7
Mediapipe + GBC	Skeleton	16-frame Video Clip	91.7	87.2
3D-CNN	RGB	16-frame Video Clip	95.4	74.1
R(2+1)D	RGB	16-frame Video Clip	96.1	83.2
UniFormer	RGB	16-frame Video Clip	94.6	88.0
본 연구	RGB	16-frame Video Clip	96.5	93.2

개선 사항이 통합된 본 연구의 최종 제안 모델은 F1-Score 93.2%를 달성하여, 표준 UniFormer 대비 F1-Score가 5.2%p 향상되었음을 확인하였다. 이러한 결과는 적용된 아키텍처 개선(SwiGLU, Group Normalization, Pre-Layer Normalization 등)과 학습 전략 최적화(Lion 옵티마이저와 Lookahead 결합, Focal Loss, OneCycleLR 등)가 UniFormer의 낙상 감지 성능을 효과적으로 개선했음을 시사한다.

이러한 결과는 원본 RGB 영상만을 사용하는 종단간 방식으로 기존 스켈레톤 정보 기반 모델 및 다른 RGB 기반 CNN/RNN 조합 모델보다 향상된 성능을 달성했다는 점에서 핵심적인 의의를 지닌다. 특히, 제안 모델은 스켈레톤 정보 기반 방식의 계산 집약적 포즈 추정 전처리 단계를 생략함으로써 시스템 파이프라인을 단순화하고 효율성을 증대시킨다. 이는 실시간 시스템 구현 시 전처리 지연 시간을 제거하고 오류 전파 위험을 최소화하여, 제한된 연산 자원을 가진 환경에서도 강건한 성능 확보 및 적용 가능성을 높인다.

나아가, 제안 모델은 스켈레톤 정보만으로는 포착하기 어려운 질감, 색상, 미세 움직임, 배경 상호작용 등 다양한 시각적 컨텍스트 정보를 직접 학습에 활용하였다. 이는 스켈레톤 정보 추출이 불안정할 수 있는 복잡한 환경에서 낙상 감지 성능 유지 및 일반화 능력 향상에 기여할 수 있다. 결론적으로, 본 연구에서 제안된 최적화된 UniFormer 기반 접근법은 AIHUB 데이터셋의 CAM1 환경에서 높은 낙상 감지 정확도와 F1-Score을 달성하였으며, 이는 아키텍처 개선 및 학습 전략의 효과를 입증한다. 동시에, 전처리 없는 종단간 학습 방식이 향후 다양한 하드웨어 플랫폼에서의 실시간 낙상 감지 응용 시스템 개발에 유망한 방향임을 시사한다.

Ⅵ. 결 론

본 연구는 복합적인 시나리오를 포함하는 AIHUB 데이터셋의 CAM1 촬영 데이터를 활용하여, 원본 RGB 영상만을 입력으로 사용하는 최적화된 UniFormer 기반 종단간 낙상 감지 모델을 제안하고 그 성능을 평가하였다. 아키텍처 개선과 최신 학습 전략을 적용한 본 모델은 기존 스켈레톤 기반 방식의 전처리 의존성을 극복하고, 낙상 감지 정확도와 F1-Score 향상을 목표로 하였다. 실험 결과, 제안 모델은 AIHUB 데이터셋의 CAM1 촬영 데이터에서 정확도 96.5%, F1-Score 93.2%를 기록하였다(〈Table 4〉 참조). 이는 AIHUB 제공 스켈레톤 정보 기반 기준선 모델(정확도 90.1%, F1-Score 90.0%) 대비 각각 6.4%p 및 3.2%p 높은 수치이다. 이러한 F1-Score의 향상은 실제 낙상을 놓치지 않는 능력(재현율)과 예측의 정확성(정밀도) 간의 균형이 효과적으로 개선되었음을 의미한다. 이를 통해 응급상황인 낙상 발생 시 신속하게 대응할 수 있게 되어, 노인 안전 관리 시스템의 실효성을 높이는 데 기여할 수 있다.

본 연구의 이론적 함의는 다음과 같다. 첫째, 기존 낙상 감지 연구들이 주로 웨어러블 센서 기반 접근 또는 스켈레톤 추정 기반 모델에 의존한 반면, 본 연구는 원본 RGB 영상을 활용한 방식으로 이를 구조적으로 해결할 수 있음을 보여주었다. 이는 인공지능 기반 감지 시스템 설계에서 데이터 전처리 의존도를 줄이고, 모델 신뢰성과 실용성을 높이는 방향성을 제시한다는 점에서 의의가 있다. 둘째, 기존 낙상 감지 프레임워크가 제한된 환경(예: 고정된 실내 공간, 정적 배경, 일정한 촬영 각도 등)에서 수집된 데이터에 기반해 설계된 반면, 본 연구는 다양한 장소, 보조기구 사용 등 현실적 조건을 반영한 복합 환경 기

반 학습을 통해 모델을 훈련하고 검증하였다. 이러한 접근은 다양한 환경에서도 일관된 성능을 유지할 수 있는 인공지능 기반 낙상 감지 시스템의 일반화 가능성을 실증적으로 보여준다. 이는 기존 연구에서 강조되어 온 도메인 일반화(domain generalization) 가능성을 제시했다는 점에서 의의가 있다(Wang et al., 2021). 셋째, 본 연구는 인공지능 기술의 조직 내 수용과 운영 메커니즘에 대한 이론적 확장을 가능하게 한다. 기존 정보시스템 연구에서는 기술 수용 모형(TAM)이나 기술-조직-환경(TOE) 프레임워크를 통해 시스템 도입 요인을 주로 탐색해왔으나(Chatterjee et al., 2021; Nevi et al., 2025), 본 연구는 별도의 장비나 센서 없이 기존 CCTV 인프라만으로 작동 가능한 AI 기반 낙상 감지 시스템을 제안하여 기술 수용의 진입장벽을 낮춘 실증 결과를 제시하였다. 이는 사용자의 인지적 부담과 조직의 초기 도입비용을 동시에 경감시키는 설계 전략이 기술 수용성과 운영 효율성에 미치는 영향을 통합적으로 조망할 수 있는 이론적 근거를 제공할 수 있다는 점에서 의의가 있다. 넷째, 본 연구는 고령자 대상 디지털 헬스케어 기술의 서비스 운영 및 가치창출 메커니즘에 대한 이론적 논의를 확장한다. 기존 연구는 기술 기반 서비스의 성과 측정에 있어 주로 비용 절감이나 사용자 만족에 국한되어 있었으나(Ashfaq et al., 2020; Wamba-Taguimdje et al., 2020), 본 연구는 낙상 감지 정확도(F1-Score)와 같은 성능 지표를 통해 실제 안전성과 돌봄 품질의 개선 효과를 실증적으로 제시하였다. 이는 기술 중심의 운영 혁신이 고객 생명과 직결된 서비스 품질에 미치는 영향을 정량적으로 입증함으로써, 정보기술(IT)이 서비스 품질 향상에 기여하는 경로를 설명할 수 있는 이론적 틀을 마련하는 데 기여할 수 있다.

본 연구의 사회적 함의는 다음과 같다. 본 연구의 사회적 함의는 낙상 감지 기술을 단순한 기술적 성능 평가를 넘어, 실제 고령자 돌봄 환경에서의 활용 가능성과 사회복지 체계에 미치는 영향까지 포괄적으로 검토했다는 데에 있다. 첫째, 본 연구에서 제안한 원본 RGB 영상 기반 낙상 감지 모델은 고가의 웨어러블 센서나 별도 장비 없이 기존 CCTV 인프라만으로 작동 가능하다는 점에서, 고령자 가정이나 요양시설 등 다양한 생활공간에 쉽게 적용될 수 있는 높은 접근성을 갖는다. 이는 경제적 여건에 따른 돌봄 서비스의 격차를 완화하고, 정보기술 기반 복지 서비스의 보편적 확산 가능성을 실질적으로 제시한 사례라 할 수 있다. 특히, 고령자의 소득 수준, 지역 격차 등에 관계없이 일정 수준의 안전 모니터링 체계를 제공할 수 있다는 점은 사회복지의 형평성 측면에서 중요한 정책적 함의를 지닌다(Richardson et al., 2022). 둘째, 본 연구는 돌봄 노동의 구조적 부담에 대한 현실적인 대안을 제시한다. 현재 요양시설과 지역사회에서의 고령자 돌봄은 대부분 인력 기반의 모니터링에 의존하고 있으나, 인력 부족과 재정 압박은 지속적으로 문제로 지적되고 있다. 본 연구의 시스템은 비대면 · 비인력 기반으로 낙상을 감지할 수 있어, 24시간 실시간 감시를 요구하는 돌봄 인력의 부담을 경감시키고, 장기요양보험 재정의 지속가능성에도 긍정적 영향을 줄 수 있다. 이는 낙상 이후 신속한 대응과 이차적 건강 피해 방지를 통해 의료비용의 사회적 부담을 줄이는 데에도 기여할 수 있다(Hawley Hague et al., 2014). 셋째, 본 연구는 낙상 감지 기술이 고령자의 자율성과 독립성을 침해하지 않으면서도 안전을 확보할 수 있는 기술 설계 방향을 제시하였다. 장비 착용 없이 비가시적으로 작동하는 본 모델은 고령자의 일상에 개입하거나 불편을 초래하지 않으며, 감시 대상이 아닌 보호의 주체로서 고령자를 존중하는 구조로 설계되었다. 실제 고

령자 대상 수용성 연구에서도, 기술이 자율성과 프라이버시를 보장하는 방식으로 구현될 경우 수용 가능성이 높아진다는 점이 확인된 바 있다(Mujirishvili et al., 2023). 이는 기술준비도(TRI) 관점에서 신기술 도입 시 느끼는 '불편함(discomfort)'과 개인정보 노출에 대한 '불안감(insecurity)'과 같은 심리적 저해요인을 최소화하는 설계 전략으로, 사용자의 자연스러운 수용을 유도하는 데 핵심적인 역할을 한다(이한신&김판수, 2019). 본 연구는 이러한 조건을 충족하는 기술을 실증함으로써, 낙상 이후 신속한 대응뿐 아니라 고령자의 심리적 안정과 삶의 질 향상에 실질적으로 기여할 수 있는 현실적 대안을 제시하였다.

다양한 함의에도 불구하고 본 연구에는 한계점 역시 존재한다. 첫째, 모델 성능 평가는 AIHUB 데이터셋의 CAM1 촬영 각도에 한정되어 이루어졌기 때문에, 촬영 조건이나 시야 각도가 달라질 경우에도 동일한 성능을 유지할 수 있는지에 대해 단언하기 어렵다. 이는 실제 환경의 다양성과 복잡성을 충분히 반영하지 못한 채 특정 조건에 편중된 결과일 수 있다는 점에서 일반화 가능성이 떨어질 수 있다. 후속 연구를 통해 다양한 촬영 각도(CAM2, CAM3 등), 조명, 배경, 가림 조건 등이 포함된 실제 환경 기반 추가 데이터셋을 확보한다면, 단일 조건에 최적화된 성능이 아닌, 다양한 현실 조건에서도 안정적인 성능을 유지하는 낙상 감지 모델을 구축할 수 있을 것으로 판단된다. 둘째, 연구에서 사용한 데이터셋은 건강한 피험자가 모사한 낙상 데이터를 기반으로 하고 있어, 파킨슨병이나 중증 관절염처럼 비정형적 움직임을 보이는 고령자의 실제 낙상 양상을 충분히 포괄하지 못한다. 따라서 특정 질환군 또는 고유한 신체 조건을 지닌 고령자에 대한 모델의 적응성과 강건성은 아직 검증되지 않았다. 파킨슨병, 관절 질환 등 특정 질환군의 고령자 데이터를 직접 수집하거나, 이와 유사한 움직임을 반영한 데이터 증강 기법을 개발하여 모델 학습에 반영한다면, 이를 바탕으로 그룹별 성능 분석을 통해 실제 돌봄 환경에서의 적용 가능성과 신뢰성을 더욱 높일 수 있을 것으로 기대된다. 나아가, 향후 연구에서는 본 모델의 기술적 성능 검증을 넘어, 박영석 외(2021)의 연구에서 강조된 '신뢰성'이나 '쾌적성'과 같은 요인이 실제 사용자의 만족도 및 심리적 안정감에 미치는 다차원적 영향을 함께 검증할 필요가 있다. 셋째, 본 연구에서는 하이퍼파라미터 설정과 옵티마이저 활용에 있어 몇 가지 제약이 있었다. 하이퍼파라미터는 기존 문헌과 사전 실험을 바탕으로 수동 조정되었으며, grid search나 random search와 같은 체계적인 탐색 기법은 적용되지 않았다. 이에 따라 최적의 설정을 충분히 탐색하지 못했을 가능성이 있으며, 향후에는 자동화된 탐색 기법을 활용한 보다 정교한 튜닝이 요구된다. 또한 본 연구는 Lion 옵티마이저를 적용하였으나, 실험 환경의 한계로 인해 비교적 작은 배치 사이즈 하에서 수행되었다. 이러한 설정은 옵티마이저의 이론적 특성과 완전히 일치하지 않을 수 있으며, 이에 따라 성능 발휘에 일정한 제약이 있었을 가능성이 있다. 향후 연구에서는 다양한 배치 크기 및 학습률 조건을 아우르는 실험 설계를 통해 최적 조합을 검증할 필요가 있다.

참고문헌

박영석, 이종섭, 연지영, 최정일 (2021). "노인장기요양시설에서의 웰니스 IT서비스 특성과 이용의도와의 관계," **경영학연구**, 제50권 1호, pp.143-171.

(Park, Y S., Lee J. S., Yeon, J. Y. and Choi J. I. (2021). "The Relationship of Wellness IT Service Characteristics and Intention to Use in Long-Term Care Facilities for the Elderly," *Korean Management Review*, 50 (1), 143-171.)

이한신 김판수 (2019). "소비자의 기술수용과 저항이 인공지능(AI) 사용의도에 미치는 영향," **경영학연구**, 제48권 5호, pp.1195-1219.

(Yi, H. S. and Kim, P. S. (2019). "The Effect of Consumer's Technology Acceptance and Resistance on Intention to Use of Artificial Intelligence (AI)," *Korean Management Review*, 48(5), pp.1195-1219.)

정옥경, 이중원, 박철 (2024). "디지털 헬스케어 고객경험이 서비스 만족과 주관적 웰빙에 미치는 영향: 원격진료 서비스를 중심으로," **경영학연구**, 제53권 3호, pp.729-759.

(Jung, O. K., Lee, J. W. and Park, C. (2024). "Effects of Customer Experience of the Digital Healthcare on Service Satisfaction and Subjective Well-Being: Focusing on Telemedicine Services," *Korean Management Review*, 53(3), pp.729-759.)

한국지능정보사회진흥원. "낙상사고 위험동작 영상-센서 쌍 데이터," AI-Hub, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71641, 2025년 4월 6일 접속.

(National Information Society Agency. "Video-Sensor Paired Data for Fall Risk Behaviors," AI-Hub, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71641, retrieved April 6, 2025.)

한국지능정보사회진흥원. "AI 허브," AI-Hub, https://www.aihub.or.kr/, 2025년 4월 2일 접속.

(National Information Society Agency. "AI Hub," AI-Hub, https://www.aihub.or.kr/, retrieved April 2, 2025.)

Aderinola, T. B., Palmerini, L., D'Ascanio, I., Chiari, L., Klenk, J., Becker, C., Caulfield, B., & Ifrim, G. (2025). "Accurate and efficient real- world fall detection using time series techniques. In G. Ifrim, T. B. Aderinola, & B. Caulfield (Eds.)," *Advanced Analytics and Learning on Temporal Data* (pp.52-79). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-53317-2_4

Alanazi, T., Babutain, K., and Muhammad, G. (2024). "Mitigating Human Fall Injuries: A Novel System Utilizing 3D 4-Stream Convolutional Neural Networks and Image Fusion," *Image and Vision Computing*, 148, 105153.

Alam, E., Sufian, A., Dutta, P., and Leo, M. (2022). "Vision-based human fall detection systems using deep learning: A review," *Computers in Biology and Medicine*, 146, 105626.

Al-qaness, M. A., Dahou, A., Abd Elaziz, M., and Helmi, A. M. (2024). "Human activity recognition and fall detection using convolutional neural network and transformer-based architecture," *Biomedical Signal Processing and Control*, 95, 106412.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF international conference on computer vision, pp.6836-6846.

Ashfaq, M., Yun, J., Yu, S., and Loureiro, S. M. C. (2020). "I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents," *Telematics and Informatics*, 54, 101473.

Assanovich, B., & Kosarava, K. (2025). "Vision-Based Fall Detector for Elderly Based on Sliding Window Approach and Feature Engineering," *Journal of Data Science and Intelligent Systems*, 3(1), pp.27-34.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization*. arXiv. https://arxiv.org/abs/1607.06450.

Bertasius, G., Wang, H., and Torresani, L. (2021). "Is space-time attention all you need for video understanding?," in proceeding of the International Conference on Machine Learning (ICML), Virtual.

Durga Bhavani, K. and Ferni Ukrit, M. (2024). "Design of inception with deep convolutional neural network based fall detection and classification model," *Multimedia Tools and Applications*, 83(8), pp.23799-23817.

Blackburn, J., Ousey, K., Stephenson, J., and Lui, S. (2022). "Exploring the impact of experiencing a long lie fall on physical and clinical outcomes in older people requiring an ambulance: A systematic review," *International Emergency Nursing*, 62, 101148.

Bui, T., Liu, J., Cao, J., Wei, G., and Zeng, Q. (2024). "Elderly fall detection in complex environment based on improved YOLOv5s and LSTM," *Applied Sciences*, 14(19), 9028.

Cao, Y., Guo, M., Sun, J., Chen, X., and Qiu, J. (2024). "Fall detection based on LCNN and fusion model of weights using human skeleton and optical flow," *Signal, Image and Video Processing*, 18(1), pp.833-841.

Carreira, J. and Zisserman, A. (2017). "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6299-6308, Honolulu, Hawaii.

Chawan, V. R., Huber, M., Burns, N., and Daniel, K. (2022). "Person identification and tinetti score prediction using balance parameters: A machine learning approach to determine fall risk," in proceeding of the 15th international conference on PErvasive technologies related to assistive environments, pp.203-212.

Chatterjee, S., Rana, N. P., Dwivedi, Y. K., and Baabdullah, A. M. (2021). "Understanding AI adoption in manufacturing and production firms using an integrated TAM-TOE model," *Technological Forecasting and Social Change*, 170, 120880.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Pham, H., Dong, X., Luong, T., Hsieh, C. J., Lu, Y., and Le, Q. V. (2023). "Symbolic Discovery of Optimization Algorithms," *Advances in Neural Information Processing Systems*, 36, pp.49205-49233.

Chu, X., Tian, Z., Zhang, B., Wang, X., & Shen, C. (2021). *Conditional positional encodings for vision transformers*. arXiv. https://arxiv.org/abs/2102.10882

Chutimawattanakul, P. and Samanpiboon, P. (2022). "Fall detection for the elderly using yolov4 and lstm," in proceeding of the 2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp.1-5.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An*

image is worth 16x16 words: Transformers for image recognition at scale. arXiv. https://arxiv.org/abs/2010.11929

El-Bendary, N., Tan, Q., Pivot, F. C., & Lam, A. (2013). "Fall Detection and Prevention for the Elderly: A Review of Trends and Challenges," *International Journal on Smart Sensing and Intelligent Systems*, 6(3), pp. 1230.

Ergüder, H., Uzun, T., & Baday, M. (2024). "Advancing Fall Detection Utilizing Skeletal Joint Image Representation and Deformable Layers," *Image Analysis and Stereology*, 43(1), pp. 97-107.

Espinosa, R., Ponce, H., Gutiérrez, S., Martínez-Villaseñor, L., Brieva, J., & Moya-Albor, E. (2019). "A Vision-Based Approach for Fall Detection Using Multiple Cameras and Convolutional Neural Networks: A Case Study Using the UP-Fall Detection Dataset," *Computers in Biology and Medicine*, 115, pp. 103520.

Feichtenhofer, C. (2020). "X3d: Expanding architectures for efficient video recognition," in proceeding of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA.

Florence, C. S., Bergen, G., Atherly, A., Burns, E., Stevens, J., and Drake, C. (2018). "Medical costs of fatal and nonfatal falls in older adults," *Journal of the American Geriatrics Society*, 66(4), pp.693-698.

Gao, M., Li, J., Zhou, D., Zhi, Y., Zhang, M., and Li, B. (2023). "Fall detection based on OpenPose and MobileNetV2 network," *IET Image Processing*, 17(3), pp.722-732.

Gutiérrez, J., Rodríguez, V., and Martin, S. (2021). "Comprehensive review of vision-based fall detection systems," *Sensors*, 21(3), pp.947.

Hawley-Hague, H., Boulton, E., Hall, A., Pfeiffer, K., and Todd, C. (2014). "Human factors and adherence to home-based exercise programs in older adults at risk of falling: A systematic review," *Physical Therapy*, 94(3), pp.319-336.

Hoang, V. H., Lee, J. W., Piran, M. J., and Park, C. S. (2023). "Advances in skeleton-based fall detection in RGB videos: From handcrafted to deep learning approaches," *IEEE Access*, 11, pp.92322-92352.

Ioffe, S. and Szegedy, C. (2015). "Batch normalization : Accelerating deep network training by reducing internal covariate shift," in proceeding of the International conference on machine learning, Lille, France.

Inturi, A. R., Manikandan, V. M., and Garrapally, V. (2023). "A novel vision-based fall detection scheme using keypoints of human skeleton with long short-term memory network," *Arabian Journal for Science and Engineering*, 48(2), pp.1143-1155.

Islam, M. A., Jia, S., & Bruce, N. D. (2020). *How much position information do convolutional neural networks encode?*. arXiv. https://arxiv. org/abs/2001.08248.

Kaur, N., Rani, S., and Kaur, S. (2024). "Real-time video surveillance based human fall detection system using hybrid haar cascade classifier," *Multimedia Tools and Applications*, 83(28), pp.71599-71617.

Keskes, O. and Noumeir, R. (2021). "Vision-based fall detection using st-gcn," *IEEE Access*, 9, pp.28224-28236.

Kim, S., Kim, S., Woo, S., Oh, J., Son, Y., Jacob,

L., Soysal, P., Park, J., Chen, L-K., and Yon, D. K. (2025). "Temporal trends and patterns in mortality from falls across 59 high-income and upper-middle-income countries, 1990 - 2021, with projections up to 2040: a global time-series analysis and modelling study," *The Lancet Healthy Longevity*, 6(1), p.100672.

Kingma, D. P. and Ba, J. L. (2015). "Adam: A method for stochastic optimization," in Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA.

Knowles, B. and Hanson, V. L. (2018). "The wisdom of older technology (non) users," *Communications of the ACM*, 61(3), pp.72-77.

Kwolek, B. and Kepski, M. (2014). "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, 117 (3), pp.489-501.

Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., & Qiao, Y. (2022). *Uniformer: unified transformer for efficient spatiotemporal representation learning*. arXiv. https://arxiv.org/abs/2201.04676

Li, X., Wang, Y., Zhou, Z., and Qiao, Y. (2020). "Smallbignet: Integrating core and contextual views for video classification," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, Venice, Italy.

Liu, W., Liu, X., Hu, Y., Shi, J., Chen, X., Zhao, J., Wang, S., and Hu, Q. (2022). "Fall detection for shipboard seafarers based on optimized BlazePose and LSTM," *Sensors*, 22(14), p. 5449.

Loshchilov, I., & Hutter, F. (2017). *Fixing weight decay regularization in Adam*. arXiv. https://arxiv.org/abs/1711.05101

Luo, B. (2023). "Human fall detection for smart home caring using Yolo networks," *International Journal of Advanced Computer Science and Applications*, 14(4), pp.53-58.

Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., and Peñafort-Asturiano, C. (2019). "UP-fall detection dataset: A multimodal approach," *Sensors*, 19(9), p.1988.

McCall, S., Kolawole, S. S., Naz, A., Gong, L., Ahmed, S. W., Prasad, P. S., and Ardakani, S. P. (2024). "Computer Vision Based Transfer Learning-Aided Transformer Model for Fall Detection and Prediction," *IEEE Access*, 12, pp.28798-28809.

Mobsite, S., Alaoui, N., Boulmalf, M., and Ghogho, M. (2023). "Semantic segmentation-based system for fall detection and post-fall posture classification," *Engineering Applications of Artificial Intelligence*, 117, 105616.

Mudiyanselage, S. P. K., Yao, C. T., Maithreepala, S. D., and Lee, B. O. (2024). "Emerging Digital Technologies Used for Fall Detection in Older Adults in Aged Care: A Scoping Review," *Journal of the American Medical Directors Association*, 26(1), 105330.

Mujirishvili, T., Maidhof, C., Florez-Revuelta, F., Ziefle, M., Richart-Martinez, M., and Cabrero-García, J. (2023). "Acceptance and Privacy Perceptions Toward Video-based Active and Assisted Living Technologies: Scoping Review,"

Journal of Medical Internet Research, 25, e45297.

Nevi, G., Pizzichini, L., Bastone, A., and Dezi, L. (2025). "Adoption of AI by micro and small health enterprises: effects of entrepreneurial orientation on the TOE model," *European Journal of Innovation Management*, forthcoming.

Núñez-Marcos, A. and Arganda-Carreras, I. (2024). "Transformer-based fall detection in videos," *Engineering Applications of Artificial Intelligence*, 132, 107937.

Ramirez, H., Velastin, S. A., Meza, I., Fabregas, E., Makris, D., and Farias, G. (2021). "Fall detection and activity recognition using human skeleton features," *IEEE Access*, 9, pp.33532-33542.

Ramirez, H., Velastin, S. A., Cuellar, S., Fabregas, E., and Farias, G. (2023). "BERT for activity recognition using sequences of skeleton features and data augmentation with GAN," *Sensors*, 23(3), 1400.

Ren, L. and Peng, Y. (2019). "Research of fall detection and fall prevention technologies: A systematic review," *IEEE Access*, 7, pp. 77702-77722.

Richardson, S., Lawrence, K., Schoenthaler, A., and Mann, D. (2022). "A framework for digital health equity," *NPJ Digital Medicine*, 5(1), 119.

Salimi, M., Machado, J. J., and Tavares, J. M. R. (2022). "Using deep neural networks for human fall detection based on pose estimation," *Sensors*, 22(12), 4544.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, Utah, USA.

Shazeer, N. (2020). *Glu variants improve transformer*. arXiv. https://arxiv.org/abs/2002.05202

Smith, L. N. (2018). *A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay*. arXiv. https://arxiv.org/abs/1803.09820

Su, C., Wei, J., Lin, D., Kong, L., & Guan, Y. L. (2024). A novel model for fall detection and action recognition combined lightweight 3D-CNN and convolutional LSTM networks. *Pattern Analysis and Applications*, 27(1), Article 3. https://doi.org/10.1007/s10044-023-01181-z

Suarez, J. J. P., Orillaza, N., and Naval, P. (2022). "AFAR: A real-time vision-based activity monitoring and fall detection framework using 1D convolutional neural networks," in Proceedings of the 2022 14th International Conference on Machine Learning and Computing, pp.555-559. https://doi.org/10.1145/3529836.3529916

Sykes, E. R. (2025). "Next-generation fall detection: harnessing human pose estimation and transformer technology," *Health Systems*, 14(2), pp.85-103.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3D convolutional networks," in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489-4497, Santiago, Chile.

Tran, D., Wang, H., Torresani, L., and Feiszli, M. (2019). "Video classification with channel-separated convolutional networks," in Proceedings of the IEEE/CVF International

Conference on Computer Vision (ICCV), Seoul, South Korea, IEEE, pp.5552-5561.

Ursul, I. (2024). "Elderly Fall Detection Using Unsupervised Transformer Model," *Electronics and information technologies/Електроніка та інформаційні технології*, 26.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is all you need," in Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, Curran Associates, Inc., pp.5998-6008.

Wamba-Taguimdje, S. L., Wamba, S. F., Kamdjoug, J. R. K., and Wanko, C. E. T. (2020). "Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects," *Business Process Management Journal*, 26(7), pp.1893-1924.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Hou, W., Chen, Y., and Yu, P. S. (2023). "Generalizing to Unseen Domains: A Survey on Domain Generalization," *IEEE Transactions on Knowledge and Data Engineering*, 35(8), pp.8052-8072.

Wang, H., Xu, S., Chen, Y., and Su, C. (2025). "LFD-YOLO: a lightweight fall detection network with enhanced feature extraction and fusion," *Scientific Reports*, 15(1), pp. 5069.

Wang, X., Ellul, J., and Azzopardi, G. (2020). "Elderly fall detection systems: A literature survey," *Frontiers in Robotics and AI*, 7, pp.71.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, pp.7794-7803.

World Health Organization, "Falls," World Health Organization, https://www.who.int/news-room/fact-sheets/detail/falls, retrieved July 2025.

Wu, Y. and He, K. (2018). "Group normalization," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, pp.3-19.

Xu, D., Wang, Y., Zhu, S., Zhao, M., and Wang, K. (2024). "Relationship between fear of falling and quality of life in nursing home residents: The role of activity restriction," *Geriatric Nursing*, 57, pp.45-50.

Xun, J., Wang, X., Wang, X., Fan, X., Yang, P., and Zhang, Z. (2025). "An efficient algorithm for pedestrian fall detection in various image degradation scenarios based on YOLOv8n," *Scientific Reports*, 15, pp.9036.

Yadav, S. K., Luthra, A., Tiwari, K., Pandey, H. M., and Akbar, S. A. (2022). "ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling," *Knowledge-Based Systems*, 239, pp.107948.

Yu, M., Gong, L., and Kollias, S. (2017). "Computer vision based fall detection by a convolutional neural network," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Glasgow, UK, Association for Computing Machinery, pp.416-420.

Yu, X., Wang, C., Wu, W., and Xiong, S. (2025). "A Real-time Skeleton-based Fall Detection Algorithm based on Temporal Convolutional Networks and Transformer Encoder," *Pervasive and Mobile Computing*, 102016.

Zahan, S., Hassan, G. M., and Mian, A. (2022). "Sdfa: Structure-aware discriminative feature

aggregation for efficient human fall detection in video," *IEEE Transactions on Industrial Informatics*, 19(8), pp.8713-8721.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. (2019). "Lookahead optimizer: k steps forward, 1 step back," in Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada, Curran Associates, Inc., pp.9188-9198.

• 저자 김준석은 현재 경북대학교 경영대학 경영정보 전공 석사과정에 재학 중이다. 경북대학교 IT대학 컴퓨터학부를 졸업하였다. 주요 연구분야는 인공지능, 정보보안 등이다.

• 저자 이새롬은 2010년 부산대학교에서 학사 학위를 받았으며, 2016년 서울대학교에서 경영정보시스템 박사학위를 받았다. 2018년부터 경북대학교 경영대학에서 부교수로 재직하고 있으며 주요 연구 관심사는 개방형 협업 및 온라인 성희롱이다.

• 저자 박종화는 현재 경북대학교 경영학부에서 조교수로 재직 중이다. 울산과학기술원 테크노경영학부 학사 및 경영과학부 박사를 취득하였다. 박사 학위 취득 이후에는 공주대학교 상업정보교육과에 재직하였다. 주요연구분야는 인공지능에 대한 사회적 영향, 플랫폼 비즈니스 등이다.