

# 실시간 쓰러짐 감지를 위한 하이브리드 시공간 융합 모델 연구

## A Hybrid Spatiotemporal Fusion Model for Real-time Fall Detection

김은영<sup>1</sup>, 이재호<sup>1,\*</sup>  
 (EunYoung Kim<sup>1</sup> and Jaeho Lee<sup>1,\*</sup>)

<sup>1</sup>Department of Software, Duksung Women's University

**Abstract:** We propose a hybrid spatiotemporal fusion model for real-time fall detection using monocular camera videos. Per-frame spatial features are extracted by a lightweight CNN with squeeze-and-excitation modules, an AutoEncoder reduces noise to generate core representations, and a two-layer GRU together with a Transformer encoder jointly learns temporal flow and global relations. Multi-task learning that combines classification error (cross-entropy) and reconstruction error (MSE) enables discrimination of subtle pre-fall motions. Evaluated on sequences (30 frames by default) compiled from multiple public datasets with a subject-wise split, the proposed model achieves 97.10% accuracy (67/69), 100% recall for falls (FN=0), and ROC-AUC of 1.00. With 0.91 GFLOPs, 3.22 M parameters, and an average inference delay of 8.16 ms per sample, its usability in edge environments is verified.

**Keywords:** fall detection, real-time, spatiotemporal fusion, CNN, GRU, Transformer, autoencoder, edge computing

### I. 서론

우리 사회는 1인 가구의 지속적 증가와 고령화로 인해, 돌봄의 사각지대를 채우는 ‘즉각성 높은 안전 모니터링’이 필수 과제가 되었다[1,2,3]. 쓰러짐은 2차 손상과 장시간 방치로 인한 위험을 동반하며, 무엇보다 대응은 골든타임을 놓치지 않는 것이 핵심이다[4,5]. 이에 따라, 본 연구는 카메라 영상만을 활용하여 실시간으로 쓰러짐을 ‘정확하고 빠르게 알리는’ 모델을 목표로 한다. 하지만, 실제 적용을 위해서는 두 가지 조건을 동시에 충족되어야 한다. 첫째, 안전을 위해 미탐(false negative) 최소화가 우선되어야 한다. 둘째, 개인정보 보호와 지연·가용성 측면에서 엣지 온디바이스 추론이 가능해야 하므로 저연산·저메모리가 필수적이다. 따라서, 모델은 높은 민감도(특히 미탐 최소화)와 함께 낮은 연산량과 메모리, 그리고 환경 변화에 대한 견고성을 동시에 고려해야 한다.

기술적으로 가장 큰 난점은 유사 동작과의 구분이다. 단일 프레임 기반 CNN은 구조가 간결하고 빠르다는 장점이 있지만, 동작의 연속성이 반영되지 않아 앓기·숙이기·물건 줍기·단순 놀기와 같은 패턴에서 거짓 경보가 발생하기 쉽다. 반면 3D-CNN이나 광류(Optical Flow)를 포함하는 구조는 시간 정보를 활용하지만, 엣지 환경에서 실시간을 보장하기 어려울 만큼 연산비용이 크거나 입력 품질 변화에 민감하다. 결국, 시간 정보의 핵심만을 효율적으로 통합하면서도 연산은 가볍게 유지하는 설계가 요구된다.

이처럼 정확도·정확도·지연·안정성 사이의 균형을 만족하는 현실적인 모델 설계가 필요하다. 이에 본 논문은 프레임 수준의 공간 단서를 경량 CNN (Convolutional Neural Network)과 SE (Squeeze-and-Excitation)로 추출하고, AutoEncoder로 노이즈를 줄여 정보를 압축한 뒤, 2층 GRU와 Transformer 인코더를 통해 근거리·원거리 시간 관계를 함께 포착하는 하이브리드 시공간 모델(CNN-AE-GRU-Transformer)을 제안한다.

여기서 쓰러짐 미탐(false negative)을 최소화하는 것을 최우선적인 원칙으로 삼고, 엣지(Edge) 환경에서 실시간 동작이 가능한 경량 비전 모델을 제시한다. 동시에, 운용성을 저해하는 오탐(false positive)을 점진적으로 낮추기 위해 데이터 확장 및 표현과 시계열 학습 강화를 종합적으로 적용하였다. 실제 테스트 환경에서의 분류 지표, ROC-AUC, 혼동행렬, 추론 지연을 통해 안정성, 정확성, 실시간성의 균형을 검증한다.

추가적으로, 모델의 적용 대상은 1차적으로 독거노인이지만, 적용 범위는 여기에 한정되지 않는다. 예를 들어 뇌전증 후유 증상, 파킨슨병, 기립성 저혈압 등 쓰러짐 발병률이 높아 위험한 환자군, 또는 재활·치료 과정에서 낙상과 같은 고위험 단계에 놓인 대상까지 다양하게 확장될 수 있다[6].

### II. 연구 배경

카메라 기반 쓰러짐 감지는 크게 프레임 단위 CNN 분류와 시계열 모델링 결합형으로 나뉜다. 단일 프레임 CNN 분류는 전처리가 간단하고 추론 속도가 빠르지만, 동작의 연속성을 반영하지 못해 앓기/숙이기/물건 줍기와 같은 바닥에서 수직적인 축에서 발생하는 자세에서 오탐이 빈번하다[7]. 이를 보완하기 위해 새로운 2-stream 네트워크인 RGB-광류(Optical Flow) 2-stream CNN 구조가 제안되었으나, 광류 계산의 불안정성과 연산량 증가로 엣지 환경에서 실시간성이 저하되는 문제가 보고되었다[8].

시간적 맥락을 처리하기 위해 CNN 특징 뒤에 LSTM/GRU를 연결하는 하이브리드 모델이 널리 연구되었다. CNN이 프

\* Corresponding Author

Manuscript received September 16, 2025; revised October 4, 2025; accepted December 25, 2025

김은영: 덕성여자대학교 소프트웨어전공 학부생(mozzi0324@duksung.ac.kr, ORCID<sup>®</sup> 0009-0001-9014-1397)

이재호: 덕성여자대학교 소프트웨어전공 교수(izeho@duksung.ac.kr, ORCID<sup>®</sup> 0000-0003-0455-9939)

레이벌 특징을 추출한 뒤 LSTM/GRU 순환모델이 시계열로 학습하는 방식은 쓰러짐의 전후의 변화를 포착하는데 효과적이지만, 긴 시퀀스에서는 여전히 장기 의존성의 소실과 윈도우 길이·패딩 전략에 민감하다는 점이 지적된다[9]. 또한, 3D-CNN은 공간-시간 커널로 동작 패턴을 직접 학습해 정확도를 높였지만, 모델 크기와 FLOPs가 커서 저 사양 엡지 디바이스 적용이 어렵고, 소규모 쓰러짐 데이터셋에서는 과적합 문제가 발생한다[10].

Skeleton/포즈 기반 + 그래프 합성곱(GCN) 접근은 개인정보 보호와 조명 변화에 강점이 있으나, 키포인트 누락·가림 시에 급격한 성능 저하가 나타나며, 카메라 각도나 촬영 위치가 바뀌면 도메인 이동 문제가 발생한다[11]. AutoEncoder나 원클래스 기반 이상탐지 방식은 정상 데이터만으로 학습해 라벨링 비용을 줄이고 미탐(FN)을 억제할 가능성이 있으나, 임계값 설정과 환경 변화에 따른 재튜닝 의존성이 크다[12,13].

최근에는 비전 기반에 오디오 정보를 융합한 멀티모달 쓰러짐 감지 연구도 제안되었다. YOLO v7-Pose로 추출한 골격 특징과 MobileNetV2 기반 음성 특징을 융합하여, 단일 영상 모달 대비 민감도를 향상시켰다. 그러나 오디오 센서 의존성으로 인해 마이크 위치 변화나 실내 환경 잡음에 따라 성능이 민감하여, 이는 카메라 단일 모달 구조 대비 배포 환경의 제약을 초래한다[14].

Transformer 계열 모델(TimeSformer, ViT 등)은 전역 어텐션을 통해 장기 문맥 포착에 강점을 보이지만, 대규모 데이터 요구와 연산비용으로 엡지 환경에서 실시간 적용에 부담이 높아지며, 소규모 쓰러짐 데이터에서는 사전학습 전이 없이는 성능적인 이점이 제한적이다. 이를 보완하고자, IMU 기반 경량 Transformer 구조를 제안했으나, 카메라 기반 시각 정보 미포함으로 인해 시각적 맥락 감지의 한계가 존재한다[15]. 반대로 경량 CNN 백본(MobileNet, ShuffleNet 등)을 활용한 실시간 연구는 추론 지연을 대폭 줄였지만, 시간적 표현력 부족으로 유사 동작에서 FN/FP 개선 폭이 제한적이다[16].

웨어러블 기반 LSTM/GRU 접근은 조명·가림에 영향을 받지 않는 장점이 있으나, 착용 순응도, 배터리 문제와 사용자별 보정 필요성이 과제로 남는다. 실제 가정 환경에서는 카메라·웨어러블 혼합 센서 융합이 제안되지만, 시스템 복잡성과 유지관리의 부담이 커진다. 또한, 실험 참가자와 실제 노인 사용자 간 조건 차이로 인해 모델 성능이 크게 저하될 수 있음을 보고했다[17]. 또한, 손목형 웨어러블 낙상 감지 시스템이 보고되어 노년층 현장 적용 가능성을 입증한 바 있다[18]. 하지만, 우리는 시각 기반 접근의 장점(비착용·공간 인식)을 유지하면서도 실제 배치 제약을 고려한 경량 모델과 추론 지연을 함께 평가하는 방향성을 가진다.

기존 연구들을 종합하면, 쓰러짐 감지에서 정확도와 실시간성을 동시에 만족하는 현실적인 해결책을 찾는 것에 한계가 있음을 알 수 있다. 따라서, 다음과 같은 요건들이 요구된다. 첫째, 시간적 맥락을 효율적으로 활용하면서도 연산량을 최소화하는 경량 아키텍처가 요구된다. 저연산, 저메모리, 저지연을 전제로 한 구조가 필요한 것이다. 둘째, 근거리와 원거리 시간 의존성을 동시에 포착할 수 있는 다층적 시계열 모델링이 필요하다. GRU의 순차적 패턴 학습과 Transformer

의 전역 어텐션을 결합하여 쓰러짐의 급격한 변화와 행동 전반의 흐름(전후 맥락)을 모두 감지해야 한다. 셋째, 노이즈가 많은 실제 환경에서도 안정적인 특징 추출을 위한 표현 학습 최적화가 필요하다. 넷째, 안전을 위해 재현율(미탐 최소화) 우선의 임계값 설계와 튜닝이 필수적이며, 데이터 측면에서는 다양한 일상 행동을 포함한 증강 전략으로 일반화를 높여야 한다. 또한, 실제 배포 환경을 고려하면 개인정보 보호와 유지보수 편의성을 위해 카메라 단일 모달을 유지해야 하며, 착용형 센서나 추가 마이크에 독립적인 구성이 바람직하다.

이와 같은 필요성과 방향성에 따라, 본 연구에서는 CNN-AE-GRU-Transformer 하이브리드 구조를 통해 공간-시간 특징을 효율적으로 학습하고, 미탐 최소화를 우선으로 하는 엡지 최적화 쓰러짐 감지 모델을 검증한다.

### III. 모델 설계 및 실험 방법론

#### 1. 데이터셋 구성 및 전처리

쓰러짐 감지를 위해, 단순히 쓰러진 한 순간의 모습을 활용해 분류하게 된다면, 해당 행동의 연속성과 시간적 흐름을 포착하기 힘들다는 한계가 있다. 쓰러짐은 시간에 따라 점진적으로 발생하는 행동 유형이고, 이를 가능한 한 빠르게 감지하는 것이 핵심이다. 따라서, 쓰러지는 과정의 전체 흐름을 담은 영상을 수집해 시계열 형태로 변환하였다. 각 원본 영상은 30프레임으로 일정하게 분할하여 하나의 시퀀스 단위를 구성하였다. 또한, 일부 영상 가운데 길이가 부족한 영상은 마지막 프레임의 복제로 채워 시퀀스의 일관성을 유지하였다. 이를 통해, 시간의 지남에 따라 변화하는 자세 및 행동 변화의 속도와 같은 동적 패턴을 학습 가능하게 한다.

이후, 데이터의 다양성 확보와 모델의 일반화 성능을 향상을 위해, 추출된 각 프레임에 다양한 전처리를 적용하였다. 다양한 공간적 특징, 조명 변화에 대응하기 위함이다. 구체적으로 RandomResizedCrop, RandomHorizontalFlip, Color Jitter, GaussianBlur, Gamma 보정, RandomErasing 등을 사용하여 입력의 다양성을 확장하였다. 모든 프레임은 224 x 224 크기로 리사이즈 후 ImageNet 기준 평균 및 표준편차로 정규화하였다. 이러한 과정은 조명, 카메라 각도, 대상의 위치 변화 등에 촉불가한 변동에서도 학습할 수 있는 기반을 제공한다.

추가적으로, 데이터셋은 다양한 연령대와 체형을 포함하도록 구축하였으며, 데이터셋 내 프레임은 품질 검사를 거쳐 흐릿하거나 노이즈가 심한 이미지를 제외하였다. 이는 학습 과정에서 특정 집단에 편향되지 않도록 방지하였으며, 모델이 핵심적인 움직임 패턴에 집중하도록 돕는다.

#### 2. Baseline 및 반복적 모델 개발

##### 2.1 초기 베이스라인 구성

본 연구의 시작점은 쓰러짐 행동의 시공간적 특성을 포착할 수 있는 기본적인 시퀀스 분류 아키텍처를 구축하는 것이었다. 초기 베이스라인은 공간 정보를 추출하는 CNN 백본과 시간적 의존성을 학습하는 순환 구조(RNN 계열), 그리고 시퀀스 전역 관계를 반영하는 Transformer Encoder를 결합한 하이브리드 형태로 설계하였다. 즉, 프레임 단위로 입력된 영상은 ResNet 계열의 CNN [25]을 통해 256차원 특징 벡터로 임베딩되며, 이 시퀀스는 LSTM으로 전달되어 일시적인 시간

표 1. 버전별 모델 구조 및 주요 변경점 요약.

Table 1. Overview of model versions and key architectural changes.

| Exp      | Key Architectural Changes                                                                                                                                                                                                                            | Key Details                                                                                                                                                                                     | Accuracy | Recall (Falling) | FLOPs /Parameters | Inference Latency | Issues                       | Improvements                                              |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|------------------|-------------------|-------------------|------------------------------|-----------------------------------------------------------|
| Baseline | 1                                                                                                                                                                                                                                                    | <ul style="list-style-type: none"> <li>• Combined CrossEntropyLoss + <math>\alpha \times \text{MSELoss}</math></li> <li>• Introduced ReduceLROnPlateau</li> </ul>                               | 97.10%   | 100%             | 4.14 G / 30.07M   | 8.71 ms           | Model is too heavy           | -                                                         |
| Exp. 1-2 | <ul style="list-style-type: none"> <li>• Custom lightweight CNN: 4 Conv layers, Global Avg Pool, feature_dim = 256</li> <li>• AutoEncoder: 256 <math>\rightarrow</math> ...</li> <li>• Two-layer LSTM + one layer Transformer (unchanged)</li> </ul> | <ul style="list-style-type: none"> <li>• First adoption of lightweight CNN</li> <li>• Same <math>\alpha</math> for classification &amp; reconstruction</li> </ul>                               | 88.41%   | 89%              | 0.90 G / 5.15 M   | 4.89 ms           | Poor Performance             | Model size $\downarrow$ to $\frac{1}{4}$ ; lightweighting |
| Exp. 3-5 | <ul style="list-style-type: none"> <li>• Custom lightweight CNN</li> <li>• Added Residual / Bottleneck / SE Blocks</li> <li>• Final feature_dim = 256</li> <li>• Replaced LSTM with two-layer GRU</li> </ul>                                         | <ul style="list-style-type: none"> <li>• Enhanced feature representation</li> <li>• Improved computational efficiency</li> <li>• <math>\alpha = 0.05</math> (classification focused)</li> </ul> | 98.55%   | 100%             | 0.91 G / 3.22 M   | 6.18 ms           | -                            | Smaller model & better accuracy                           |
| Exp. 6   | <ul style="list-style-type: none"> <li>• Same as Exp.5</li> <li>• Re-inject final softmax probabilities into AC latent (feedback_fc)</li> </ul>                                                                                                      | <ul style="list-style-type: none"> <li>• Attempted reconstruction improvement</li> <li>• Introduced Early Stopping</li> </ul>                                                                   | 95.65%   | 97%              | 0.91 G / 3.22 M   | 6.70 ms           | Lower performance than Exp.5 | Underperformed vs. Exp.5                                  |
| Exp. 7~  | <ul style="list-style-type: none"> <li>• Same as Exp.5</li> </ul>                                                                                                                                                                                    | <ul style="list-style-type: none"> <li>• Dataset expanded with additional non-fall &amp; daily-activity videos;</li> </ul>                                                                      | 97.10%   | 100%             | 0.91 G / 3.22 M   | 8.16 ms           | -                            | Robustness $\uparrow$ (t-SNE separation, ROC-AUC = 1.00)  |

흐름을 요약한 뒤, Transformer 인코더를 통해 전체 시퀀스 내 상호작용을 보완적으로 융합하여 최종 분류기로 연결하였다. 이 구조에는 단순 분류 손실(CrossEntropyLoss)에 더해 AutoEncoder 기반 재구성 손실(MSE Loss)을 병합한 멀티태스크 학습을 적용하여 특징 표현력을 높였고, 손실 조합은 다음과 같이 정의하였다:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{MSE}}, \quad \alpha = 0.05 \quad (1)$$

여기서  $\alpha$ 는 재구성 손실의 가중치로, 검증셋 기반 그리드 서치로 선정하였다.  $\alpha$  후보 {0, 0.01, 0.03, 0.05, 0.10, 0.20} 를 비교한 결과,  $\alpha = 0.05$ 가 F1과 ROC-AUC가 가장 높고, 쓰러짐 Recall = 100%(FN = 0)를 유지하였다.  $\alpha \geq 0.10$ 은 재구성 항의 영향이 커져 FP가 증가했고,  $\alpha \leq 0.03$ 은 AE의 표현 보강 효과가 약해 소폭의 성능 저하가 관찰되었다. 따라서 본 논문에서는  $\alpha = 0.05$ 를 기본값으로 사용한다. 이 초기 모델은 쓰러짐 행동을 일정 수준 이상 감지하였으나, 모델 파라미터 수와 연산량이 높은 편으로 측정되어 모델의 무게가 무거운 편에 속했다. 즉, 실시간 적용성에 한계를 보였다.

## 2.2 초기 베이스라인 구성

대규모 ResNet 기반 백본은 높은 표현력을 제공하지만, 실시간 처리와 엣지 디바이스 배포를 위해 연산량과 모델 크기를 줄여야 했다. 따라서 첫 번째 실험 단계에서는 기존의 무거운 CNN 백본을 자체 경량 CNN으로 대체하는 것을 목표

로 하였다. 먼저, 특징 추출부를 직접 설계한 4단계 단순 컨볼루션 블록으로 대체하였다. 각 블록은 ‘3x3 컨볼루션 - BatchNorm - ReLU’ 순으로 구성하며, 채널 수를 64  $\rightarrow$  128  $\rightarrow$  128  $\rightarrow$  256으로 점진적 증감을 시켜 주요 공간 정보를 보존하도록 하였다. 이 과정에서 채널 차원을 일시적으로 축소하는 1x1 bottleneck 구조는 제외하며 구현 복잡성을 줄였으며, 그 결과 모델 파라미터 수는 약 40% 감소하고 연산량(FLOPs)은 약 35% 절감되었다.

그러나 이렇게 경량화된 백본을 그대로 사용한 모델은 검증 정확도가 약 9% 정도 하락하였다. 즉, 정상/쓰러짐 클래스 분류에서 미세한 움직임을 포착하는 데 필요한 낮은 수준의 공간 특징이 일부 손실된 것으로 분석되었다. 따라서, 이 실험은 ‘효율 VS 정확도’ 사이의 균형점을 탐색하는 중요한 계기를 심어주었다. 즉, 단순 경량화로는 실시간 적용성과 검출 성능을 동시에 만족하기 어렵기에, 표현력 저하를 보완하기 위한 후속 단계가 필요하다는 결론을 얻었다.

## 2.3 표현력 강화 및 시계열 처리 최적화

경량화 실험의 성능 저하(표현력 손실 문제)를 보완하기 위해, 공간적 표현력 강화 및 시계열 의존성 모델링 최적화라는 두 가지 측면에서 아키텍처를 보강하였다.

우선, 공간적 표현력 강화를 위해, Bottleneck과 SE 블록을 통합하였다. 경량 백본의 단순 컨볼루션 구조를 개선하여, 채널 축소-확장 형태의 bottleneck 블록에 Squeeze-and-Excitation

(SE) 모듈[26]을 결합하였다. 각 Bottleneck 블록은 '1×1 축소 → 3×3 핵심 연산 → 1×1 컨볼루션 확장' 단계로 이루어져 연산 효율을 높인다. 그 뒤에 글로벌 평균풀링을 활용한 SE 블록이 채널별 중요도를 동적으로 재조정하여 시각적 단서를 구성하는데 기여한다. 이 조합을 통해, 경량 구조 대비 약 9% 정확도 개선이 이루어졌다.

두 번째 관점인 시계열 의존성 모델링 강화 측면의 개선을 위해서는 시계열 처리 부분을 GRU로 교체할 시도하였다. 기존 LSTM 기반의 시간적 의존성 학습부에서는 긴 시퀀스에서 정보 소실과 계산 지연이 발생하는 문제점이 있었다. 이에 GRU (Gated Recurrent Unit) 구조[27]로 전환하였다. GRU는 업데이트 게이트와 리셋 게이트 2개만으로 이루어진 단순 구조로, LSTM 대비 내부 파라미터 수가 약 25% 적고 학습 및 추론 속도가 평균 15% 빨라지는 이점을 보인다. 본 실험에서는 은닉 상태의 크기를 256으로 유지하면서도 학습 속도가 20% 빨라졌다. 또한, 초기 베이스라인 대비 모델 크기와 추론 지연은 줄이면서도 정확도, 재현율 측면에서는 유지 또는 향상되는 결과를 얻었다.

결과적으로, Bottleneck+SE 모듈과 GRU 전환을 결합한 이중 개선 실험은 '경량성 유지 + 표현력 보장 + 시계열 의존성 강화' 라는 3가지의 포인트를 잡은 것이다. 추가적으로, 반복 실험에서는 이전 체크포인트를 재사용하여 학습을 이어가는 전략을 병행하였다. 각 버전의 구조, 주요 변경점, 그리고 초기/중간/개선 모델 성능은 표 2에 요약하여 제시한다.

### 3. 성능 정제 및 학습 전략

#### 3.1 손실 함수 및 멀티태스크 설계

모델 구조 개선과 병행하여 손실 설계도 정제하였다. 본 연구는 단일 분류 손실만으로는 쓰러짐 전조의 미세한 패턴을 민감하게 잡아내기 어려운 점을 고려하여, CNN이 추출한 특징을 AutoEncoder 모듈을 통해 잠재 표현으로 압축 및 재구성하는 구조를 포함시켰다. 해당 재구성 과정에서 발생하는 MSE 손실을 분류 손실과 결합함으로써, 모델이 정상/쓰러짐 상태 간의 특징 차이를 조금 더 뚜렷히 구분할 수 있도록 유도하였다. MSE는 표현력의 정제, CrossEntropy는 클래스 결정력 역할을 하며, 두 손실의 가중치는  $\alpha = 0.05$ 로 설정하여 분류 성능을 지나치게 저하시키지 않으면서도 표현력이 보장되도록 균형을 유지하였다.

학습률 스케줄링은 초기 정적 학습률인 ReduceLROn Plateau 대신 OneCycleLR로 전환하였다. 초기 학습률  $2 \times 10^{-5}$ , 최대 학습률  $1 \times 10^{-4}$ , pct\_start=0.1, div\_factor=25, final\_div\_factor=1e4라는 설정으로 전체 에폭 동안 학습률은 선상승-후감소하였다. 이 스케줄러 조합은 실험에서 보다 빠른 수렴과 검증 성능 안정화로 이어졌다.

#### 3.2 Mixed Precision, Early Stopping, Regularization

연산 효율과 수치 안정성을 동시에 확보하기 위해 mixed precision 학습을 적용하였다. Autocast 컨텍스트 내에서 주요 연산을 반정밀도(float16)로 수행하고, GradScaler를 활용하여 스케일된 손실 기반 역전파를 적용함으로써 underflow/overflow 문제 없이 학습 안정성을 유지하였다. 옵티마이저에는 Adam을 사용했고, weight decay를  $1e-5$  수준으로 소량 적용하여 과적합을 억제하였다. 또한, 학습 루프(train\_epoch)와 검증 루프

(val\_epoch)에서는 각각 CrossEntropy와 MSE 손실, 정확도 (Accuracy)를 실시간으로 모니터링하며, 검증 손실(Validation CE)이 이전 최저치보다 개선될 때마다 모델 가중치를 체크포인트로 저장한다. 이를 통해, 불필요한 과적합과 계산 낭비를 방지하면서 최적의 모델을 확보하는데 기여한다.

전체 훈련 파이프라인은 다음과 같이 진행된다. 먼저 모델 가중치와 학습 상태를 점검하여 기존 체크포인트(.pth)가 존재하면 이를 불러와 학습을 이어가고, 없을 경우 무작위 초기화로 시작한다. 이어 분류 손실(CrossEntropy)과 재구성 손실(MSE)을 정의한 뒤 Adam 옵티마이저와 OneCycleLR 스케줄러를 초기화한다. Mixed precision 학습을 적용하기 위해 autocast와 GradScaler를 활성화하고 학습을 진행한다. 각 epoch에서는 train\_epoch 단계에서 미니배치별 손실 계산과 역전파로 가중치를 갱신하고, 곧바로 val\_epoch에서 검증 CE/MSE와 정확도를 측정하여 수렴 여부를 확인한다. 검증 CrossEntropy가 직전 최저치를 갱신하면 해당 시점의 파라미터를 최적 모델로 저장하며, 일정 횟수 동안 개선이 없으면 조기 종료 기준에 따라 학습을 마무리한다. 이러한 절차는 그림 1의 흐름도로 요약된다.

#### 4. CNN/AE/GRU/Transformer 확정 하이브리드 모델 구조

제안하는 하이브리드 모델은 영상 프레임 기반의 공간적 특징과 시계열적 흐름을 결합하여 쓰러짐 행동을 보다 정교하게 분류하도록 설계되었다. 전체 아키텍처 구조는 네 개의 주요 블록 -1) CNN 백본, 2) 오토인코더(AE), 3) GRU 시퀀스 인코더, 4) Transformer 인코더-으로 구성되며, 각 블록은 잔차 연결(residual connection), 정규화(normalization), 그리고 과적합을 억제하는 드롭아웃(dropout)을 포함한다. 이는 정보 손실을

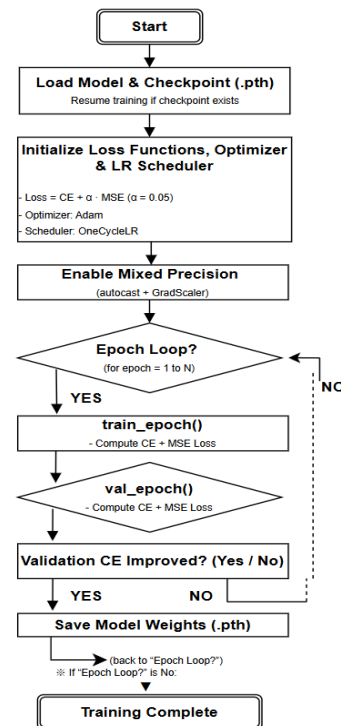


그림 1. 전체 훈련 과정.

Fig. 1. Entire training process.

최소화하고 일반화 성능을 높이는 것으로 상호 보완적으로 결합된다.

#### 4.1 CNN 기반 특징 추출

입력 시퀀스(기본 30프레임)는 개별 프레임 단위로 경량 백본에 공급된다. 이 백본은 Bottleneck 구조와 SE (Squeeze-and-Excitation) 모듈을 결합하여, 계산량은 최소화하면서도 유의미한 256차원 특징 벡터를 생성한다. 구체적으로, 각 컨볼루션 블록 이후에 채널별 가중치를 동적으로 조정하는 SE 블록을 삽입해, 쓰러짐과 비슷한 미세한 움직임 패턴이 가진 시각적 단서를 놓치지 않도록 고려하였다. 본 연구에서 사용된 CNN 블록은 경량화된 커스텀 구조로, 각 Bottleneck(+SE) 모듈의 출력을 GAP (Global Average Pooling)하여 256차원 per-frame feature로 요약하였다. 이 feature는 AE (AutoEncoder) 입력으로 전달되어 이후 단계의 정제 및 시계열 학습을 위한 공간 단서로 활용된다.

## 4.2 AutoEncoder 기반 표현 정제 보강

CNN이 추출한 256차원 feature는 그대로 사용하는 대신, AutoEncoder의 인코더를 거쳐 128차원 잠재 벡터로 압축된다. 본 연구에서는 VAE (Variational AutoEncoder) 대신 표준 AE (AutoEncoder)를 채택하여, KL 항 계산 및 추가 파라미터로 인한 연산 지연을 방지하고 실시간 추론의 안정성을 확보하였다. 이 과정은 노이즈를 차단하고, 쓰러짐 동작에서 핵심적인 특징만을 남기는 역할을 한다. 즉, AE는 배경·조명 등 비의도적 변동을 정규화하여 프레임 간 특성의 일관성을 높이고, 이후 시계열 모듈이 동작 자체의 변화에 집중하도록 돕는다. 인코더에 이어지는 디코더는 입력을 복원하며, 재구성 손실(MSE)을 분류 손실과 함께 최적화하여 표현을 매끈하게(regularize)하고 정상/비정상 패턴의 경계를 선명하게 만든다. (참고로 추론 단계에서는 재구성 손실을 사용하지 않으므로 AE로 인한 추가 지연은 최소화이다.) 결과적으로,  $CrossEntropy + \alpha \cdot MSE$ 의 멀티태스크 학습은 오답/미탐을 유발하는 배경 변동에 둔감한 표현을 형성하여, 단일 분류 손실만 사용한 경우보다 쓰러짐 전조 동작에 더 견고해진다. 이렇게 얻은 128차원 정제 임베딩은 GRU에서 근거리 시계열 의존을 안정적으로 학습하도록 기초 표현으로 사용된다.

### 4.3 GRU를 통한 시계열 의존성 학습

정제된 잠재 벡터 시퀀스(배치 크기  $B \times$  시퀀스 길이  $S \times 128$ 차원)는 2층 GRU 블록으로 전달된다. GRU는 LSTM보다 파라미터와 연산량이 적으면서도 과거 타임스텝 정보를 효과적으로 유지/갱신하는 구조를 갖고 있다. GRU의 동작 원리에 관한 식은 아래와 같다.

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1}), \quad r_t = \sigma(W_r x_t + U_r h_{t-1}), \\ \tilde{h}_t &= \tanh(W_h x_t + U_h (r_t \odot h_{t-1})), \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned} \quad (2)$$

본 연구에서는 은닉 상태 차원을 256으로 설정하고, 2개의 층을 쌓아 시퀀스 전반에 걸친 중·단기 의존성을 유지하면서 계산 효율을 확보하였다. 이를 통해 쓰러짐 동작 직전의 미세한 자세 변화나 균형 상실 패턴을 시계열 관점에서 정밀히 포착할 수 있다. GRU는 단·중거리 시계열 의존성 학

습에 특화되어 있으며, 이후 단계의 Transformer는 이를 기반으로 장거리 전역 관계를 보장하여 두 모듈이 중복되지 않고 상호 보완적으로 작동하도록 설계하였다.

#### 4.4 Transformer 기반 전역 관계 통합

GRU를 거친 출력( $B \times S \times 256$ 차원)은 Transformer 인코더 레이어로 전달된다. 본 연구의 Transformer는 ViT (Vision Transformer)가 아닌 표준 Transformer encoder이며, 1-layer,  $d_{model}=256$ ,  $nhead=4$ ,  $FFN=1024$ 의 최소 구성으로 전역 시간 관계를 보장한다. GRU가 단-중거리 시계열 의존성을 요약하고, Transformer (1-layer)가 원거리 전역 관계를 보장함으로써 두 모듈이 중복되지 않고 상호보완적으로 작동한다. 이 구조는 GRU의 지역적 맥락 요약 능력과 Transformer의 장기 의존 통합 능력을 결합하여, 앞기/숙이기/단순 놀기 등 유사 동작 경계 구분의 정확성을 향상시키는데 기여한다.

이때 Transformer는 GRU가 생성한 시계열 표현을 입력으로 받아, 각 타임스텝 간의 상호 연관성을 전역적으로 재해석한다. 본 모델은 Multi-Head Self-Attention 메커니즘을 통해 전체 시퀀스 내 모든 타임스텝 간에도 직접적 상관관계를 모델링할 수 있기 때문에, GRU가 장기 의존성을 완전하게 포착하기 어려운 긴 시퀀스에서도 중요한 전역 패턴의 손실 없이 작동한다. Transformer의 전역 어텐션 메커니즘 핵심 식은 아래와 같다[30].

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

구조는 1-layer standard encoder로, d\_model(차원) = 256, nhead(헤드수) = 4, feedforward 차원 = 1024의 단일 레이어를 사용하였다. 레이어 내 LayerNorm과 Dropout(p=0.1)을 포함하여 안정적인 학습을 돕는다. Transformer 출력 시퀀스 중 마지막 타임스텝 벡터(전역 통합 표현)를 최종 분류기(fully connected + softmax)에 입력하여, 쓰러짐/미쓰러짐 확률을 산출한다.

위 모델에 사용된 추가 최적화 기법을 서술한다.

- Residual Connection: CNN → AE → GRU → Transformer  
각 블록 사이에 잔차 연결을 도입하여, 정보가 깊은 네트워크 층을 통과할 때 손실되지 않게 하였다.
- Layer Normalization & Dropout: 각 단계별 활성화 값 분포를 정규화하고, 과적합을 줄이기 위해 드롭아웃을 병행 적용하였다.
- 멀티태스킹 학습: 분류 손실과 재구성 손실을 병합하여, 분류 정확도와 이상 패턴 탐지 감도를 동시에 향상시켰다.

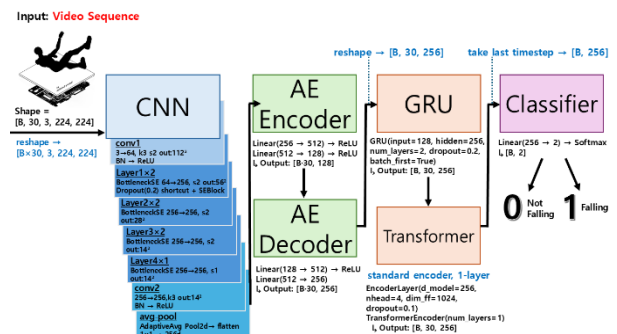


그림 2. CNN-AE-GRU-Transformer 모델 구조.

Fig. 2. CNN-AE-GRU-Transformer model architecture.

이처럼 각 블록이 서로의 약점을 보완하여 정보를 단계별로 정제·융합하기 때문에, 단일 아키텍처만 사용하는 경우와 대비하여 한계점을 극복할 수 있다. CNN은 공간적 신호를 풍부히 추출하고 AutoEncoder는 중요한 패턴을 강조하며 GRU는 시간적 흐름을 기억하고 Transformer는 전역적 상호작용을 종합한다. 이 복합적 처리 파이프라인은 단일 구조에서 흔히 겪는 과적합, 지역적 시간 의존성 한계, 전역 패턴 학습 부족 등의 문제를 동시에 해결하는데 기여한다.

#### IV. 성능 평가

##### 1. 실험 데이터

본 연구는 공개 벤치마크 기반의 쓰러짐(Fall) 및 일상행동(Daily Activity) 영상 데이터를 활용하였다. 학습과 검증에는 공개된 해당 데이터셋을 가공하여 사용하였으며, 쓰러짐과 정뿐만 아니라 이와 유사한 걷기, 앉기, 숙이기, 구부리기 등 일반적인 일상 행동을 포함하도록 구성하였다. 데이터셋은 전처리 과정에서 224x224 정규화, Train/Validation/Test = 7:2:1 분할, 그리고 Random Flip·Crop·Erase 등 데이터 증강 기법을 적용하여 데이터의 다양성과 일반화 성능을 확보하였다. 데이터셋 구성 및 전처리 요약은 표 2에 제시하였다.

본 데이터셋은 공개 벤치마크 데이터를 기반으로 전처리·가공된 형태로 재구성하였으며, 조명·각도 등 다양한 환경 변화에 대응할 수 있도록 구성되었다. 이러한 데이터 구성은 학습 과정에서 환경적 요인(밝기, 시야각 등)에 대한 견고성(robustness)를 높이고, 실시간 추론 환경에서도 안정적인 일반화 성능을 확보할 수 있도록 돕는다.

##### 2. 평가 지표

본 연구는 미탐(false negative) 최소화를 최우선 목표로 한다. 오탐(false positive)을 줄이는 것도 물론 중요하지만, 실제 쓰러짐을 놓치는 미탐은 인명 사고로 직결될 수 있는 가장 치명적인 위험이기 때문이다. 이에 따라 쓰러짐 클래스의 재현율(Recall)을 가능한 한 높게 유지하는 임계값 범위를 먼저 탐색하고, 그 범위 내에서 정밀도(Precision) 손실이 최소가 되는 점을 선택하였다. 검증 단계에서 선택한 것은 테스트 단계에도 동일하게 적용하여 일관된 기준을 유지하였다.

성능 평가는 subject-wise split 방식으로 분리된 테스트 셋에서 수행하였다. 모델의 성능을 다각도로 분석하기 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score, ROC-AUC, 혼동행렬(Confusion Matrix), 그리고 평균 추론 지연 시간(Avg. Inference Delay)을 사용하였다. Accuracy는 전체 예측 중 정답의 비율을 나타내며, 불균형이 크지 않은 이진 분류에서 전반적인 성능을 직관적으로 보여준다. Precision과 Recall은 특히 쓰러짐(Falling) 클래스에서의 오탐(False Positive)과 미탐(False Negative)을 분석하는 핵심 지표로, 안전을 위한 모니터링 환경에서 재현율을 높이는 것이 필수적이다. F1-score는 정밀도와 재현율의 조화 평균으로, 클래스별 편향 없는 균형 성능을 나타낸다. ROC-AUC는 임계값 전 범위에서 민감도와 특이도의 변화를 반영하여 분류기의 전반적 판별력을 평가한다[29]. Confusion Matrix는 두 클래스 간 오분류 패턴을 시각적으로 확인할 수 있게 하며, Avg. Inference Delay는 단일 샘플 처리 소요 시간을 측정하여 실시간성 여부를 판단하는 근거로 설정하였다.

표 2. 데이터셋 구성 및 전처리.

Table 2. ataset & Preprocessing.

|                         | Dataset and Preprocessing                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                       |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                         | Fall Image Dataset                                                                                                                                                                                                                                                                                           | Daily Activity Image Dataset                                                                                                                                                                                                                                                                                                                          |
| Source                  | <ul style="list-style-type: none"> <li>- UR Fall Dataset[19]</li> <li>- Multiple Cameras Fall Dataset[20]</li> <li>- Abnormal Behavior Video of Seniors (AI Hub)[21]</li> <li>- UMAFall: Fall Detection Dataset[22]</li> <li>- Video and Sensor Data of Fall and Hazardous Movements (AI Hub)[23]</li> </ul> | <ul style="list-style-type: none"> <li>- UR Fall Dataset</li> <li>- UMAFall: Fall Detection Dataset (Universidad de Malaga)</li> <li>- Abnormal Behavior Video of Seniors (AI Hub)</li> <li>- Video and Sensor Data of Fall and Hazardous Movements (AI Hub)</li> <li>- HMDB51 Dataset[24]</li> <li>- Extracted and edited via YouTube API</li> </ul> |
| Scale                   | 370 videos ▶ 4,012 images extracted                                                                                                                                                                                                                                                                          | 1,418 videos ▶ 19,402 images extracted                                                                                                                                                                                                                                                                                                                |
| Characteristics         | Sequences capturing the falling process                                                                                                                                                                                                                                                                      | Sequences of daily behaviors (e.g., walking, sitting, bending, crouching, etc.)                                                                                                                                                                                                                                                                       |
|                         | Diverse in age, gender, and camera angles, the dataset is enriched with variations in lighting and both indoor and outdoor daily activities                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                       |
| Augmentation Techniques | RandomResizedCrop, RandomHorizontalFlip, ColorJitter, RandomRotation, GaussianBlur, RandomErasing, Cutout                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                       |
| Preprocessing & Split   | ToTensor ▶ Normalize   Train / Val / Test = 7:2:1                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                       |

##### 3. 단계별 개선 결과

본 절에서는 앞서 제시한 Table 2의 실험 순서(Baseline → Exp.1-2 → Exp.3-5 → Exp.6 → Exp.7)에 따라 모델이 어떻게 개선되었는지를 서술한다. 모든 비교는 4.1 도입부에서 서술한 목표성-쓰러짐 미탐 최소화(Recall 우선)-을 동일하게 적용한 상태에서 이루어졌다.

Baseline의 초기 기준선은 ResNet-50 계열의 비교적 큰 백본과 AutoEncoder, 그리고 LSTM + Transformer로 구성되었다. 이 조합은 테스트 셋에서 높은 정확도 97.10%와 안정적인 쓰러짐 재현율 100%를 보였으나, 연산량과 파라미터 규모가 각각 4.14 G / 30.07 M에 달해 평균 추론 지연이 8.71ms로 측정되었다. 즉, 인식 성능은 우수하지만 실시간 적용성 측면에서 한계가 드러났다.

Exp.1-2: 1차 경량화에서는 대규모 백본을 경량 CNN으로 교체해 0.90 G / 5.15 M까지 줄였고, 지연도 4.89 ms로 절반 수준이 되었다. 하지만 쓰러짐 시퀀스의 국소적인 움직임이나 자세 변화를 충분히 포착하는 표현력이 부족해 정확도 88.41%, 쓰러짐 Recall 89%로 하락했다. 단순 축소만으로는 미세 동작 단서가 소실된다는 어려움을 확인했다.

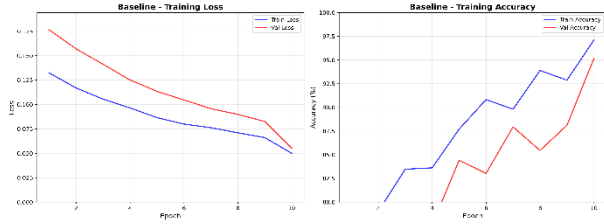


그림 3. Baseline 모델의 훈련 과정.

Fig. 3. Training process of baseline Model.

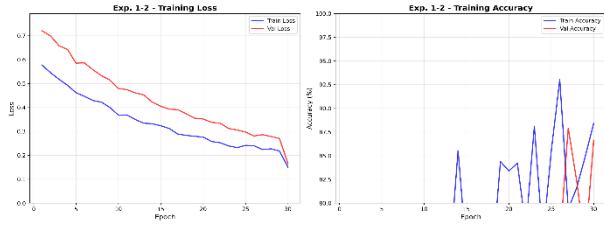


그림 4. Exp. 1-2 모델의 훈련 과정.

Fig. 4. Training process of Exp. 1-2 model.

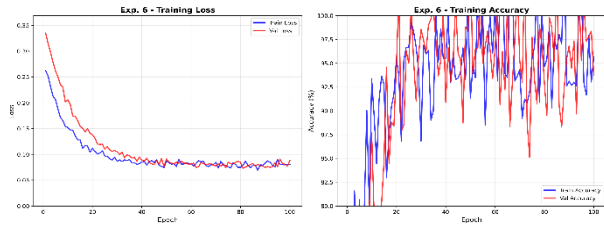


그림 5. Exp. 3-5 모델의 훈련 과정.

Fig. 5. Training process of Exp. 3-5 model.

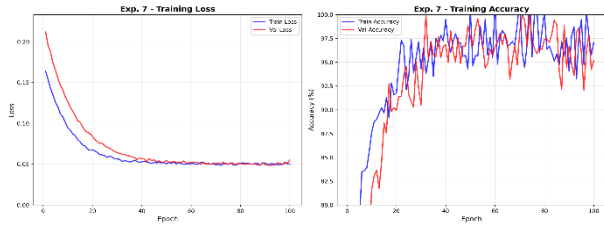


그림 6. Exp. 6 모델의 훈련 과정.

Fig. 6. Training process of Exp. 6 model.

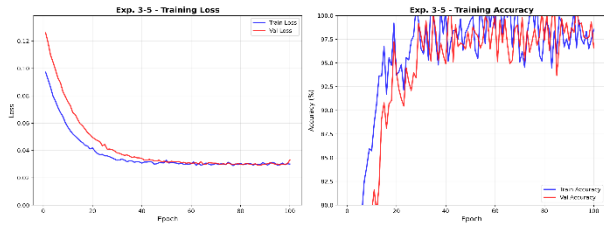


그림 7. Exp. 7 모델의 훈련 과정.

Fig. 7. Training process of Exp. 7 model.

Exp.3-5: 표현력 · 시계열 강화에서는 경량 구조를 유지하면서 Residual/Bottleneck+SE로 공간 표현을 보강하고, 순환 모듈을 2-layer GRU로 전환했다. 그 결과 정확도 98.55%, 쓰러짐 Recall 100%, FLOPs/파라미터 0.91 G / 3.22 M, 지연 6.18 ms를 달성해, 경량성을 유지하면서 성능을 향상시켰다. t-SNE

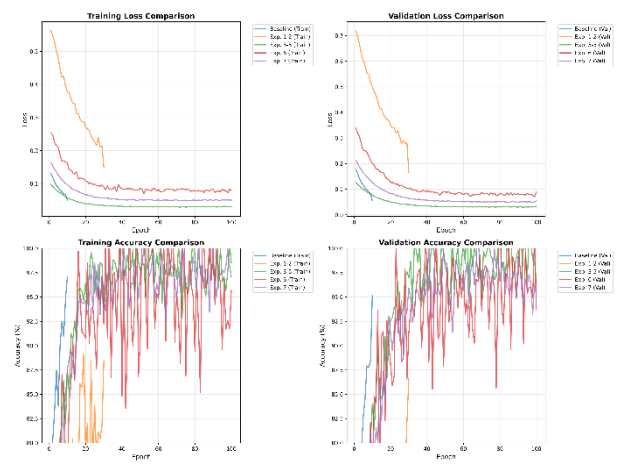


그림 8. 모든 실험의 훈련 곡선 비교.

Fig. 8. Training curves comparison of all experiments.

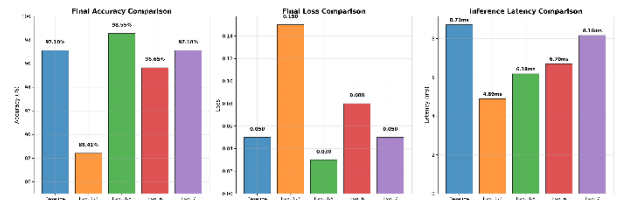


그림 9. 최종 성능 요약.

Fig. 9. Final performance summary.

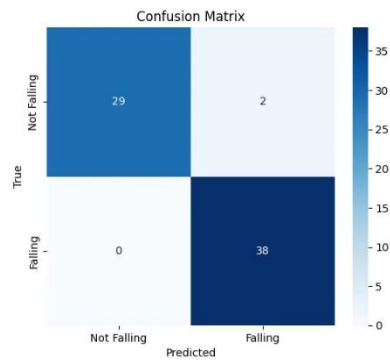


그림 10. 혼동행렬

Fig. 10. Confusion matrix.

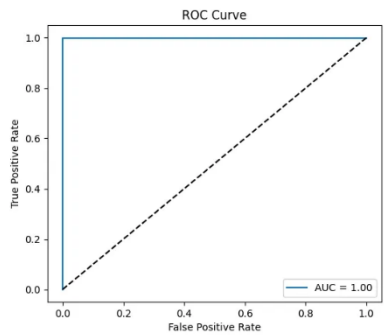


그림 11. ROC 곡선.

Fig. 11. ROC curve.

에서 두 클래스의 경계가 뚜렷해지고, 단계별 ROC도 전반적인 분별력 향상을 시사했다.

표 3. 분류 리포트.

Table 3. Classification report.

|              | Method | Accuracy | F1-score | Support |
|--------------|--------|----------|----------|---------|
| Not Falling  | 1.00   | 0.94     | 0.97     | 31      |
| Falling      | 0.95   | 1.00     | 0.97     | 38      |
| Accuracy     |        |          | 0.97     | 69      |
| Macro avg    | 0.97   | 0.97     | 0.97     | 69      |
| Weighted avg | 0.97   | 0.97     | 0.97     | 69      |

표 4. 평균 추론 지연 및 정확도 표.

Table 4. Average inference delay and test accuracy.

| Metric                  | Value  | Description           |
|-------------------------|--------|-----------------------|
| Test Accuracy           | 97.10% | Computed on Test Set  |
| Average Inference Delay | 8.16ms | Per sample (average)  |
| Total Samples           | 69     | Test set sample count |

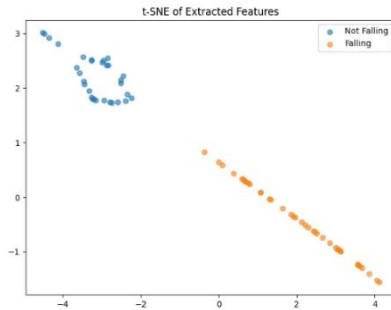


그림 12. t-SNE 임베딩.

Fig. 12. t-SNE embedding of features.

Exp.6: 확률 피드백 시도에서는 softmax 확률을 AE 잠재에 재주입(feedback\_fc)하여 재구성과 분류를 정렬하려 했으나, 테스트 기준 정확도 95.65%, 쓰러짐 Recall 97%, 지연 6.70 ms로 Exp.5 대비 성능 하락과 지연 증가가 관찰되었다. 추론 단계에서 클래스 확률 변동이 잠재코드에 과도하게 영향을 미치며 특징 분포의 수축과 일부 경계 샘플의 혼선이 발생했다. 따라서, 단일 스트림에서의 확률 피드백은 위와 같은 이유로 채택하지 않았다. 이는 분류 신호를 표현에 직접 재주입할 때 발생하는 오버피팅 및 의존성 증폭 위험성을 시사한다.

Exp.7: 데이터 확장 재학습(아키텍처 고정)에서는 Exp.5 구조로 고정하고, 비쓰러짐 및 일상 동작 영상을 추가하여 데이터 분포를 확장시키고 재분할하였다. ROC-AUC = 1.00로 분별력이 유지되었고, 혼동행렬에서 FN=0으로 미탐 최소화 목표를 충족했다. 또한, t-SNE에서도 클래스 간 분리가 보다 안정적으로 나타나 데이터 다양화가 일반화와 견고성에 기여했음을 보여준다. 다만 데이터 파이프라인의 변화로 평균 지연이 8.16ms로 소폭 증가했지만, 실사용 허용 범위를 벗어나지 않으며 Recall 우선 목표성 하에서 합리적인 교환으로 판단된다.

요약하면, 본 개선 과정은 (a) 단순 경량화의 성능 저하 → (b) 경량 상태에서의 표현력 · 시계열 강화로 성능 회복 및 향상 → (c) 새로운 추가 시도 → (d) 데이터 다양성 확대를 통한 일반화 보강의 순서로 수렴하였다. 최종적으로 Exp.5

구조 + 데이터 추가 보강(Exp.7 설정)가 미탐 최소화 · 경량성 · 실시간성의 균형을 가장 만족한다.

#### 4. 확정 모델 평가

본 확정 모델은 Exp.5 아키텍처(경량 CNN + AutoEncoder + 2-layer GRU + 1-layer Transformer)를 유지한 채, 확장 데이터셋(Exp.7 설정)으로 학습한 구성이다. 그 결과로, 표 3과 같이 정확도 97.10%를 기록했으며(67/69), 쓰러짐 Recall 100%로 미탐(FN) 0건을 달성했다. 쓰러짐 클래스의 Precision 0.95, F1-score 0.97이며, 비쓰러짐 클래스는 Precision 1.00, Recall 0.94, F1-score 0.97로 나타났다. 혼동행렬 기준 수치는 그림 3과 같이 TN=29, FP=2, FN=0, TP=38으로, 비쓰러짐에 대한 거짓 경보율(FPR)은  $2/31 \approx 6.5\%$  수준이다. ROC-AUC는 그림 4와 같이 1.00으로, 임계값 전 범위에서 높은 분별력을 확인하였다. 또한, t-SNE에서 그림 5와 같이 두 클래스가 명확히 분리되어, 데이터 확장 이후에도 표현 공간의 경계 안전성이 유지됨을 보여준다. 표 4와 같이 실시간성 측면에서, 평균 추론 지연은 8.16 ms/sample 형태로 측정되었다. 이는 연산량 0.91 G FLOPs / 파라미터 3.22 M과 결합될 때 엣지 환경 배포에 요구되는 응답성을 만족시킨다.

#### 5. 비교실험

본 절에서는 제안한 모델과 최신 메이저 컨퍼런스 기반 방법을 비교한다. 비교 대상은 Skeleton 기반 동작 인식 모델인 BlockGCN [31]와, 최근 CVPR 2024에서 공개된 포즈 추정 기반 방법 FinePOSE [32]이다. 두 기법은 공개 저장소를 통해 재현이 가능하며, 우리의 데이터 분할(Subject-wise), 시드 (0/1/2 평균), 예폭(동일), 입력 전처리(프레임 30장) 및 평가 절차를 동일하게 적용하여 공정성을 확보하였다. 본 표의 수치는 동일 프로토콜(동일 분할 · 시드 · 예폭, MPS, ms/샘플, 분류기 지연)로 재측정한 결과이며, 4.2절 기준 값과 장치 · 기준 차이로 상이할 수 있다.

속도 비교는 하드웨어/백엔드 의존성이 큰 전처리(키폰트 추출)를 분리하여, 모든 모델에 대해 공통 지표인 매개변수 수(Params), FLOPs, 분류기 지연시간(Classifier Latency, ms)을 보고한다. 본 실험에서는 포즈 기반 파이프라인의 end-to-end 시간(키폰트+분류기)을 측정하고, 분류 성능 평가는 Accuracy, F1, Latency(효율)로, 효율 평가는 매개변수 수(Parameters)와 분류기 지연시간(Classifier Latency, ms)으로 설정하였으며, 결과는 표 5와 같다.

표 5는 제안한 모델과 BlockGCN, FinePOSE (분류기 헤드 결합)의 성능과 효율을 요약한다. 포즈 기반 파이프라인의 전체 지연은 키폰트 추출 시간이 대부분을 차지하여 장치·

표 5. 비교실험: 성능과 효율.

Table 5. Comparative experiments: performance and efficiency.

|                      | Accuracy | F1   | Latency (ms) |
|----------------------|----------|------|--------------|
| BlockGCN (CVPR 2024) | 0.84     | 0.97 | 0.12         |
| FinePOSE (CVPR 2024) | 0.83     | 0.97 | 0.46         |
| Ours                 | 0.99     | 0.99 | 150.28       |

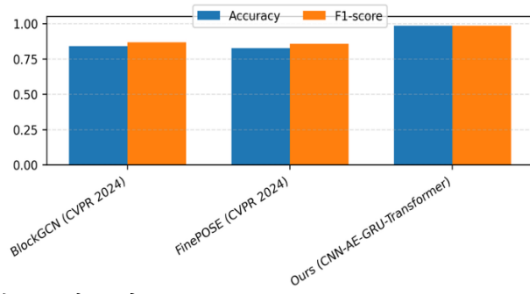


그림 13. 성능 비교(Accuracy / F1-score).

Fig. 13. Performance comparison (Accuracy / F1-score).

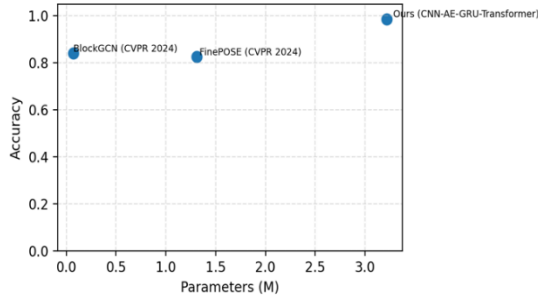


그림 14. 효율 비교(Accuracy-parameters).

Fig. 14. Efficiency comparison (Accuracy vs Parameters).

백엔드 환경에 따라 변동이 크다. 공정한 비교를 위해 본 표에는 분류기 지연(ms/샘플, MPS)만 표기하였다. 동일 환경에서 확인한 키포인트 추출 시간은 약 4.40 s/샘플이며, 분류기 시간은 BlockGCN 0.12 ms, FinePOSE 0.46 ms로 측정되었다. 따라서 두 방법의 end-to-end 시간은 각각 약 4.40012 s, 4.4004 6 s로 해석할 수 있다.

#### 6. 견고성 평가

실제 환경 변동을 모사하기 위해 테스트 분할에 한해 카메라 각도, 조명, 가림(occlusion)을 합성하여 분류 성능을 평가하였다. 이에 각도는  $-3^{\circ}/0^{\circ}/+3^{\circ}$ 로, 조명은 Dark (x0.97) /Medium (x1.00) /Bright (x1.03)로 설정하였고, 가림은 프레임 내 임의 위치에 사각형 cutout을 5%/10%/12% 비율로 적용하였다. 모든 평가는 동일 프로토콜(30-frame, 동일 전처리 · 시드 · 장치)에서 수행되었으며, 변동은 추론 단계에서만 적용하였다. 회전 손실 완화를 위해 스케일→회전→센터크롭을 사용하고, 표준화를 위해 정규화와 autocontrast+gamma (0.9-1.1)를 적용하였다.

본 표는 실제 운용대역(각도  $\pm 3^{\circ}$ , 조명  $\times 0.97$ –1.03, 가림 5–12%)에서의 결과를 요약한 것이다. 운용대역을 벗어난 과

표 6. 견고성 요약(각도 $\times$ 조명, 가림 강도별 F1-score).Table 6. Robustness summary (angle $\times$ illumination & occlusion F1-score).

| Condition              | Dark (x0.97) | Medium (x1.00) | Bright (x1.03) | Occl 5% | 10%  | 12%  |
|------------------------|--------------|----------------|----------------|---------|------|------|
| Angle $-3^{\circ}$     | 0.77         | 0.78           | 0.76           | -       | -    | -    |
| Angle $0^{\circ}$      | 0.76         | 0.76           | 0.76           | -       | -    | -    |
| Angle $+3^{\circ}$     | 0.76         | 0.76           | 0.76           | -       | -    | -    |
| Occlusion (all angles) | -            | -              | -              | 0.79    | 0.66 | 0.61 |

도한 변동(ex.각도  $\pm 15^{\circ}$  이상, 조명  $\times 0.9$  미만 ·  $\times 1.1$  초과, 가림  $\geq 20\%$ )에서는 F1 저하가 명확히 관찰되었다. 이는 본 모델이 조명, 가림 등에 본질적으로 민감하기 때문에 후속 연구에서 극한 상황에 대한 견고성을 추가로 개선할 계획이다.

#### V. 결론

본 연구는 경량 CNN-AE-GRU-Transformer로 구성된 시공간 융합 모델을 제안하며, 쓰러짐을 실시간으로 보다 신속하게 감지하기 위한 방법론을 제시하였다. 제안 모델은 작은 연산량 (약 0.91G FLOPs, 3.22M 파라미터)과 짧은 지연(8.16 ms/샘플)을 유지하면서, 테스트 셋에서 정확도 97.10%, F1-score 0.97, ROC-AUC 1.00을 보였고, 특히 쓰러짐 미탐(FN) 0 건으로 재현율(Recall) 100%를 달성했다. 이는 쓰러짐 대응에서 알림 속도가 핵심이라는 점을 뒷받침한다.

물론, 데이터의 규모와 다양성이 아직 충분하지 않고, 앉기 · 숙이기 등 유사 동작에서 오탐(FP)이 드물게 발생한다. 향후에는 실제 가정 환경 영상을 더 수집하여 다각적인 조명, 카메라 각도 사례를 증가시키고, 경량 어텐션이나 확장된 시퀀스 학습을 도입하되 모델 크기는 유지하는 형태의 연구를 지속할 예정이다. 또한, 정량화 · 프루닝 등 온디바이스 최적화를 통해 추가 경량화를 추진할 예정이다. 이를 통해 ‘가볍고 빠르면서도 놓치지 않는’ 카메라 기반 쓰러짐 감지 모델로 발전시키고자 한다.

#### REFERENCES

- [1] Statistics Korea, 2024 Census of Population and Housing, Daejeon, Korea: Statistics Korea, Jul. 2025.
- [2] Statistics Korea, 2024 Statistics on the Aged, Daejeon, Korea: Statistics Korea, Sept. 2024.
- [3] S. E. Lee, “Risk factors of falls according to living arrangements in older adults: comparison between living with others and living alone,” *J. Korea Academia-Industrial Cooperation Soc.*, vol. 25, no. 6, pp. 114-123, Jun. 2024.  
<https://doi.org/10.5762/KAIS.2024.25.6.114>
- [4] K. M. Cho, Factors influencing injuries after falls in the elderly, M.S. thesis, Dept. of Nursing Science, Graduate School, Univ. of Ulsan, Ulsan, Korea, 2017.  
<https://oak.ulsan.ac.kr/bitstream/2021.oak/6572/2/200000008897.pdf>
- [5] Korea Fire Institute, Golden time in emergency situations, Korea Fire Institute, 2020.  
<http://www.kifire.kr/board05/view.php?idx=1905&page=2&search=&find=>
- [6] T. Montero-Odasso, R. van der Velde, M. Alexander, et al., “World guidelines for falls prevention and management for older adults,” *Age and Ageing*, vol. 51, no. 9, afac205, 2022.  
<https://doi.org/10.1093/ageing/afac205>
- [7] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, “Vision-based fall detection with convolutional neural networks,” *Wireless Commun. Mob. Comput.*, Article ID 9474806, 2017.  
<https://doi.org/10.1155/2017/9474806>
- [8] S. Jiao, G. Li, G. Zhang, J. Zhou, and J. Li, “Multimodal fall detection for solitary individuals based on audio–video fusion,” *Comput. Methods Programs Biomed. Update*, art. 100532, 2024.  
<https://doi.org/10.1016/j.cmpbup.2024.100532>

- [9] P. Sitpasert, A. Namburi, and W. Sa-Ngiamvibool, "Vision-based fall detection using convolution neural network and long short-term memory," *ICIC Express Lett. Part B: Appl.*, vol. 18, no. 3, pp. 209–218, Mar. 2024.
- [10] T. Alanazi and G. Muhammad, "Human fall detection using 3D multi-stream convolutional neural networks with fusion," *Diagnostics*, vol. 12, no. 12, art. 3060, Dec. 2022. <https://doi.org/10.3390/diagnostics12123060>
- [11] J. Yang, Y. He, J. Zhu, Z. Lv, and W. Jin, "Fall detection method for infrared videos based on spatial-temporal graph convolutional network," *Sensors*, vol. 24, no. 14, art. 4647, 2024. <https://doi.org/10.3390/s24144647>
- [12] F. A. Elshwemy, "A new approach for thermal vision based fall detection using autoencoder and ConvLSTM," *Proc. 2020 INASS*, 2020.
- [13] I. Kareem, S. F. Ali, M. Bilal, and M. S. Hanif, "Exploiting the features of deep residual network with SVM for fall detection," *PLoS ONE*, vol. 19, no. 4, e0314959, Apr. 2024. <https://doi.org/10.1371/journal.pone.0314959>
- [14] A. Núñez-Marcos and I. Arganda-Carreras, "Transformer-based fall detection in videos," *Eng. Appl. Artif. Intell.*, vol. 132, Art. 107937, Jun. 2024. <https://doi.org/10.1016/j.engappai.2024.107937>
- [15] J. Fernandez-Bermejo, J. Martinez-del-Rincon, J. Dorado, X. del Toro, M. J. Santofimia, and J. C. Lopez, "Edge computing transformers for fall detection in older adults," *Int. J. Neural Syst.*, vol. 34, no. 5, art. 2450026, 2024. <https://doi.org/10.1142/S0129065724500266>
- [16] M. Gao, "Fall detection based on OpenPose and MobileNetV2 network," *IET Image Process.*, 2023. <https://doi.org/10.1049/ipr2.12667>
- [17] J. A. Santoyo-Ramón, E. Casilari-Pérez, and J. M. Cano-García, "A study on the impact of the users' characteristics on the performance of wearable fall detection systems," *Sci. Rep.*, vol. 11, art. 23011, Dec. 2021. <https://doi.org/10.1038/s41598-021-02537-z>
- [18] J.-H. Hong, J.-H. Lim, H.-B. Park, and D. Shin, "Design of flexible fall detection wristband and fall notification system considering wearability for the elderly," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 29, no. 7, pp. 550–555, 2023.
- [19] UR Fall Detection Dataset (URFD), University of Rzeszów, Rzeszów, Poland, 2014.
- [20] Le2i/IMVIA Fall Detection Dataset, Université de Bourgogne, Dijon, France, 2020.
- [21] Abnormal Behavior of Seniors (Video) Dataset, National Information Society Agency (AI Hub), Daejeon, Korea, 2020.
- [22] E. Casilari, J. A. Santoyo-Ramón, and J. M. Cano-García, "UMAFall: A multisensor dataset for the research on automatic fall detection," *Procedia Comput. Sci.*, vol. 110, pp. 32–39, 2017.
- [23] Video and Sensor Data of Fall and Hazardous Movements, National Information Society Agency, Daejeon, Korea, 2020.
- [24] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," *Proc. ICCV*, 2011, pp. 2556–2563.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, Dec. 2015. <https://arxiv.org/abs/1512.03385>
- [26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *arXiv preprint arXiv:1709.01507*, 2017. <https://arxiv.org/abs/1709.01507>
- [27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014. <https://arxiv.org/abs/1412.3555>
- [28] P. Mikićević, S. Narang, J. Alben, et al., "Mixed precision training," *ICLR*, 2018. <https://arxiv.org/abs/1710.03740>
- [29] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [30] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," *Proc. NeurIPS*, 2017. <https://doi.org/10.48550/arXiv.1706.03762>
- [31] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua, "BlockGCN: Redefining topology awareness for skeleton-based action recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Code: <https://github.com/ZhouYuxuanYX/BlockGCN>
- [32] J. Xu, Y. Guo, and Y. Peng, "FinePOSE: Fine-grained prompt-driven 3D human pose estimation via diffusion models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 561–570, 2024. Code: [https://github.com/PKU-ICST-MIPL/FinePOSE\\_CVPR\\_2024](https://github.com/PKU-ICST-MIPL/FinePOSE_CVPR_2024)
- [33] Y. M. Kim, and J. E. Ha, "Comparison of CPU and GPU-based image processing performance in OIS inspection equipment," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 31, no. 9, pp. 985–989, 2025.
- [34] J. Lee, H. J. Lee, S. K. Arachchige, N. Kim, H. Heo, and K. Lee, "Search operations with geolocation estimation of missing persons based on real-time drone images," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 8, pp. 890–896, 2024.



**김 은 영**

2025년 덕성여자대학교 소프트웨어전공 학사 재학중. 관심분야는 Deep Learning, Machine Learning, Computer Vision, Web/App Development, Interactive Data Visualization.



**이 재 호**

2005년 고려대학교 전자컴퓨터공학과 석사. 2008년~2013년 동 대학 전기전자 전과공학과 박사. ~2015년 LG전자 차세대표준연구소 선임연구원. ~2019년 서원대학교 정보통신공학과 조교수. ~2024년 덕성여자대학교 소프트웨어전공 조교수. 2025년~현재 덕성여자대학교 소프트웨어전공 부교수. 관심분야는 WPAN, MAC, Bluetooth, Wi-Fi, Localization, NLP, ML.