Smart Media Journal / Vol.15, No.6 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.6.130

스켈레톤 기반 ST GCN과 다중 인스턴스 학습을 이용한 감시 영상 이상행동 탐지

(Surveillance video anomaly detection using skeleton based ST GCN and multi instance learning)

변무혁*, 양현성*, 정세훈**, 심춘보***

(Mu-Hyeok Byun*, Hyun-Sung Yang*, Se-Hoon Jung**, Chun-Bo Sim***)

요 약

본 논문에서는 감시 영상에서 이상행동을 탐지하기 위한 스켈레톤 기반 약지도 학습 프레임워크를 제안한다. 기존 RGB(Red Green Blue) 외관 기반 방법은 외관 변화에 취약하고 개인정보 노출 위험이 있다는 한계가 존재한다. 제안 방법은 인체의 관절 좌표로부터 ST GCN(Spatial Temporal Graph Convolutional Network)으로 시공간 특징을 학습하고, 영상 수준 라벨만을 사용하는 다중 인스턴스 학습(Multiple Instance Learning, MIL)으로 모델을 훈련한다. 또한 기존 RGB 환경에서 정립된 MIL 손실 함수가 스켈레톤 입력 데이터의 특성을 온전히 반영하지 않음을 ablation study로 입증하고 이를 재설계한 손실 함수를 제안한다. 실험 결과, 제안 방법은 UCF Crime 데이터셋에서 프레임 수준 AUC ROC(Area Under the Receiver Operating Characteristic Curve) 0.8368을 달성하여 I3D 기반 베이스라인의 0.7935를 능가했으며, 최신 약지도 베이스라인과의 비교에서도 더 높은 성능을 보였다.

■ 중심어 : 영상 이상행동 탐지 ; 스켈레톤 기반 행동 인식 ; ST GCN ; 다중 인스턴스 학습 ; 약지도 학습

Abstract

This paper proposes a skeleton-based weakly supervised learning framework for detecting abnormal behavior in surveillance footage. Existing RGB (Red Green Blue) appearance-based methods have limitations, such as vulnerability to changes in appearance and the risk of personal information exposure. The proposed method learns spatiotemporal features from human joint coordinates using a Spatial Temporal Graph Convolutional Network (ST GCN) and trains the model using Multiple Instance Learning (MIL), which utilizes only image-level labels. Furthermore, through an ablation study, we demonstrate that the MIL loss function established in the existing RGB environment does not fully reflect the characteristics of skeleton input data, and we propose a redesigned loss function. Experimental results show that the proposed method achieved a frame-level AUC ROC (Area Under the Receiver Operating Characteristic Curve) of 0.8368 on the UCF Crime dataset, surpassing the I3D-based baseline's 0.7935 and demonstrating higher performance compared to the latest weakly supervised baseline.

■ keywords : Video Anomaly Detection ; Skeleton-Based Behavior Recognition ; ST GCN ; Multiple Instance Learning ; Weakly Supervised Learning

Ⅰ. 서 론

최근 공공, 교통, 상업 등 다양한 환경에서 폐쇄회로 텔레비전(CCTV)을 비롯한 감시 카메라 보급이 빠르게 확대되고 있다. 공공기관이 설치 및 운영하는 CCTV는 2021년 약 146만 대에서 2023년 약 177만 대로 매년 증가하는 추세이며 [1], 이에 따라 영상 기반 이상행동 탐지 기술에 대한 수요도 함께 증가하고 있다. 그러나 실제 사고나 범죄 영상을 사람이 직접 모니터링하는

* 정회원, 국립순천대학교 멀티미디어공학부
** 정회원, 국립순천대학교 컴퓨터공학과
*** 정회원, 국립순천대학교 인공지능공학부
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25434896)
접수일자 : 2026년 05월 20일 게재확정일 : 2026년 06월 02일
교신저자 : 심춘보 e-mail : cbsim@sunchon.ac.kr

방식은 영상의 양이 방대하고 이상 사건의 발생 빈도가 낮아 효율적이지 못하다. 실제로 일부 통합관제센터에서는 관제 요원 한 명이 수백 대의 CCTV를 감시해야 하는 인력 부족 문제가 보고되고 있어, 사람에 의한 상시 모니터링만으로는 이상 사건을 놓치기 쉽다[2]. 따라서 방대한 감시 영상 속에서 폭행, 절도, 방화 등 이상 사건이 발생한 구간을 효율적으로 탐지하는 기술이 필수적으로 요구된다. 그러나 실제 감시 환경에서는 정상 영상과 이상 영상의 비율이 극심하게 불균형하며, 이상 사건은 종류가 다양하고 사전에 모든 유형을 정의하기 어렵다는 점에서 일반적인 행동 인식 문제보다 어려운 과제다.

이러한 어려움 때문에 초기 연구는 정상 데이터만으로 학습한 후 정상 분포에서 벗어나는 표본을 이상으로 판정하는 단일 클래스 학습 방식이 주를 이뤘다. 그러나 이 접근법은 정상 행동의 다양성을 충분히 포착하지 못해, 학습 단계에서 보지 못한 정상 패턴까지 이상으로 오탐지하는 한계가 있다. 이를 해결하기 위해 Sultani[3]는 영상 수준의 약한 라벨만 사용하여 학습 가능한 다중 인스턴스 학습 기반의 이상행동 탐지 프레임워크를 제안했고, I3D(Inflated 3D ConvNet)[4] 또는 C3D(Convolutional 3D) RGB(Red Green Blue) 특징을 입력으로 사용하여 UCF Crime 데이터셋에서 AUC ROC(Area Under the Receiver Operating Characteristic Curve) 0.7935의 성능을 보고했다. 이후 다양한 후속 연구가 MIL 프레임워크를 확장했으나 대부분 RGB 외관 정보에 의존하기 때문에 조명, 배경, 의상 등 외관 변화에 취약하고 영상 데이터의 개인정보 노출 위험이 존재한다는 한계가 있었다.

이러한 한계를 극복하기 위한 대안으로 스켈레톤(skeleton) 기반 행동 인식이 주목받고 있다. 인체 골격 정보는 외관 변화에 강건하고 개인정보를 직접 노출하지 않으면서도 행동의 본질적 특징인 자세와 움직임 패턴을 표현할 수 있다는 장점을 가진다. 그러나 기존의 스켈레톤 기반 방법은 대부분 완전 지도학습 환경에서 짧고 정렬된 행동 클립을 대상으로 평가됐으며, 영상 수준 라벨만 사용하는 약지도 학습 환경에서의 이상 행동 탐지에는 적용된 사례가 제한적이다.

본 논문에서는 이러한 한계를 극복하기 위해 스켈레톤 기반 약지도 학습(weakly supervised) 프레임워크를 제안한다. 다중 인체의 관절 좌표로부터 시공간 특징을 학습하고 영상 수준 라벨만을 사용하는 MIL 랭킹 손실로 모델을 학습하며, 기존 RGB 환경에서 정립된 MIL 손실 함수가 스켈레톤 입력 환경의 특성을 온전히 반영하지는 못해 손실 함수를 재설계한다.

제안하는 프레임워크는 크게 세 단계로 구성된다. 먼저 YOLOv11 Pose[5]를 이용하여 입력 영상의 매 프레임에서 다중 인체의 관절 좌표를 추출하고, 길이가 가변적인 영상을 일정 개수의 스니펫으로 분할하여 시간 축을 정규화한다. 다음으로 각 스니펫의 골격을 ST GCN[6]에 입력하여 공간 그래프 합성곱으로 인접 관절 사이의 구조적 관계를, 시간 합성곱으로 인접 프레임 사이의 움직임을 함께 학습함으로써 시공간 특징을 추출한다. 마지막으로 완전 연결 계층과 sigmoid를 거쳐 스니펫별 이상 점수를 산출하며, 한 스니펫에 여러 인체가 존재하는 경우 가장 높은 점수를 대표 점수로 선택한 뒤 MIL 랭킹 손실로 이상 영상과 정상 영상을 구분하도록 학습한다. 특히 기존 RGB 기반 MIL 손실의 강한 이상 희소성 정규화가 스켈레톤 입력에서는 이상 점수를 과도하게 억제하는 문제를 확인하고, 이상 희소성 가중치를 축소하는 한편 정상 영상의 점수를 낮추는 정규화 항을 추가하여 정상과 이상의 점수 분리를 개선하도록 손실 함수를 재설계한다.

본 논문의 주요 기여는 다음과 같다. 첫째, 스켈레톤 입력과 MIL을 결합한 약지도 학습 기반 이상행동 탐지 프레임워크를 제안하고 UCF Crime 데이터셋에서 기존 베이스라인을 능가하는 성능을 달성했다. 둘째, 기존 MIL 손실 함수의 정규화 가중치가 스켈레톤 입력 환경에 그대

로 적용하기에는 한계가 있음을 ablation study로 정량적으로 입증하고, 이를 재설계한 손실 함수를 제안했다. 셋째, 손실 함수 구성, 입력 채널, 시간 풀링 방식, 옵티마이저의 설계 요소에 대한 ablation study를 수행하여 각 설계 선택의 정량적 기여도를 검증했다. 넷째, 스켈레톤 기반 접근의 강점과 한계를 이상행동 클래스별로 분석하여 향후 연구 방향을 제시했다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 감시 영상 이상행동 탐지, 스켈레톤 기반 행동 인식, 인체 자세 추정에 관한 관련 연구를 소개한다. Ⅲ장에서는 제안하는 스켈레톤 기반 약지도 학습 프레임워크의 전체 구조, ST GCN 네트워크 구성, MIL 손실 함수 설계 및 학습 설정을 상세히 기술한다. Ⅳ장에서는 UCF Crime 데이터셋에서 제안 방법의 성능을 베이스라인과 비교하고 클래스별 분석 및 ablation study 결과를 제시한다. 마지막으로 Ⅴ장에서 결론과 향후 연구 방향을 논한다.

Ⅱ. 관련 연구

1. 감시 영상 이상행동 탐지

감시 영상에서 이상행동 탐지는 정상 영상과 이상 영상이 극심하게 불균형하고 이상 사건의 정의가 모호하며 영상이 길고 다양한 시각적 조건을 포함한다는 점에서 일반적인 행동 인식 문제보다 어렵다. 초기 연구는 정상 데이터만으로 학습한 후 정상 분포에서 벗어나는 표본을 이상으로 판정하는 단일 클래스 방식이 주를 이루었으나[7, 8], 정상 행동의 다양성을 모두 포착하기 어려워 실제 감시 환경에서의 일반화 성능에 한계가 있다.

Sultani는 영상 수준의 약한 라벨만을 사용하여 학습 가능한 다중 인스턴스 학습 기반 이상행동 탐지 프레임워크를 제안했다. 이 방법은 영상을 일정 개수의 스니펫(snippet)으로 분할하고 각 영상을 스니펫의 집합인 bag로 간주한다. 이후 이상 영상 bag 내에서 이상 점수가 가장 높은 스니펫의 값이 정상 영상 bag 내 최고 점수보다 크도록 hinge 형태의 랭킹 손실로 학습한다. 또한 실세계 환경의 13개 이상행동 클래스로 구성된 대규모 UCF Crime 데이터셋도 함께 공개했으며, I3D 또는 C3D 기반의 RGB 특징을 입력으로 받아 AUC ROC 0.7935의 성능을 보였다. 이후 다양한 후속 연구가 발표됐으나[9, 10], 대부분 RGB 특징에 의존하기 때문에 외관 정보가 부족한 환경에서는 강건성이 떨어지는 한계가 있다.

2. 스켈레톤 기반 행동 인식

RGB 영상 대신 인체 골격을 입력으로 사용하는 행동 인식 방법은 조명, 배경, 의상 등 외관 변화에 강건하며 개인 정보 보호 측면에서도 유리하다는 장점을 가진다. 초기 스켈레톤 기반 방법은 관절 좌표 시계열을 RNN(Recurrent Neural Network) 또는 CNN(Convolutional Neural Network)으로 처리했으나, 인체 골격이 가지는 그래프 구조를 명시적으로 활용하는 데 제약이 따른다.

ST GCN은 인체 골격을 그래프 로 표현하고, 공간 축에서는 그래프 합성곱으로 인접 관절 간 정보를 교환하여 시간 축에서는 시간 합성곱으로 인접 프레임 간 정보를 교환하는 시공간 합성곱 신경망이다. ST GCN은 행동 인식 벤치마크에서 우수한 성능을 보이며, 이후 2s-AGCN[11], CTR GCN[12], MS-G3D[13]의 베이스라인으로 사용되었다.

3. 스켈레톤 기반 행동 인식을 위한 인체 자세 추정

스켈레톤 기반 행동 인식의 전제는 정확한 인체 자세 추정이다. OpenPose[14], AlphaPose[15], HRNet[16], Mask R-CNN 기반 자세 추정[17] 등의 자세 추정 모델이 정확도 측면에서 우수한 성능을 보이나, 감시 영상 환경에서 요구되는 다중 인체 검출과 실시간 처리를 동시에 만족시키기에는 한계가 있다. 본 논문에서는 객체 검출과 자세 추정을 단일 단계로 통합한

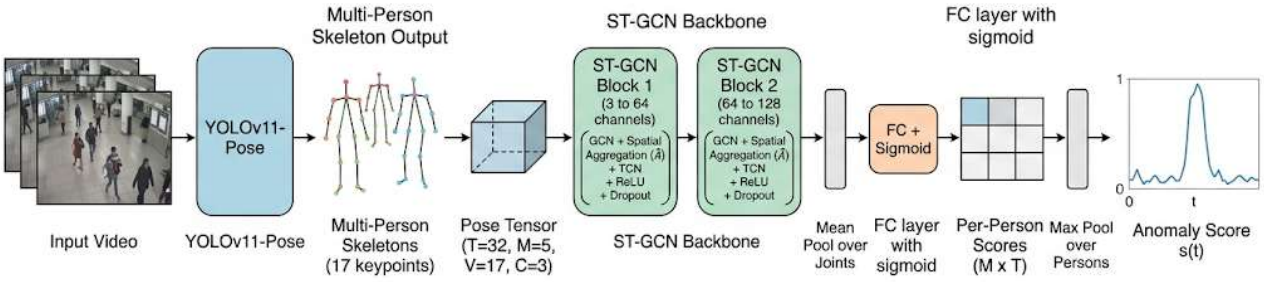

그림 1. 제안하는 스켈레톤 기반 이상행동 탐지 프레임워크의 전체 구조

YOLOv11 Pose를 사용한다. YOLOv11 Pose는 COCO[18] keypoint 규약에 따라 인체당 17개 관절을 x, y, confidence의 3차원 정보로 출력하며, 한 프레임 내 다중 인체를 동시에 검출할 수 있어 군중 환경의 감시 영상에 적합하다.

Ⅲ. 제안 방법

1. 전체 구조

본 논문에서는 감시 영상에서 이상행동을 탐지하기 위해 스켈레톤 기반 약지도 학습 프레임워크를 제안한다. 제안하는 방법은 먼저 YOLOv11 Pose를 통해 다중 인체 스켈레톤을 추출하고, ST GCN으로 시공간 특징을 학습하며, MIL 랭킹 손실로 영상 수준 이상행동을 분류하는 세 단계로 구성된다.

그림 1은 제안하는 방법의 전체 구조다. 입력 영상으로부터 매 프레임마다 다중 인체의 17개 관절 좌표를 추출하고, 영상을 T개의 균등 스니펫으로 분할한다. 각 스니펫은 ST GCN을 통해 시공간 특징으로 변환되며, 완전 연결 계층과 sigmoid 활성화 함수를 거쳐 [0, 1] 범위의 이상 점수(anomaly score)로 출력된다. 한 프레임 내 다수의 인체가 존재할 경우, 최대 풀링(max pooling)을 적용하여 가장 높은 이상 점수를 산출한 객체의 점수를 해당 스니펫의 대표 점수로 산출한다.

2. 다중 인체 스켈레톤 추출

입력 영상의 매 프레임에서 YOLOv11 Pose 모델을 이용하여 인체 스켈레톤을 추출한다. YOLOv11 Pose는 각 관절에 대해 x, y, confidence의 3차원 특징을 출력한다. confidence 값은 해당 관절의 검출 신뢰도를 의미하며, 미검출 또는 가림(occlusion) 상황에서 학습에 유용한 보조 신호로 활용된다.

다중 인체 환경을 고려하여 프레임당 최대 $M=5$명까지 검출하고, 검출 인원이 5명 미만인 경우 0으로 패딩한다. 영상의 길이가 가변적이기 때문에 시간 축 정규화를 위해 영상을 32개의 균등 스니펫으로 분할하고, 각 스니펫 내 프레임 특징을 평균하여 단일 표현으로 집약한다. 따라서 영상 한 편의 입력 텐서 형상은 $(T, M, V, C) = (32, 5, 17, 3)$이며, T는 스니펫 수, M은 프레임당 최대 검출 인원 수, V는 관절 수, C는 채널 수다.

3. ST GCN 구조

ST GCN은 인체 골격을 그래프 $G=(V,E)$로 표현하여 공간 그래프 합성곱(Graph Convolutional Network, GCN)과 시간 합성곱(Temporal Convolutional Network, TCN)을 결합한 신경망이다. V는 17개 관절 노드 집합이고, E는 인접한 관절을 잇는 골격 간선 집합이다.

COCO 17 keypoint Topology에 따라 18개의 골격 간선과 self loop를 포함하는 인접 행렬 A를 구성하고, 차수 행렬(degree matrix) D를 이용하여 식 (1)과 같이 정규화한다.

$$\tilde{A} = D^{-1}A \quad (1)$$

$\tilde{A}$는 정규화된 인접 행렬이고, D는 인접 행렬 A의 각 행의 합을 대각 성분으로 갖는 대각 차수 행렬이다. 하나의 ST GCN 블록은 GCN 계층과 TCN 계층의 순차 결합으로 구성되며 식 (2)와 같다.

$$f_{out} = \mathrm{Dropout}(\mathrm{ReLU}(\mathrm{TCN}(\tilde{A} \cdot \mathrm{GCN}(f_{\in})))) \quad (2)$$

GCN은 커널 크기 1×1의 합성곱으로 채널 변환을 수행하고, $\tilde{A} \cdot \mathrm{GCN}(f_{\in})$는 정규화된 인접 행렬과의 텐서 곱 연산을 통해 인접 관절 간 정보를 교환한다. TCN은 시간 축 커널 크기 3의 1차원 합성곱으로 인접 스니펫 간 시간적 상호작용을 학습한다. 이후 3→64→128개로 확장된 채널은 2개의 ST GCN 블록으로 구성된다. 마지막 블록의 출력 $(T, V, 128)$에 대해 관절 축으로 평균 풀링을 적용하여 스니펫별 128차원 임베딩을 얻고, 완전 연결 계층을 통해 1차원 점수로 투영한 후 sigmoid를 거쳐 [0, 1] 범위의 이상 점수로 변환한다.

다중 인체 처리를 위해, 영상 내 M명의 인체 각각에 대해 ST GCN을 적용한 후 검출된 골격에 대해 max pooling을 수행한다. t번째 스니펫에서의 영상 점수 $s(t)$는 식 (3)과 같다.

$$s(t) = \max_m(\sigma(W h_m(t))) \quad (3)$$

$h_m(t)$는 t번째 스니펫에서 m번째 인체의 128차원 임베딩, σ는 sigmoid 함수, W는 완전 연결 계층의 가중치다. 이는 MIL의 기본 가정인 bag 내에서 이상 점수가 가장 높은 instance가 bag를 대표한다는 원리를 다중 인체 환경으로 확장한 것이다.

4. MIL 손실 함수

UCF Crime 데이터셋은 영상 수준의 정상 및 이상 라벨만 제공하며 프레임 수준의 이상 구간 정보는 학습에 사용하지 않는 약지도 학습 환경이다. 이를 위해 본 논문에서는 MIL 랭킹 손실을 기본 구조로 채택했다. 그러나 기존 손실 함수에 포함된 강한 이상 희소성 가중치가 스켈레톤 입력 환경에서는 오히려 이상 점수를 과도하게 억제하여 정상 및 이상 데이터의 분리를 저해한다는 점을 확인했다. 이에 따라 이상 희소성 가중치를 1/8 수준으로 축소하고, 정상 영상 점수 분포를 약하게 누르기 위한 정상 영상 점수 억제 항(normal sparsity)을 보조 정규화 항으로 추가하여 손실 함수를 재설계한다.

영상 한 편을 T개의 스니펫으로 구성된 bag로 간주하고, 이상 영상 bag의 최고 점수가 정상 영상 bag의 최고 점수보다 높아지도록 훈련한다. 전체 손실 함수는 식 (4)와 같이 네 개 항의 가중합으로 정의된다.

$$L = L_{rank} + \lambda_s L_{smooth} + \lambda_n L_{normal} + \lambda_a L_{anomaly} \quad (4)$$

각 항은 식 (5)~(8)과 같다.

$$L_{rank} = \max(0, 1 - \max_t s_a(t) + \max_t s_n(t)) \quad (5)$$

$$L_{smooth} = \frac{1}{T-1}\sum_{t=1}^{T-1}(s_a(t+1) - s_a(t))^2 \quad (6)$$

$$L_{normal} = \frac{1}{T}\sum_{t=1}^{T} s_n(t) \quad (7)$$

$$L_{anomaly} = \sum_{t=1}^{T} s_a(t) \quad (8)$$

식 (4)~(8)에서 사용되는 기호는 다음과 같다. $s_a(t)$와 $s_n(t)$는 각각 이상 영상과 정상 영상의 t번째 스니펫 점수고, $T = 32$는 영상당 스니펫 개수다. 손실 함수 가중치는

$\lambda_s = 8 \times 10^{-5}$, $\lambda_n = 1 \times 10^{-4}$, $\lambda_a = 1 \times 10^{-5}$로 설정한다.

L_{rank}는 MIL의 핵심 가정으로, 이상 영상의 최대 점수가 정상 영상의 최대 점수보다 hinge margin 1 이상 크도록 강제한다. L_{smooth}는 인접 스니펫 간 점수 차이를 작게 유지하여 시간적 일관성을 학습한다. L_{normal}은 정상 영상의 전체 점수 평균을 낮추는 정규화 항으로 정상 점수 분포를 직접 억제하여 정상 및 이상 점수 분리를 개선한다. $L_{anomaly}$는 이상 영상에서도 일부 스니펫만이 실제 이상 구간이라는 사전 지식을 반영하여 이상 점수의 희소성을 유도한다.

Ⅳ. 실험 결과 및 고찰

1. 데이터셋 및 평가 환경

본 논문에서 제안하는 방법의 성능을 평가하기 위해 UCF Crime 데이터셋을 활용했다. UCF Crime은 대규모 이상행동 탐지 데이터셋으로 13개 이상행동 클래스와 정상 영상으로 구성된다. 전체 약 1,900편의 영상 중 훈련 데이터로 정상 영상 800편과 이상 영상 810편을 사용했고, 테스트 데이터로 정상 영상 150편과 이상 영상 140편을 사용했다. 모든 영상은 30 FPS(Frames Per Second)이며, 테스트 데이터에 대해서는 프레임 수준의 이상 구간 시작, 종료 프레임이 주석으로 제공된다. 실험 환경은 표 1과 같다.

표 1. 실험 환경 구성

H/W		S/W	
CPU	Intel I7-13700K	OS	Windows 11
GPU	NVIDIA RTX 4080 Super	Programming language	Python 3.10.18
RAM	DDR5 64 GB RAM	Deep learning framework	PyTorch 2.5.1

안정적인 MIL 학습을 위해 옵티마이저로 AdaGrad[19]를 사용한다. 배치 크기는 16으로 설정하고, 매 미니배치마다 정상 영상 8개와 이상 영상 8개를 쌍으로 구성하여 식 (5)의 랭킹 손실을 계산한다. 학습률 스케줄러는 patience를 5로 설정한 ReduceLROnPlateau를 사용하며 손실이 개선되지 않을 때 학습률을 0.5배로 감쇠시킨다. 최대 200 에폭 학습하되 학습 손실이 15 에폭 동안 개선되지 않으면 학습을 조기 종료한다. MIL 환경에서는 검증 손실과 테스트 AUC의 상관성이 낮은 점을 고려하여 학습 손실 최저값을 기준으로 최적 가중치를 저장한다.

본 논문에서는 이상행동 탐지의 다각적 평가를 위해 8개 성능 지표를 사용한다. 평가는 UCF Crime 표준 평가 방식과 동일하게 프레임 수준에서 수행한다. 각 영상에서 산출한 스니펫별 이상 점수를 해당 구간의 프레임 점수로 확장하고, 테스트 영상에 제공되는 프레임 수준 이상 구간 주석을 정답으로 사용하여 전체 테스트 프레임에 대해 지표를 계산한다. AUC ROC는 ROC 곡선 하부 면적으로 임계값에 무관한 분류 성능을 측정하는 지표이며, AUC PR(Area Under the Precision-Recall Curve)은 Precision Recall 곡선 하부 면적으로 클래스 불균형 환경에서 AUC ROC보다 더 민감하게 작동한다. Accuracy는 전체 표본 중 올바르게 분류된 비율을 의미하고, Precision은 이상으로 예측한 표본 중 실제 이상의 비율을 의미하며, Recall은 실제 이상 표본 중 올바르게 탐지된 비율로 민감도다. F1 Score는 Precision과 Recall의 조화 평균으로 두 지표를 균형 있게 반영하는 지표이며, Specificity는 실제 정상 표본 중 올바르게 분류된 비율이다. FAR(False Alarm Rate)은 정상 표본을 이상으로 잘못 탐지한 비율로 1에서 Specificity를 뺀 값이다.

임계값 의존 지표인 Accuracy, Precision, Recall, F1 Score, Specificity, FAR에 대해서는 ROC 곡선상에서 TPR(True Positive Rate)과 FPR(False Positive Rate)의 차이를 최대화하는 Youden's J 통계량 기반 최적 임계값을 적용했다.

2. 실험 결과

제안 방법의 성능을 검증하기 위해 두 가지 베이스라인과 비교했다. 첫 번째는 Sultani가 제안한 MIL 기반 모델로, UCF Crime 데이터셋을 처음 공개하며 함께 제시한 표준 베이스라인이자 이후 다수의 약지도 이상행동 탐지 연구가 공통 비교 기준으로 채택해 온 모델이기 때문에 선택했다. 이 모델은 영상에서 추출한 1024차원 I3D RGB 특징을 입력으로 받아 512, 32, 1차원으로 이어지는 3계층 완전 연결 신경망과 sigmoid를 거쳐 스니펫별 이상 점수를 산출하며, MIL 랭킹 손실로 학습한다. 두 번째는 RTFM(Robust Temporal Feature Magnitude learning)으로, 다중 스케일 시간 네트워크와 feature magnitude 기반 스니펫 선택을 결합한 최신 약지도 이상행동 탐지 모델이다. 표 2는 두 베이스라인과 제안하는 모델의 프레임 수준 전체 성능을 비교한 결과다.

표 2. 베이스라인과 제안하는 ST GCN 모델의 전체 성능 비교

Metric	Baseline	RTFM	Proposed
AUC ROC	0.7935	0.7932	**0.8368**
AUC PR	0.2160	0.2037	0.0221
Accuracy	0.6871	0.7593	**0.8345**
Precision	0.1693	0.2018	0.0255
Recall	0.8002	0.7360	**0.9138**
F1 Score	0.2795	0.3168	0.0497
Specificity	0.6778	0.7612	**0.8341**
FAR	0.3222	0.2388	**0.1659**

표 2에서 제안 방법은 AUC ROC에서 0.8368을 달성하여 Sultani 베이스라인의 0.7935 대비 0.0433 향상을 보인다. 최신 약지도 베이스라인인 RTFM은 동일한 I3D 특징을 입력으로 사용했을 때 AUC ROC 0.7932를 기록했다. Accuracy는 0.1474, Recall은 0.1136, Specificity는 0.1563, FAR은 0.1563의 차이로 베이스라인을 능가했다. 이러한 성능 향상은 스켈레톤 입력이 조명, 의상, 배경 등 외관 노이즈로부터 자유롭고 인체의 자세와 움직임 자체를 직접 표현하기 때문에 이상행동의 본질적 패턴을 학습하기에 유리하다는 점에서 기인한다. 또한 ST GCN은 인체 골격의 그래프 구조를 명시적으로 활용하여 관절 간 시공간 의존성을 효과적으로 학습하므로, 전역 외관 특징을 평면적으로 처리하는 I3D 기반 베이스라인 대비 행동 패턴의 표현력이 높다. 특히 Recall이 0.8002에서 0.9138로 향상되고 FAR이 0.3222에서 0.1659로 낮아진 점은, 제안 방법이 실제 이상 사건의 누락을 줄이는 동시에 정상 영상에 대한 오탐지 또한 효과적으로 억제하고 있음을 보여준다.

반면 AUC PR, Precision, F1 Score에서는 두 RGB 베이스라인이 모두 제안 방법보다 우세했으며, Sultani 베이스라인을 기준으로 각각 0.1939, 0.1438, 0.2298의 차이를 보였다. 이는 모델의 본질적 성능 열위라기보다는 데이터의 극심한 불균형 구조에서 기인하는 수학적 결과로 해석할 수 있다. 테스트 데이터는 정상 영상 150편과 이상 영상 140편으로 구성되며, 이상 영상에도 정상 구간이 상당 부분 포함되어 있어 프레임 수준 이상 구간 주석을 기준으로 집계하면 전체 테스트 프레임 가운데 실제 이상 프레임은 약 7.6%에 불과하다. 이처럼 정상 프레임이 전체의 90% 이상을 차지하는 극심한 클래스 불균형 환경에서는 모델이 FAR을 낮추더라도 절대적인 정상 프레임 수가 압도적으로 많기 때문에 오탐지 절대 수는 여전히 높게 유지된다. 제안 방법은 FAR을 0.3222에서 0.1659로 절반 가까이 낮추면서도 Recall을 0.8002에서 0.9138로 높여, 정상 영상의 오탐지는 줄이고 실제 이상 영상의 탐지율은 향상시키는 데 성공했다. 이는 실제 감시 환경에서 운영자의 알람 피로도(alarm fatigue)를 줄이면서 이상 사건 미탐지율을 동시에 낮추는 실용적 의미를 갖는다.

13개 이상행동 클래스 각각에 대한 AUC ROC 비교 결과는 표 3과 같다.

표 3. UCF Crime 테스트 셋에서의 클래스별 AUC ROC 비교

Class	Baseline	RTFM	Proposed
Abuse	0.7749	0.8510	**0.8797**
Arrest	0.7672	0.7504	**0.9038**
Arson	0.9196	0.9616	0.8743
Assault	0.9717	0.9908	0.8323
Burglary	0.8825	0.8625	**0.8956**
Explosion	0.8585	0.8672	**0.8715**
Fighting	0.8773	0.9535	0.703
RoadAccidents	0.8396	0.8257	**0.8721**
Robbery	0.8374	0.9309	**0.9332**
Shooting	0.8958	0.8629	0.8574
Shoplifting	0.8378	0.9077	0.6275
Stealing	0.9063	0.8520	**0.9218**
Vandalism	0.9430	0.9653	0.9098
Overall	0.7935	0.7932	**0.8368**

제안 방법은 13개 클래스 중 Abuse, Arrest, Burglary, Explosion, RoadAccidents, Robbery, Stealing을 포함한 7개 클래스에서 Sultani 베이스라인 이상의 성능을 보이며, 특히 Arrest에서 0.1366, Robbery에서 0.0958, Abuse에서 0.1048의 큰 폭으로 개선됐다. 이들 클래스는 인체 자세나 움직임 패턴 자체가 이상 신호로 작용하는 특성을 가지므로 스켈레톤 기반 접근의 강점이 잘 드러난다. 구체적으로 Arrest는 손을 머리 위로 들거나 무릎을 꿇는 등 일상적 자세에서 명확히 벗어나는 정형화된 동작이 강한 신호로 작동하며, Robbery와 Abuse 또한 가해자와 피해자 간 자세 비대칭이나 위협적인 신체 움직임이 두드러져 외관 정보 없이도 충분한 식별이 가능하다. 베이스라인의 경우 이러한 자세 단서를 RGB 픽셀 분포 안에서 간접적으로만 학습할 수 있어 동일한 신호의 활용에 구조적 한계가 있는 반면, 제안 방법은 관절 좌표를 직접 입력으로 사용하므로 자세 기반 이상 신호를 보다 명확하게 포착한다.

반면 Assault에서 0.1394, Fighting에서 0.1743, Shoplifting에서 0.2103의 차이로 베이스라인 대비 성능 하락이 나타났다. Assault의 경우 칼이나 무기와 같은 외관 단서가 RGB 특징으로 직접 포착되는 반면, 스켈레톤만으로는 이러한 보조 객체 정보를 직접적으로 활용하기 어렵다. Fighting은 두 사람 간의 물리적 상호작용이 핵심 신호지만, 제안 모델은 각 인체를 독립적으로 처리한 후 max pooling으로 집약하므로 인체 간 상호작용 신호를 직접 학습하기 어렵다. Shoplifting은 미세한 손 동작과 소지품 은닉이 핵심이지만, 17개 관절만으로는 손가락 수준의 정교한 동작을 표현하기 어렵다. 실제로 RGB 기반 RTFM은 이 세 클래스에서 각각 0.9908, 0.9535, 0.9077의 높은 AUC ROC를 기록하여, 외관 단서와 보조 객체 정보가 해당 이상행동의 판별에 결정적으로 작용함을 뒷받침한다. 이러한 한계는 스켈레톤 기반 접근의 구조적 특성에서 비롯되며, 향후 RGB 특징과의 융합 또는 관절 표현의 확장을 통해 개선할 수 있다.

손실 함수에 대해서는 식 (4)의 항을 단계적으로 분해하여 다음 네 가지 변형으로 비교했다. 은 랭킹 항만 포함하는 가장 단순한 구성이고, 는 에 smoothness 항을 추가한 구성이며, 는 에 이상 희소성 가중치 를 추가한 기존 MIL 손실 함수다. 제안 손실 함수는 L3의 이상 희소성 가중치를 수준인 로 축소하면서 정상 점수 억제 항을 보조 정규화로 추가한 최종 구성이다.

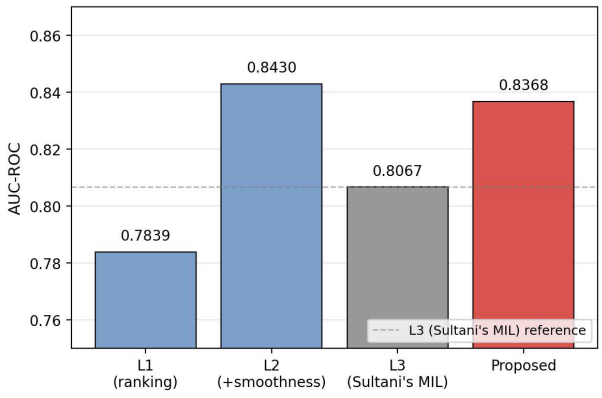

그림 2. 베이스라인 모델과 제안 모델의 AUC ROC 비교 결과

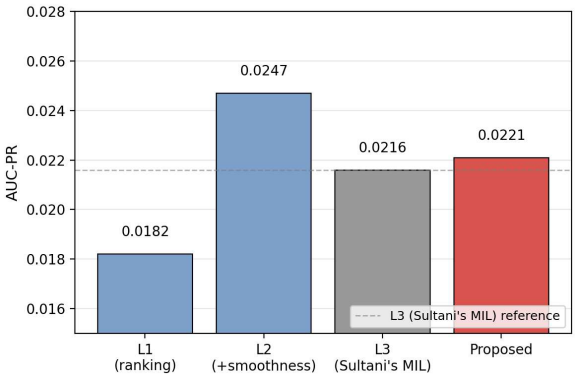

그림 3. 베이스라인 모델과 제안 모델의 AUC PR 비교 결과

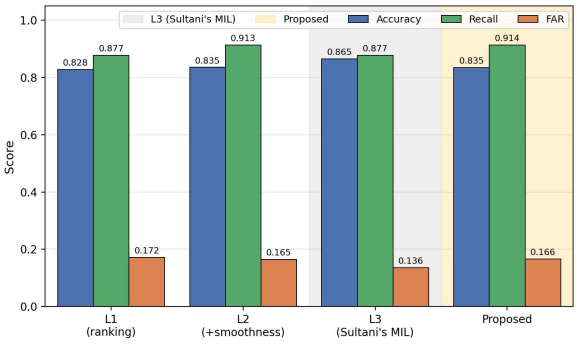

그림 4. 베이스라인 모델과 제안 모델의 Accuracy, Recall, FAR 성능 비교

그림 2는 AUC ROC, 그림 3은 AUC PR을 비교하며, 그림 4는 Accuracy, Recall, FAR의 세 지표를 비교한다. 그림 2와 3에서 L_3는 회색으로 표시하고, 본 논문의 제안 손실은 붉은색으로 표현했다. 제안하는 손실 재설계 효과는 다음과 같다.

첫째, smoothness 항이 시간적 일관성 학습에 결정적으로 기여한다. L_1과 L_2를 비교하면 AUC ROC가 0.7839에서 0.8430으로 0.0591만큼 향상됐으며, 이는 인접 스니펫 간 점수 일관성 학습이 시간적 노이즈를 효과적으로 억제함을 의미한다.

둘째, 베이스라인의 원 손실 L_3가 스켈레톤 입력 환경에서는 부적합함을 정량적으로 확인했다. L_2 대비 L_3는 이상 희소성 가중치 $8 \times 10^{(-5)}$만 추가한 것임에도 AUC ROC가 0.8430에서 0.8067로 0.0363만큼 하락했다. 이는 RGB 외관 특징과 달리 스켈레톤 입력은 이상 신호 자체가 약하게 표현되기 때문에, 베이스라인이 RGB 환경에서 채택한 강한 희소성 정규화가 오히려 이상 점수를 과도하게 억제하여 정상 및 이상 분리를 저해함을 시사한다.

셋째, 본 논문이 제안하는 손실 재설계의 효과를 입증했다. 동일한 데이터 및 구조 조건에서 기존 손실 L_3를 그대로 적용한 변형 대비 제안 손실 함수는 AUC ROC 0.8067에서 0.8368로 0.0301의 향상을 달성했다. 이는 이상 희소성 가중치를 1/8 수준으로 축소한 결정이 스켈레톤 기반 MIL 학습 환경에 적합함을 정량적으로 보여주며, 본 논문이 주장하는 손실 재설계의 핵심 근거다.

넷째, 정상 점수 억제 항 L_{normal}의 보조 정규화 역할은 제한적이다. L_{normal}을 배제한 L2 모델과 제안 모델의 AUC ROC 차이는 0.0062로 훈련 무작위성 범위 내에 있다.

기존 손실 L_3와 제안 손실 사이의 클래스별 효과 차이는 표 4와 같다.

표 4. 베이스라인 원 손실 L3와 제안 손실의 클래스별 AUC ROC 비교

Class	L3	Proposed
Abuse	0.9078	0.8797
Arrest	0.8561	**0.9038**
Arson	0.904	0.8743
Assault	0.8365	0.8323
Burglary	0.9166	0.8956
Explosion	0.8753	0.8715
Fighting	0.4515	**0.703**
RoadAccidents	0.8952	0.8721
Robbery	0.9071	**0.9332**
Shooting	0.8357	**0.8574**
Shoplifting	0.5800	**0.6275**
Stealing	0.8041	**0.9218**
Vandalism	0.9061	**0.9098**
Overall	0.8067	**0.8368**

표 4와 같이 제안 손실 재설계는 13개 이상행동 클래스 중 7개 클래스에서 기존 손실 대비 개선을

보이며, 특히 Fighting에서 AUC ROC가 0.2515, Stealing에서 0.1177의 큰 폭으로 개선이 두드러진다. 이들 클래스는 이상 구간이 영상 전체에 광범위하게 분포하거나 정상 행동과의 경계가 모호한 특성을 가지며, 베이스라인의 강한 이상 희소성이 정상 부분의 학습까지 억제하여 성능이 저하되는 대표적 사례다. 반면 Abuse, Arson 등 외관 단서가 강하고 이상 구간이 짧고 명확한 클래스에서는 두 손실 간 차이가 0.03 이내로 미세하다. 이러한 결과는 본 논문의 손실 재설계가 모든 클래스에 균일하게 작용하기보다는 스켈레톤 기반 신호가 약하거나 이상 구간이 광범위한 클래스에서 특히 효과를 발휘함을 의미한다.

표 5는 입력 채널, 시간 풀링 방식, 옵티마이저의 세 가지 구조적 설계 요소에 대한 실험 결과다.

표 5. 입력 채널, 시간 풀링, 옵티마이저에 따른 성능 비교 결과

Variant	AUC ROC	AUC PR	Accuracy	Recall	FAR
Input 2ch	0.8180	0.0195	0.8305	0.8944	0.1698
Pool mean	0.6590	0.0074	0.4983	0.8056	0.5032
Optim Adam	0.4890	0.0066	0.7800	0.4264	0.2184
Proposed	**0.8368**	**0.0221**	**0.8345**	**0.9138**	**0.1659**

세 요소 중 옵티마이저의 영향이 가장 크게 나타났다. 옵티마이저를 AdaGrad에서 Adam[20]으로 교체한 경우 AUC ROC가 0.4890으로 무작위 분류 수준으로 성능이 저조했다. 이는 MIL 랭킹 손실이 max 연산을 포함하여 매끄럽지 않은 손실 표면을 가지며, Adam의 적응적 모멘텀 추정이 이러한 환경에서 불안정한 갱신을 유발함을 시사한다.

시간 축 풀링 또한 결과에 큰 영향을 미쳤다. 풀링 방식을 max에서 mean으로 교체한 경우 AUC ROC가 0.6590으로 크게 하락하고 FAR이 0.5032로 증가했다. 이는 MIL의 기본 가정인 이상 점수가 가장 높은 스니펫이 bag의 대표 점수를 결정해야 한다는 원리를 온전히 반영하기 어렵기 때문이다. 평균을 취할 경우 다수의 정상 스니펫이 소수의 이상 스니펫 신호를 희석시켜 정상과 이상의 구분 능력이 상실된다.

한편 confidence 채널의 효과는 소폭의 기여로 확인됐다. 입력을 2채널로 축소한 경우 AUC ROC가 0.0188 하락했으며 미검출, 오검출 관절 슬롯이 다수 존재하는 환경에서 confidence가 신뢰도 보조 신호로 작동함을 의미한다.

V. 결 론

본 논문에서는 감시 영상에서 영상 수준의 약한 라벨만을 사용하여 이상행동을 탐지하기 위한 스켈레톤 기반 약지도 학습 프레임워크를 제안했다. 제안하는 방법은 YOLOv11 Pose를 이용한 다중 인체 스켈레톤 추출, ST GCN을 이용한 시공간 특징 학습, MIL 랭킹 손실에 정상 영상 점수 억제 항을 추가한 개선된 손실 함수의 세 가지 구성 요소로 이루어진다. UCF Crime 데이터셋에서의 실험 결과, 제안 방법은 AUC ROC 0.8368을 달성하여 베이스라인 대비 0.0433의 향상을 보였으며, Accuracy, Recall, Specificity, FAR 등 정상, 이상 분류 능력을 측정하는 지표 전반에서 베이스라인을 능가했다.

향후 연구에서는 스켈레톤과 RGB 외관 특징을 융합한 멀티모달 접근으로 외관 정보가 중요한 클래스의 성능을 보완하고, 인체 간 상호작용을 명시적으로 모델링하여 Fighting과 같은 상호작용 기반 이상행동의 탐지 성능을 개선하며, 손 및 얼굴 keypoint를 포함한 확장 모델로 미세 동작 표현 능력을 강화할 계획이다.

REFERENCES

[1] *공공기관 CCTV 설치 및 운영대수*(2025), https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=2855 (accessed May., 29, 2026).
[2] 곳곳에 CCTV 늘어서 안심? 477대 지켜보는 눈은 단 '한 명'(2025), https://www.mt.co.kr/society/2025/03/10/2025030711595172247 (accessed May., 29, 2026).

[3] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 6479-6488, Salt Lake City, USA, June 2018.

[4] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 6299-6308, Honolulu, USA, July 2017.

[5] Ultralytics YOLO11 (2024), https://github.com/ultralytics/ultralytics (accessed May., 20, 2026).

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *Proc. AAAI Conf. Artificial Intelligence*, vol. 32, no. 1, pp. 7444-7452, New Orleans, USA, Feb. 2018.

[7] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning Temporal Regularity in Video Sequences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 733-742, Las Vegas, USA, June 2016.

[8] W. Liu, W. Luo, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection - A New Baseline," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 6536-6545, Salt Lake City, USA, June 2018.

[9] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp. 4975-4986, Virtual, Oct. 2021.

[10] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph Convolutional Label Noise Cleaner: Train a Plug-and-Play Action Classifier for Anomaly Detection," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 1237-1246, Long Beach, USA, June 2019.

[11] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 12026-12035, Long Beach, USA, June 2019.

[12] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition," *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp. 13359-13368, Virtual, Oct. 2021.

[13] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 143-152, Virtual, June 2020.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using PartAffinity Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, Jan. 2021.

[15] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, et al., "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 7157-7173, June 2023.

[16] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 5693-5703, Long Beach, USA, June 2019.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Computer Vision*, pp. 2961-2969, Venice, Italy, Oct. 2017.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., "Microsoft COCO: Common Objects in Context," *Proc. European Conf. Computer Vision*, pp. 740-755, Zurich, Switzerland, Sept. 2014.

[19] J. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, vol. 12, no. 7, pp. 2121-2159, July 2011.

[20] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proc. Int. Conf. Learning Representations*, San Diego, USA, May 2015.

저 자 소 개

변무혁(정회원)
2026년 국립순천대학교 컴퓨터공학과 졸업(공학사)
2026년～현재 국립순천대학교 멀티미디어공학과 석사과정
<주관심분야 : 컴퓨터비전, 딥러닝, 객체탐지, 행동분석>

양현성(정회원)
2022년 국립순천대학교 멀티미디어공학과 졸업(공학사)
2026년～현재 국립순천대학교 스마트융합학부 멀티미디어 공학 전공 석·박사통합과정 수료
<주관심분야 : 컴퓨터 비전, 딥러닝, 객체 추적>

정세훈(정회원)
2012년 국립순천대학교 멀티미디어공학과 졸업(공학석사)
2017년 국립순천대학교 멀티미디어공학과 졸업(공학박사)
2018년 영산대학교 빅데이터융합전공 조교수
2020년 국립안동대학교 창의융합학부 조교수
2022년 국립순천대학교 컴퓨터공학과 조교수
2024년～현재 국립순천대학교 컴퓨터공학과 부교수
<주관심분야 : 블록체인, 딥러닝, 생성모델, 빅데이터 분석 및 예측>

심춘보(정회원)
1996년 전북대학교 컴퓨터공학과 졸업(공학사)
1998년 전북대학교 컴퓨터공학과 졸업(공학석사)
2003년 전북대학교 컴퓨터공학과 졸업(공학박사)
2005년～현재 국립순천대학교 인공지능공학부 교수
<관심분야 : 빅데이터, 블록체인, 딥러닝, 생성모델, 자연어 처리, 강화학습>