AIFFEL

모두의연구소

절도 모니터링 도우미

실시간 CCTV 영상 분석을 통한 지능형 동작 인식 시스템

RealCvongE

윤여원, 김찬중, 이재영, 전휘호

INDEX

01 프로젝트 팀 구성 및 역할

02 프로젝트 수행 절차

03 프로젝트 배경

04 프로젝트 발전 과정

05 프로젝트 결과

06 프로젝트 한계점 , Future Work

07 Appendix

08 Reference

01 프로젝트 팀 구성 및 역할

훈련생	역할
윤여원 (팀 리더)	▪ 추론 서비스, 모바일 플랫폼, 벡엔드 서비스 설계 및 구현, PoseC3D 전처리
김찬중 (팀원)	▪ Pickle 데이터 전처리(균등화, frame 보간, combine), 영상 보간 및 feature 추출 ▪ 모델 학습 및 평가, 학습 결과 시각화
이재영 (팀원)	▪ BN - wvad 동작 구현 및 성능 향상 ▪ pickle file 내 사람 좌표 visualize
전휘호 (팀원)	▪ HR-Pro ▪ 데이터 포맷 변경

02 프로젝트 수행 절차

[일정 및 계획]

구분	기간	활동	비고
사전 기획	• 2/26(월) ~ 3/2(목)	• 프로젝트 기획 및 주제 선정 • 기획안 작성	• 아이디어 선정
데이터 수집	• 3/4(월) ~ 3/8(금)	• 데이터 수집	•
모델 탐색	• 3/1(월) ~ 3/15(금)	• 논문 리뷰 • 오픈소스 코드 리뷰	• 모델 선정 및 분석
데이터 전처리	• 3/16(월) ~ 4/5(금)	• 샘플 전처리 데이터 탐색 • I3D 피쳐 추출, 라벨 데이터 만들기	•
모델 학습 및 실험	• 3/25(월) ~ 4/9(화)	• PoseC3D, HR-Pro, BN-WVAD 모델 학습 • 학습된 모델 테스트 및 실험	•
서비스 구축	• 3/16(월) ~ 4/12(금)	• 플러터 활용 모바일 플랫폼 설계 및 구현 • Firebase, FAST API 활용 벡엔드 설계 및 구현 • 모델들을 통합하여 추론 서비스 구현	• 최적화, 오류 수정, 추론속도 실험
총 개발기간	• 2/26(월) ~ 4/12(금)(총 7주)	•	•

03 프로젝트 배경

[프로젝트 주제 및 선정 배경]

- **프로젝트 주제 : 실시간 CCTV 영상 분석을 통한 지능형 동작 인식 시스템 개발**

- **프로젝트 선정 배경**
  - 도시화 및 공공장소의 안전 관리에 대한 수요 증가
  - 기존 CCTV 시스템의 한계점을 극복하고, 보다 효율적인 동작 인식 기술 필요성 인식
  - 딥러닝 및 영상 처리 기술의 발전을 통한 새로운 솔루션 제공 가능성

[프로젝트 목적]

- **실시간 동작 인식을 통한 안전 사고 및 범죄 예방**

- **AI 기술을 활용하여 인간 감시원의 부담 경감**

- **비용 효율적인 보안 시스템 구현**

03 프로젝트 배경

[프로젝트 개요]

- **컨셉 : 실시간 영상 분석과 딥러닝을 결합한 행동 인식**
- **훈련 내용과의 관련성 : 영상 데이터셋을 활용한 딥러닝 모델의 지속적인 훈련 및 개선**
- **개발 환경**
  - Python, PyTorch, OpenCV 등을 활용한 개발 환경 구축
  - 고성능 컴퓨팅 환경과 클라우드 리소스 활용

[프로젝트 구조]

- **데이터 수집 : 매장내 다양한 각도의 CCTV 카메라에서 수집된 다양한 행동 영상 데이터**
- **데이터 처리 : 영상에서 동작을 인식하고 분류하는 딥러닝 모델**
- **결과 출력 : 인식된 동작에 대한 실시간 알림 및 로깅**

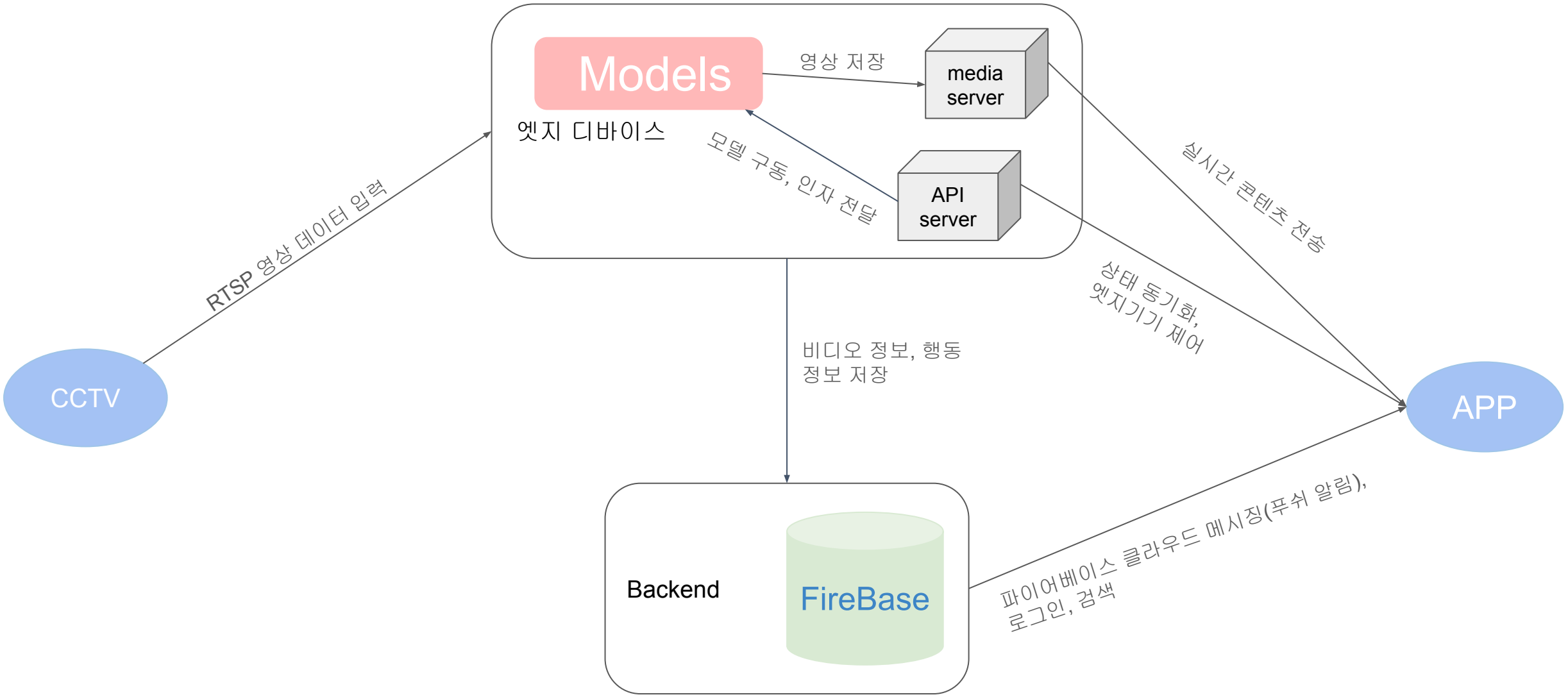
03 프로젝트 배경

[기대 효과]

- 실시간 대응을 통한 사건 사고의 신속한 처리
- 지속적인 학습과 데이터 업데이트를 통한 인식률의 지속적 향상
- 인공지능을 통한 장기적인 보안 관리 시스템의 혁신
- 사회 안전성 향상과 범죄 예방에 기여
- 운영 비용 절감과 인력 자원의 효율적 재배치
- 실시간 대응 시스템을 통한 공공 서비스의 질적 수준 제고

04 프로젝트 발전과정

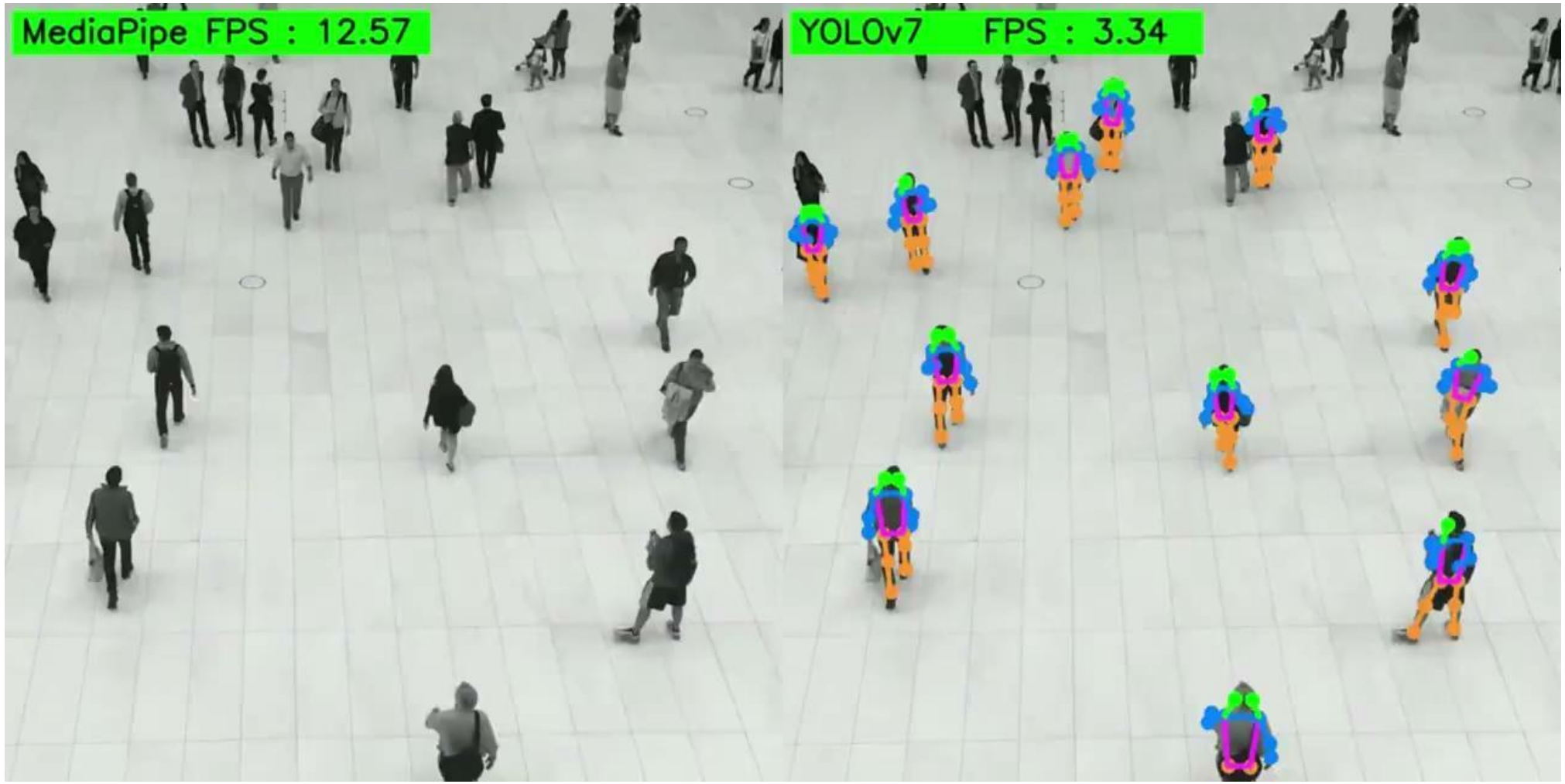
시스템 아키텍쳐 구상도



서비스 구축 Task

04 프로젝트 발전 과정

YOLO vs Mediapipe



04 프로젝트 발전 과정

PoseC3D vs ST-GCN

Table 1: **Differences between PoseConv3D and GCN.**

	Previous Work	PoseConv3D
Input	2D / 3D Skeleton	2D Skeleton
Format	Coordinates	3D Heatmap Volumes
Architecture	GCN	3D-CNN

GCN based methods 의 한계

1. Robustness

- 좌표의 분포변화에 크게 영향을 받아 좌표의 소폭 변화에도 결과가 크게 변할 수 있다

2. Interoperability

- 기존 행동 인식은 RGB, 광학 흐름, 스켈레톤 등을 효과적으로 조합하여 성능을 향상
- 스켈레톤의 그래픽 형태는 조합 자체가 어려워, 이 방법을 사용하는 데 한계가 있음

3. Scalability

- GCN은 모든 인간 관절을 노드로 다루기 때문에, GCN의 복잡성은 사람 수에 선형적으로 증가
- 여러 사람이 포함된 그룹 활동 인식과 같은 경우, 제약이 발생할 수 있음

04 프로젝트 발전 과정

mmaction2 & HR - pro & BN - wvad

1. mmaction2

- yoloV8과 결합하여 webcam을 통해 skeleton based action recognition 진행

2. HR-pro

- snippet level 과 instance level에서 Temperal Action Localization을 진행 confidence score 추출

3. BN-WVAD

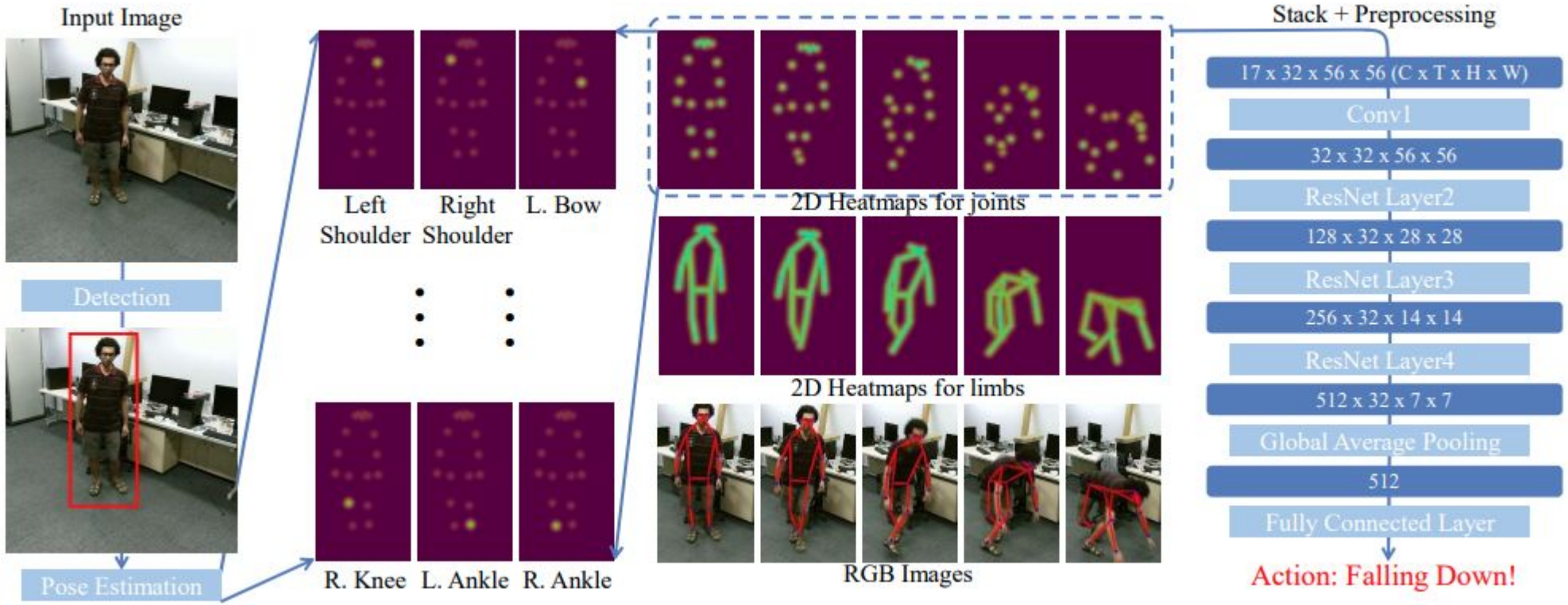
- Weakly Supervised Video Anomaly Detection 을 진행하여 anomaly score 추출

05 프로젝트 수행 결과

모델 개요

PoseC3D(Skeleton-based Action Recognition**)**

3D-CNN을 활용한 뼈대 기반 행동 인식 모델. 추론된 포즈 시퀀스를 넣으면 행동 확률이 나온다.



05 프로젝트 수행 결과

결과 제시 1. 탐색적 분석 및 전처리

데이터 구성

- 구매행동: 시험, 구매, 반품, 비교, 선택 ,매장이동으로 구성된 6개의 행동 Untrimmed 영상과 라벨 (926,15 GB)
- 절도 행동: Untrimmed 영상과 라벨 (253.17GB)

정제 전 1분 영상

정제 후 26초로 축약된 영상

Skeleton based Action Recognition에서는 일반적으로 행동의 시작과 끝이 trimmed된 데이터를 사용하여 추론

- **인식 정확도 향상**
- **기존 데이터셋과의 호환성**
- **행동의 명확한 정의**

데이터 출처 : AIHUB 의 실내 구매행동, 실내 이상행동 데이터 중 절도

05 프로젝트 수행 결과

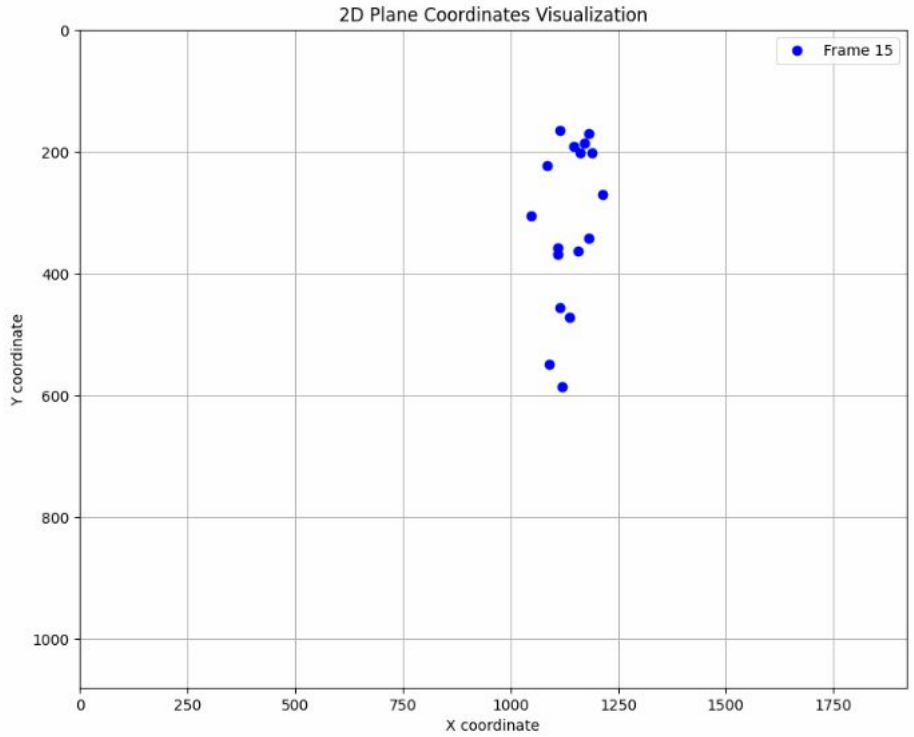
결과 제시 1. 탐색적 분석 및 전처리

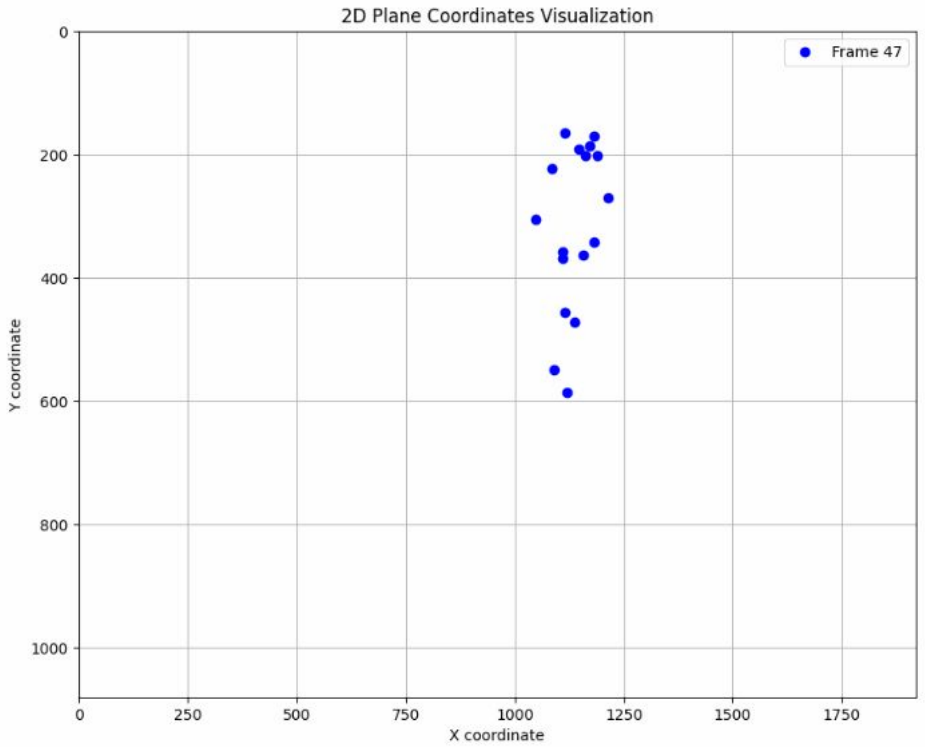
1.Faster-RCNN(human Detector) , HRNET(pose estimator) 를 이용

- 영상데이터에서 훈련에 필요한 skeleton 데이터를 추출
- (배치 크기, 프레임 수, 키포인트 수, 좌표) 차원을 가진 NumPy 배열

2. 데이터 보완

- 절도와 매장이동 영상은 3FPS
- 나머지 구매행동은 10FPS





05 프로젝트 수행 결과

결과 제시 1. 탐색적 분석 및 전처리

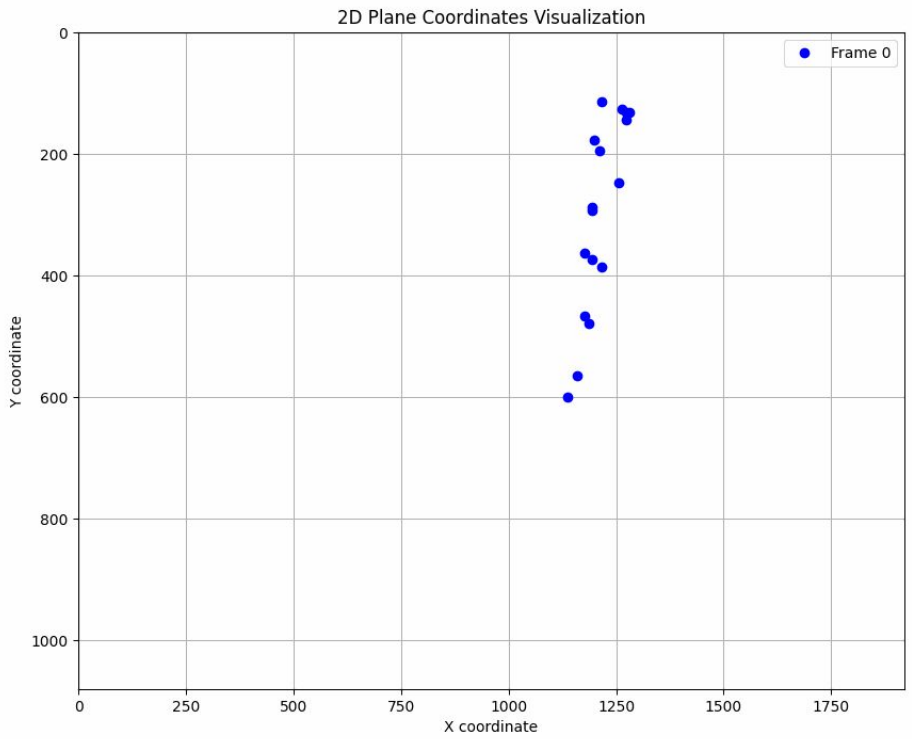
1.Faster-RCNN(human Detector) , HRNET(pose estimator) 를 이용

- 영상데이터에서 훈련에 필요한 skeleton 데이터를 추출
- (배치 크기, 프레임 수, 키포인트 수, 좌표) 차원을 가진 NumPy 배열

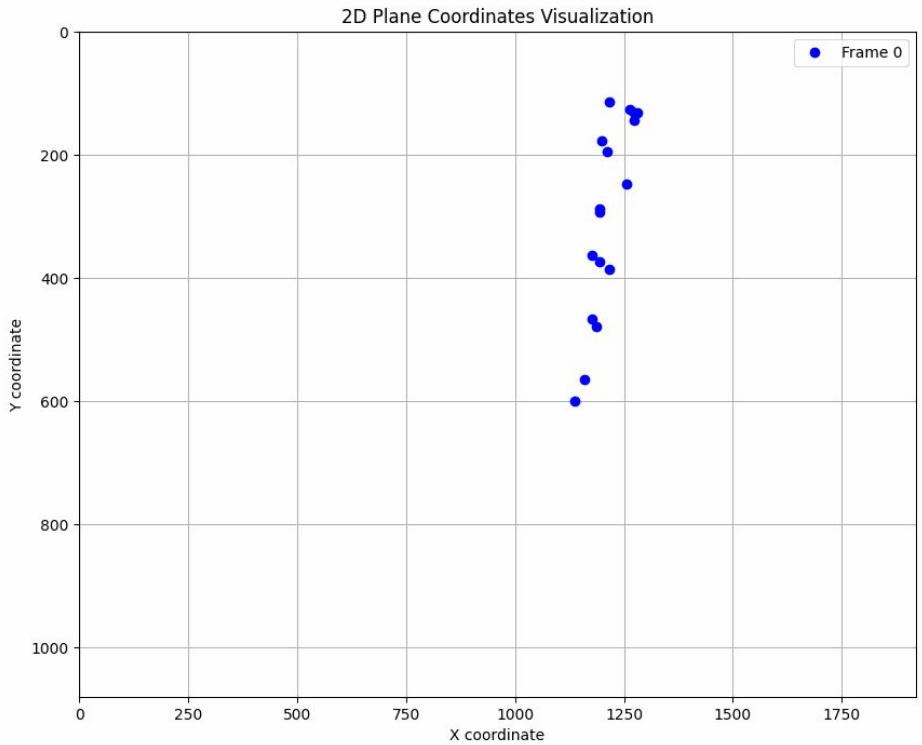
2. 데이터 보완

- 백터 선형보간을 통해 기존 데이터를 10FPS로 보완

3 FPS 영상



10 FPS 보완 영상



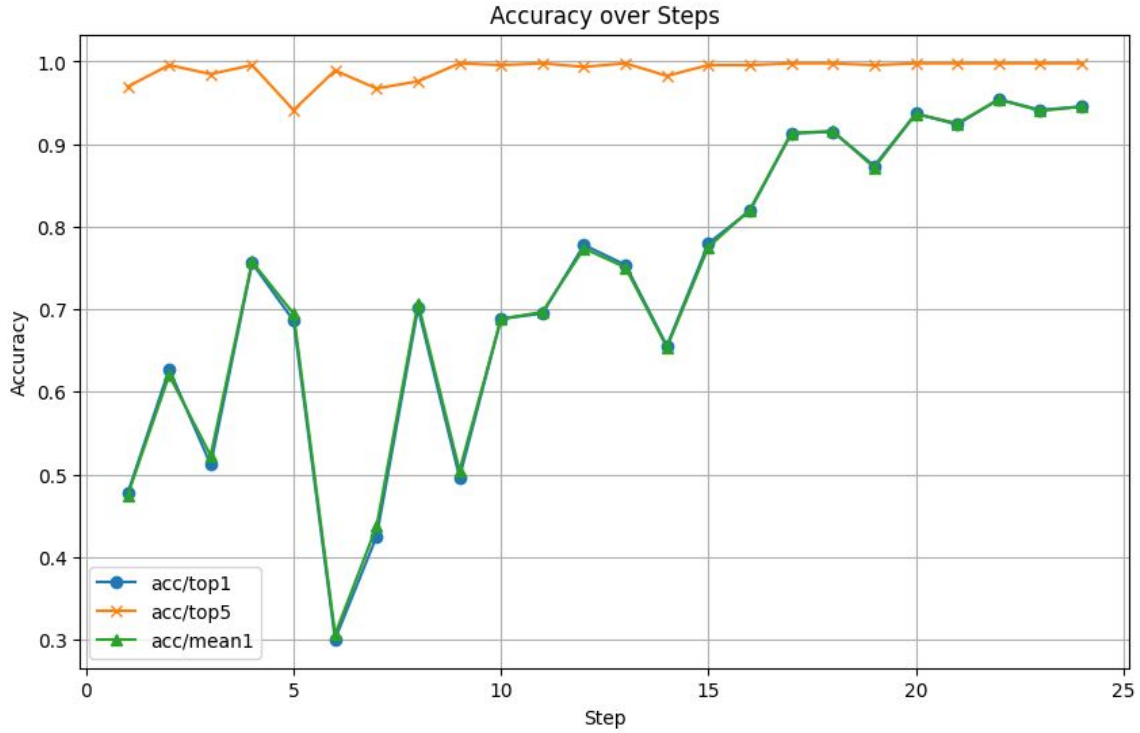
05 프로젝트 수행 결과

결과 제시 2. 모델 훈련

PoseC3D 데이터 증강

1. RandomResizedCrop: 56% ~100% 범위로 무작위 crop
2. Flip: 50%로 좌우 반전

결과 제시 2. 모델 결과

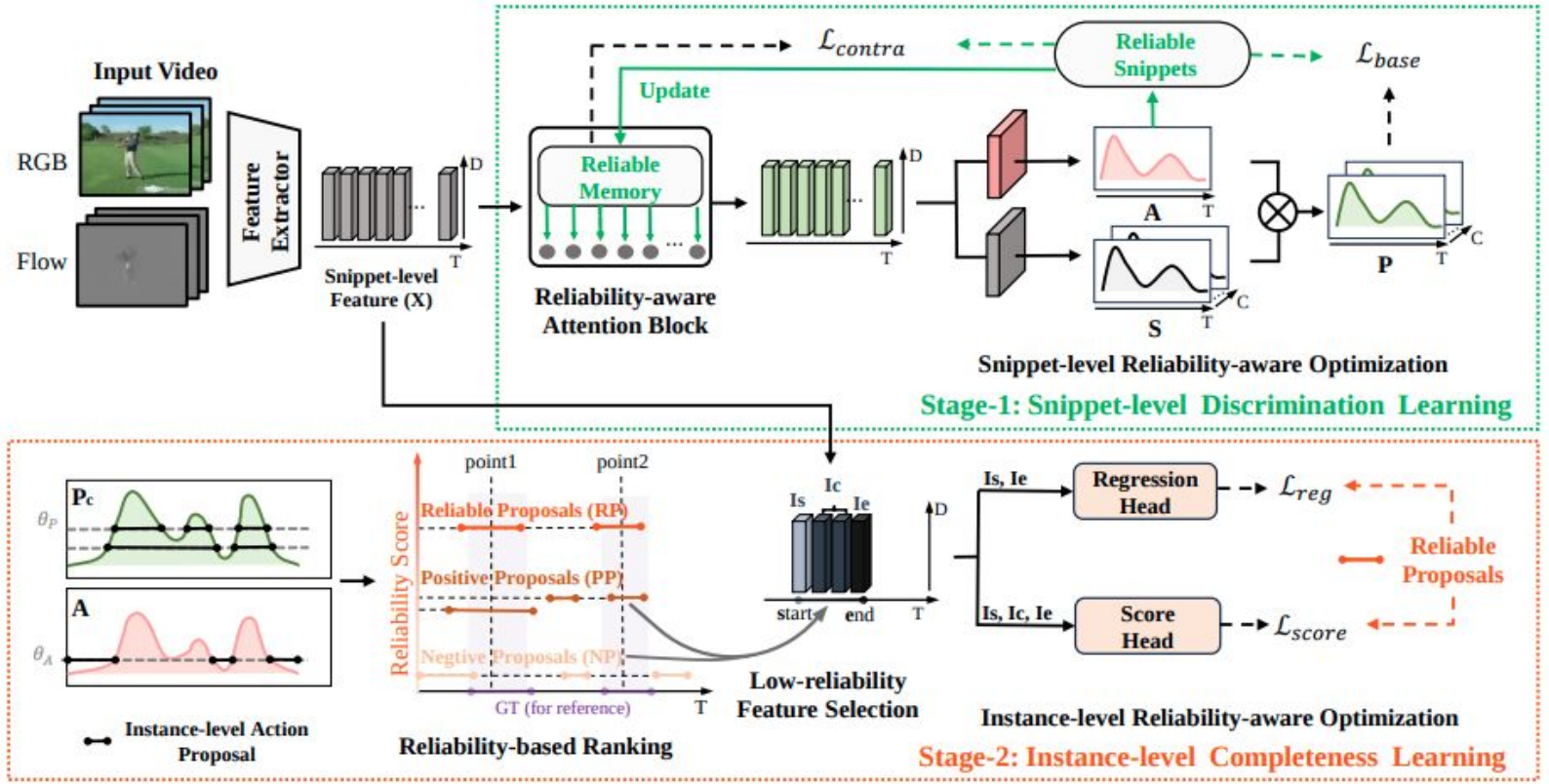


05 프로젝트 수행 결과

결과 제시 2. 모델 개요

HR-Pro(Point-supervised Temporal Action Localization**)**

행동 포인트를 학습해서 행동 구간을 예측하는 모델. 영상을 넣으면 (라벨, 신뢰도 점수, 구간)을 담은 리스트를 준다.



05 프로젝트 수행 결과

결과 제시 1. 탐색적 분석 및 전처리

영상: Untrimmed 영상 데이터 → RGB I3D(프레임수/16 , 1024(비디오 특징)) 차원을 가진 NumPy 배열과 Optical Flow I3D(프레임수/16 , 1024(비디오 특징)) 차원을 가진 NumPy 배열을 합친 배열

라벨: Start Frame ,End Frame → Point = (Start Frame + End Frame) / 2

결과 제시 2. 훈련 결과

```
-------------------------------------------------------------------------------
|t-IoU |0.100||0.200||0.300||0.400||0.500||0.600||0.700|
-------------------------------------------------------------------------------
|mAP   |0.714||0.491||0.447||0.430||0.386||0.231||0.067|
-------------------------------------------------------------------------------
|Average-mAP: 0.3950 Average mAP[0.1:0.5]:0.4934 Average mAP[0.3:0.7]:0.3120
-------------------------------------------------------------------------------
```

05 프로젝트 수행 결과

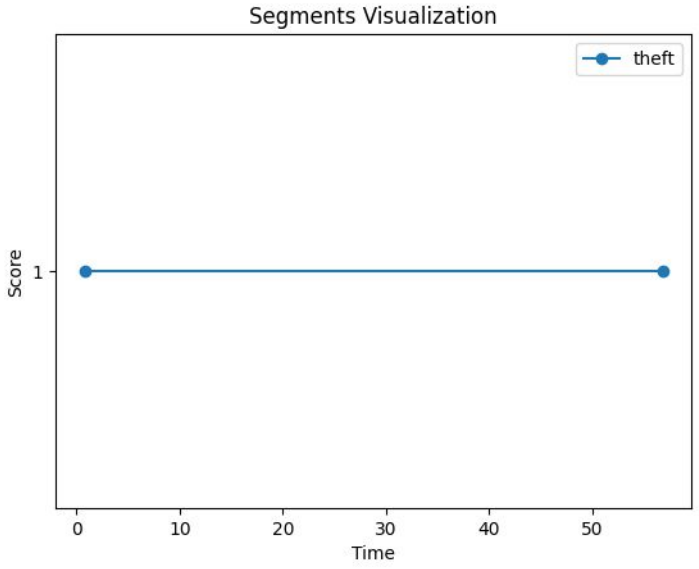
결과 제시 2. 결과 시각화

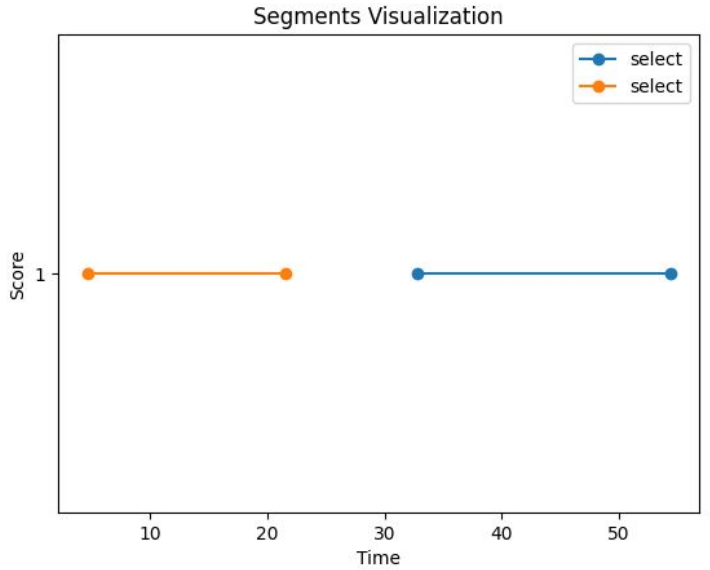
- 사람이 움직인 구간을 행동 구간으로 예측

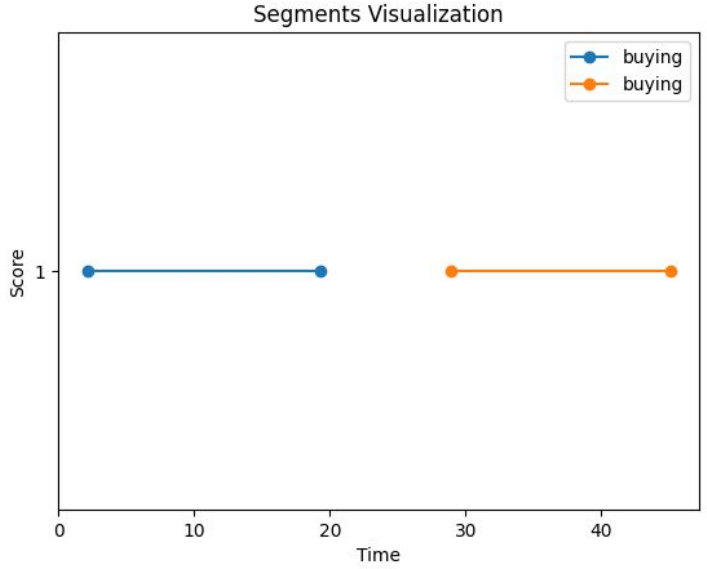
라벨 예측 성공(절도)

라벨 예측 실패(시험 -> 선택)

라벨 예측 성공(구매)





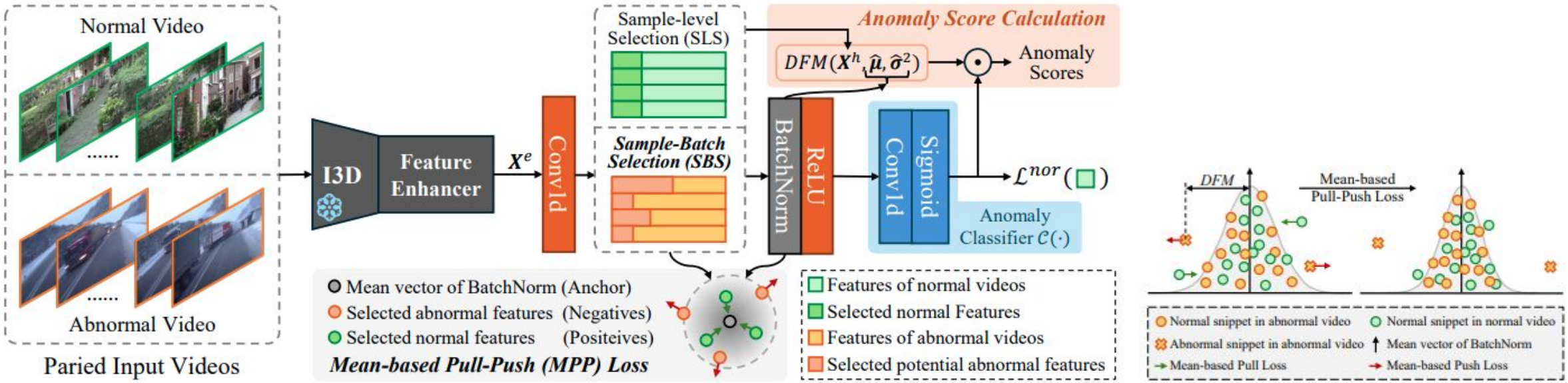


05 프로젝트 수행 결과

결과 제시 1. 모델 개요

BN-WVAD(Weakly Supervised Video Anomaly Detection**)**

비정상, 정상 영상을 구분하는 모델. 영상을 넣으면 프레임 수준의 이상점수 준다.



05 프로젝트 수행 결과

결과 제시 2. 모델 결과

영상: Untrimmed 영상 데이터 → RGB I3D(프레임수/16 , 1024(비디오 특징)) 차원의 Numpy 배열

abnormal : normal	AUC	AP
641 : 4959	99.46	55.41
641 : 659	99.55	69.52
1282 : 1298	99.53	65.43

영상 시퀀스 데이터 종류	이상 점수 예측값
normal	1 ~ 2
abnormal	10 ~ 17

05 프로젝트 수행 결과

결과 제시 1. 전체 추론 과정

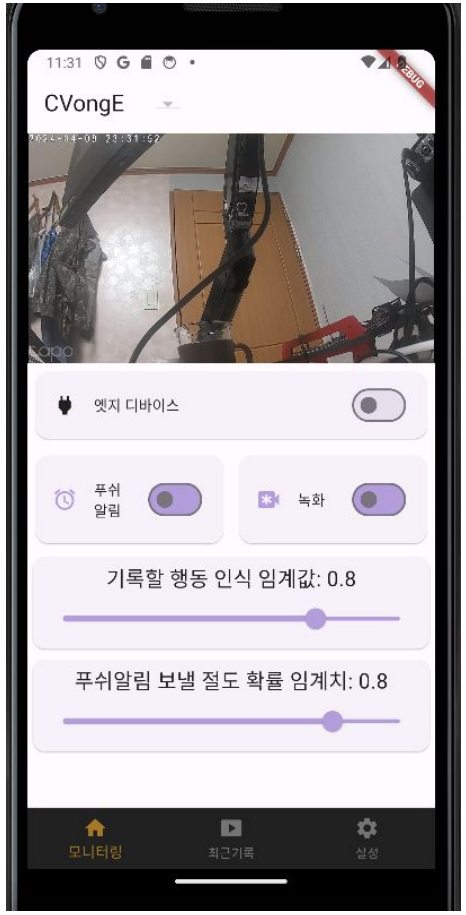
영상데이터로 부터 행동 확률을 구하는 전체 프로세스

05 프로젝트 수행 결과

결과 제시 2. 화면구성

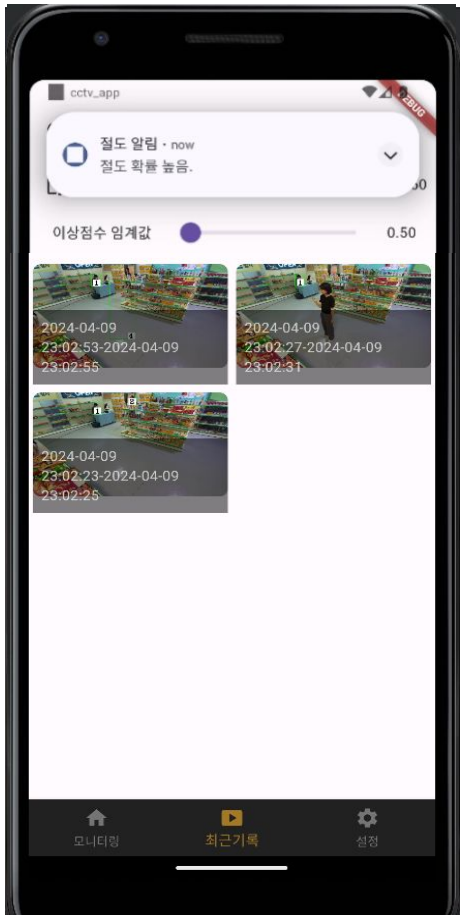
CCTV 등록

기기 제어,모니터링

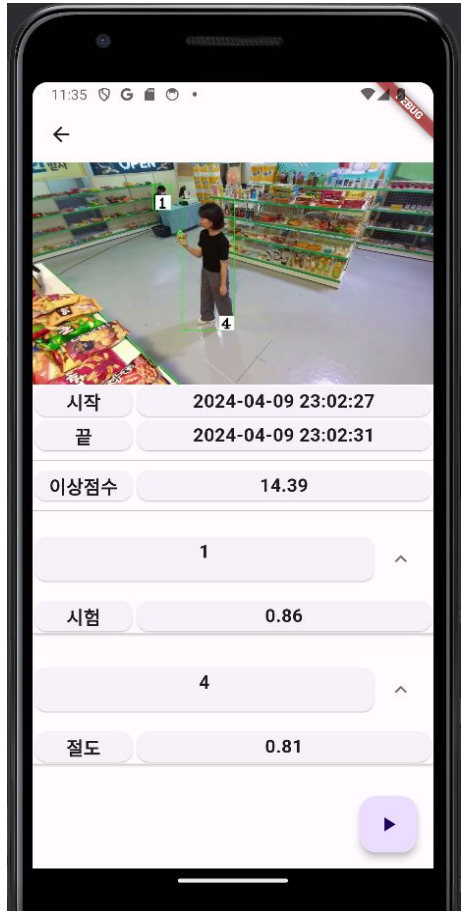


녹화 목록 검색

절도 푸쉬 알림



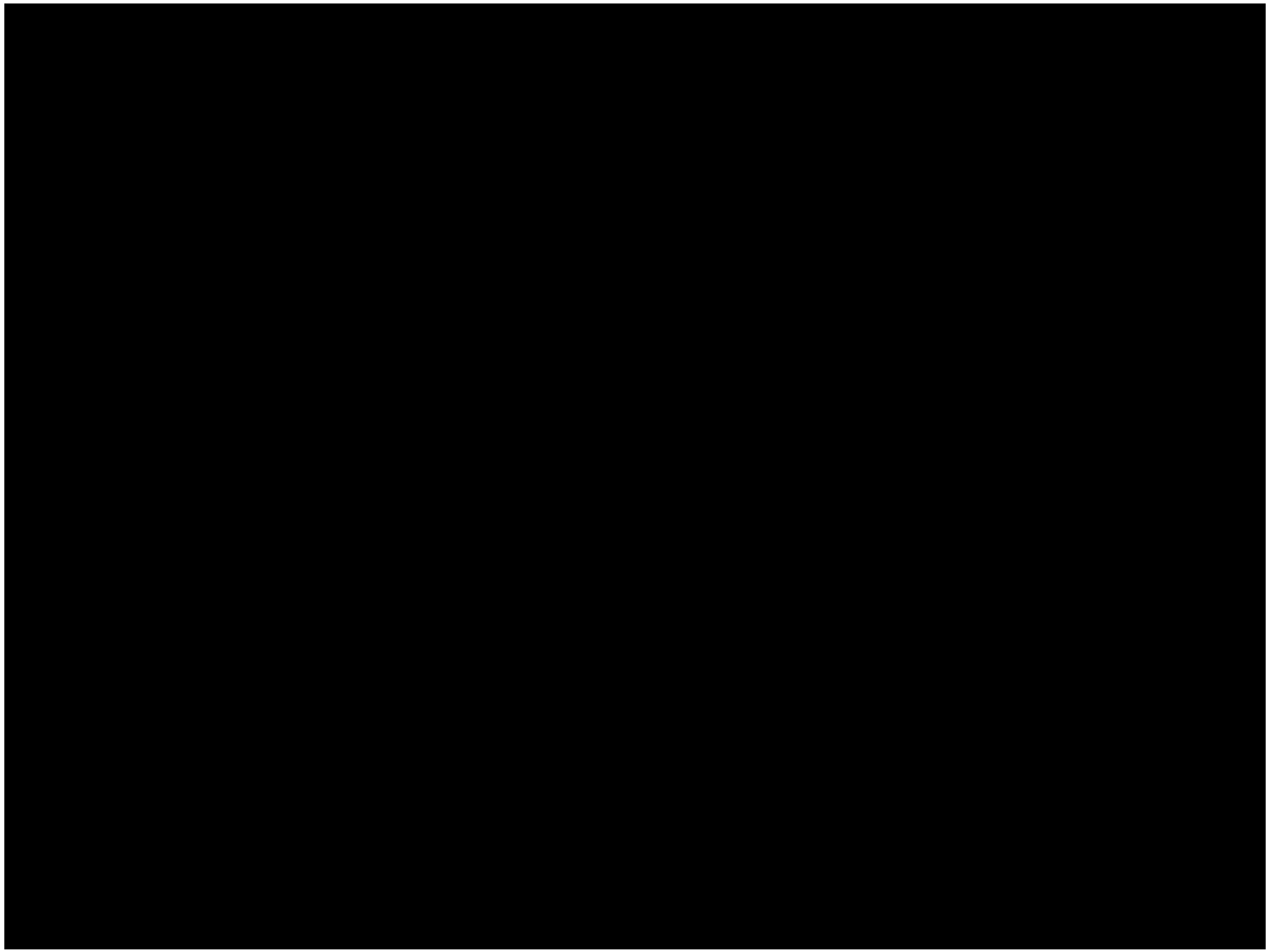
행동 인식, 이상점수 로그



05 프로젝트 수행 결과

결과 제시 3. 시연영상

이상 점수, 행동인식 상세 페이지



06 프로젝트 한계점 , Future Work

[프로젝트 한계점]

- 추론 딜레이가 있다.
- 강건하지 않다.
- 사용자 친화적이지 않다.
- 범용성을 위해선 온디바이스에 적용해야겠지만 온디바이스에 적용하기엔 무겁다.

[Future Work]

- 두 모델의 출력 값을 **Score**로 환산하여, **Task**에 더 적합한 모델에 더 많은 가중치를 두어 합산하는 방법.
- 모델을 하나로 통합하여 시간 구간까지 예측하는 뼈대 기반 행동 지역화 모델을 만드는 방법.
- 모델을 경량화하여 온디바이스에 적용할 수 있는 방법.

Q&A

AIFFEL 모두의연구소

07 Appendix

A. PoseC3D Trimmed Data로 만드는 이유

1. **인식 정확도 향상** : 행동의 핵심 부분에 집중할 수 있어 불필요한 프레임이나 노이즈가 제거되므로 인식 정확도가 향상됩니다.
2. **기존 데이터셋과의 호환성** : 대부분의 공개 행동 인식 데이터셋은 trimmed 데이터 형태로 제공됩니다. 따라서 trimmed 데이터를 사용하면 기존 데이터셋 및 벤치마크와 직접 비교가 가능해집니다.
3. **행동의 명확한 정의** : 시작과 끝점이 명확히 정의된 trimmed 데이터를 사용하면 인식하고자 하는 행동을 명확히 규정할 수 있습니다.

B. PoseC3D 데이터 증강을 위해 RandomResizedCrop과 Flip만을 사용한 이유

1. **증강 기법의 효과성** : RandomResizedCrop과 Flip은 간단하면서도 효과적인 증강 기법으로 알려져 있습니다. 이 두 가지 기법만으로도 모델의 일반화 능력을 상당히 향상시킬 수 있기 때문에, 다른 증강 기법을 추가하지 않고도 충분한 성능 향상을 기대할 수 있습니다.

2.**계산 효율성** : 증강 기법을 추가할수록 데이터 전처리에 필요한 계산량이 증가하게 됩니다. RandomResizedCrop과 Flip은 상대적으로 계산 부담이 적은 편이므로, 이 두 가지 기법만을 사용함으로써 계산 효율성을 높일 수 있습니다.

3. **데이터의 특성** : PoseC3D는 인체 동작 인식을 위한 데이터셋으로, 인체의 크기나 위치, 방향 등의 변화에 강건해야 합니다. RandomResizedCrop과 Flip은 이러한 변화를 효과적으로 다룰 수 있는 증강 기법이므로, 이 데이터셋의 특성에 잘 부합합니다.

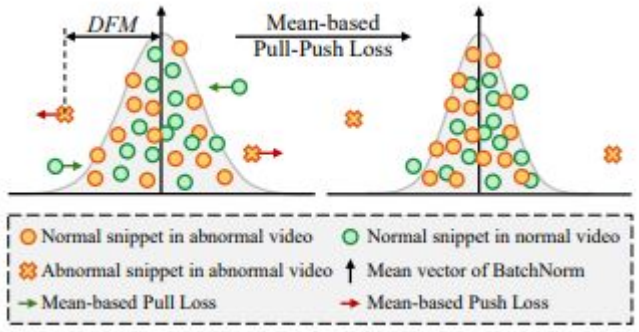
4. **증강 기법 간의 중복 피하기** : 어떤 증강 기법들은 서로 유사한 효과를 가질 수 있습니다. 예를 들어, RandomResizedCrop은 크기 변화와 위치 변화를 모두 다루므로, 별도의 위치 변화 증강 기법을 추가하는 것은 중복될 수 있습니다. 중복을 피함으로써 증강의 다양성을 유지하면서도 계산 부담을 줄일 수 있습니다.

따라서, PoseC3D 데이터 증강에 RandomResizedCrop과 Flip만을 사용한 것은 효과성, 효율성, 데이터 특성, 중복 방지 등을 고려한 선택이라고 볼 수 있습니다. 이 두 가지 증강 기법으로도 충분한 성능 향상을 얻을 수 있을 것으로 기대되며, 필요에 따라 다른 증강 기법을 추가하는 것도 고려해 볼 수 있습니다.

07 Appendix

C. Anomaly Score 정상 데이터는 1~2, 비정상 데이터는 10~20 이라는 결과가 나오는 이유

BN-WVAD에서 이상점수는 DFM(Divergence of Feature from Mean vector) 으로 각 데이터 포인트의 특징 벡터와 평균 벡터 사이의 차이(발산, Divergence)를 계산합니다. 이 차이를 DFM이라고 합니다. DFM이 크다는 것은 해당 데이터 포인트가 정상 데이터의 평균적인 패턴에서 벗어나 있다는 것을 의미합니다. 즉, 이상치일 가능성이 높습니다.



D. I3D(Inflated 3D ConvNet) 란

- 비디오 데이터에서 시공간 정보를 추출하기 위해 사용되는 심층 학습 기반 특징 표현입니다.
- 인플레이션(Inflation) 기법 사용: 사전 학습된 2D CNN 모델(예: ImageNet에서 학습한 Inception-v1)의 필터를 '인플레이션'하여 3D 커널로 확장합니다. 이를 통해 대규모 이미지 데이터셋에서 학습한 특징을 비디오 도메인으로 전이할 수 있습니다.
- '인플레이션(Inflation)'은 i3d 아키텍처에서 사용되는 핵심 기법 중 하나로, 2D 컨볼루션 필터를 3D 컨볼루션 필터로 확장하는 과정을 의미합니다.

08 Reference

[1] Yixuan Zhou(2023). “BatchNorm-based Weakly Supervised Video Anomaly Detection”. https://arxiv.org/pdf/2311.15367.pdf

[2] Huaxin Zhang(2024). “HR-Pro: Point-supervised Temporal Action Localization via Hierarchical Reliability Propagation”. https://arxiv.org/pdf/2308.12608.pdf

[3] Haodong Duan(2022). “Revisiting Skeleton-based Action Recognition”. https://arxiv.org/pdf/2104.13586.pdf

[4] https://github.com/orgs/RealCVongE/repositories