AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, etc…

2023.01.10 논문 리뷰

배성훈

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Research Background**:
  - **Action Recognition**:
    - Video에서 수행되는 Action을 **감지**하고 **분류**하는 과정
    - Human Action의 복잡성, Appearance와 motion의 변화, Background와 Lighting condition의 변화에 대한 challenging problem을 가짐

  - **Traditional Dataset**:
    - KTH, Weizmann, HMDB51, Hollywood-2, UCF101
      (짧은 clip들로 구성, 하나의 action만 포착, 수동으로 trimming)
      *__Trimming__: 확대 시 화질 저하되지 않는 해상도 높은 사진을 이용하는 기법, 자신이 원하는 부분을 잘라내고, 그 부분 확대
    - 기존의 데이터셋은 인간 Action의 다양성과 복잡성을 적절하게 다루지 못함
    - 또한 **Coarse level에만 annotation**을 달기 때문에 fine-grained action recognition을 연구하기 어려움

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Research Background**:
  - **Action Vocabulary**:
    - 80개의 서로 다른 atomic visual action으로 구성
    - 기존 데이터셋의 경우 제한된 수의 복합적인 action들로 구성되어 action이 dense하게 구성되지 못함

  - **Spatio-temporal localization**:
    - 최근의 접근법은 **2-stream variant** (RGB와 Optical flow data를 별도로 처리하는 방법)
    - Frame level에서 Action classes를 구별하도록 학습된 Object Detector에 의존
    - Multi-Frame 접근법
      - Tubelets: Multi-Frame에서 localization과 classification 공동 추정
      - T-CNN: 3D convolution 사용해 Short tubes 추정
        Tubes: Atomic visual action을 가지는 frame에 annotation된 영역
      - Micro-tubes: 2개의 연속적인 frame에 의존
    - 저자는 spatio-temporal tubes의 아이디어를 기반으로 하지만, SOTA인 I3D convolution과 Faster R-CNN의 Region proposal를 사용
      ***Spatio-temporal tubes**: atomic visual action이 수행된 위치와 특정 시간 간격에 따라 Frame에 annotation된 영역

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Research Background**:
  - **Spatio-temporal localized atomic visual actions**:
    - 1. 영상, 비디오 등 시각적인 것에 담긴 사람을 bounding box로 localized
    - 2. Bounding box 내 atomic actio에 대해 labeling
    - 이때 actio은 spatio-temporal에 따라 localized

  - **Action**은 계층으로 구성
    Finest level은 atomic body movement로 구성
    Coarser level은 Goal-directed behavior

Left: Sit, Ride, Talk to; Right: Sit, Drive, Listen to

Finest Level action: 주황색 글씨('Sit')

Coarser Level action: 빨간색 글씨('Drive, Ride'),
파란색 글씨("Talk to, Listen to")

빨간색 글씨: Interaction with objects
파란색 글씨: Interaction with other person

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Motivation**:
  - Spatio-temporal localized atomic visual action을 위한 대규모 비디오 데이터 세트의 부족을 해결하기 위해 만들어짐

    ***Atomic visual actions**: 더 이상 나눠질 수 없는 기본적인 action, 더 복잡한 action의 구성 요소
  - Video action recognition의 SOTA 발전, Action recognition algorithm 평가를 위한 벤치마크를 제공

  - Real world에서는 모든 Action에는 Atomic Action들의 연속된 Annotation 이 필요하기 때문에 higher-level events가 필요
    -> 이러한 점이 motivate되어 AVA는 15-minute clip들을 labeling

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **AVA Dataset**:
  - 437개의 서로 다른 영화를 15분에서 30분 간격으로 frame 선정
  - 1Hz sampling frequency가 한 영화당 900개의 keyframe을 제공
    1Hz의 keyframe에 annotation을 달은 이유: 정밀한 시간 annotation을 요규하지 않으면서 action의 의미론적 내용을 포착할 수 있을 만큼 dense함
  - 각 keyframe은 AVA vocabulary에 따라 영상 내 모든 사람 객체의 action을 labeling

Left: Sit, Talk to, Watch; Right: Crouch/Kneel, Listen to, Watch

Left: Stand, Carry/Hold, Listen to; Middle: Stand, Carry/Hold, Talk to; Right: Sit, Write

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Data Collection**:
  - Ground truth를 만드는 과정으로, AVA dataset의 annotation은 5 단계로 진행된다.
    1. Action Action Vocabulary Generation
    2. Movie and Segment Selection
    3. Person Bounding box annotation
    4. Person linking
    5. Action annotation

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Data Collection**:
  - **Action Vocabulary Generation**
    - 3가지 원칙
      1. **Generality**
         - 일상생활 영상에서 Generic action을 수집 (Movie)
      2. **Atomicity**
         - Action classes는 명확한 visual signature를 가지고 Interacted object와 독립
         - (어떤 object를 보유할지 특정하지 않고 보유) -> list를 짧게, 완전하게 유지
      3. **Exhaustivity (완전도)**
         - 이전 dataset의 knowledge를 사용해 list 초기화하고, annotator에 의해 label이 지정된 AVA dataset의 action이 ~99%를 포함할 때까지 여러 round에서 list를 반복
    - In AVA vocabulary:
      - **14 Pose classes, 49 Person-Object interaction classes, 17 Person-Person interaction classes**

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Data Collection**:
  - **Movie and Segment Selection**
    - AVA dataset의 raw video conten는 YouTube를 통해 얻음
    1. 여러 국가의 탑 배우들의 list 모음
    2. List내 이름 YouTube search query 에 적용 -> 2000개의 result 검색
       (Film or television annotation, 30분 이상의 running time, 업로드 1년 이상, 조회수 1000회 이상)
       (흑백, 저해상도, 애니메이션, 만화 및 게임 비디오 제외)
    3. Movie 영상에서 15min ~ 30min 사이의 sub-part 만 label 지정
       (제목 트레일러 annotation 하지 않기 위해 시작 부분 제외)
    4. Label을 지정한 후, 각각의 15min clip은 1초의 stride로 900 개의 overlapping 3s movie segment로 분할

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Data Collection**:
  - **Person bounding box annotation**
    - Bounding box로 사람과 actino을 localization
    - Keyframe에 여러 대상이 있다면, 각 주체는 action annotation을 위해 별도로 annotator에게 표시되어 이들의 action label이 다를 수 있음
    - Bounding box annotation은 수동으로 집약해 hybrid 접근법 선택

    1. Faster R-CNN의 person detector를 사용해 초기 bounding box set을 만듬
       높은 정밀도를 위해 operation point 설정
    2. Annotators는 Detector가 놓친 나머지 bounding box를 annotate

    - 이러한 접근법은 bounding box 전체의 recall을 보장

    - **정확하지 않은 bounding box**는annotator에 의해 **표시**되고, action annotation의 **다음 단계 annotator에 의해 제거**된다.

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

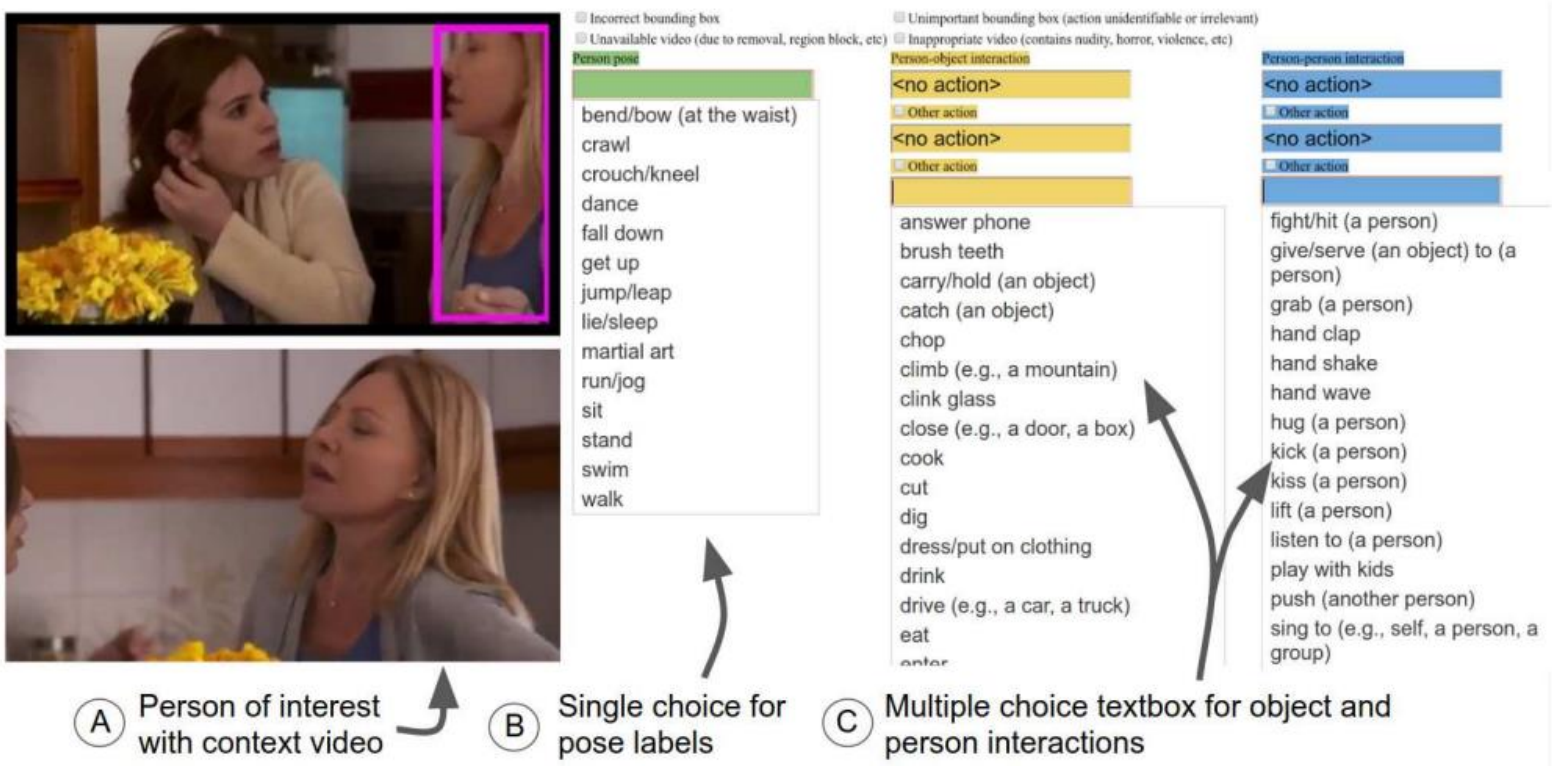
- **Data Collection**:
  - **Person link annotation**
    - 동일한 사람이 수행하는 개별 atomic visual action에 대한 annotation들을 연결하는 과정
    - Ground truth person tracklets을 얻기 위해 짧은 시간동안의 bounding box들을 연결
    - Person embedding을 사용해 인접한 keyframes의 bounding box 사이의 pairwise similarity를 계산하고 Hungarian algorithm을 통해 최적의 matching으로 해결
      * **Pairwise similarity**: atomic visual action 쌍 간의 유사성 계산
      * **Hungarian Algorithm**: Annotation의 충돌이나 불일치 해결에 사용
    - 자동으로 matching을 해주는 Hungarian algorithm은 좋은 방법이지만, 각각의 match를 검증하는 human annotator를 사용해 false positive를 추가로 제거
    - 이러한 절차를 몇 초에서 몇 분 사이의 81,000개의 tracklets 만듬

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Data Collection**:
  - **Action Annotation**
    - Action label은 crowd-sourced annotators에 의해 생성

    A. Target segment의 middle frame과 반복되는 embedded video로써의 segment

    B, C. 7개의 action label로 채워진 text boxes, 1 pose action, 3 person-object interations, 3 person-person interactions로 구성 (list에 없는 action인 경우 "other action" checkbox 체크)



AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Data Collection**:
  - **Action Annotation**

    2 Stage Action annotation pipeline

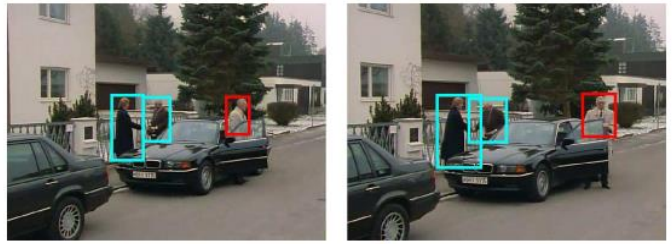
    1. **Action Proposal**
       공동의 proposal를 통해 높은 recall
    2. **Action Verfication**

    - 적은 예시를 가진 action에 대해 recall 성능 향상
    - 각각의 video clip은 3명의 독립적인 annotators에 의해 annotate
    - 최소 2명의 annotators로부터 검증되는 경우에만 action label을 ground truth로 간주

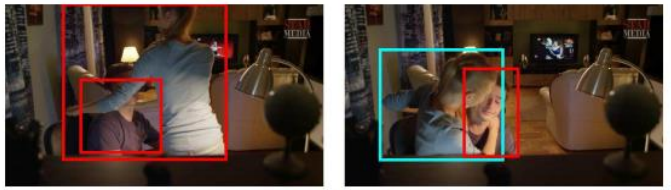
AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Training, Validation, Test set**:
  - **Action Training, validation, test set들은 Video level에서 분할**된다.
  - **하나의 비디오의 모든 segments**가 오직 **one split에서만 나타나도록** 함
  - 437 videos 은 239 개 의 training set, 64 개 의 validation set, 134 개 의 test set 으 로 나 뉨 (55:15:30) (215000, 57000, 120000 segments)
  - **Temporal context를 활용하는 이유**:
    - 연속적인 Segment 동안 변화하는 atomic aciton의 예시를 보면, Action classes가 다양한 contex에 따라 표현하는 것이 달라짐으로 이러한 미세한 차이를 구별해야 하고, 이를 위해 시간의 흐름에 따라 바뀌는 actio을 잘 구별해야한다.


clink glass → drink


open → close


turn → open


grab (a person) → hug


look at phone → answer phone


fall down → lie/sleep

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Temporal Structure**:
  - Segments 간의 person link를 하는 동안, 동일한 사람이 수행하는 action 의 쌍을 보면 공통된 연속적인 action 이 발견됨
  - 이러한 쌍을 분류하기 위해 **NPMI**(**normalized pointwise mutual information**)을 사용

    *NPMI: 두 사건이 함께 발생할 정도를 판단, 서로 정보량이 다른 사건을 비교할 때 그 값의 스케일이 다르기 때문에 제대로 된 비교가 어렵다는 점이 있음. 따라서 범위를 [-1, 1]로 일정하게 정규화 할 필요가 있음

$$NPMI(x,y) = (\ln \frac{p(x,y)}{p(x)p(y)})/(-\ln p(x,y))$$

    - 값은 -1,1 사이에 존재,
    - 절대 동시에 발생하는 단어가 아닌 경우 -1
    - 항상 동시에 발생하는 단어인 경우 1
    - 독립 쌍인 경우 0

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Temporal Structure**:
- "look at phone" → "answer phone", "fall down"→ "lie", or "listen to" → "talk to"
  - Action이 전환될 때 **공통된 temporal patterns**을 가짐.
- "ride" ↔ "drive", "play music" ↔ "listen", or "take" ↔ "give/serve"
  - 비슷한 의미를 가지는 action 쌍

동일한 사람에 대한 연속적인 1초 segments에서

상위 NPMI를 가진 action 쌍

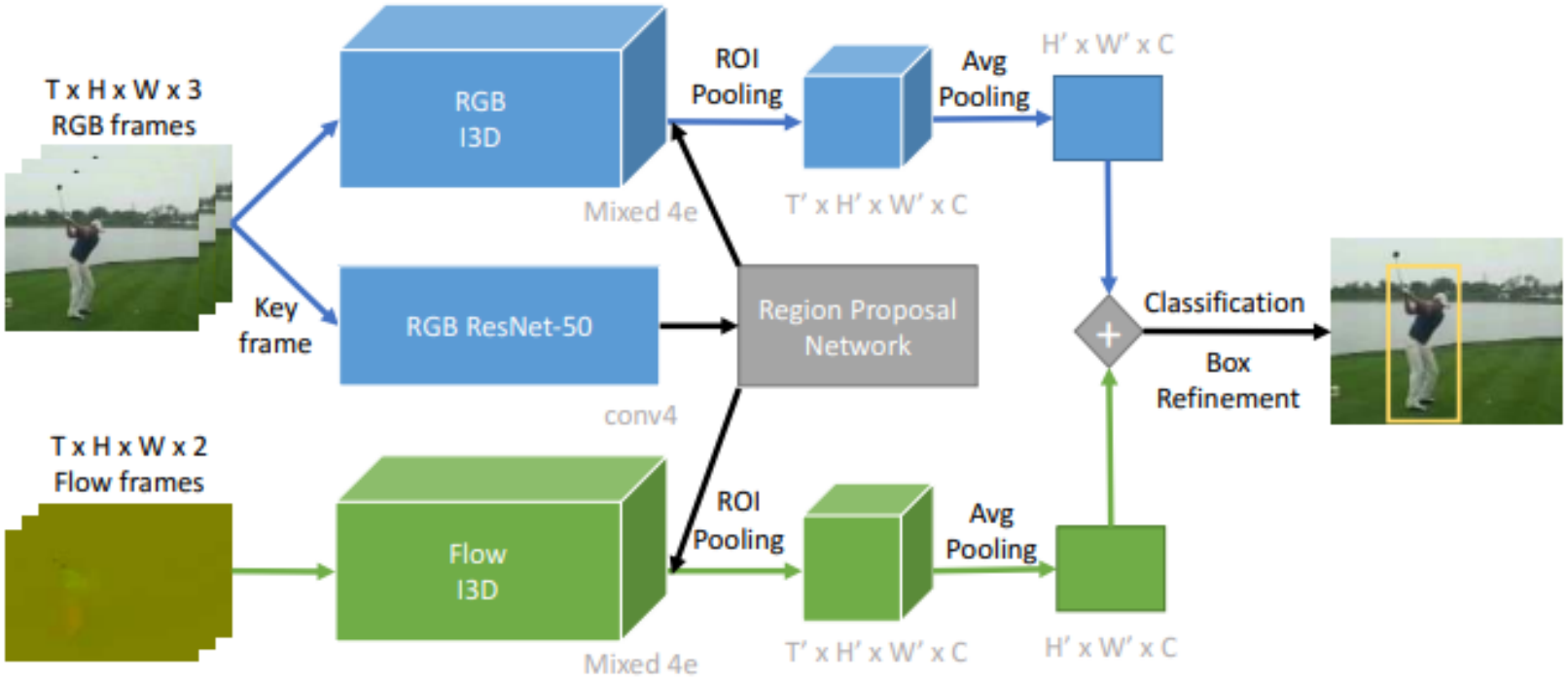
First Action	Second Action	NPMI
ride (eg bike/car/horse)	drive (eg car/truck)	0.68
watch (eg TV)	work on a computer	0.64
drive (eg car/truck)	ride (eg car bike/car/horse)	0.63
open (eg window/door)	close (eg door/box)	0.59
text on/look at a cellphone	answer phone	0.53
listen to (person)	talk to (person)	0.47
fall down	lie/sleep	0.46
talk to (person)	listen to (person)	0.43
stand	sit	0.40
walk	stand	0.40

여러 사람이 동시에 수행하는 action 쌍

Person 1 Action	Person 2 Action	NPMI
ride (eg bike/car/horse)	drive (eg car/truck)	0.60
play musical instrument	listen (eg music)	0.57
take (object)	give/serve (object)	0.51
talk to (person)	listen to (person)	0.46
stand	sit	0.31
play musical instrument	dance	0.23
walk	stand	0.21
watch (person)	write	0.15
walk	run/jog	0.15
fight/hit (a person)	stand	0.14

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- Action recognition task는 영상 내 **Actor가 다수**이거나, **이미지 크기가 작거나**, **미묘하게 다른 action**을 취하거나 **background scenes**가 무슨 일이 벌어지는지 **충분히 묘사되지 않는다면** 분류에 어려움이 생긴다.
- 저자는 Multi-frame temporal 정보를 사용하는 spatio-temporal action localization에 대한 최근의 접근법에 영감을 받아 SOTA action localization 접근법을 발전
- Action Detection을 위해 I3D 기반의 더 큰 Temporal context의 영향에 의존

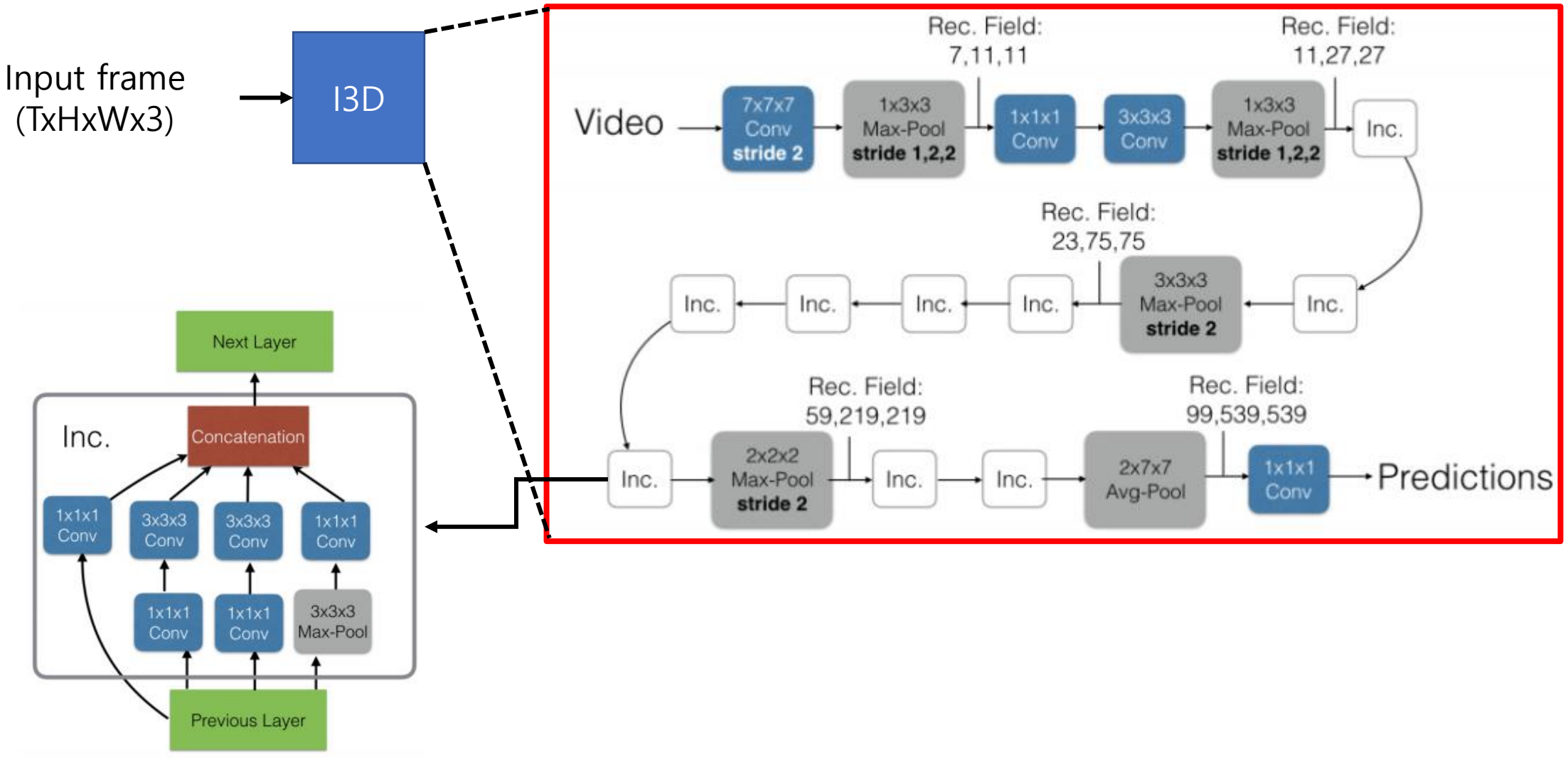


AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- Action의 End-to-End classification, localization 을 위해 Faster R-CNN을 사용
- 하지만 이러한 접근법은 시간이 지나메 따라 Multi-Frame의 입력 채널이 연결되는 **첫 번째 layer에서 temporal 정보를 잃어버림**
- 이를 **해결**하기 위해 **Temporal context 를 Model 하는 I3D 구조** (Inception 3D)를 제안
- I3D 구조는 Inception 구조를 기초로 설계했고, 기존의 2D conv를 3D conv로 변환
- I3D 구조는 temporal 정보를 유지

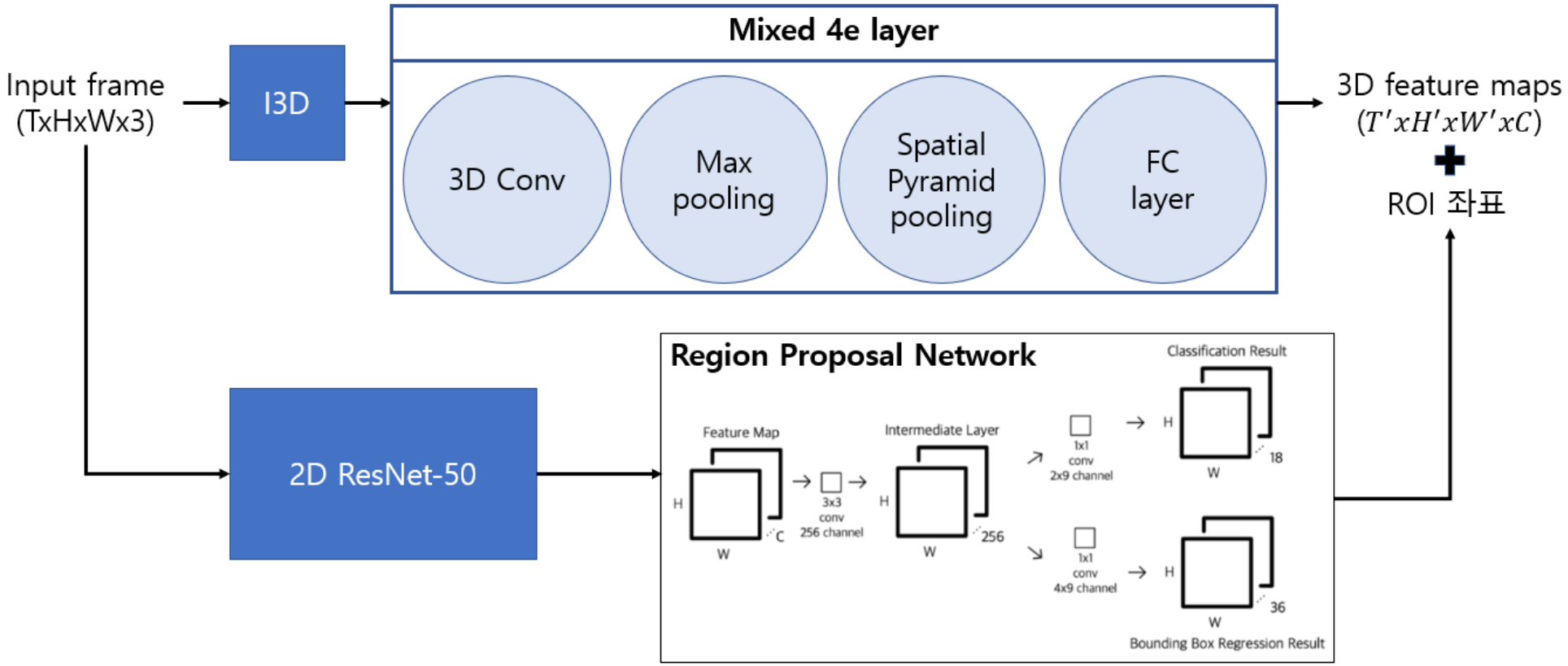
AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- I3D 와 Faster R-CNN을 같이 사용하기 위해, 모델 변환



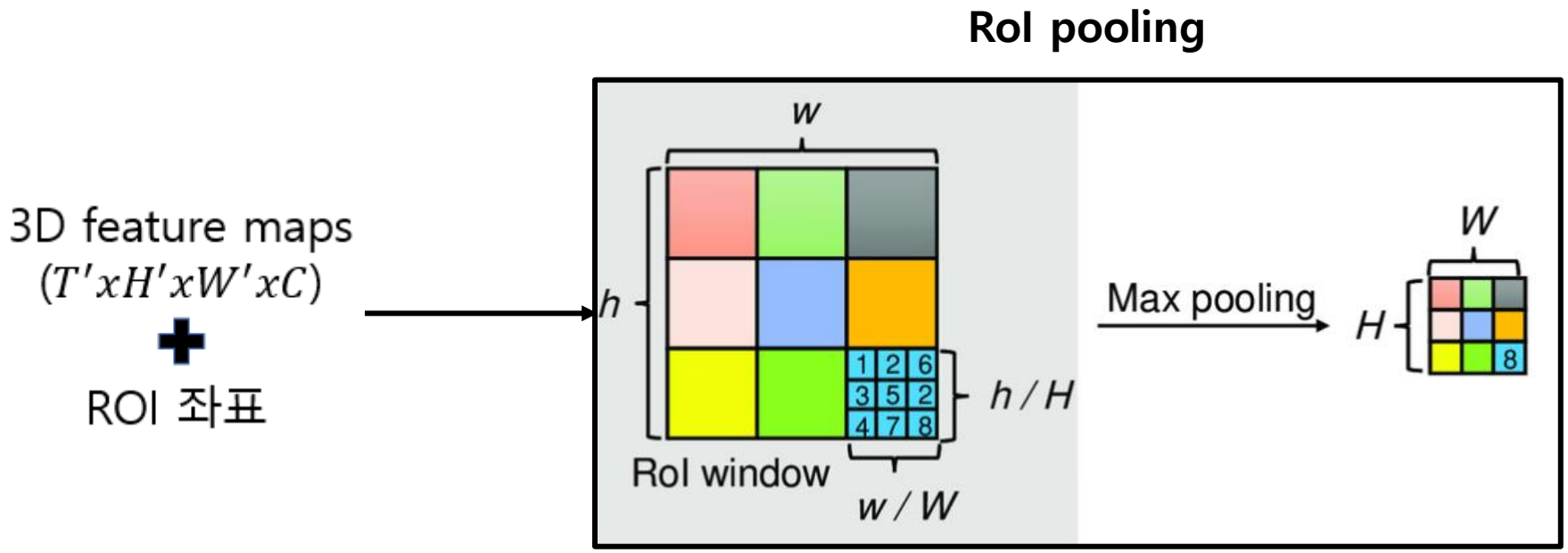
AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- **Spatial stream**
- Input length가 다른 I3D가 생성된 Action proposal의 quality에 영향을 미치지 못하도록 ResNet-50을 RPN 입력으로 사용
- Mixed 4e layer: Low & High level 정보를 사용해 Action classification 을 통해 frame에서 특징 추출



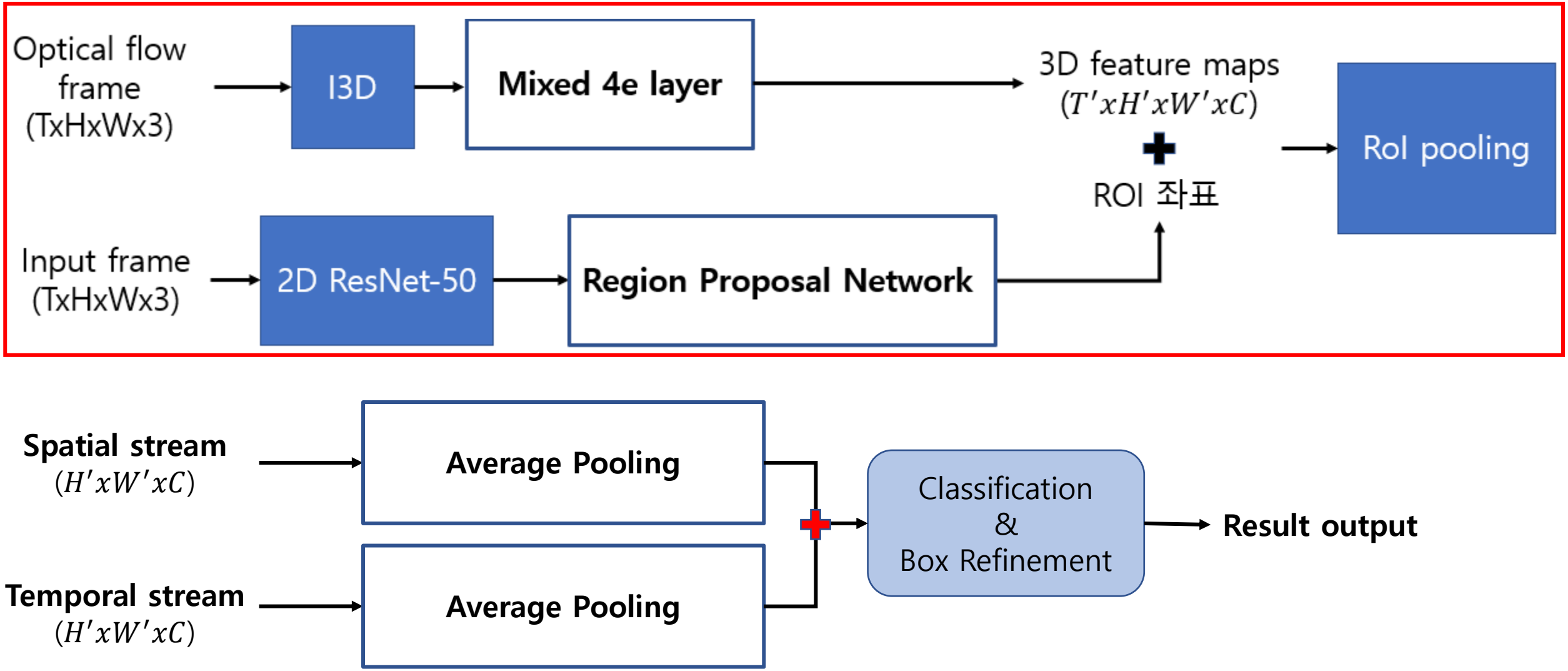
AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- **Spatial stream**
- RoI pooling을 통해 네트워크가 frame의 가장 유용한 영역에 집중할 수 있게 해주는 동시에 관련성이 적은 정보를 무시할 수 있게 해 네트워크 성능 향상



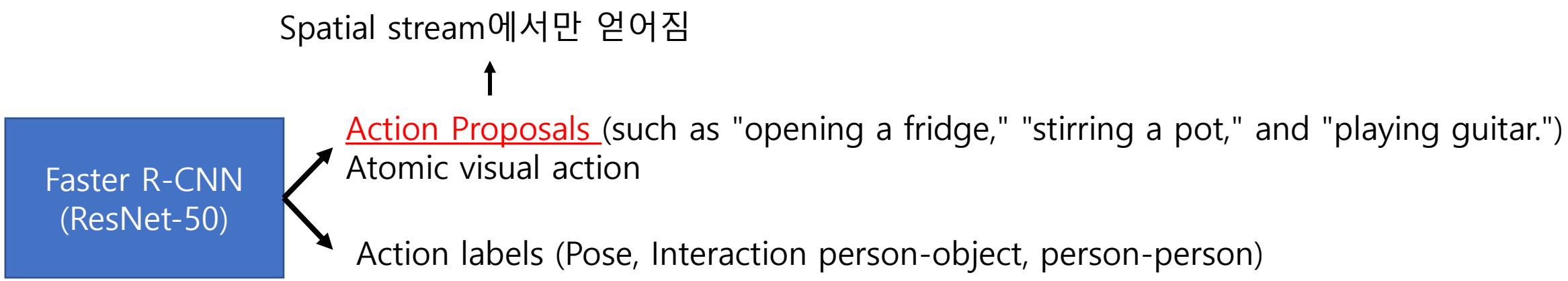
AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- **Temporal stream**



AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- **Baseline:**
  - Iteration: 600k ~ 1,000k
  - Asynchronous SGD
  - Input resolution: 320 -> 400
  - ResNet-50은 imageNet으로 사전 학습된 모델로 초기화
  - I3D network는 Kinetics datset으로 사전 학습된 모델 초기화 (Spatial, Temporal stream 모두)
  - Post-processing: Output frame-level detections를 threshold=0.6으로 Non-Maximum suppression
  - 각 class 별 하나씩 binary sigmoid losses의 합으로 대체 -> softmax X



AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Action Localization Model**:
- **Linking:**
  - Frame-level별 detections를 얻으면, Action tubes에 Link
  - 얻어진 Tubes의 average score를 기준으로 video-level 성능 구함
  - Detection link와 ground truth link 사이의 classes의 IoU 점수 계산 시, class에 의해 label이 지정된 tube segments만 고려

- **Linking:**
  - IoU 성능을 frame level과 video level에서 측정
  - IoU threshold=0.5로 설정해 Average precision (AP) 구함

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Result**:
  - Frame-mAP:
    - JHMDB: 73.3%
    - UCF101-24: 76.3%
    - SOTA 달성
  - Video-mAP:
    - JHMDB: 78.6%
    - UCF101-24: 59.9%
    - SOTA 달성

Frame-mAP	JHMDB	UCF101-24
Actionness [42]	39.9%	-
Peng w/o MR [30]	56.9%	64.8%
Peng w/ MR [30]	58.5%	65.7%
ACT [41]	65.7%	69.5%
Our approach	**73.3%**	**76.3%**

Video-mAP	JHMDB	UCF101-24
Peng w/ MR [30]	73.1%	35.9%
Singh *et al.* [38]	72.0%	46.3%
ACT [41]	73.7%	51.4%
TCNN [16]	76.9%	-
Our approach	**78.6%**	**59.9%**

  - AVA
    - Frame-mAP: 15.8%
    - Video-mAP: 12.3% (0.5 IoU), 17.9% (0.2 IoU)

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Action

- **Result**:
  - AVA에 대한 고찰
  - AVA의 성능이 좋게 나오기 위해서는 fine-grained details와 풍부한 temporal model을 인식해야 함
  - Temporal window의 length를 늘리면 모든 데이터셋에서 3D two-stream 모델에 도움이 된다.
  - RGB와 optical flow를 결합하는 방식이 더욱 성능을 향상
  - AVA는 기존 데이터셋 대비 더 큰 temporal context 이점을 가짐
  - AVA는 다른 데이터셋과 다르게 spatial, temporal stream에 포화가 발생하지 않고 계속해서 성능 증가

Model	Temp.+ Mode	JHMDB	UCF101-24	AVA
2D	1 RGB + 5 Flow	52.1%	60.1%	14.2%
3D	5 RGB + 5 Flow	67.9%	76.1%	13.6%
3D	10 RGB + 10 Flow	73.4%	78.0%	14.2%
3D	20 RGB + 20 Flow	76.4%	78.3%	14.8%
3D	40 RGB + 40 Flow	76.7%	76.0%	**15.8**%
3D	50 RGB + 50 Flow	-	73.2%	15.7%
3D	20 RGB	73.2%	77.0%	14.6%
3D	20 Flow	67.0%	71.3%	10.1%