Deeper Action

湃道智能 AI PRIME

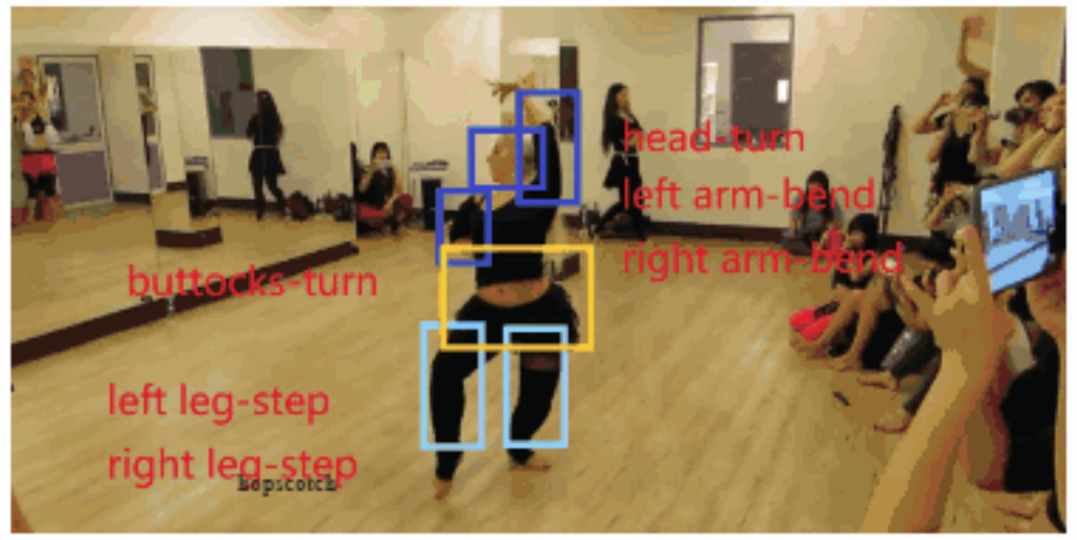
ECCV TEL AVIV 2022

Kinetics-TPS Track on Part-level Action Parsing and Action Recognition
Tech Report

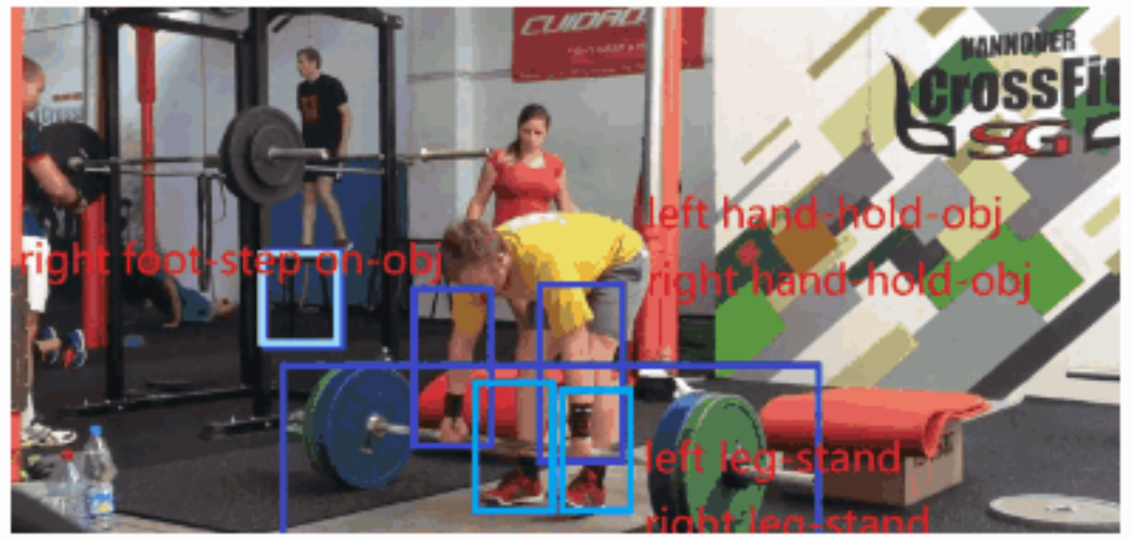
Jiawei Dong[1], Yuliang Chen[2], Shuo Wang[1]

1. Shanghai Paidao Intelligent Technology Co., Ltd.
2. Chongqing Jiaotong University

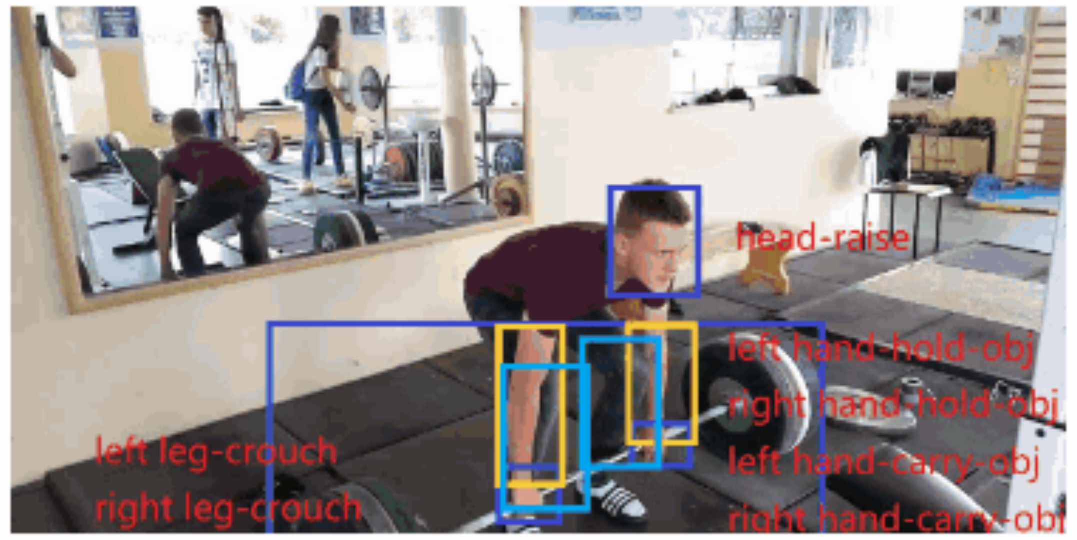
Dataset Introduction and Statistics



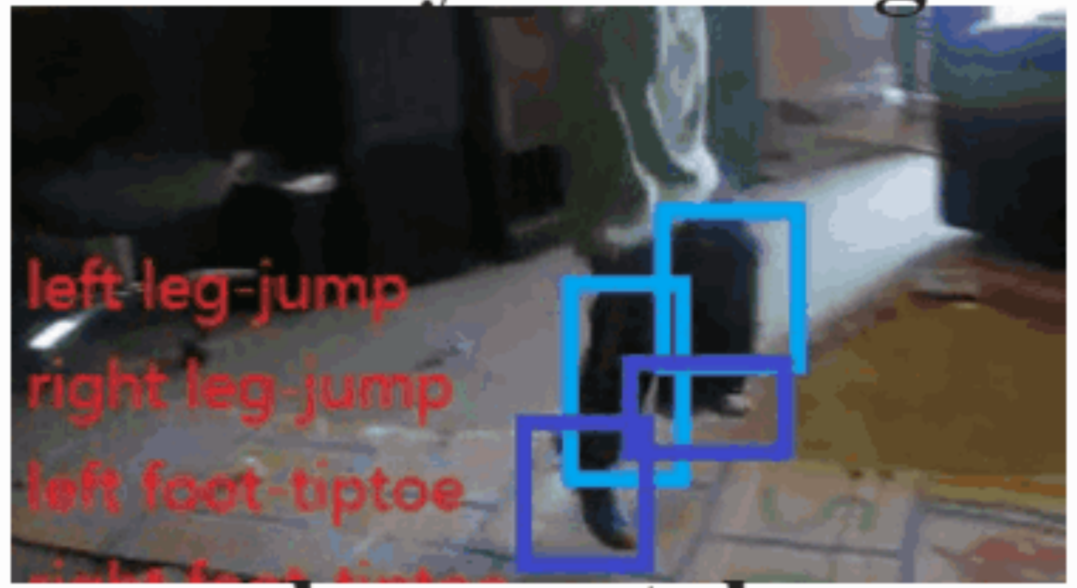
belly_dancing



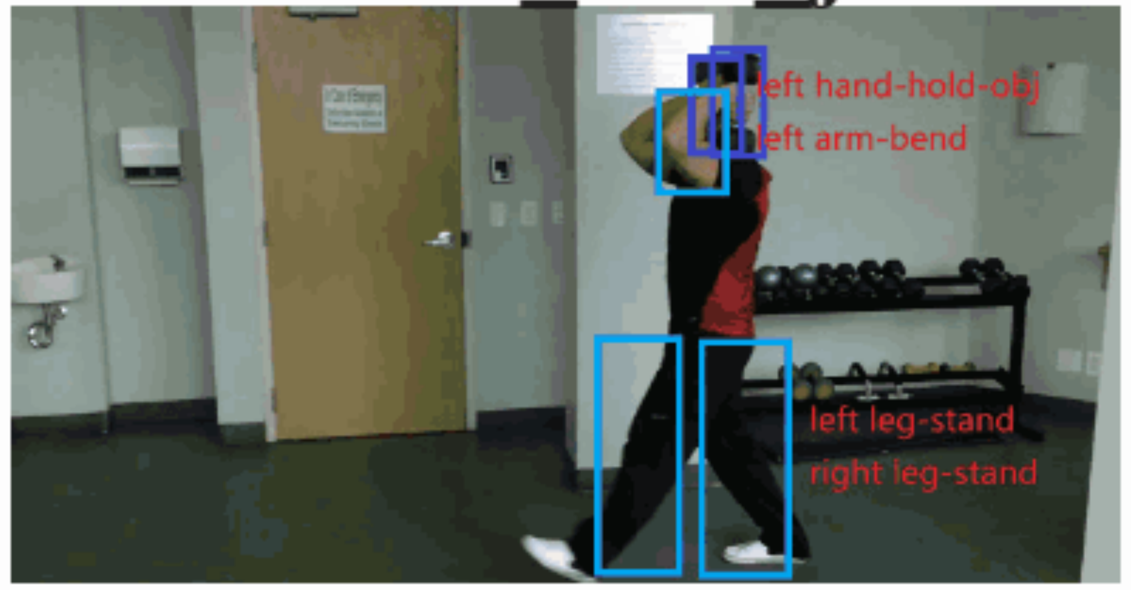
clean_and_jerk



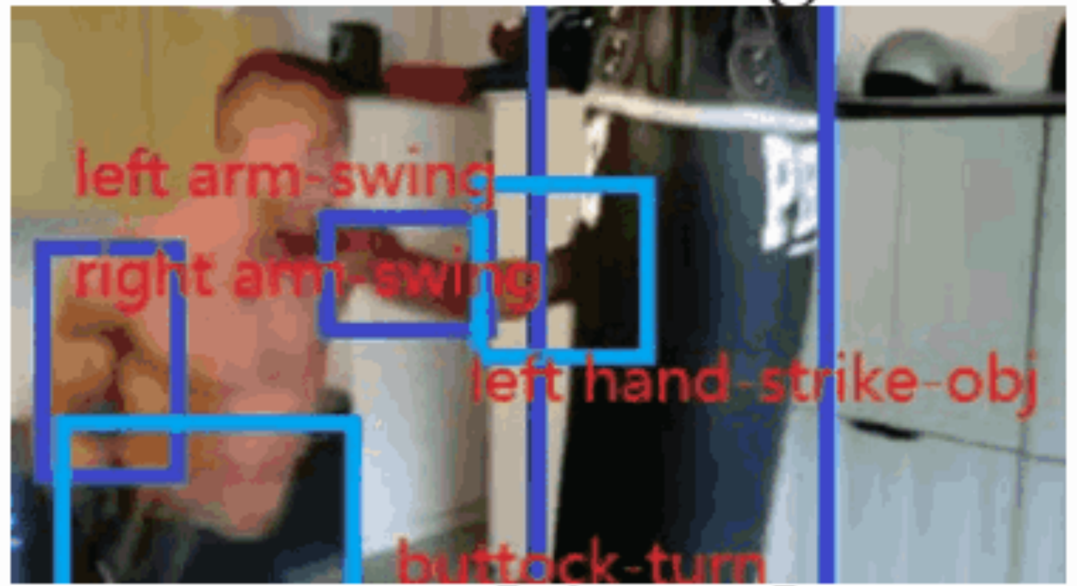
deadlifting



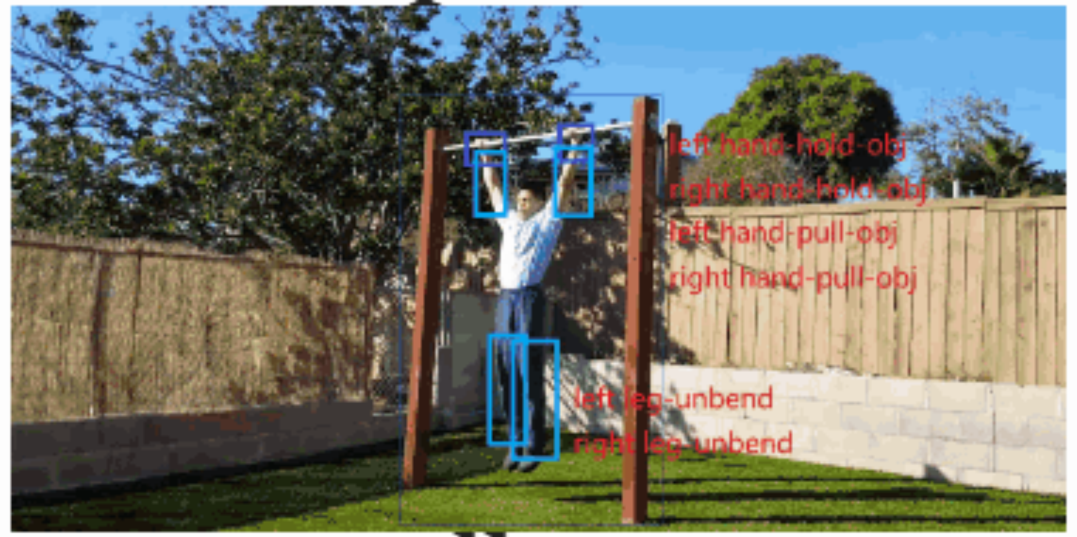
hopscotch



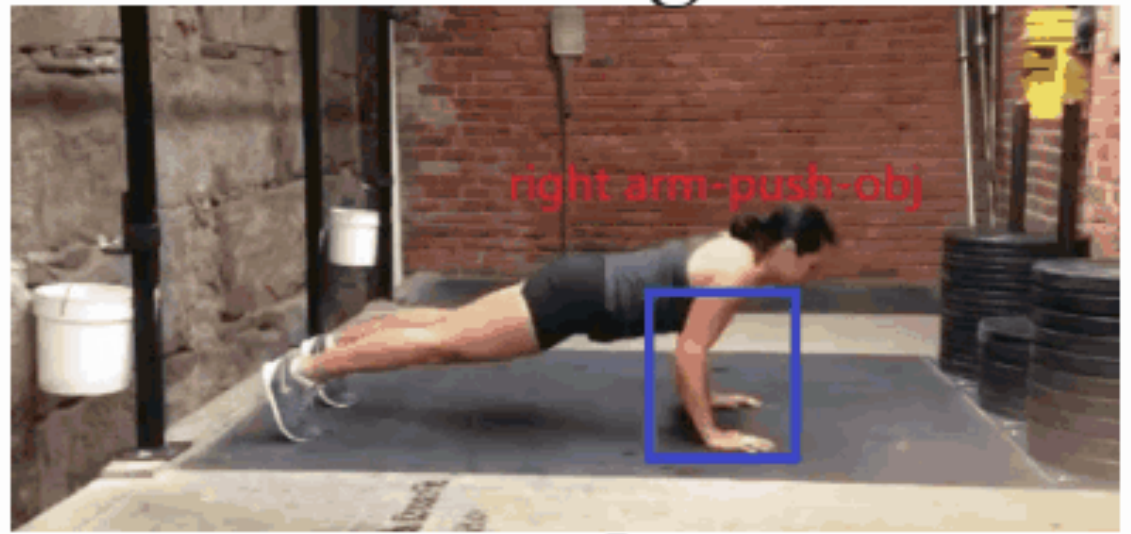
lunge



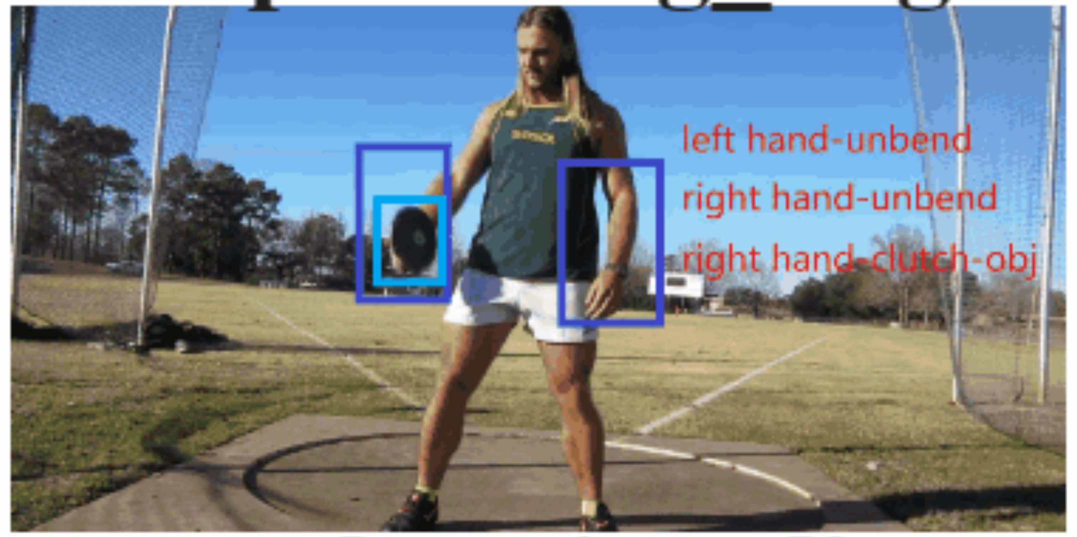
punching_bag



pull_ups



push_up



throwing_discus

Needs to predict:

- Human bounding box
- Body part bounding box
- Frame-level part state
- Video-level human action

Dataset Introduction and Statistics

Kinetics-TPS contains 4740 videos

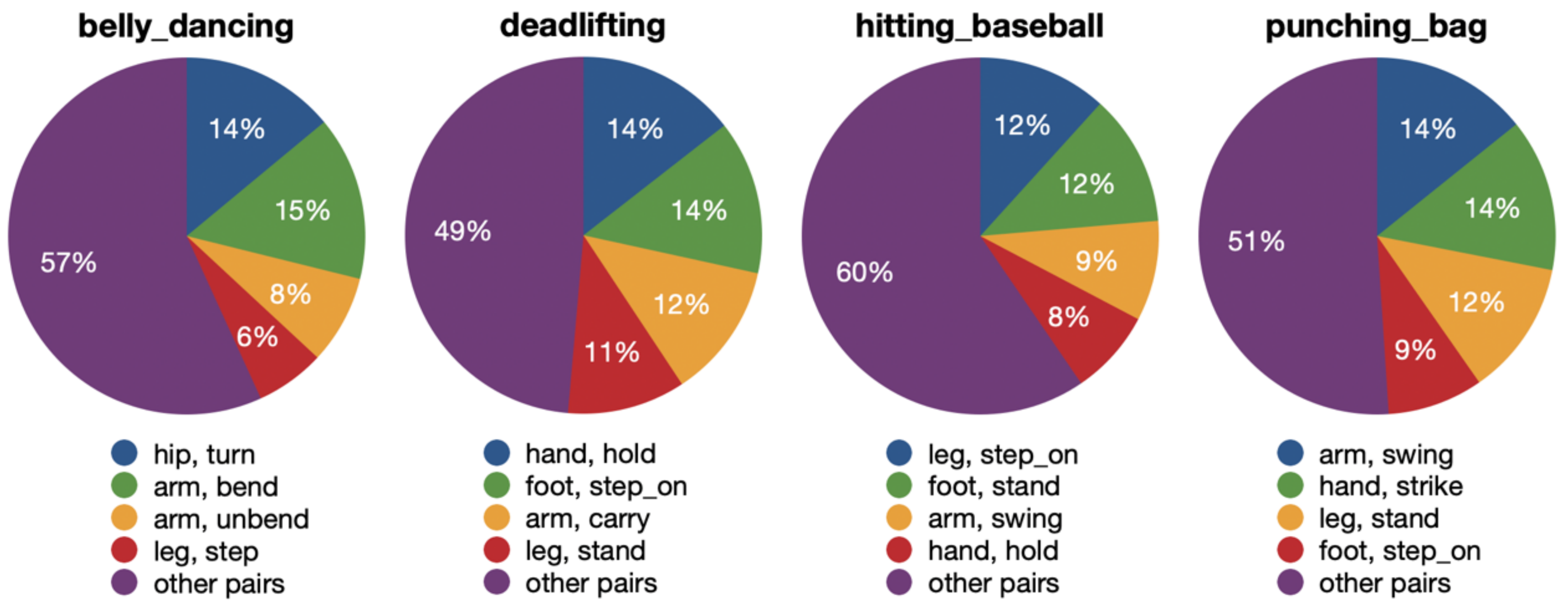
1) Bounding boxes of human instances: 1.6 M

2) Bounding boxes of body parts: 7.9M

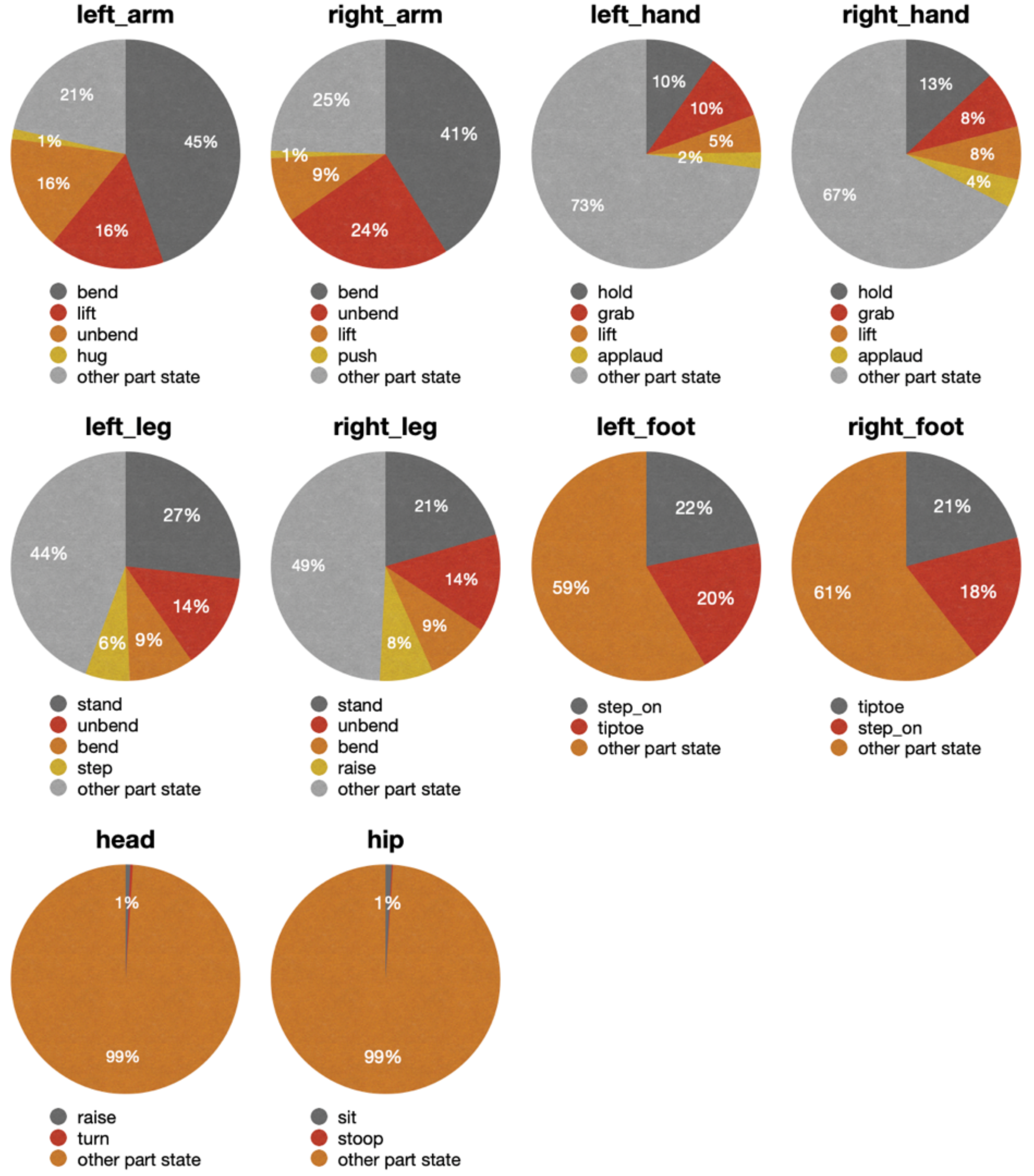
3) Part state tags of each annotated part: 7.9M

4) Bounding boxes and tags of objects: 0.5 M

5) 'body part，part state' pairs of four exemplar classes in Kinetics-TPS



6) Top-5 'part state' tags of each body part in Kinetics-TPS



Data Preprocess

Frame extracting

- We extract 574,851 labeled frames from 3,809 thousand training videos, extract 48,655 frames from 932 testing set videos with 5 frames interval. The extracted frame images retain the original resolution.

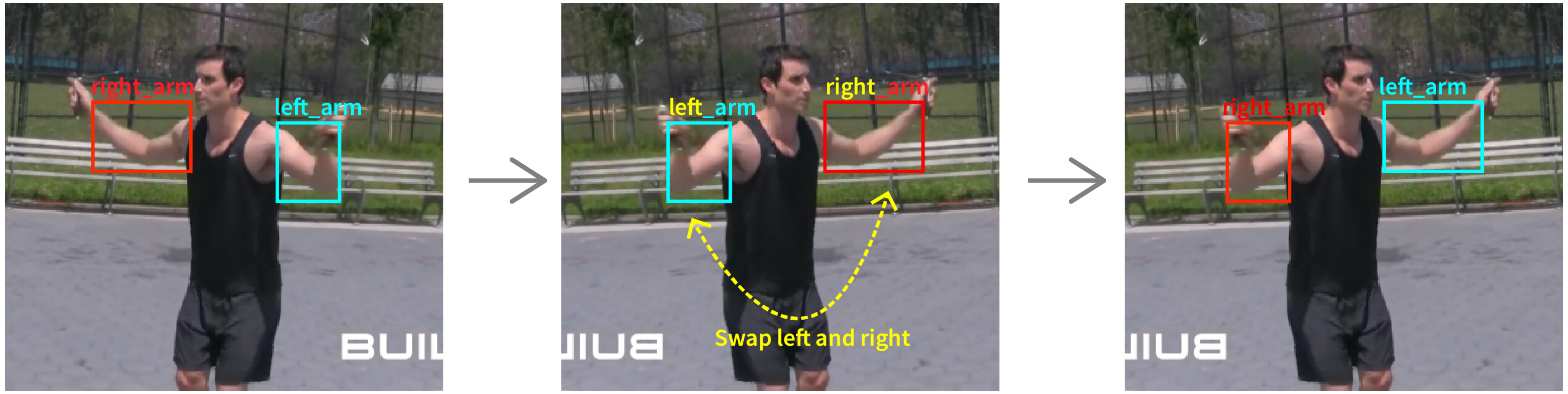
Split	Video Number	Frame Number
Training Set	3,809	574,851
Testing Set	932	48,655

Data Augmentation

- **Object detection:** mixup, mosaic, label swap, rotation, perspective, scale and shear.
- **Action recognition network:** label swap, rotation and scale.

Label Swap

We considered that the task required to distinguish the left and right parts of the human body, so when horizontal flipping is used, we needed to swap the label with "left" and "right".



Original Image

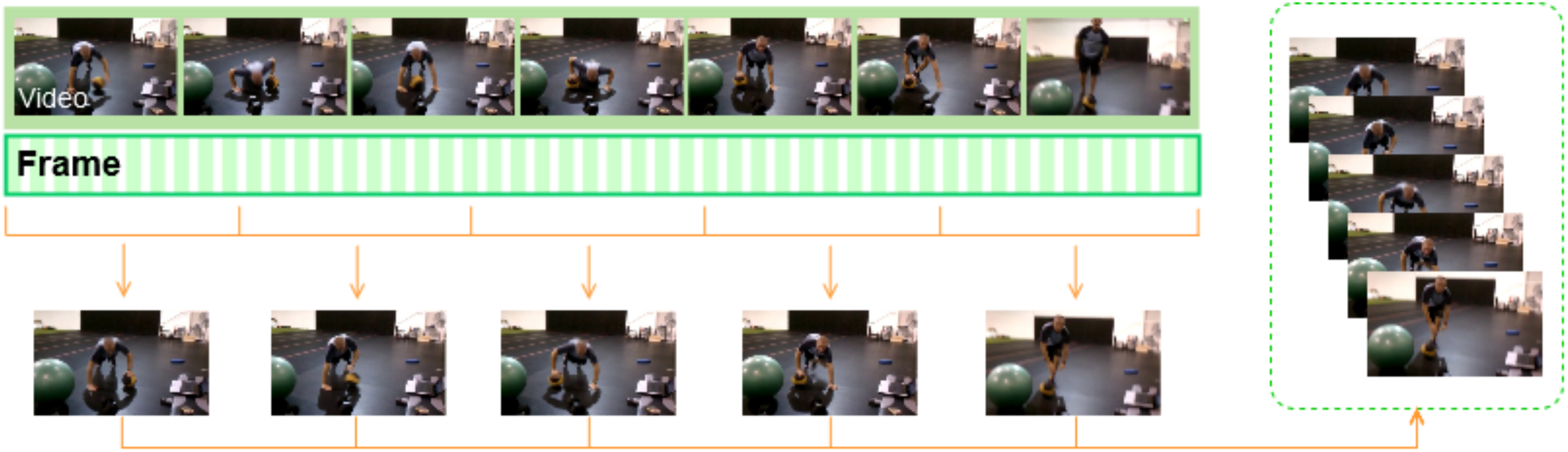
Horizontal Fliped Image

Horizontal Fliped Image + Label Swap

Data Preprocess

Uniform Sample

When *n* frames is required for sampling from a video, the video is divided into *n* segments of equal length, for each segment, there is only one frame is sampled in random position.

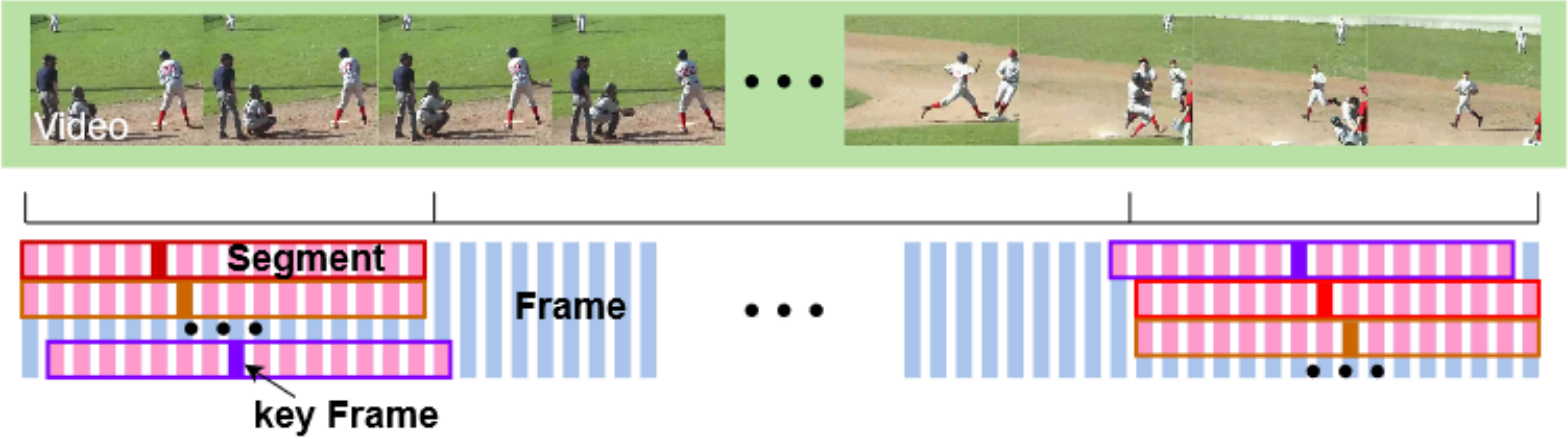


- **Advantage:** no matter how long the video duration is, uniform sampling can avoid missing key information.
- **Disadvantage:** the sampled frames may lack continuous information for videos with long video duration or short duration of key actions.

Data Preprocess

Dense Sample

For a video, we sample one segment with fixed length, and the length of this segment is determined by the number of sampling frames and frame interval. For each segment, the label of start frame or middle frame will be used as the label of the segment, and we used padding for the beginning and end of the video.

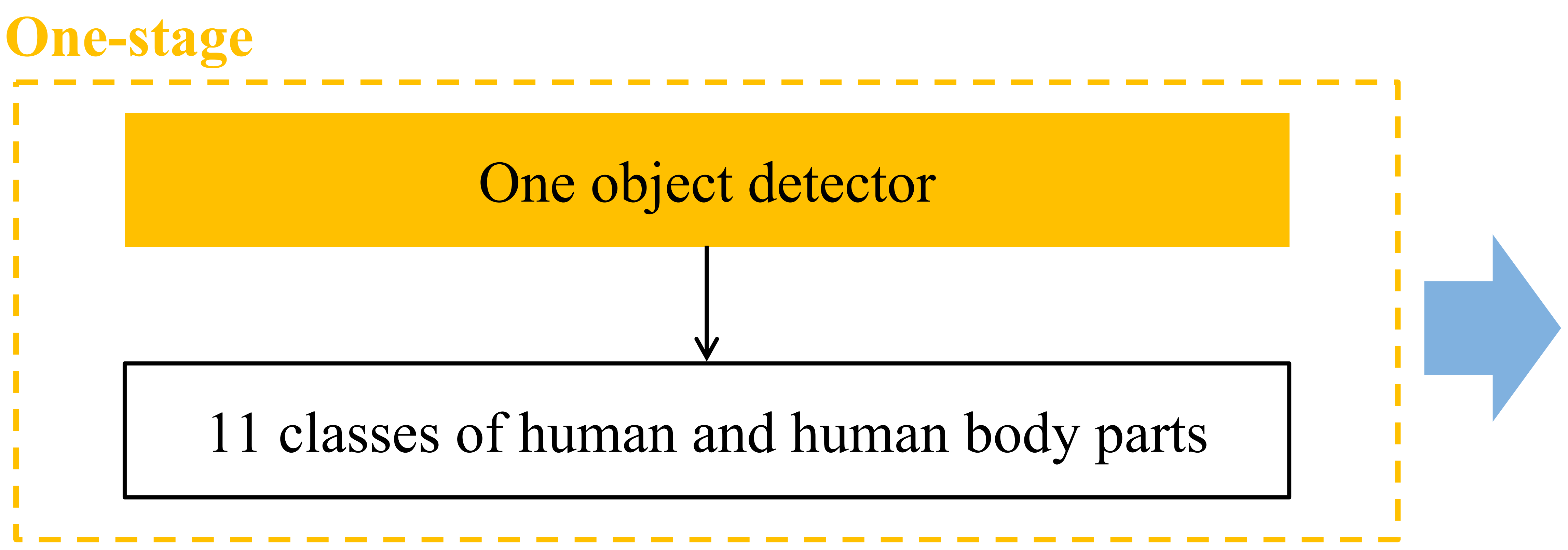


- **Advantage:** strengthening the recognition of action with short duration. All frames in the segment have strong temporal information due to their small frame interval.
- **Disadvantage:** the number of sampling frames directly affects the performance of action recognition network, which requires manual adjustment.

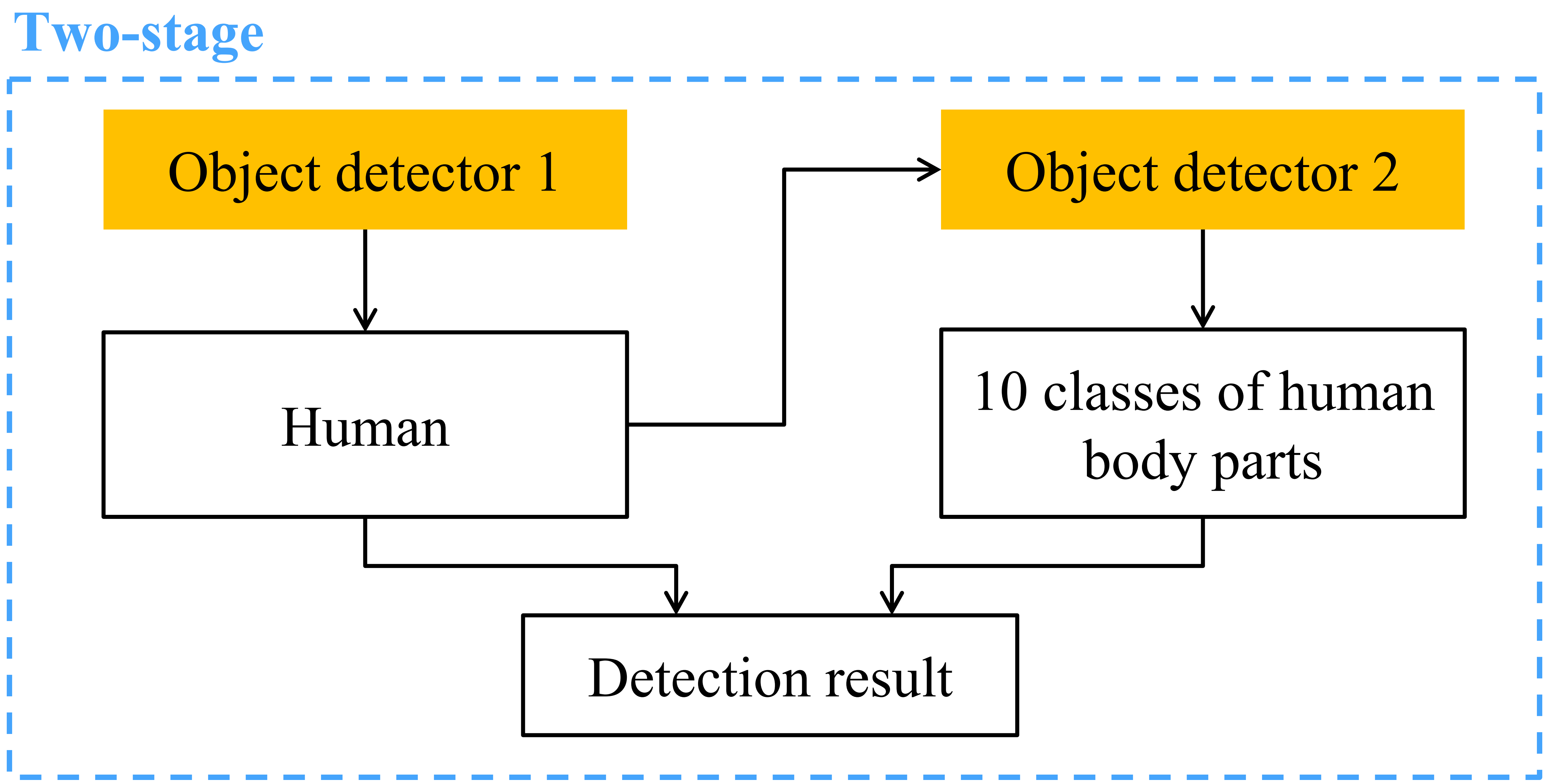
Our Method

The methods we used are composed of three parts: **human and body parts detection, video action recognition and part state recognition**. All the methods share the same detection and video action recognition block, the only difference between methods is part state recognition block.

Human and Body Parts Detection



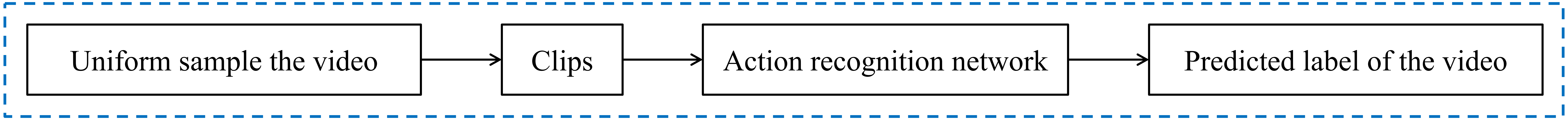
- Training a object detector with total of 11 classes of human and human body parts.



- Training a detector that only detects the human body.
- Cropping the RGB image of the person according to the person's bounding box.
- Detecting human body part of 10 classes.

Our Method

Video Action Recognition



Uniform sample the video → Clips → Action recognition network → Predicted label of the video

push_up

front_raises

skateboarding

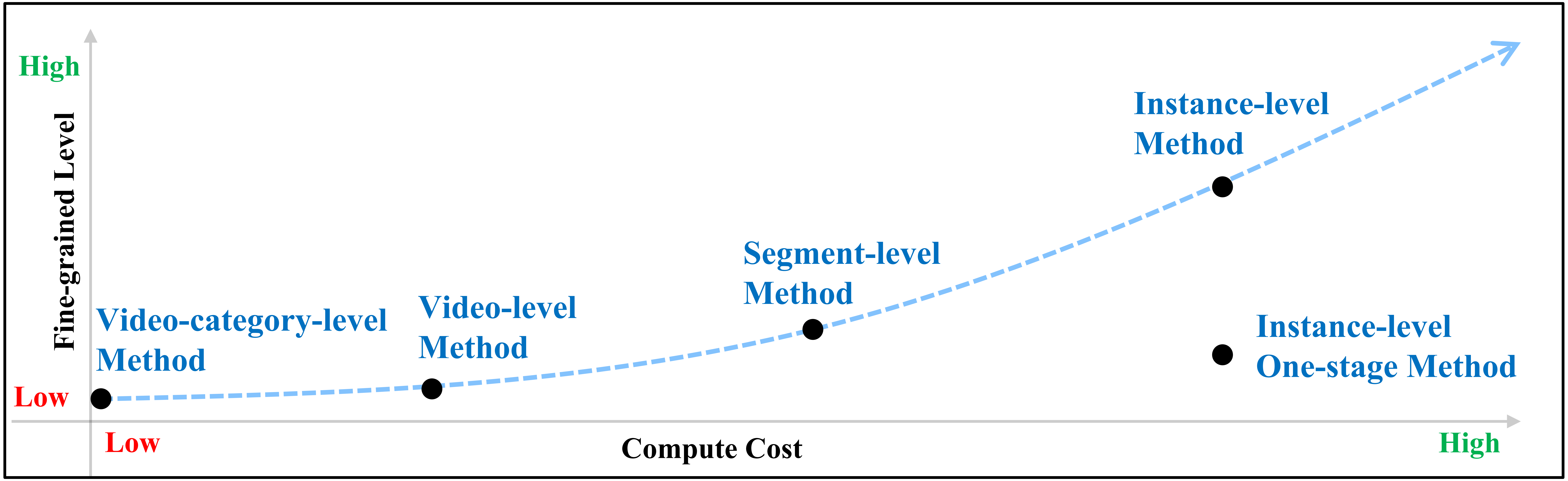
riding_mechanical_bull

Category Num	24
Model	Video Swin Transformer
Epoch	80
Batch-size	2
Clip Length	32
Video Resolution	360
Learning Rate	0.0003
Optimizer	AdamW
Pre-trained	ImageNet
Val Top1 ACC	99%

Our Method

Part State Recognition

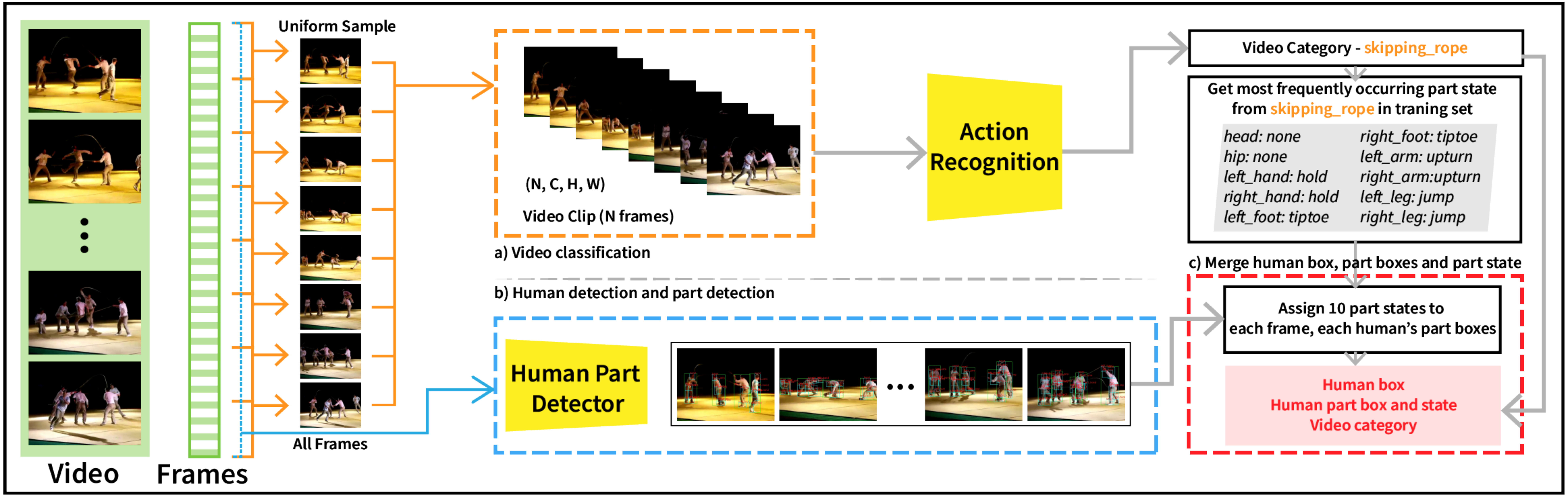
Action recognition of human body parts is critical step of this challenge. According to the fine-grained level, from low to high, we propose **video-category-level**, **video-level**, **segment-level** and **instance-level methods**.



Our Method

Video-Category-Level Method（0.4834）

- Counting the part state in each category, and obtain the most frequently occurring part state in each video category.
- For a given video, according to the predicted video category , the most frequently occurring part states of the video category are assigned to the part states of each person in each frame of the video.



Uniform Sample

Video

Frames

All Frames

(N, C, H, W)

Video Clip (N frames)

a) Video classification

Action Recognition

Video Category - skipping_rope

Get most frequently occurring part state from skipping_rope in traning set

head: none
hip: none
left_hand: hold
right_hand: hold
left_foot: tiptoe
right_foot: tiptoe
left_arm: upturn
right_arm:upturn
left_leg: jump
right_leg: jump

c) Merge human box, part boxes and part state

b) Human detection and part detection

Human Part Detector

Assign 10 part states to each frame, each human's part boxes

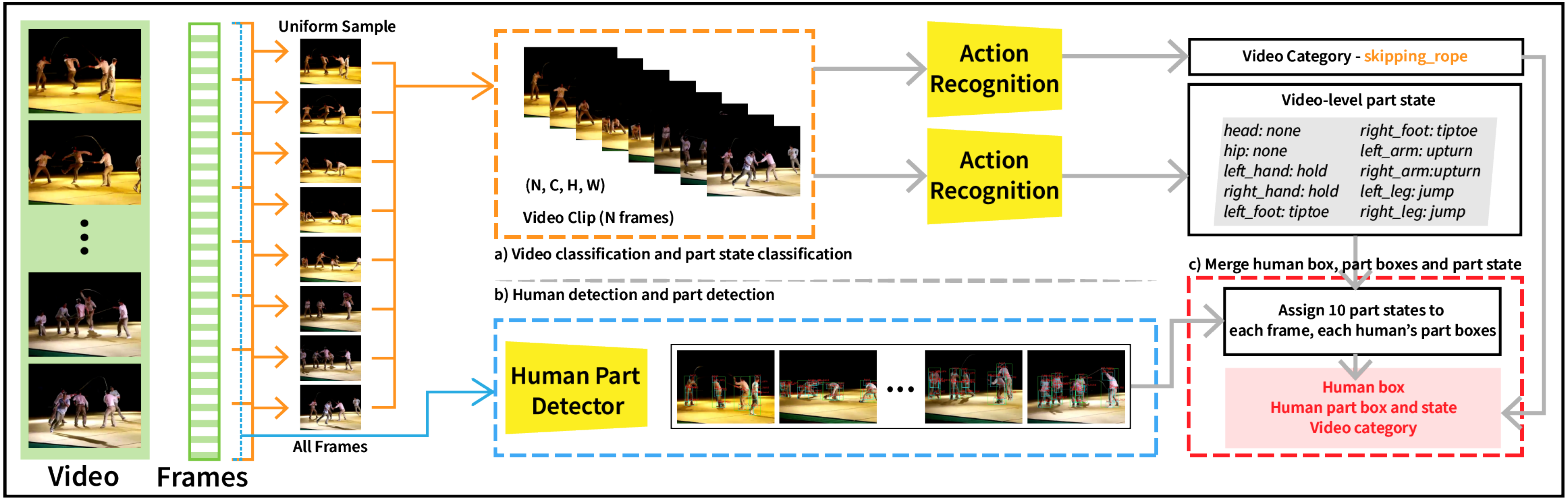
Human box
Human part box and state
Video category

Category-level method

Our Method

Video-Level Method（0.5911）

- Counting the most frequently occurring part state of each part of each video in training data set.
- Using the most frequently occurring part state of each part as its labels for training.
- Assigning the predicted label to each human of each frame in this video.



Uniform Sample

(N, C, H, W)

Video Clip (N frames)

Action Recognition

Action Recognition

Video Category - skipping_rope

Video-level part state

head: none
hip: none
left_hand: hold
right_hand: hold
left_foot: tiptoe
right_foot: tiptoe
left_arm: upturn
right_arm:upturn
left_leg: jump
right_leg: jump

a) Video classification and part state classification

b) Human detection and part detection

c) Merge human box, part boxes and part state

Human Part Detector

Assign 10 part states to each frame, each human's part boxes

Human box
Human part box and state
Video category

All Frames

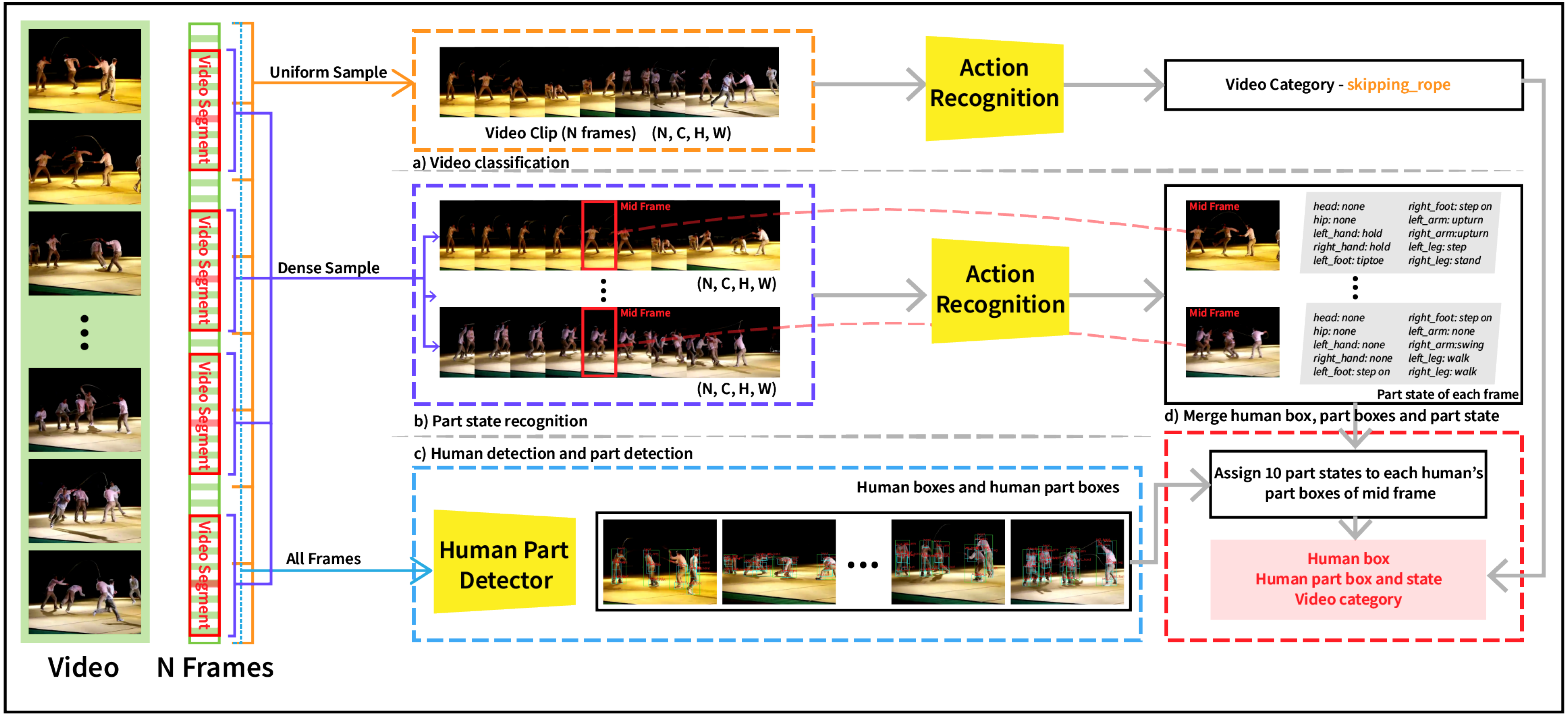
Video

Frames

Video-level method

Our Method

Segment-Level Method（0.560093）



Segment-level method

Our Method

Segment-Level Method（0.560093）

Segment-level method experiment results

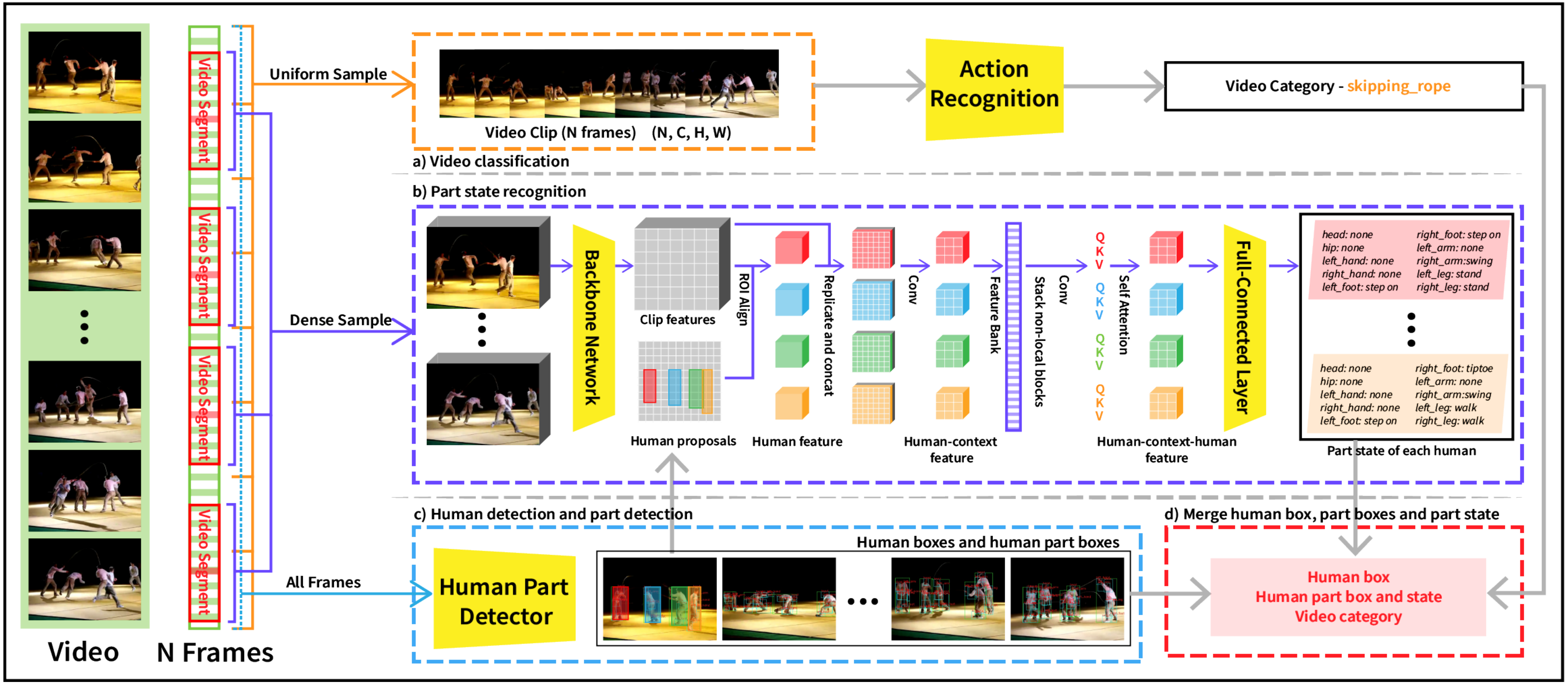
Model	Backbone	Clip Length	Lr	Epoch	Leaderboard Score
ir-CSN	ResNet3dCSN	16	5.12E-04	58	0.549715
ir-CSN	ResNet3dCSN	32	5.12E-04	58	-
ir-CSN	ResNet3dCSN	32	2.56E-04	58	**0.560093**

The multi-label action recognition network we used is ir-CSN. We train for 80 epochs with batch size 2, labels num 108, segment length 32, video resolution 320, base learning rate 0.000256, one-cycle scheduler and AdamW optimizer. We used IG-65M pre-trained model for training.

Using this method, our score on leaderboard can reach up to 0.560093.

Our Method

Instance-Level Method（0.662429）



Instance-level method

Our Method

Instance-Level Method（0.662429）

Instance-level method experiment results

Relation Model	Backbone	Clip Length	Epoch	Det-Threshold	Leaderboard score
Person-Person	Slowfast-Resnet101	16	3	0.1	0.395823
	Slowfast-Resnet101	16	4	0.1	0.554853
	Slowfast-Resnet101	16	5	0.1	**0.558903**
	Slowfast-Resnet101	16	6	0.1	0.554733
Person-Context-Person	Slowfast-Resnet101	16	6	0.1	0.620262
	Slowfast-Resnet101	32	6	0.1	**0.626608**
	Slowfast-Resnet101	32	6	0.01	**0.662429**

We have experimented another method which is focusing on modeling person-person relation, inspired by AIA. We only need to replace the person-context-person module in the part state recognition part with person-person module.

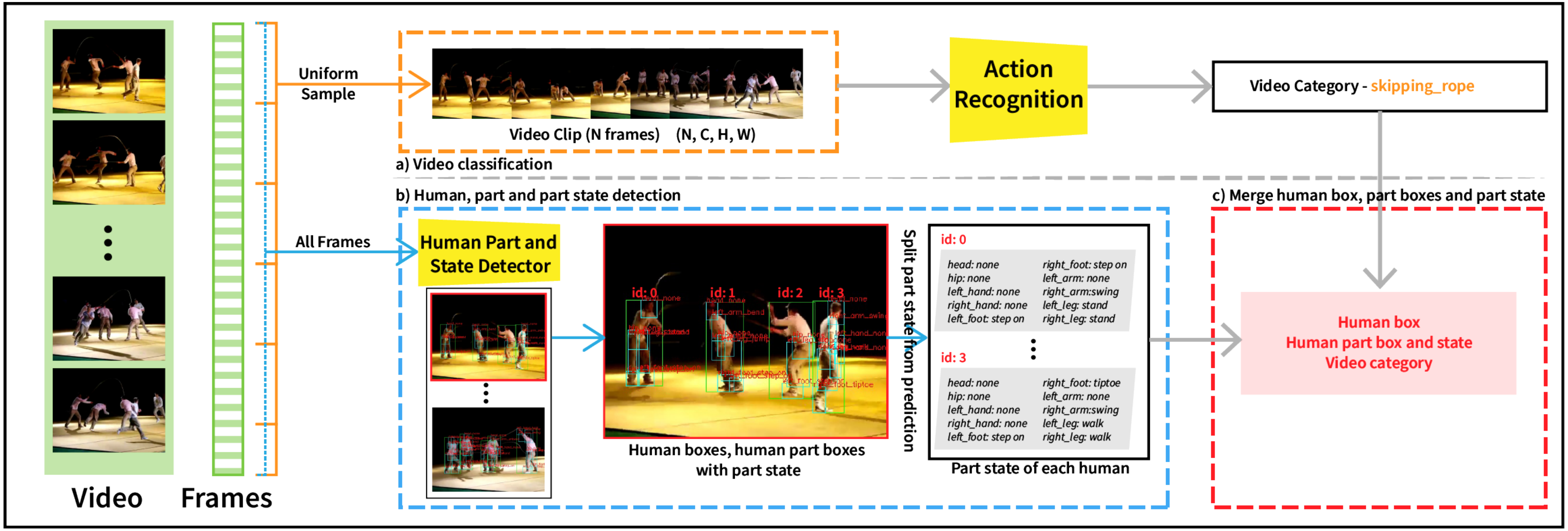
Here is our experiment result, result shows that the person-context-person modeling can obtain better detection result than person-person modeling.

Our Method

Instance-Level One-stage (0.6597)

Some actions may be accurately identified without considering their temporal characteristics.

- In training: Concatenating the part name and part state into a new label.
- In inferring: Getting predictions, we can easily split part name and part state form the predicted labels of bounding box.

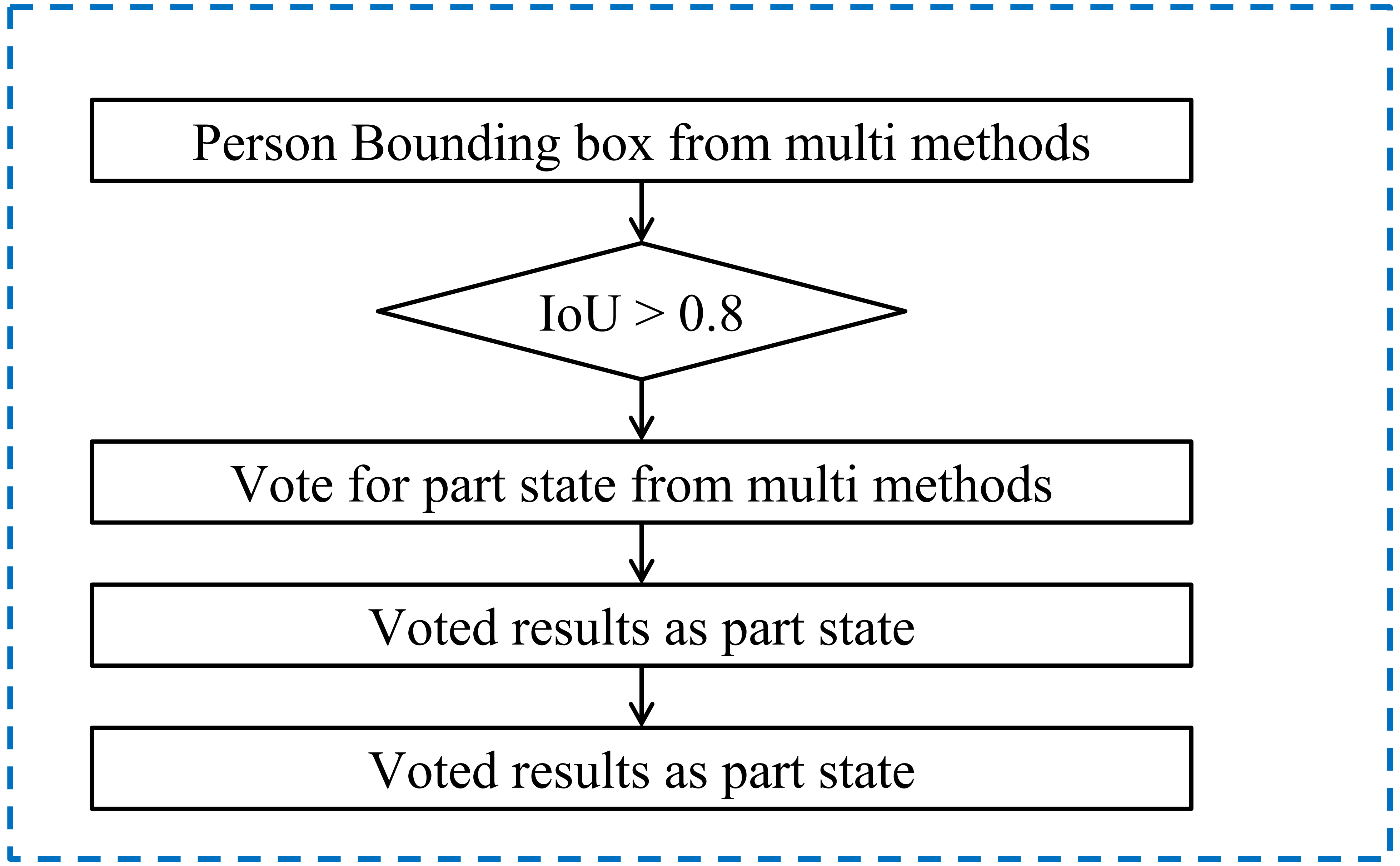


Instance-level method – One stage

Our Method

Ensemble by Voting (0.6824)

- Traverse the bounding boxes of all people in all frames under all methods.
- If the IoU of multiple bounding boxes is larger than 0.8, it is assumed multiple bounding boxes are referring to the same person.
- Count the state of each part predicted by different methods
- Take the part state with the largest count number as the part state of our ensemble result.



Person Bounding box from multi methods

IoU > 0.8

Vote for part state from multi methods

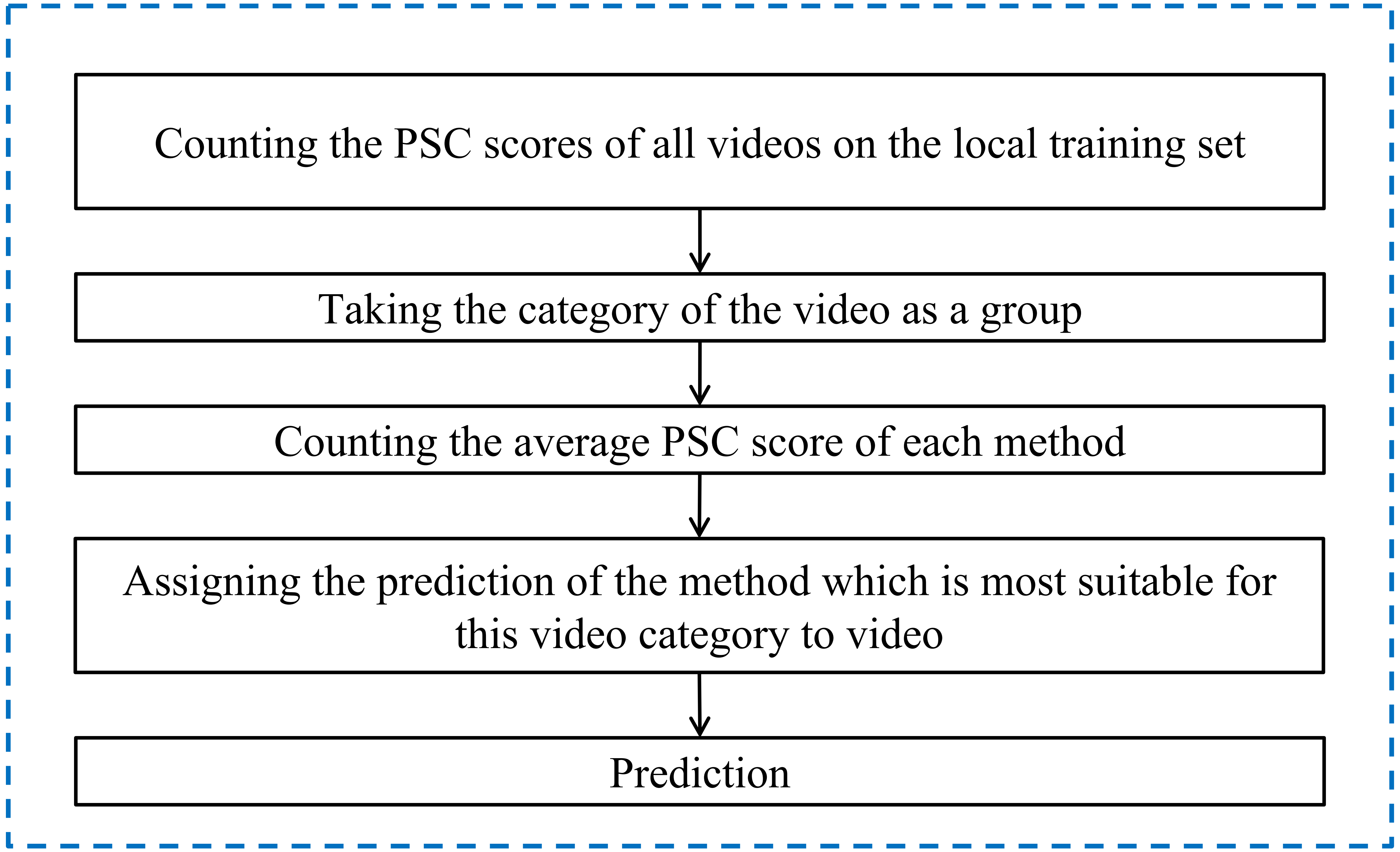
Voted results as part state

Voted results as part state

Ensemble by voting

Ensemble by Video Category (0.7389)

- Calculate the Part State Correctness(PSC) scores of all videos on the local training set with all methods .
- Get the most suitable method for each video category.
- On the testing set, according to the predicted video categories, assign the prediction of the method which is most suitable for this video category.



Counting the PSC scores of all videos on the local training set

Taking the category of the video as a group

Counting the average PSC score of each method

Assigning the prediction of the method which is most suitable for this video category to video

Prediction

Ensemble by video category

Conclusion

- To improve detection performance, we use two detectors to detect human and parts separately, bypassing the process of assigning parts to human

- We propose a data augmentation method called label swap

- To improve granularity of part state prediction, we Propose **video-category-level, video-level, segment-level and instance-level methods**. Moreover, it is verified that person-context-person relationship modeling can effectively improve the recognition ability of the network for complex actions, and it is more efficient than the traditional person-context and person-person modeling

- Although temporal information is critical in part state recognition, but even if the temporal information is discard, high PSC scores can be obtained with only two detectors, which may be due to the long-tailed distribution of the dataset

- Methods designed with different structures are good at different category of videos in the prediction of part states, so ensemble multiple results of methods can greatly improve the score

Thank you for listening!