School of Computer Science and Engineering

Intelligent Surveillance System: the gap between now and future

-Using CCTV for human behavior analysis

Dr. Xun Li
Research Associate
Xun.li1@unsw.edu.au

Overview

- Intelligent Surveillance System, the gap between now and future
- Introduction to Action Recognition
- Workflow Overview
- Pedestrian Detection and Tracking
- Action Recognition Overview
- Skeleton-based Action Recognition
- Demos

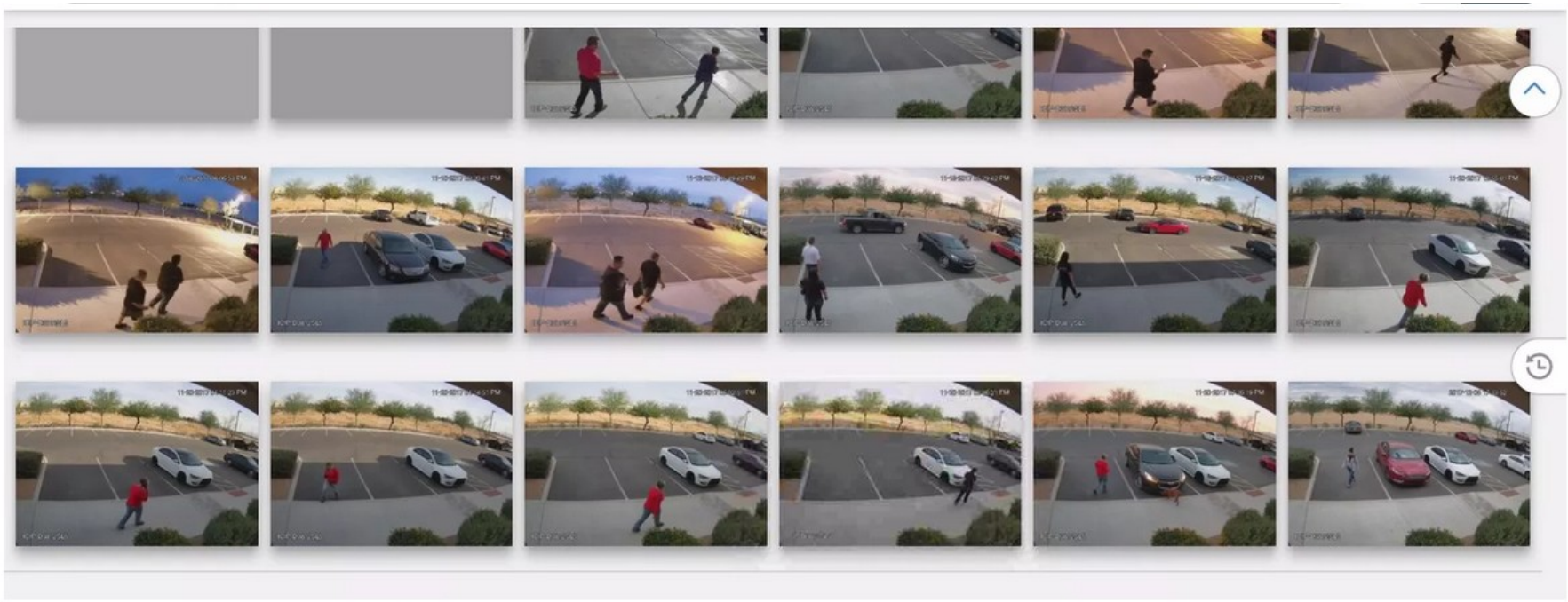
Introduction to Intelligent Surveillance System

- Traditional CCTV with manual reading

- “Smart” security cams with basic functions

- Combining artificial intelligence with surveillance:

  Today: detection, identification, recognition. But what next? ***Behavior Understanding***



A screenshot showing Ella being used to search for people wearing red. | Image: IC Realtime

Ref: https://www.theverge.com/2018/1/23/16907238/artificial-intelligence-surveillance-cameras-security

Introduction to human action recognition

Applications:

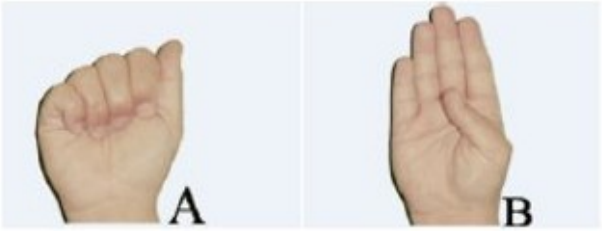
- Human Gesture based Control, Robotics, Human-Machine interaction, Medical Systems, and **Intelligent Surveillance System**

Picture from third party

Introduction to human action recognition

(1) Human actions:

- primitive posture/gesture
- unit action (walk, stand, sit, etc.)
- human activity
- multiple human interaction
- group activity



Pictures from third party

UNSW SYDNEY

Introduction to human action recognition

(2) Context:

- Recognize single person , unit action in trimmed videos: a classification problem;

- Recognize multiple humans' actions in untrimmed videos taken at natural environment: action detection + classification;

- For a Intelligent Surveillance System, three levels [1] of vision processing are required: human detection (low-level vision), human tracking (intermediate-level vision), and behaviour understanding methods (high-level vision)

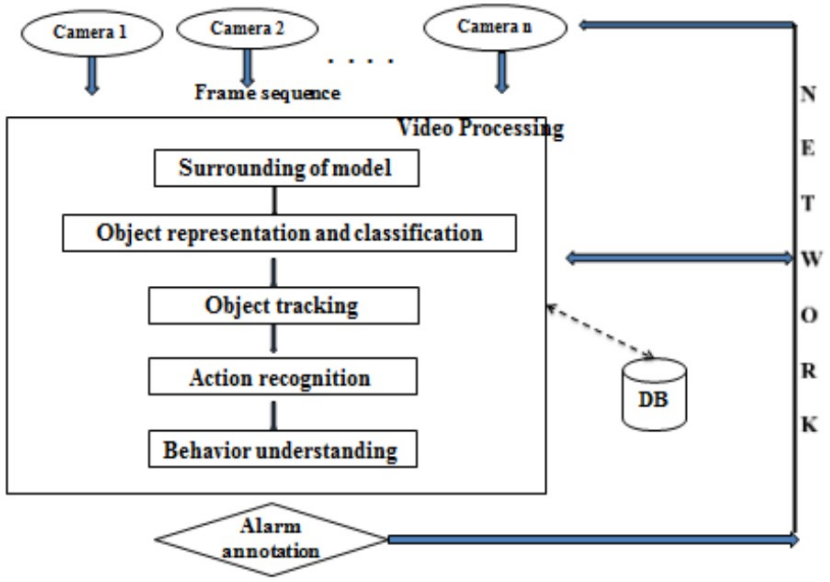
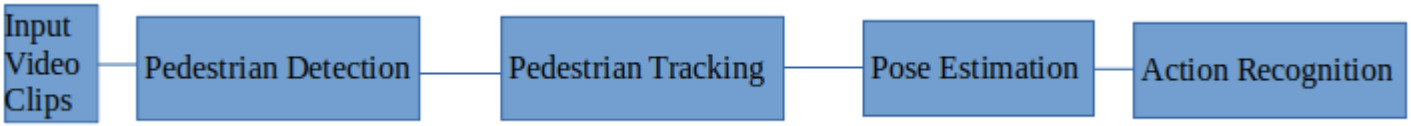


Fig1 . From Revathi, A.R., 2012. A Review of Human Activity Recongnition and Behavior Understanding in Video Surveillance. Academy and Industry Research Collaboration Center (AIRCC), pp. 375–384. doi:10.5121/csit.2012.2337

Workflow Overview

- Detection
- Tracking
- Action Recognition



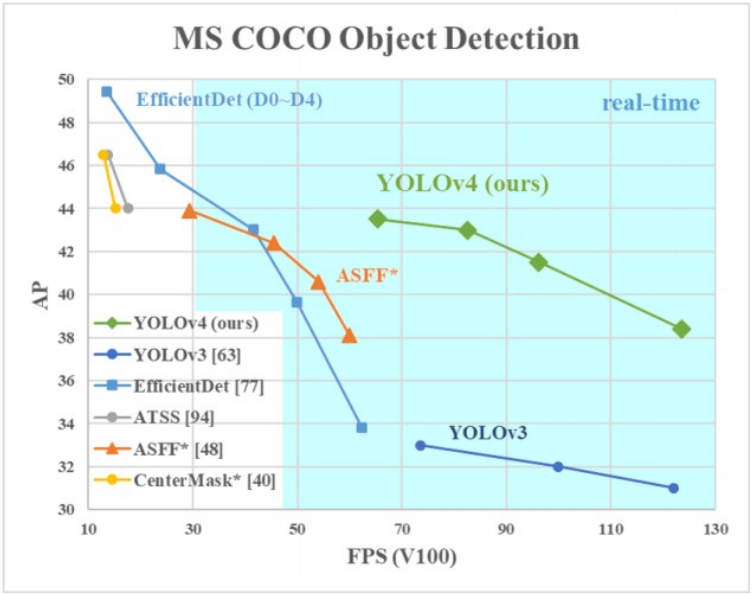
Module 1: Pedestrian Detection

Target objects are detected on each frame and output in the form of a bbox. Detectors are always trained in advance. Common pedestrian detector are:

(1) Non-deep learning based methods: color histogram, HOG+image pyramid+sliding window+SVM classifier. Problem for these methods are the handling of occlusion and deformation during detection and tracking.

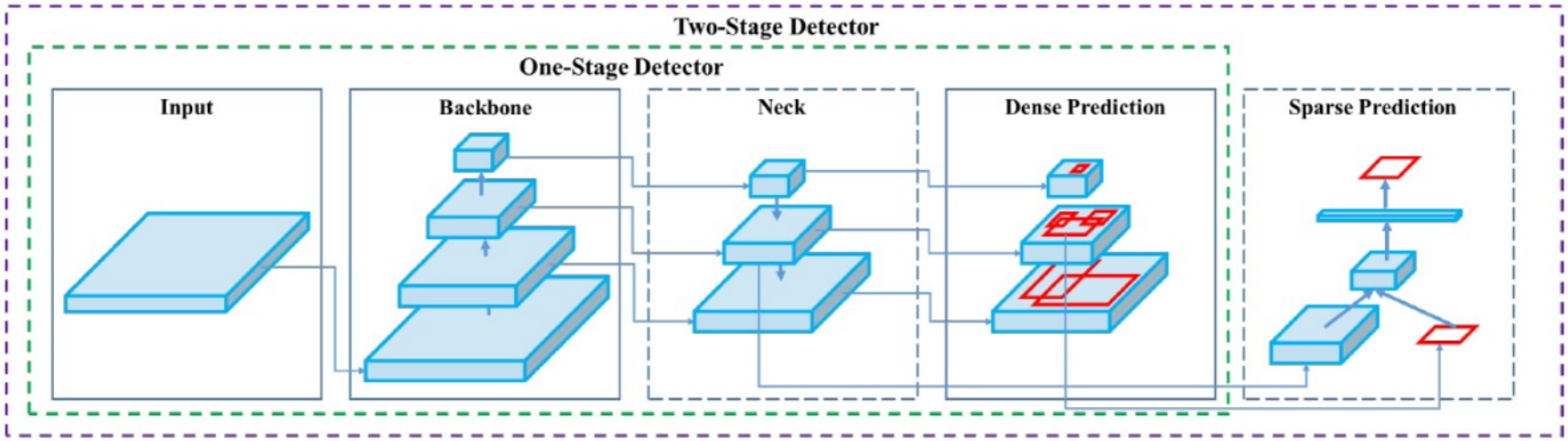
(2) Deep-learning based methods: Faster RCNN , SSD , YOLO series, CenterNet etc.

(3) We are currently using YOLO series as it is a good balance between accuracy and speed, which was pre-trained on MSCOCO [4] dataset.



Picture from third party

Module 1: Pedestrian Detection



Input: { Image, Patches, Image Pyramid, … }

Backbone: { VGG16 [68], ResNet-50 [26], ResNeXt-101 [86], Darknet53 [63], … }

Neck: { FPN [44], PANet [49], Bi-FPN [77], … }

Head:

Dense Prediction: { RPN [64], YOLO [61, 62, 63], SSD [50], RetinaNet [45], FCOS [78], … }

Sparse Prediction: { Faster R-CNN [64], R-FCN [9], … }

Picture from third party

Module 1: Pedestrian Detection Demo

Module 2: Pedestrian Tracking

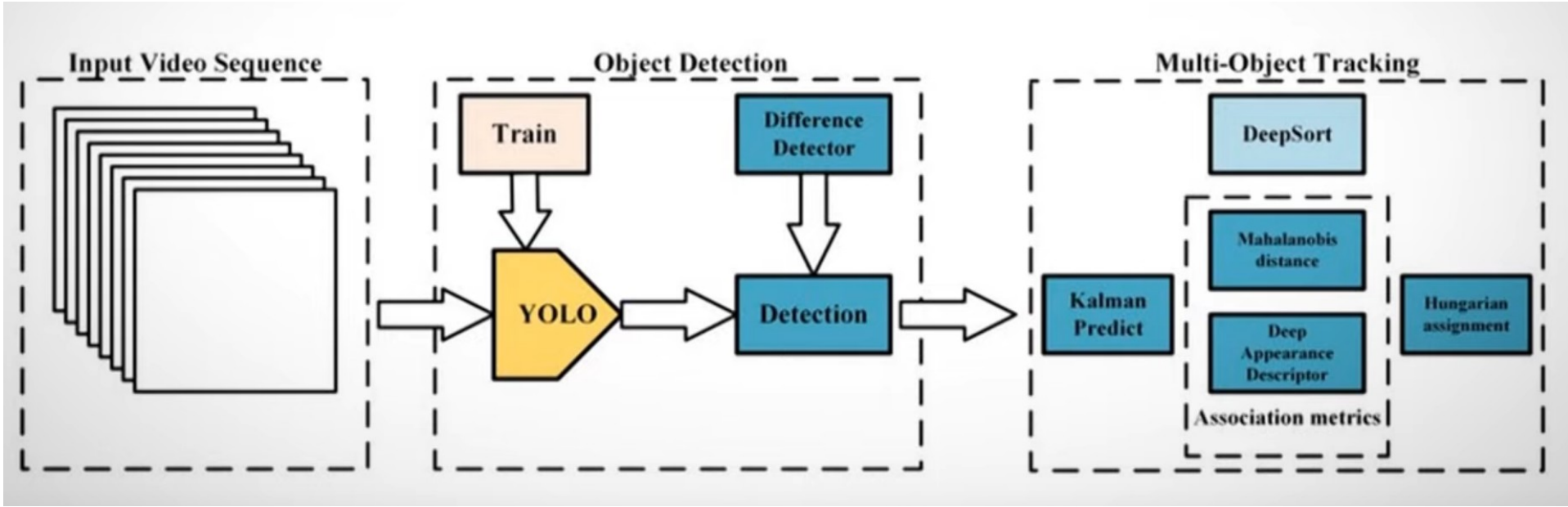
1. What is tracking:

Tracking is locating objects in successive frames of a video. Two main substreams are VOT and MOT:

(1) VOT : focus on target re-localisation

(2) MOT: focus on the data association of multiple detected targets.

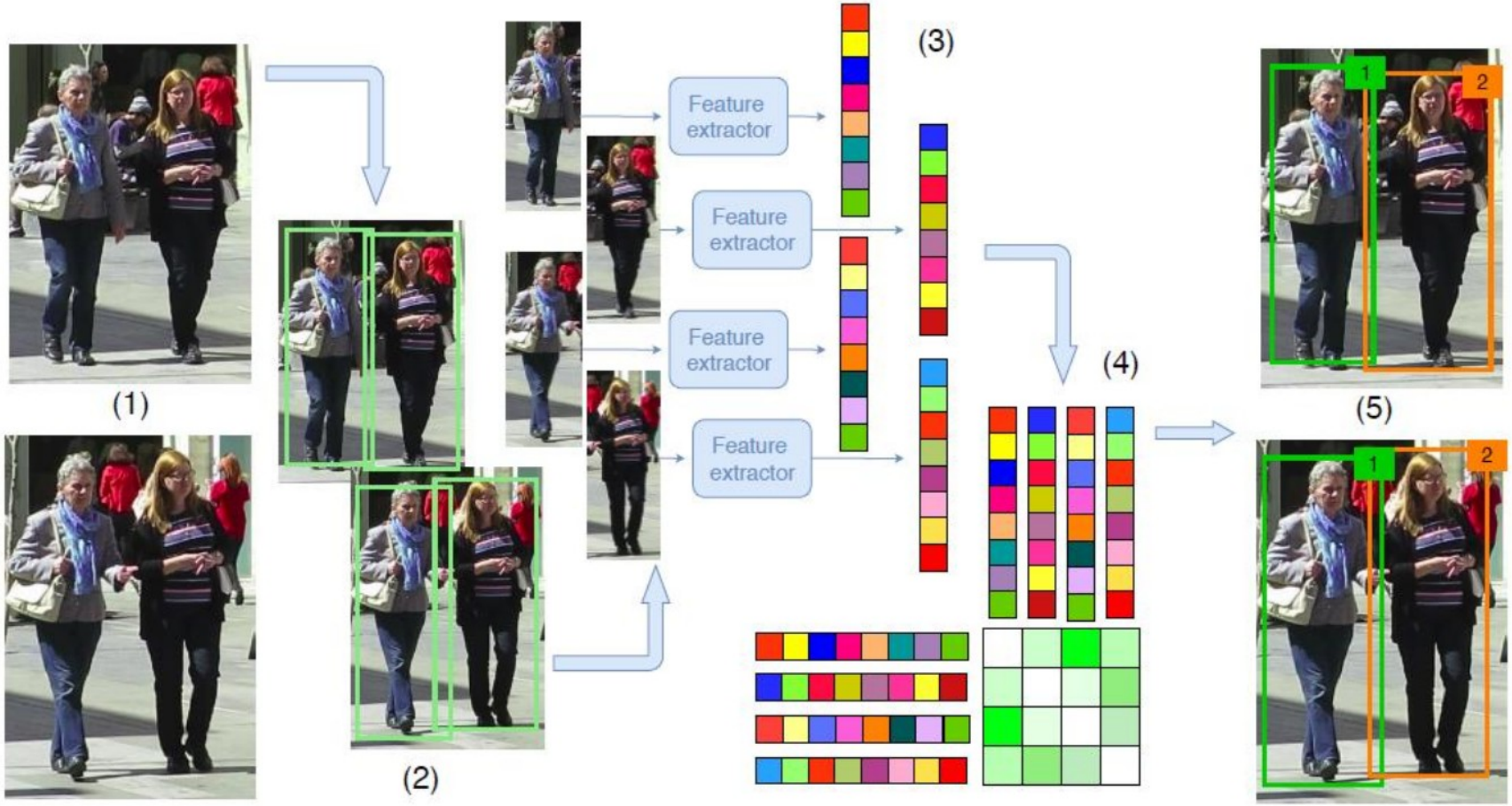
(3) Classical detection-based tracking methods：



Picture from third party

Module 2: Pedestrian Tracking Methodology

MOT



Picture from third party

Module 2: Pedestrian Tracking Demo

Action Recognition Overview

- Traditional Methods
- Deep Learning based Methods with Spatial-temporal Feature Representations

  -Two Stream

  - C3D

  - LSTM
- Skeleton-based Methods

Traditional Methods

Before the age of deep learning, traditional action recognition studies often extract local high-dimensional visual features (HOG, dense trajectories etc.), then encode these features at video-level via bag of visual words and finally train a classifier like SVM of RF on bag of visual words.

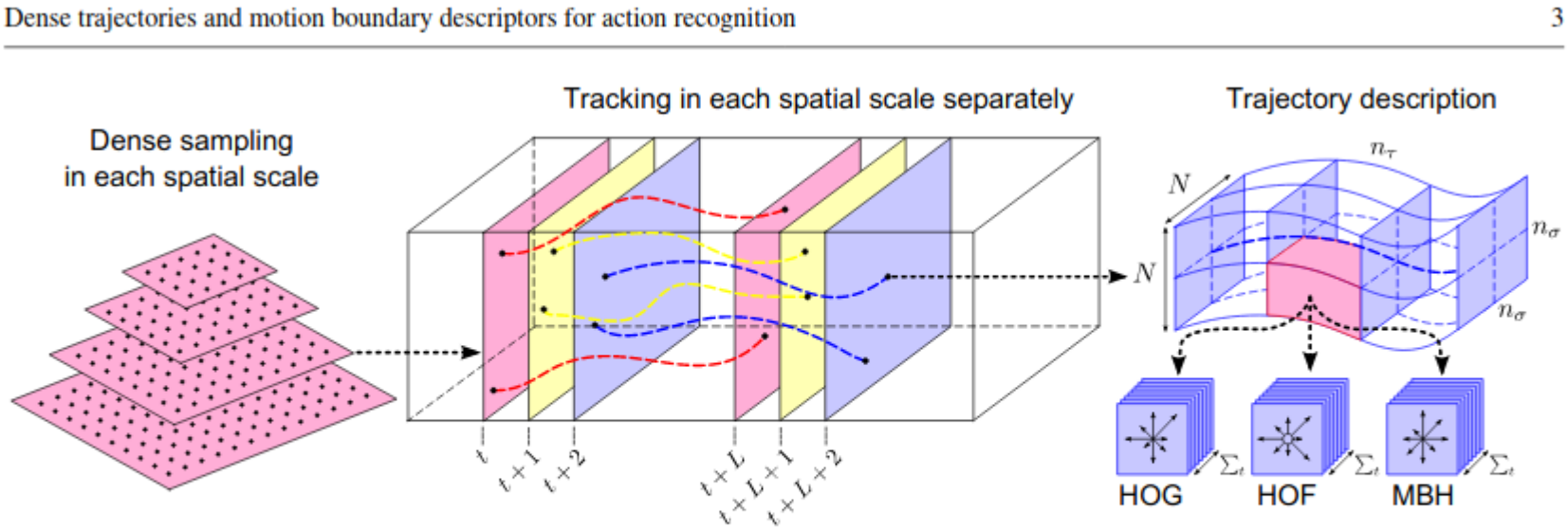


Fig. 2 Illustration of our approach to extract and characterize dense trajectories. Left: Feature points are densely sampled on a grid for each spatial scale. Middle: Tracking is carried out in the corresponding spatial scale for L frames by median filtering in a dense optical flow field. Right: The trajectory shape is represented by relative point coordinates, and the descriptors (HOG, HOF, MBH) are computed along the trajectory in a $N \times N$ pixels neighborhood, which is divided into $n_\sigma \times n_\sigma \times n_\tau$ cells.

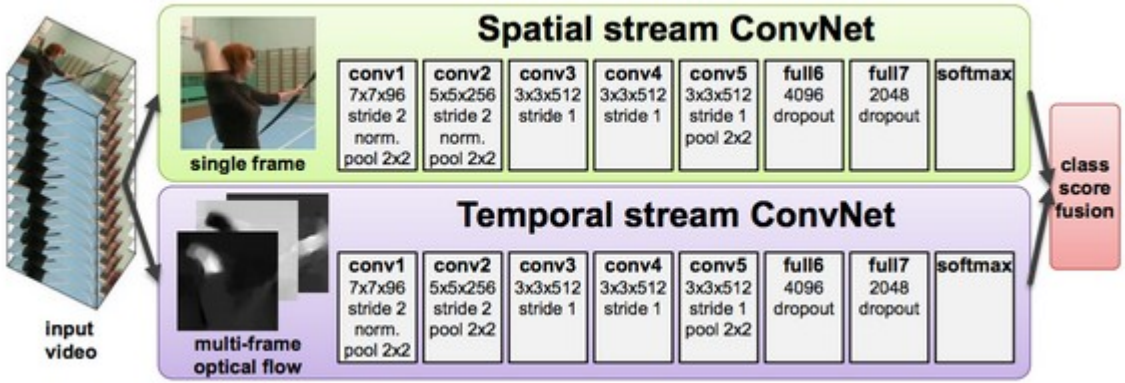
SOTA:iDT

H. Wang, A. Kl ̈aser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition.IJCV,103(1):60–79, 2013.

Deep Learning based Methods with Spatial-temporal Feature Representations: <1> Two stream approach

Two stream methods use two branches of convolution networks for feature extraction:

- a convolution network to extract spatial features from single RGB frame;
- a convolution network to extract temporal features from multiple successive frames, and usually optical flow features are used.

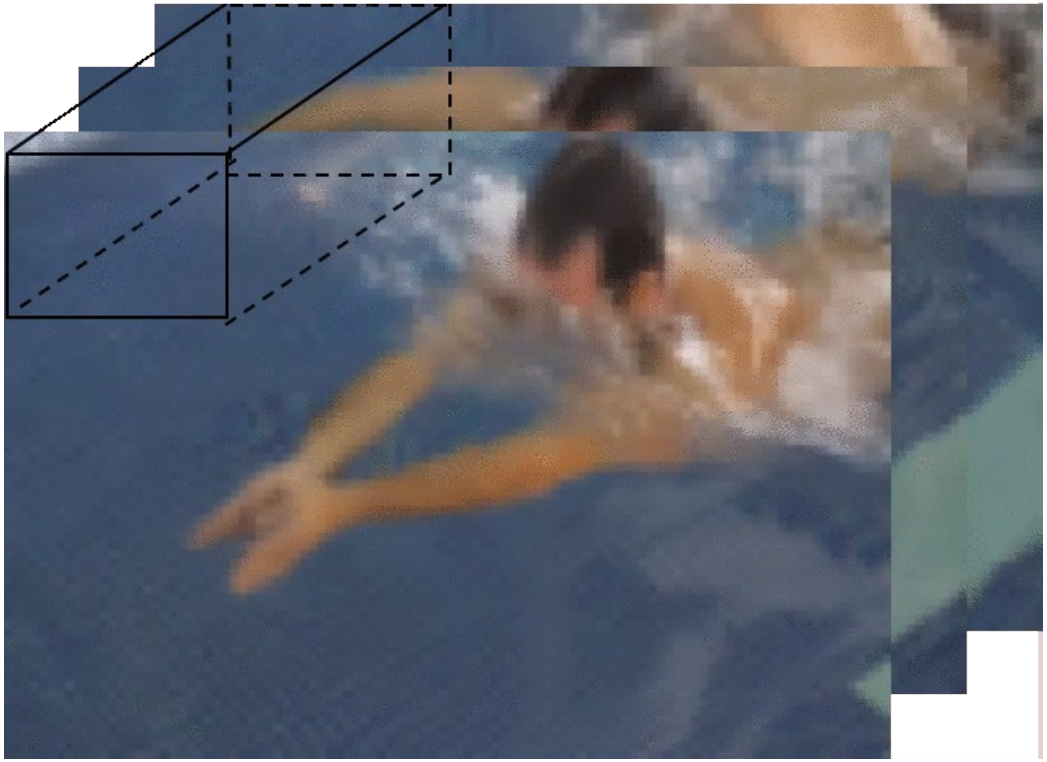


[TwoStreamCNN] K. Simonyan and A. Zisserman. Two-stream convolutional networks recognition in videos. arXiv preprintarXiv:1406.2199, 2014.

Deep Learning based Methods with Spatial-temporal Feature Representations: <2> C3D

This approach extends the 2D convolution to 3D (adding temporal dimension) to directly extract both spatial and temporal features. Classical paper:

[C3D] *Learning Spatiotemporal Features with 3D Convolutional Networks, Du Tran et al. ICCV2015.*



Deep Learning based Methods with Spatial-temporal Feature Representations: <3> LSTM and RNN based approaches

Methods in this domain uses CNN to extract spatial features while using RNN (including LSTM) to extract temporal features. ***[LRCN] Long-term recurrent convolutional networks for visual recognition and description. CVPR2015***

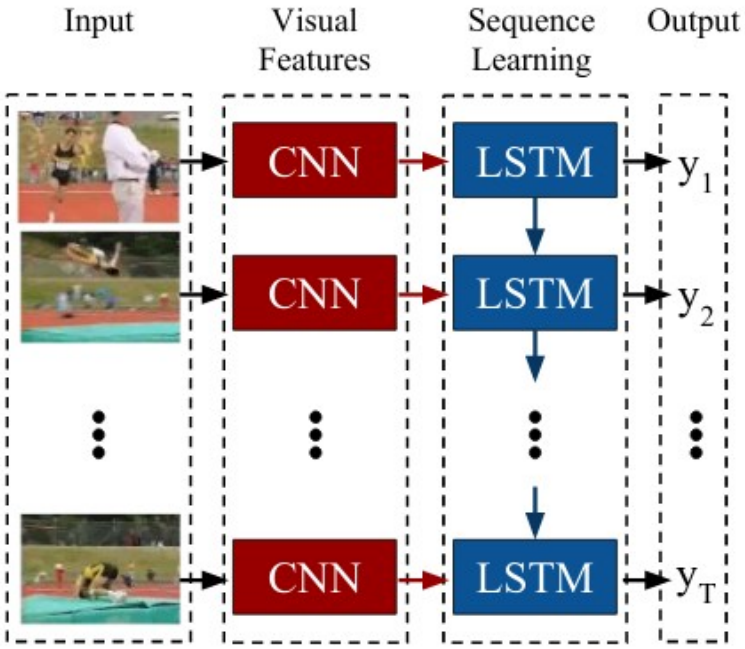
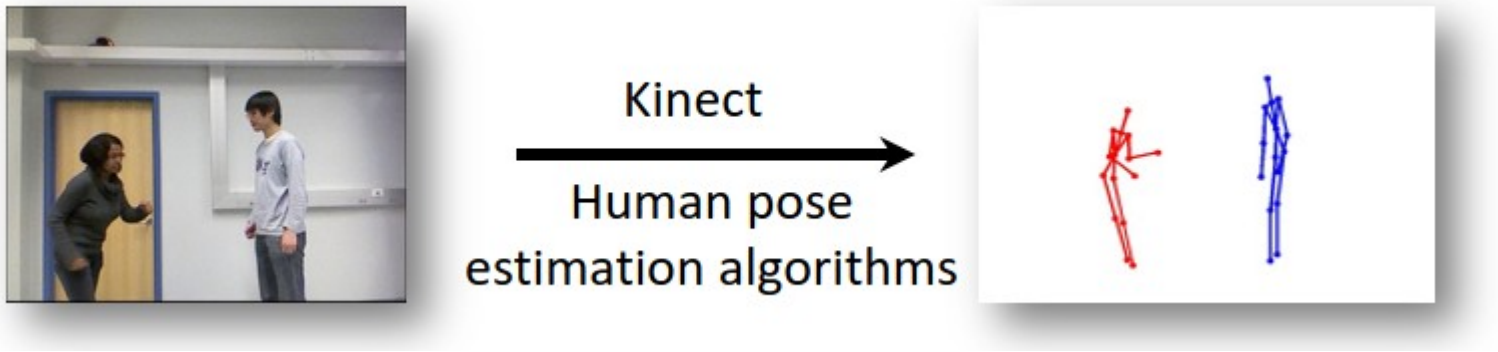


Fig. 1. We propose *Long-term Recurrent Convolutional Networks* (LRCNs), a class of architectures leveraging the strengths of rapid progress in CNNs for visual recognition problems, and the growing desire to apply such models to time-varying inputs and outputs. LRCN processes the (possibly) variable-length visual input (left) with a CNN (middle-left), whose outputs are fed into a stack of recurrent sequence models (*LSTMs*, middle-right), which finally produce a variable-length prediction (right). Both the CNN and LSTM weights are shared across time, resulting in a representation that scales to arbitrarily long sequences.

Skeleton-based Action Recognition

Input: skeleton sequence from RGB (2D), RGB-D (3D), or wearable sensors.
Output: action classification results



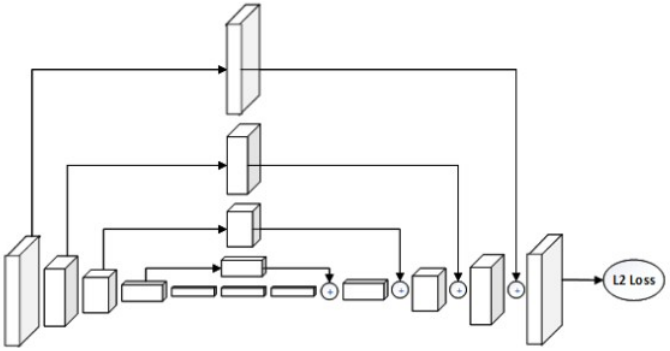
-Since skeleton sequence does not contain color information, it is not affected by the limitations of RGB video,like background clutter, illumination changes, appearance variation, etc.

-Skeleton information are higher level features, and such robust representation allows to model more discriminative temporal characteristics about human actions.

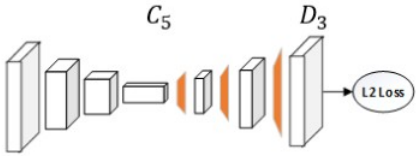
Picture from third party

Module 3: Pose Estimation

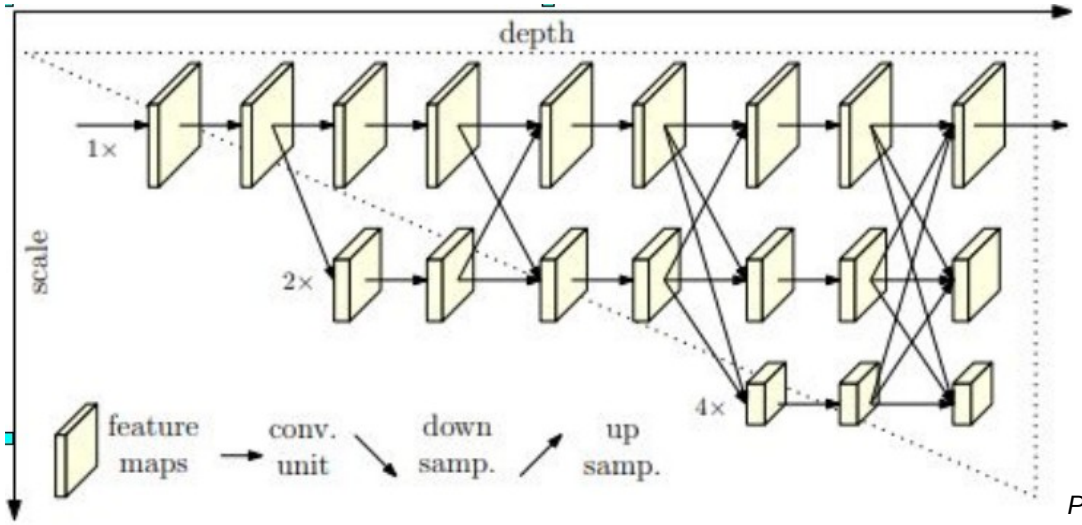
Human Pose Estimation is the problem of localization of human joints (also known as keypoints) in images or videos. Two main directions are single person pose estimation and multiple people pose estimation with either top-down or bottom-up design. We use HRNet [9] (SOTA) in our approach.



(a) Hourglass



(c) Simple Baselines



Pictures from original papers

Module 4: Skeleton-based action recognition

Main different approaches: hand crafted features, CNN-based, RNN-based, GCN-based

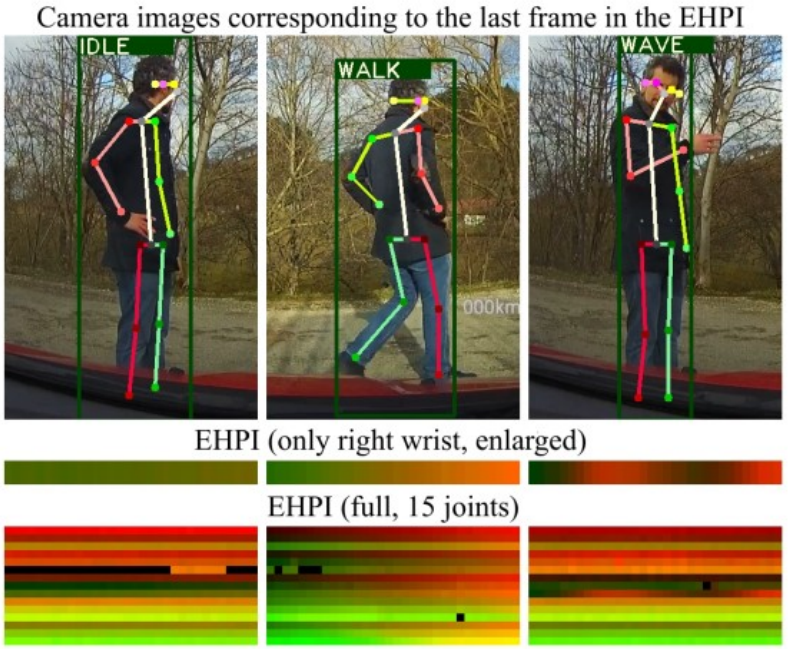


Fig. 4. EHPI examples of different actions. The example of the right wrist, which is explicitly shown at three times its height, clearly shows that a smooth color gradient is visible in the idle action, a color gradient from green to orange is visible during walking and a repetitive gradient from green to red is observable during waving.

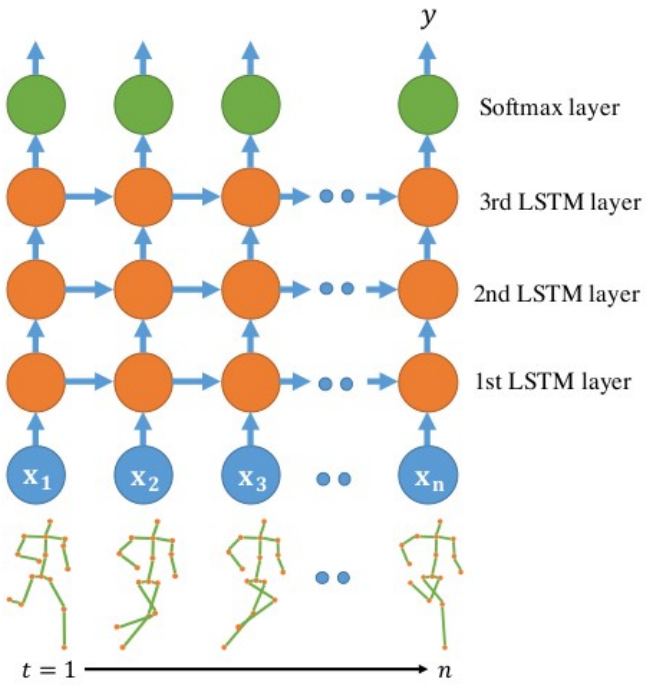


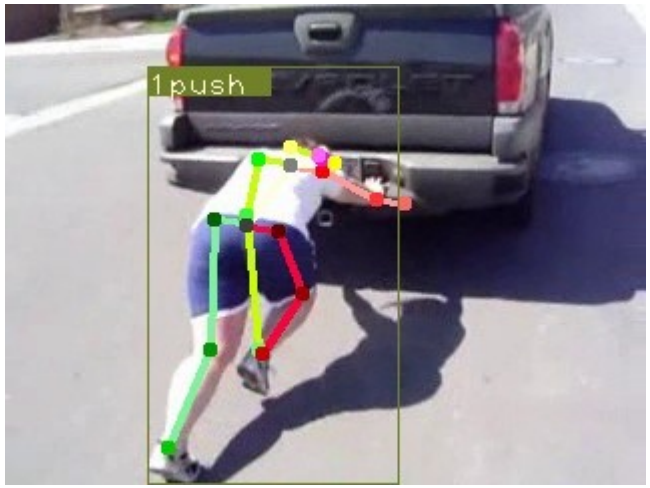
Figure 3: The LSTM architecture in our approach, where each orange dot is one LSTM layer as Fig. 2.

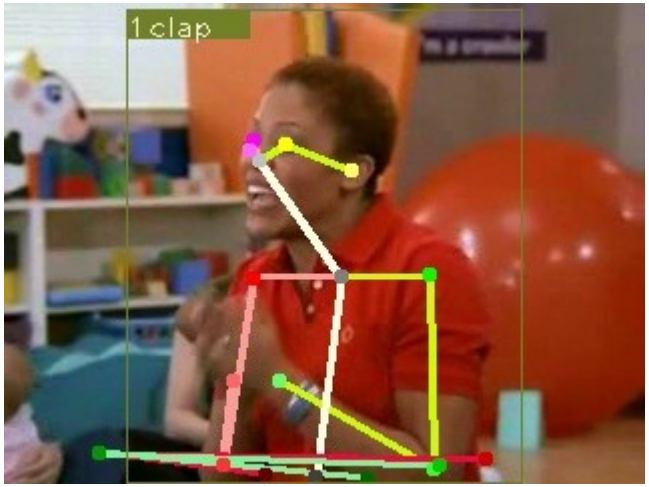
Ref:

Dennis Ludl, Thomas Gulde, and Crist'obal Curio. 2019. Simple yet efficient real-time pose-based action recognition. In ITSC.

Zhang, S.; Liu, X.; and Xiao,J.2017.On geometric features for skeleton-based action recognition using multilayer lstm networks. InWACV.IEEE.

Module 4: Skeleton-based Action Recognition Demo





Module 4: Skeleton-based Action Recognition Demo

References

[1] W Luo, X Zhao, and TK Kim. Multiple object tracking: A Literature Review. arXiv preprint arXiv:1409.7618, 2014.
[2] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in 2016 IEEE Int. Con on Image Proces., 2016, pp. 3464–3468.(SORT)
[3] R. E. Kalman. A new approach to linear filtering and predic-tion problems.Transactions of the ASME–Journal c BasicEngineering, 82(Series D):35–45, 1960.
[4] N. J. Gordon, D. Salmond, and A. Smith. Novel approachto nonlinear/non-Gaussian Bayesian stat estimation.IEEProceedings F (Radar and Signal Processing), 140:107–113,1993.
[5] Gioele Ciaparrone, Francisco Luque S´anchez, Siham Tabik,Luigi Troiano, Roberto Tagliaferri, and Francisc Herrera.Deep learning in video multi-object tracking: A survey.CoRR, abs/1907.12740, 2019
[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards real-time object detection wit region proposal networks.” In Advances in neural information processing systems (NIPS), pp. 91-99. 2015.
[7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander CBerg Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer,2016.
[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time objec detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.
[9] https://cv-tricks.com/object-detection/faster-r-cnn-yolo-ssd/
[10] H. W. Kuhn. The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly, 2:83–97,1955
[11] Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric CoRRabs/1703.07402(2017)
[12] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and patter recognition, pages 7263–7271, 2017.
[13] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
[14] Hourglass: A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.
[15] B. Xiao, H. Wu, and Y. Wei, Simple baselines for human pose estimation and tracking, in ECCV, 2018
[16] HRNet: Sun, K., Xiao, B., Liu, D., & Wang, J. Deep high resolution representation learning for human pose estimation. In CVPR, 2019

Questions and Discussion