Case Studies
Smart Sport with AI

Chia-Chi Tsai (蔡家齊)
cctsai@gs.ncku.edu.tw
AI System Lab
Department of Electrical Engineering
National Cheng Kung University

Outline

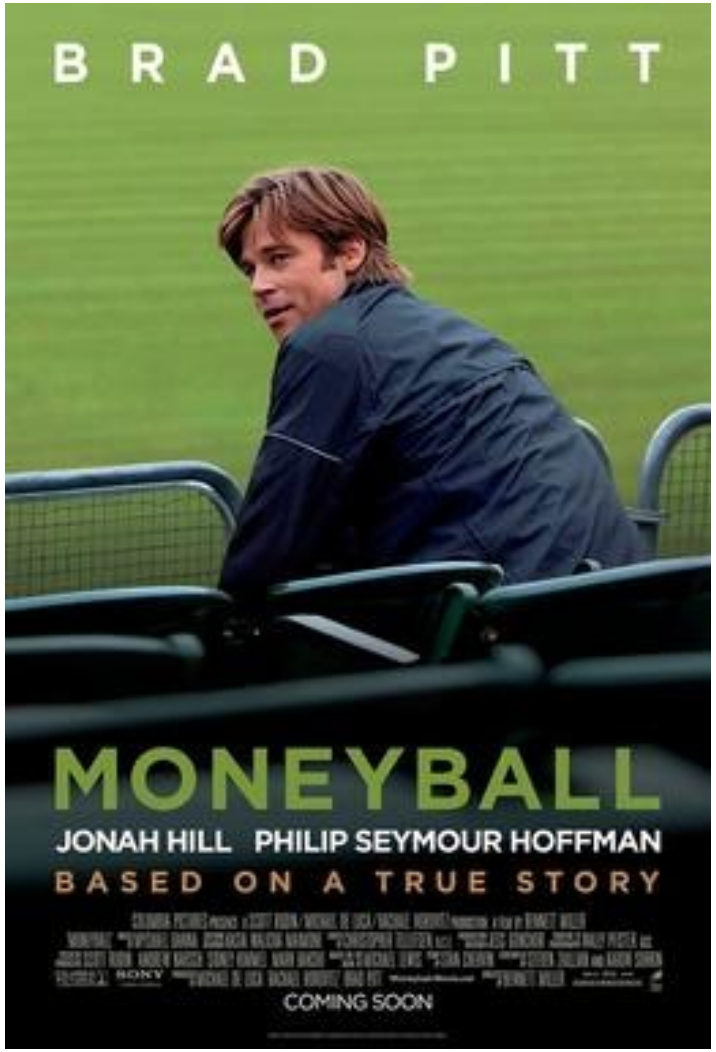
- Overview of Smart Sport
- Pose Estimation
- Case Studies

Outline

- Overview of Smart Sport
- Pose Estimation
- Case Studies

Sabermetrics

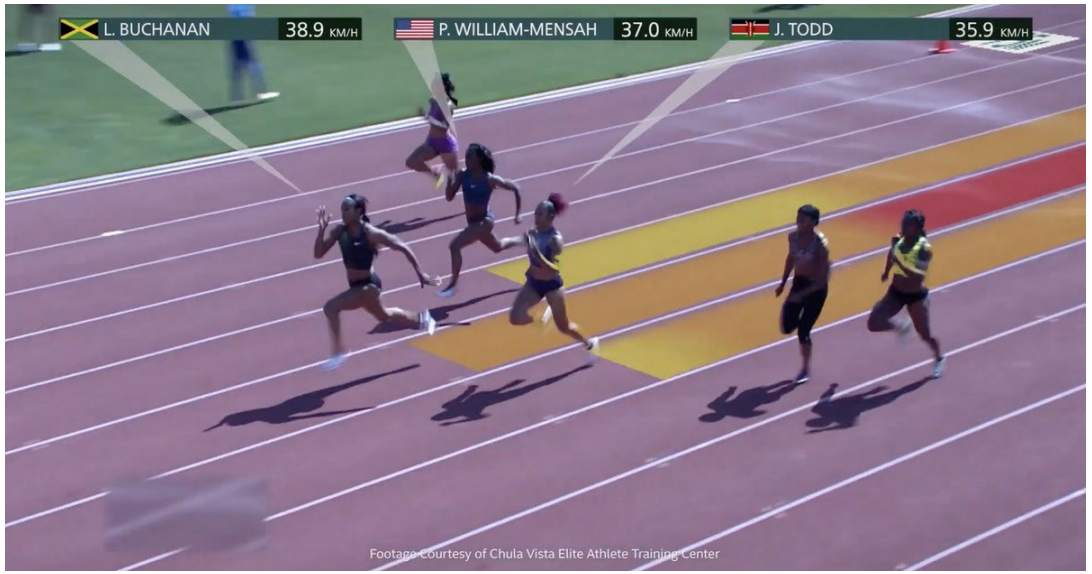
- Moneyball
  - General manager Billy Beane, focus on an analytical, evidence-based, sabermetrics approach to assembling competitive baseball team in 2002
- Before sabermetrics
  - Dependent on the skills of their scouts
  - Speed, quickness, arm strength, hitting ability and mental toughness
- Sabermetrics
  - Empirical analysis of baseball statistics that measures in-game activity
  - Runs Created, Extrapolated Runs, Park Factor

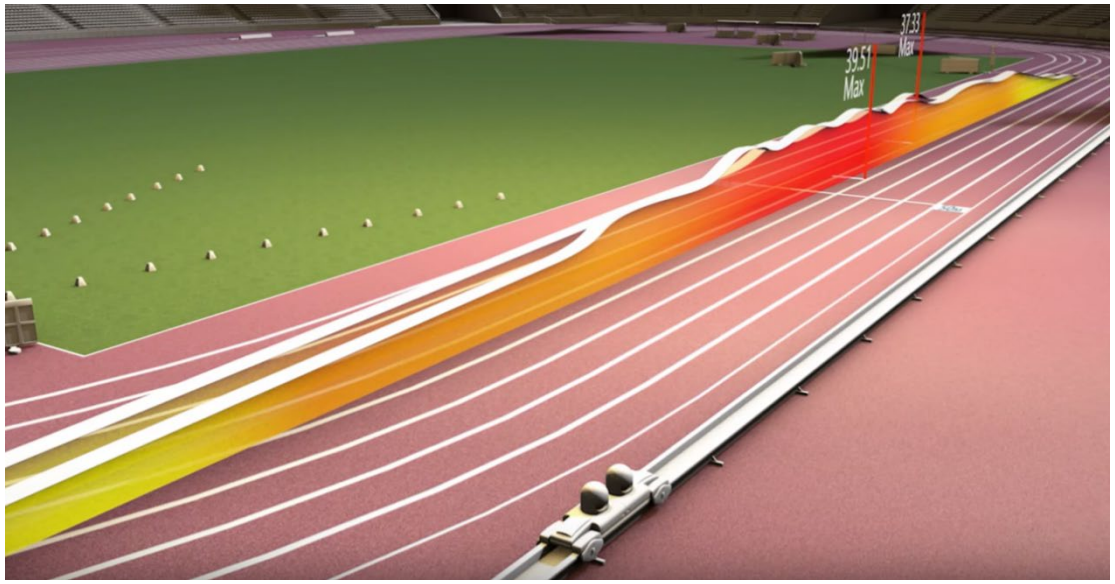


Smart Sport


Strength and Conditioning
Performance Psychology
Research and Development
Medicine
Sport Science
Performance Analysis
Biomechanics
Performance Nutrition
Physiology

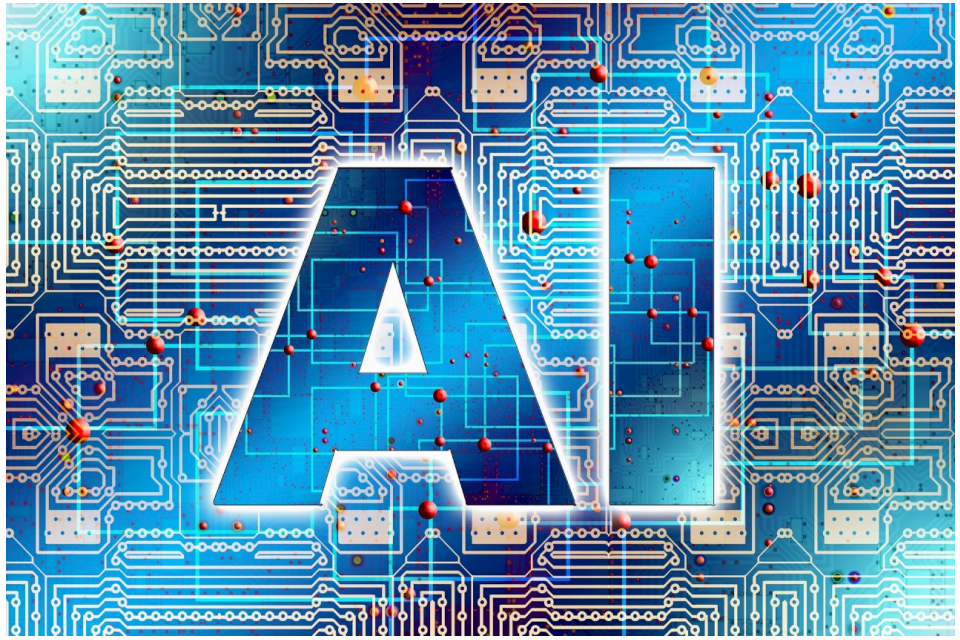
Athlete/Team Live or Statistic Analysis





What is Artificial Intelligence ?

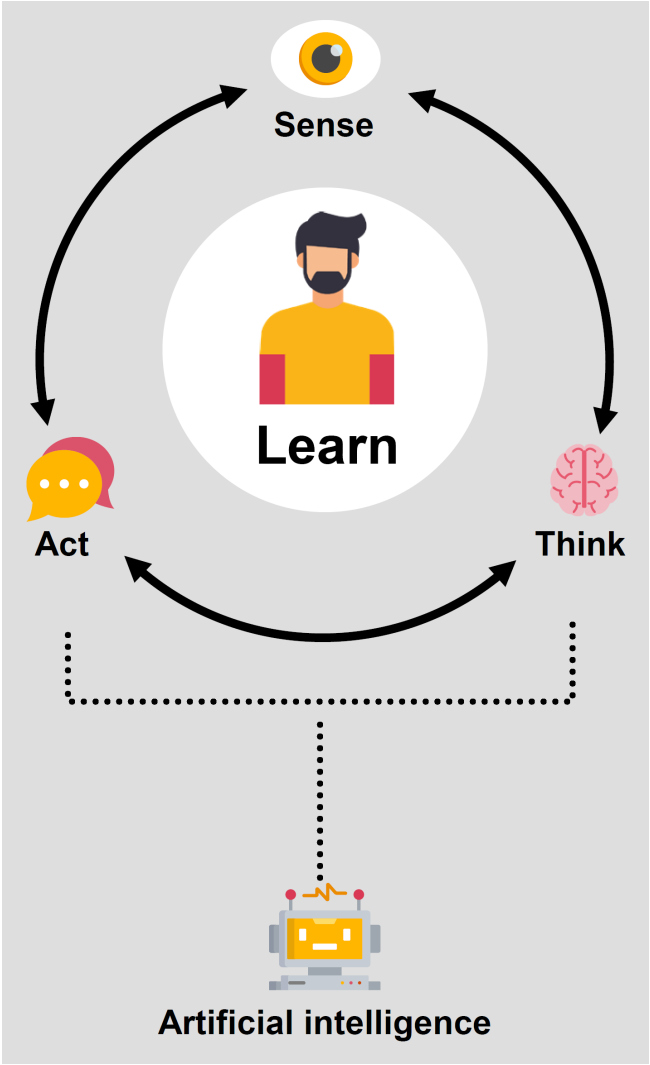
- Covering a variety of **Smart Technologies**
- Everyday practical AI
  - Learning, growing, making decisions
- Higher level of AI
  - Self-driving car, drones
- Four basic AI
  - Automated intelligence
    - Perform automated tasks
  - Assisted intelligence
    - Assist with better decision
  - Augmented intelligence
    - Automate decision-making process
  - Autonomous intelligence
    - Mimic human's ability to **Sense**, **Think** and **Act**



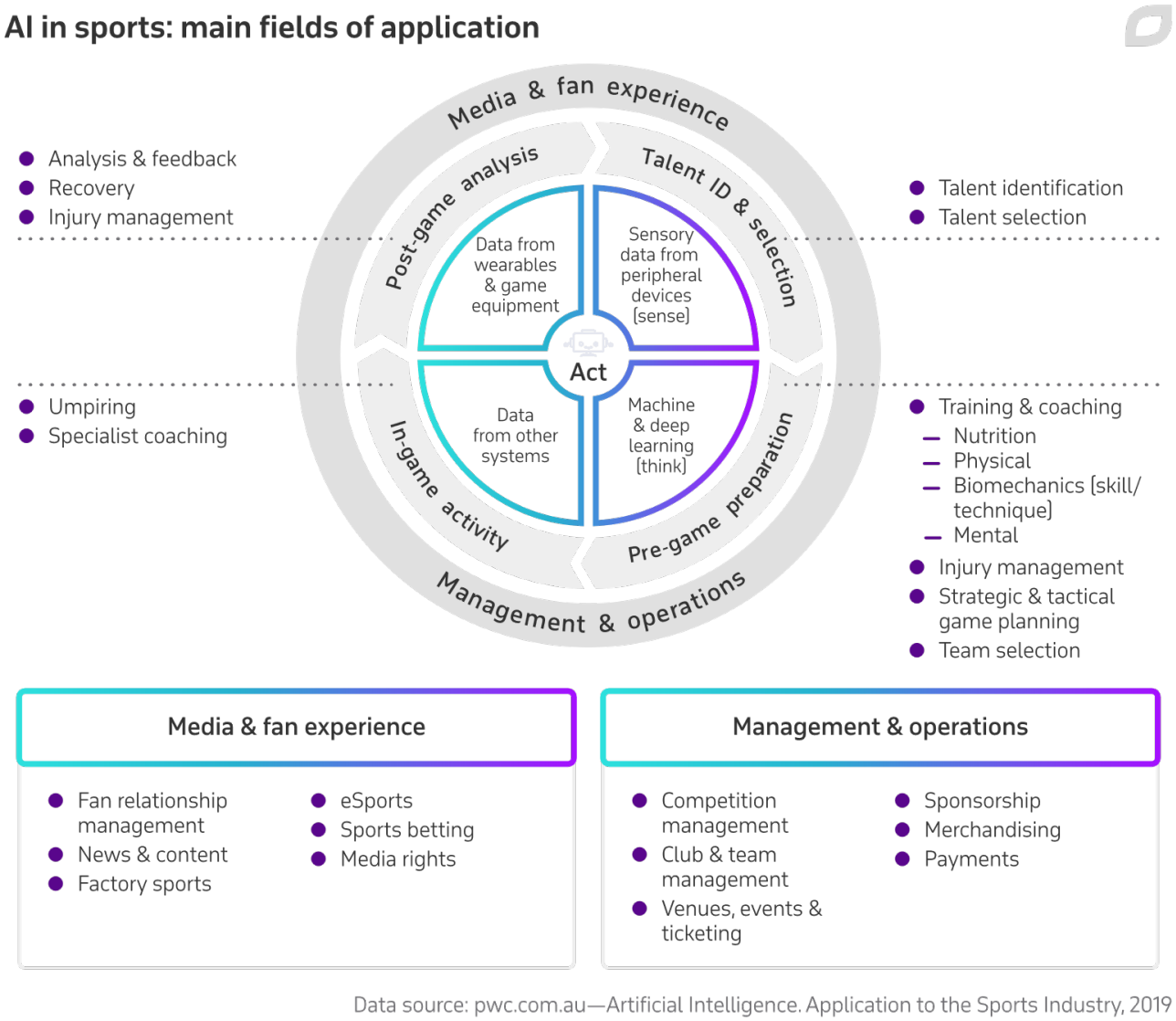
AI - Sense, Think and Act

- Sense: sensory AI
  - Image and video analysis
  - Facial recognition
  - Speech analytics
  - Text analytics
- Think: cognitive AI
  - Machine learning methodology
  - Reinforcement learning
  - Decision making
- Act: executable AI
  - Natural language generation
  - Image synthesis

- Sense, think, and act: artificially intelligent solutions
  - AI-enhanced analytics solutions
  - Conversational service solutions
  - Intelligent research solutions
  - Intelligent recommendation solutions
  - Pretrained vertical solutions

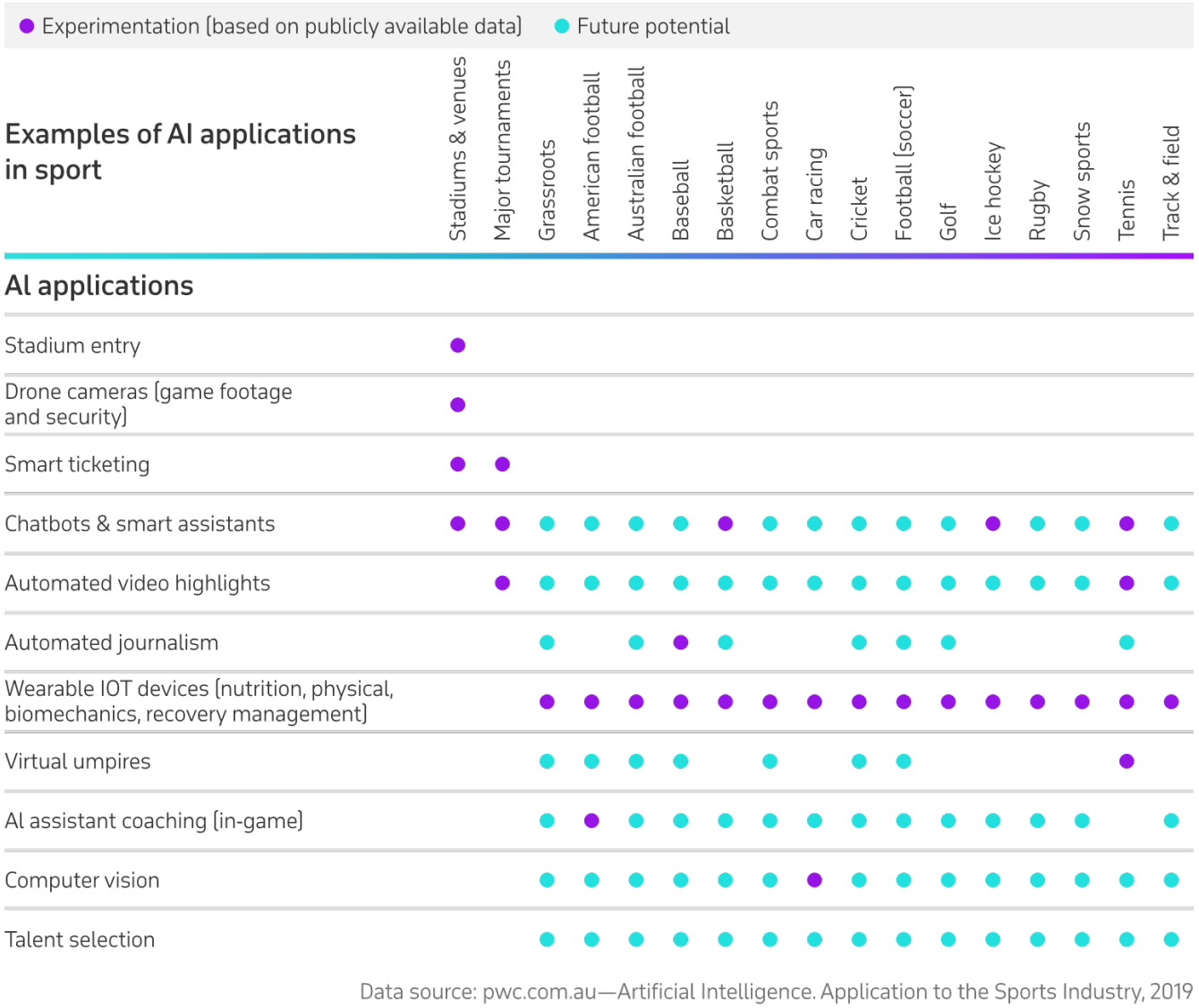


AI Technology for Sports



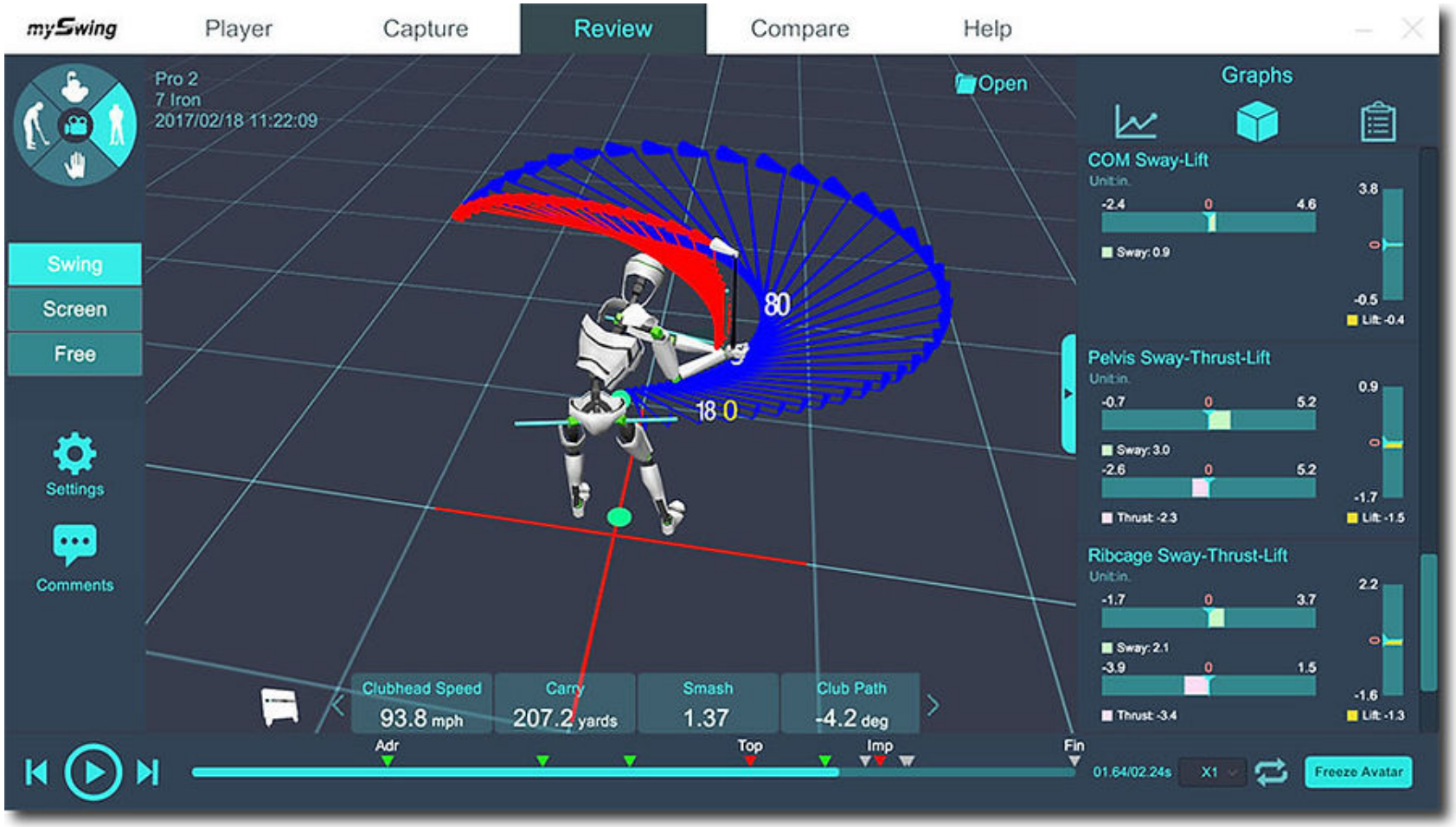
Data source: pwc.com.au—Artificial Intelligence. Application to the Sports Industry, 2019

AI-based Technologies in Sports



Data source: pwc.com.au—Artificial Intelligence. Application to the Sports Industry, 2019

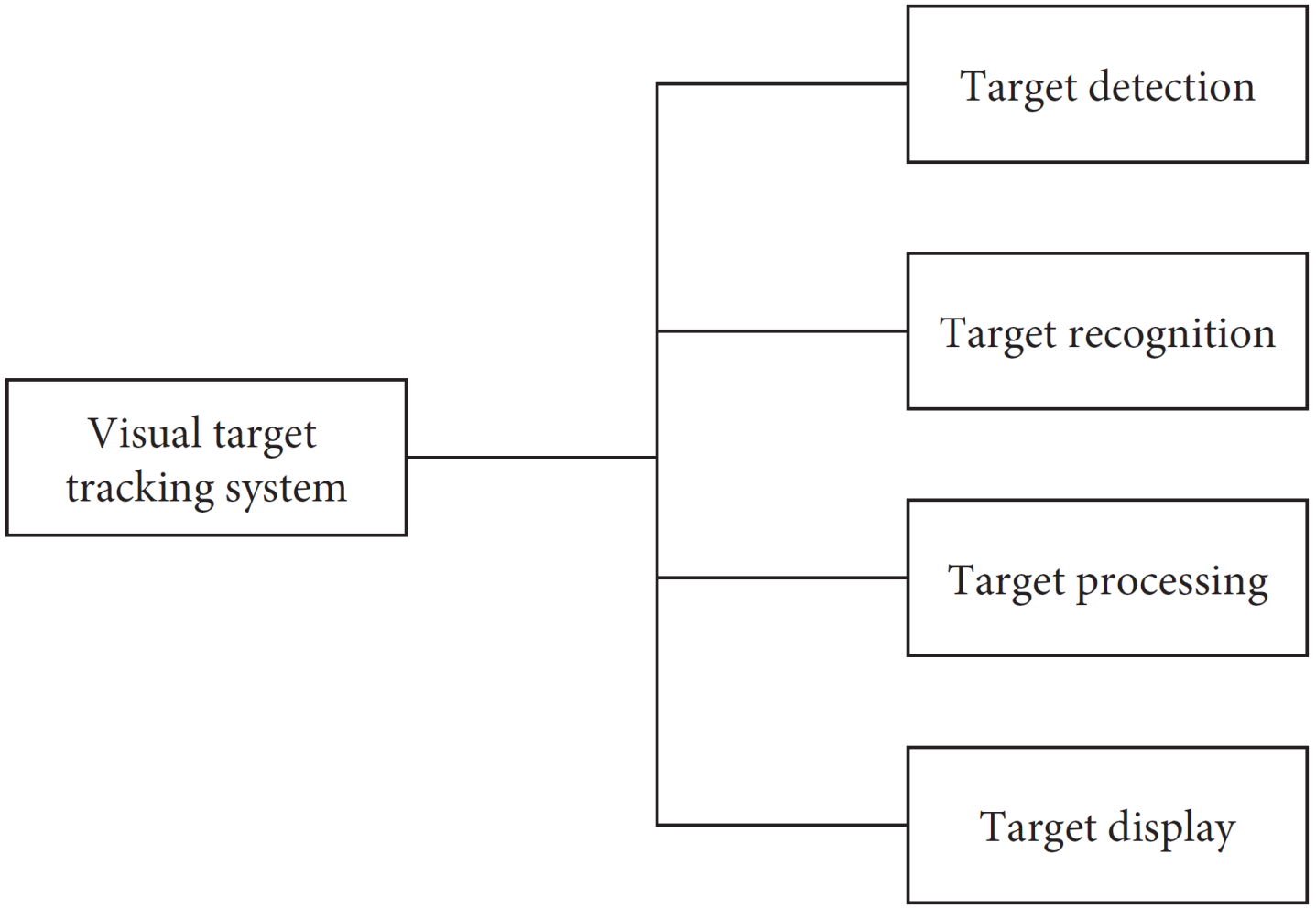
MySwing Professional



MySwing - Wearable AI Devices

- Motion Capture
  - Full-body wireless motion capture suit
  - Record the player’s movements in precise details
- Playback Analysis
  - Analyze the angle and acceleration of their swings through playback
  - 3D model, analysis chart, and other auxiliary reference tools
- Storage Comparison
  - Upload a golfer’s movement data from every session to a cloud-based server
  - Keep track of their levelchanges through the data and compare it to the movement data of professional golfers

Visual Target Tracking System



SportVU

SportVU

- Three high-definition cameras or sensors are placedat the console, each recording one-third of the playing field.
- The live video is recorded into the computer system, where the lines of sight of the three cameras converge to form the playing field
- Any object on the field (player, referee, and ball) appears as a dot on the operator's computer screen.
- Track players' movements through these dots

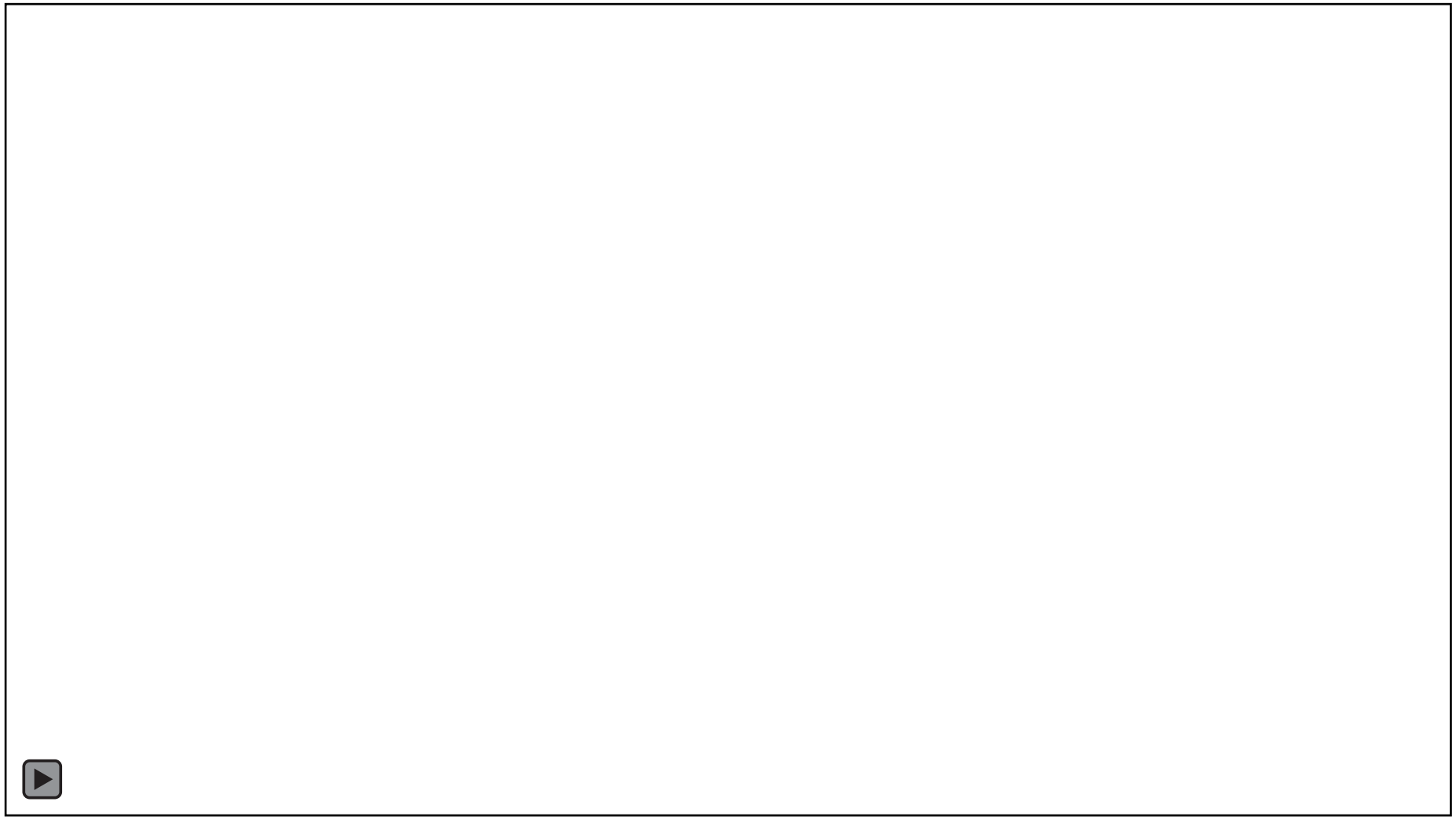
Virtual Reality Technology

- Building Virtual Training Scenarios
  - Virtual training scenarios and virtual training equipment need to be modeled
  - Simulate the real training scenes and provide the most basic virtual training scenes for athletes
- Capturing Motion Data
  - Capture data from real human movements
  - Wearable tracking devices or vision tracking algorithm
- Collection of Physiological and Psychological Data
  - Important reflections of the athletic state of the athlete
  - Pulse rate, blood pressure, athletes' mood, etc.
- Action Replay
  - Create a virtual world with rich visual and auditory information
  - Immersive and non-immersive systems
  - Experience realistic stereo vision and stereo hearing to interact with the virtual environment

QB SIM

- Built in a real rugby stadium
- Combining OptiTrack motion capture technology and VR virtual reality technology

Motive + OptiTrack



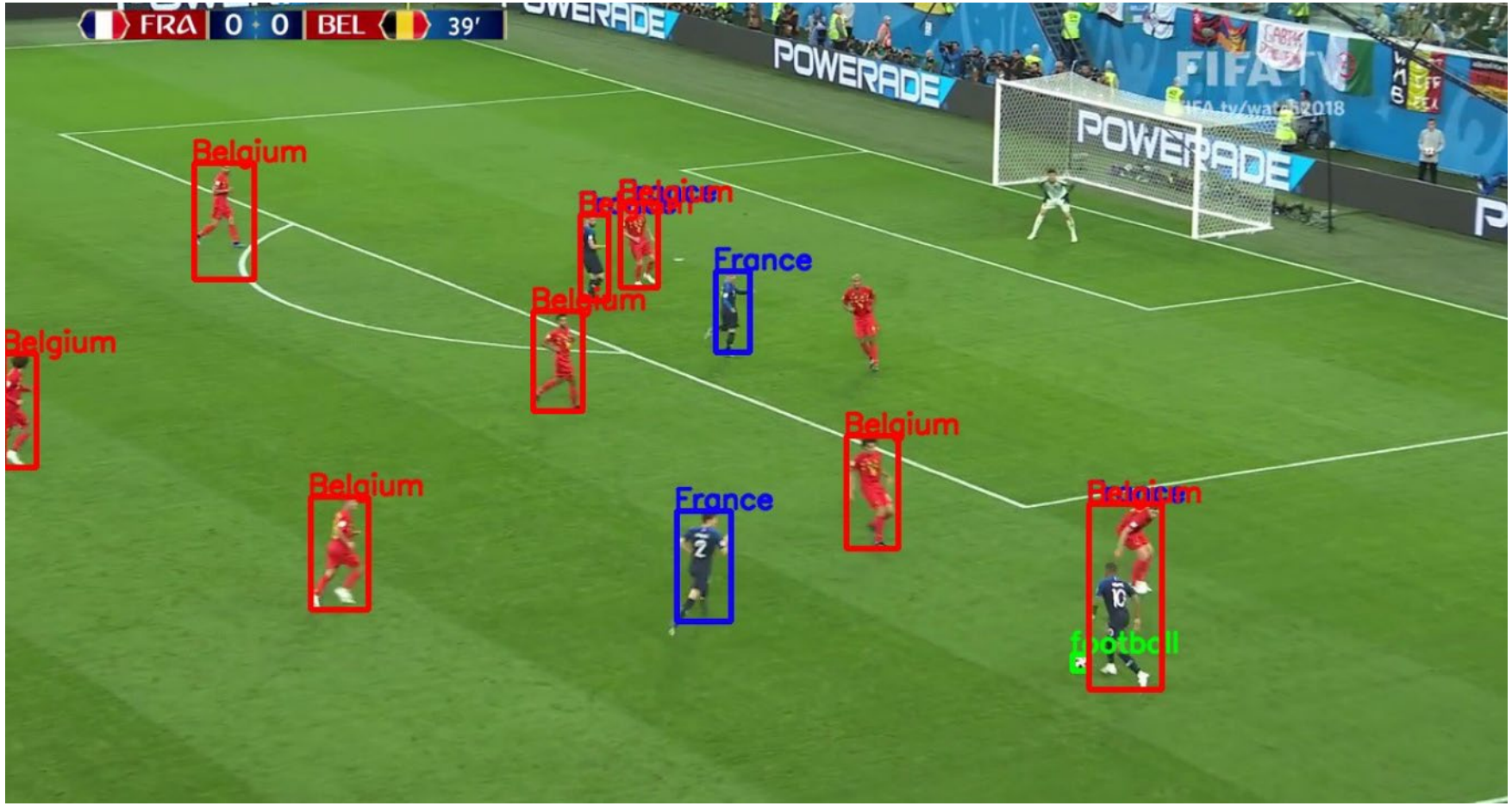
Intel's 3DAT


x velocity
6.90
m/s
sway
0.00
m
angle of attack
46.18
deg
avg stride length
2.36
m
avg stride freq
3.52
strides/s
Minimize Small
Video Graph Metrics
1
2
3
4
intel
00:06:27
376/750 frames

AI Sensing Technologies

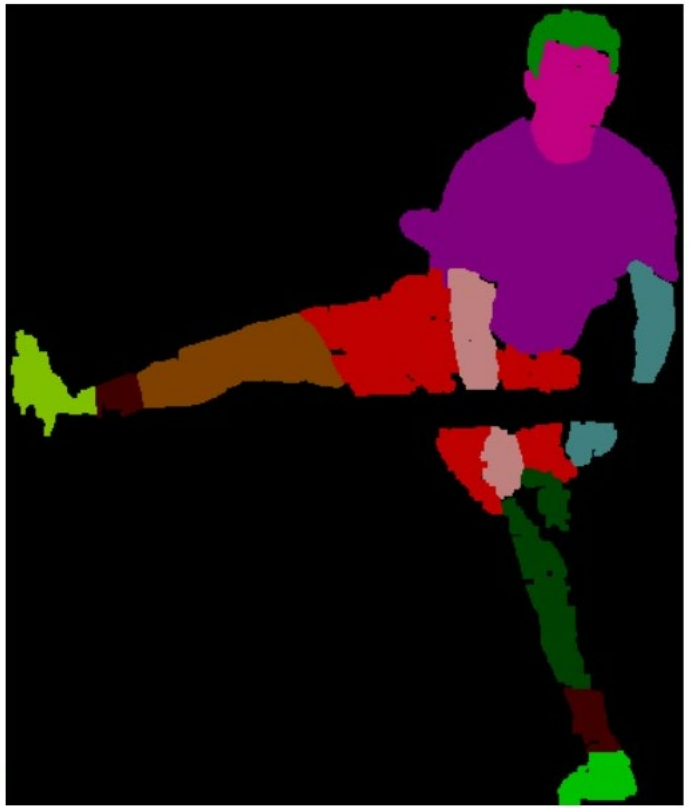
- Human Detections
- Instance Segmentations
- Human Parsing
- Pose estimation
- Body mesh Estimation
- Human Tracking

Human Detection

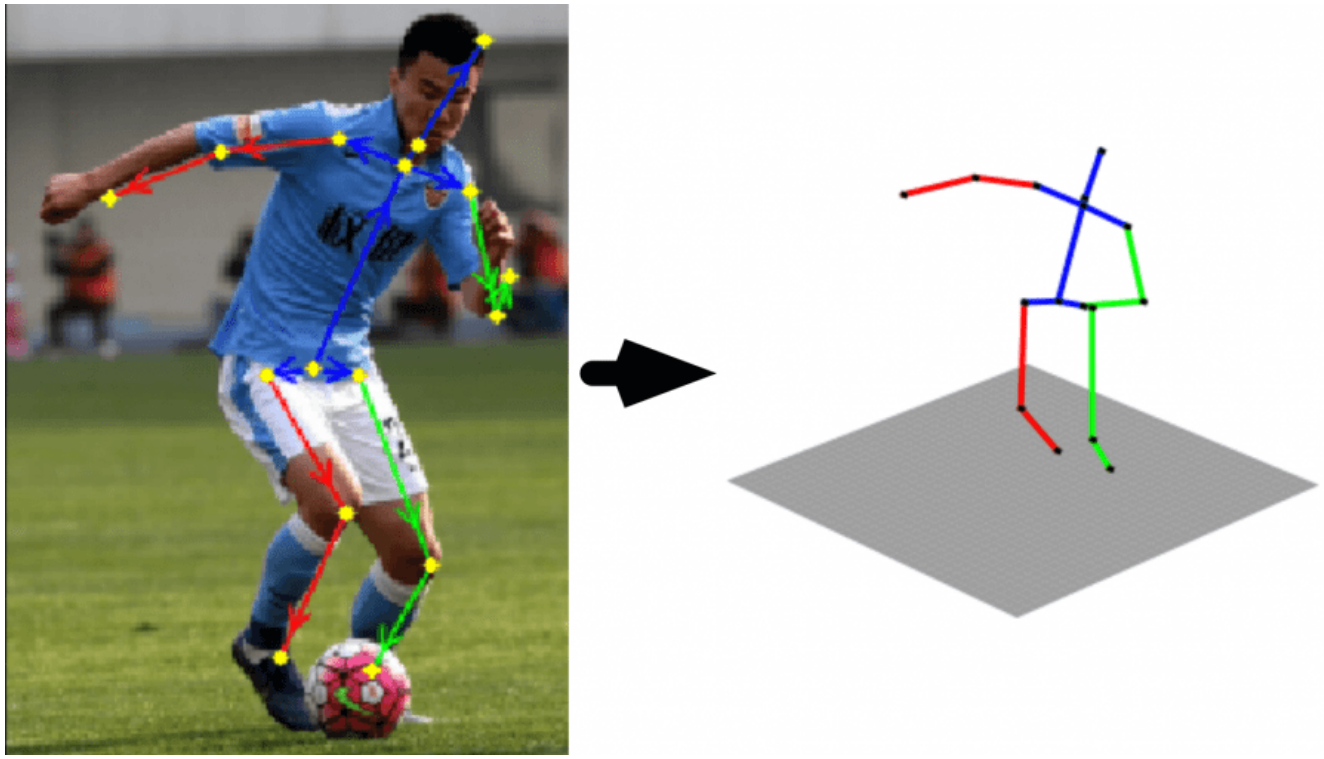


Instance Segmentation

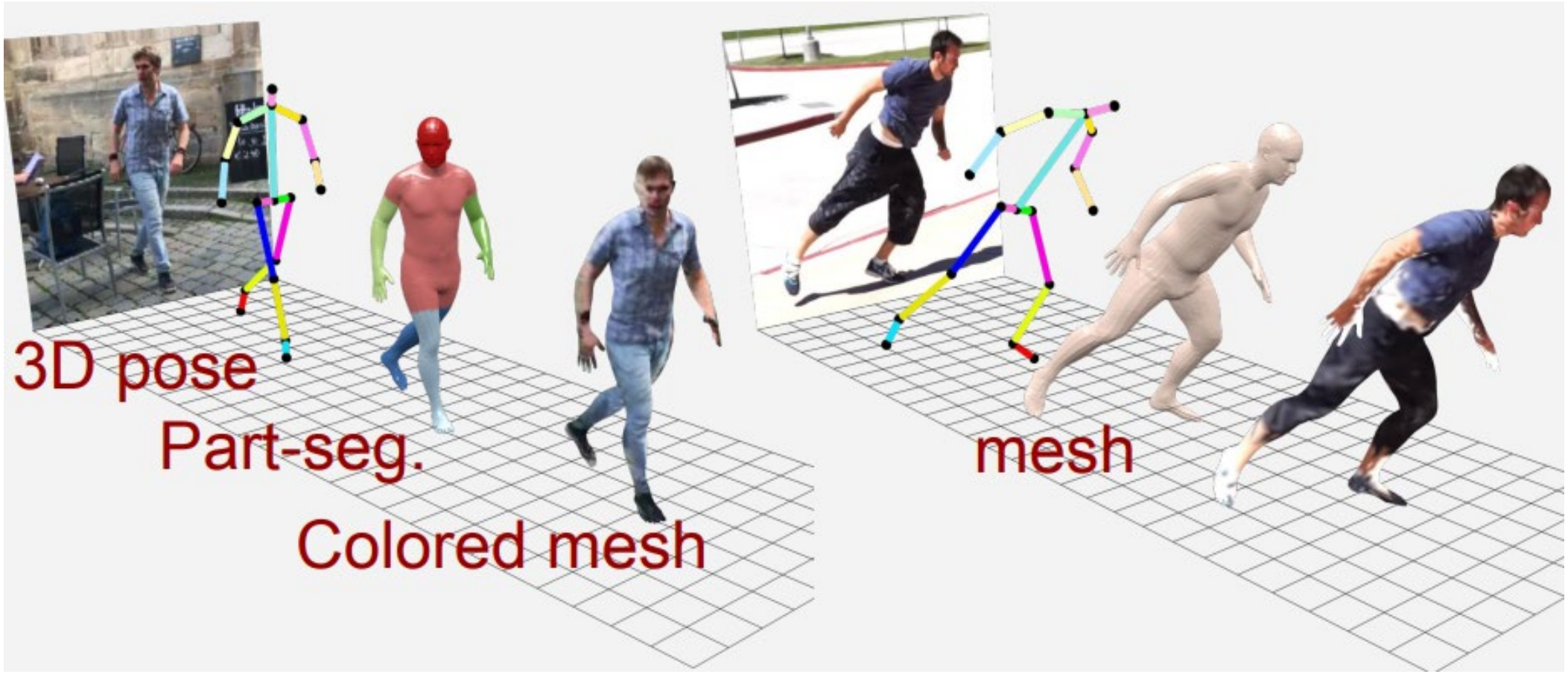
Human Parsing



Pose Estimation



Body Mesh Estimation



Human Tracking

Media and the Fan Experience

- Chatbots & Smart Assistants
  - NBA and NHL use virtual assistants to respond to fan inquiries
    - Includes live game information, team stats, ticketing, parking and arena logistics
  - Wimbledon have overlayed Augmented Reality (AR) features within their Chatbot technology
    - Identify players, key statistics and hotspots in real-time
- Video Highlights
  - Wimbledon used IBM Watson to assist production teams to create video highlights
    - Analyses player emotion, movement, and crowd noise in order to determine the most interesting and must-see moments
- Automated Journalism
  - Translate hard data from baseball into narratives
  - Providing content to fans including coverage over the minor leagues

Before the Game - Training & Coaching

- AI application

Current knowledge on a particular technique or game tactic = A huge number of training and game performances data + Knowledge from several expert sports scientists and coaches

As body of knowledge on the particular technique and tactic grows and develops, knowledge base of the AI application will be updated

Before the Game – Analysis and Tracking

- Apps and wearable devices
  - Physical activity analysis
  - Physical statistics
  - Physical stat tracking
- Beyond tracking with AI and ML
  - Performance evaluation
  - Training recommendations
  - Nutrition suggestions
- From elite level down to grassroot users
  - Cheaper and more accessible

In-Game – Virtual Umpires

- When dealing with umpire uncertainty
  - Slow motion reply or Hawk-eye
  - Decision Review System (DRS) and Video Assistant Referee (VAR)

- AI involved umpire
  - E.g. line umpires -> speed, placement of tennis shots
  - Eliminate time spent on reviewing decisions

- Real Umpire
  - Towards on-field player behavior management

However, do fans really want the perfect decision making in real-time?

In-Game – AI Assistant Coaches

- How AI can improve overall performance for sport teams?
  1. With huge number of video data and statistical data
  2. Analyzing common mistakes
  3. Improve game strategies and tactics

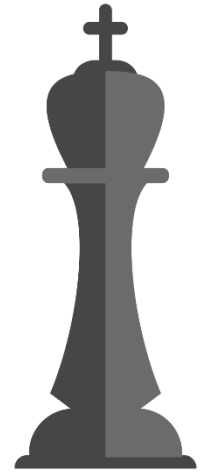
  Keep improving with more data collected and more game played

- Moreover
  - Measuring and predicting player contribution throughout the game
  - Dynamically altering game plans based on what is happening on-field
  - Uncover strategic insights that may not have been previously achievable

- What's the jobs for real coach
  - Communicate and enforce the decision with players
  - With added emotions

Beyond Physical Sports - Man vs. Machine

It took
47 years for the computer to win a match
67 years for the computer to be unbeatable



1950 - 1979

1950
First chess program is written

1967
First chess computer to play in a tournament

1970
The first all computer championship was held in New York

1977
The first microcomputer chess playing machine created

1980 - 1989

1980
Edward Fredkin created the Fredkin Prize for Computer Chess. The award came with $100,000 for the first program to beat a reigning world chess champion.

1981
First computer to beat a chess master in tournament play and gain a Master rating (Elo rating: 2,258)

1983
First microcomputer to beat a chess master in tournament play

1985
Kasparov played 15 of the top chess computers and won every game (32-0)

First computer to gain an Elo rating greater than 2,400 (2,530)

1988
First grandmaster to lose to a computer in a major tournament

First chess computer to gain a Grandmaster rating

1989
Deep Thought the best computer at the time is defeated by Kasparov

1990 - 1997

1996
Deep Blue wins one game against Kasparov but is defeated overall (first computer to win a game against a reigning world champion under regular time controls)

1997
Deep Blue is upgraded and defeats Kasparov in a six-game match (first computer to win a match against a reigning world champion under regular time controls).

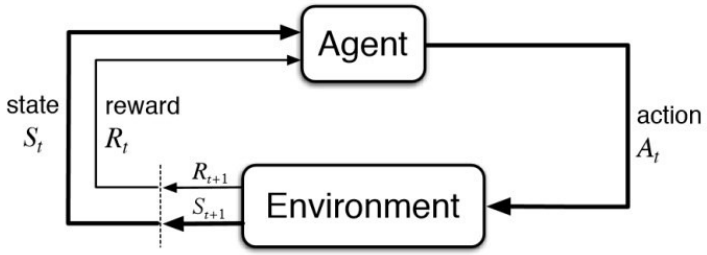
The $100,000 Fredkin Award went to the inventors of Deep Blue

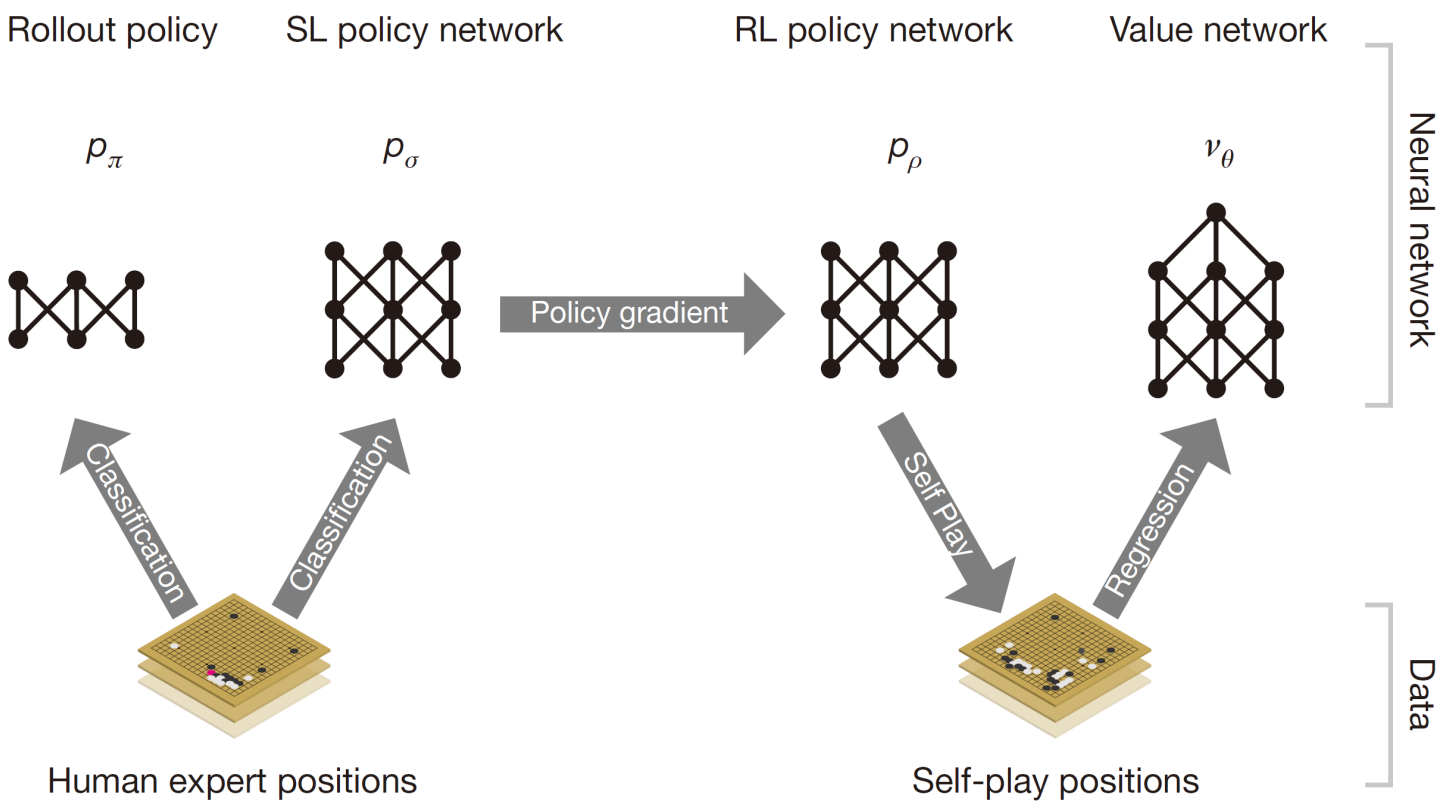
2017 - current

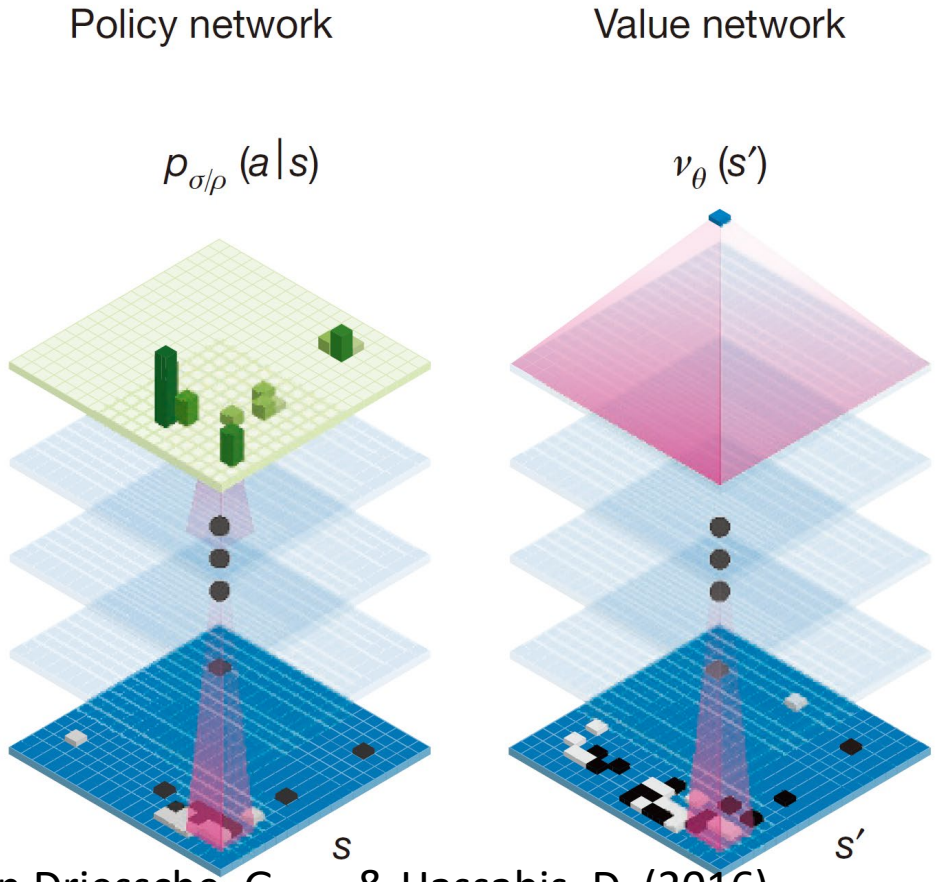
2017
AlphaZero is released and has been unbeatable (Elo rating: 3,600)

AlphaGo – Exceed Human Expert

AlphaGo - Mastering the Game of Go

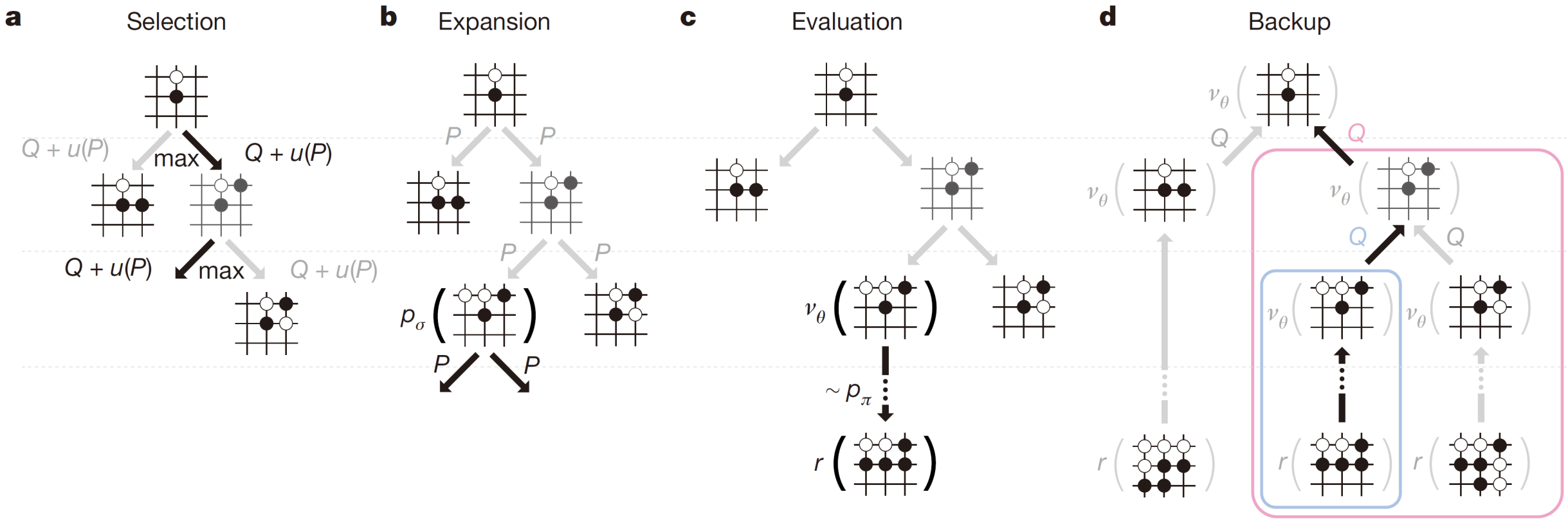






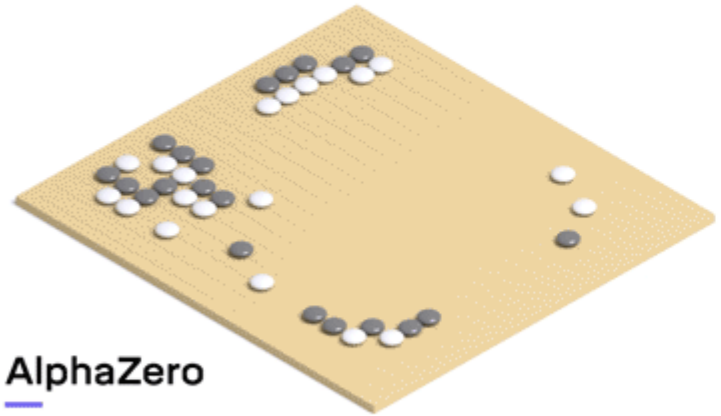
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *nature*, *529*(7587), 484-489.

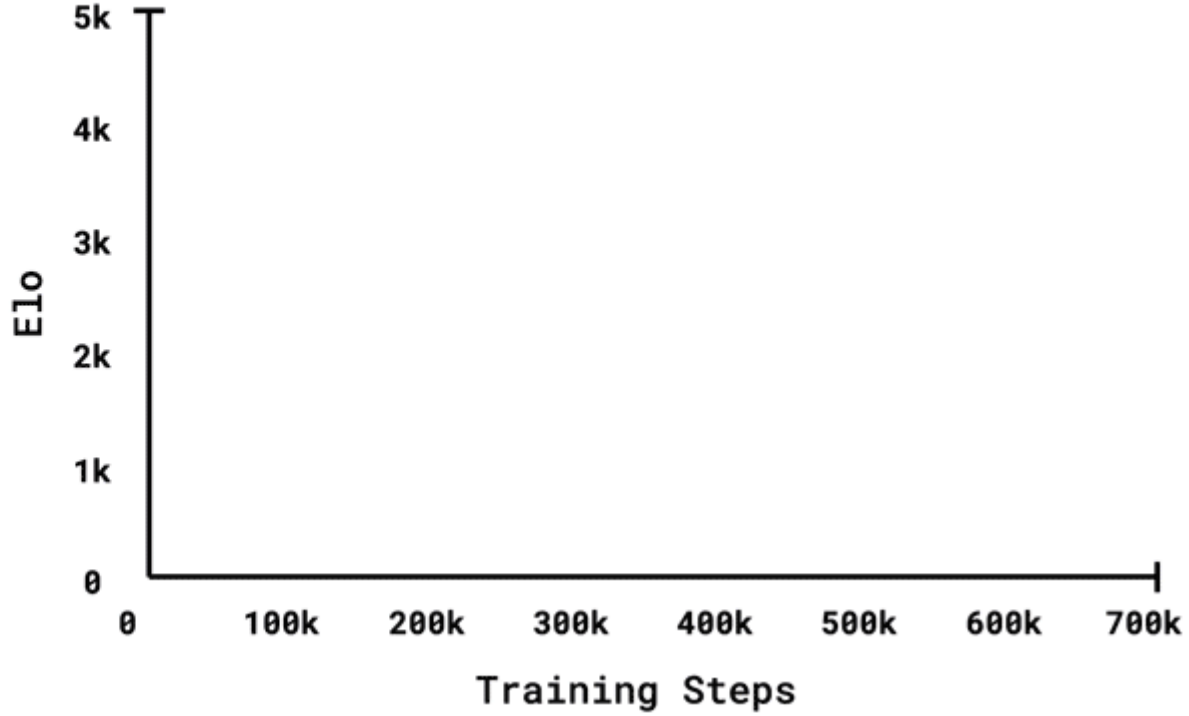
Monte Carlo Tree Search in AlphaGo



AlphaZero - Taught Itself from Scratch without Human Play

- **Moves** made by the machine shows new methods and strategies that were **unexplored before**





Back to Sport

- Unlike board games where movement of games pieces are strict, sport involves human movement and positioning which is complex and dynamic

- Some questions left?
  - Can certain sporting codes or specific tactics / components within sporting games utilize the learnings from AlphaZero and AlphaGo?
  - Can self-play be employed to uncover things that players and coaches never knew about their sport?
  - What is the investment and when will a sporting team obtain an ROI(Return on Investment)?

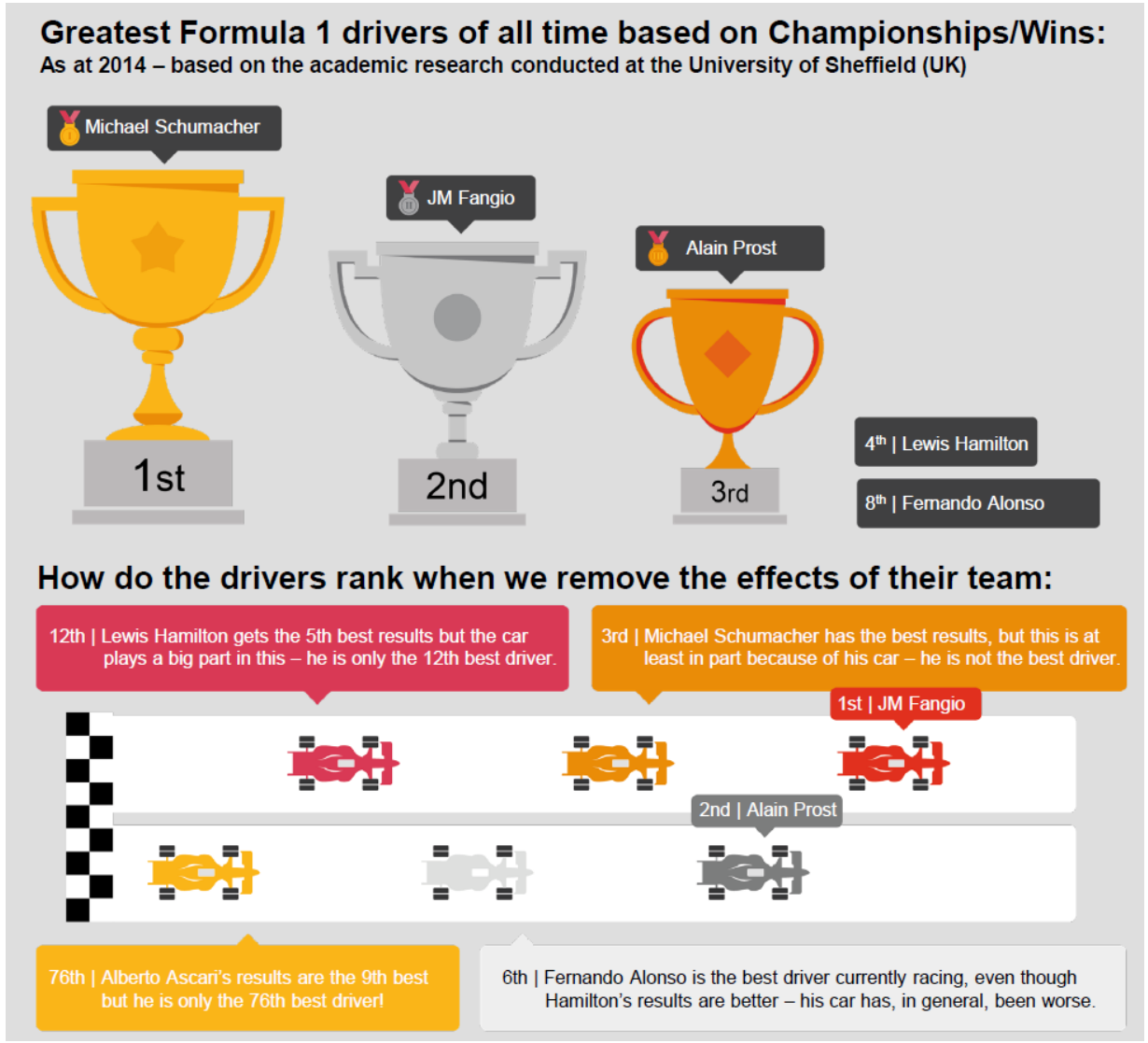
Formula One – Driver vs. Car

- Detailed statistical data were already recorded
  - Suspension deflection
  - Microliter of fuel used
  - Temperature of all main components
  - Instantaneous acceleration in any direction
- When every race is in progress, massive computer simulations are being run for every possible scenario

Contribution to race outcome	Driver performance	Car/Team Performance
In 1980	30%	70%
In 2018	10%	90%

"Drivers are just interchangeable light bulbs – you plug them in and they do the job"
Teddy Mayer, McLaren's hard-headed team principal

Formula One – Driver vs. Car



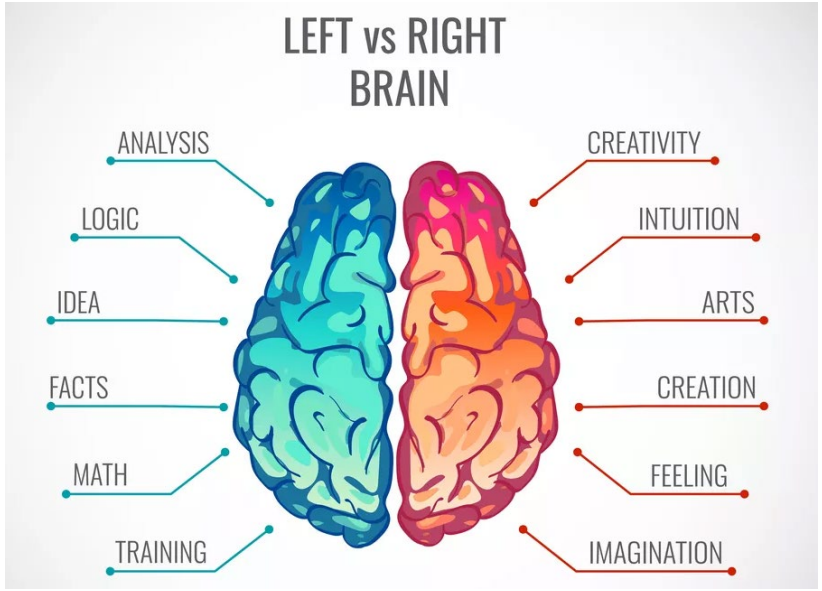
Daniel Ricciardo, F1 driver from McLaren-Mercedes's team, believed that

"Formula One is still too much about who has the fastest car rather than who has the most natural talent"

Bell, A., Smith, J., Sabel, C. E., & Jones, K. (2016). Formula for success: multilevel modelling of Formula One driver and constructor performance, 1950–2014. *Journal of Quantitative Analysis in Sports*, *12*(2), 99-112.

Ethical Considerations of AI

- F1's strategy is approaching every sport industries
- Coaching is an **imperfect** science
  - Human experiences
  - Intuition without the need for conscious reasoning
- When AI assisted coach or AI virtual coach in every play
  - Played in the most **efficient** and **effective** way
  - Will this make sport boring?



Imperfection makes things fun !

Things to Consider

- Regulatory Compliance
  - i.e. technology cap vs salary cap
- Privacy and Data Leakage
- Cyber Security
- Resilience -Business Continuity, Disaster Recovery and Crisis Management
- Third party dependencies
- Strategic and Operational Risk Management

- Integrity of Sport – Using athlete data to determine if games have been “thrown”
- In-game betting if athlete and game data is exposed
- In-game activities if security is compromised (i.e. AI Assistant Coaches, Virtual Umpires, etc.)
- Talent selection during the drafting process if talent identification applications are compromised
- Stadium and venue security if drone surveillance is compromised
- Impact on fans if chatbots and smart assistants are compromised

AI is the Future for Sport Industry

- AI is impacting nearly every professional sport and is now also filtering through to grassroots participants.
- Umpires/referees require assistance to make the right decisions in moments that matter.
- Fans are demanding more personalized experiences and greater connectivity.

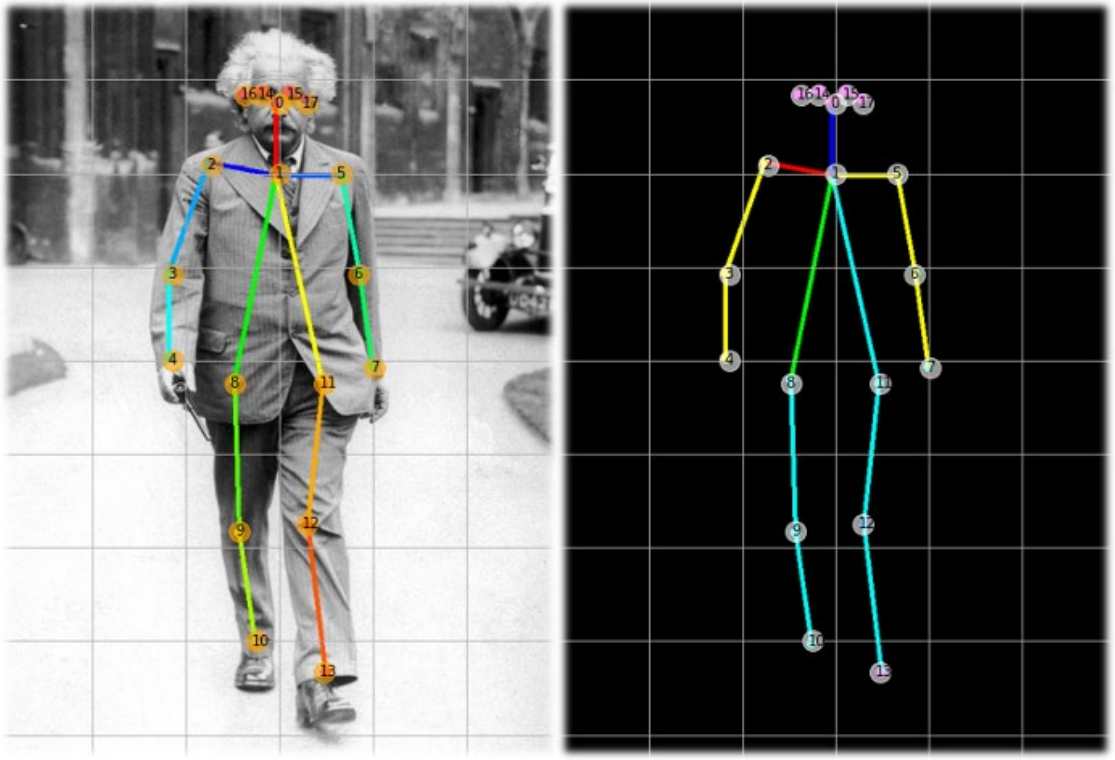
It's already difficult to compare sporting teams and players from different eras, but this will become increasingly more difficult with AI in the mix… only time will tell

Outline

- Overview of Smart Sport
- Pose Estimation
- Case Studies

Pose Estimation

- Pose estimation is a popular task in Computer Vision

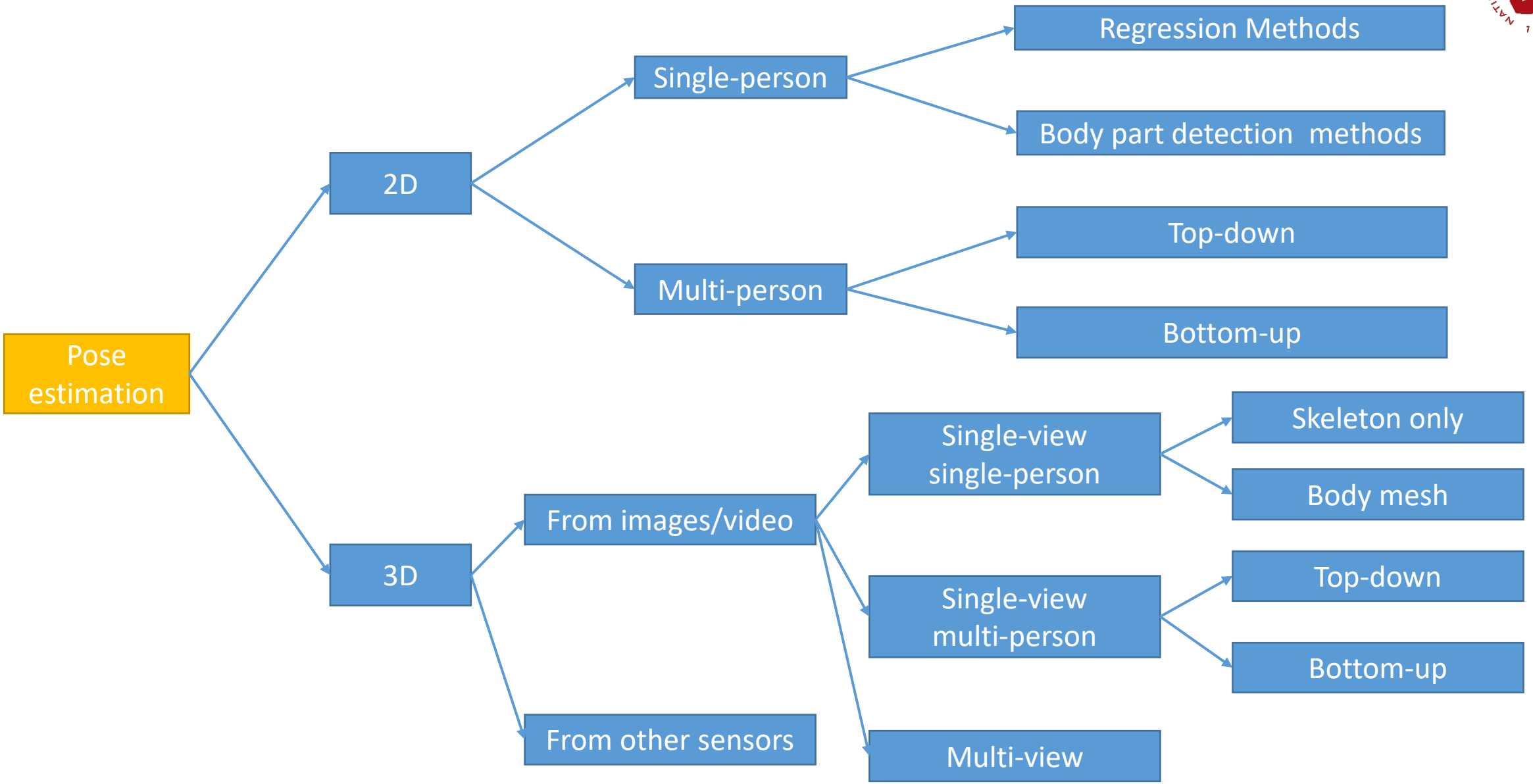


Pose Estimation

- To estimate the 2D or 3D position or spatial location of human body
  - From visual, such as images and videos

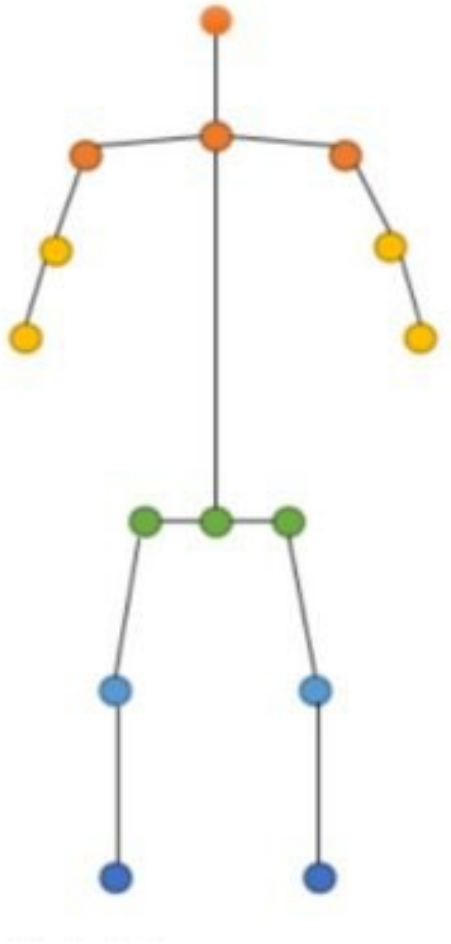
- Challenges
  - Unknown number of people
  - People can appear at any pose or scale
  - People contact and overlapping
  - Runtime complexity grows with the number of people

Taxonomy of Pose Estimation



Kinematic Model

- Also called skeleton-based model
- For 2D and 3D pose estimation
- A sets of joint position and limb orientation to represent the human body structure
- Skeleton pose are used to capture the relations between different body parts
- Limited in representing texture or shape information

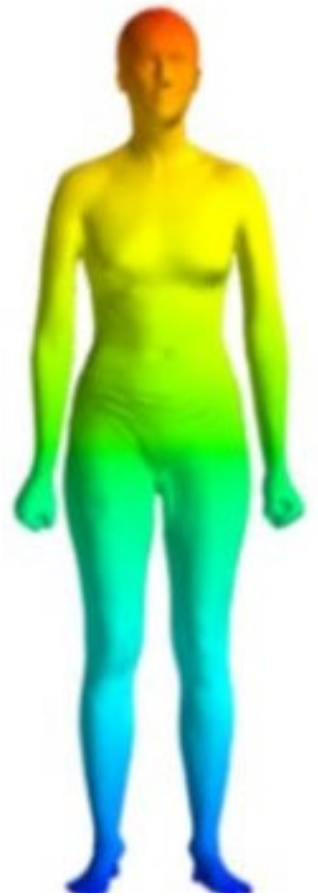


Planar Model

- Also-called contour-based model
- Used for 2D pose estimation
- Represent the appearance and shape of a human body
- Body parts are usually represented by rectangles approximating the human body contours
- Example – Cardboard model
  - Composed of body part rectangular shapes representing the limbs of a person

Volumetric Model

- Used for 3D pose estimation
- Based on 3D human pose estimation for recovering 3D human mesh
- Example - Skinned Multi-Person Linear (SMPL) model
  - Widely used model in 3D HPE
  - Modeled with natural pose-dependent deformations exhibiting soft-tissue dynamics
  - 1786 high-resolution 3D scans of different subjects of poses with template mesh



2D Human Pose Estimation

- Estimate the 2D position or spatial location of human body keypoints
  - From images or videos
- Traditional 2D HPE methods
  - Adopt different hand-crafted feature extraction techniques for body parts
  - Describe human body as a stick figure to obtain global pose structures
- Deep learning-based approaches
  - Achieved a major breakthrough in HPE
  - improving the performance significantly
- Regression methods
- Heatmap-based methods

Single-Person 2D HPE

- Regression method
  - Direct learn a mapping from original images to the kinematic body model
  - Produce joint Coordinates
  - A good feature that encodes rich pose information is critical for regression-based methods
  - One popular strategy to learn better feature representation is multi-task learning
    - Sharing representations between related tasks
    - E.g., Pose estimation and pose-based action recognition

- Heatmap-based method
  - Predict body joint locations using the supervision of heatmaps
  - The goal is to estimate K heatmaps $\{H_1, H_2, \ldots H_K\}$for a total of K keypoints
    - The target (or ground-truth) heatmap is generated by a 2D Gaussian centered at the ground-truth joint location
    - The pixel value $H_i(x, y)$in each keypoint heatmap indicates the probability that the keypoint lies in the position (x, y)
    - Minimizing the discrepancy (e.g., the Mean Squared-Error) between the predicted heatmaps and target heatmaps

Regression Methods

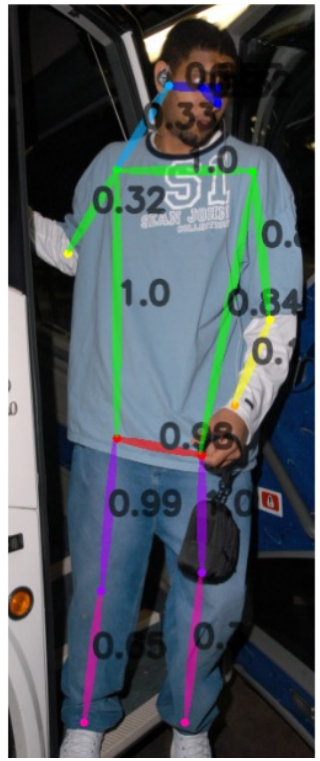
Input image

Pose Regression Network

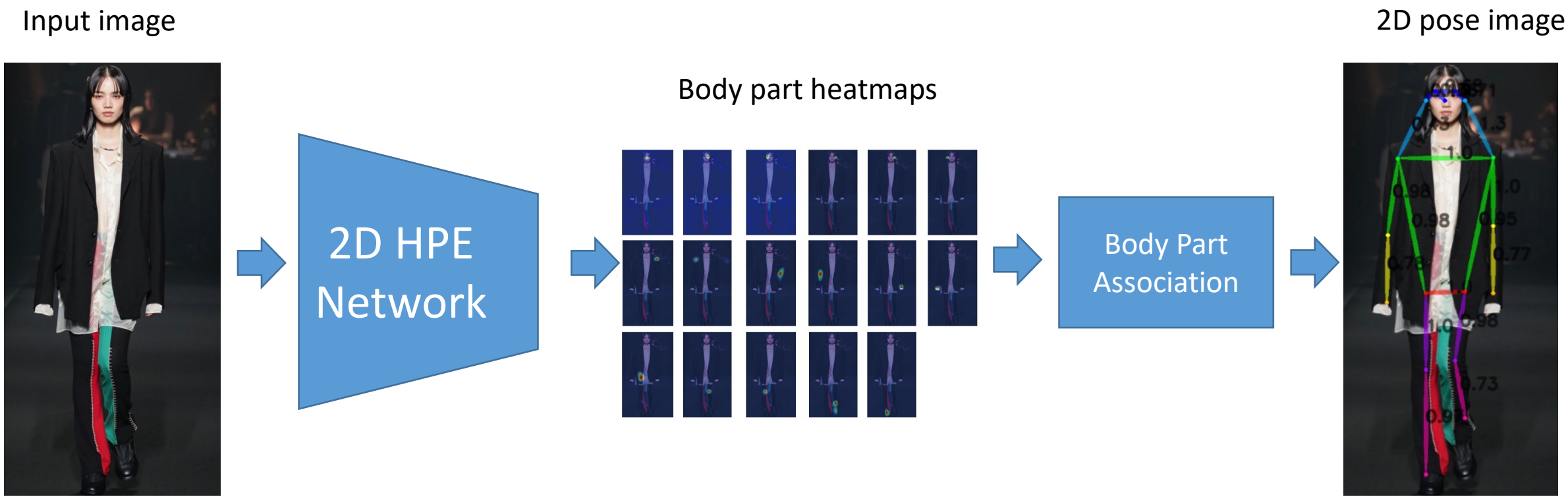
Keypoints (coordinates)

$$[[x_1, y_1],$$
$$[x_2, y_2],$$
$$\ldots$$
$$[x_n, y_n]]$$

2D pose image



Heatmap-based Methods

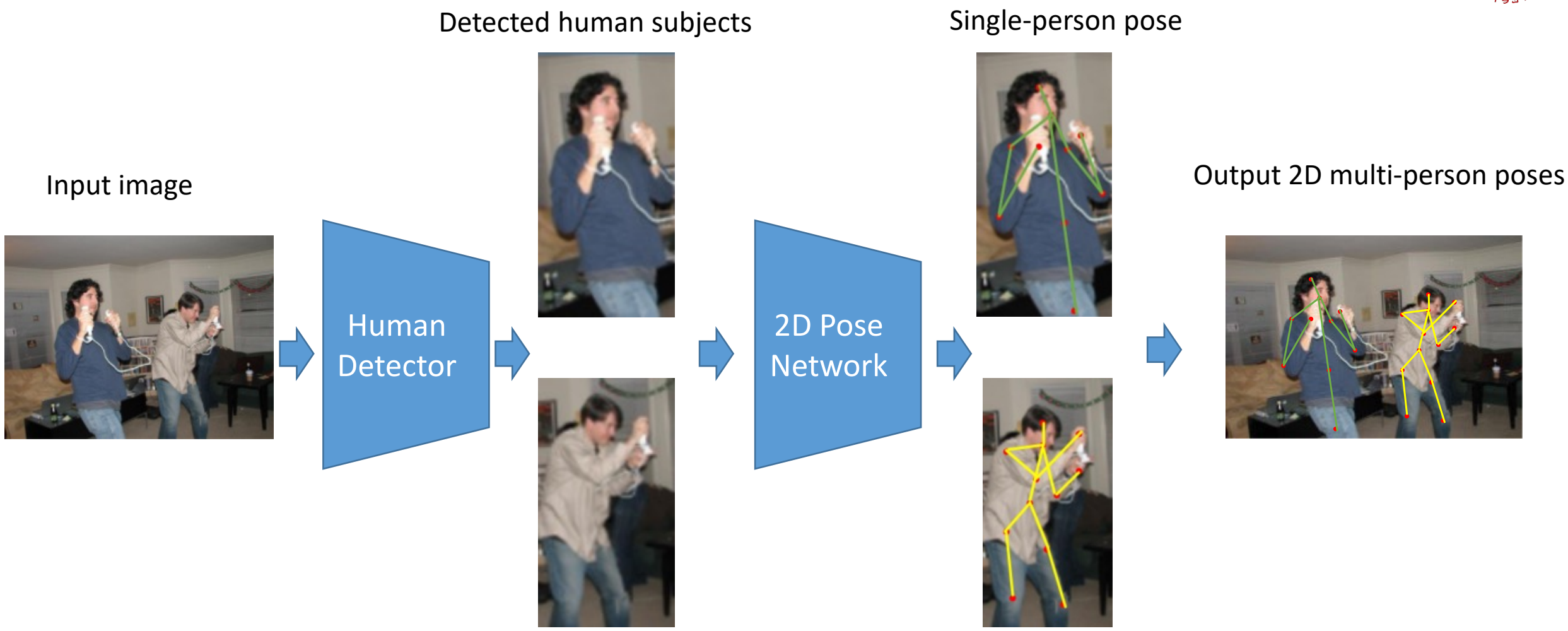


Multi-Person 2D HPE

- Top-down
  - Detect then single person pose estimation
  - Complexity proportional to number of people
  - Accuracy affected by detector performance

- Bottom-up
  - Pose estimate then clustering
  - Complexity not directly proportional to number of people
  - Real-time speed
  - Can take advantage of multi-task structures
    - Example – PersonLab combine the pose estimation and person segmentation module for keypoints detection and association

Top-Down Approaches



Bottom-Up Approaches

Challenges in 2D HPE

- Reliable detection of individuals under significant occlusion
  - Human detectors may fail in the first step of top-down pipeline due to occlusion
  - Difficulty of keypoint association is more pronounced for bottom-up approaches in occluded scenes
- Computation efficiency
  - Deploy the networks on resource-constrained devices
  - Real-world applications require more efficient HPE methods on commercial devices
    - Bring better interaction experience for users
- Limited data for rare poses
  - It might be enough for the normal pose estimation
  - For unusual pose, e.g.,falling, it's still limited data currently
  - Develop effective data generation or augmentation techniques to generate extra pose data for training more robust models

3D Human Pose Estimation

- Predict locations of body joints in 3D space
- Provide extensive 3D structure information related to human body
- From monocular images/videos
  - Ill-posed and inverse problem
  - Due to projection of 3D to 2D
  - 2D pose annotation can be easily obtained
- From multiple camera, IMU, or even Liar
  - Well-posed problem
  - Employ information fusion techniques
  - 3D pose annotation is time-consuming and manual labeling is not practical
- Challenges for in-the-wild data with unusual poses and occlusions

Single-view Single-person 3D HPE

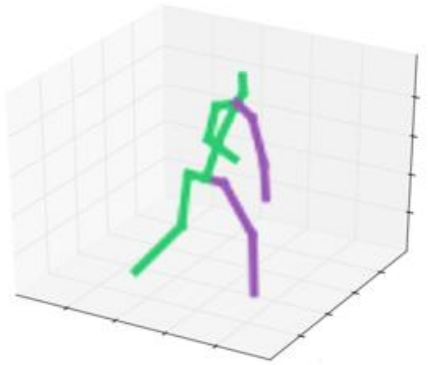
- Skeleton-only
  - Estimate 3D human joints as the final output
  - Do not employ human body models to reconstruct 3D human mesh representation
  - Direct estimation
    - Infer the 3D human pose from 2D images
    - Without intermediately estimating 2D pose representation
  - 2D to 3D lifting
    - Infer 3D human pose from the intermediately estimated 2D human pose
    - Benefiting from the excellent performance of state-of-the-art 2D pose detectors
    - Generally, outperform direct estimation approaches
- The kinematic model
  - An articulated body representation by connected bones and joints with kinematic constraints
  - Leverage prior knowledge based on the kinematic model
    - Skeletal joints connectivity information
    - Joints rotation properties
    - Fixed bone-length ratios
- Videos can provide temporal information to improve accuracy and robustness of 3D HPE

Skeleton-only – Direct Estimation

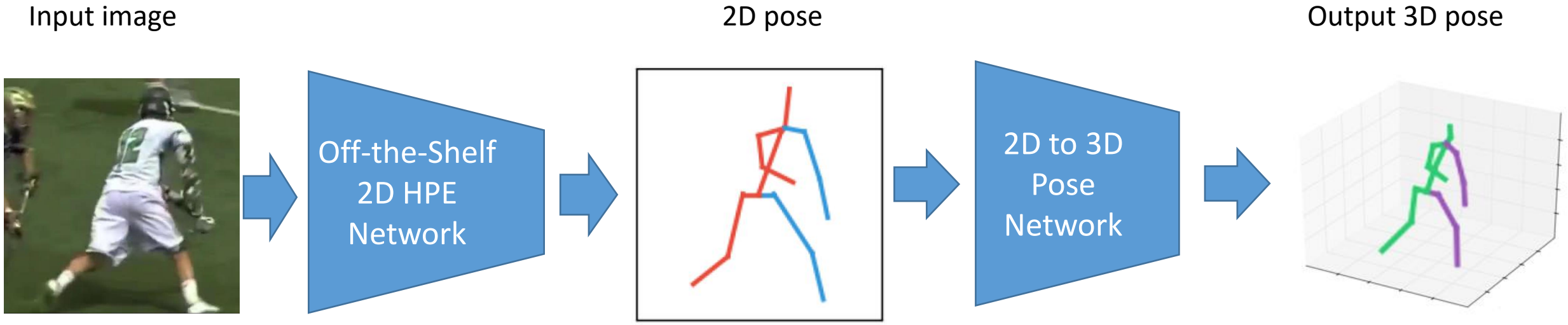
Input image

End-to-end network

Output 3D pose



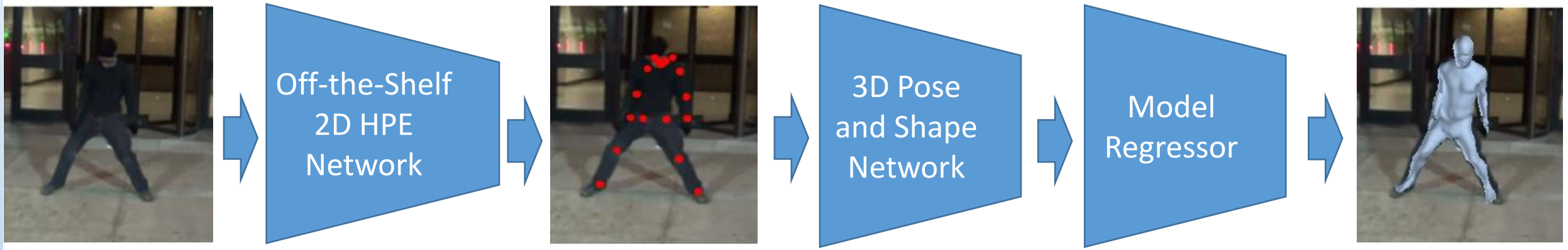
Skeleton-only – 2D to 3D Lifting



Human Mesh Recovery (HMR)

- HMR methods incorporate parametric body models to recovery human mesh
- The 3D pose can be obtained by using the model-defined joint regression matrix
- Volumetric models
  - Used to recover high-quality human mesh
  - Providing extra shape information of human body
- SMPL model has been widely used
  - Compatible with existing rendering engines
  - Reconstruct 3D human mesh from SMPL parameters
- Extended SMPL-based models
  - Address the limitations of the SMPL model
    - High computational complexity
    - Lack of hands and facial landmarks
  - SMPLify
    - An optimization method which fits the SMPL model
    - To the detected 2D joints and minimizes the re-projection error
- Cylinder Man Model
  - Generate occlusion labels for 3D data and performed data augmentation

Human Mesh Recovery



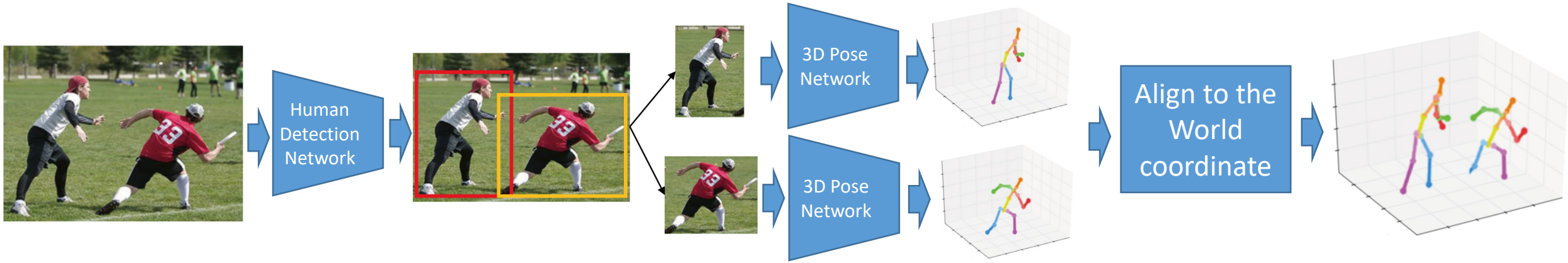
Single-view Multi-person 3D HPE

- 3D multi-person HPE from monocular RGB images or videos
- Top-down approaches
  1. Perform human detection to detect each individual person
  2. For each detected person, absolute root (center joint of the human) coordinate and 3D root-relative pose are estimated by 3D pose networks
  3. Based on the absolute root coordinate of each person and their root-relative pose, all poses are aligned to the world coordinate
  - Usually achieve promising results by relying on the state-of-the-art person detection methods and single-person HPE methods
  - Excessive Computational complexity and the inference time
    - With the increase in the number of humans
  - Global information in the scene may get neglected
    - The estimated depth of cropped region may be inconsistent with the actual depth
  - For body mesh reconstruction
    - Can be easily recovered by incorporating the 3D single-person human mesh recovery method

Single-view Multi-person 3D HPE

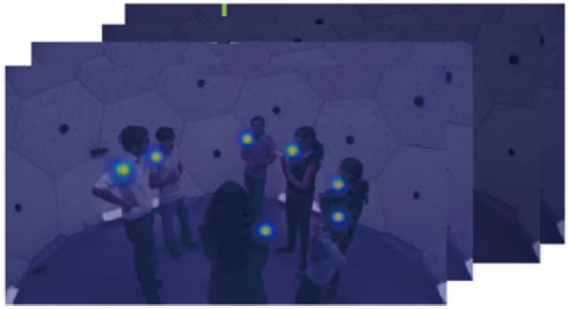
- Bottom-up Approaches
  1. Produce all body joint locations and depth maps
  2. Associate body parts to each person according to the root depth and part relative depth
  - Linear computation and time complexity
  - Challenge
    - How to group human body joints belonging to each person
    - Occlusion
  - For body mesh recovery
    - Not straightforward for the bottom-up approaches to reconstruct human body meshes
    - Additional model regressor module is needed to reconstruct human body meshes based on the final 3D poses

Top-Down Approaches

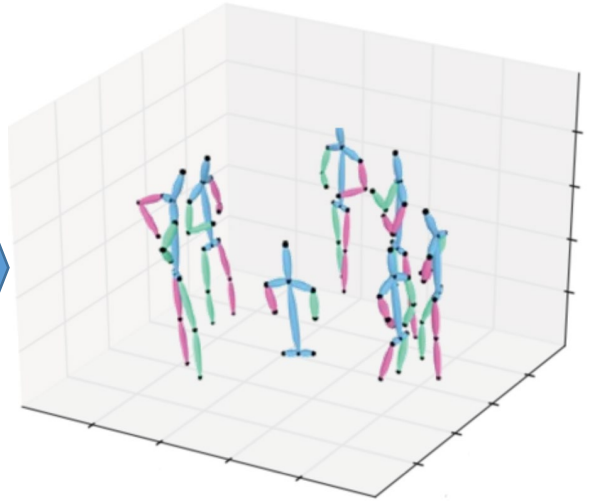


Bottom-Up Approaches

3D Pose Network



Depth-Aware Body Part Association



Multi-view 3D HPE

- Partial occlusion is a challenging problem for 3D HPE in the single-view setting
- Natural solution to overcome this problem - Estimate 3D human pose from multiple views
  - Occluded part in one view may become visible in other views
- Association of corresponding location between different cameras needs to be resolved
- Existing approaches
  - Tackle the association problem by optimizing model parameters to match the model projection with the 2D pose
    - Need large memory and expensive computational cost, especially for multi-person 3D HPE under multi-view settings
  - Employ multi-view consistency constraint in the network training
    - Requires a large amount of 3D ground-truth training data
  - Semi-supervised training
    - Learn the geometry-aware 3D latent representation from multi-view images and background segmentation without 3D annotations
  - Multi-view matching frameworks to reconstruct 3D human pose across all viewpoints with consistency constraints
- Besides accuracy, take consideration of
  - Lightweight architecture, fast inference time, and efficient adaptation

3D HPE from Other Sensors

- Depth and point cloud sensors
  - Depth sensors
    - Low-cost and increased utilization
    - The depth ambiguity problem can be alleviated by using depth sensors
      - One of the key challenges in 3D HPE
  - Point clouds can provide more information
- Inertial Measurement Units(IMUs) with monocular images
  - Track the orientation and acceleration of human body parts
    - by recording motions without object occlusions and clothes obstructions
  - The drifting problem may occur overtime when using IMUs
- Radio frequency (RF) device
  - The ability to traverse walls and to bounce off human bodies in the WiFi range
    - Without carrying wireless transmitters
  - Privacy can be preserved due to non-visual data
  - Relatively low spatial resolution compared to visual camera images and the RF systems
- Other sensors
  - Non-line-of-sight (NLOS) imaging system
  - Fish-eye camera
  - Pressure sensing mat

Datasets for 2D HPE

- Max Planck Institute for Informatics (MPII) Human Pose Dataset
  - Includes around 25,000 images containing over 40,000 individuals with annotated body joints
  - Covers 410 human activities and all the images are labeled

Datasets for 2D HPE

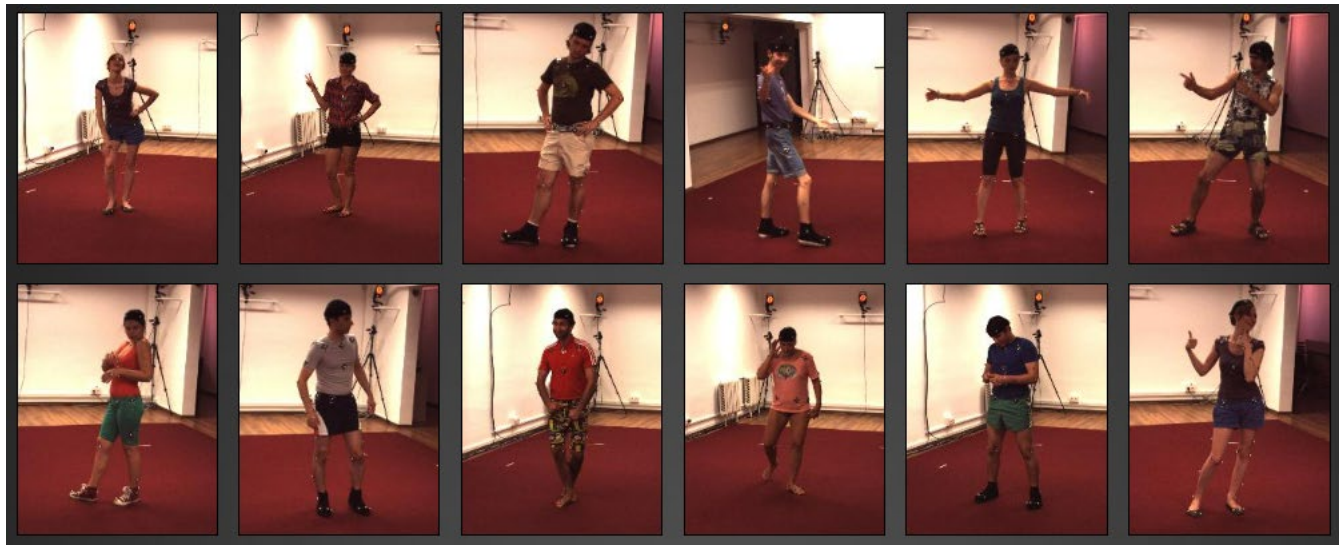
- Microsoft Common Objects in Context (COCO) Dataset
  - More than 330,000 images and 200,000 labeled subjects with keypoints
  - Each individual person is labeled with 17 joints

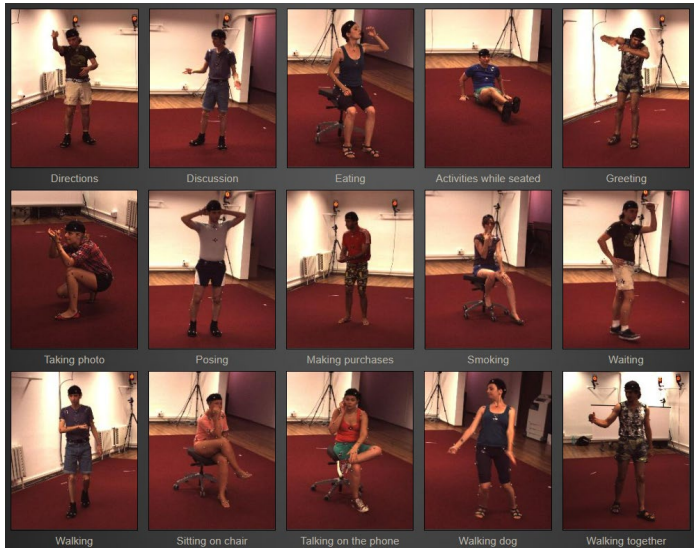
Evaluation Metrics for 2D HPE

- Percentage of Correct Parts (PCP)
  - Evaluates stick predictions to report the localization accuracy for limbs
- Percentage of Correct Keypoints (PCK)
  - Measure the accuracy of localization of different keypoints within a given threshold
- Average Precision (AP)
  - Measure the accuracy of keypoints detection according to precision
  - The ratio of true positive results to the total positive results
- Average Recall (AR)
  - Measure the accuracy of keypoints detection according to recall
  - The ratio of true positive results to the total number of ground truth positives

Datasets for 3D HPE

- Human3.6M
  - Indoor dataset for 3D HPE from monocular images and videos
  - 11 professional actors performing 17 activities
  - From 4 different views in an indoor laboratory environment
  - Contains 3.6 million 3D human poses with 3D ground truth annotation
    - Captured by accurate marker-based MoCap system





Datasets for 3D HPE

- MPI-INF-3DHP
  - Captured by a commercial marker-less MoCap system in a multi-camera studio
  - 8 actors performing 8 human activities
  - More than 1.3 million frames from 14 cameras were recorded in a green screen studio
    - Allows automatic segmentation and augmentation

Datasets for 3D HPE

- MuPoTS-3D
  - Multi-person 3D test set
  - Ground-truth 3D poses were captured by a multi-view marker-less MoCap system
  - Containing 20 real-world scenes
  - More than 8,000 frames were collected in the 20 sequences by 8 subjects
  - Challenging samples includes occlusions, drastic illumination changes, and lens flares

Evaluation Metrics for 3D HPE

- MPJPE (Mean Per Joint Position Error)
  - The most widely used metric
  - Computed by using the Euclidean distance between the estimated 3D joints and the ground truth positions

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} \|J_i - J_i^*\|_2$$

- PMPJPE
  - Also called Reconstruction Error
  - The MPJPE after rigid alignment by a post-processing between the estimated pose and the ground truth pose
- NMPJPE
  - The MPJPE after normalizing the predicted positions in scale to the reference

Evaluation Metrics for 3D HPE

- MPVE (Mean Per Vertex Error)
  - Measures the Euclidean distances between the ground truth vertices and the predicted vertices

$$MPVE = \frac{1}{N}\sum_{i=1}^{N} \|V_i - V_i^*\|_2$$

- 3DPCK
  - 3D extended version of the Percentage of Correct Keypoints (PCK) metric
  - An estimated joint is considered as correct if the distance between the estimation and the ground-truth is within a certain threshold
    - Generally the threshold is set to $150mm$

Outline

- Overview of Smart Sport
- Pose Estimation
- Case Studies

Case Studies on 2D Pose

- Top-down
  - DeepPose
  - Convolutional Pose Machines (CPM)
  - Hourglass
  - Simple Baselines
  - MSPN
- Bottom-up
  - Openpose
  - Associative Embedding
  - HigherHRNet

DeepPose - Estimation as Regression

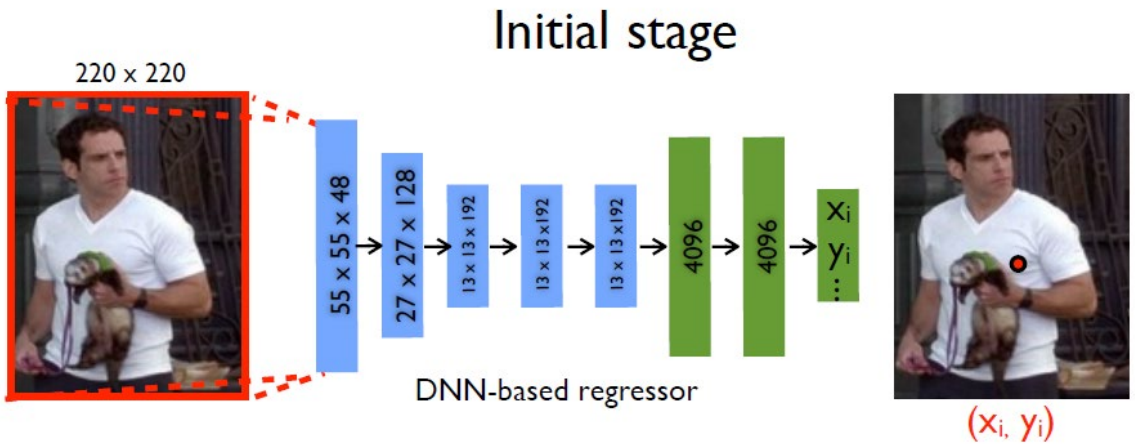
- Target
  - Predict a pose vector $\hat{y} = \left[y_1^T, \ldots, y_i^T \ \ldots, y_k^T\right]^T, i \in \{1, \ldots, k\}$
  - Each y_i contains the x and y coordinate of the i^{th} joint
- Normalize x, y according to the bounding box which contain this person
  - Bounding box $b = (b_c, b_w, b_h)$,
    - b_c: bounding box center, b_w: bounding box width b_h: bounding box height

$$N(y_i; b) = \begin{bmatrix} \frac{1}{b_w} & 0 \\ 0 & \frac{1}{b_h} \end{bmatrix} (y_i - b_c)$$

- Prediction the pose vector with a CNN network $\psi(x; \theta)$

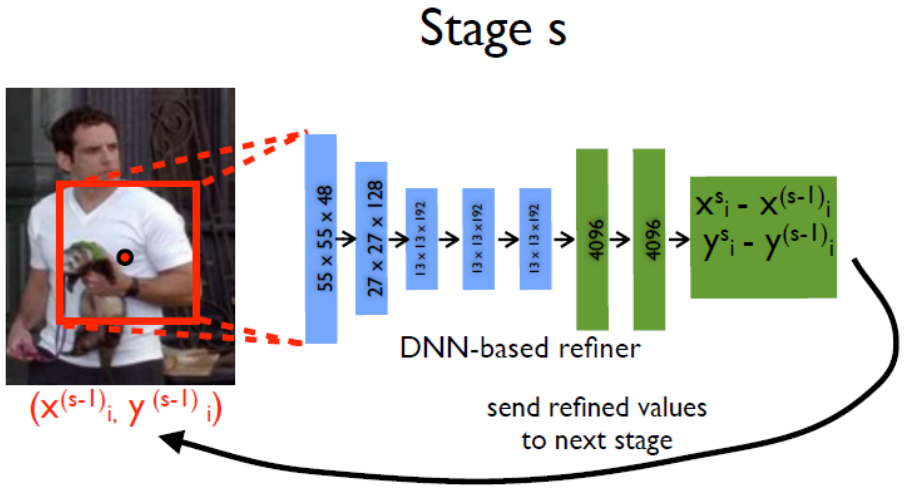
$$\hat{y} = N^{-1}\left(\psi(N(x); \theta)\right)$$

DeepPose – Iterative Refinement



Stage 1:

$$y^1 \leftarrow N^{-1}(\psi(N(x; b^0); \theta_1); b^0)$$



Stage s:

$$y_i^s \leftarrow y_i^{(s-1)} + N^{-1}(\psi_i(N(x; b); \theta_s); b)$$

joint bounding box b_i:

for $b = b_i^{(s-1)}$

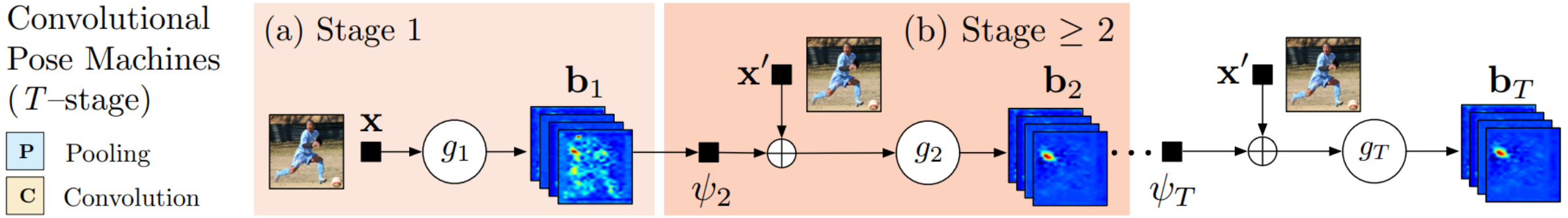
$$b_i^s \leftarrow (y_i^s, \sigma diam(y^s), \sigma diam(y^s))$$

Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1653-1660).

CPM – Heatmap-based Estimation

- Target:
  - Predict the image location $Y = \{Y_1, ... Y_P\}$ for all P part in an image
  - $Y_P \in \mathcal{Z} \subset \mathbb{R}^2$, $\mathcal{Z}$ is the set of all (u, v) locations in an image

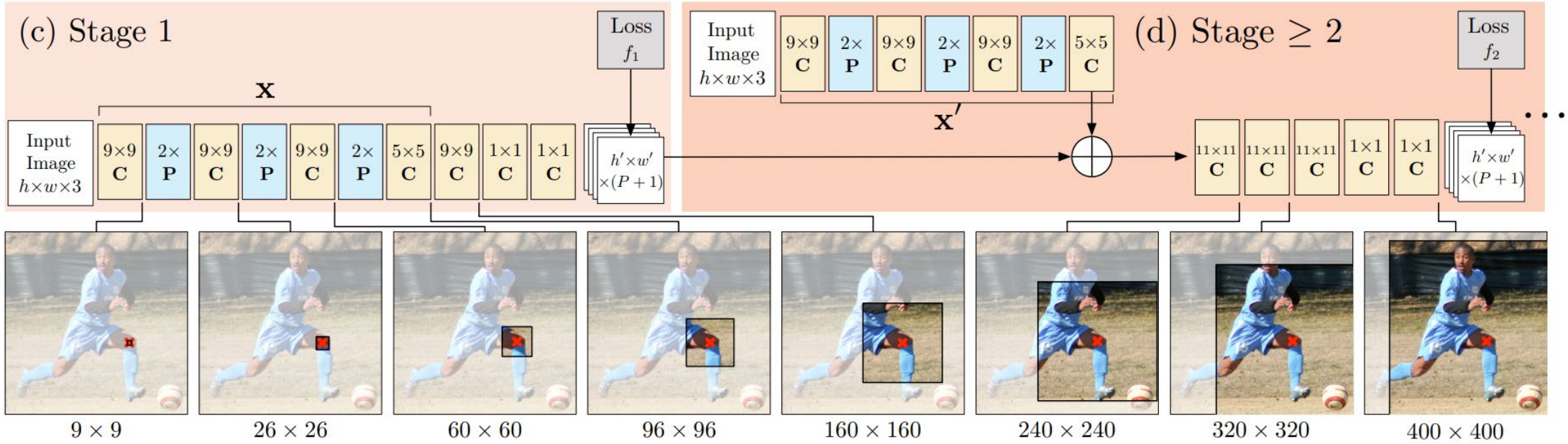
- Multi-stage refinement



Wei, S. E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 4724-4732).

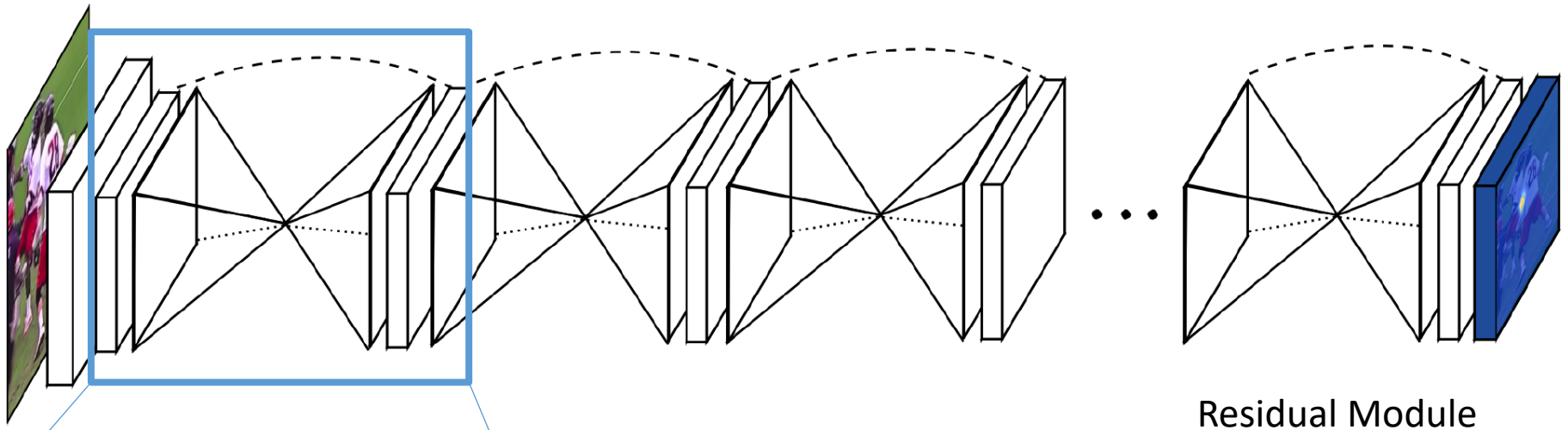
CPM

- Effective Receptive Field

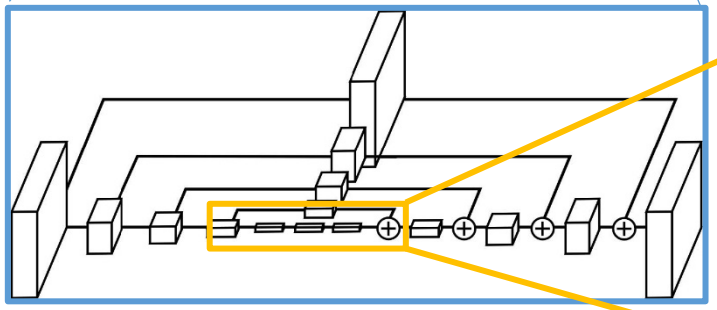


Hourglass

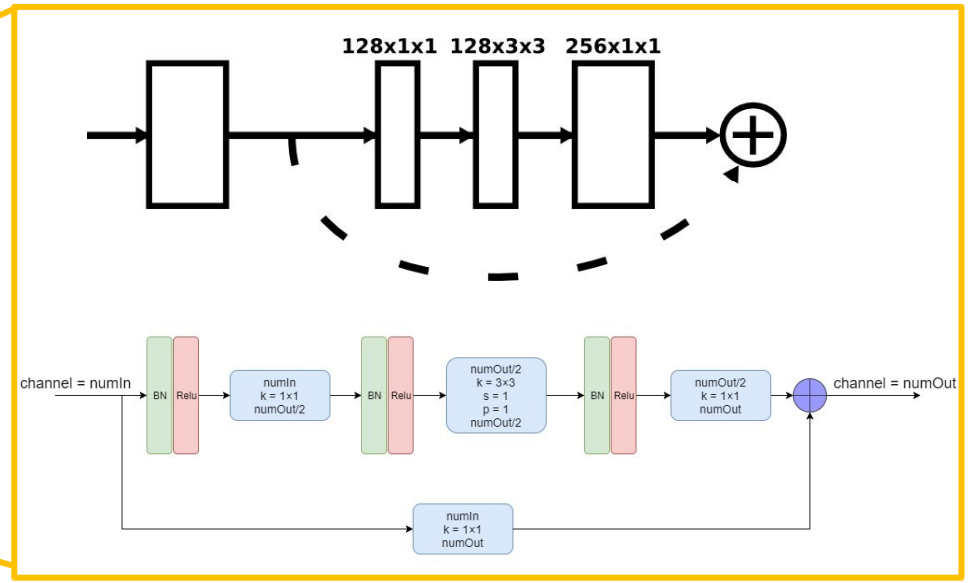
Use Nearest Neighbor for up-sampling



Hourglass Module



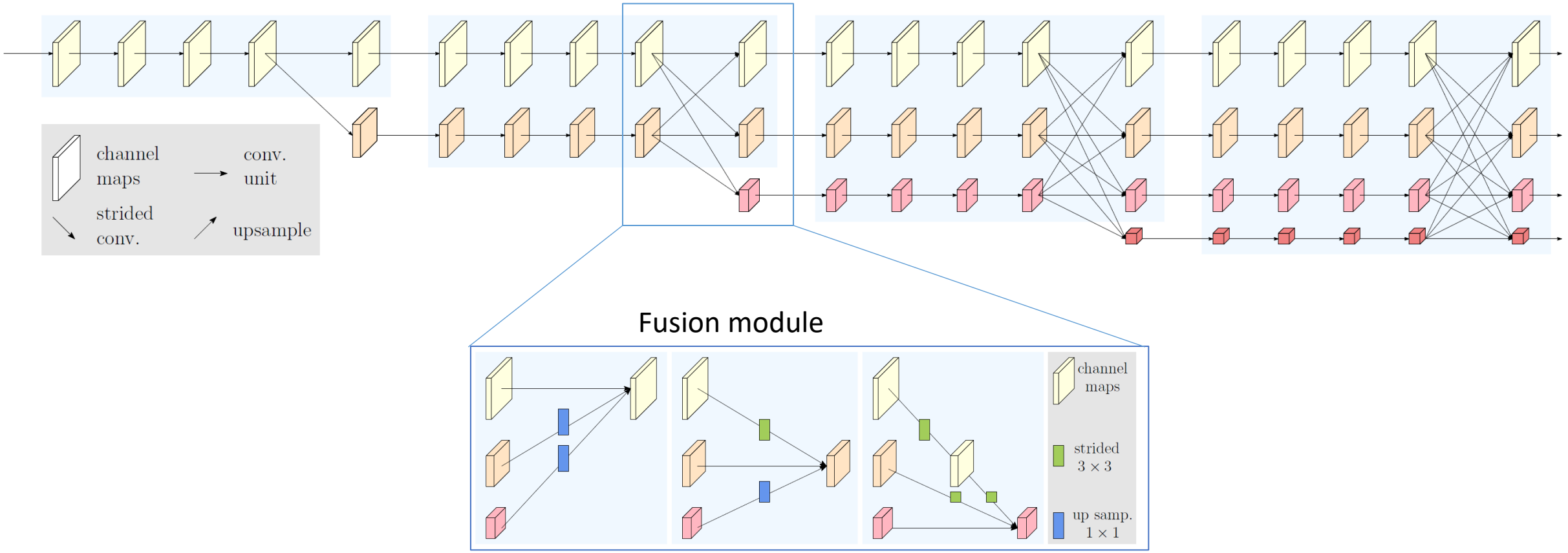
Residual Module



Newell, A., Yang, K., & Deng, J. (2016, October). Stacked hourglass networks for human pose estimation. In *European conference on computer vision* (pp. 483-499). Springer, Cham.

HRNet

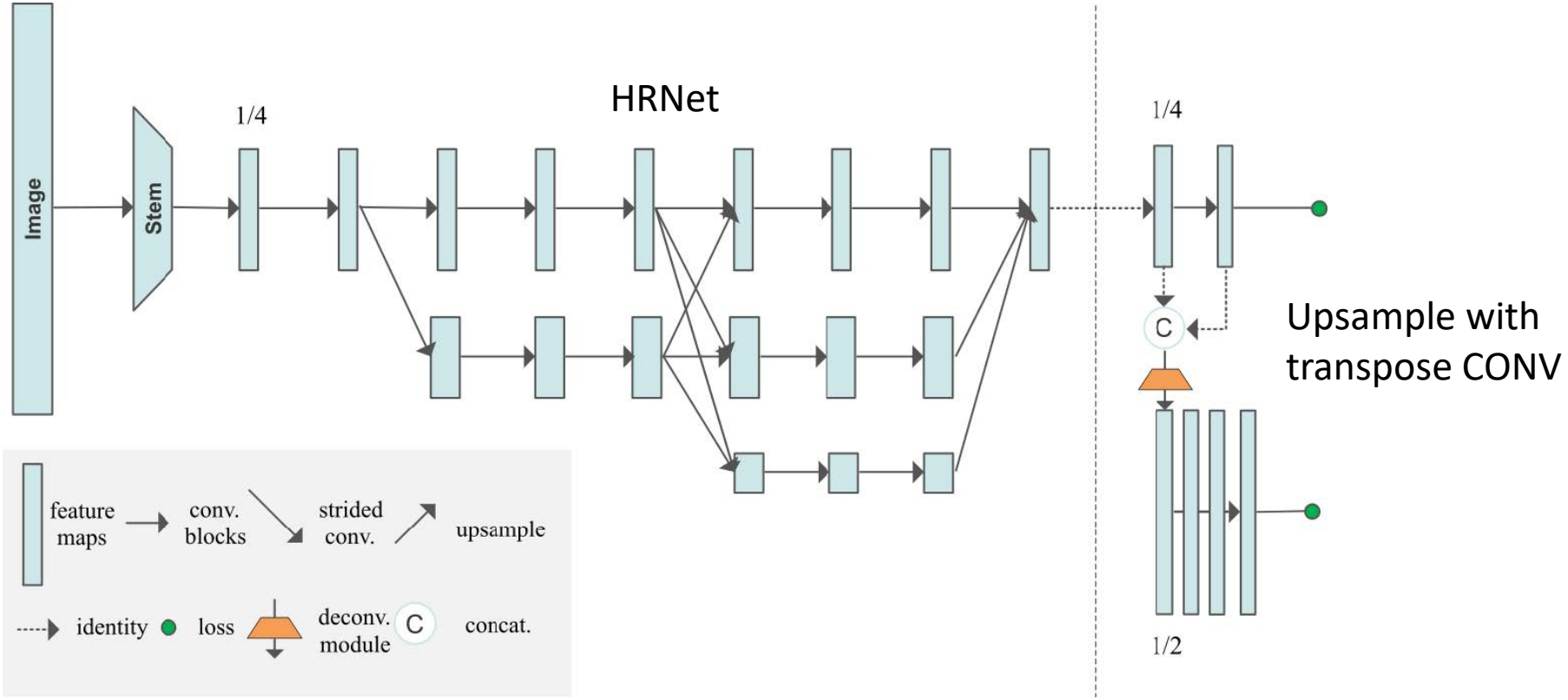
- Parallel multi-resolution sub-networks



Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3349-3364.

Higher HRNet

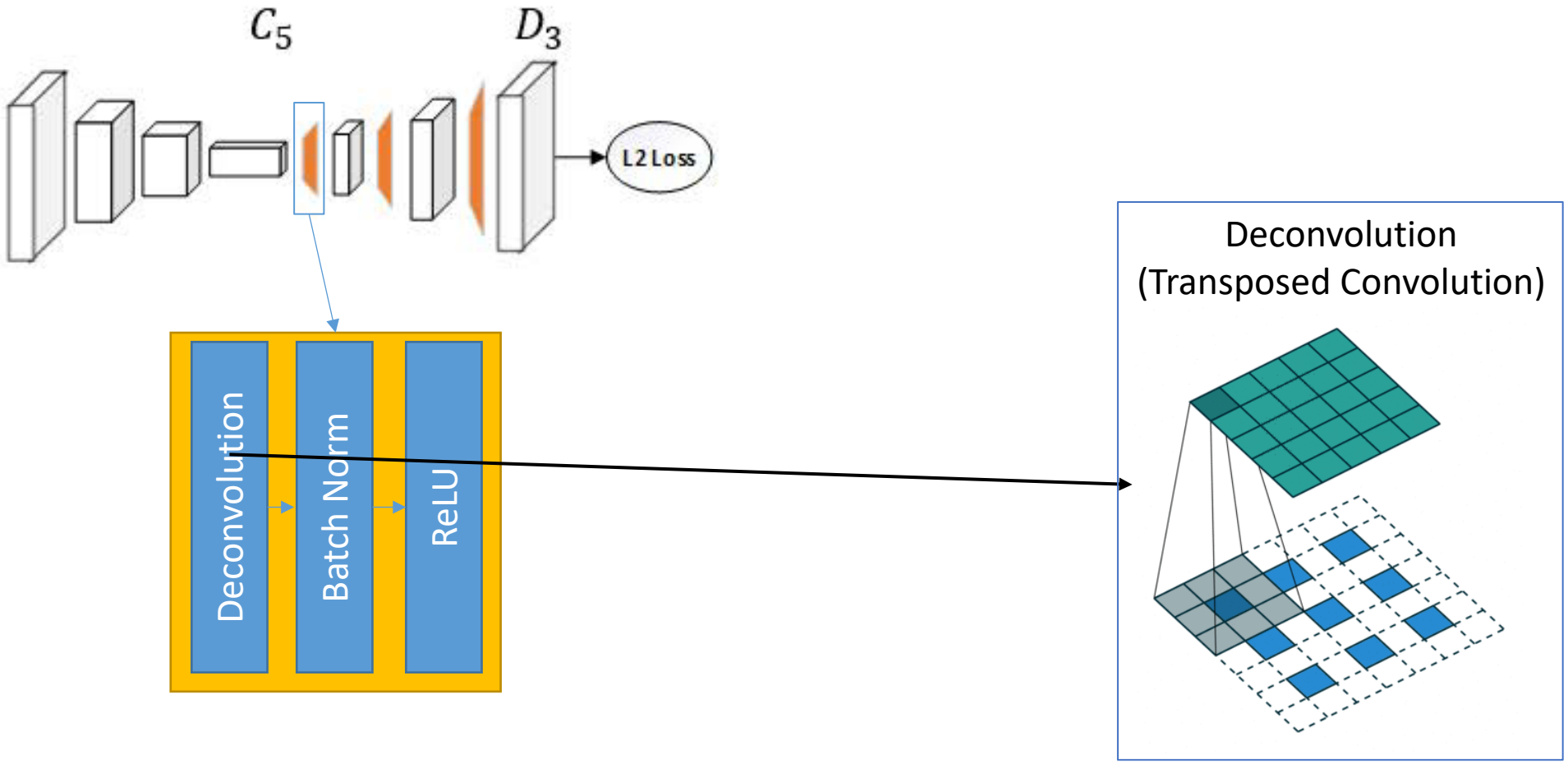
- ¼ Resolution is not accurate enough
- Upscaling using transpose convolution



Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5386-5395).

Simple Baselines

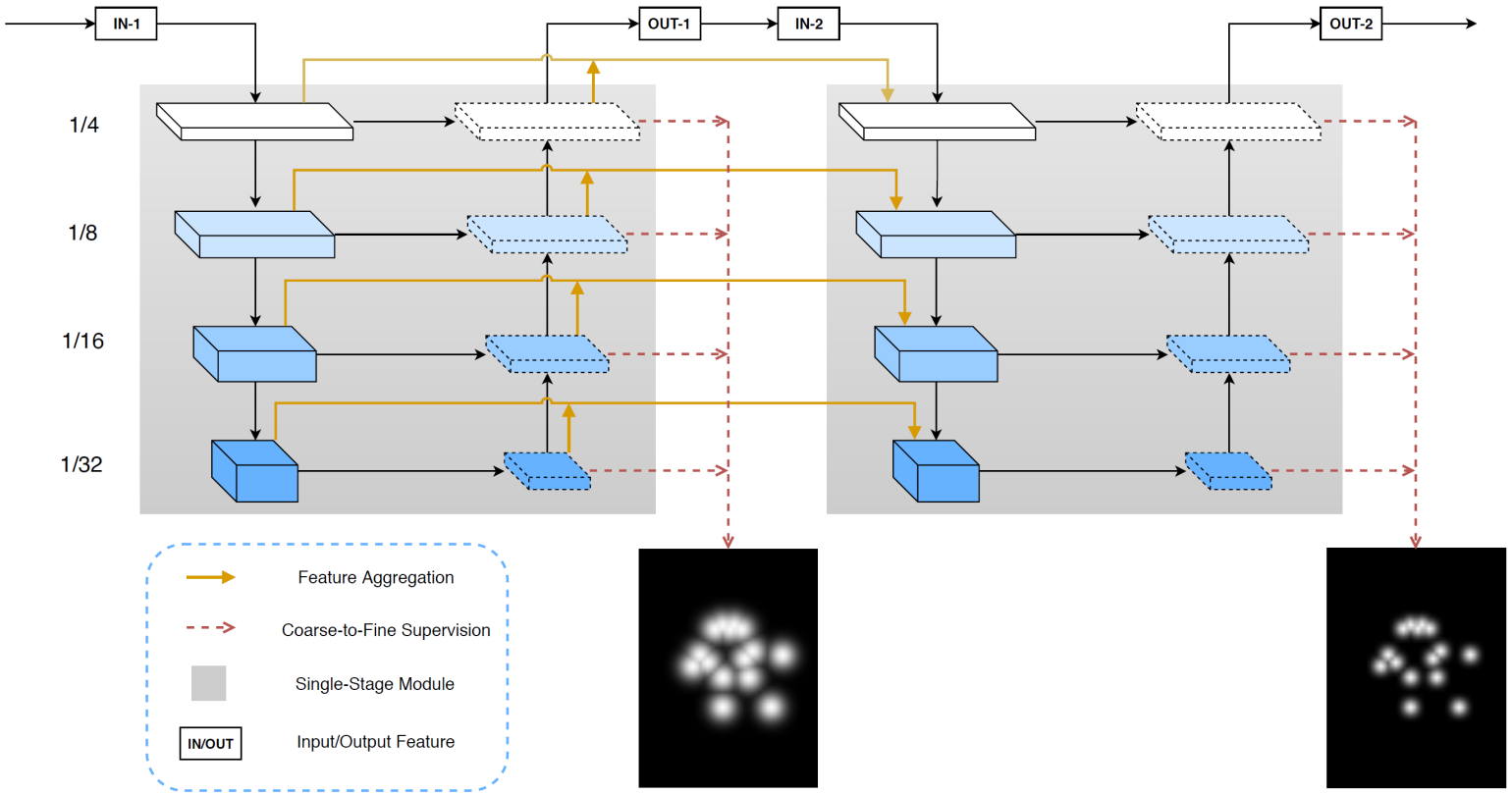
- Estimation Using A Deconvolution Head Network



Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 466-481).

MSPN

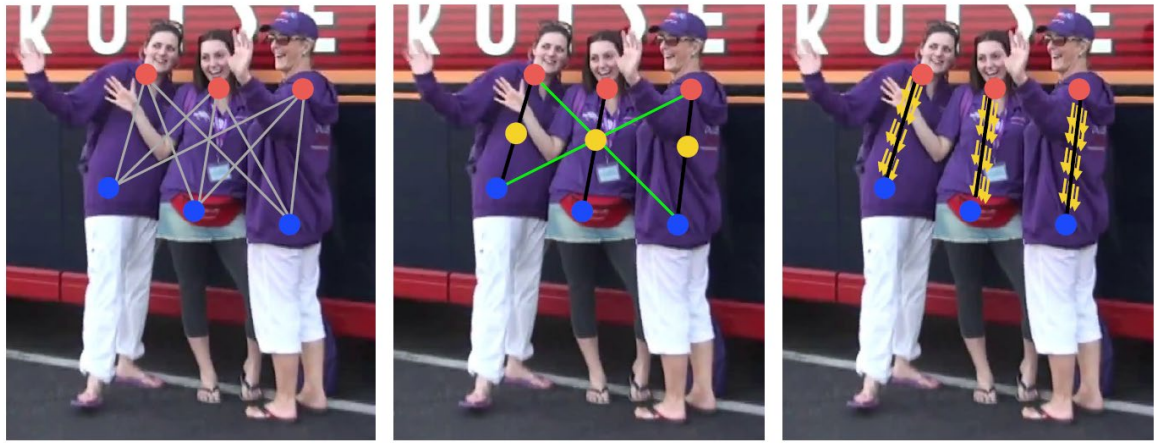
- Multi stage coarse-to-fine supervision



Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., ... & Sun, J. (2019). Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*.

Openpose

- Connect each body joint by Part Affinity Fields(PAF)
- Bipartite graph problem

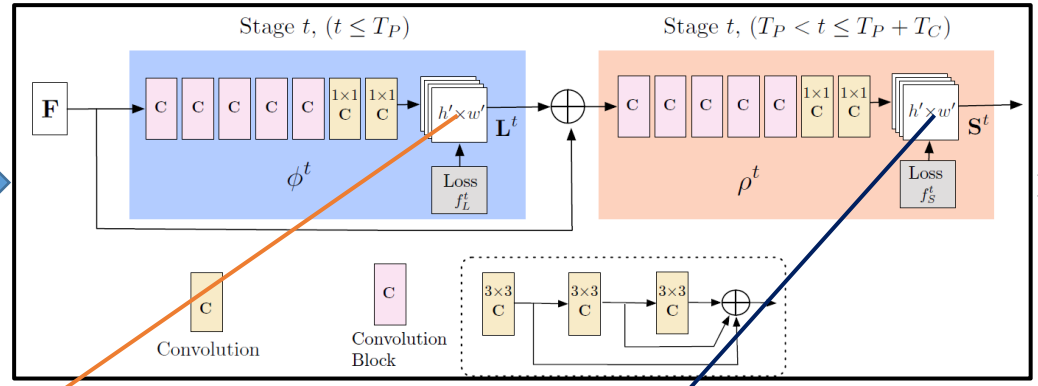


Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291-7299).

Openpose

Input Image

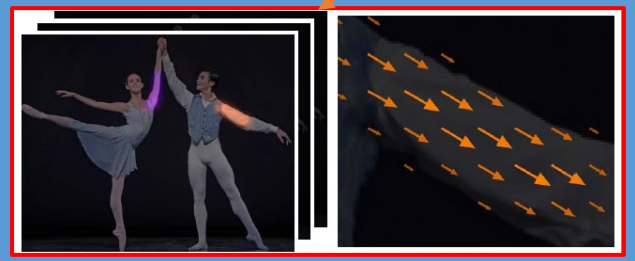
Multi-stage Estimation



$\times N$ *stages*

Parsing Results

Bipartite Matching

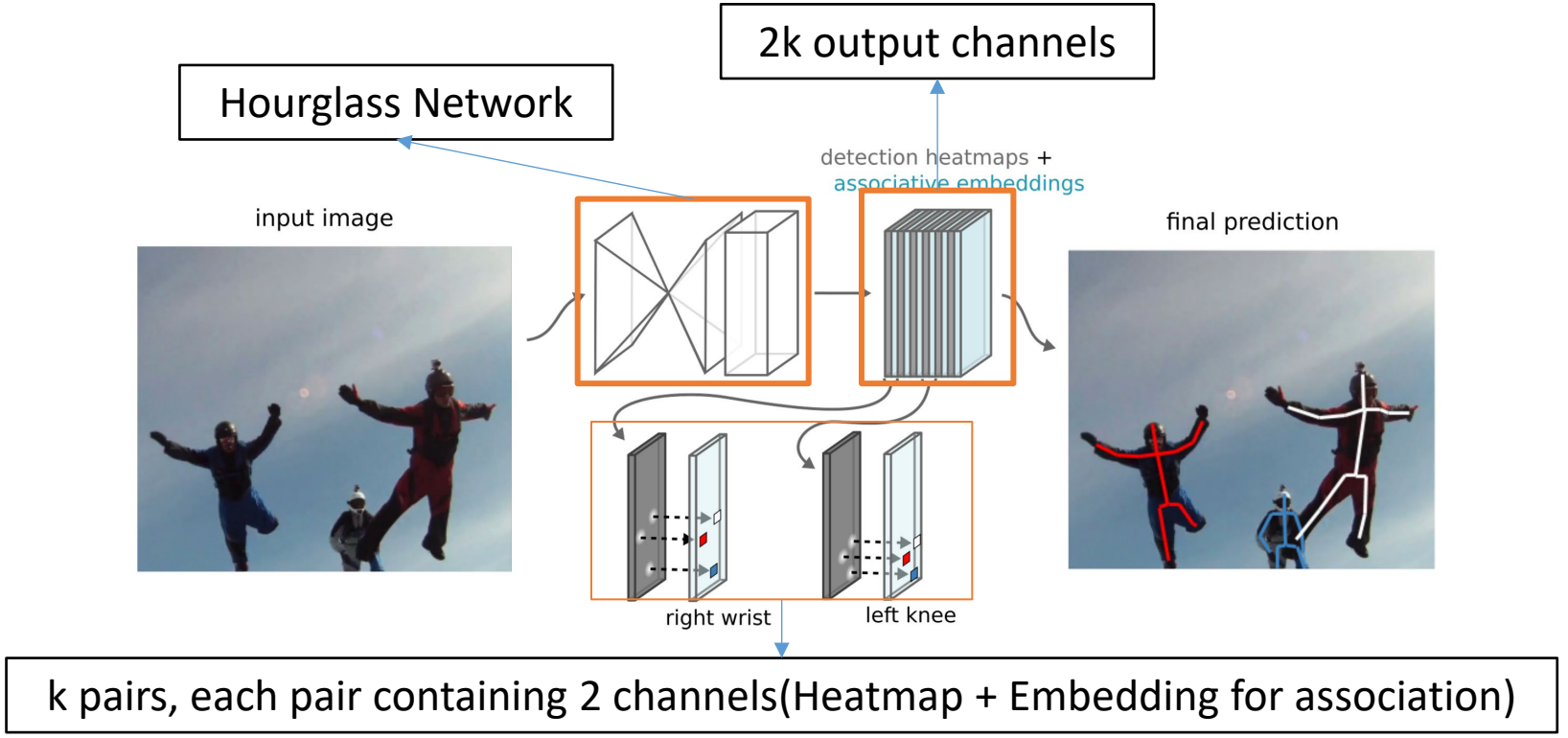


Part Affinity Fields

Part Confidence Maps

Hourglass + Associate Embedding

- Not only predict key points, but also their corresponding embedding
  - For k body joints => predicting 2k output channels
    - k channels for body joint heatmap , k channels for embedding



Newell, A., Huang, Z., & Deng, J. (2017). Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, *30*.

1D Embedding Association

- Group detections across body parts by comparing the tag values of detections and matching up those that are close enough