

University of Wrocław

# Human Keypoint Detection

Marcin Banak, Maciej Dengusiak, Patryk Flama, Szymon Fica

Wrocław June 12, 2025

# Task goal

Given a short video of human, we want to be able to detect what action <sup>1</sup> is being performed by human.

Based on previous idea, and if succeeded, we decided to extend it a little with such concept:

Given a video feed from a camera, we want to be able to extract skeletons <sup>2</sup> of all the people in it.

Then, based on those skeletons we would detect action performed by each human in the frame.

---

<sup>1</sup>action - some human-like activity that can be observed in a video

<sup>2</sup>skeleton - nodes (such as knee, elbow, hand, head) connected by edges (arm, leg, etc) representing human body parts

# Data

We will use the COCO 2020 Keypoint Detection Dataset (<https://cocodataset.org/#keypoints-2020>) that contains images and images for them. This would allow us to train such model, by extracting frame-by-frame images from videos.



# Data

Additionally we will use HMDB51 from torchvision to provide videos labeled with actions.



brush  
hair



cartwheel



catch



chew



clap



climb



climb  
stairs



dive



draw  
sword



dribble



drink



eat



fall  
floor



fencing



flic  
flac



golf



hand  
stand



hit



hug



jump



kick



kick  
ball



kiss



laugh



pick



pour



pullup



punch

The first part of the task is classical Computer Vision problem, so we will use similar approach.

The second part is classification, but of a video, probably of unknown length. That's why we are going to use recurrent NN for it.

# Additional Experiments

After training the skeleton-detection model we want to:

- Check convolution layers to see what kernels have been trained
- Check last convolution layers to see how they react for some input data
- Try to extract some images that optimize human features (such as "perfect head")