



Lifting 2D Keypoints to 3D Human Pose Estimation

Amoh Benjamin, Vasu Patel, Jiwen Chen

Spring 2022 CS 231n: Convolutional Neural Networks for Visual Recognition

TA Mentor: Sumith Kulal



Motivation

Problem: Non-trivial mapping between 2D poses and 3D poses due to the depth ambiguity and the occlusion of the input image

Goal: Improve existing transformer-based models to lift the 2D joints from multi-view images to the 3D position

Dataset

Human Pose Estimation: Human 3.6M dataset

Key Features:

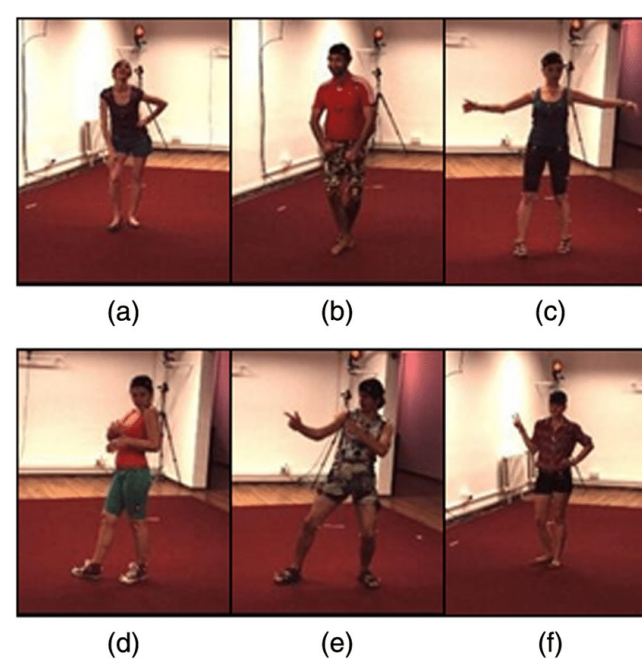
- Largest human action dataset
- 15 everyday activities by 15 subjects
- 50 Hz 4-point video camera

Training: S1, 5, 6, 7, 8; total 3,119,552 frames

Validation: S9, 11; total 543,360 frames

Data Pre-Processing:

- Obtain 2D key-point detection using CPN network
- 440 X 440 pixel around the center of person



Evaluation Metrics

Mean Per Joint Position Error (MPJPE)

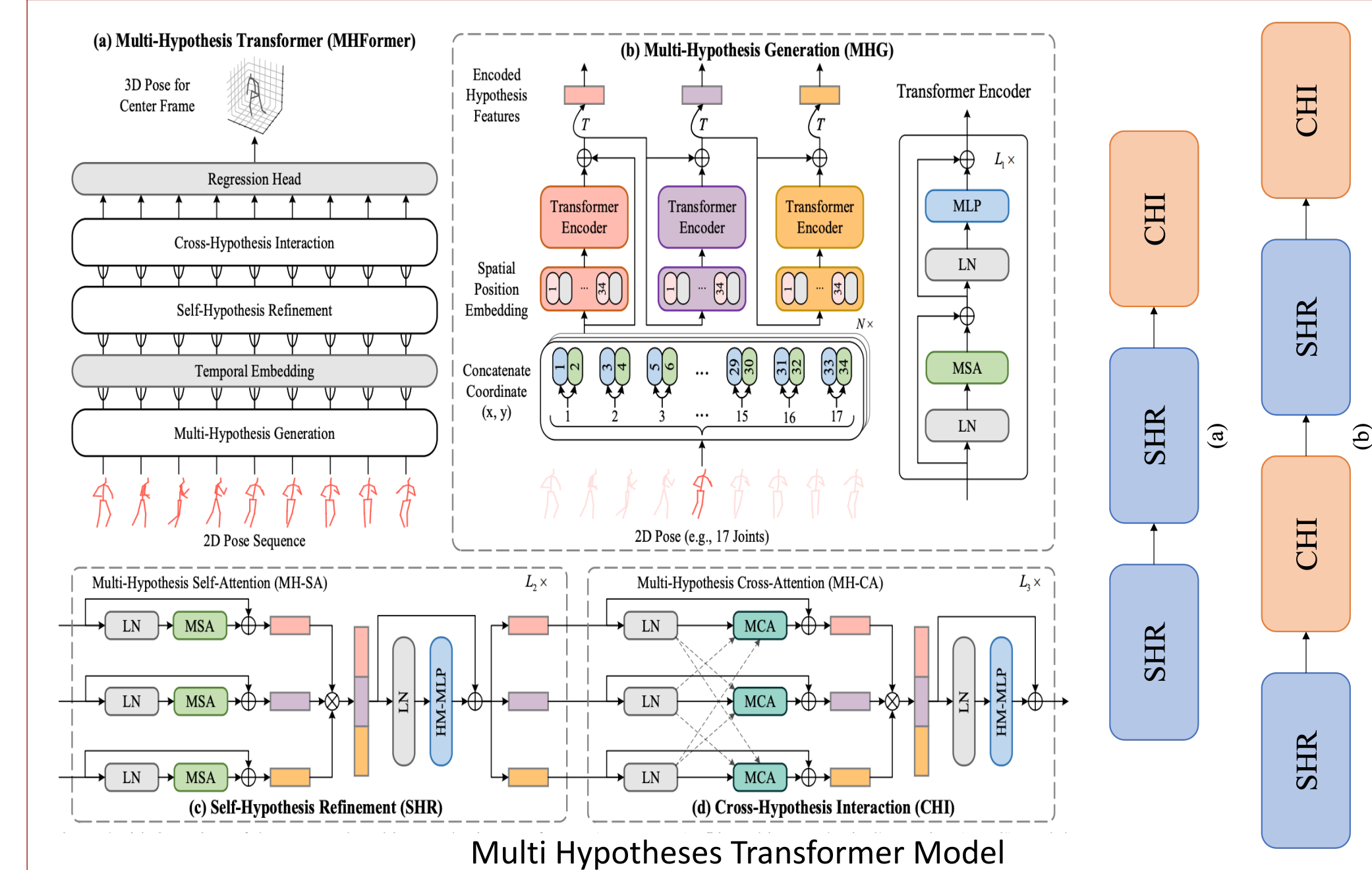
$$MPJPE(f, s) = \frac{1}{N_s} \sum_{n=1}^{N_s} ||P_{f,s}^{(f)}(i) - P_{gt,s}^{(f)}(i)||_2$$

f = frame index | s = corresponding skeleton

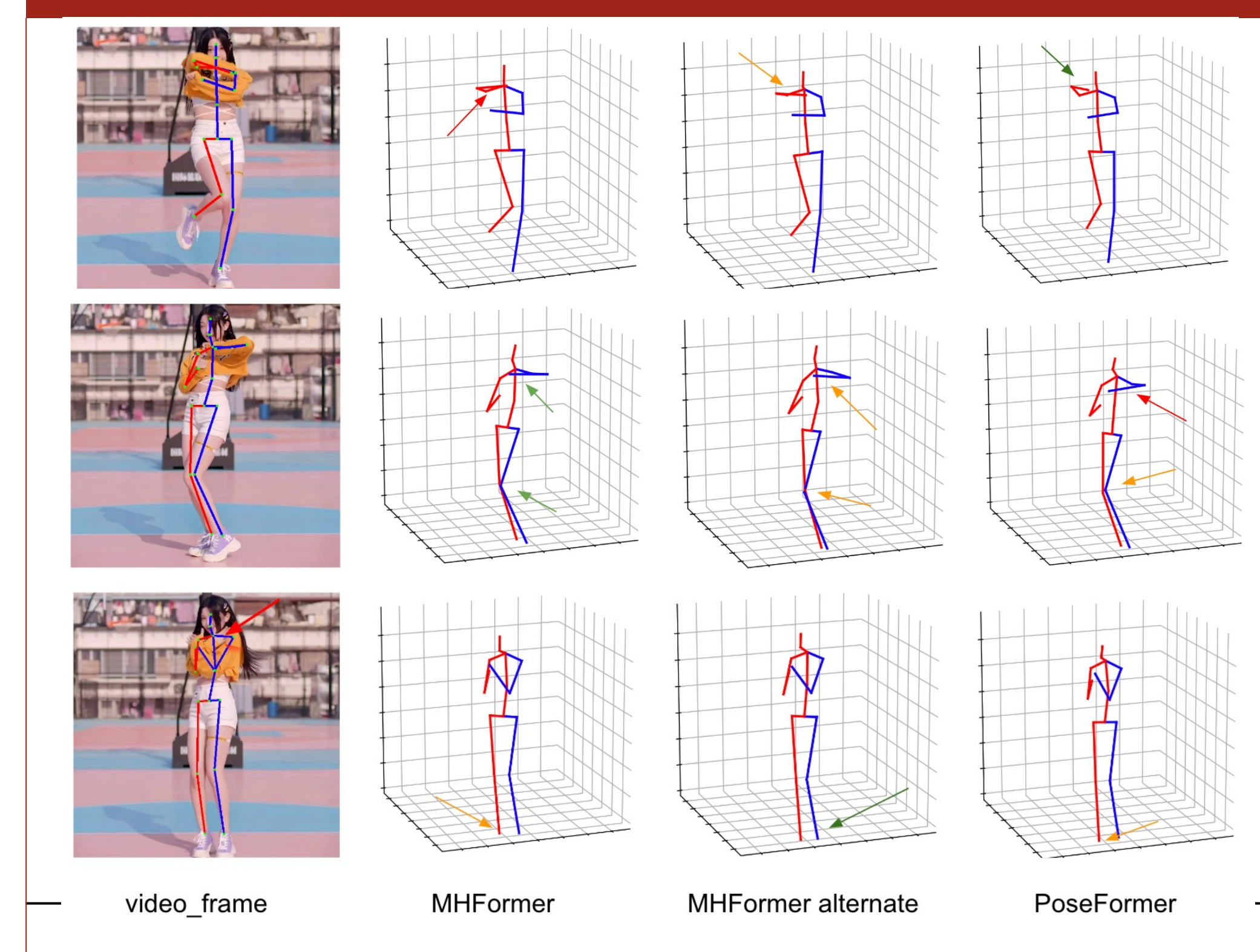
$P_{gt,s}^{(f)}$ – estimated ground truth.

$P_{f,s}(i)$ – estimated position of joint i

Model Architecture



Qualitative Results



Quantitative Results

Actions	MHF-Original	MHF-Alt	PF-Original	Actions	MHF-Original	MHF-Alt	PF-Original
Directions	48.58	52.48	48.77	Directions	36.32	36.58	37.25
Discussion	49.99	51.35	59.05	Discussion	39.35	40.63	42.82
Eating	45.86	48.25	44.98	Eating	35.94	36.78	35.72
Greeting	51.16	54.02	54.41	Greeting	39.46	40.83	42.03
Phoning	55.15	58.82	61.39	Phoning	42.67	44.52	45.78
Photo	65.45	70.09	61.14	Photo	43.82	46.83	47.19
Posing	49.87	51.47	47.5	Posing	37.03	37.34	37.04
Purchases	52.82	52.9	50.25	Purchases	36.57	37.16	38.44
Sitting	69.49	73.4	76.21	Sitting	56.83	57.82	59.29
SittingDown	72.22	73.58	85.07	SittingDown	57.21	57.09	64.64
Smoking	54.34	56.55	56.36	Smoking	41.98	42.87	42.99
Waiting	52.36	51.64	51.02	Waiting	38.95	37.92	38.34
WalkDog	54.6	55.55	54.39	WalkDog	40.17	41.63	43.76
Walking	44.71	47.09	39.85	Walking	31.1	32.17	32.98
WalkTogether	45.59	50.58	43.98	WalkTogether	32.23	33.67	37.23
Average	54.15	56.52	55.59	Average	40.64	41.59	43

Left: MPJPE vs. Different models trained after two epochs.

Right: P-MPJPE vs. Different models trained after two epochs

Discussion

- MHFormer performs better on actions with heavy occlusions, for example, sitting and phoning. This is because MHFormer generates multiple hypotheses
- PoseFormer performs better on actions with less occlusions, for example, walking and posing. This is because single hypothesis is capable of predicting less occlusive actions while multi-hypothesis adds variance
- For all models, actions with heavier occlusion have worse performance compared with actions with less occlusion
- Transformer models need at least the root location (trunk) for inferencing 3D pose estimation
- Input data accuracy (2D joint position) is crucial to the performance of all the transformer models

Future Works

- Train more epochs until convergence
- Try alternate model architecture to leverage MHFormer multiple hypotheses along with PoseFormer to generate 3D HPE for occluding situations
- Potentially fuse Transformer-XL for increasing the long-term dependency