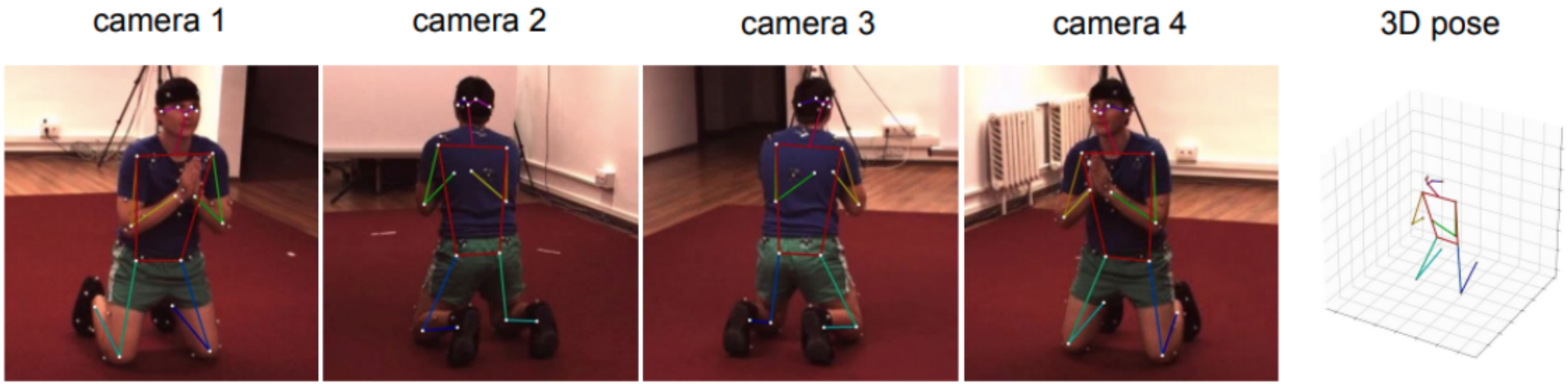
TransFusion:
Cross-view Fusion with Transformer for 3D Human Pose Estimation

Paper ID: 0016

Haoyu Ma[1], Liangjian Chen[2], Deying Kong[1], Zhe Wang[1], Xingwei Liu[1],
Hao Tang[1], Xiangyi Yan[1], Yusheng Xie[3], Shih-Yao Lin[4], Xiaohui Xie[1]
[1]University of California, Irvine, [2]Facebook, [3]Amazon, [4]Sony

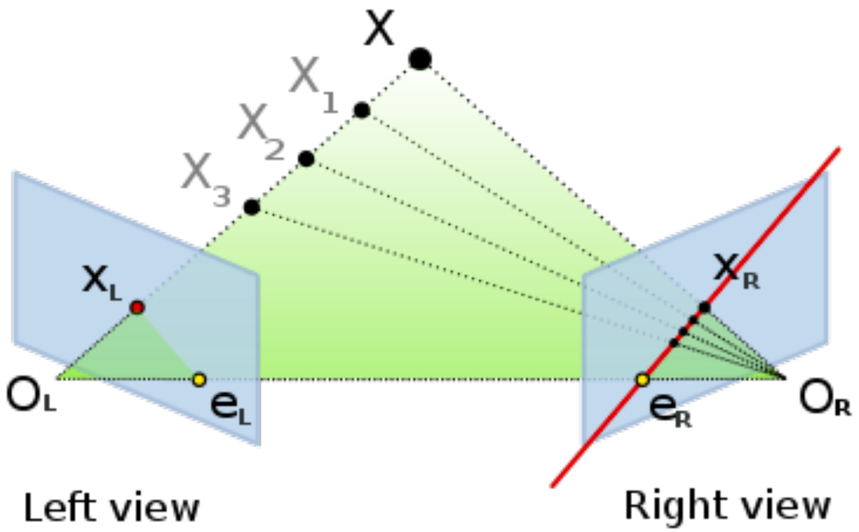
Background: Multi-view 3D Pose Estimation

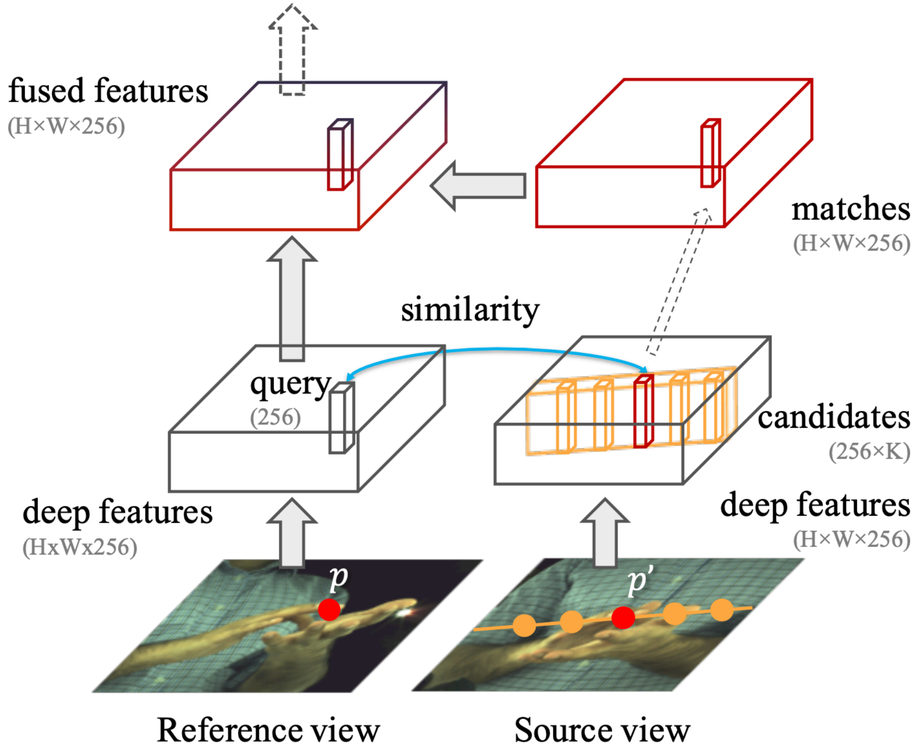
- Apply a 2D pose detector on each view separately to localize 2D joints
- Perform triangulation to acquire the 3D position of joints



Related Work: Epipolar Transformers

fuse information from different views





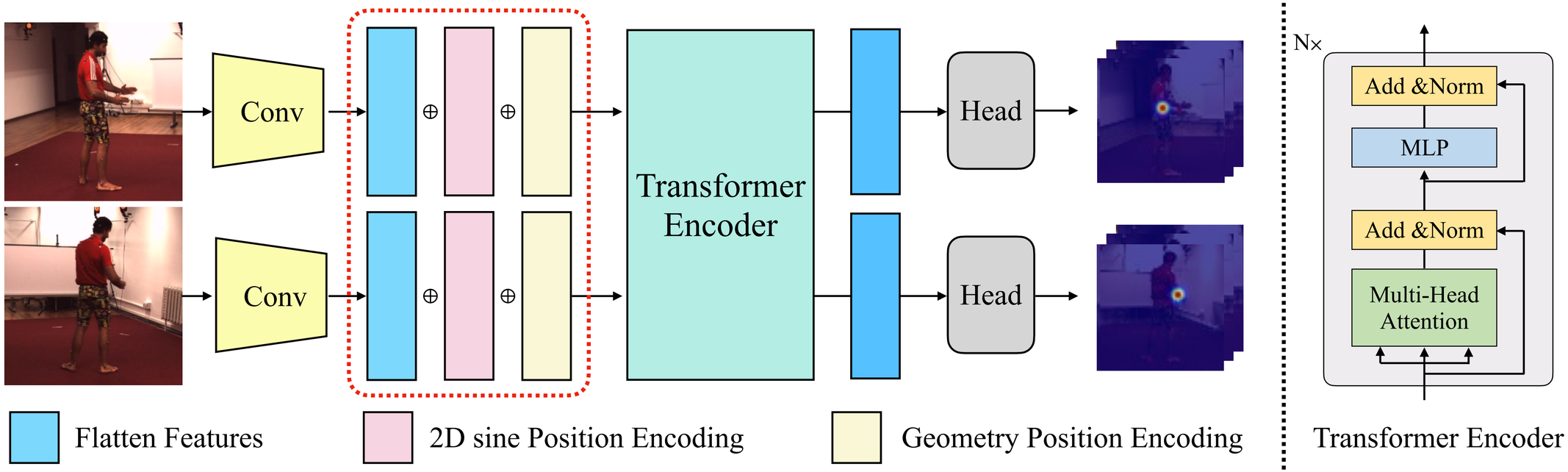
- Semantic information off the epipolar line is discarded !

He, Yihui, et al. "Epipolar transformers." *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020.

https://en.wikipedia.org/wiki/Epipolar_geometry

TransFusion: Framework

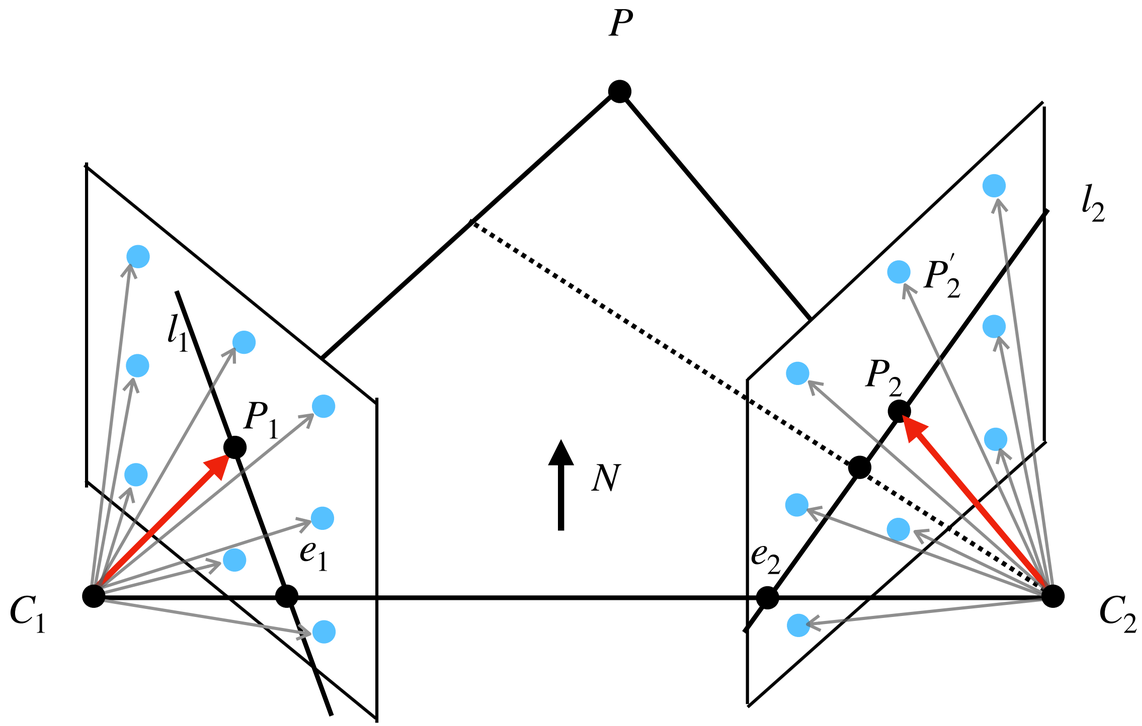
Global attention between two views



Geometry Positional Encoding

Encode 3D location information

- Use the direction vector to encode the relative 3D location

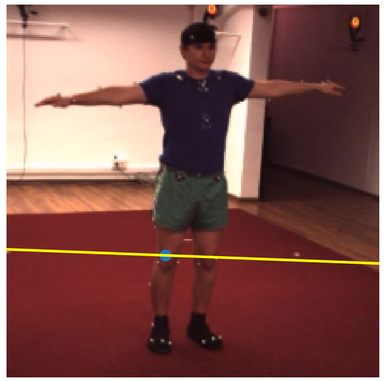


Epipolar Field

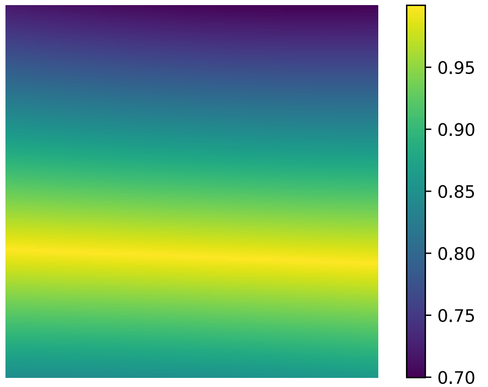
Encode epipolar geometric constraints through position encoding

- Relationship between the pixel in the first view and *ALL* pixels in the second view
- The cross-view attention map is biased toward the epipolar line, but can drift away

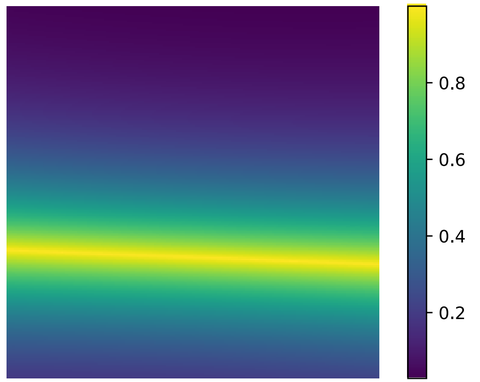
Query Pixel



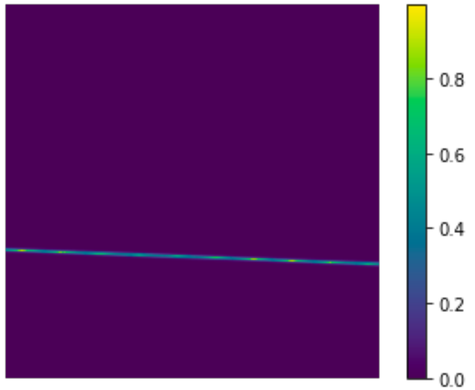
Epipolar Line



Epipolar Field ($\gamma = 1$)

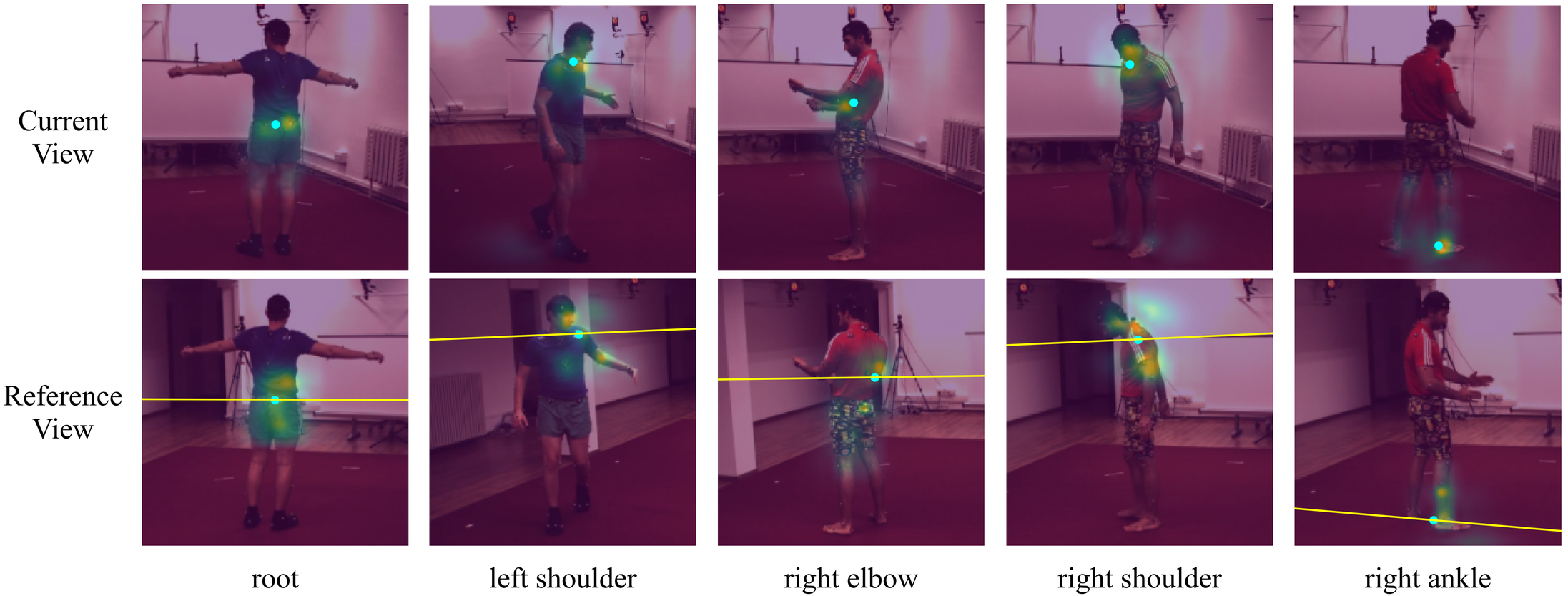


Epipolar Field ($\gamma = 10$)



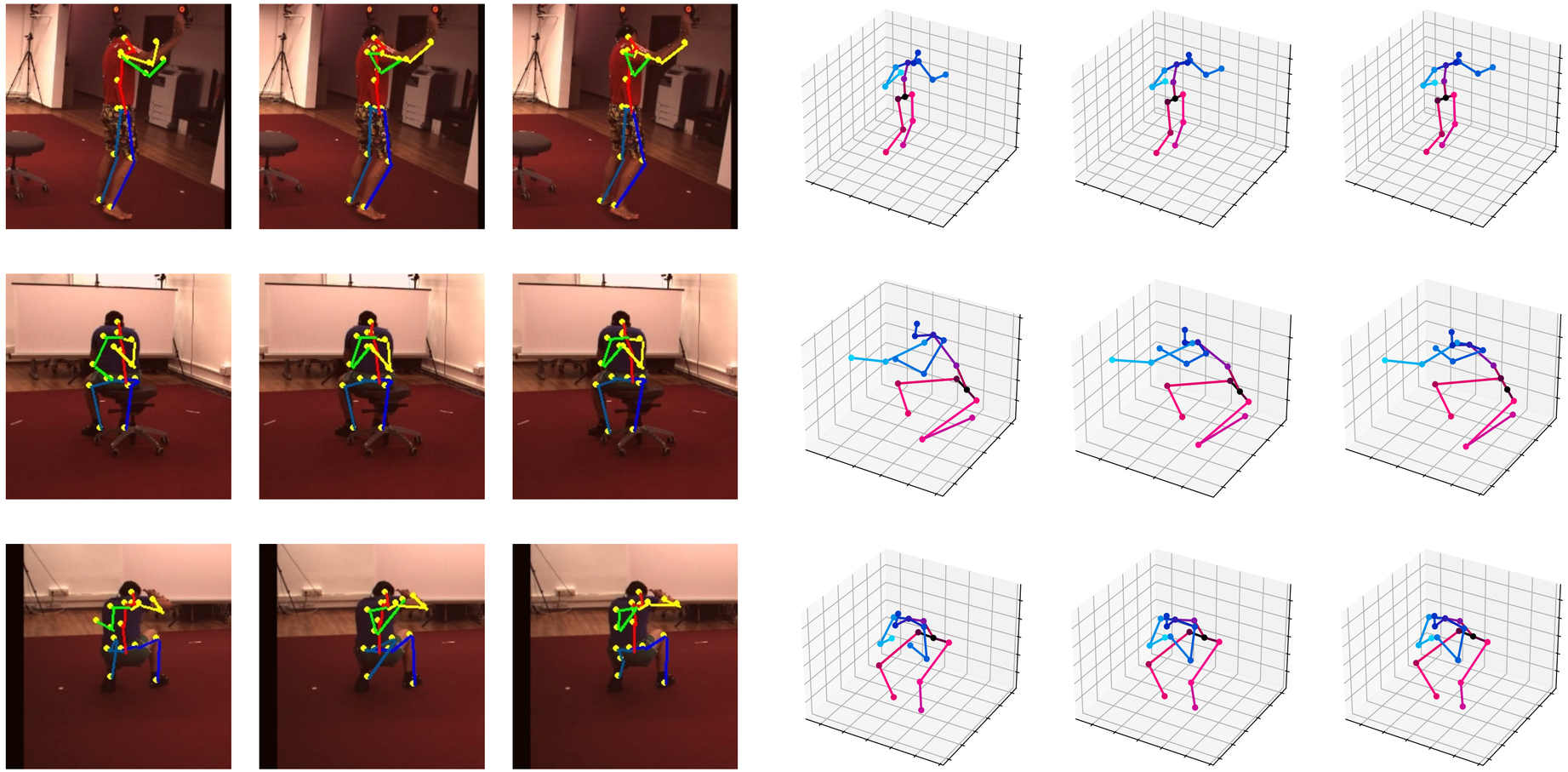
Epipolar Field ($\gamma = 1000$)

Results: Attention Map



Qualitative Results

Baseline (2D) | Ours (2D) | Ground Truth (2D) | Baseline (3D) | Ours (3D) | Ground Truth (3D)



Thanks!