Microsoft

VoxelPose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment

Chunyu Wang
Microsoft Research Asia

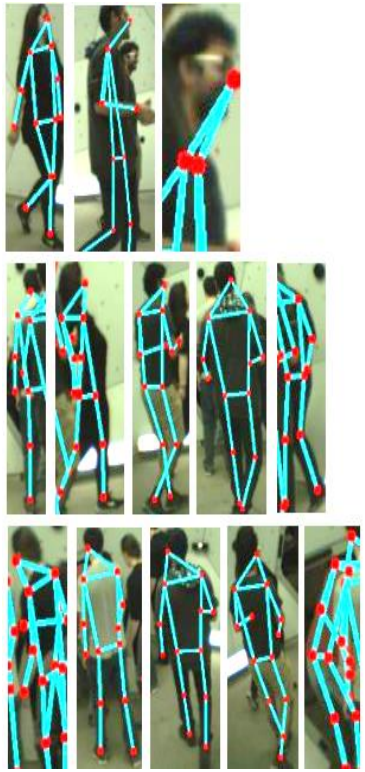
https://github.com/microsoft/voxelpose-pytorch

Broad Impact

- Intelligent retail (Microsoft Connected Store)
- Sports broadcasting/training/judging
- Human-robot interaction
- Augmented/virtual reality

Previous Work

Multiview Images → 2D Keypoint Estimation → Cross view Matching → Triangulation or Pictorial Model → Multi-person 3D Poses



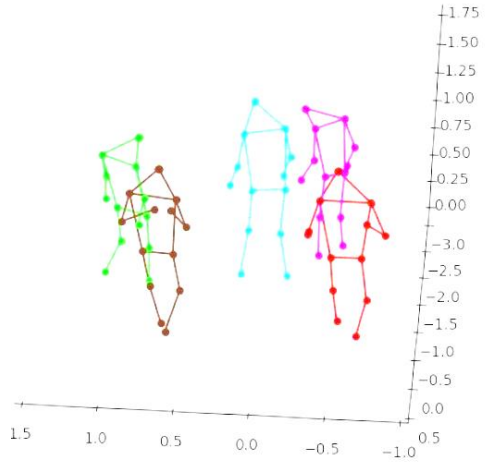
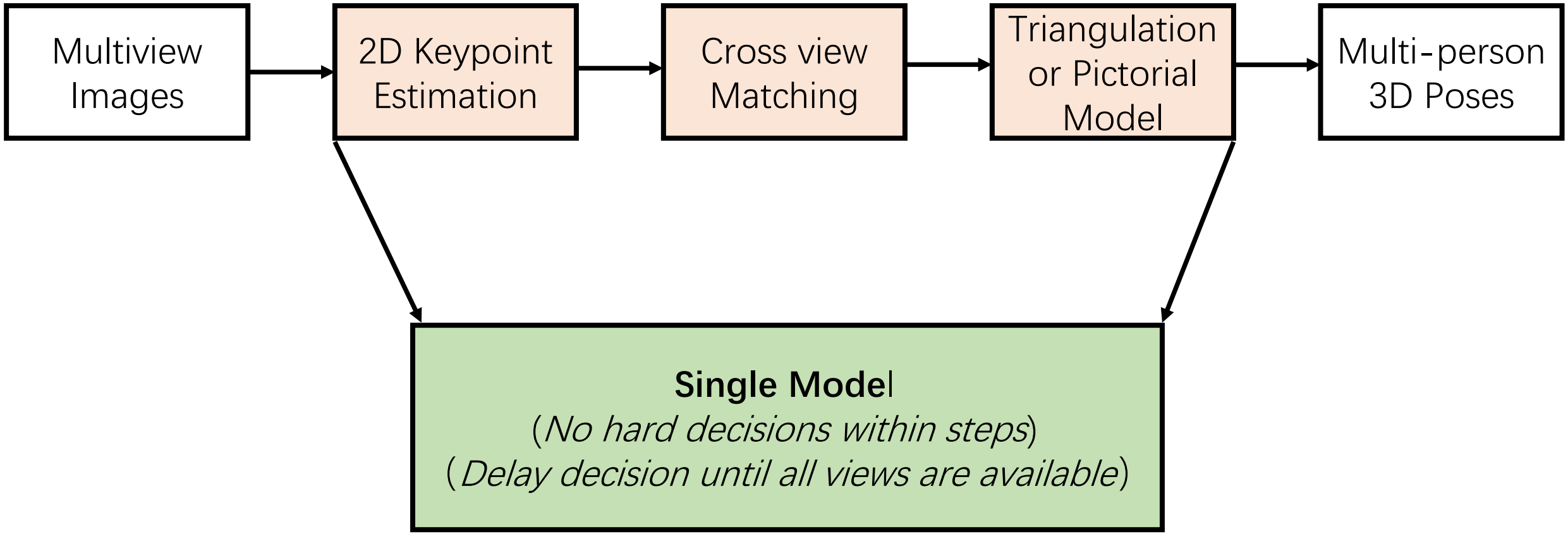


Image Credit: Dong, Junting, et al. “Fast and robust multi-person 3d pose estimation from multiple views.”, CVPR 2019

VoxelPose


Multiview Images
2D Keypoint Estimation
Cross view Matching
Triangulation or Pictorial Model
Multi-person 3D Poses
Single Model
(No hard decisions within steps)
(Delay decision until all views are available)

VoxelPose

VoxelPose

1. Discretize 3D Space by Voxels

VoxelPose

1. Discretize 3D Space by Voxels

2. Compute a feature for each voxel by inversely projecting 2D features to 3D

VoxelPose

1. Discretize 3D Space by Voxels
2. Compute a feature for each voxel by inversely projecting 2D features to 3D
3. The resulting feature is robust to occlusion

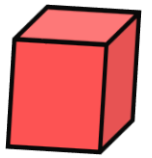
VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

VoxelPose

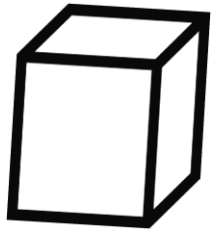
1. Discretize 3D Space by Voxels
2. Compute a feature for each voxel by inversely projecting 2D features to 3D
3. The resulting feature is robust to occlusion
4. Predict whether each voxel contains body joints

VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

Hybrid Model- (1) Human Detection



(300mm x 300mm x 300mm)



(2000mm x 2000mm x 2000mm)

The proposals need not to be very precise since we will refine them in the following step.

VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

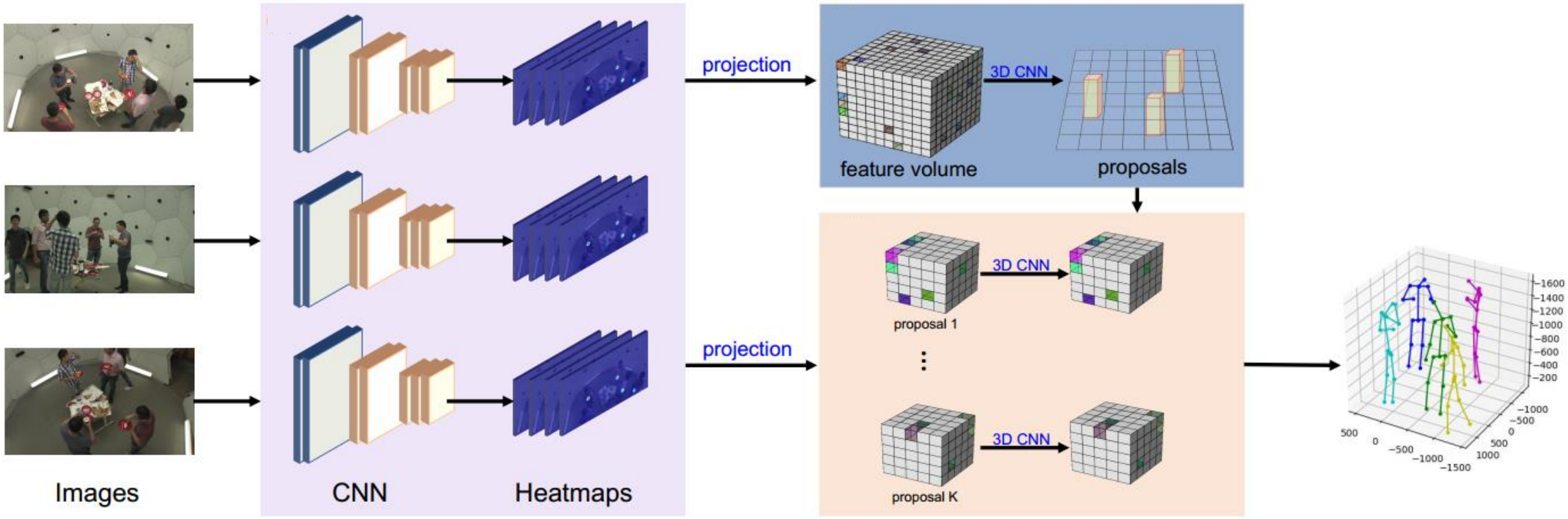
Hybrid Model- (2) Joint Detection

(30mm x 30mm x 30mm)

This is sufficiently accurate for body joint localization.

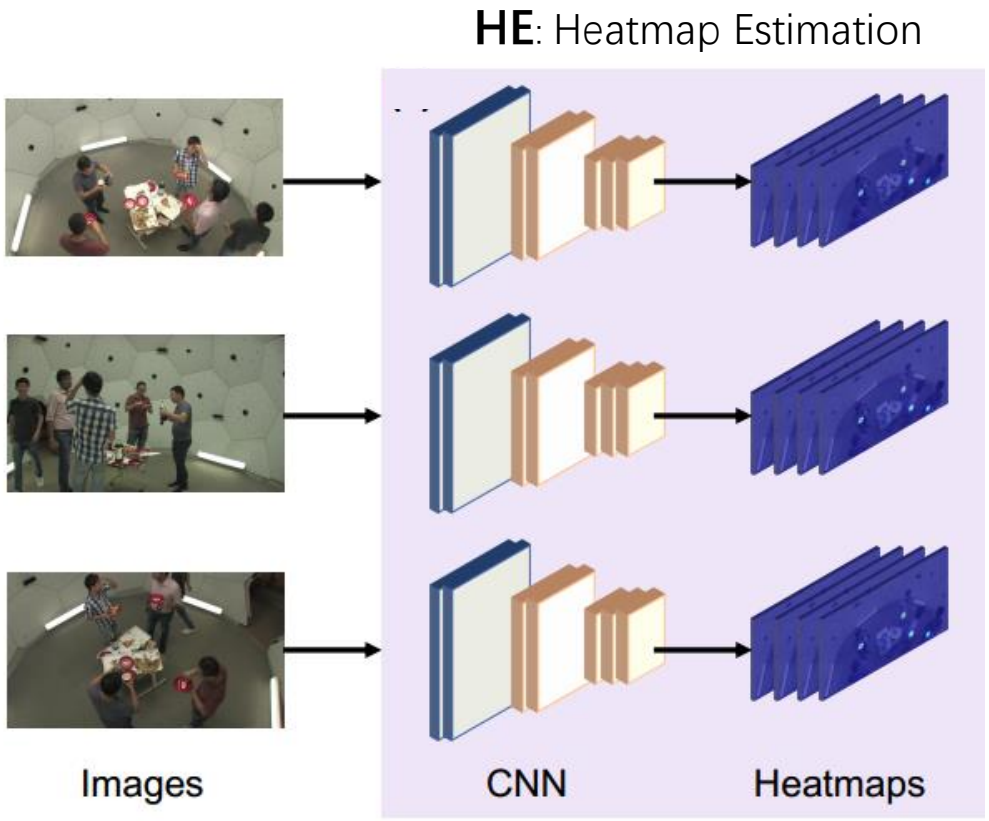
VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

Technical Details of VoxelPose



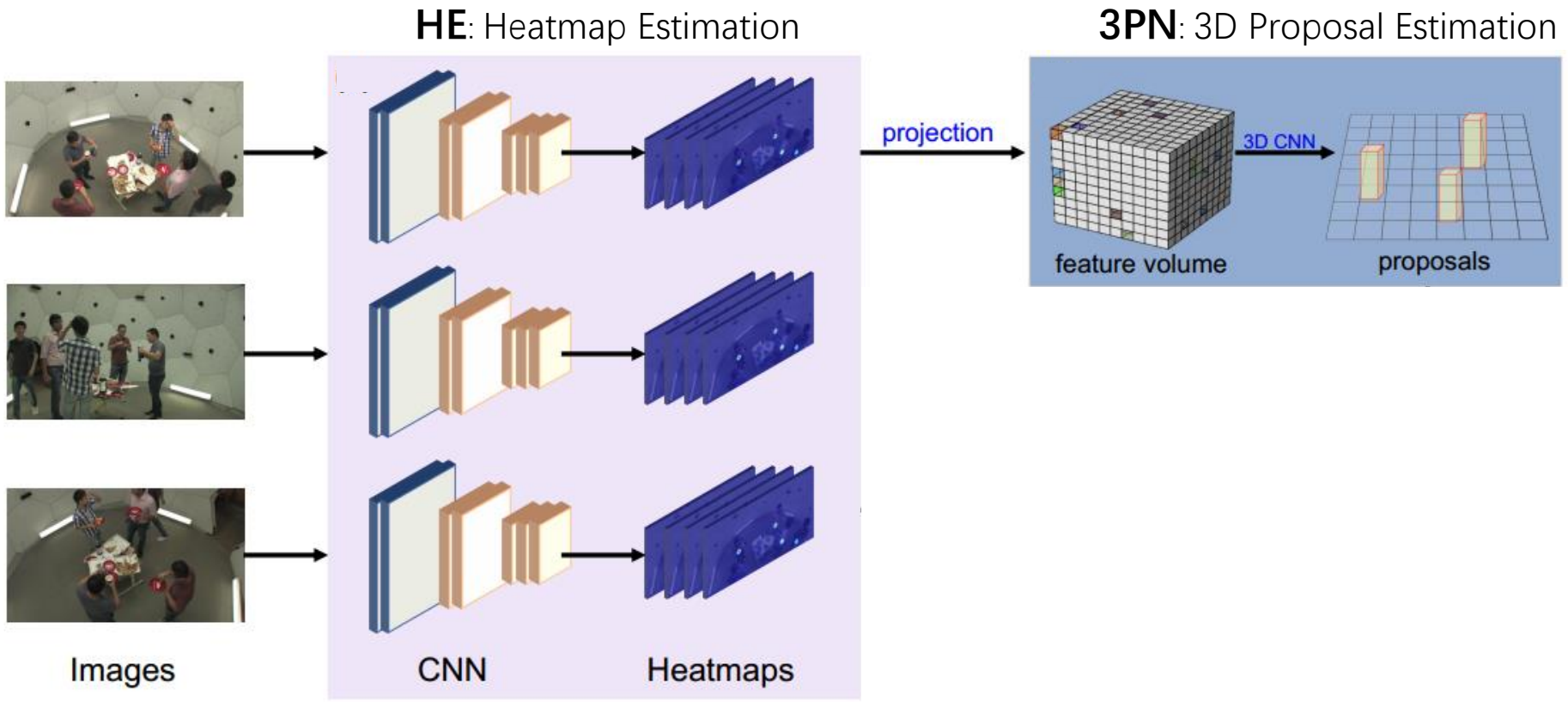
VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

Step 1: 2D Heatmap Estimation

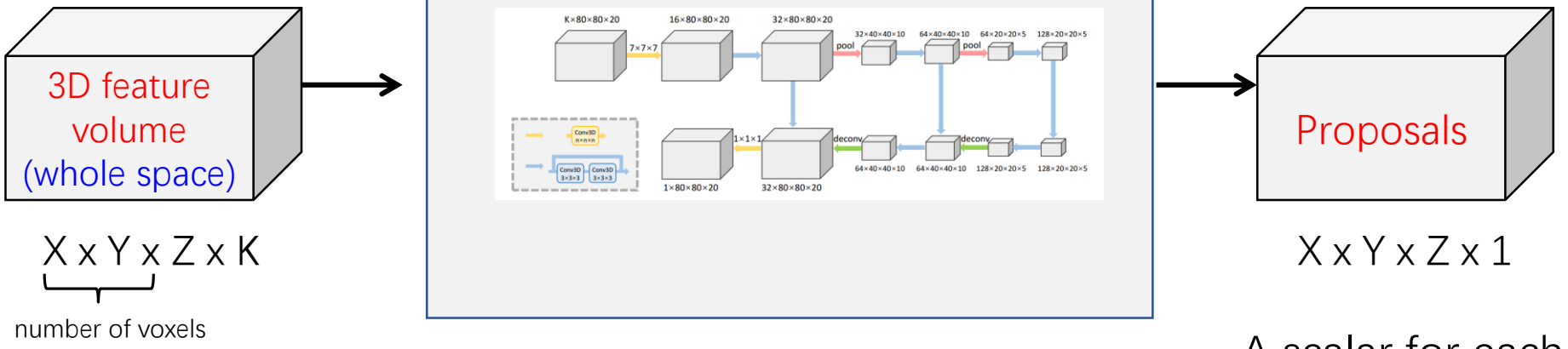


It can use the existing methods such as OpenPose, HRNet and AlphaPose.

Step 2: 3D Person Detection



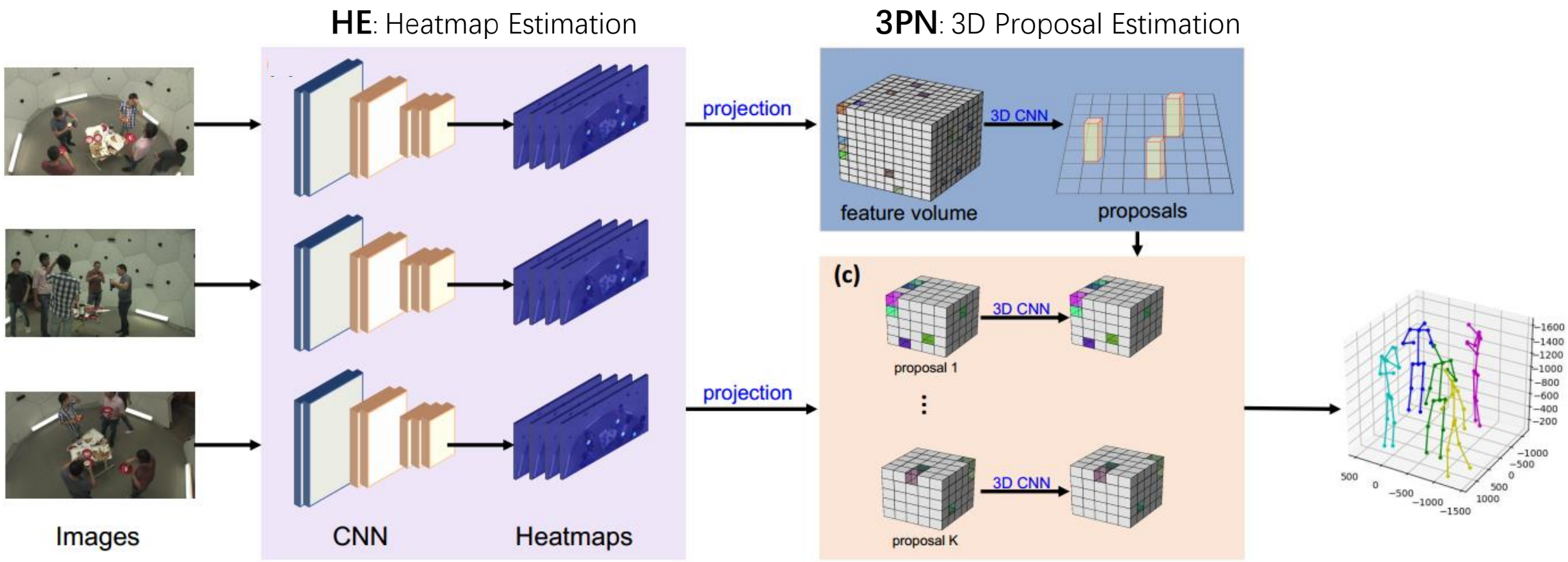
Step 2: 3D Person Detection



A scalar for each voxel:
The likelihood of having a people centered at the voxel

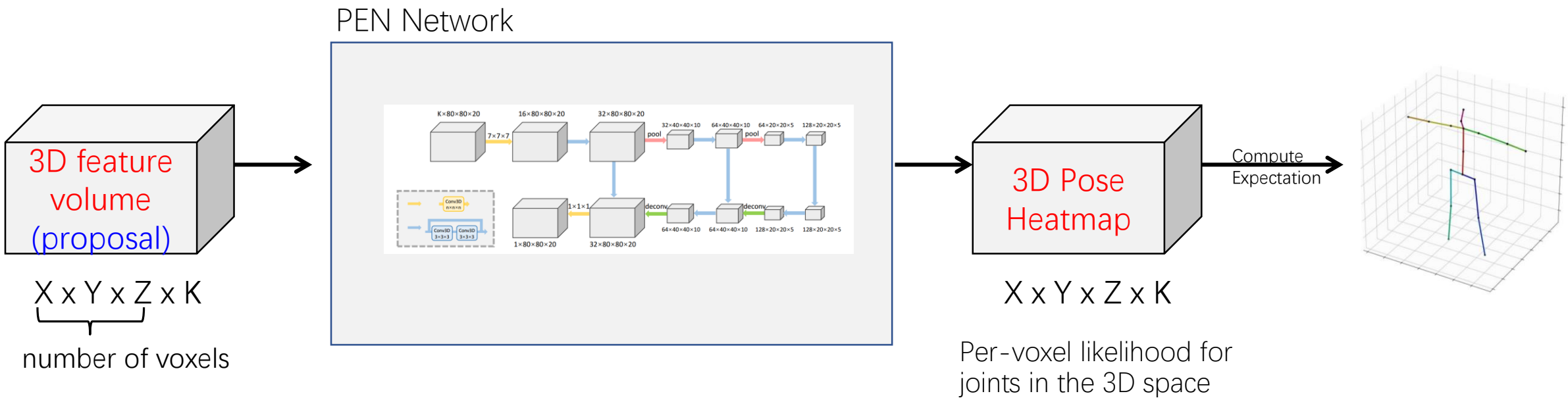
We keep K largest voxels (proposals) after NMS

Step 3: 3D Joint Detection

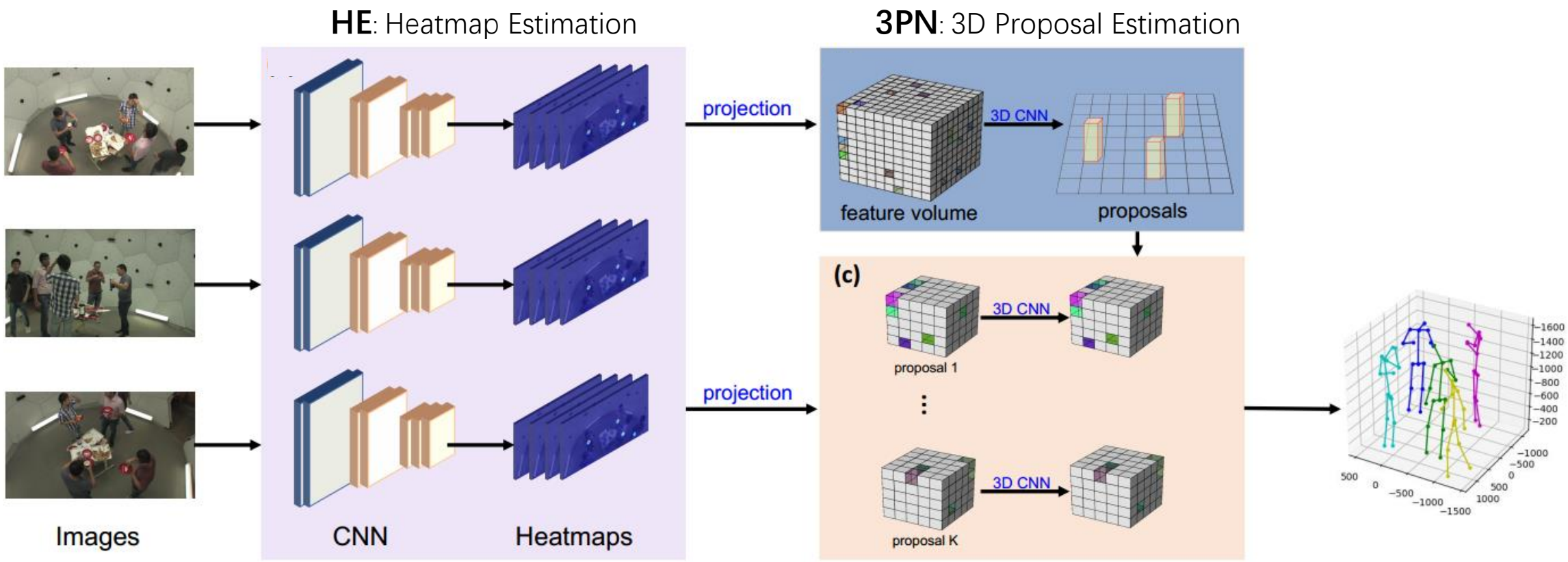


VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

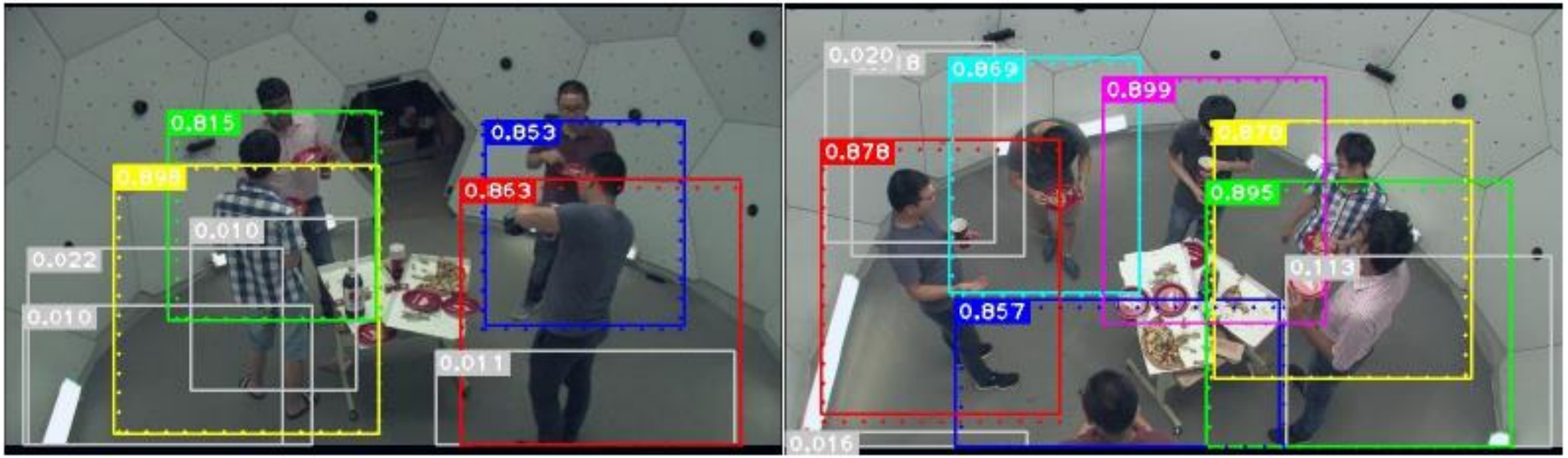
Step 3: 3D Joint Detection



Joint Training

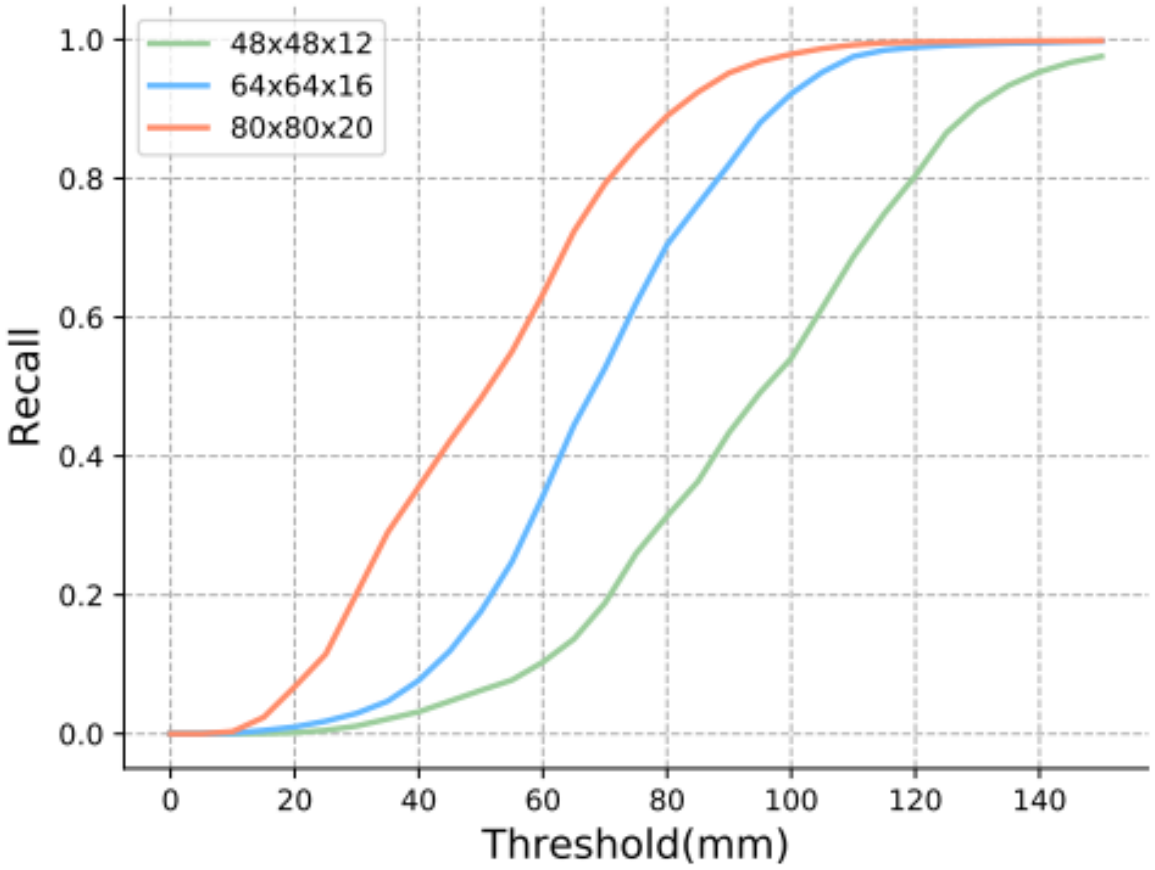


Proposal Quality



We project the 3D proposals to 2D for visualization. Colored boxes represent their estimated confidence is larger than 0.1.

Proposal Recall Rate



- When the threshold is 140mm, we get about 95% recall when voxel size is 300mm
- This is sufficient for 3D pose estimation
- Using a smaller voxel improves the precision

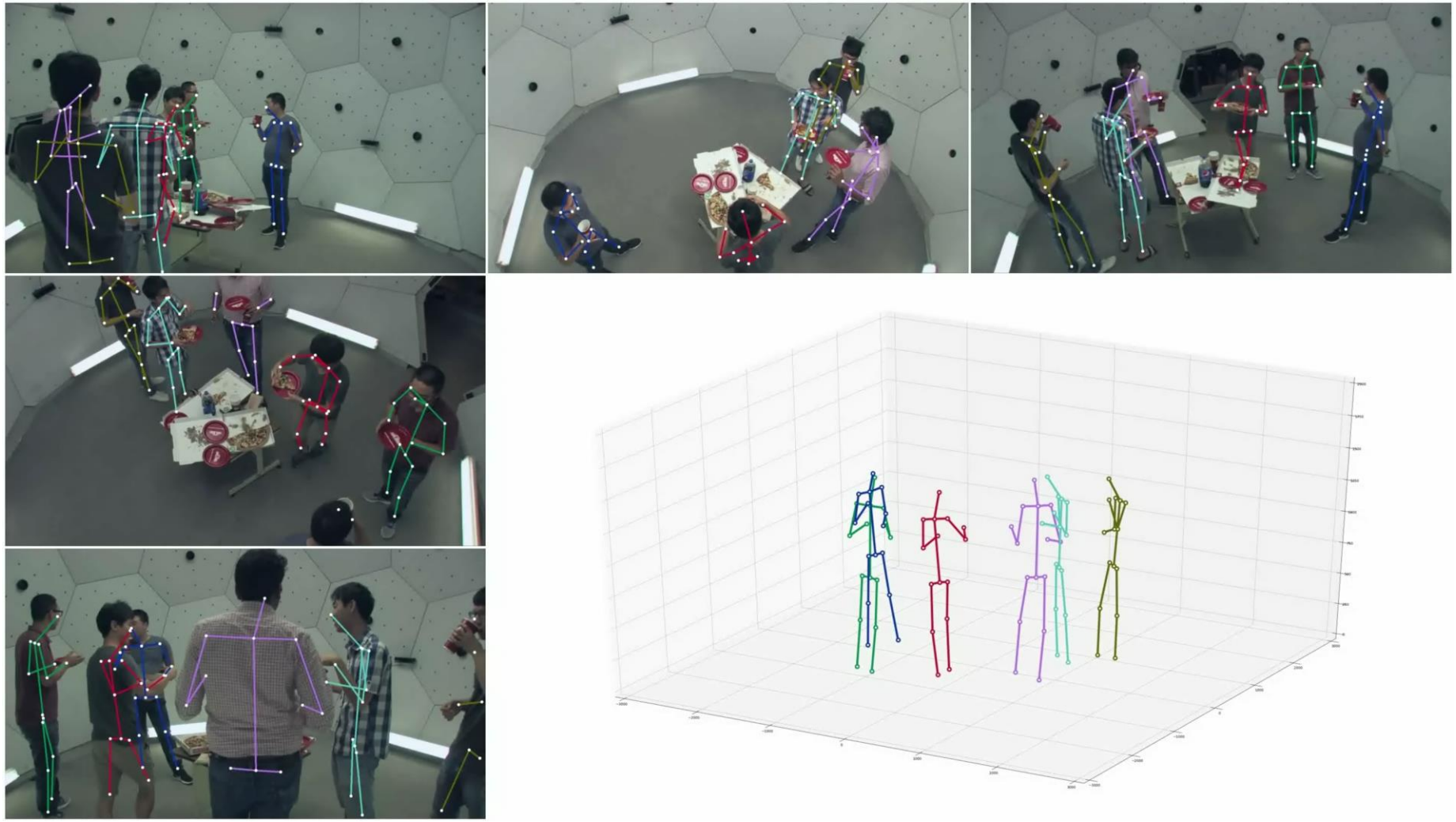
Impact of Camera Number

Camera Number	AP^{25} ↑	AP^{50} ↑	AP^{100} ↑	AP^{150} ↑	MPJPE ↓
5	83.59	98.33	99.76	99.91	17.68mm
3	58.94	93.88	98.45	99.32	24.29mm
1	0.860	23.47	80.69	93.32	66.95mm
5*	50.91	95.25	99.36	99.56	25.51mm

* means training/testing on different cameras.

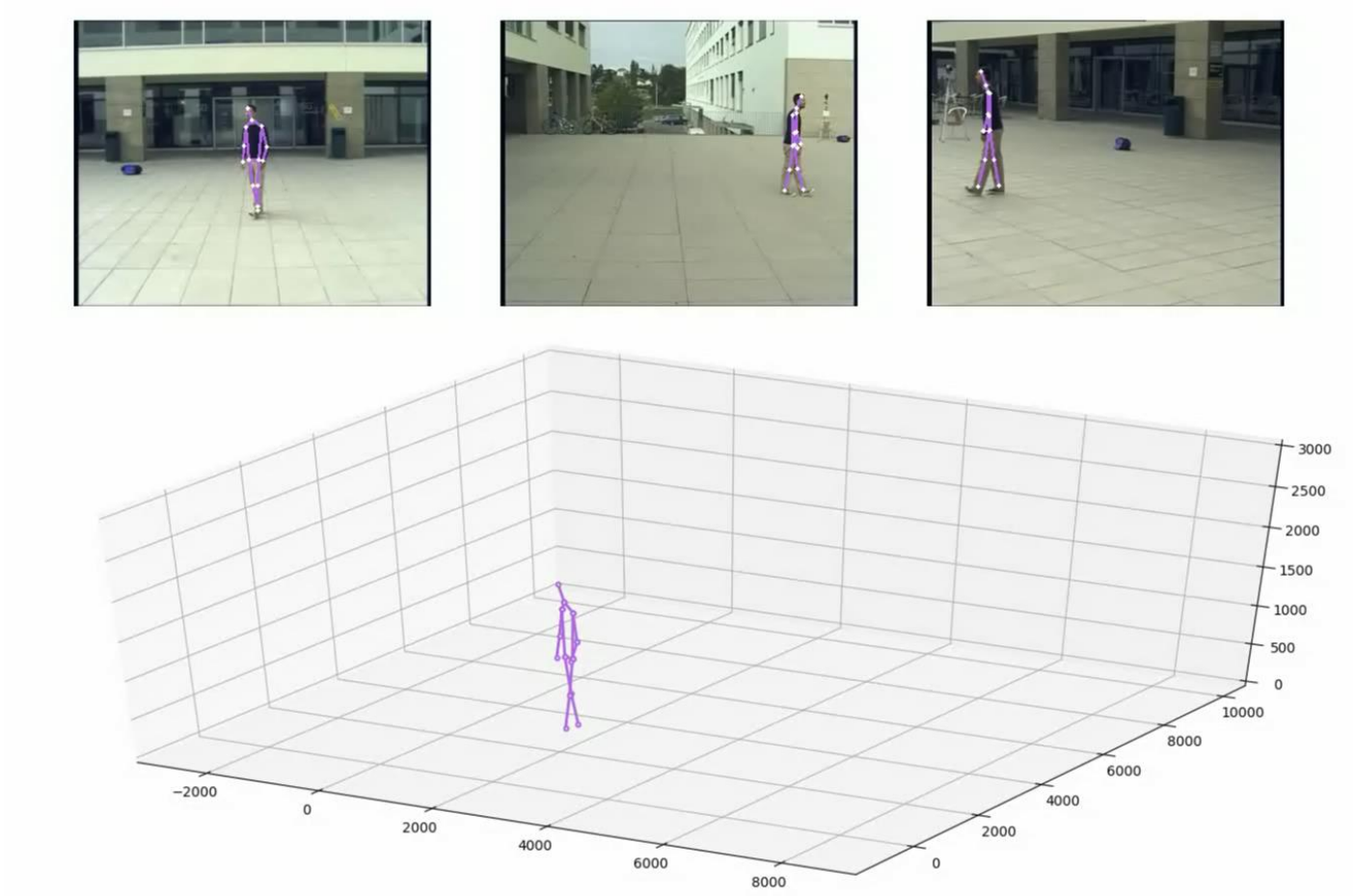
- The error increases mildly when we decrease the number from 5 to 3.
- The error increases notably when using only one camera.
- It generalizes to different camera configurations.

Demo



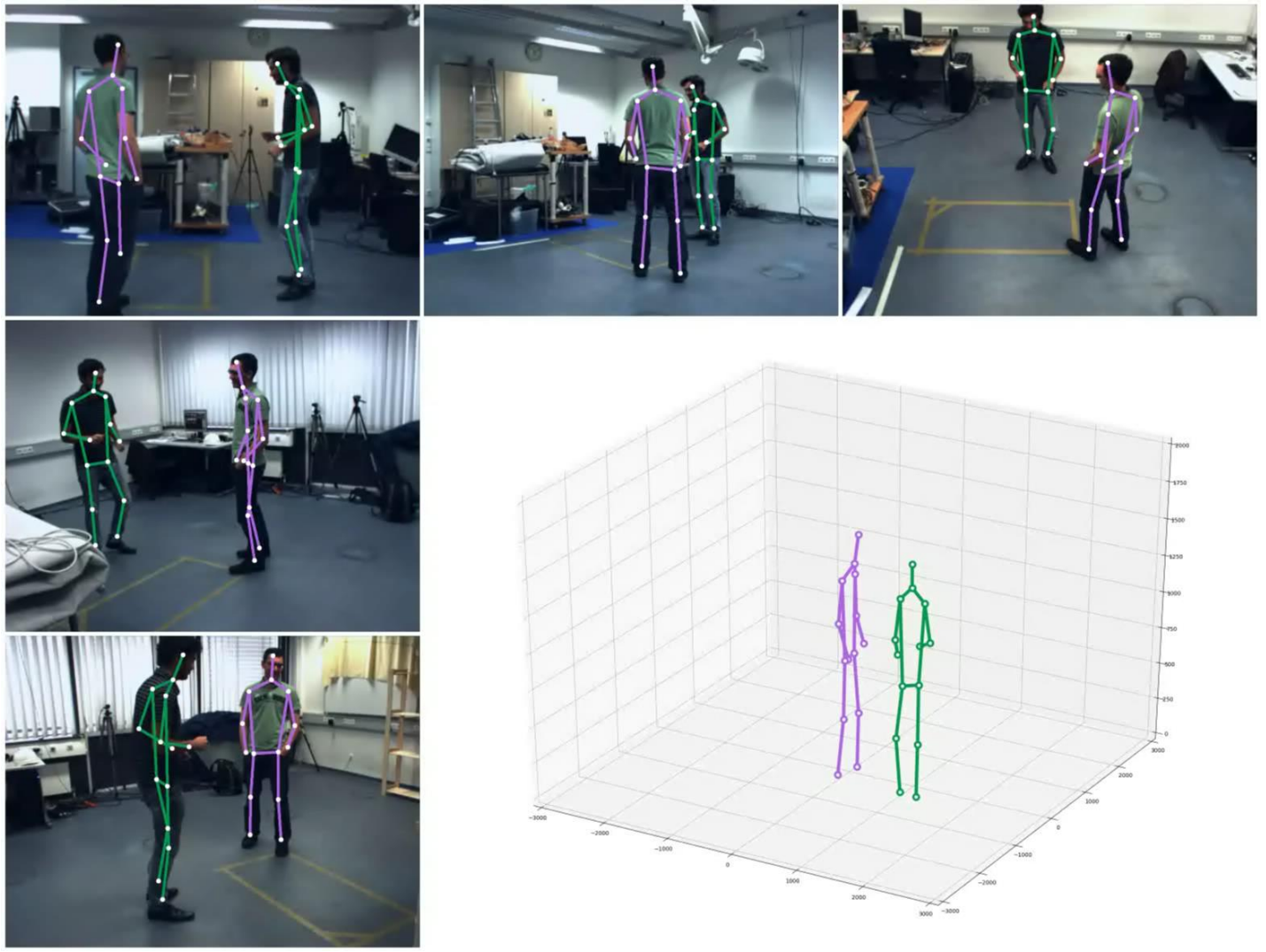
VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

Demo



VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng

Demo



VoxelPose: Hanyue Tu, Chunyu Wang, Wenjun Zeng