Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views

董峻廷

个人简介：

董峻廷，浙江大学硕士生，指导老师为周晓巍教授，研究方向为计算机视觉，主要专注于3D vision，特别是**3D human pose estimation**，个人主页：http://jtdong.com/

Pipeline：

1. Background
2. Related work
3. Our approach
4. Results

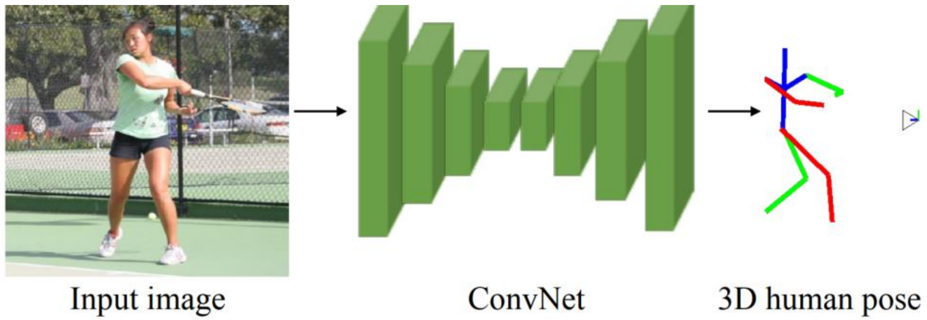
1. Background

3D human pose estimation的**定义**:

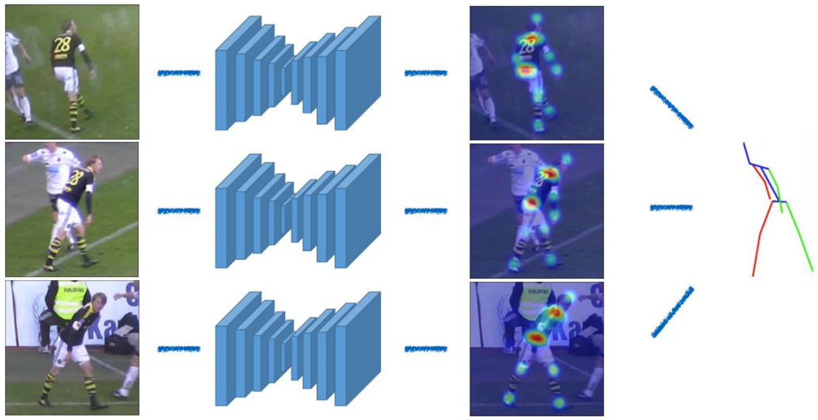
Input： images

Output： 3D human pose（N*3的一组关键点）

3D human pose from single view

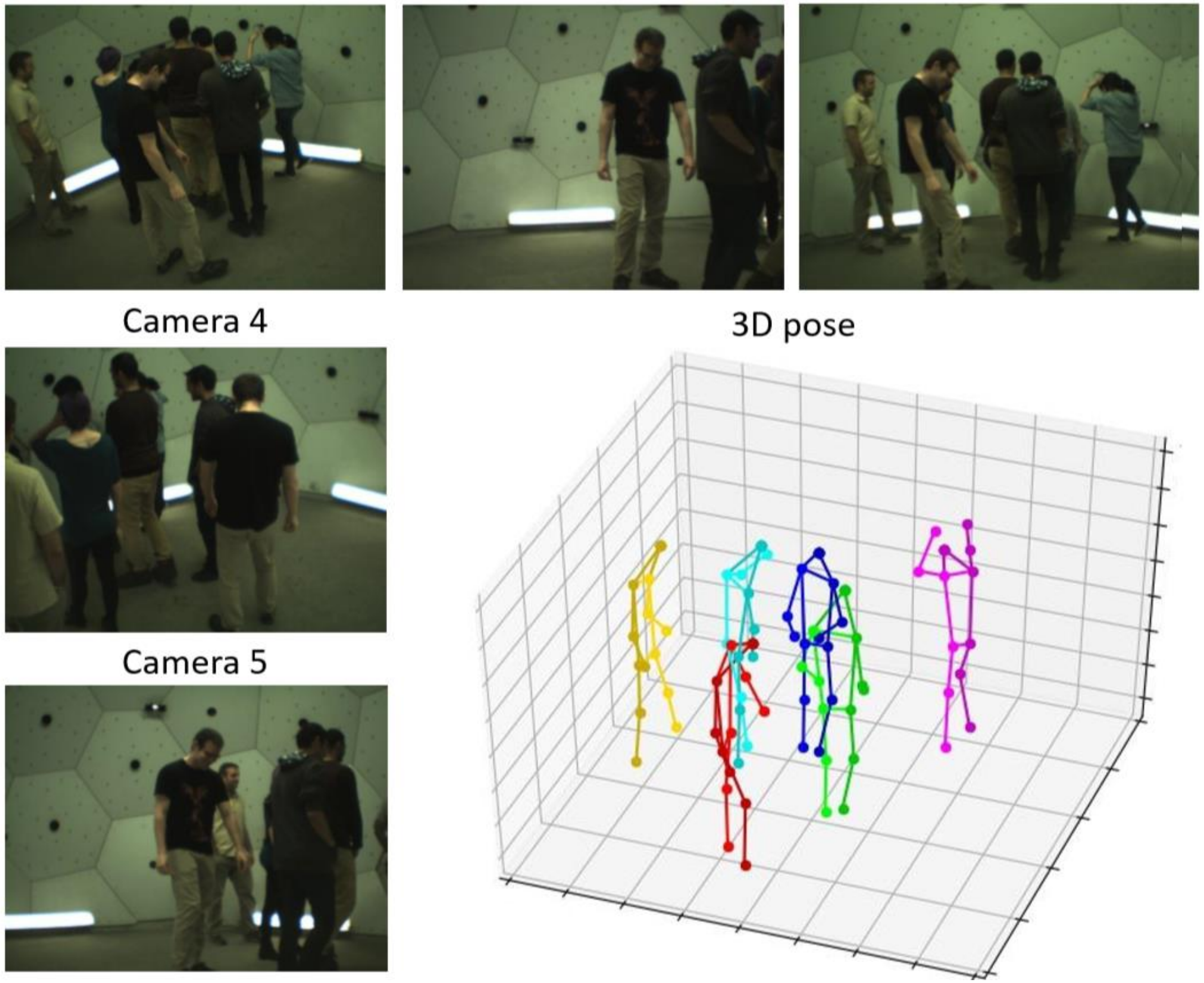


3D human pose from multiple views



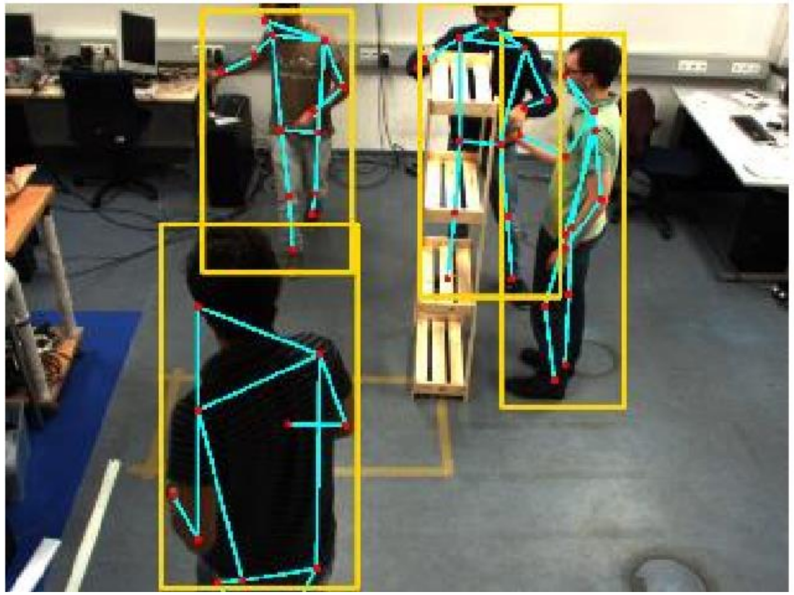
Harvesting Multiple Views for Marker-less 3D Human Pose Annotations. CVPR 2017

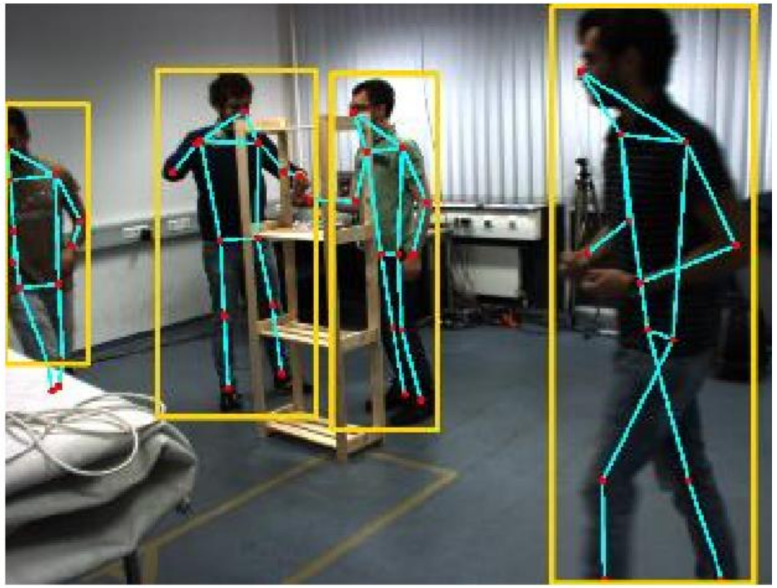
Crowd scene



1. Background

Main challenge： **Finding correspondence is hard!**

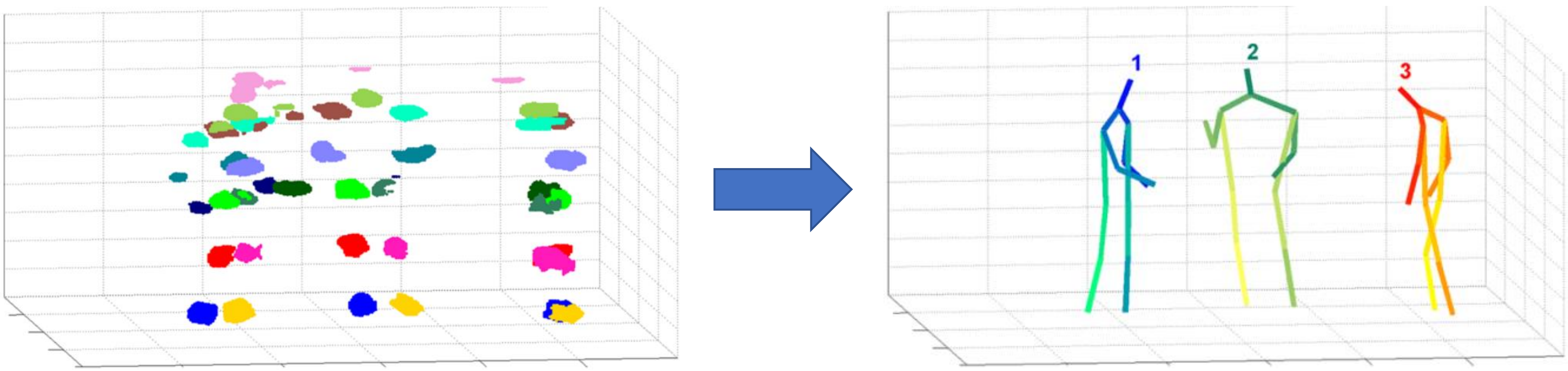




2. Related work

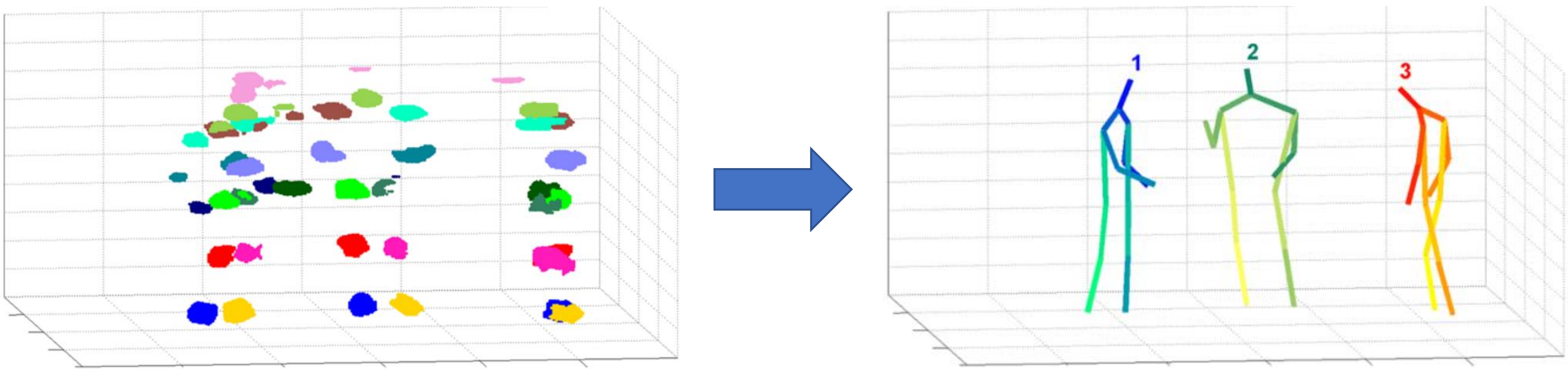
之前方法:

1. 构建一个所有人的common state space
2. 使用3D pictorial structure去做inference



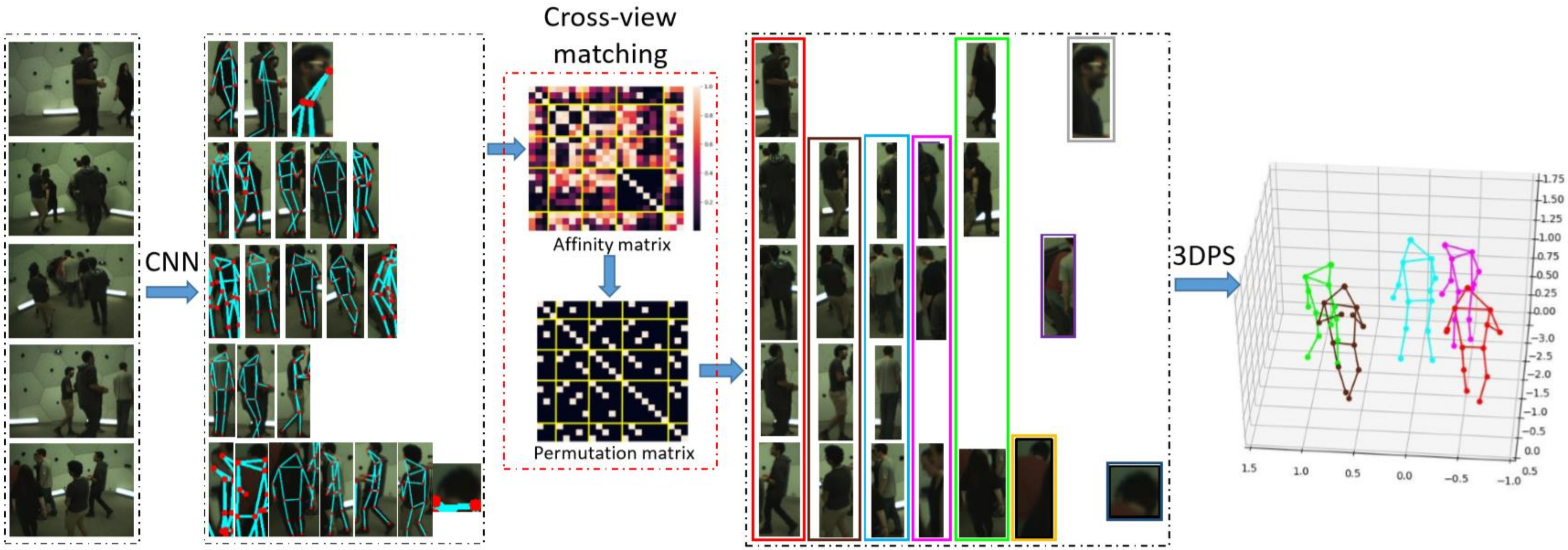
缺点：

1. State space太大，inference速度很慢
2. 只利用几何约束去找correspondence，不够鲁棒



3. Our approach

Pipeline:

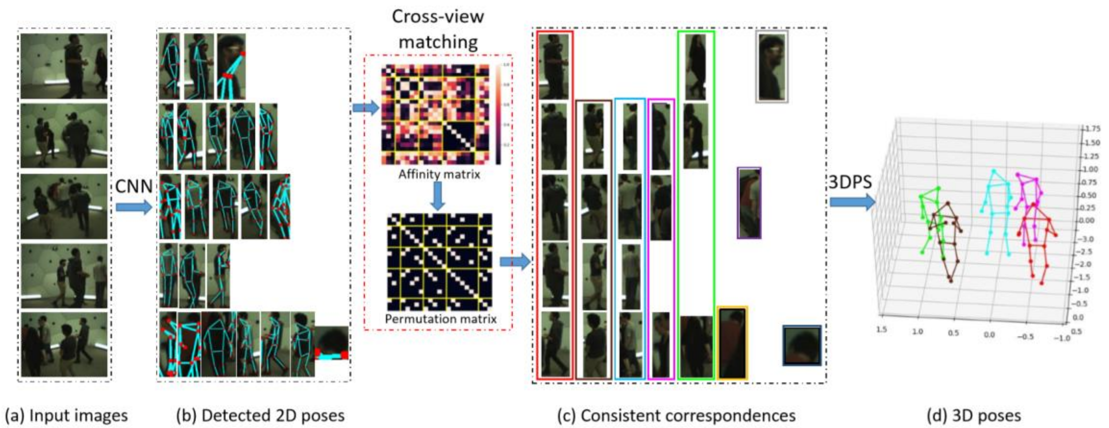


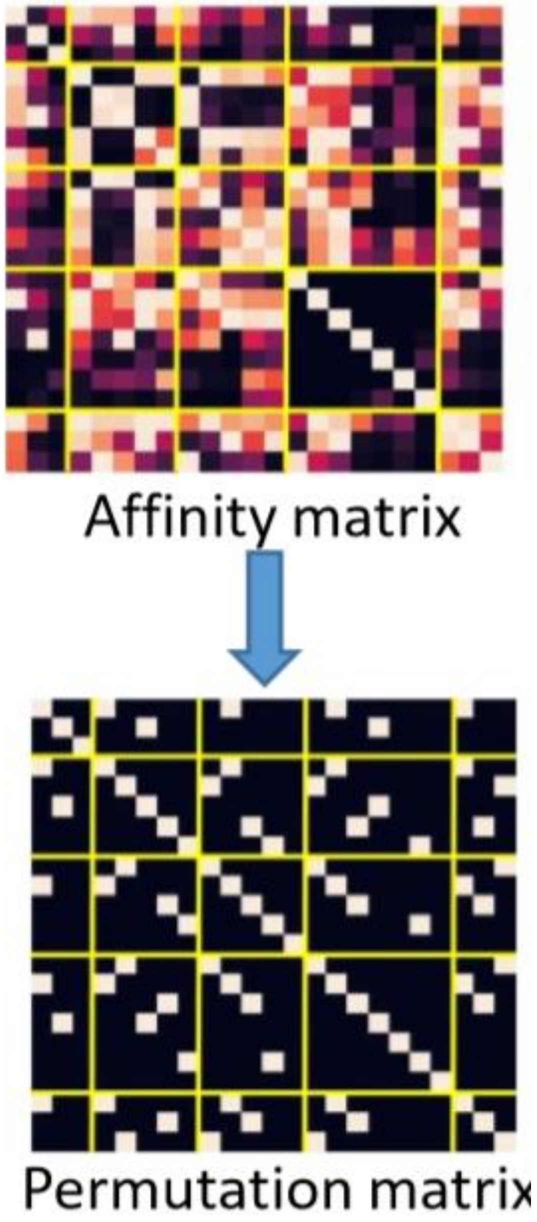
(a) Input images (b) Detected 2D poses (c) Consistent correspondences (d) 3D poses

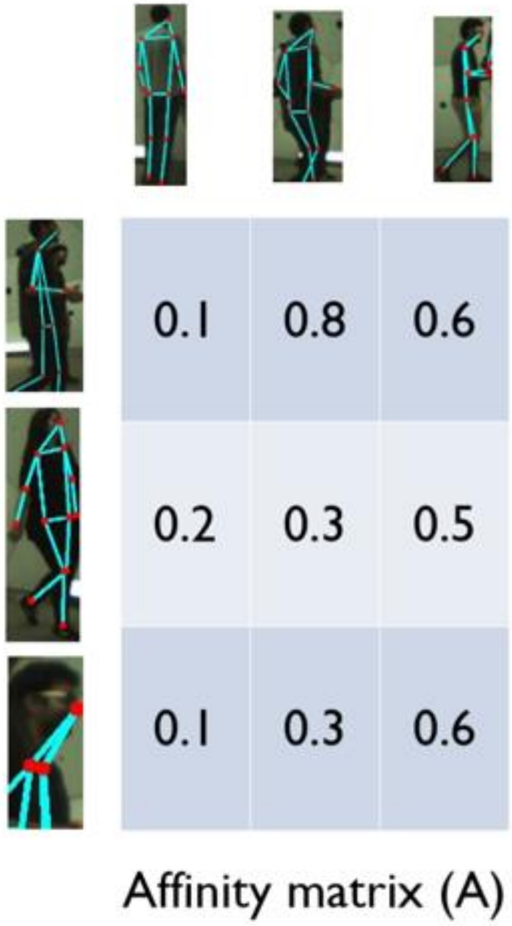
3. Our approach

Construct the Affinity matrix:

Idea: combining appearance and geometry



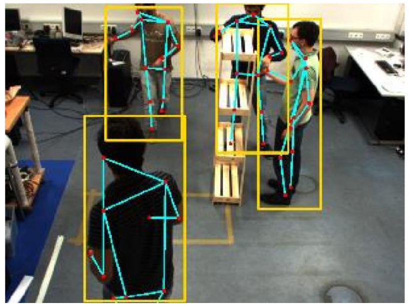


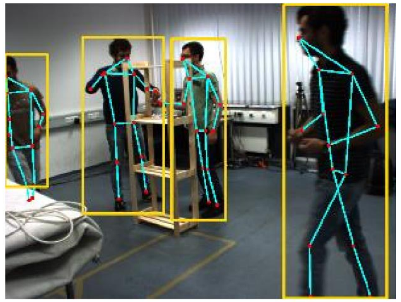


3. Our approach

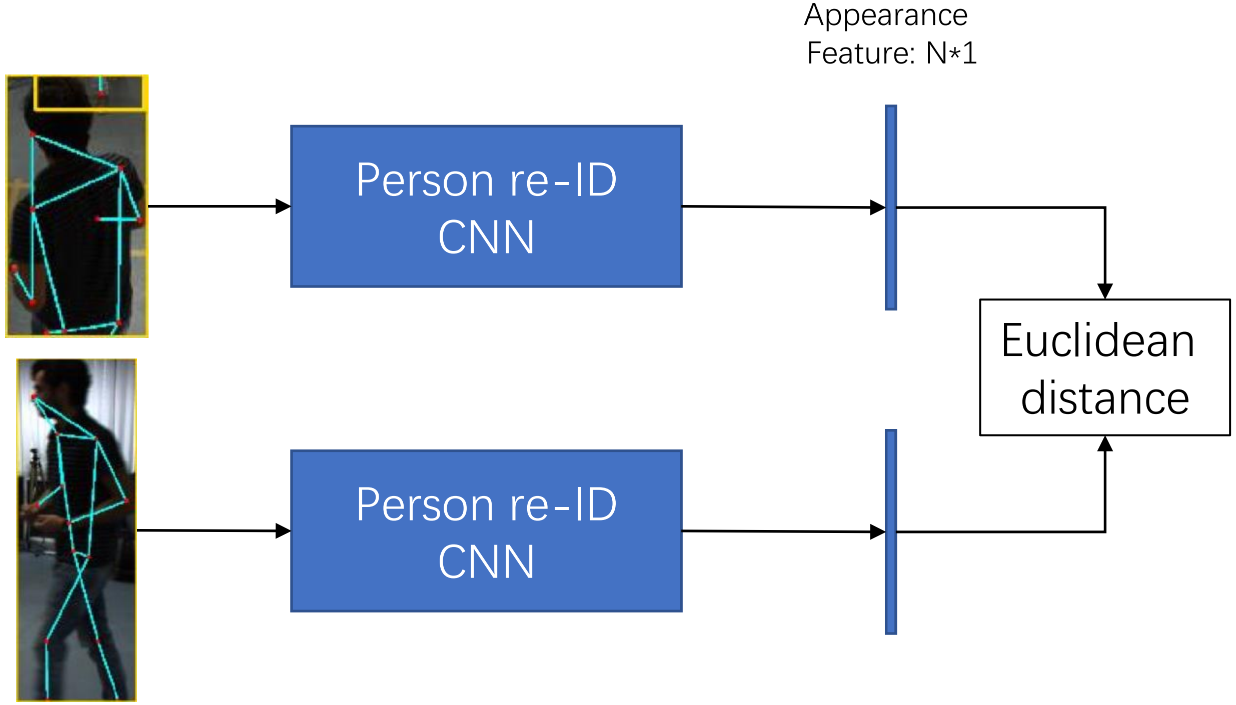
Construct the Affinity matrix

Idea: combining appearance and geometry



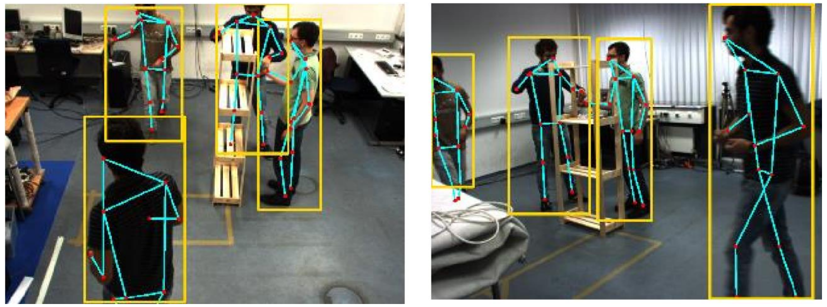


Use re-identification network to measure appearance consistency

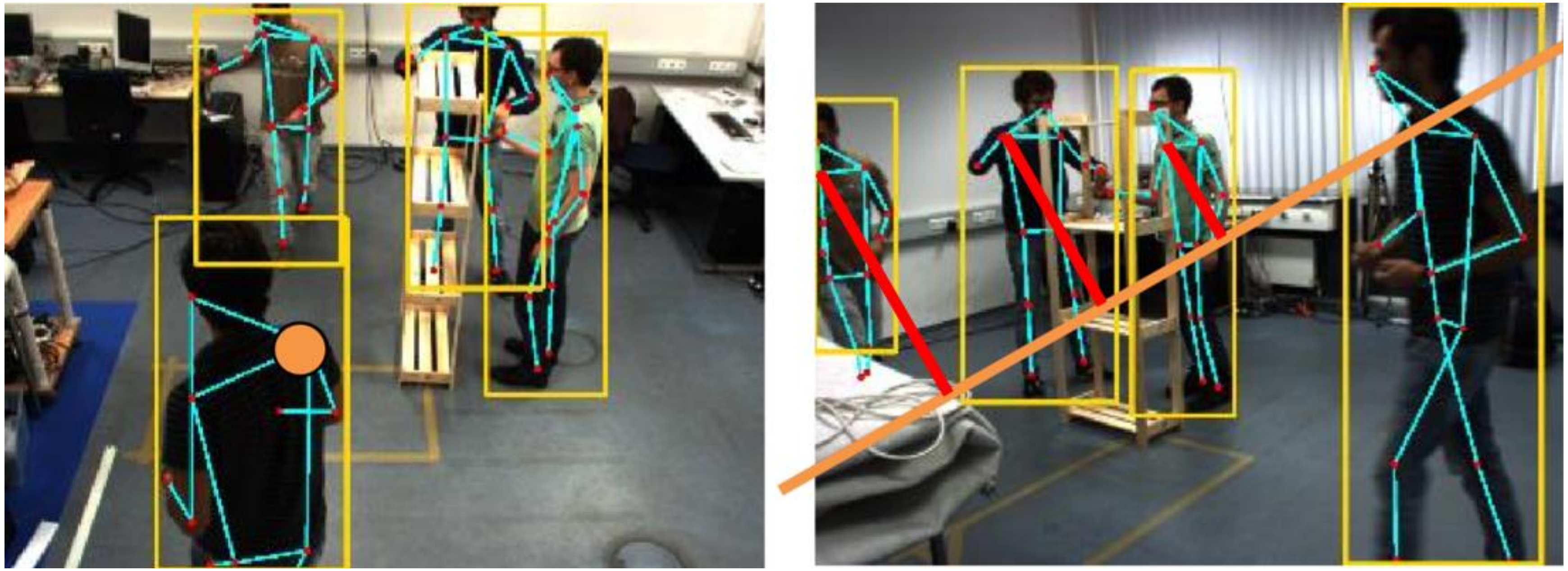


Construct the Affinity matrix

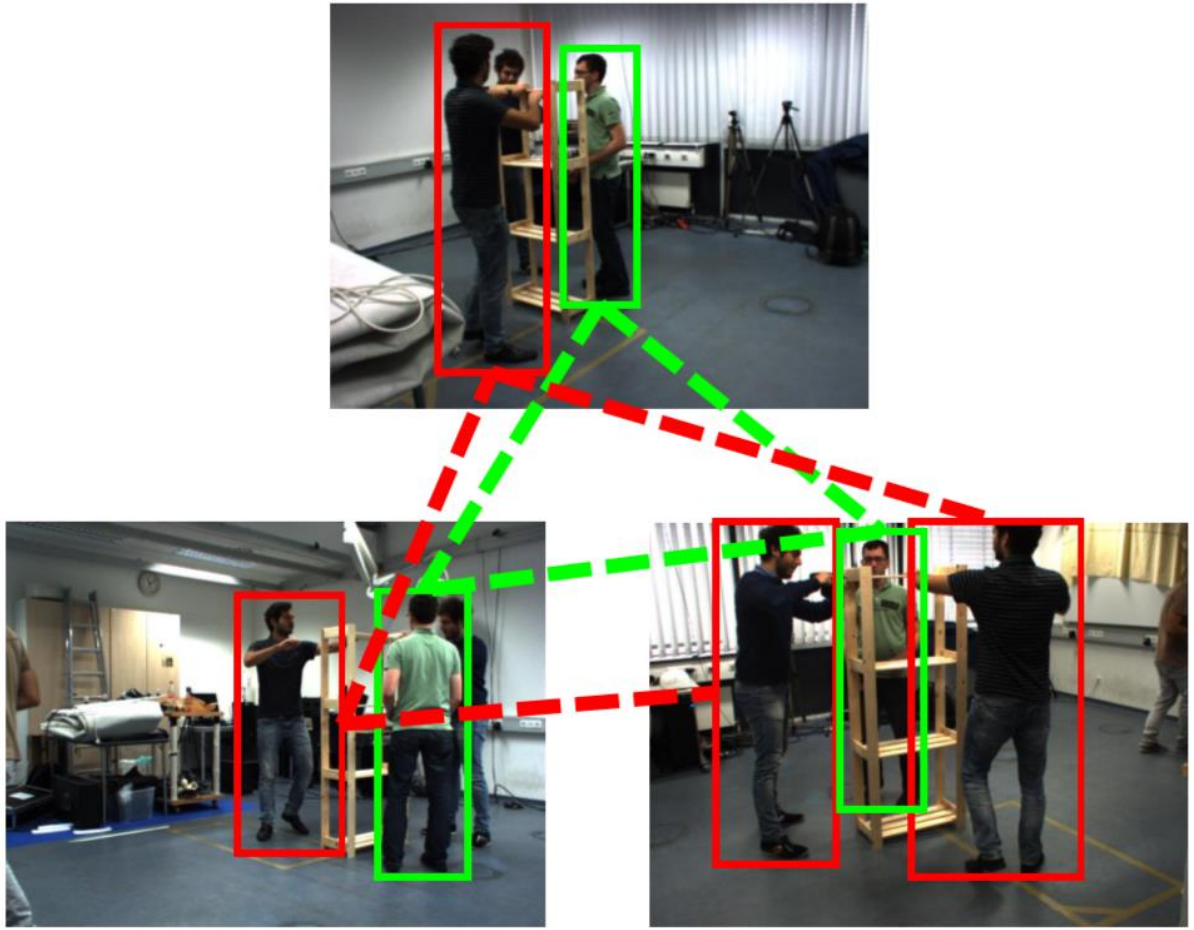
Idea: combining appearance and geometry



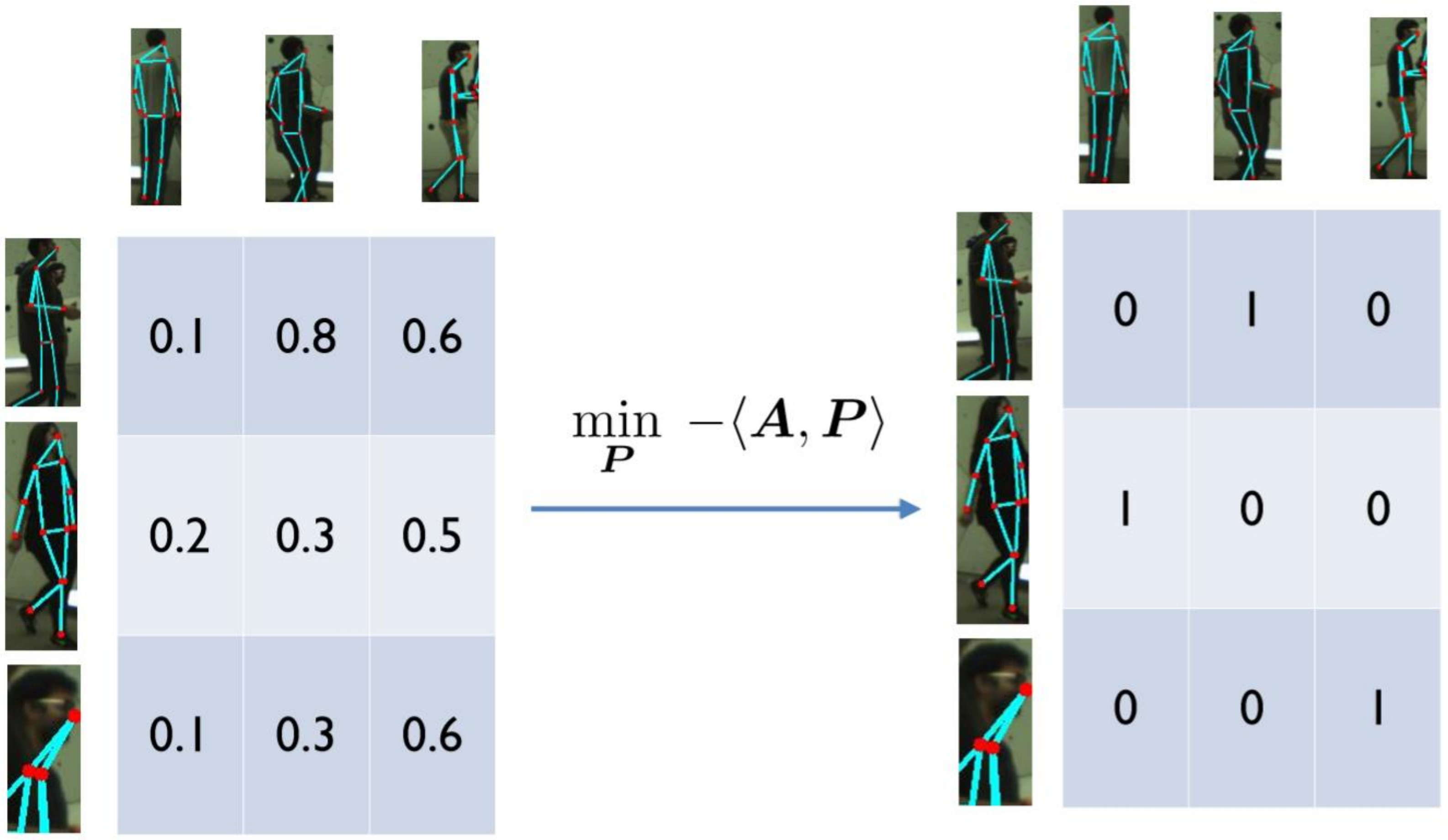
Use epipolar constraint to measure geometric consistency



Idea: using cycle-consistency constraint



Matching two views

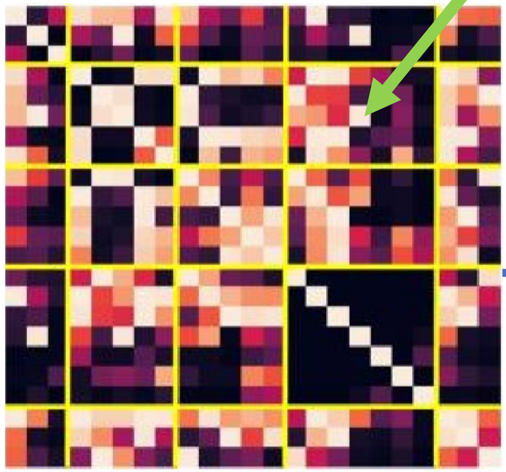


Affinity matrix (A)

Permutation matrix (P)

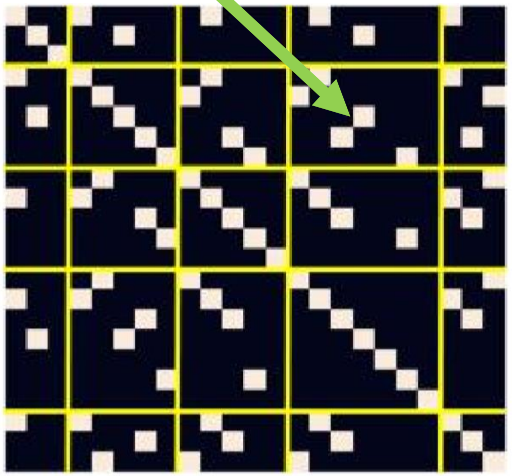
Matching multiple views

Each block is an affinity or permutation matrix between two images



Affinity matrix

$$\min_{\boldsymbol{P}} \; -\langle \boldsymbol{A}, \boldsymbol{P} \rangle + \lambda \|\boldsymbol{P}\|_*$$



Permutation matrix

This should a low-rank matrix if the cycle consistency is satisfied [Huang et al. 2013]

3. Our approach

求解优化问题：

$$
\begin{aligned}
\min_{\boldsymbol{P}} \quad & -\langle \boldsymbol{A}, \boldsymbol{P} \rangle + \lambda \|\boldsymbol{P}\|_*, \\
\text{s.t.} \quad & \boldsymbol{P} \in \mathcal{C},
\end{aligned}
$$

Rewrite as follows by introducing an auxiliary variable **Q**

$$
\begin{aligned}
\min_{\boldsymbol{P}, \boldsymbol{Q}} \quad & -\langle \boldsymbol{A}, \boldsymbol{P} \rangle + \lambda \|\boldsymbol{Q}\|_*, \\
\text{s.t.} \quad & \boldsymbol{P} = \boldsymbol{Q}, \ \boldsymbol{P} \in \mathcal{C}.
\end{aligned}
$$

3. Our approach

求解优化问题：

The augmented Lagrangian is:

$$\mathcal{L}_\rho(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{Y}) = -\langle \boldsymbol{A},\boldsymbol{P}\rangle + \lambda\|\boldsymbol{Q}\|_* + \langle \boldsymbol{Y},\boldsymbol{P}-\boldsymbol{Q}\rangle + \frac{\rho}{2}\|\boldsymbol{P}-\boldsymbol{Q}\|_F^2,$$

Optimization:

Algorithm 1: Consistent Multi-Way Matching

Input: Affinity matrix $\boldsymbol{A}$

Output: Consistent correspondences $\boldsymbol{P}$

randomly initialize $\boldsymbol{P}$ and $\boldsymbol{Y} = \boldsymbol{0}$;

while *not converged* **do**

    $\boldsymbol{Q} \leftarrow \mathcal{D}_{\frac{\lambda}{\rho}}(\frac{1}{\rho}\boldsymbol{Y}+\boldsymbol{P})$;

    $\boldsymbol{P} \leftarrow \mathcal{P}_{\mathcal{C}}(\boldsymbol{Q}-\frac{1}{\rho}(\boldsymbol{Y}-\boldsymbol{A}))$;

    $\boldsymbol{Y} \leftarrow \boldsymbol{Y}^k + \rho(\boldsymbol{P}-\boldsymbol{Q})$;

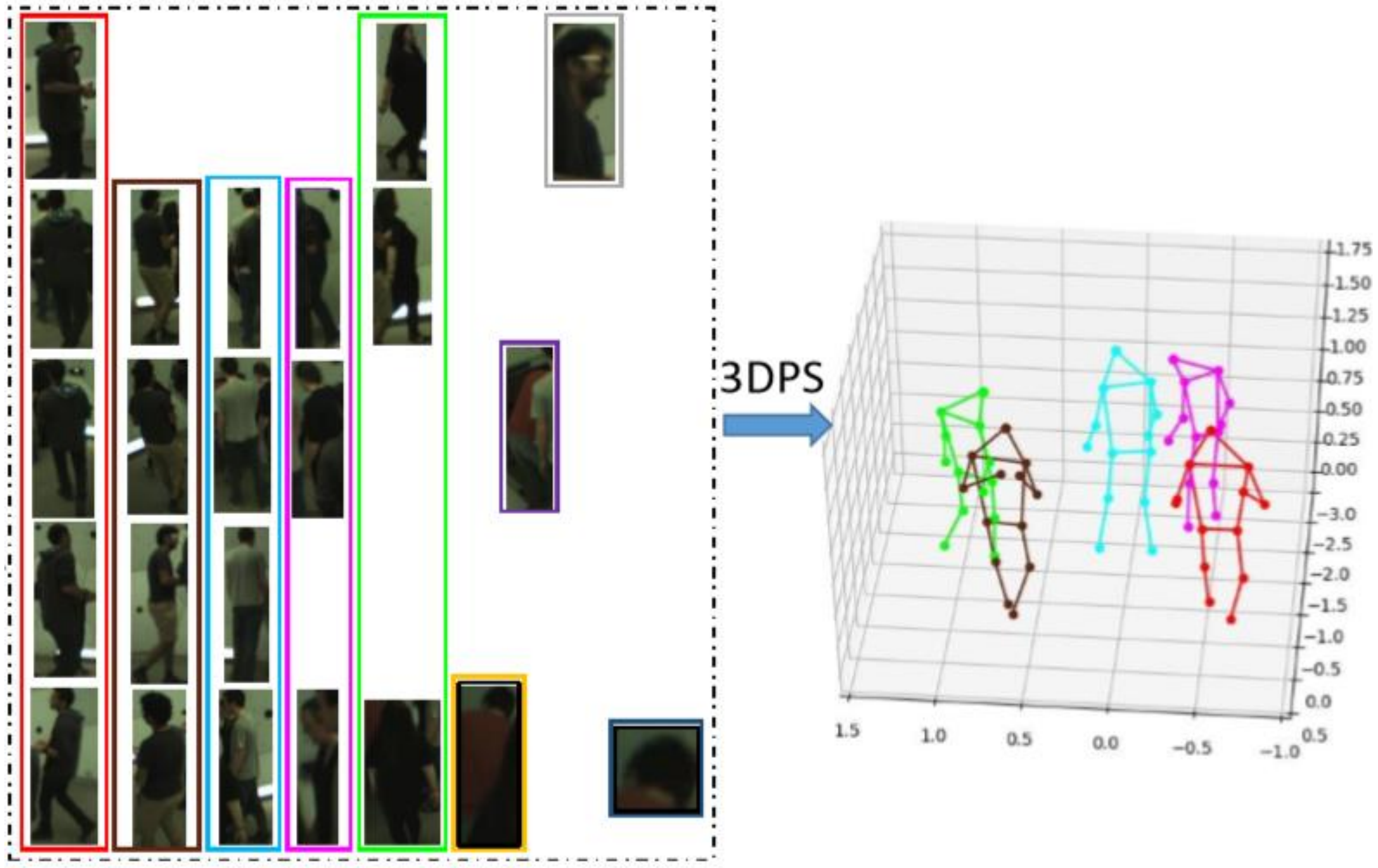
end

quantize $\boldsymbol{P}$ with a threshold equal to 0.5.

$\mathcal{D}$ denotes the operator for singular value thresholding

$\mathcal{P}_{\mathcal{C}}(\cdot)$ denotes the orthogonal projection to $\mathcal{C}$

3D pictorial structure (3DPS)



3D pictorial structure (3DPS)

3D pictorial structure: We use a joint-based representation of 3D poses, i.e., $\boldsymbol{T} = \{\boldsymbol{t}_i | i = 1, ..., N\}$, where $\boldsymbol{t}_i \in \mathbb{R}^3$ denotes the location of joint i. Given 2D images from multiple views $I = \{I_v | v = 1, ..., V\}$, the posterior distribution of 3D poses can be written as:

$$p(\boldsymbol{T}|I) \propto \prod_{v=1}^{V} \prod_{i=1}^{N} p(I_v|\pi_v(\boldsymbol{t}_i)) \prod_{(i,j)\in\varepsilon} p(\boldsymbol{t}_i, \boldsymbol{t}_j), \qquad (12)$$

where $\pi_v(\boldsymbol{t}_i)$ denotes the 2D projection of $\boldsymbol{t}_i$ in the v-th view and the likelihood $p(I_v|\pi_v(\boldsymbol{t}_i))$ is given by the 2D heat map output by the CNN-based 2D pose detector [10], which characterizes the 2D spatial distribution of each joint.

The prior term $p(\boldsymbol{t}_i, \boldsymbol{t}_j)$ denotes the structural dependency between joint $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$, which implicitly constrains the bone length between them. Here, we use a Guassian distribution to model the prior on bone length:

$$p(\boldsymbol{t}_i, \boldsymbol{t}_j) \propto N(\|\boldsymbol{t}_i - \boldsymbol{t}_j\| | L_{ij}, \sigma_{ij}), \qquad (13)$$

where $\|\boldsymbol{t}_i - \boldsymbol{t}_j\|$ denotes the Euclidean distance between joint $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$, L_{ij} and σ_{ij} denote the mean and standard deviation respectively, learned from the Human3.6M dataset [19].

Comparison with state-of-the-art

Campus	Actor 1	Actor 2	Actor 3	Average
Belagiannis *et al*. [1]	82.0	72.4	73.7	75.8
Belagiannis *et al*. [3]	83.0	73.0	78.0	78.0
Belagiannis *et al*. [2]	93.5	75.7	84.4	84.5
Ershadi-Nasab *et al*. [12]	94.2	92.9	84.6	90.6
Ours w/o 3DPS	90.6	89.2	97.7	92.5
Ours	**97.6**	**93.3**	**98.0**	**96.3**
Shelf	**Actor 1**	**Actor 2**	**Actor 3**	**Average**
Belagiannis *et al*. [1]	66.1	65.0	83.2	71.4
Belagiannis *et al*. [3]	75.0	67.0	86.0	76.0
Belagiannis *et al*. [2]	75.3	69.7	87.6	77.5
Ershadi-Nasab *et al*. [12]	93.3	75.9	94.8	88.0
Ours w/o 3DPS	97.9	89.5	**97.8**	95.1
Ours	**98.8**	**94.1**	**97.8**	**96.9**

Table 2: Quantitative comparison on the Campus and Shelf datasets. The numbers are percentage of correctly estimated parts (PCP). The results of other methods are taken from respective papers. 'Ours w/o 3DPS' means using triangulation instead of the 3DPS model to reconstruct 3D poses from matched 2D poses.

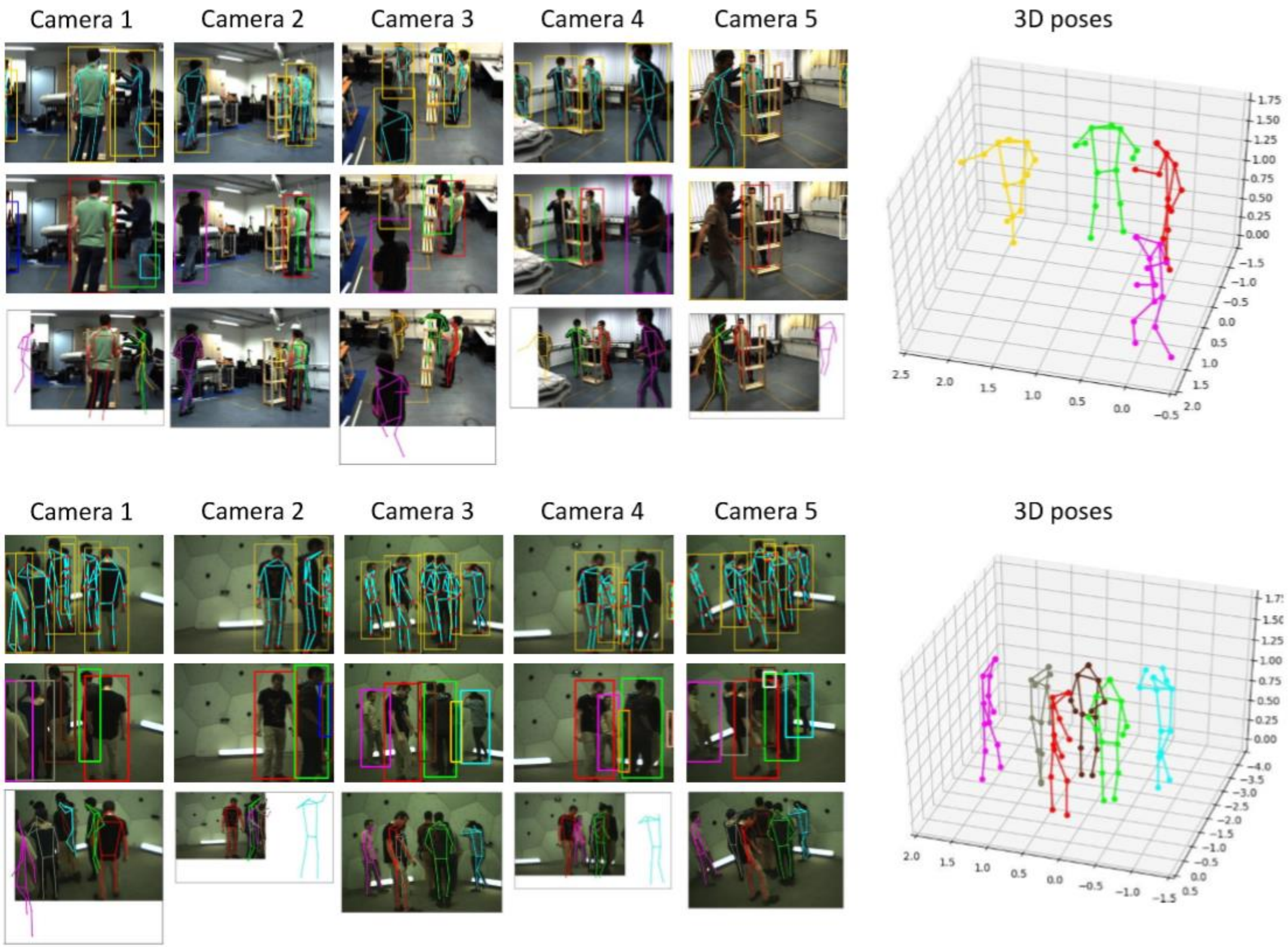
Ablation analysis

1. Appearance or geometry?
2. Direct triangulation or 3DPS?
3. Matching or no matching?

Campus	Actor 1	Actor 2	Actor 3	Average
Ours	**97.6**	**93.3**	**98.0**	**96.3**
Appearance	**97.6**	**93.3**	96.5	95.8
Geometry	97.4	90.1	89.4	92.3
No 3DPS	90.6	89.2	97.7	92.5
No matching	84.8	89.0	71.5	81.8
Shelf	**Actor 1**	**Actor 2**	**Actor 3**	**Average**
Ours	**98.8**	**94.1**	**97.8**	**96.9**
Appearance	98.6	60.5	94.3	84.5
Geometry	97.2	79.5	96.5	91.1
No 3DPS	97.9	89.5	**97.8**	95.1
No matching	98.1	91.1	92.8	94.0

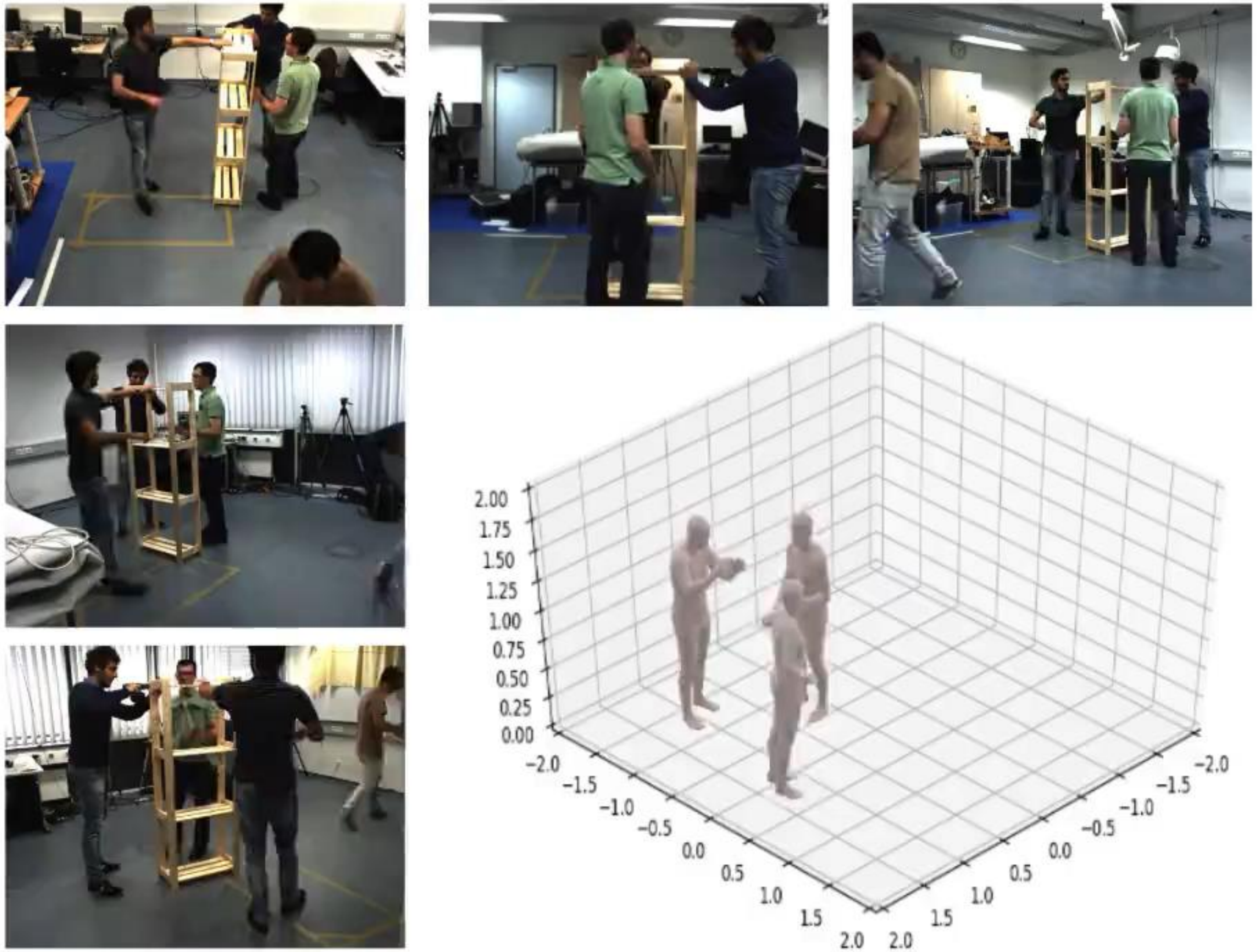
4. Result

Qualitative evaluation



4. Result

Demo



Running time

We report running time of our algorithm on the sequences with four people and five views in the Shelf dataset, tested on a desktop with an Intel i7 3.60 GHz CPU and a GeForce 1080Ti GPU. Our unoptimized implementation on average takes 25 ms for running reID and constructing affinity matrices, 20 ms for the multi-way matching algorithm, and 60 ms for 3D pose inference. Moreover, the results in Table 2 show that our approach without the 3DPS model also obtains very competitive performance, which is able to achieve real-time performance at $>$ 20fps.

THANK YOU!

Q & A