Learnable Triangulation of Human Pose

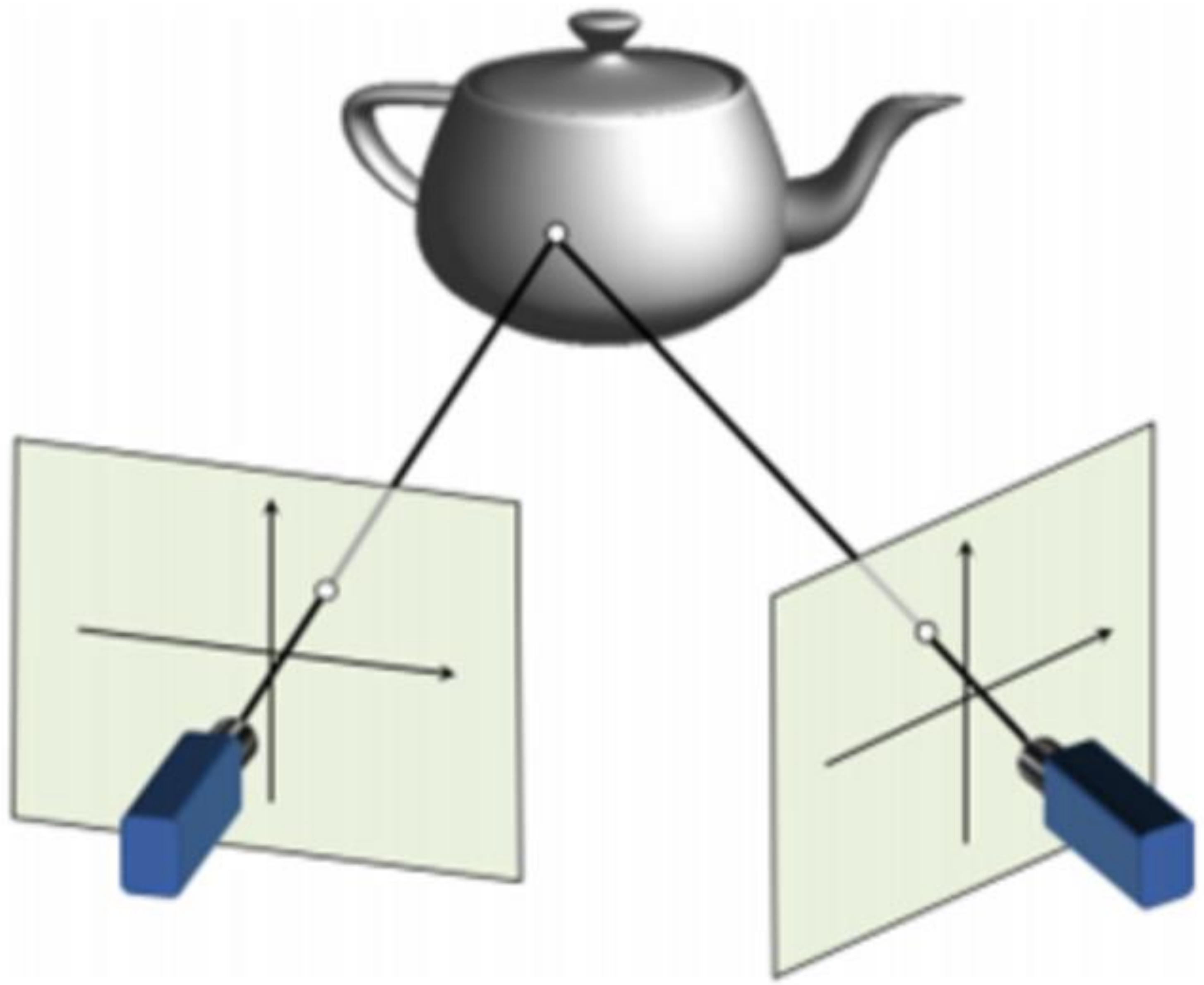
Karim Iskakov[1] **Egor Burkov**[1,2] **Victor Lempitsky**[1,2] **Yury Malkov**[1]
[1] **Samsung AI Center, Moscow**
[2] **Skolkovo Institute of science and Technology, Moscow**

2023. 8. 14.

경영과학연구실 전재현

Backgrounds

- Triangulation



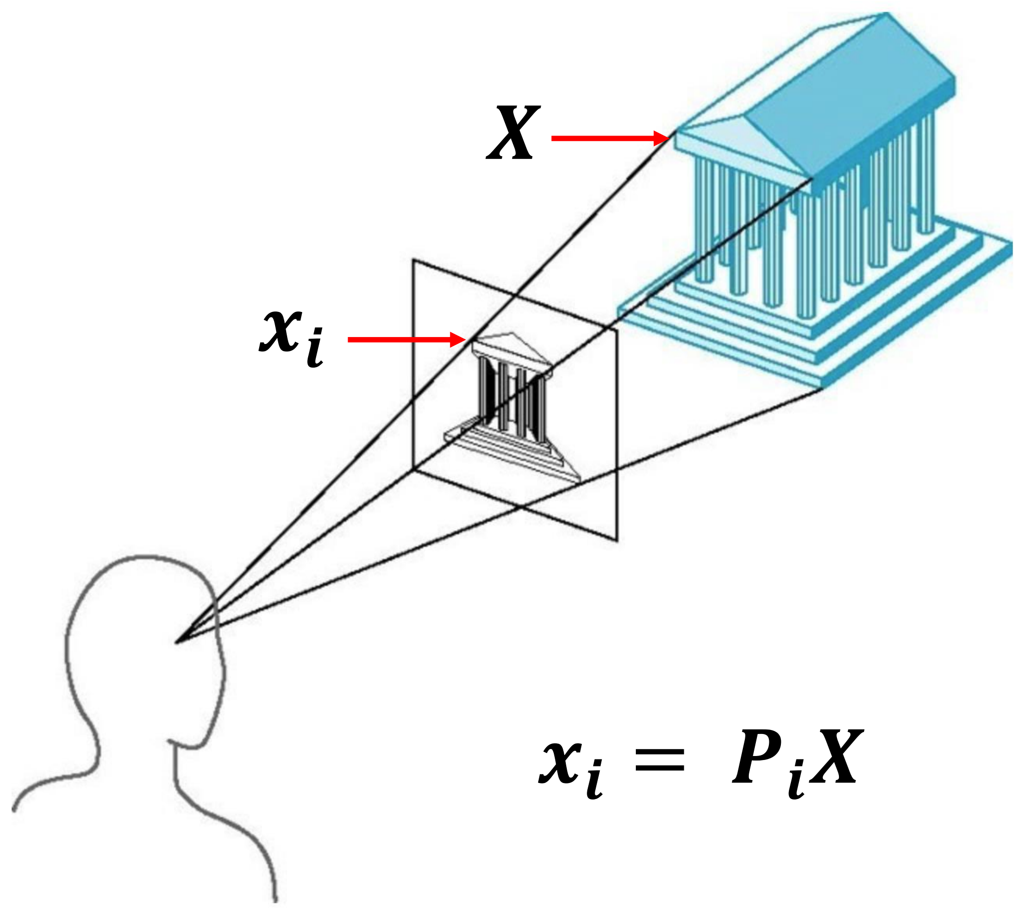
- **Technique used to estimate the position of a point in 3D space based on observations from two or more cameras**

- **The position of the 3D point is determined by finding the intersection of the lines formed by connecting the observed points from each camera**

- **Using camera parameters and corresponding points to determine the 3D position**

Backgrounds

- How to solve?
  - Find the X that satisfies the AX=0 equation
  - X : 3D homogenous coordinate (4 x 1)
    X = (U, V, W, 1)
  - A : coefficient matrix (2C x 4)
    The matrix A can be determined using the camera projection matrix $\boldsymbol{P_i}$ and the projected point $\boldsymbol{x_i}$
    A = $\begin{bmatrix} \boldsymbol{u_1 P_{1,3} - P_{1,1}} \\ \boldsymbol{v_1 P_{1,3} - P_{1,2}} \\ \boldsymbol{u_2 P_{2,3} - P_{2,1}} \\ \boldsymbol{v_1 P_{2,3} - P_{2,2}} \end{bmatrix}$
    $\boldsymbol{P_{i,j}}$ = the j-th row of the projection matrix $\boldsymbol{P_i}$ (3 x 4)
    $\boldsymbol{x_i}$ $(\boldsymbol{u_i}, \boldsymbol{v_i}$, 1) = 2d homogenous coordinate of X projected by $\boldsymbol{P_i}$



Problem Statement & Key Idea

- Problem Statement
  - **When utilizing the multi-view to determine 3D human pose, heatmaps of poor quality due to occlusions or noise can influence the results**

- Key Idea
  - **To reflect the quality of each view, a learnable weight is added**
  - **Estimate the 3d pose by applying learnable weights to both algebraic and volumetric triangulation**

Related works

- Single view 3D pose estimation
  - **A simple yet effective baseline for 3d human pose estimation(J.Martinez et al. 2017)**
    proposed to lifting the 2D coordinates to 3D via deep neural networks.

  - **Integral human pose regression(X. Sun et al. 2018)**
    proposed to infer the 3D coordinates directly from the images using convolutional neural networks.

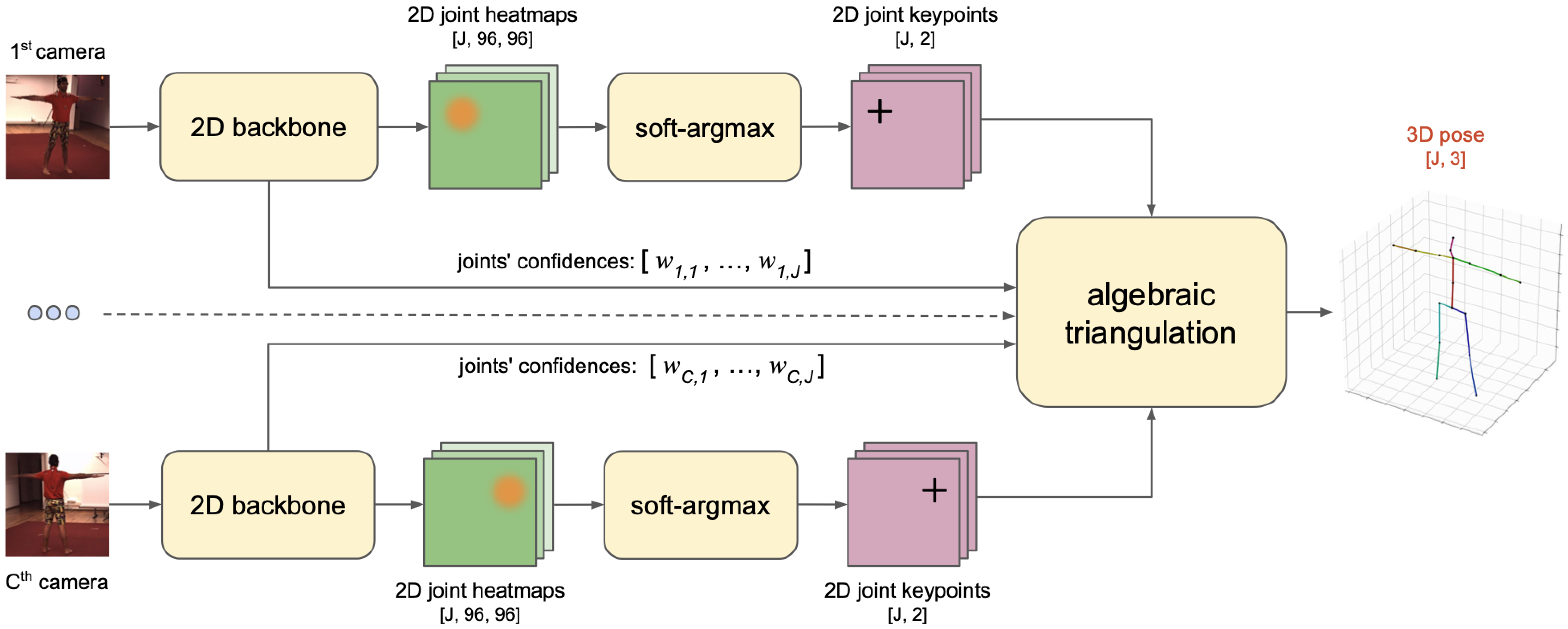
- Multi-view 3D pose estimation
  - **A generalizable approach for multi-view 3D human pose regression**
    (A. Kadkhodamohammadi and N. Padoy. 2018)
    proposed concatenating joints' 2D coordinates from all views into a single batch as an input to a fully connected network

  - **Panoptic studio : A massively multiview system for social interaction capture**
    (H. Joo et al. 2015)
    utilized unprojection of 2D keypoint probability heatmaps to volume with subsequent non-learnable aggregation

- Multiple view geometry
  - **Mutiple view geometry in computer vision(R. Hartley et al. 2003)**
    described the geometric relationships in multiple view for computer vision

Method

- Algebraic Triangulation(baseline)
  - **Using synchronized video streams from C cameras with known projection matrices P_c**
  - **For each timestamp, the frames are processed independently(not using temporal information)**
  - **Process each joint independently of each other**
  - **Using heatmaps to infer the 2D location of the joint**
  - **Then proceeding with triangulation using camera parameters to find the 3D points ($A_j X = 0$)**



Method

- Problem from determining 2d points
  - **The accuracy of the 2D extracting algorithm is high**
  - **There are times when 2d points are not accurate in the event of occlusions**
  - **Cannot assign the same weight and consider all views equally**

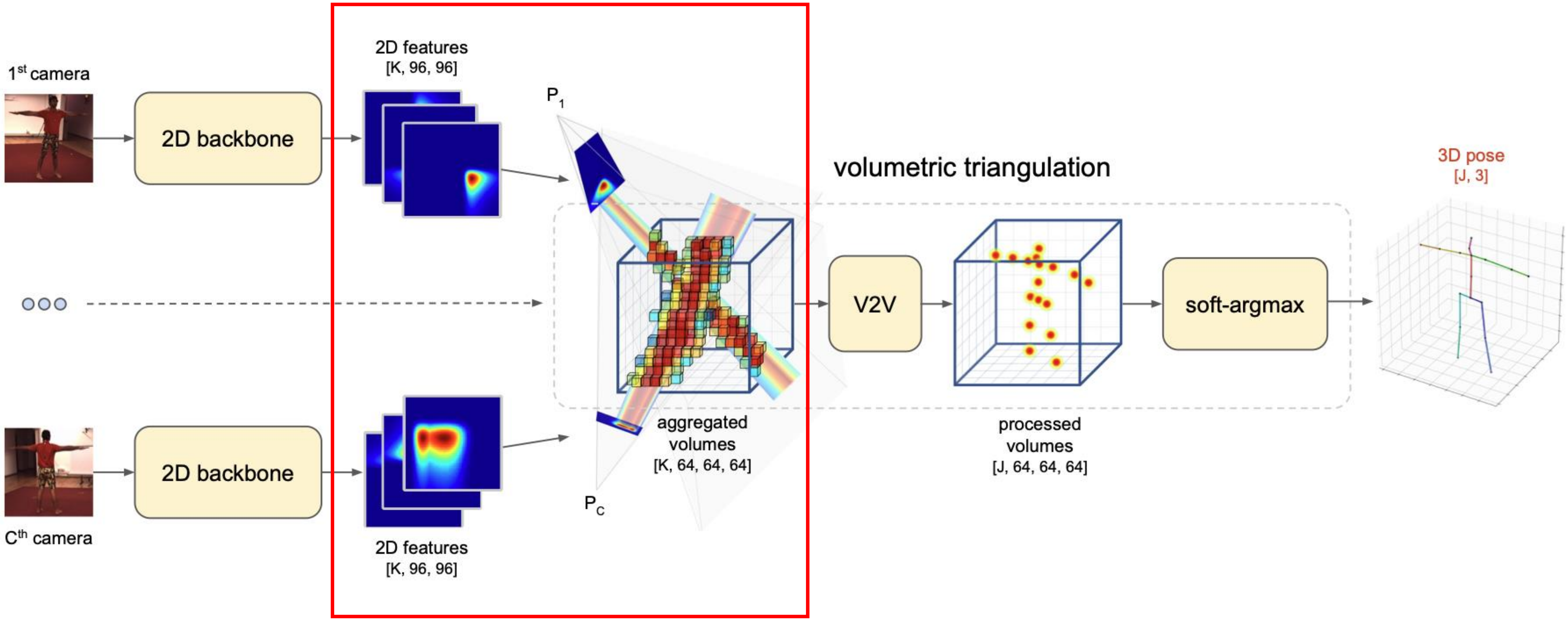
- RANSAC
  - **RANdom SAmple Consensus**
  - **Statistical method used for estimating models, especially in situations where there are outliers in the data**
  - **If using RANSAC, there is a drawback that the model cannot learn from outlier cameras**

Method

- Learnable camera-joint Confidence Weights
  - **Apply learnable weights $\boldsymbol{w_c}$ meaning contribution of camera c**
  - **By using learnable weights, become more robust against joints that are incorrectly estimated due to noise or occlusions**
  - **In scenes with sever occlusions, the heatmap is spread out evenly, so learnable weights $\boldsymbol{w_c}$ is measured to be small**
  - $\boldsymbol{w_j} = (\boldsymbol{w_{1,j}}, \boldsymbol{w_{1,j}}, \boldsymbol{w_{2,j}}, \boldsymbol{w_{2,j}}, \dots, \boldsymbol{w_{C,j}}, \boldsymbol{w_{C,j}})$
  - $\boldsymbol{w_{i,j}}$ **means weight of the j-th joint captured by the i-th camera**
  - **Solve the equation which satisfies $\boldsymbol{w_j} \circ \boldsymbol{A_j}\, \boldsymbol{X}$ = 0**
    (∘ means Hadamard product)

Method

- Volumetric Triangulation
  - **Unproject the feature maps produced by the 2D backbone into 3D volumes**
  - **Filling a 3D cube around the person via projecting output of the 2D network along projection rays inside the 3D cube size LxLxL**
  - **The cubes obtained from multiple views are then aggregated together and processed**



Method

- 3 methods for the aggregation
  - **Raw summation of the voxel data :**
    Simply add the heatmap values from all cubes

$$V_k^{\text{input}} = \sum_c V_{c,k}^{\text{view}}$$

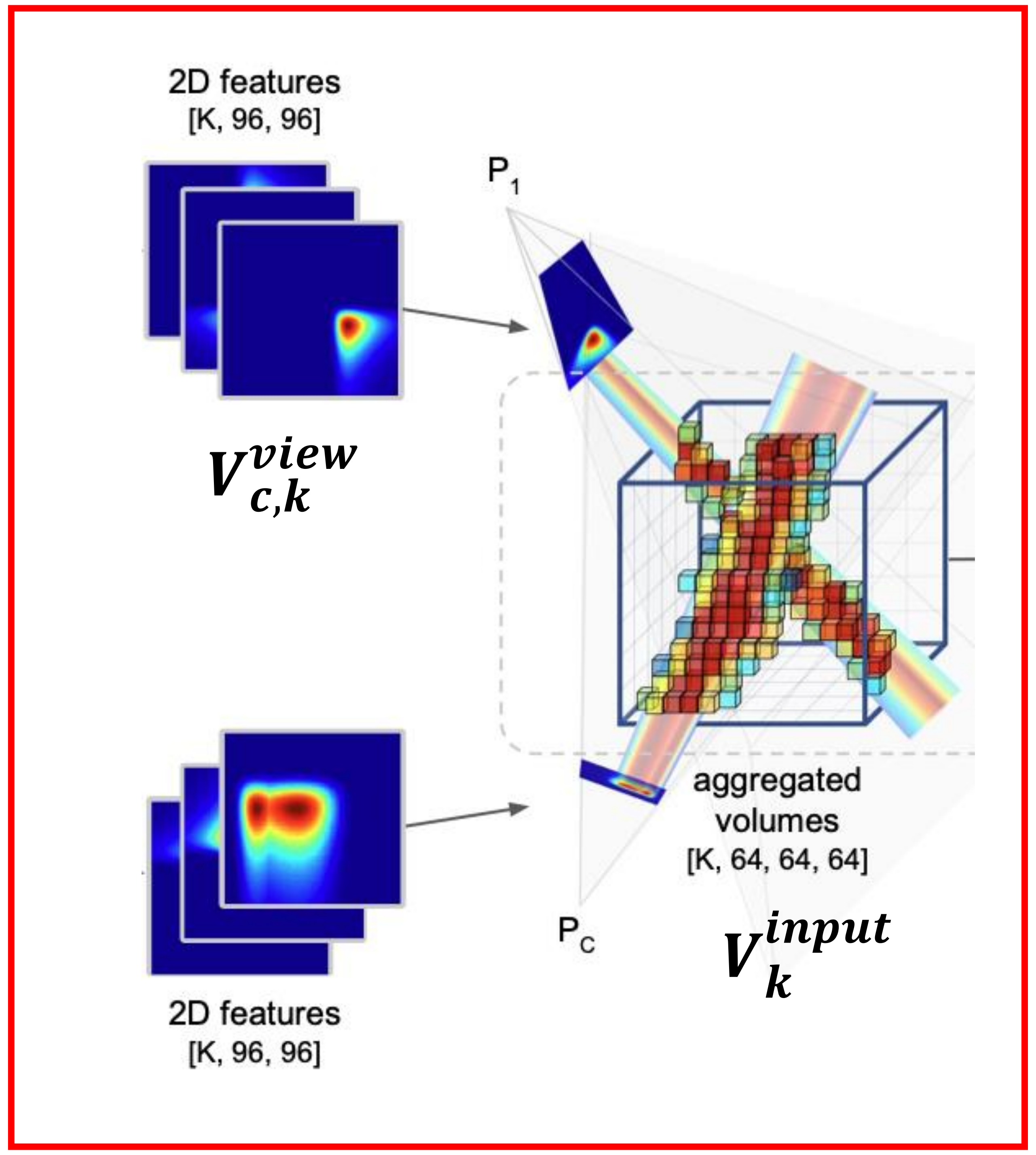
  - **Summation of the voxel data with normalized confidence multipliers d_c :**
    Weighted sum of heatmap values using the learnable weight d_c

$$V_k^{\text{input}} = \sum_c \left(d_c \cdot V_{c,k}^{\text{view}} \right) / \sum_c d_c$$

  - **Calculating a relaxed version of maximum :**
    Weighted sum of heatmap values using the softmax function

$$V_{c,k}^{w} = \exp(V_{c,k}^{\text{view}}) / \sum_c \exp(V_{c,k}^{\text{view}})$$

$$V_k^{\text{input}} = \sum_c V_{c,k}^{w} \circ V_c^{\text{view}}$$



Experiment

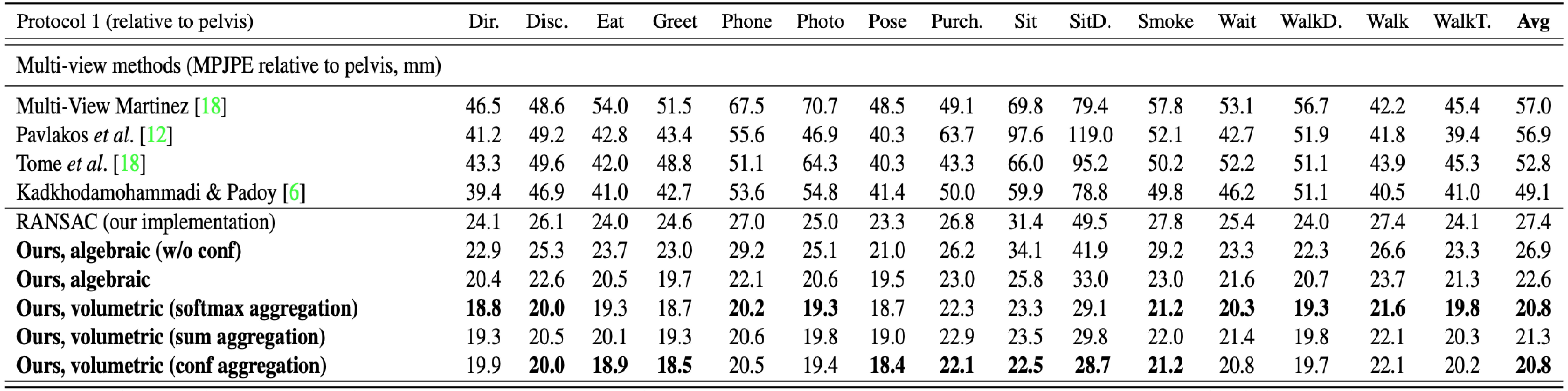
- Experiment Details
  - **Used Human3.6M and CMU Panoptic datasets**
  - **The size of volumetric cube L : 2.5m**
  - **The number of output channels from the 2D backbone : K=32**
  - **2D backbone : ResNet-152 network**

Results

- Experimental Results

  - **Comparison between other algorithms and proposed methods with human 3.6m dataset**
  - **Volumetric methods performs the best, providing about 30% reduction in the error to the RANSAC**
  - **Used 4 cameras for this experiment**

*** MPJPE relative to pelvis :**
Mean Per Joint Position Error from the pelvis(mm)

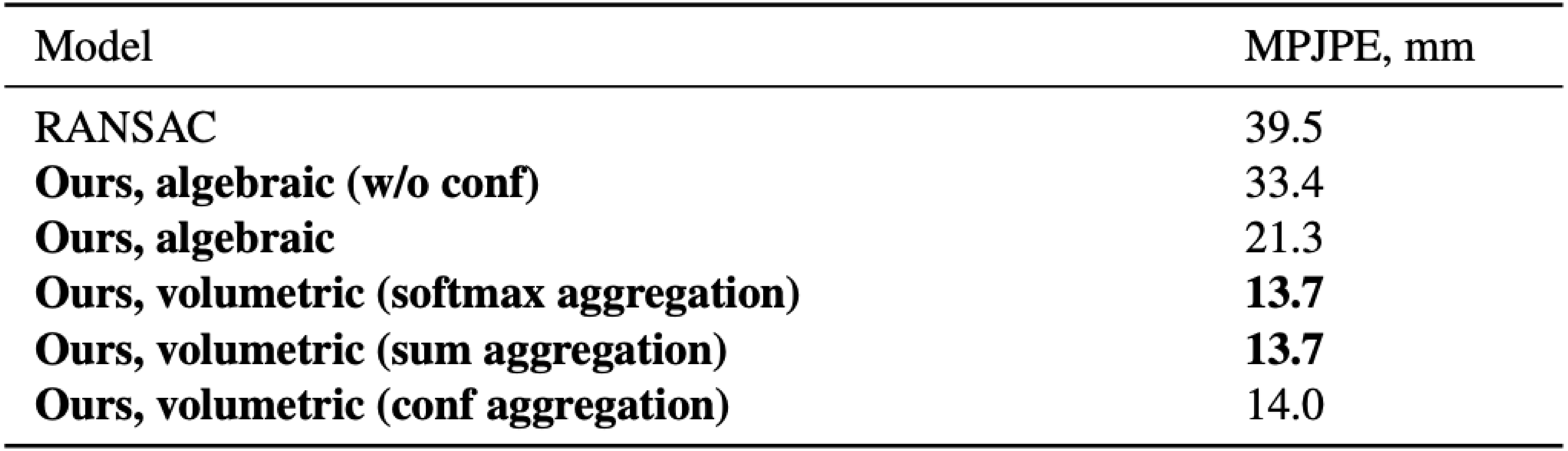


Protocol 1 (relative to pelvis)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	**Avg**
Multi-view methods (MPJPE relative to pelvis, mm)																
Multi-View Martinez [18]	46.5	48.6	54.0	51.5	67.5	70.7	48.5	49.1	69.8	79.4	57.8	53.1	56.7	42.2	45.4	57.0
Pavlakos *et al.* [12]	41.2	49.2	42.8	43.4	55.6	46.9	40.3	63.7	97.6	119.0	52.1	42.7	51.9	41.8	39.4	56.9
Tome *et al.* [18]	43.3	49.6	42.0	48.8	51.1	64.3	40.3	43.3	66.0	95.2	50.2	52.2	51.1	43.9	45.3	52.8
Kadkhodamohammadi & Padoy [6]	39.4	46.9	41.0	42.7	53.6	54.8	41.4	50.0	59.9	78.8	49.8	46.2	51.1	40.5	41.0	49.1
RANSAC (our implementation)	24.1	26.1	24.0	24.6	27.0	25.0	23.3	26.8	31.4	49.5	27.8	25.4	24.0	27.4	24.1	27.4
Ours, algebraic (w/o conf)	22.9	25.3	23.7	23.0	29.2	25.1	21.0	26.2	34.1	41.9	29.2	23.3	22.3	26.6	23.3	26.9
Ours, algebraic	20.4	22.6	20.5	19.7	22.1	20.6	19.5	23.0	25.8	33.0	23.0	21.6	20.7	23.7	21.3	22.6
Ours, volumetric (softmax aggregation)	**18.8**	**20.0**	19.3	18.7	**20.2**	**19.3**	18.7	22.3	23.3	29.1	**21.2**	**20.3**	**19.3**	**21.6**	**19.8**	**20.8**
Ours, volumetric (sum aggregation)	19.3	20.5	20.1	19.3	20.6	19.8	19.0	22.9	23.5	29.8	22.0	21.4	19.8	22.1	20.3	21.3
Ours, volumetric (conf aggregation)	19.9	**20.0**	**18.9**	**18.5**	20.5	19.4	**18.4**	**22.1**	**22.5**	**28.7**	**21.2**	20.8	19.7	22.1	20.2	**20.8**

Results

- Experimental Results
  - **Comparison between RANSAC method and proposed methods with CMU dataset**
  - **Volumetric approach has a dramatic advantage over the algebraic one**
  - **Used 4 cameras for this experiment**

*** MPJPE(mm) : Mean Per Joint Position Error**

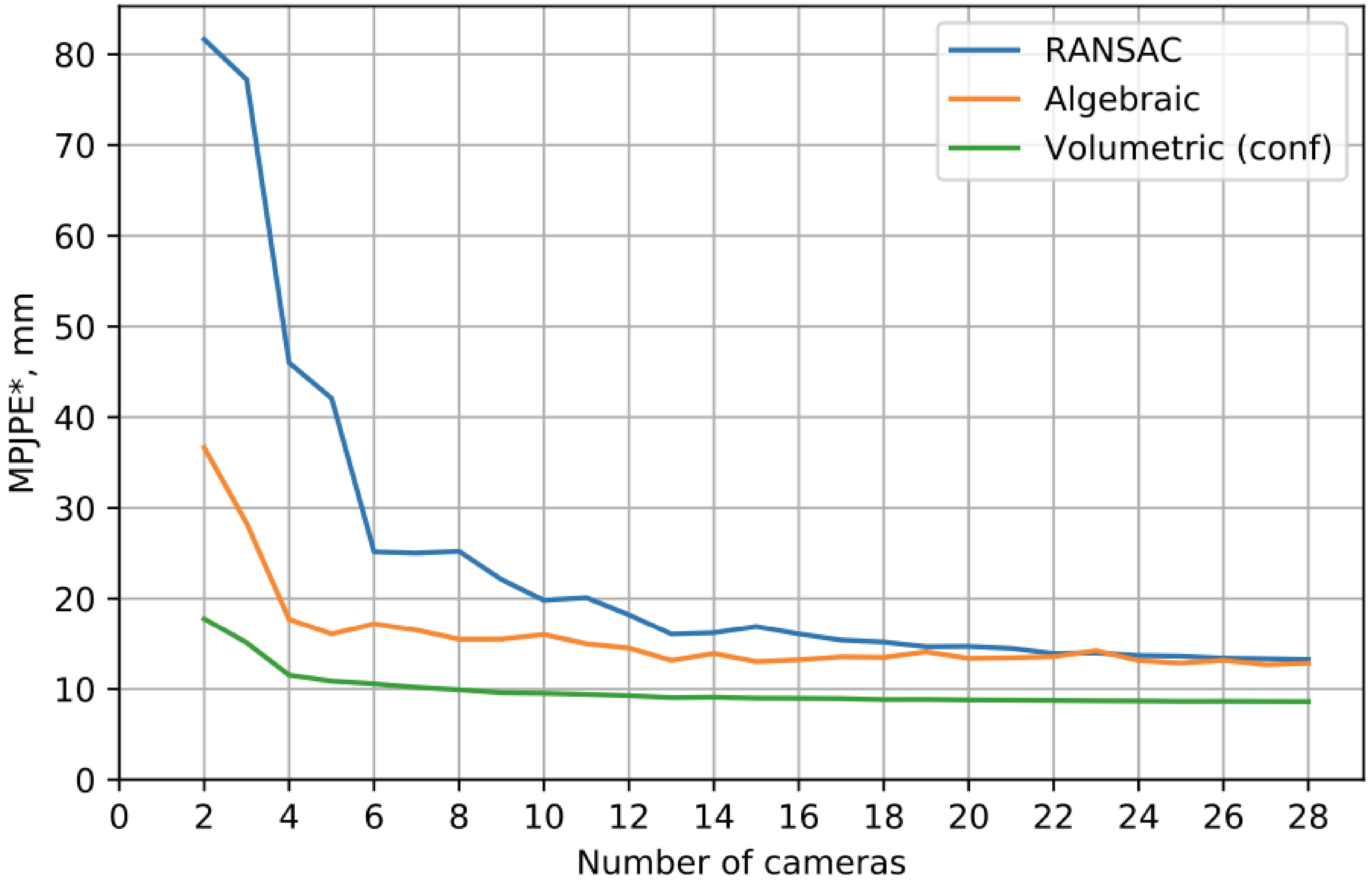


Model	MPJPE, mm
RANSAC	39.5
Ours, algebraic (w/o conf)	33.4
Ours, algebraic	21.3
Ours, volumetric (softmax aggregation)	**13.7**
Ours, volumetric (sum aggregation)	**13.7**
Ours, volumetric (conf aggregation)	14.0

Results

- Experimental Results
  - **Error versus the numbers of used cameras with CMU Panoptic dataset**
  - **Volumetric triangulation methods drastically reduced the number of cameras in real-life setups**
  - **The error of RANSAC approach with 28 cameras > The error of Volumetric approach with 4 cameras**

*** MPJPE(mm) : Mean Per Joint Position Error**



Conclusion

- Conclusion

  - **Applying the confidence weight of each feature maps, they achieved better 3D estimation results**
  - **Using volumetric triangulation method, they reduced the number of views needed to achieve high accuracy**
  - **The limitation of this algorithm is that it supports only a single person in the scene**