Fusing Wearable IMUs with Multi-View Images for Human Pose Estimation : A Geometric Approach

Zhe Zhang Chunyu Wang Wenhu Qin Wenjun Zeng
Southeast University, Nanjing, China Microsoft Research Asia, Beijing, China

2023. 7. 17.

경영과학연구실 전재현

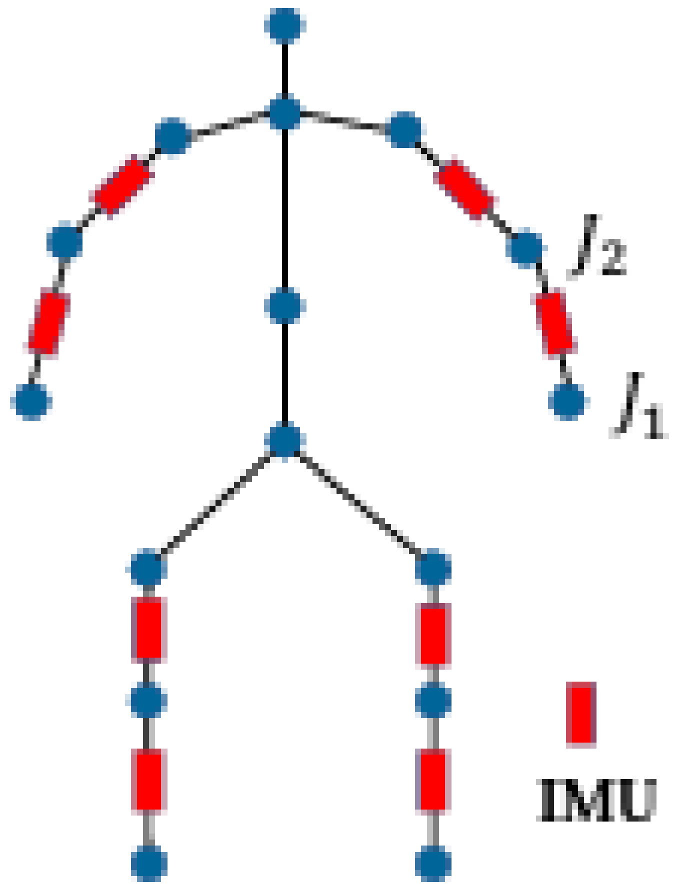
- IMU Sensor
  - **Inertial Measurement Unit**
  - **A device combines multiple sensors like accelerometers, gyroscopes, and magnetometers**
  - **Using the information mentioned above, after calibrating the initial position of the sensor, it is possible to estimate the position of the sensor**
  - **Advantages when using only IMU sensor : Robustness in certain environments(occlusion, low light conditions)**
  - **Drawbacks when using only IMU sensor : Calibration error**
Drift phenomenon
Difficult to apply in real-world situations

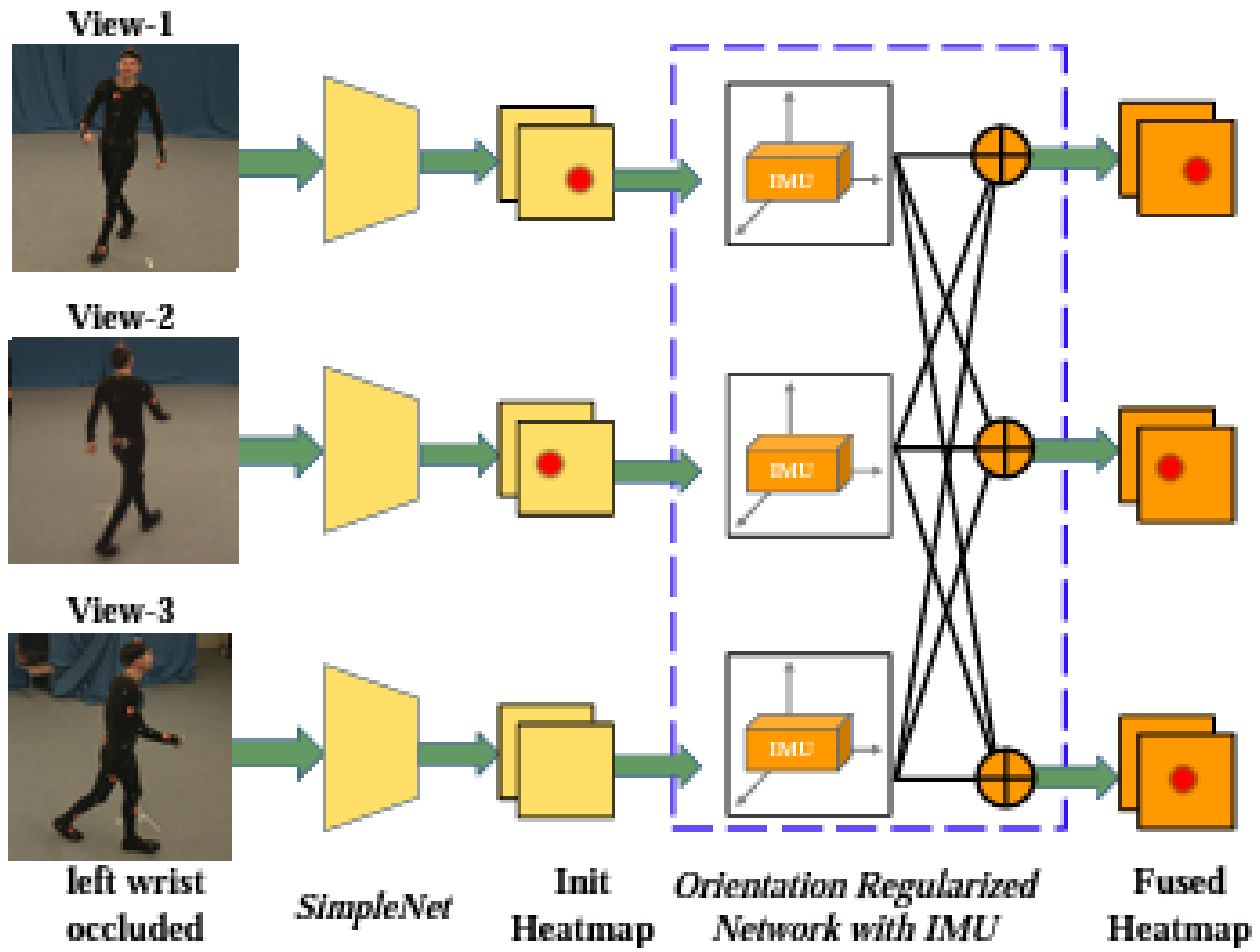
Problem Statement & Key Idea

- Problem Statement
  - **Estimating 3D human pose from a multi-view image using orientation data from IMUs**
- Key Idea
  - **Instead of estimating 3D poses or pose embeddings from images and IMUs separately and then fusing them in the late stage, they fuse IMUs and image features in a very early stage with the aid of 3D geometry**
  - **Use the orientation of the limb, when constructing 3d human pose**

- Images-based
  - Haibo Qiu et al. Cross view fusion for 3d human pose estimation
    proposed to first estimate 2D pose for every camera view, and then estimate the 3D pose by triangulation
  - Helge Rhodin et al. Learning monocular 3d human pose estimation from multi-view images
    proposed a method to estimate camera pose jointly with human pose, which allows to utilize multi-view images where calibration is difficult
- IMUs-based
  - Denis Time et al. Rethinking pose in 3D : Multi-stage refinement and recovery for markerless motion capture
    proposed to reconstruct human pose from 5 accelerometers by retrieving prerecorded poses
  - Daniel Roetenberg et al. Xsens mvn : full 6dof human motion tracking using miniature inertial sensors
    used 17 IMUs equipped with 3D accelerometers, gyroscopes and magnetometers and all the measurements are fused using a Kalman Filter
- “Images+IMUs”-based
  - Matthew Trumble et al. Total capture : 3D human pose estimation fusing video and inertial sensors
    proposed a two stream network to concatenate the pose embeddings separately derived from images and IMUs for regressing the final pose

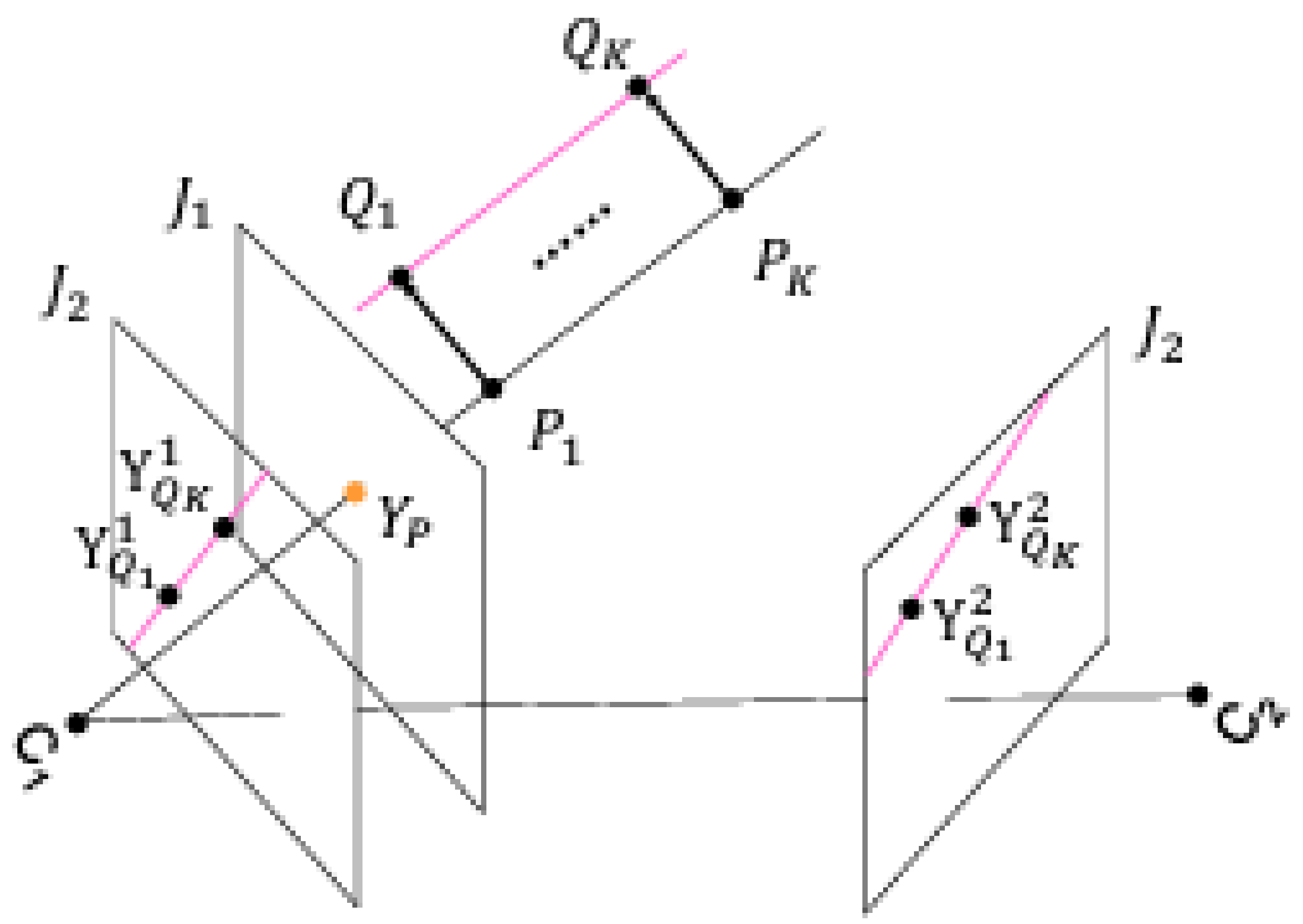
- ORN for 2D Pose Estimation
  - **ORN : Orientation Regularized Network**
  - **Takes multi-view images as input and estimates initial heatmaps**
  - **With the aid of IMU orientations, fuses the heatmaps of the linked joints(Same-View Fusion)**
  - **Also fuses the heatmaps across all views(Cross-View Fusion)**





- Same-View Fusion

  - **Helps to accurately localize the occluded joints based on their neighbors**
  - **Determine the relative positions between each pair of joints in the images using orientation data**



(a)

$$Q_k = P_k + o * l \quad \forall k = 1, \cdots, K$$

Q_k **: Possible 3D point candidate of J_2 using IMU orientation and J_1**

P_k **: Possible 3D point candidate of Y_p**

o **: orientation vector from IMU**

l **: average limb length**

- Same-View Fusion

  - **Enhance the heatmap value using linked joints**

$$H_1(Y_P) \leftarrow \lambda H_1(Y_P) + (1-\lambda) \max_{k=1\cdots K} H_2(Y_{Q_k})$$

0	0.5	0.7	0.8	0
0	0.4	0.9	0.8	0
0	0.4	0.4	0.6	0
0	0	0	0	0
0	0	0	0	0

H1
(Heatmap of J1)

0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0.4	0.5	0.7
0	0	0.3	0.9	0.8

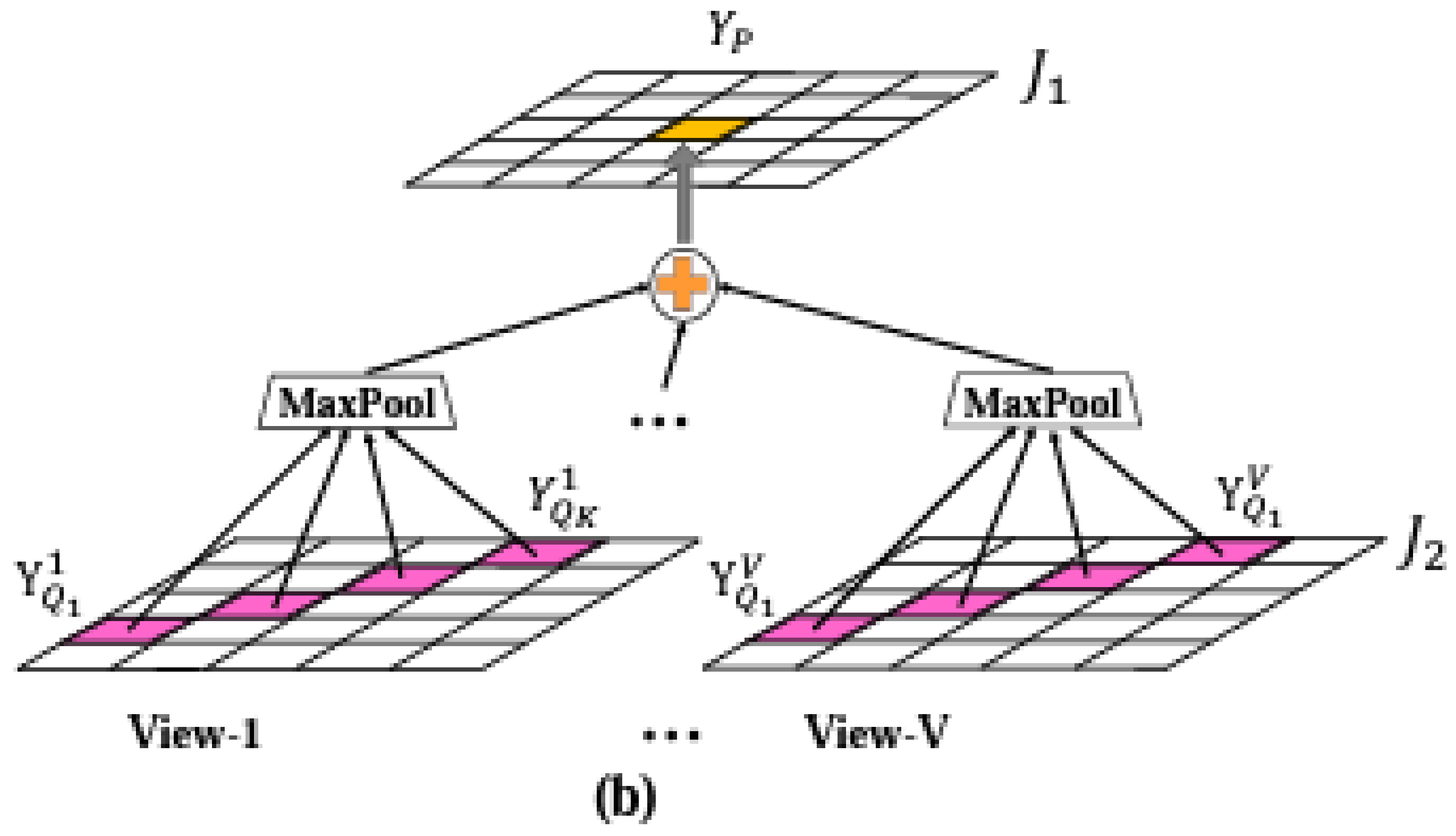
H2
(Heatmap of J2)

0	0.2	0.3	0.4	0
0	0.1	0.95	0.3	0
0	0.2	0.1	0.3	0
0	0	0	0	0
0	0	0	0	0

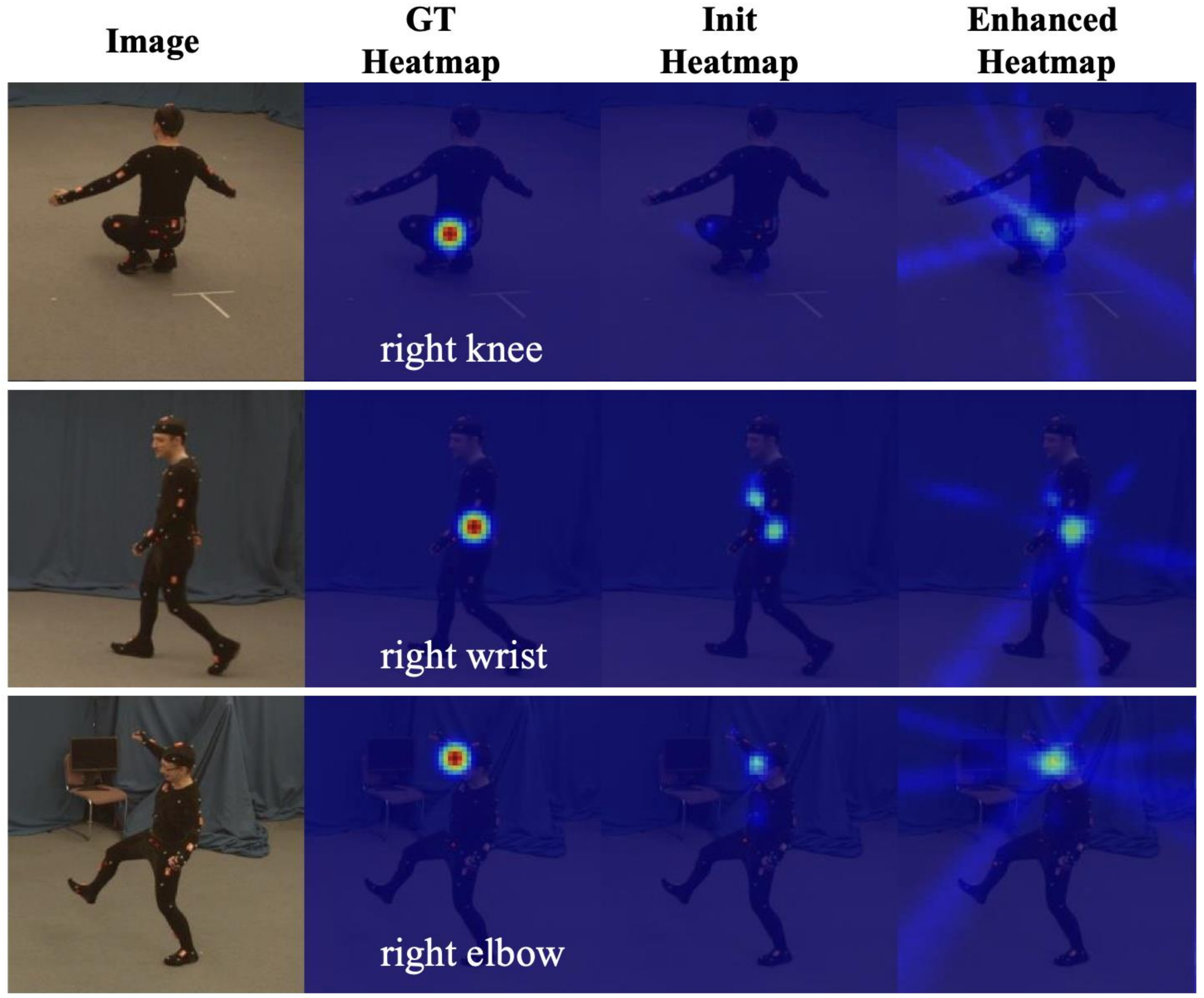
Enhanced H1

- Cross-View Fusion

  - **Some non-corresponding locations are mistakenly enhanced in Same-View Fusion**
  - **Performs fusion across multiple views simultaneously**



$$H_1(Y_P) \leftarrow \lambda H_1(Y_P) + \frac{(1-\lambda)}{V} \sum_{v=1}^{V} \max_{k=1\cdots K} H_2^v(Y_{Q_k}^v)$$



Example of Cross-View Fusion

- ORPSM for 3D Pose Estimation

  - **ORPSM : Orientation Regularized Pictorial Structure Model**
  - **Pictorial Structure Model : Modeling the inter-relationship between joints to estimate the pose**
  - **Objective Function :**

Maximize $$p(\mathcal{J}|\mathcal{F}) = \frac{1}{Z(\mathcal{F})} \prod_{i=1}^{M} \phi_i^{\text{conf}}(J_i, \mathcal{F}) \prod_{(m,n)\in\mathcal{E}_{limb}} \psi^{\text{limb}}(J_m, J_n) \prod_{(m,n)\in\mathcal{E}_{IMU}} \psi^{\text{IMU}}(J_m, J_n)$$

- ORPSM for 3D Pose Estimation

  - **Objective Function :**

Maximize $$p(\mathcal{J}|\mathcal{F}) = \frac{1}{Z(\mathcal{F})} \prod_{i=1}^{M} \phi_i^{\text{conf}}(J_i, \mathcal{F}) \prod_{(m,n)\in\mathcal{E}_{limb}} \psi^{\text{limb}}(J_m, J_n) \prod_{(m,n)\in\mathcal{E}_{IMU}} \psi^{\text{IMU}}(J_m, J_n)$$

  - **Unary Potential : Average response over all camera views**

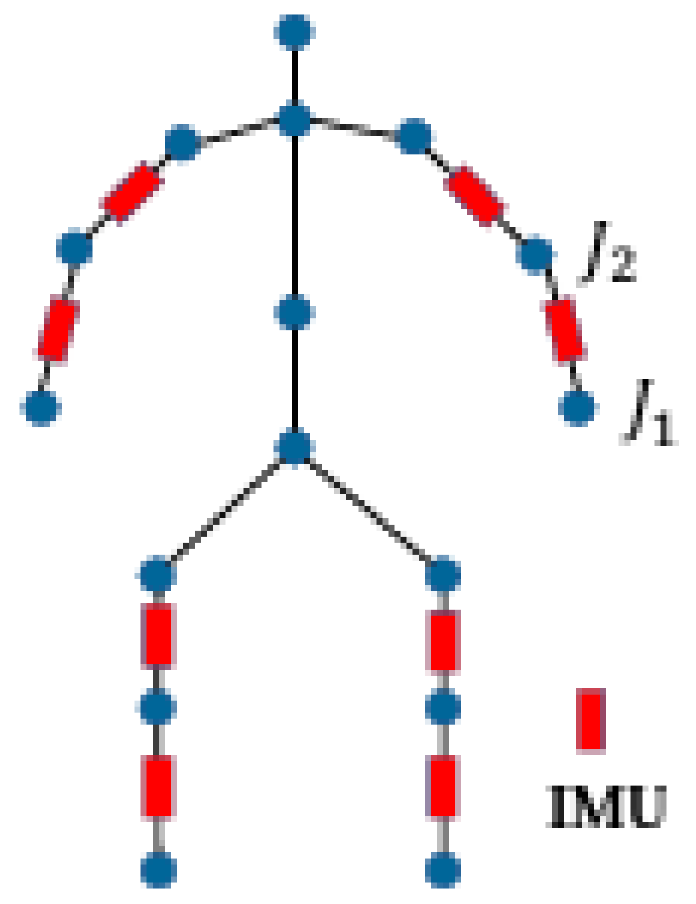
$$\phi_i^{\text{conf}}(J_i, \mathcal{F})$$

  - **Limb Length Potential :**

$$\psi^{\text{limb}}(J_m, J_n) = \begin{cases} 1, & \text{if} \quad |l_{m,n} - \tilde{l_{m,n}}| \leq \epsilon, \\ 0, & \text{otherwise} \end{cases}$$

  - **Limb Orientation Potential:**

$$\psi^{\text{IMU}}(J_m, J_n) = \frac{J_m - J_n}{\|J_m - J_n\|_2} \cdot o_{m,n}$$



- Experiment Details

  - **Used Total Capture, Human3.6M(3d) dataset**
  - **Total Capture(2D, 3D) : Dataset with images, IMUs and ground truth 3D pose**
  - **Human3.6M(3D) : Dataset with images and ground truth 3D pose**

- Experimental Results

  - **2D Pose Estimation Result using Total Capture Dataset**
  - **SN : Simple Network(ResNet50)**
  - ORN^{same} **: Using only Same-View Fusion**
  - ORN **: Using Cross-View Fusion**

*** PCKh @ : The Percentage of Correct Keypoints**

Methods	PCKh@	Hip	Knee	Ankle	Shoulder	Elbow	Wrist	*Mean (Six)*	Others	Mean (All)
SN	1/2	99.3	98.3	98.5	98.4	96.2	95.3	97.7	99.5	98.1
ORN^{same}	1/2	99.4	99.0	98.8	98.5	97.7	96.7	98.3	99.5	98.6
ORN	1/2	**99.6**	**99.2**	**99.0**	**98.9**	**98.0**	**97.4**	**98.7**	99.5	98.9
SN	1/6	97.5	92.3	92.5	78.3	80.8	80.0	86.9	95.4	89.1
ORN^{same}	1/6	97.2	94.0	93.3	78.1	83.5	82.0	88.0	95.4	89.9
ORN	1/6	**97.7**	**94.8**	**94.2**	**81.1**	**84.7**	**83.6**	**89.3**	95.4	90.9
SN	1/12	**87.6**	67.0	68.6	47.4	50.0	49.3	61.7	78.1	65.8
ORN^{same}	1/12	81.2	70.1	68.0	43.9	51.6	50.1	60.8	78.1	65.2
ORN	1/12	85.3	**71.6**	**70.6**	**47.7**	**53.2**	**51.9**	**63.4**	78.1	67.1

- Experimental Results
  - 3D Pose Estimation Result using Total Capture Dataset
  - LSTM-AE[26] : Has benefits when it is applied to periodic actions

* MPJPE(mm) : Mean Per Joint Position Error

Approach	IMUs	Temporal	Aligned	Subjects(S1,2,3)			Subjects(S4,5)			Mean
				W2	A3	FS3	W2	A3	FS3	
PVH [27]				48.3	94.3	122.3	84.3	154.5	168.5	107.3
Malleson *et al.* [15]	✓	✓		-	-	65.3	-	64.0	67.0	-
VIP [28]	✓	✓	✓	-	-	-	-	-	-	26.0
LSTM-AE [26]		✓		**13.0**	23.0	47.0	**21.8**	40.9	68.5	34.1
IMUPVH [6]	✓	✓		19.2	42.3	48.8	24.7	58.8	61.8	42.6
Qiu *et al.* [19]				19.0	21.0	28.0	32.0	33.0	54.0	29.0
SN + PSM				14.3	18.7	31.5	25.5	30.5	64.5	28.3
SN + PSM			✓	12.7	16.5	28.9	21.7	26.0	59.5	25.3
ORN + ORPSM	✓			14.3	**17.5**	**25.9**	23.9	**27.8**	**49.3**	**24.6**
ORN + ORPSM	✓		✓	12.4	14.6	22.0	19.6	22.4	41.6	20.6

- Experimental Results

  - **3D Pose Estimation Result using Human 3.6M dataset**
  - **No IMU data in Human 3.6M dataset**
    - ➔ **Created limb orientations using the ground truth 3D poses**

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

Methods	Hip	Knee	Ankle	Shoulder	Elbow	Wrist	*Mean (Six)*	Others	Mean (All)
noFusion (SN + PSM)	23.2	28.7	49.4	29.1	28.4	32.3	31.9	18.3	27.9
ours (ORN + ORPSM)	**20.6**	**18.6**	**28.2**	**25.1**	**21.8**	**24.2**	**23.1**	18.3	21.7

- Conclusion

  - **Using orientation of limbs and cross-view fusion, the accuracy of the 2D pose estimation increased**
  - **By using more accurate 2D heatmaps, the accuracy of 3D pose estimation has also increased**
  - **But in some cases, the accuracy was lower than the method using sequential information**