VIMCAN: Visual-Inertial 3D Human Pose Estimation with Hybrid Mamba-Cross-Attention Network

Supplementary Material

1. Implementation of Cross-Attention

As illustrated in Fig. 1, we adopt a standard Cross-Attention mechanism to integrate multimodal features. For each group g, visual features Y_g^V are projected into queries Q_g^V, while inertial features Y_g^I are projected into keys K_g^I and values V_g^I via separate linear layers. To preserve the skeletal structural information, a residual connection is applied exclusively to the visual queries.

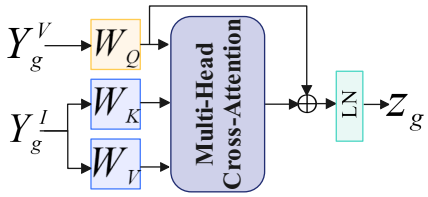

Figure 1. The architecture of the Cross-Attention module.

2. Group Partition Strategy

As shown in Table 1, Group 0 uses all available keypoints and IMUs without any partitioning. Group 3 divides the human body into three regions and Group 5 further refines the human body into five parts.

This hierarchical partitioning facilitates more structural and modality-specific feature extraction, which is essential for robust human pose estimation.

Table 1. The group configurations and their corresponding components for visual keypoints and inertial sensors. The indice is presented in Fig.3 of the main paper.

Groups	Partition	Visual	Inertial
0	-	All keypoints	All IMUs
3	Torso	$V_{0,7,8,9,10}$	$I_{0,1}$
	Upper	$V_{11,12,13,14,15,16}$	$I_{4,5}$
	Lower	$V_{1,2,3,4,5,6}$	$I_{2,3}$
5	Torso	$V_{0,7,8,9,10}$	$I_{0,1}$
	Left Arm	$V_{0,7,8,11,12,13}$	$I_{0,1,4}$
	Rigth Arm	$V_{0,7,8,14,15,16}$	$I_{0,1,5}$
	Left Leg	$V_{0,4,5,6}$	$I_{0,2}$
	Rigth Leg	$V_{0,1,2,3}$	$I_{0,3}$

3. Data Preprocessing for 2D Keypoints

The input keypoints are derived from 2D detectors (*i.e.*, MediaPipe and SimpleNet) and mapped to a predefined set of $J = 17$ body joints. For joints with direct correspondences (*e.g.*, LeftShoulder, RightKnee), we use the respective landmark coordinates. For composite joints such as Hip, Spine, Spine3, and Neck, we apply geometric computations to approximate their positions:

- **Hips**: Computed as the midpoint between the left and right hip landmarks.
- **Spine**: Interpolated 25% of the way from the hip center to the shoulder center.
- **Spine3**: Interpolated 75% of the way from the hip center to the shoulder center.
- **Neck**: Defined as the point 33% of the way from the shoulder center to the nose.

To achieve scale and translation invariance, we normalize the 2D keypoints as follows. First, we compute the bounding box scale as the maximum of the width and the height of the keypoint set. Each keypoint is then divided by this scale and is represented in the form of root-relative coordinates.

4. Impact of 2D Detector

To quantitatively assess the influence of upstream 2D pose detection on our framework, we compare MediaPipe and SimpleNet on TotalCapture testing set. Table 2 reports their per-joint and overall performance in terms of MPJPE (in pixels) and Percentage of Correct Keypoints (PCK) at two thresholds (*i.e.*, 25, 50). SimpleNet achieves a lower MPJPE compared to MediaPipe, indicating higher detection accuracy. This performance gap is further reflected in the PCK metrics, where SimpleNet consistently outperforms MediaPipe at both thresholds. These improvements in the quality of 2D detection directly contribute to the enhanced performance of 3D pose estimation, as observed in Table 1 of the main paper. This analysis confirms that while our framework demonstrates robustness to varying levels of 2D detection noises, the overall system performance benefits significantly from higher-quality upstream detection.

Table 2. The comparison of the per-joint and overall performance for 2D pose detectors on TotalCapture testing set. MP: MediaPipe. SN: SimpleNet. MPJPE: Average MPJPE (pixel, lower is better). PCK: Percentage of Correct Keypoints (percentage, higher is better). 50 and 25 are thresholds.

Joint	MPJPE ↓		PCK@50 ↑		PCK@25 ↑	
	MP	SN	MP	SN	MP	SN
Hips	33.2	**16.2**	98.6	98.6	86.2	**90.1**
RightUpLeg	32.9	**18.1**	98.0	**98.1**	85.2	**89.1**
RightLeg	35.5	**17.9**	96.4	**97.4**	87.2	**90.7**
RightFoot	37.8	**20.2**	94.9	**95.7**	86.8	**90.5**
LeftUpLeg	34.3	**19.6**	98.4	98.4	84.8	**86.1**
LeftLeg	32.8	**15.1**	96.6	**97.6**	88.2	**92.8**
LeftFoot	35.2	**17.2**	95.3	**96.2**	88.5	**92.2**
Spine	39.5	**25.1**	97.4	97.4	75.7	**77.1**
Spine3	38.1	**19.3**	94.2	**96.6**	80.8	**87.4**
Neck	41.0	**25.4**	96.1	96.1	64.8	**66.7**
Head	34.8	**19.5**	98.7	**98.8**	85.9	**86.9**
LeftArm	44.7	**32.7**	94.0	**98.8**	40.4	**49.2**
LeftForeArm	35.8	**20.3**	98.1	**98.2**	78.7	78.7
LeftHand	37.6	**19.0**	94.5	**95.9**	76.7	**86.1**
RightArm	40.3	**21.1**	91.1	**93.8**	74.4	**85.2**
RightForeArm	34.9	**19.0**	97.6	**98.2**	85.7	**85.9**
RightHand	37.8	**18.8**	95.7	**97.5**	81.5	**88.7**
Overall	36.9	**20.3**	96.2	**96.7**	80.5	**82.1**