Multimodal Human Motion Capture Based on Pre-Trained Models

Mengyin Liu 23020241154425[1], Zekai Wu 23020241154456[2],
Zhengzhu Liu 23020241154353[1], Qi Li 23020241154413[1], Lin Xie 23020241154458[1]

[1]Information Class
[2]AI Class

Abstract

Human motion capture involves collecting data on human movement using sensors and converting it into digital representations. This technology is widely used in fields like film production, video game development, sports analysis, and virtual reality. Traditional methods rely on dedicated sensor systems such as motion capture devices, RGB cameras, LiDAR, and IMUs. However, these approaches, whether used individually or in simple combinations, face limitations in flexibility and real-world applicability.To overcome these challenges, this project focuses on multimodal data fusion using pre-trained models, particularly those developed for image-based feature extraction. By extracting features from modalities like images and point clouds, and then fusing these features, the aim is to improve the accuracy and robustness of motion capture systems. Experimental validation will be conducted on multiple datasets to assess the performance of the proposed approach.This study seeks to develop a more flexible and reliable method for human motion capture, addressing the limitations of traditional systems and better meeting the demands of practical applications.

Introduction

This project aims to address the limitations of traditional human motion capture systems by developing a multimodal motion capture algorithm based on pre-trained models. Originally utilized in film and animation, human motion capture technology has broadened its applications to include sports analysis and medical rehabilitation due to its ability to digitally record and store human movements. However, conventional systems, such as optical and inertial motion capture methods, encounter significant drawbacks, including high costs, complicated setups, and limited real-time applicability. Furthermore, these systems often require subjects to wear specialized equipment, which can interfere with the naturalness of their movements.

With rapid advancements in deep learning and the emergence of powerful pre-trained models, there is now an opportunity to create more flexible and efficient motion capture systems that do not depend on intricate sensor configurations or optical markers. This transition could lower system costs, simplify setups, and broaden the range of scenarios suitable for motion capture.

The motivation behind this research stems from the growing demand across various industries for more robust, flexible, and cost-effective motion capture solutions. While traditional methods perform well in controlled settings, they struggle to adapt to real-world environments where issues such as occlusion, diverse surroundings, and the interaction between the human body and the environment complicate the capture process. By leveraging pre-trained models for feature extraction, this project seeks to explore how multimodal data fusion—incorporating both image and point cloud data—can effectively tackle these challenges.

Ultimately, we will rigorously test and refine the proposed model using established datasets, aiming to enhance accuracy, robustness, and adaptability in human motion capture across various applications.

Related work

Traditional Human Motion Capture

Traditional motion capture primarily includes optical and inertial systems. Optical systems use motion capture suits with optical markers and multiple infrared cameras to estimate human posture by tracking the markers. Inertial systems equip subjects with inertial measurement units (IMUs) to collect movement data. Examples include Optitrack and Motion Analysis for optical systems, and Xsens and Noitom for inertial systems. However, optical systems require complex setups and extensive data processing, leading to high costs. Both systems also necessitate wearing specialized equipment, which can impact the authenticity and range of motion.

Single-modality Human Motion Capture

In contrast, image-based motion capture methods do not require markers for calibration, nor do they require complex post-processing, making them more cost-effective and simpler to use, thus gaining wider research and application. HMR(Kanazawa et al. 2018) uses a parameterized SMPL model to describe the human body, minimizing the re-projection error of key points. It uses a large three-dimensional human motion dataset to verify the authenticity of the generated human motion, i.e., the authenticity of the model-generated posture parameters. Unlike HMR, VIBE(Kocabas, Athanasiou, and Black 2020) focuses on

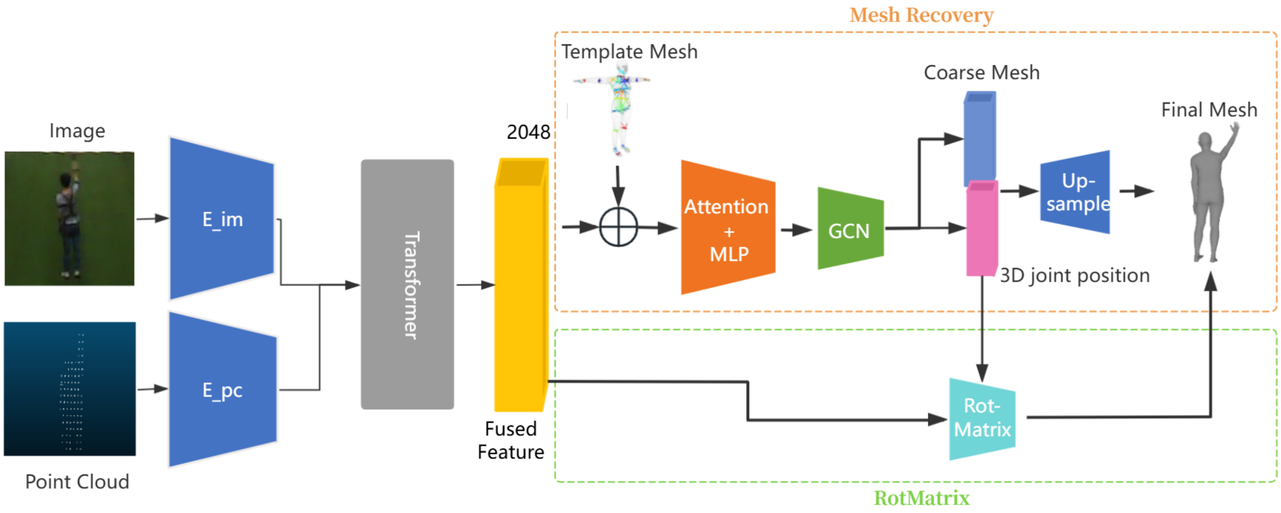

Figure 1: The overall architecture of the Multi-modal Fusion Human Pose Estimation(MFHPE)

motion capture from RGB videos. It utilizes existing large-scale motion capture datasets (such as AMASS(Mahmood et al. 2019)) and two-dimensional keypoint annotations. Additionally, VIBE includes an adversarial learning framework based on self-attention mechanisms to distinguish between real human motion and motion generated by VIBE. HuMoR(Rempe et al. 2021) is a three-dimensional human motion model based on RGBD images, used for robust estimation of action sequences and body shape. Despite substantial progress in estimating three-dimensional human motion and shape through dynamic observation, recalculating a reasonable posture sequence in the presence of noise and occlusion remains a challenge. Therefore, HuMoR proposes a generative model in the form of a conditional variational autoencoder, which learns the distribution of posture changes at each step in the motion sequence.

Multimodal Human Motion Capture

Single-modality data has significant limitations, and multimodal work is underway. TotalCapture(Joo, Simon, and Sheikh 2018) collects human body posture using multi-view cameras and IMUs in a studio setting. 3DPW(Von Marcard et al. 2018) records pedestrian postures in urban environments through IMUs and handheld cameras. PedX(Kim et al. 2019) documents pedestrian postures using stereo images and LiDAR point clouds. FusionPose(Cong et al. 2023) reconstructs human posture using RGB and LiDAR body point clouds. LIP(Ren et al. 2023) reconstructs human posture using sparse IMUs and LiDAR. In addition to using monocular cameras and IMUs, LiDARHuman26M(Li et al. 2022), HSC4D(Dai et al. 2022), SLOPER4D(Dai et al. 2023), and CIMI4D(Yan et al. 2023) collect human posture information through static monocular LiDAR, body-carried LiDAR, head-mounted LiDAR, and LiDAR, respectively. ImmFusion(Chen et al. 2023) reconstructs human posture using RGB cameras and millimeter wave radar point clouds. Its algorithmic framework, as shown in Figure 1, includes feature extraction, multimodal data fusion, and human mesh recovery. The feature extraction part uses a pre-trained HR-Net network to extract image features, and PointNet++ is used for point cloud feature extraction. During multimodal data fusion, a human template mesh is incorporated, followed by the application of self-attention layers and feed-forward network layers, and finally, a graph convolutional neural network module is used to predict a rough 3D human mesh.

Pre-trained models

Pre-trained models are neural network models that have been trained on large-scale datasets, containing a vast number of parameters that can fully learn the characteristics of the input data. The main idea behind pre-trained models is to leverage massive datasets and computational resources to pre-train models in data-rich environments for subsequent fine-tuning or transfer learning tasks.

Pre-trained models can take various neural network architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers. These network structures are typically deep and highly parameterized, capable of learning rich feature representations.

Proposed Solution

The overall architecture of the Multi-modal Fusion Human Pose Estimation(MFHPE) algorithm based on image pre-trained models is shown in Figure 1. It mainly consists of three modules: the feature extraction module, the multi-modal feature fusion module, and the human mesh recovery module. The feature extraction module is responsible for extracting features from each frame of images and point clouds, returning the respective feature encoding vectors. Pre-trained models are utilized for image feature extraction. The multi-modal feature fusion module uses the features of the same frame extracted by the feature extraction module from images and point clouds. The human mesh recovery module, which includes two algorithm branches. The first branch regresses to obtain the vertex coordinates of the human mesh, and the second branch predicts the joint rotations in the SMPL parametric human model. By constraining the error between the predicted and ground-truth rotation matrices, this branch better guides the reconstruction of the human mesh.

Feature extraction

The input to the feature extraction module consists of corresponding images and point clouds from the same frame. The module includes two feature extractors: E_{im} and E_{pc} , which extract features from images and point clouds, respectively. The image feature extractor E_{im} utilizes pre-trained models whose network parameters do not require further training. In subsequent experiments, DINOv2, HRNet, and ResNet were used as pre-trained models. The point cloud feature extractor E_{pc} employs the PointNet++ model, whose parameters require training. The output of the feature extraction module includes Image Features $\in R^{batch\_size \times 1024}$ and Point Cloud Features $\in R^{batch\_size \times 1024}$, representing the fine-grained features extracted from the images and point clouds. These features capture the information of the corresponding frame's image and point cloud data. For image feature extraction, the process involves loading a pre-trained model and performing on-the-fly feature inference for each image before further operations. However, this approach requires repeating the inference process during every training session, resulting in high non-reusability and slower training speeds due to the inference overhead. Therefore, the approach was improved by pre-inferring the image features and saving the feature vectors as npy files, which can be loaded as needed during training. In this thesis experiment, DINOv2, HRNet, and ResNet were used to extract image features in advance. The extracted features represent an abstract expression of the information describing each image. The features extracted by different pre-trained models are stored in separate file paths.

Multi-modal feature fusion

This module first concatenates the two feature vectors using the concatenate operation and then feeds them into the Transformer(Vaswani 2017) module ϕ_T for feature fusion, as shown in the formula(1).

$$\text{Fused Feature} = \phi_T(\text{Image Feature}, \text{Point Cloud Feature}) \tag{1}$$

The Transformer module includes multi-head attention, feedforward networks, residual connections, and normalization layers. Multi-head attention enables learning diverse feature representations by combining results from different attention heads. The feedforward network enhances feature mapping and representation. Residual connections and normalization layers after each attention and feedforward module address gradient issues and accelerate training. The module's output is the Fused Feature $\in R^{batch\_size \times 1024}$, representing the feature vectors obtained by fusing the image and point cloud features.

Human mesh recovery

The first branch attaches the human template mesh vertices to the feature vectors, as shown in formula(2).

$$F' = concat(J^{Template}, V^{Template}, F) \tag{2}$$

Here, F represents the Fused Feature, F' represents the feature vector after the concatenation operation, $J^{Template} \in R^{24 \times 3}$,and $V^{Template} \in R^{431 \times 3}$. These are synthesized from the SMPL model with pose and shape parameters set to zero and obtained through two down-sampling operations. Concat refers to the concatenation operation. The concatenated features are then sequentially fed into the Attention+MLP network ϕ_{AM} (with a structure similar to a Transformer) and a Graph Convolutional Network (GCN). The GCN represents the human body joints as a graph structure, enabling regression of the human body mesh vertex coordinates, as shown in formula(3).

$$F'' = GCN(\phi_{AM}(F')) \tag{3}$$

After that, we obtain a coarse set of human body mesh vertex coordinates $V \in R^{batch\_size \times 455 \times 3}$ and joint position coordinates J $\in R^{batch\_size \times 24 \times 3}$ are obtained. The coarse vertex coordinates are then processed through two up-sampling networks to produce the predicted full human body mesh vertex coordinates $V \in R^{batch\_size \times 6890 \times 3}$. The second branch directly combines the predicted human joint coordinates with the feature vectors and uses a RotMatrix Regressor to jointly predict the human joint rotation matrices, as shown in formula (4).

$$R = RotMatrix(J, Upsample(F'')) \tag{4}$$

The predicted rotation matrix $R \in R^{\text{batch\_size} \times 24 \times 3 \times 3}$. The loss between the predicted rotation matrix and the ground truth is also included in the total loss function to guide the synthesis of the final human mesh. The RotMatrix Regressor mainly consists of a Graph Convolutional Network (GCN) structure module. Its input is the concatenated human joint coordinates and the multimodal fused feature vectors, and its output is a compressed feature representing the rotation vector parameters. With the joint coordinates and vertex coordinates of the human mesh, the SMPL human body parametric model can be used to render the 3D human mesh and visualize it.

Experiments

Experimental Setup

We evaluate the performance of our proposed Multimodal Fusion for Human Pose Estimation (MFHPE) framework on two publicly available datasets: LiDARHuman26M and RELI11D. These datasets provide multimodal data, including RGB images and LiDAR point clouds, along with ground-truth human pose annotations.

Implementation Details

The MFHPE framework is implemented using PyTorch 1.11.0, and trained on a system equipped with an Intel Xeon Silver 4216 CPU, 512GB of RAM, and an NVIDIA GeForce RTX 3090 GPU with 24GB memory, running Ubuntu 18.04.6. Training is conducted with a batch size of 48 over 50 epochs, using the Adam optimizer and an initial learning rate of 1e-4. A dynamic learning rate adjustment strategy is employed to progressively decrease the learning rate as training advances, ensuring stable convergence. Key metrics, including joint position loss, mesh vertex loss, and joint rotation loss, are logged to TensorBoard. The best-performing model on the validation set is saved, with checkpoints stored every 10 epochs to facilitate model evaluation and resumption. Training on the LiDARHuman26M dataset requires 7–10 days on the RTX 3090 GPU, highlighting the need for optimization to reduce computational costs.

Pre-trained Models

The feature extraction module in MFHPE leverages three pre-trained models:

- DINOv2: A self-supervised model that captures robust visual features without labeled data.
- HRNet: A high-resolution network that maintains detailed representations for pose estimation.
- ResNet: A widely-used residual learning framework for image feature extraction.

Evaluation Metrics

We employ the following metrics to quantify the performance of our framework:

- Mean Per Vertex Error (MPVE): Measures the average Euclidean distance between predicted and ground-truth mesh vertices.
- Mean Per Joint Position Error (MPJPE): Calculates the Euclidean distance between predicted and ground-truth joint positions.
- Procrustes Aligned MPJPE (PA-MPJPE): Evaluates MPJPE after rigid alignment to account for translation and rotation errors.

Quantitative Results

Results on LiDARHuman26M The quantitative results on the LiDARHuman26M dataset are presented in table 1. Among the tested pre-trained models, ResNet achieves the best performance with an MPJPE of 114.80 mm and a PA-MPJPE of 42.20 mm, outperforming both DINOv2 and HRNet. This demonstrates the effectiveness of ResNet in extracting image features for human pose estimation.

Results on RELI11D Table 2 summarizes the performance of MFHPE on the RELI11D dataset. While ResNet achieves the lowest MPVE of 213.77 mm, DINOv2 excels in MPJPE and PA-MPJPE metrics, with values of 85.22 mm and 51.90 mm, respectively. This highlights the robustness of DINOv2 in scenarios with sparse LiDAR data.

Table 1: Quantitative Comparison Results of Testing LiDARHuman26M

Model	MPVE	MPJPE	PA-MPJPE
DINOv2	247.15	134.02	50.77
HRNet	255.17	126.05	48.01
ResNet	249.47	114.80	42.20

Table 2: Quantitative Comparison Results of Testing RELI11D

Model	MPVE	MPJPE	PA-MPJPE
DINOv2	359.99	85.22	51.90
HRNet	235.74	135.20	65.81
ResNet	213.77	138.95	63.86

Figure 2 illustrates the loss curves during training. The steady decrease in training loss indicates that the model is converging effectively.

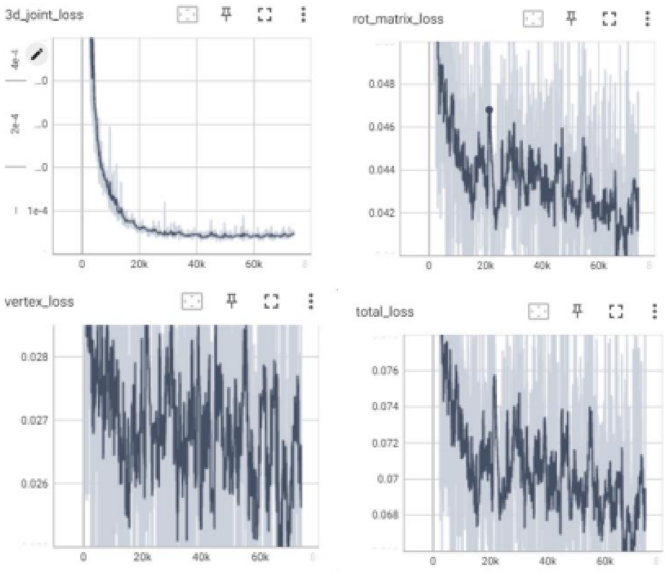

Figure 2: Curve of Loss Value Changes During Model Training

Figure 3 and Figure 4 show the input sequences for LiDARHuman26M and RELI11D, including both RGB images and LiDAR point clouds. These visualizations provide insight into the data used by the model for pose estimation.

Ablation Studies

In this section, we present ablation studies to evaluate the impact of the RotMatrix regressor. Figure 5 and table 3 present qualitative results of the ablation experiment, comparing human mesh recovery with and without the RotMatrix branch. The addition of the RotMatrix regressor improves joint rotations and overall pose estimation.

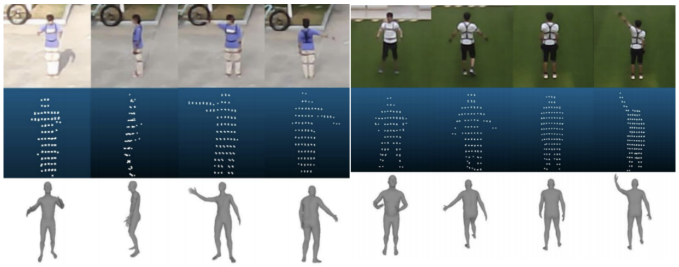

Figure 3: Partial Results Visualization on the LiDARHuman26M Test Set

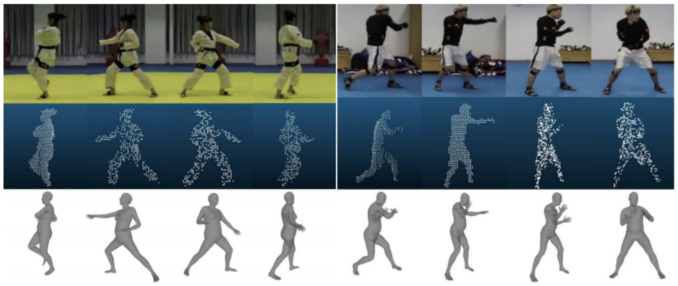

Figure 4: Partial Results Visualization on the RELI11D Test Set

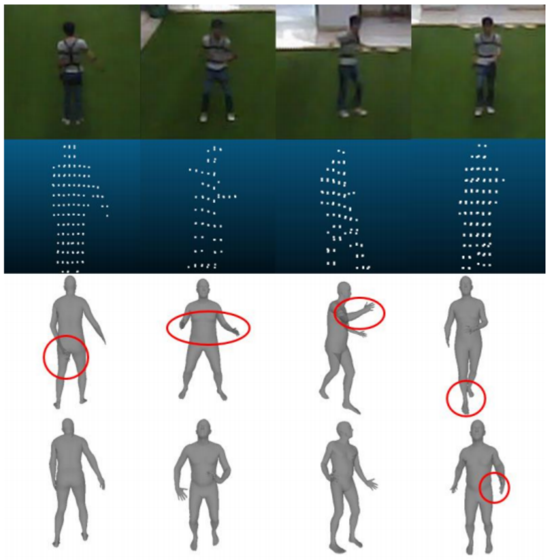

Figure 5: Ablation Study Results – Visual Comparison The first two rows show the inputs (RGB images and LiDAR point clouds). The third row displays results without the RotMatrix regressor, while the fourth row shows results with it. Red circles highlight anatomically incorrect poses, underscoring the importance of the RotMatrix regressor for realistic human motion capture.

Table 3: Quantitative Ablation Study Results

Configuration	MPVE	MPJPE	PA-MPJPE
With RotMatrix Regressor	134.02	50.77	247.15
Without RotMatrix Regressor	135.50	51.33	305.00

The inclusion of the RotMatrix regressor leads to a 19.0% reduction in MPVE, highlighting its importance in improving the accuracy of joint rotations and enhancing pose estimation performance.

Discussion

The experimental results validate the effectiveness of the MFHPE framework.

1. Model Comparison: ResNet consistently achieves better performance on LiDARHuman26M, while DINOv2 excels on RELI11D in challenging outdoor scenarios.
2. Ablation Insights: The RotMatrix regressor plays a crucial role in refining joint rotation predictions, with its inclusion leading to a 19.0% improvement in MPVE on LiDARHuman26M.
3. Dataset Observations: The variation in performance across datasets underscores the importance of selecting suitable pre-trained models for specific data characteristics.

Conclusion

Summary

This paper introduces MFHPE, a novel multimodal fusion approach for human pose estimation, leveraging pre-trained models to extract image features and fuse them with point cloud data. The framework utilizes DINOv2, HRNet, and ResNet for feature extraction and is tested on the LiDARHuman26M and RELI11D datasets. Results show that DINOv2 and ResNet perform best in terms of MPVE, MPJPE, and PA-MPJPE.

The human mesh recovery module features a dual-branch structure with a RotMatrix regressor, which improves the capture of realistic human motion. Ablation studies demonstrate that the RotMatrix regressor enhances the model's accuracy in joint rotations and overall pose estimation.

Improvements and Future Work

Although the proposed framework shows promising results, several improvements are needed:

Training Efficiency: Training on the LiDARHuman26M dataset takes 7–10 days on a NVIDIA GeForce RTX 3090 GPU, and further training could lead to better convergence. Future work will explore optimizing training strategies to reduce computational costs and improve efficiency.

Frame Consistency: Currently, the model processes each frame independently, leading to discontinuous motion. Incorporating temporal models or smoothness constraints between frames could address this issue.

SMPL Model Limitations: The use of the SMPL model for 3D visualization limits the capture of fine-grained motions (e.g., hand gestures). Replacing SMPL with SMPLX, which includes detailed hand and facial expression parameters, would enhance the model's performance.

References

Chen, A.; Wang, X.; Shi, K.; Zhu, S.; Fang, B.; Chen, Y.; Chen, J.; Huo, Y.; and Ye, Q. 2023. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2752–2758. IEEE.

Cong, P.; Xu, Y.; Ren, Y.; Zhang, J.; Xu, L.; Wang, J.; Yu, J.; and Ma, Y. 2023. Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 461–469.

Dai, Y.; Lin, Y.; Lin, X.; Wen, C.; Xu, L.; Yi, H.; Shen, S.; Ma, Y.; and Wang, C. 2023. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 682–692.

Dai, Y.; Lin, Y.; Wen, C.; Shen, S.; Xu, L.; Yu, J.; Ma, Y.; and Wang, C. 2022. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6792–6802.

Joo, H.; Simon, T.; and Sheikh, Y. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8320–8329.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kim, W.; Ramanagopal, M. S.; Barto, C.; Yu, M.-Y.; Rosaen, K.; Goumas, N.; Vasudevan, R.; and Johnson-Roberson, M. 2019. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4(2): 1940–1947.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5253–5263.

Li, J.; Zhang, J.; Wang, Z.; Shen, S.; Wen, C.; Ma, Y.; Xu, L.; Yu, J.; and Wang, C. 2022. Lidarcap: Long-range markerless 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20502–20512.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.

Rempe, D.; Birdal, T.; Hertzmann, A.; Yang, J.; Sridhar, S.; and Guibas, L. J. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11488–11499.

Ren, Y.; Zhao, C.; He, Y.; Cong, P.; Liang, H.; Yu, J.; Xu, L.; and Ma, Y. 2023. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE Transactions on Visualization and Computer Graphics*, 29(5): 2337–2347.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.

Yan, M.; Wang, X.; Dai, Y.; Shen, S.; Wen, C.; Xu, L.; Ma, Y.; and Wang, C. 2023. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12977–12988.