HSLR: Multi-stage Temporal 3D Human-Scene Reconstruction Based on the Point Cloud Sequence

Ming Yan
24520220157282
School of Medicine
Class Information

Yan Zhang
23020221154181
School of Information
Class Information

Shuqiang Cai
23020221154070
School of Information
Class Information

Jie Yang
24520220157283
School of Medicine
Class Information

Li Lin
23020221154099
School of Information
Class Information

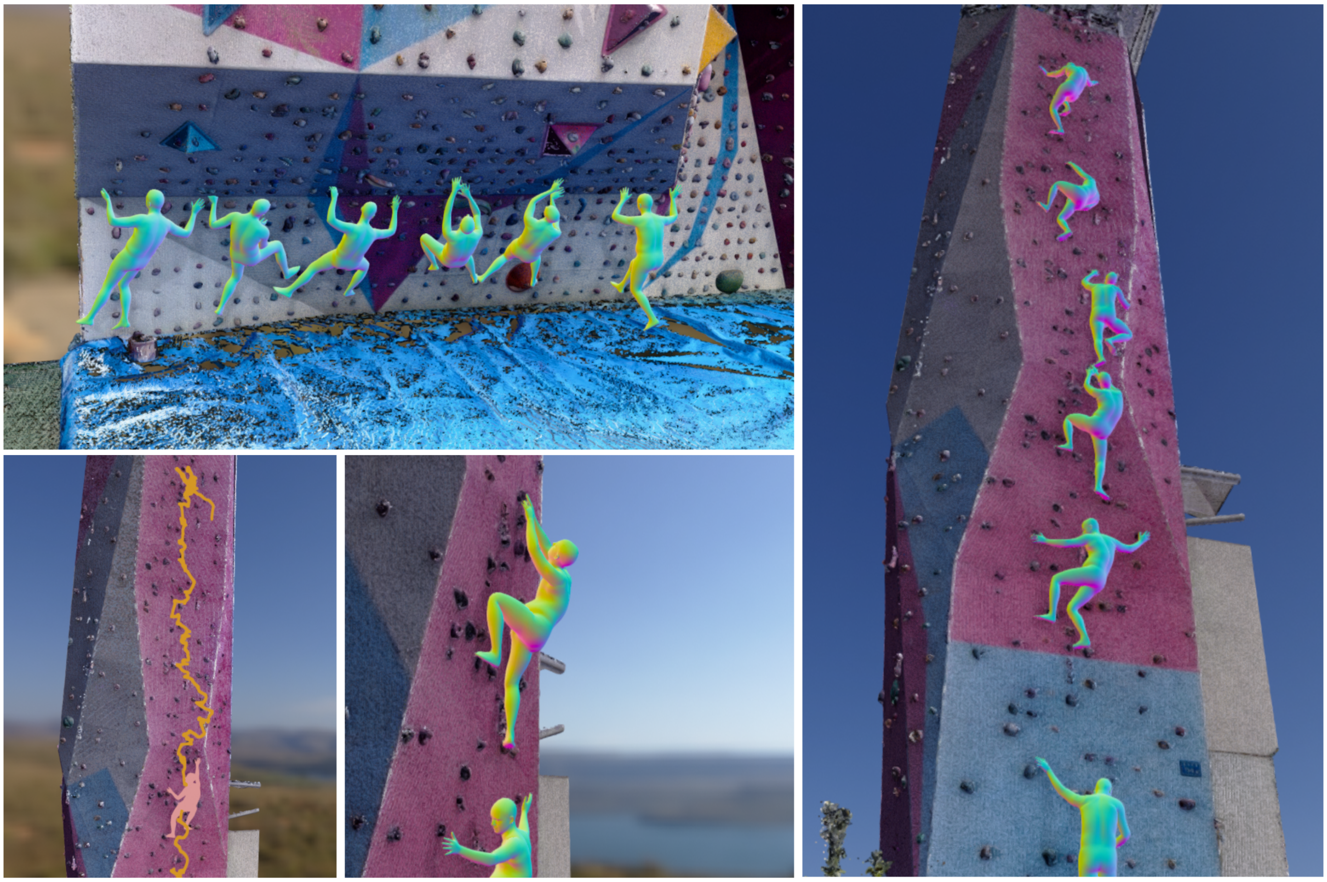

Figure 1. Overview of HSLR: We propose a method for 3D reconstruction of the point cloud sequence of the human body and the scene. The PointNet++ encoder is used to encode the point cloud sequence in time series, input the Point4Transformer network to resample the features, and then use ST-GCN to predict the based on the context information. Human body SMPL of point cloud sequence, and use a multi-stage optimized generation method to reconstruct the human body and the scene hybrid. At the same time, in order to test the generalization of HSLR, we also proposed a Human-Scene data set based on complex rock climbing.

Abstract

Most of the existing 3D HPE methods are limited to RGB(Standard Red Green Blue) cameras, and utilizing RGBD(RGB-Depth) adds additional data overhead. We propose HSLR (Human Scene Lidar Reconstration), a method that uses an efficient point cloud encoder for feature extraction, Transformer for resampling, and finally human SMPL(Skinned Multi-Person Linear Model) reconstruction using a time series graph convolutional neural network. Furthermore, in order to make the generated data closer to the real scene, we use a trajectory optimization method to estimate the global translation of the reconstructed human SMPL based on the accurate global 3D localization of the point cloud. We will conduct experiments to demonstrate the effectiveness of HSLR. ***This is a final paper for deep learning courses in 2022.***

1. Introduction

Humans live in a three-dimensional world, no matter when and where, the human body will always be in constant contact with the scenes or objects around them. Recovering human actions from scenes is critical for understanding human behavior, human-scene interaction synthesis, and virtual avatar creation. In recent years, many works have made progress in human and scene understanding [8,9,12,19,37], however, these datasets or methods are only aimed at the reconstruction of human daily actions. Limited by factors such as venue and equipment, many existing datasets can only complete the collection of simple indoor or outdoor scenes and simple actions. If you want to capture human movements in large scenes that are difficult to move, such as mountains in nature, huge rock climbing walls and other complex scenes, the previous devices are often powerless. In addition, most of the daily actions of the existing datasets are in contact with the ground under natural gravity. The interaction between the human body and the environment is not complicated, and most of them only have contact with the feet and other parts intermittently. The capture and reconstruction of complex human-scene brings new challenges to today's computer vision.

Most of the existing 3D HPE methods are limited to monocular or multi-purpose RGB or RGBD data. RGBD cameras add unnecessary overhead for reconstructing the human body, and only the acquisition of depth information is not as accurate as radar equipment. Therefore, we need to propose a human body and scene reconstruction algorithm based on radar point cloud, which can be robust to the 3D reconstruction of human body and scene under the premise of using the most simplified hardware equipment.

In order to solve the above problems, the contributions of this work are as follows: **1. A dataset for fine-grained reconstruction of people and scenes is proposed. 2. A robust method for 3D human pose estimation from point clouds is proposed.**

2. Related Work

2.1. Human Pose Datasets

Recently, deep neural network-based approaches have made significant progress in estimating the pose and shape of human from images, video and inertial measurement units (IMU).

As deep neural network approaches are data-driven. The focus of human pose estimation research is partially driven by the design of datasets. To recover 2D pose from RGB videos, PennAction [42] and PoseTrack [1] are the two datasets with ground-truth annotations. Kinetics-400 [4] and InstaVariety [14] are created through 2D keypoint detectors. The label provided by them is pseudo ground-truth. SURREAL [32] provides a synthetically-generated human pose data rendered from human motion capture data.

For 3D human pose estimation, researchers have collected multiple datasets. HumanEva [29] contains 4 subjects performing a set of predefined actions within indoor scenarios, and with static background. It provides the community with synchronized motion capture and multi-view video data. The actions in HumanEva contain walking, jogging, throw/catch, gesture, boxing, etc. All these actions are ground-base actions.

3DPW [33] is an in-the-wild 3D dataset that collected through a set of IMU sensors and a hand-held camera. It contains 51,000 video frames of several outdoor and indoor activities performed by 7 actors. The activities of 3DPW contain walking, sitting, going up-stairs, and taking bus. The PedX [16] is pedestrian pose dataset. It consists of 5,000 pairs of stereo images and LiDAR data. It provides 3D pseudo label through a 3D model fitting algorithm.

AMASS [22] is a large-scale MoCap dataset. Spans over 300 subjects and contains 40 hours of motion sequences. It is widely used as motion priors for pose estimation task.

The LiDARHuman26M [20] is a multi-modality dataset which consists of LiDAR point clouds, RGB videos, and IMU data. It records 13 actors performing 20 daily activities (e.g., walking, running, phoning, bowling) from long distance in 2 controlled scenes. HSC4D [5] is a human-centered 4D scene capture dataset for human pose estimation and localization. It is collected by body-mounted IMU and LiDAR through walking in 3 scenes.

Based on the discussion above, none the above datasets contains the climbing actions, which is covered by this work.

2.2. Pose Estimation Methods

Extensive work has focused on estimating the pose, shape, and motion of human from pure vision-base data.

SIMPLify-X [23] compute the human pose, hand pose, and face expression from a single monocular image. PARE [18] address the occlusion issues through learning body-part-guided attention mask. FuturePose [34] model movement features using optical flow, and predict the movement of skeleton human joints through a LSTM module. GLAMR [39] estimate global human mesh with dynamic cameras. PiFu [26] and PiFuHd [27] estimate clothed human from RGB images using implicit representations. ICON [35] improves them through using the SMPL priors [2]. S3 [38] represents human pose, shape, and clothing as neural implicit function, and estimate them from a single image or a single LiDAR sweep.

RobustFusion [30] proposes a robust human volumentric capture system using a single RGBD camera without prescanned template. EventCap [36] uses a CNN-based human pose detection module and a optimization method to capture human motion using an event camera. LiDARCap [20] estimate 3D human pose using LiDAR point clouds with PointNet-based neural networks.

Human pose priors are used in pose estimation tasks [41]. Most of them learn prior from the AMASS dataset [22]. HMR [13] regresses human poses and shape from RGB images in a end-to-end fashion. VIBE [17] estimate human pose from RGB videos through using AMSSS dataset as a prior. HMR and VIBE both use discriminators to discrimate between poses from AMASS and poses estimated by them. VPose [24] transforms the pose space into Gaussian spaces through variational autoencoder. HuMOR [25] learns the pose change distribution based on conditional variational autoencoder. Pose-NDF [31] learns a continuous model for plausible human pose based on neural distance fields.

A few work considers human scene interactions, PROX [10] estimates human poses with scene constraints. POSER [11] populates scenes with realistic human poses. LEMO [41] learns a motion smoothness prior, and considers the contacts among humans and scenes.

3. Method

In order to accurately evaluate the performance of various methods under the premise of complex actions, we also collected the corresponding high -difficult rock climbing data set and high -precision three -dimensional reconstruction scene while proposing the HSLR algorithm. The content of this section will be explained from data calibration and specific model details.

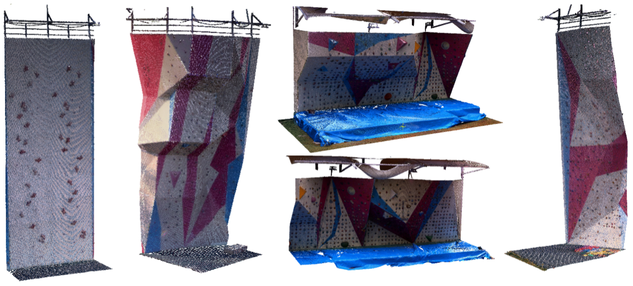

Figure 2. HSLR provides high-quality 3D reconstruction of RGB point cloud scenes.

3.1. Coordinates

Coordinate Systems. We define three coordinate systems: 1) IMU coordinate system $\{I\}$: origin is at the pelvis joint of the first SMPL model, and $X/Y/Z$ axis is pointing to the right/upward/forward of the human. 2) LiDAR Coordinate system $\{L\}$: origin is at the center of the LiDAR, and $X/Y/Z$ axis is pointing to the right/forward/upward of the LiDAR. 3) Global/World coordinate system $\{W\}$: the scene's coordinate we manually define. We use the right subscript $k, k \in \ Z^+$ to indicate the index of a frame, and the right superscript, I or L or W (default to W), to indicate the coordinate system that the data belongs to. For example, the 3D point cloud frames from LiDAR is represented as $P^L = \{P_k^L, k \in \ Z^+\}$

Coarse calibration. Before data capturing, the actor stands facing or parallel to a large real-world object with a flat face, such as a wall or a square column. His right/front/up is regarded as the scene's $X/Y/Z$ axis direction, and the midpoint of his ankles' projection on the ground is set as the origin. After the data are collected, we manually find the first frame's ground plane and the object's plane, and then calculate their normal vector $g = [g_1, g_2, g_3]^\top$ and $m = [m_1, m_2, m_3]^\top$, respectively. The coarse calibration matrix R_{WL} from the LiDAR starting position to the world coordinate $\{W\}$ is calculated as:

$$R_{WL} = \begin{bmatrix} e_1 & e_2 & e_3 & 0 \\ m_1 & m_2 & m_3 & 0.2 \\ g_1 & g_2 & g_3 & h \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where $[e_1, e_2, e_3]^\top \ = \ m \times g$ and h is the height of the LiDAR from the ground. Based on the definition of IMU coordinate system $\{I\}$, the coarse calibration matrix R_{WI} from $\{I\}$ to $\{W\}$ is defined as: $R_{WI} \ = \left[(1, 1, -1)(2, 3, 1)(3, 2, 1)(4, 4, 1)\right]_{triad}$

3.2. Notation

We use the right subscript $k, k \in \ Z^+$ to indicate the index of a frame, and the right superscript, I or L or W

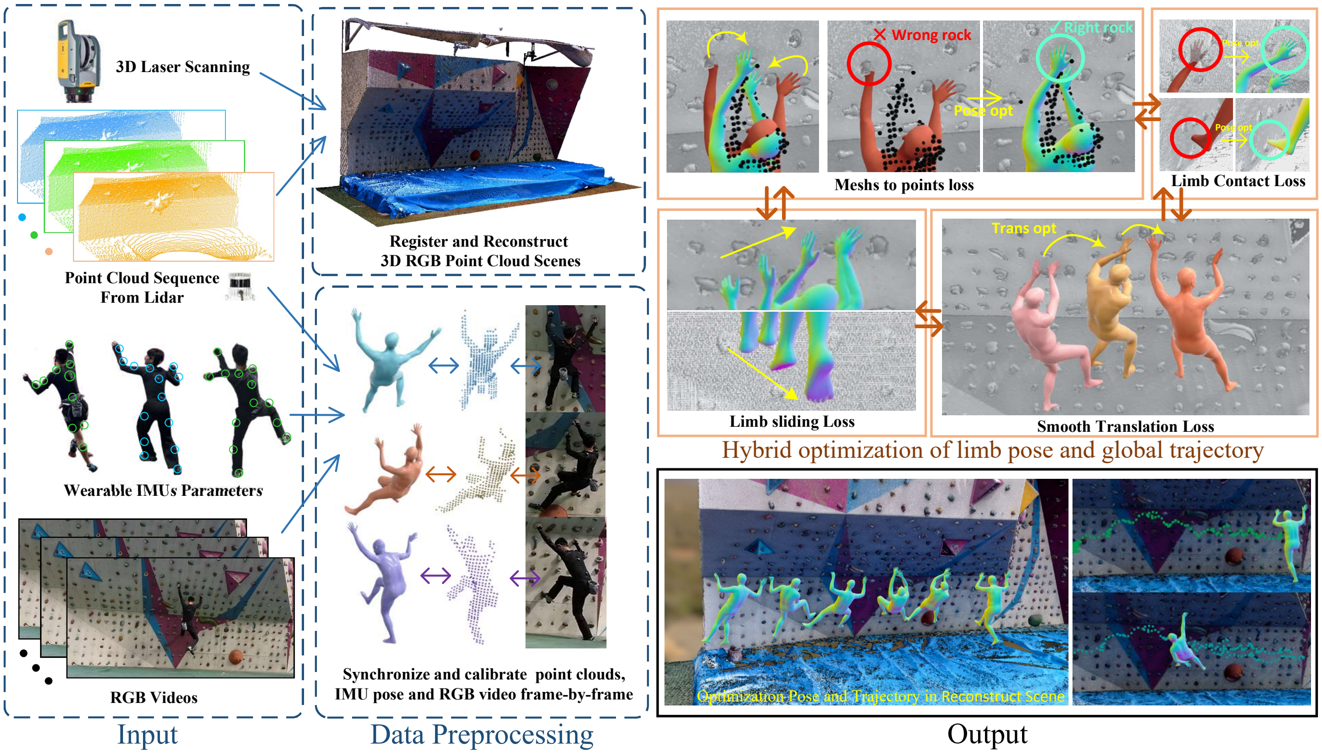
3D Laser Scanning
Point Cloud Sequence From Lidar
Wearable IMUs Parameters
RGB Videos
Input
Register and Reconstruct 3D RGB Point Cloud Scenes
Synchronize and calibrate point clouds, IMU pose and RGB video frame-by-frame
Data Preprocessing
Wrong rock
Right rock
Pose opt
Meshs to points loss
Pose opt
Pose opt
Limb Contact Loss
Trans opt
Limb sliding Loss
Smooth Translation Loss
Hybrid optimization of limb pose and global trajectory
Optimization Pose and Trajectory in Reconstruct Scene
Output

Figure 3. **Overview of main annotation pipeline.** The blue arrows indicate data flows, and the yellow arrows represent the direction of optimization. **Dotted box:** The input of each scene consists of RGB videos, point cloud sequence, IMU measurements, and 3D laser scanning data. Data pre-processing stage calibrates and synchronizes different modalities. **Solid box:** The blending optimization stage optimize including the pose and translation based on multiple constraint losses.

(default to W), to indicate the coordinate system that the data belongs to. For example, the 3D point cloud frames from LiDAR is represented as $P^L = \{P^L_k, k \in Z^+\}$ and the 3D scene is represented as $\boldsymbol{S}$. M^W_k indicates the k-th frame in human motion $M = (T, \theta, \beta)$ in world coordinate system, where T is the $N \times 3$ translation parameter, θ is the $N \times 24 \times 3$ pose parameter, and β is the $N \times 10$ shape parameter. We use the Skinned Multi-Person Linear (SMPL) [21] body model Φ to map k-th frame's motion representation M_k to its triangle mesh model, $V_k, F_k = \Phi(M_k)$, where body vertices $V_k \in \mathbb{R}^{6890\times3}$ and faces $F_k \in \mathbb{R}^{13690\times3}$.

3.3. Reconstruction Scene

In the research of human-scene interaction, we consider that accurate scene reconstruction is vital for the method understanding. Previous works reconstruct scenes using depth cameras [7, 10, 40, 41] with much lower accuracy than LiDAR and cannot check large scenes. HSLR uses Trimble X7 to scan 3D scene information and rebuild the precisely measured scene in space. We provide 7 high-precision reconstruction scenes with a total point cloud amount of 40M, as shown in Fig. 2.

3.4. Blending Optimization Loss

We utilize scene and physical constraints to perform a blending optimization of pose and translation to obtain accurate and scene-natural human motion $\boldsymbol{M^W}$ annotation. The following constraints are used: the limb contact constraint $\mathcal{L}_{ct}$ encourages reasonable hand and foot contact with the scene mesh without penetrating. The limb sliding constraint $\mathcal{L}_{sld}$ eliminates the unreasonable slippage of the limbs during climbing. The smoothness constraint $\mathcal{L}_{smt}$ makes the translation, orientation, and joints remain temporal continuity. The mesh to point constraints $\mathcal{L}_{m2p}$ minimizing the distance between constructed SMPL vertices to the point clouds of human body. As shown in 3

The optimization is expressed as:

$$\begin{aligned} \mathcal{L} &= \lambda_{ct}\mathcal{L}_{ct} + \lambda_{sld}\mathcal{L}_{sld} + \mathcal{L}_{smt} + \lambda_{m2p}\mathcal{L}_{m2p} \\ \boldsymbol{M} &= \arg\min_{\boldsymbol{M}} \mathcal{L}(\boldsymbol{M}|T^W, \theta^I, R^W, \boldsymbol{S}) \end{aligned} \quad (2)$$

where λ_{ct}, λ_{sld}, λ_{smt}, λ_{m2p} are coefficients of loss terms. $\mathcal{L}$ is minimized with a gradient descent algorithm that optimize $\boldsymbol{M^W} = (T, \theta)$. $\boldsymbol{M^W}$ is initialized according to Paper Sec 3.3.

Limb contact Loss. This loss is defined as the distance from a stable foot or hand to its nearest neighbor in the scene vertices. First, we detect the foot and hand state based on its movements. The movement is calculated based on the set of vertices of hands and feet. One limb is marked as stable if its movement is smaller than $3cm$ and smaller than another limb (foot or hand)'s movement. We obtain the contact environment near the stable limb through a neighbor search. The limb contact loss is $\mathcal{L}_{ct} = \mathcal{L}_{\text{ct}_{feet}} + \mathcal{L}_{\text{ct}_{hand}}$.

$$\mathcal{L}_{\text{ct}_{feet}} = \frac{1}{l}\sum_{j=1}^{l-1}\sum_{v\in VF^{\mathcal{SF}_j}}\frac{1}{|VF^{\mathcal{SF}_j}|}\|v_f - \widetilde{v_f}\cdot pf_j\|_2 \quad (3)$$

$$\mathcal{L}_{\text{ct}_{hand}} = \frac{1}{l}\sum_{i=1}^{l-1}\sum_{v\in VH^{\mathcal{SH}_i}}\frac{1}{|VH^{\mathcal{SH}_i}|}\|v_h - \widetilde{v_h}\cdot ph_i\|_2 \quad (4)$$

where $\widetilde{v_f}$ and $\widetilde{v_h}$ is homogeneous coordinate of v_f and v_h. $VF^{\mathcal{SF}_j}$ and $VH^{\mathcal{SH}_i}$ are the sets of the vertices of a stable foot $\mathcal{SF}_j$ and a stable hand $\mathcal{SH}_i$. The loss is average over all frames of a sequence with length l.

Limb sliding Loss. This loss reduces the motion's sliding on the contact surfaces, making the motion more natural and smooth. The sliding loss is defined as the distance of a stable limb over every two successive frames: $\mathcal{L}_{sld} = \mathcal{L}_{\text{sld}_{feet}} + \mathcal{L}_{\text{sld}_{hands}}$.

$$\mathcal{L}_{\text{sld}_{feet}} = \frac{1}{l}\sum_{j=1}^{l-1}\|\mathbb{E}(VF^{\mathcal{SF}_{j+1}}) - \mathbb{E}(VF^{\mathcal{SF}_j})\|_2 \quad (5)$$

$$\mathcal{L}_{\text{sld}_{hands}} = \frac{1}{l}\sum_{i=1}^{l-1}\|\mathbb{E}(VH^{\mathcal{SH}_{i+1}}) - \mathbb{E}(VH^{\mathcal{SH}_i})\|_2 \quad (6)$$

where $\mathbb{E}(\cdot)$ calculates the center of the vertices list.

Smooth Loss. The smooth loss includes the translation term $\mathcal{L}_{trans}$ and the joints term $\mathcal{L}_{joints}$.

$$\mathcal{L}_{smt} = \lambda_{trans}\mathcal{L}_{trans} + \lambda_{joints}\mathcal{L}_{joints} \quad (7)$$

The $\mathcal{L}_{trans}$ smooths the trajectory T of human (the translation of the pelvis) through minimizing the difference between LiDAR and a human's translation difference. The smooth term is as follows:

$$\mathcal{L}_{trans} = \frac{1}{l}\sum_{j=1}^{l-1}\max(0, \|T^L_{j+1} - T^L_j\|_2 - \|T_{j+1} - T_j\|_2) \quad (8)$$

where T^L_k is the translation of LiDAR at k-th frame, and T_k is the translation we optimized for. The $\mathcal{L}_{joints}$ is the term that smooths the motion of body joints in global 3D space, which minimizes the mean acceleration of the joints. For this loss, we only consider stable joints on the torso and the neck. Let $\delta^s_j = J^s_j - J^s_{j-1}$ represent the difference of joints between consecutive frame. $\mathcal{L}_{joints}$ is defined as follows.

Constraint term			Scene			
$\mathcal{L}_{cont}$	$\mathcal{L}_{smt}$	$\mathcal{L}_{m2p}$	Vertical 1	Vertical 2	Horizontal 1	Horizontal 2
✗	✗	✗	48.28	60.04	59.83	47.74
✓	✓	✗	22.64	28.33	41.67	26.64
✓	✗	✓	33.48	40.44	44.77	31.44
✗	✓	✓	24.64	38.37	42.07	30.08
✓	✓	✓	**16.24**	**23.46**	**34.34**	**20.21**

Table 1. Loss of the optimization stage for different constraints

$$\mathcal{L}_{joints} = \frac{1}{l}\sum_{j=1}^{l-1}\|\delta^s_{j+1} - \delta^s_j\|_2 \quad (9)$$

Since the static scenes are collected in Paper Sec 3.1, we design a method to segment human point clouds as annotation data. For each frame of dynamic LiDAR output, we manually register to the same coordinate system of the IMU to obtain the RT matrix. Next, the human body in the multi-frame dynamic scene is manually removed to generate a sparse static scene. For each frame of point cloud, the points within the threshold range of the sparse scene are eliminated to obtain the segmented human point cloud $\mathcal{P}_i$. For each segmented human point cloud $\mathcal{P}_i$.

Mesh to point loss. For each estimated human meshes, we use Hidden Points Removal (HPR) [15] to remove the invisible mesh vertices from the perspective of LiDAR. Then, we use Iterative closest point (ICP) [28] to register the visible vertices to $\mathcal{P}$, which is segmented human point clouds. We re-project the human body mesh in the LiDAR coordinate to select the visible human body vertices V'. For each frame, We use $\mathcal{L}_{m2p}$ to minimize the 3D Chamfer distance between human points $\mathcal{P}_i$ and vertices ${V'}_i$. More details about loss terms definition are given in the appendix. For each frame, the $\mathcal{L}_{m2p}$ constraint is regularized with the following equation:

$$\begin{aligned}\mathcal{L}_{m2p} = \frac{1}{|\mathcal{P}|}\sum_{p_i\in\mathcal{P}}\min_{v_i\in V'}\|p_i - v_i\|^2_2 +\\ \frac{1}{|V'|}\sum_{v_i\in V'}\min_{p_i\in\mathcal{P}}\|v_i - p_i\|^2_2\end{aligned} \quad (10)$$

3.5. Quantitative evaluation.

To understand the impact of different constraints used in the optimization stage, we conduct ablation study of 3 different constraints: L_{cont}, L_{smt} and L_{m2p}. Tab. 1 shows the loss of using different combinations of constraints for motions from 4 scenes. The loss is an indicator of violation of motion constraints. Without using any term, the loss

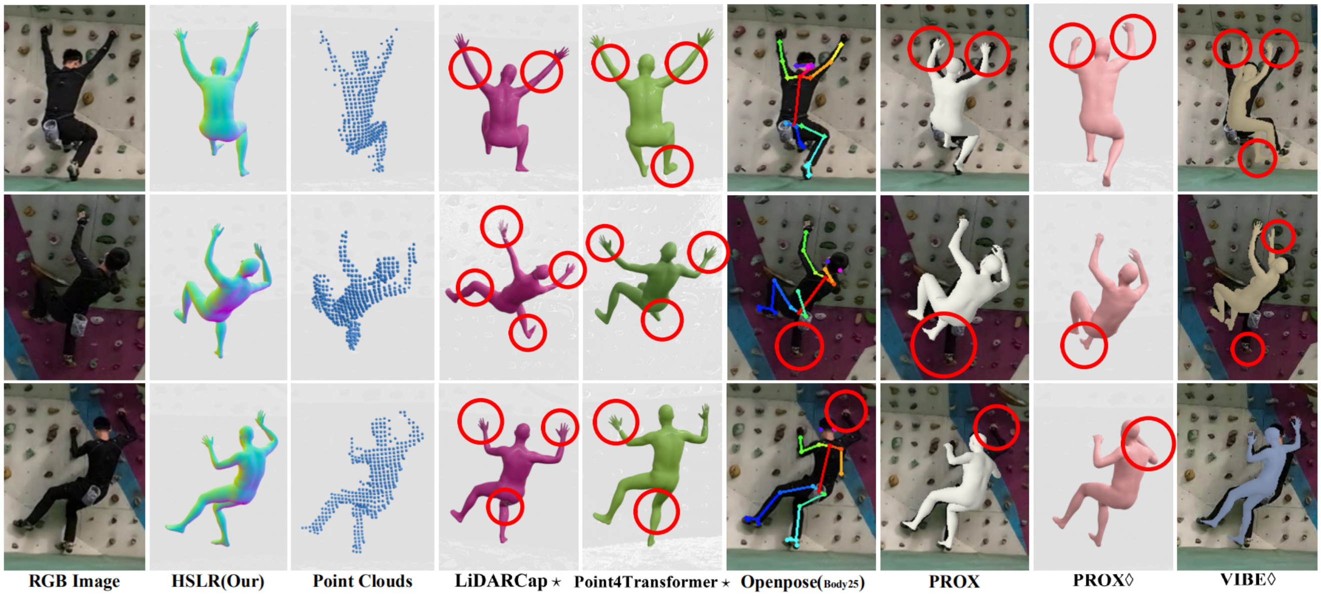

Figure 4. Qualitative results of several algorithms on the HSLR dataset. It is challenging to reconstruct a climbing pose with high ductility, even if algorithms are re-trained (marked by ⋆) or fine-tuned (marked by ◇) based on HSLR. As indicated by the red circles, all these methods have artifacts for limbs. As a scene-aware method, PROX performs better than other methods that do not use scene constraints. This suggests that it is necessary to consider the human-scene interaction annotation provided in HSLR.

is largest, which suggests that motions may seem unnatural. The L_{ct} and L_{smt} terms can reduce total loss, which indicates that they can improve the overall quality of data. Combing L_{m2p} can further improve the quality of motions. Overall, all constraint terms are necessary to produce accurate and smooth human pose and translation.

4. Experiment

Evaluation metrics In this section, we report Procrustes-Aligned Mean Per Joint Position Error (PMPJPE), Mean Per Joint Position Error (MPJPE), Percentage of Correct Keypoints (PCK), Per Vertex Error (PVE), and Acceleration error(m/s^2) (ACCEL). Except ACCEL, error metrics are measured in millimeters.

Input	Method	ACCEL↓	PMPJPE↓	MPJPE↓	PVE↓	PCK0.5↑
LiDAR	LiDARCap	12.39	222.11	358.13	422.65	0.50
	LiDARCap⋆	2.59	86.38	115.93	136.83	0.90
	P4Transformer⋆	3.32	100.58	130.99	156.27	0.87
RGB	VIBE	68.02	770.77	287.14	857.83	0.17
	VIBE◇	57.88	161.21	116.78	187.70	0.76
	MAED	33.61	135.57	472.82	515.46	0.25
	MAED◇	17.50	135.57	170.43	197.66	0.74
Scene	PROX	-	109.34	265.34	279.50	0.53
	PROX◇	-	109.33	147.41	165.12	0.79
LiDAR&Scene	HSLR(Ours)	**0.62**	**80.73**	**70.06**	**94.84**	**0.95**

Table 2. Comparison of pose estimation by SOTA on different modal data. ⋆ indicates training based on the HSLR dataset. ◇ denotes fine-tuned based on the HSLR dataset. Other experiments used the pretrained model of the original method.

Pose Estimation In this task, the poses of climbing humans are estimated from RGB imagery or LiDAR point clouds based on the HSLR dataset. For the methods evaluated in this section, VIBE [17] estimate poses from RGB images, while LiDARCap and P4Transformer [6] recover the motions from point clouds. The qualitative results of pose estimation are depicted in Fig. 4. As it is pointed out by the red circles in this figure, all these methods have artifacts. The quantitative results are depicted in Tab. 2. The pretrained model of LiDARCap does not perform well (PCK0.5= 0.46) on HSLR. Further, we train LiDARCap and P4Transformer from scratch based on HSLR. Compared with the indicators of the original paper, their performance is also not satisfactory. The RGB-based approach (VIBE) does not perform well on this dataset too. After fine-tuning on CIME4D, the performance of VIBE is improved. However, the performance is still poor compared to the original paper. Overall, the error metrics for all these methods are high. This indicates that HSLR is a challengding dataset for human pose estimation.

Pose Estimation with Scene Constraints PROX [10] is the most frequently used dataset for estimating the human body in the Scene currently. However, because most of the motion in PROX are daily movements, such as walking, standing and sitting, there are no complex and high stretch actions. In this task, we choose PROX estimates human poses from RGB images with 3D scene constraints. PROX obtains human skeleton information from monocular RGB images using openpose [3]. We convert the mesh scene provided by HSLR into *sdf* form to test PROX. Due to the inaccurate positioning of PROX in 3D space, we use HSLR finetunes PROX to focus on estimating human poses. Tab. 2 shows in the quantitative evaluation results, and Fig. 4 is a qualitative comparison. We observe that due to the defects of RGB images, when the color of the background is similar to the texture and the clothes of the volunteers, the current state-of-the-art 2D joint point detection algorithm cannot detect the human skeleton in rock climbing well. We fine-tune PROX, but the results show that in the HSLR scene, the human joints reconstructed by PROX have serious deviations, and the movements of the volunteers are not correctly restored.

From the experimental results of qualitative and quantitative analysis, we can observe that we propose that the HSLR method has the most advanced performance than SOTA, and the entire action has the highest degree of reduction and the smoothest. It can be seen from the results of the analysis matrix that compared to LiDARCap and pure Point4Transformaer, our PVE indicator exceeded the 100 mark for the first time. In the result of visualization, compared with other comparison methods, the HSLR proposed by our proposal can accurately estimate the position of the bone point, and the hand point and foot point at the end of the limb can be relatively accurate.

5. Conclusion

This work proposes a point cloud-based human pose estimation method, which uses additionally collected high-precision scene data to estimate more accurate pose and trajectory. Ablation experiments demonstrate the benefits of multiple optimizations in the HSLR approach. In the qualitative and quantitative experiments, compared with the current SOTA method based on point cloud and RGB, it can be seen from multiple scenes and actions that our proposed method has the best effect, and can accurately restore the movement details of the extremities.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 3

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2

[5] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. 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 (CVPR)*, pages 6792–6802, June 2022. 2

[6] Hehe Fan, Yi Yang, and Mohan S. Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14199–14208, 2021. 6

[7] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 4

[8] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):14:1–14:17, 2019. 2

[9] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[10] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2282–2292. IEEE, 2019. 3, 4, 6

[11] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14708–14718. Computer Vision Foundation / IEEE, 2021. 3

[12] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11400–11411, 2021. 2

[13] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 3

[14] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5607–5616, 2019. 2

[15] Sagi Katz, Ayellet Tal, and Ronen Basri. Direct visibility of point sets. In *ACM SIGGRAPH 2007 papers*, pages 24–es. 2007. 5

[16] Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charlie Barto, Ming-Yuan Yu, Karl Rosaen, Nicholas

Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4:1940–1947, 2019. 2

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 3, 6

[18] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11127–11137, October 2021. 3

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[20] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20502–20512, 2022. 2, 3

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015. 4

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 3

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, June 2019. 3

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 3

[25] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11488–11499, October 2021. 3

[26] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2304–2314. IEEE, 2019. 3

[27] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 81–90. Computer Vision Foundation / IEEE, 2020. 3

[28] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009. 5

[29] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 2009. 2

[30] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *ECCV*, 2020. 3

[31] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, October 2022. 3

[32] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 2

[33] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2

[34] Erwin Wu and Hideki Koike. Futurepose - mixed reality martial arts training using real-time 3d human pose forecasting with a RGB camera. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1384–1392. IEEE, 2019. 3

[35] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: implicit clothed humans obtained from normals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13286–13296. IEEE, 2022. 3

[36] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Ming Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4967–4977, 2020. 3

[37] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):27:1–27:15, 2018. 2

[38] Ze Yang, Shenlong Wang, Siva Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 3

[39] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: global occlusion-aware human mesh recovery with dynamic cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11028–11039. IEEE, 2022. 3

[40] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 180–200, Cham, 2022. Springer Nature Switzerland. 4

[41] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11323–11333. IEEE, 2021. 3, 4

[42] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. *2013 IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 2