Check for updates

DOI: 10.1111/cgf.14961
Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

MOVIN: Real-time Motion Capture using a Single LiDAR

Deok-Kyeong Jang[†1,2], Dongseok Yang[†1,2], Deok-Yun Jang[†1,3], Byeoli Choi[†1,2], Taeil Jin[2] and Sung-Hee Lee[‡2]

[1]MOVIN Inc.
[2]Korea Advanced Institute of Science and Technology (KAIST)
[3]Gwangju Institute of Science and Technology (GIST)

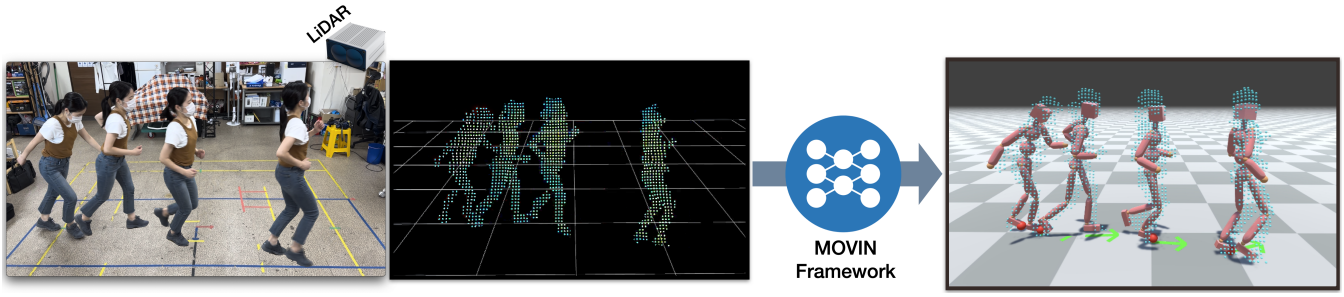


Figure 1: *Our MOVIN framework enables real-time full-body motion capture with global translation from 3D LiDAR point cloud.*

Abstract

Recent advancements in technology have brought forth new forms of interactive applications, such as the social metaverse, where end users interact with each other through their virtual avatars. In such applications, precise full-body tracking is essential for an immersive experience and a sense of embodiment with the virtual avatar. However, current motion capture systems are not easily accessible to end users due to their high cost, the requirement for special skills to operate them, or the discomfort associated with wearable devices. In this paper, we present MOVIN, the data-driven generative method for real-time motion capture with global tracking, using a single LiDAR sensor. Our autoregressive conditional variational autoencoder (CVAE) model learns the distribution of pose variations conditioned on the given 3D point cloud from LiDAR. As a central factor for high-accuracy motion capture, we propose a novel feature encoder to learn the correlation between the historical 3D point cloud data and global, local pose features, resulting in effective learning of the pose prior. Global pose features include root translation, rotation, and foot contacts, while local features comprise joint positions and rotations. Subsequently, a pose generator takes into account the sampled latent variable along with the features from the previous frame to generate a plausible current pose. Our framework accurately predicts the performer's 3D global information and local joint details while effectively considering temporally coherent movements across frames. We demonstrate the effectiveness of our architecture through quantitative and qualitative evaluations, comparing it against state-of-the-art methods. Additionally, we implement a real-time application to showcase our method in real-world scenarios. MOVIN dataset is available at https://movin3d.github.io/movin_pg2023/.

CCS Concepts

• ***Computing methodologies*** → ***Motion capture; Motion processing; Neural networks;***

1. Introduction

With the increasing demand for immersive and interactive experiences in the fields of filming, animation, and the metaverse, real-time motion capture has become an essential technology for animating virtual characters to realize interactions with the virtual environment and between the users. However, state-of-the-art motion capture technologies, including optical and inertial sensors, are hardly affordable for general users for their price and inconvenience.

To address the limitations, researchers focused on utilizing com-

† Equal contribution
‡ Corresponding author

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

monly accessible sensors, such as cameras and VR tracking devices, to achieve high-quality motion capture in real-time. Recent deep learning approaches robustly predict the full-body pose, leveraging single RGB video stream [LLZ*22], sparse sets of IMU sensors [HKA*18], or sparse VR tracker configurations [YKL21]. Despite the promising results, there still remains a significant gap that requires improvement; 2D image-based methods suffer from inaccurate global translation, and sparse trackers provide only underdetermined constraints that cannot disambiguate different poses with the same tracker configuration. Furthermore, these sensors often face inherent limitations including magnetic interference from surrounding electronics and optical occlusion in complex indoor environments, which can degrade the accuracy and reliability of the captured motion data.

The recent success of LiDAR sensors in object and human detection for autonomous driving [ZTJF21] demonstrates the potential of LiDAR to significantly improve the performance and usability of current motion capture technologies. Unlike traditional 2D camera systems, LiDAR sensors can provide reliable and precise 3D positions of the tracking target in the form of the point cloud. Moreover, this LiDAR-generated point cloud can provide full-body information about the subject, which is not available for sparse configuration of wearable trackers. Pioneering works already proved that LiDAR sensors can assist existing motion capture technologies and increase accuracy for long-range human pose detection [LZW*22]. Furthermore, as LiDAR technology has found applications across various industrial sectors, including security and smart cities, increasing demand has prompted mass production, lowering LiDAR prices.

This paper introduces MOVIN, a novel framework for real-time motion capture using a single LiDAR sensor, as illustrated in Fig. 1. To the best of our knowledge, our framework is the first LiDAR-based real-time full-body motion capture with global translation.

Our model employs an autoregressive conditional variational autoencoder (CVAE) architecture to establish the relationship between the input point cloud and the output full-body motion, considering the previous output motion. The encoder component, based on the Transformer architecture, maps encoded features to a multivariate Gaussian distribution. Meanwhile, the decoder component follows a Mixture-of-Expert architecture, generating output features by sampling from the distribution while incorporating condition features.

To address the distinct characteristics of the input and conditions, we have carefully designed input/output and feature embedding/expanding modules. The input 3D point cloud integrates current and subsampled data from the past 1-second interval to ensure temporal coherence of the output motion. We process the condition, which represents the output of the previous frame, separately for local and global pose features. The local pose feature of joint local transformations is processed via a skeleton-aware Graph Convolutional Network that preserves inherent body part structure. The global pose feature, which includes root position, rotation, and foot contacts, is handled using Multilayer Perceptrons. Such designs enable our framework to effectively represent and integrate diverse input sources of 3D point cloud, skeletal poses, and global translation and contacts.

For training and evaluation of our framework, we collected a precisely synchronized dataset comprising LiDAR point cloud and optical motion capture data. The dataset involved 10 subjects with varied body shapes and motion styles, engaging in a wide range of action categories of in-place movements and locomotion.

To validate the effectiveness of our method, we conducted comprehensive testing on unseen subjects, performing not only daily activities but also challenging motions such as lunging, sitting on the floor, and squatting. Furthermore, we highlight the practicality of our framework by showcasing its real-time application leveraging a single LiDAR sensor, implemented with a commercial game engine.

In summary, this paper presents the following main contributions:

- The real-time full-body motion capture framework based on a single LiDAR, incorporating global translation tracking.
- A novel design for feature encoding and decoding from different input sources, utilizing an autoregressive conditional variational autoencoder (CVAE) architecture to generate full-body poses from 3D point cloud data.
- A high-quality dataset featuring diverse subjects, containing synchronized LiDAR point cloud and optical motion capture data for a wide range of actions.

2. Related Work

2.1. Motion Capture

High-quality motion capture techniques using optical markers [opt09, vic10, VAV*07] and inertial measurement units (IMUs) [xse11] have emerged as leading solutions in the industry, offering precise and reliable data for human motion analysis and character animation. For end-users, Vive trackers [viv11] offers a cost-effective solution. However, current technologies require a large number of markers or sensors on the body and a time-consuming setup process. Therefore, researchers explored alternatives with a sparse setup of IMUs [VMRBPM17, YZX21, JYG*22, YZH*22] and trackers [ACB*22, JSQ*22, WWY22]. Despite their promising results, these methods still have limitations in the tracking accuracy and coverage of motion categories.

Markerless motion capture techniques have been extensively explored [BM98, ATS*08, HTTM12] to enhance the accessibility of motion capture technology, by reducing the cost and improving usability. While multi-view camera algorithms [AARS13, BSC13, DFJ*22] have achieved higher accuracy, they often require laborious camera system calibration. Mono-camera approaches with optimization techniques [BKL*16, KPD19] and neural networks [PZDD17, WLLL22, HPY*22] lack depth information and struggle to track global translations. Despite offering an additional depth channel, RGBD-based solutions [BMB*11, MSS*17, YZ21] are hindered by limited camera resolution and a field of view (FOV), which makes them impractical for product-level applications.

2.2. LiDAR-based 3D Human Pose Estimation

Recent advancements in 3D human pose estimation have seen the emergence of image-based methods like VIBE [KAB20] and Mo-

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

tionBERT [ZML*22]. These methods follow a two-stage process: extracting 2D keypoints and fitting the SMPL model [LMR*15] to estimate 3D keypoints. However, relying on 2D fitting poses limitations that compromise the accuracy of 3D keypoints. Additionally, the absence of depth information presents challenges for accurate global tracking.

To tackle these challenges, researchers have explored the integration of LiDAR sensors in 3D human pose estimation. LiDARs offer precise depth measurements, making them well-suited for large-scale environmental measurements for autonomous driving scenes [LVC*19, SGJ*20]. Recent studies have delved into utilizing LiDAR for capturing detailed 3D human poses [RZH*23]. Moreover, sensor fusion approaches combining LiDAR and cameras have been proposed [CXR*22, RZH*23] to leverage the complementary strengths of these sensors. However, these methods primarily focus on scene-level tasks like human detection and segmentation, rather than capturing skeletal motions with precise global translation.

2.3. Neural Generative Models for Motion Synthesis

Motion synthesis has been a prominent area of research, aimed at generating high-quality motion with minimal effort. Initial studies utilized probabilistic methods including principal component analysis (PCA) [SHP04, CH05, LZWM06], Gaussian mixture models (GMMs) [MC12], and Gaussian processes [GMHP04, WFH08, LWH*12].

Generative neural networks have recently gained substantial attention due to their impressive results in character motion synthesis. Multiple methods adopted Generative Adversarial Networks (GANs [GPAM*14]) for speech-to-gesture synthesis [FNM19], motion control [WCX21], and generation from a single motion clip [LAZ*22]. Variational Autoencoder (VAE) is another commonly used architecture that enables random sampling from a specified distribution. Furthermore, conditional VAE(CVAE) [SLY15] based methods use constraints such as motion history [LZCVDP20a], motion categories [PBV21], and speech [LYL*19, LYC*20] for generation. Henter et al. [HAB20] proposed utilizing normalizing flow for motion generation, enabling efficient training with exact maximum likelihood. Aliakbarian et al. [ACB*22] extended the work with an additional latent region approximator model. Inspired by the recent accomplishments of diffusion models in computer vision research, Tevet et al. [TRG*23] and Zhang et al. [ZCP*22] proposed language-driven motion synthesis techniques, while Tseng et al. [TCL23] focused on synthesizing dance motion from music.

3. MOVIN Dataset

While most publicly available motion datasets, such as Human3.6M [IPOS13] and 3DPW [VMHB*18], are primarily designed for 3D pose estimation from 2D images, PROX [VMHB*18] and LH26M [LZW*22] provide depth data from RGBD cameras and LiDAR sensors, respectively. However, the depth data in PROX are relatively noisy for sensor limitations and the point cloud in LH26M tends to be sparse for being captured from a far distance. In this work, we provide the MOVIN dataset with synchronized pairs of 3D point cloud and motion data, designed for full-body motion capture from 3D point cloud data.

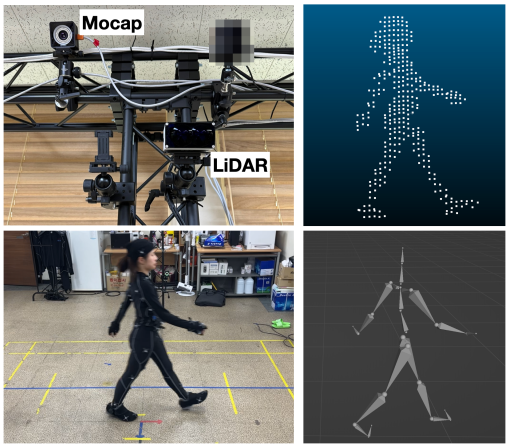


Figure 2: *The integrated optical motion capture and LiDAR system, with representations of resulting point cloud data and skeletal motion capture.*

	Train set		Test set	
# Subjects	8		2	
Motion type	Static	Locomotion	Static	Locomotion
# Frames	56,535	75,134	12,472	17,038
Elapsed time$_{min}$	47	62	10	14

Static	Locomotion
T-pose, A-pose, Idle, Hands on waist	Walking
Elbows bent up, down	Jogging
Bow, Look, Roll head	Running
Windmill arms, Touch toes	Crouching
Twist torso, Hula hoop, Lean	Transitions
Lunge, Squat, Jumping Jack	Moving backward
Kick, Turn	Jumping
Walk / Run in place	Sitting on the floor

Table 1: *Dataset composition details and motion categories.*

Motion capture System. We employed the OptiTrack system [opt09] comprising 21 PRIME 13 high-speed infrared cameras to capture human motion. By tracking passive reflective markers positioned on the subject's body keypoints, the system accurately records joint positions and orientations. Using an optical-based motion capture system offers the advantage of avoiding global location errors that may arise during extended recording sessions, which is a common issue with IMU-based motion capture systems. Given that our Movin dataset requires prolonged collections of motion data, an optical-based motion capture system is highly suitable for this purpose.

LiDAR sensor. LiDAR sensors emit laser light to accurately measure distances and generate high-resolution 3D maps of the surrounding environment and objects inside. The resulting 3D point cloud data provides detailed information about objects' geometries.

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

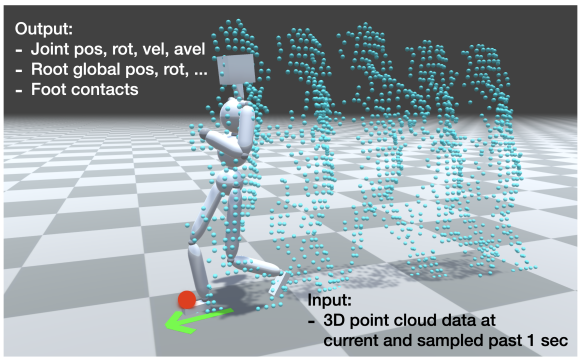


Figure 3: *Illustration of the input and output in inference time. The gray skeleton represents the output joint positions, rotations, and velocities at the current time frame. Blue particles represent the 3D point cloud sampled in the current and the past 1-second time window. The green arrow and red sphere on the ground denote the global translation of the root and foot contacts, respectively.*

To capture 3D point cloud data of moving subjects, we utilized the ML-X model [SOS23] by SOSLAB, a high-performance solid-state LiDAR operating at a frequency of 20 Hz. The ML-X model offers a wide field of view with a 120 degree horizontal and 35 degree vertical coverage, generating a detailed point cloud at a resolution of 56×192. This point cloud provides precise 3D global coordinates along with light intensity values. One notable advantage of this solid-state LiDAR system is capturing depth and intensity in both 3D and 2D image formats; this allows for treating the point cloud data similar to images in image-motion datasets. Moreover, the ML-X model exhibits minimal distortion compared to other spinning-type LiDAR systems, making it well-suited for capturing fast and complex movements with body overlaps.

Data acquisition. Motion and point cloud data are captured simultaneously, as shown in Figure 2. Subjects performed two main types of static movements and locomotion. Details on the dataset composition and motion categories are provided in Table 1.

To extract only the points belonging to human subjects, we applied background filtering to the captured point cloud. The refined point cloud contains approximately 200 to 300 points per frame. We recorded only the 3D position data and excluded the intensity data, as it is not significant among subjects who wear identical black motion capture suits. Furthermore, such intensity data does not provide information about typical human clothing.

After aligning the captured motion and point cloud data to a shared global coordinate frame, we synchronized the time frames and downsampled the motion data to 20Hz, which matches the operating frequency of the LiDAR sensor.

4. Input/Output representation

Figure 3 illustrates the input and output of our framework for a current frame t in inference time. The input consists of the current 3D point cloud data as well as subsampled past 1-second point cloud data (at a frequency of 20 Hz). Each frame contains 256 3D points. We include four past point clouds sampled from 5, 10, 15, and 20 frames prior to the current frame. To ensure consistent input dimensions, we randomly discard points beyond 256 and perform zero-position padding for frames with fewer than 256 points. This results in the input $\mathbf{p}_t = [p_t, p_{t-5}, p_{t-10}, p_{t-15}, p_{t-20}]$, where $p_t \in \mathbb{R}^{256\times 3}$. Considering historical point cloud significantly improves the quality of results (Sec. 7.2).

The output for a current frame t consists of a global pose feature $\mathbf{g}_t$ and a local pose feature $\mathbf{x}_t$. The global pose feature includes the character's root position, rotation, velocity, angular velocity, and foot contacts, represented as $\mathbf{g}_t = [r^l, r^r, \dot{r}^l, \dot{r}^r, c] \in \mathbb{R}^{17}$, where $r^l \in \mathbb{R}^3$, $r^r \in \mathbb{R}^6$, $\dot{r}^l \in \mathbb{R}^3$, $\dot{r}^r \in \mathbb{R}^3$, and $c \in \mathbb{R}^2$. The local pose feature consists of joint local positions, rotations, velocity, and angular velocity relative to the parent joint, denoted as $\mathbf{x}_t = [x^l, x^r, \dot{x}^l, \dot{x}^r] \in \mathbb{R}^{n_j\times 15}$, where $x^l \in \mathbb{R}^{n_j\times 3}$, $x^r \in \mathbb{R}^{n_j\times 6}$, $\dot{x}^l \in \mathbb{R}^{n_j\times 3}$, and $\dot{x}^r \in \mathbb{R}^{n_j\times 3}$. Here, n_j represents the number of joints.

5. MOVIN Framework

The MOVIN framework, illustrated in Fig. 4, is based on an autoregressive conditional variational autoencoder (CVAE) architecture. During the training phase, MOVIN reconstructs the current global and local pose features, denoted as $\mathbf{g}_t$ and $\mathbf{x}_t$, respectively, using the 3D point cloud history $\mathbf{p}_t$, as well as the previous global and local pose features $\mathbf{g}_{t-1}$ and $\mathbf{x}_{t-1}$. In addition, the model is trained to shape the latent variable z into a Gaussian distribution. The framework comprises two components: the feature encoder, responsible for mapping input features to the latent distribution, and the pose generator, which generates global and local pose features.

During the inference phase, the embedding modules of the feature encoder and pose generator are used exclusively to generate the output pose features $\hat{\mathbf{x}}_t$ and $\hat{\mathbf{g}}_t$ for the current time step. The input consists of a randomly sampled latent variable z, while the conditions include the history of 3D point cloud $\mathbf{p}_t$ and the output pose features from the previous frame, $\hat{\mathbf{x}}_{t-1}$ and $\hat{\mathbf{g}}_{t-1}$.

5.1. Feature Encoder

The feature encoder takes the previous output pose features $\mathbf{g}_{t-1}$, $\mathbf{x}_{t-1}$, the history of 3D point cloud data $\mathbf{p}_t$, and the current target pose features $\mathbf{g}_t$, $\mathbf{x}_t$ as inputs and encodes them to a latent variable z in a Gaussian distribution $\mathcal{N}(\mu, \sigma)$. The feature encoder is composed of embedding modules that individually embed the input features, and a transformer encoder that captures the relationships between the embedded vectors, allowing the model to learn a distribution of possible pose features for the current time frame.

Embedding modules. To capture contextual information at different scales in the input 3D point cloud data, we utilize PointNet++ [QYSG17] for extracting the embedded vector $f_t^{\mathbf{p}}$ from the history of point cloud $\mathbf{p}_t$:

$$f_t^{\mathbf{p}} = \text{PointNet++}(\mathbf{p}_t) \in \mathbb{R}^{5\times C}, \tag{1}$$

where C denotes the number of channels.

For pose features, we employ separate embeddings for global

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

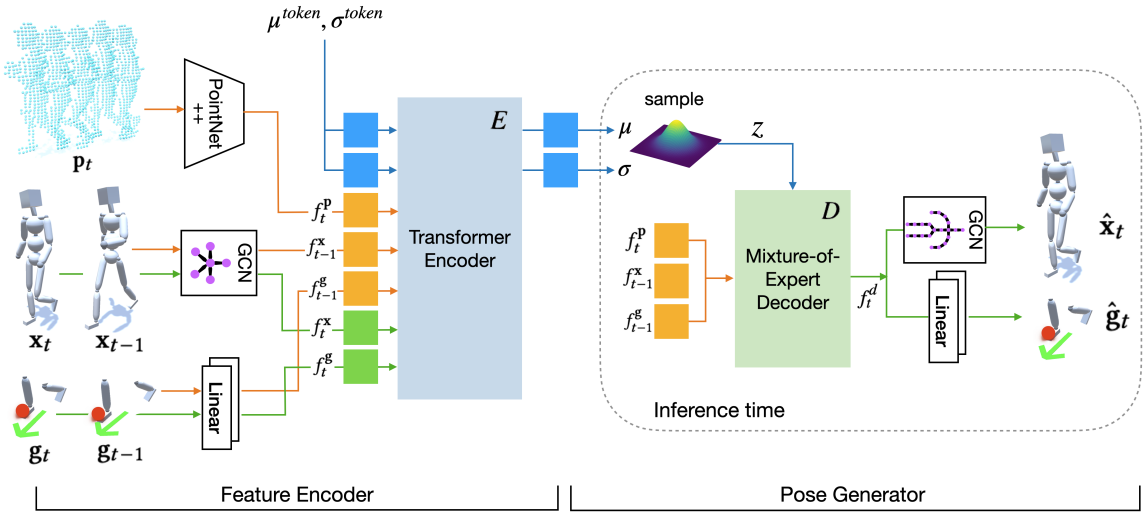


Figure 4: *Overview of MOVIN framework. The model separates into the Feature Encoder and the Pose Generator. At inference time, only the Pose Generator and the embedding modules of Feature Encoder are used. Given the sampled point cloud sequence* $\mathbf{p}_t$*, our model generates current global and local pose features* $\hat{\mathbf{g}}_t$*,* $\hat{\mathbf{x}}_t$*, which is used as a condition the next time frame.*

and local pose features (as mentioned in Sec.4). The Graph Convolution Network (GCN) [YLX*19, LCC*19, SZCL19, JPL22] is used to reduce the spatial resolution of the input local features $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ while preserving the body part structure. Additionally, a two-layer MLP is utilized to embed the global pose features $\mathbf{g}_t$ and $\mathbf{g}_{t-1}$. The pose feature embedding process can be defined as follows:

$$f_t^{\mathbf{x}} = \mathrm{GCN}(\mathbf{x}_t) \in \mathbb{R}^C, \quad f_t^{\mathbf{g}} = \mathrm{MLP}(\mathbf{g}_t) \in \mathbb{R}^C \tag{2}$$

By applying the same procedure to $\mathbf{x}_{t-1}$ and $\mathbf{g}_{t-1}$, we obtain $f_{t-1}^{\mathbf{x}}$ and $f_{t-1}^{\mathbf{g}}$.

After feature embedding, we have five embedded vectors: $f_t^{\mathbf{p}}$, $f_t^{\mathbf{x}}$, $f_{t-1}^{\mathbf{x}}$, $f_t^{\mathbf{g}}$, and $f_{t-1}^{\mathbf{g}}$. These vectors serve as input for the subsequent transformer encoder. Moreover, $[f_t^{\mathbf{p}},\ f_{t-1}^{\mathbf{x}},\ f_{t-1}^{\mathbf{g}}]$ are used as conditions for the Pose Generator.

Transformer encoder. To model the correlation between human joints and local clusters in the corresponding point cloud, we utilize a transformer architecture [VSP*17]. The transformer encoder E takes learnable tokens $[\mu^{token}, \sigma^{token}]$ and concatenated embedded vectors $[f_t^{\mathbf{p}},\ f_{t-1}^{\mathbf{x}},\ f_{t-1}^{\mathbf{g}},\ f_t^{\mathbf{x}},\ f_t^{\mathbf{g}}]$ as inputs. These inputs are encoded to obtain the parameters of a Gaussian distribution $\mathcal{N}(\mu, \sigma)$. The reparameterization trick is then applied to transform these parameters and obtain the decoder input distribution $z \in \mathbb{R}^C$:

$$E(z|f_t^{\mathbf{p}},\ f_{t-1}^{\mathbf{x}},\ f_{t-1}^{\mathbf{g}},\ f_t^{\mathbf{x}},\ f_t^{\mathbf{g}}) = \mathcal{N}(z;, \mu, \sigma) \tag{3}$$

5.2. Pose Generator

Given the sampled latent variable z, the pose generator is an autoregressive model that generates current global and local pose features, $[\hat{\mathbf{x}}_t,\ \hat{\mathbf{g}}_t]$, conditioned on the embedded vectors of the sampled point cloud history and the previous pose features $[f_t^{\mathbf{p}},\ f_{t-1}^{\mathbf{x}},\ f_{t-1}^{\mathbf{g}}]$. Since a single LiDAR sensor often suffers from self-occlusions between body parts, it increases ambiguity between the obtained point cloud and the ground truth full-body pose. To address this, we sample a latent vector z from a prior distribution and use the point cloud as the condition to generate plausible body motion.

Inspired by motionVAE [LZCVDP20b], we incorporate a Mixture-of-Expert (MoE) decoder, which we have observed empirically to enhance pose construction and reduce visual artifacts. MoE methods are often used to divide the problem space into distinct partitions assigned to a fixed number of neural network experts. A gating network is then employed to determine the relative contribution of each expert's prediction when computing the final output or prediction. In our framework, the MoE decoder generates an output, and this output is further expanded using expanding modules to obtain the final full-body pose and foot contacts. These expanding modules use inverse forms of the embedding modules found in the feature encoder.

Mixture-of-Expert decoder The MoE decoder D consists of eight expert networks with identical structures. A single shared gating network is incorporated to blend the weights of the experts, thereby determining the weights of the decoder network. Given the latent variable z and the set of condition features $[f_t^{\mathbf{p}},\ f_{t-1}^{\mathbf{x}},\ f_{t-1}^{\mathbf{g}}]$, the MoE decoder D computes the output f_t^d as follows:

$$f_t^d = D(z,\ f_t^{\mathbf{p}},\ f_{t-1}^{\mathbf{x}},\ f_{t-1}^{\mathbf{g}}) \in \mathbb{R}^{2C}, \tag{4}$$

where $f_t^{\mathbf{p}}$ represents the embedded vector of the current point cloud, and $f_{t-1}^{\mathbf{x}}$, $f_{t-1}^{\mathbf{g}}$ represent the embedded vectors of the previous pose features.

Expanding modules. The output of the MoE decoder f_t^d is further processed by De-GCN and De-MLP modules, which have architectures symmetric to the embedding modules. These modules expand the dimensions of the output to obtain the final global and local

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

pose features $\hat{\mathbf{g}}_t$ and $\hat{\mathbf{x}}_t$ as follows:

$$\hat{\mathbf{x}}_t = \text{De-GCN}(f_t^d[:C]), \quad \hat{\mathbf{g}}_t = \text{De-MLP}(f_t^d[C:]) \tag{5}$$

6. Training

The overall model is trained by minimizing the reconstruction $\mathcal{L}_{rec}$ and KL-divergence $\mathcal{L}_{kl}$ losses. The reconstruction loss comprises both the local and global pose feature reconstruction losses. The local reconstruction loss quantifies the L1 errors in joint space with respect to the parent and character space with respect to the character's root. Similarly, the global reconstruction loss measures the L1 errors between the generated and ground truth global root position, rotation, velocity, and foot contacts. In addition, the KL-divergence loss $\mathcal{L}_{kl}$ regularize distribution of latent variable z to be near the prior distribution $\mathcal{N}(0,I)$.

The total loss function is thus:

$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{rec} + w_{kl}\mathcal{L}_{kl} \\ \mathcal{L}_{rec} &= \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_1 + \|FK(\hat{\mathbf{x}}_t) - FK(\mathbf{x}_t)\|_1 + \|\hat{\mathbf{g}}_t - \mathbf{g}_t\|_1,\end{aligned} \tag{6}$$

where the first and second terms of $\mathcal{L}_{rec}$ denote local reconstruction loss, and the last term is for global reconstruction loss. w_{kl} is weight of KL-divergence loss.

Implementation details. The AdamW optimizer was used over 120 epochs, with a learning rate of 10^{-4}. Loss weight w_{kl} was set as 1. In the embedding module, the GCN layer comprises 1 spatial convolution layer along with body part pooling. The 2-layer MLP comprised a feed-forward network with 256 hidden units and ReLU activation. Meanwhile, PointNet++ is made up of 3 set abstraction layers. Transformer encoder E comprised 2 layers of 64 channels with 4 heads, and the MoE decoder D consists of 8 identically structured expert networks and a single gating network. The gating network is also a 3-layer feed-forward network with 256 hidden units. Expanding module have architectures symmetric to the embedding modules. To prevent the covariate shift during autoregressive inference, we set the prediction length as 8 frames for training. Scheduled sampling was also utilized in our model to enable long-term generation by making the model robust to its own errors. With four 12GB 2080ti GPUs, training took around 60 hours in an end-to-end manner.

7. Evaluation and Experiments

To validate the effectiveness of our method, we conducted comprehensive quantitative and qualitative evaluations against state-of-the-art methods. We selected VIBE [KAB20] and MotionBERT [ZML*22] as representative baselines, which are vision-based approaches. To match the skeleton hierarchy for comparison, we applied BVH conversion and an optimization-based retargeting [JKL18] to the output parameters of the baseline methods. Furthermore, we downsampled the retargeted outputs to 20 fps to align with our output framerate. We disabled any postprocessing for all methods to ensure a fair comparison of the network architectures. For visual animation results, please refer to the supplementary video.

	MJPE$_{cm}$	MJRE°	MJLVE$_{cm}$	MJAVE°	Jitt.
GT	–	–	–	–	446.87
VIBE	10.86	18.39	2.39	3.16	1103.15
MotionBERT	10.62	18.05	**1.75**	**2.24**	**395.11**
MOVIN	**6.21**	**10.12**	1.89	2.75	871.53
	MPPE$_{cm}$	MPRE°	MPLVE$_{cm}$	MPAVE°	Cont.$_{\%}$
MOVIN	4.42	11.64	2.46	4.94	94.28

Table 2: *Quantitative measures of MOVIN and state-of-the-art methods. Pelvis (P) errors are only measured for MOVIN since the baselines cannot accurately capture global translation.*

Due to the unavailability of public datasets containing synchronized video, LiDAR point cloud, and motion capture data, our experiments were conducted solely on our held-out test set. The test set consists of two subjects with heights of 162 cm and 170 cm and each subject performed motion categories of static movement and locomotion. The length of the entire test set is around 25 minutes.

Additionally, we performed ablation studies to investigate the impact of our design choices, including the utilization of point cloud history and the implementation of the autoregressive scheme.

The quantitative metrics include mean position error (M*PE), rotation error (M*RE), linear velocity error (M*LVE), and angular velocity error (M*AVE), for the Pelvis (P) and other body Joints (J). Joint position and linear velocity errors are calculated using forward kinematics in the pelvis coordinate frame. In addition for MOVIN, we assessed contact accuracy by comparing ground truth and predicted contact labels obtained by applying a threshold of 0.5 to the predicted contact probabilities.

Lastly, we showcase a real-time motion capture demo on the wild unseen subject and discuss about effect of post-processing.

7.1. Comparison with State-of-the-art methods

We conducted inference for VIBE and MotionBERT using their public code and note that these baselines are offline methods with fixed input sequence lengths of 16 and 243, respectively. In contrast, our model, MOVIN, performed per-frame prediction with a sliding window size of 1 to simulate real-time usage. We specifically measured pelvis errors for MOVIN, the errors significantly impact the quality of the output full-body pose. Since explicit global localization is not supported by the baselines, we did not measure pelvis errors for them and provided them with ground truth pelvis trajectory for qualitative analysis.

Table 2 presents the quantitative evaluation results of MOVIN and the state-of-the-art methods. MOVIN demonstrated a significant improvement over MotionBERT in terms of average joint position and rotation errors, with margins of approximately 4.41 cm and 7.93 degrees, respectively. However, for joint linear and angular velocities, MotionBERT exhibited slightly better performance. This advantage can be attributed to MotionBERT's utilization of a longer input window of 243 frames, enabling it to maintain continuity and achieve smoother transitions in the output. Notably, the

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

	MJPE$_{cm}$	MJRE°	MJLVE$_{cm}$	MJAVE°	Jitt.
GT	–	–	–	–	446.87
w/o past pcd	6.70	11.49	1.95	2.76	919.15
w/ past poses	7.25	12.67	**1.80**	2.44	**708.08**
w/o autoreg.	**6.09**	**9.71**	2.07	3.15	1118.68
512 points	6.39	10.15	1.91	**2.74**	929.60
MOVIN	6.21	10.12	1.89	2.75	871.53
	MPPE$_{cm}$	MPRE°	MPLVE$_{cm}$	MPAVE°	Cont.%
w/o past pcd	4.98	12.03	1.67	6.76	92.79
w/ past poses	5.44	12.34	1.58	5.68	93.93
w/o autoreg.	4.45	**11.39**	**1.48**	5.83	**94.43**
512 points	**4.35**	11.83	1.51	5.90	94.38
MOVIN	4.42	11.64	1.50	**4.94**	94.28

Table 3: *Quantitative measures of MOVIN and ablation models. The term "w/o past pcd" denotes the variant trained without incorporating point clouds sampled from a previous time window. "w/ past poses" refers to the version that includes past poses. "w/o autoreg." signifies the variant that employs a non-autoregressive pose generator. Lastly, "512 points" designates the model variant that utilizes a point cloud consisting of 512 points.*

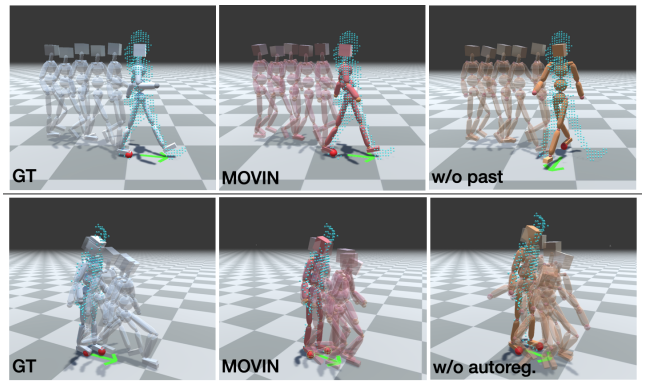


Figure 5: *Visual comparisons of ablation models. Without past point clouds as input, the model exhibits abrupt changes in the global heading direction and incorrect movement. The model with a non-autoregressive Pose Generator produces outputs with unrelated poses between frames.*

output motions of MotionBERT displayed lower jitter values compared to the ground truth, indicating an over-smoothing effect that can be clearly observed in the supplementary video. VIBE showed similar position and rotation errors to MotionBERT but suffered from severe jittering in the output poses, as indicated by the large jitter value in Table 2.

Figure 8 presents two sets (170 cm male and 162 cm female) of four-column images representing the ground truth (GT), and output full-body motions from MOVIN, VIBE, and MotionBERT, respectively. The robustness of MOVIN is evident as it generates plausible full-body motion, regardless of the subject body shape and across diverse action categories. In general, MOVIN preserved details in the ground truth and maintained temporal continuity in the output full-body motion. Please refer to our supplementary video for a comprehensive evaluation comparing MOVIN with the baselines.

Regarding the global localization performance, MOVIN exhibited average pelvis position and rotational errors of 4.42 cm and 11.64 degrees, respectively. The snapshots from our real-time application, depicted in Figure 6, reveal that the output global trajectory is well aligned with that of the user.

7.2. Ablation study

The aim of our ablation study is to validate our design choices of utilizing point cloud history as input, excluding past poses from input, selecting an optimal number of input points, and implementing autoregression in both training and inference phases. Table 3 presents quantitative metrics for five different ablation models: one without past point cloud input, one with past poses as input, one with 512 points as input, one without autoregression, and our proposed model, MOVIN.

Past point cloud sequence and poses. The model without a past point cloud sequence underperformed compared to the proposed model. it showed an increase of 0.5 cm in average joint position error and 1.4 degrees in rotation error. Specifically for the pelvis joint, the average position error and angular velocity error increased by 0.5 cm and 1.8 degrees, respectively. In the output motion sequence, we observed that this model often fails to maintain a temporal continuity, especially for rapid movements or cases when certain body parts are occluded by others (i.e. walking sideways or sitting) as shown in Figure 5 (1st row); this results in abrupt changes in the global heading direction and body poses.

Providing past poses performed the worst among the methods. We hypothesize that simply providing previously generated poses leads to the model suffering from accumulated errors in the autoregressive input, thereby making it challenging to recover from inaccurate predictions.

Autoregression. The model without autoregression shows no significant differences compared to the proposed model in terms of position and rotation errors. However, there are noticeable increases in linear and angular velocity errors for the joints, particularly for the pelvis where the angular velocity error rises by approximately 0.9 degrees. Additionally, the jitter value increases by around 250. These findings suggest that incorporating autoregression and exposing the model to accumulated prediction errors during training, enables it to robustly handle such errors during inference and produce stable and continuous poses in the output sequence. The result in Figure 5 (2nd row) shows that the ablated model produces the discontinuous poses during the Lunge.

Number of the input points. Doubling the number of input points (from 256 to 512) increases computation time proportionally but does not bring a significant improvement in performance. To achieve real-time inference at the pace of a 20Hz LiDAR sensor, we opted for 256 points as input.

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

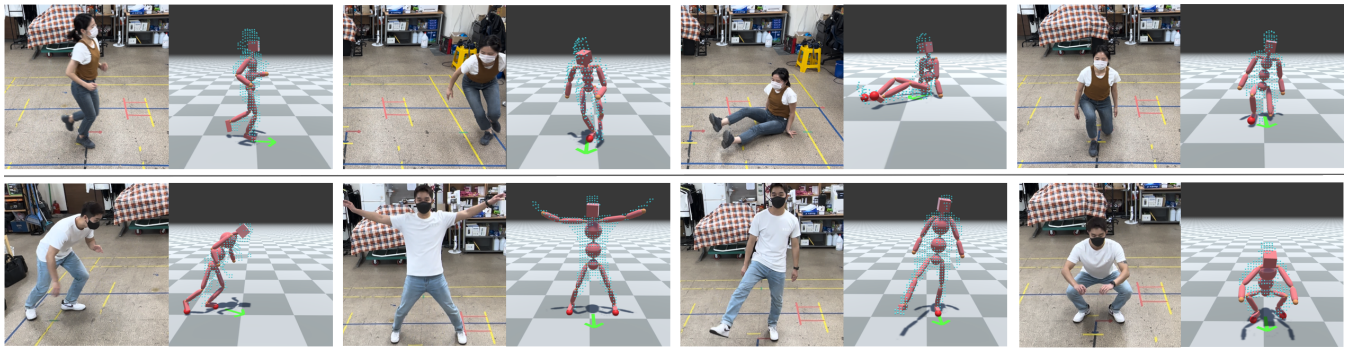


Figure 6: *Real-time motion capture results. For each user's pose, the left shows reference images and the right shows the corresponding captured full-body pose. Our framework robustly captures both body dimensions and a wide range of motions.*

7.3. Real-time Motion Capture Demo

Figure 6 showcases snapshots from our real-time motion capture system with a single LiDAR, implemented in Unity3D. Compared to recent methods that use multiple RGB cameras, our system does not require offline calibration and captures the subject's motion in real-time, allowing users to check the results immediately. In the first row, a female subject with a height of 159cm performs challenging motions such as sitting on the floor and lunging, which are accurately captured by our system. The second row highlights the application's ability to capture dynamic actions from a male subject with a height of 175cm, including running, jumping jacks, kicking, and squats. Our framework not only accurately tracks these diverse movements but also effectively captures the subjects' body dimensions. Please refer to our supplementary video for a detailed demonstration of our model's real-time performance.

7.4. Effect of Post-processing

As the real-time pose generation method cannot consider future poses, the output motion may exhibit foot sliding. To address this issue, we employ predicted contact labels and utilize inverse kinematics to correct foot positions. The target foot position is determined by interpolating between the previous and the output foot positions. Our supplemental video demonstrates the impact of this post-processing by comparing output motions with and without it.

8. Limitations and Future Work

While our proposed model, MOVIN, successfully captures diverse motions in real-time, it is important to acknowledge the existing limitations for future research.

One limitation of our model arises when encountering unseen motions, as demonstrated by the windmill motion example shown in Figure 7. In these cases, the generated pose does not align with the corresponding point cloud, highlighting the need for improved generalization to novel or uncommon movements. Expanding the size of the training dataset by incorporating a wider range of motion variations is a potential approach to address this limitation.

Another is the relatively low frames-per-second (fps) performance of our current implementation, primarily due to the operating frequency of the LiDAR sensor. To overcome this, future work could explore techniques such as point cloud upsampling or hardware improvements to enhance the fps rate. Upsampling the point cloud data can provide denser and more frequent input information, resulting in smoother output motion and a higher frame rate.

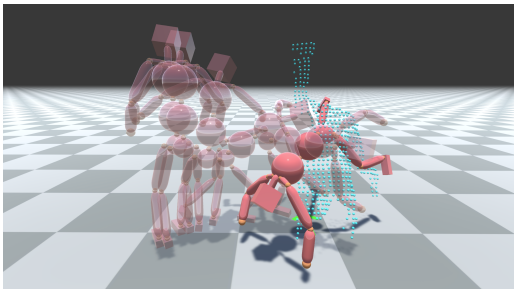


Figure 7: *An example of failure case. When the user performs an unseen motion (windmill), the model generates a pose that does not align with the corresponding point cloud.*

Additionally, while our model demonstrates reasonable handling of self-occlusions between body parts, it struggles in cases of severe occlusions, such as when a subject curls up or the environment is cluttered with objects. To improve performance in such scenarios, considering the use of multiple LiDAR sensors positioned from different angles could be a viable solution. By capturing aligned point clouds from multiple perspectives, the model can access more detailed information and enhance its capture performance.

9. Conclusion

We present MOVIN, the first data-driven generative model for real-time full-body motion capture using a single LiDAR sensor. Our approach addresses the challenges of full-body tracking by eliminating the need for body-worn suits and devices while maintaining high-quality motion capture. MOVIN utilizes an autoregressive CVAE model to learn the distribution of pose variations from 3D point cloud data. By separately embedding global and local pose features, our model effectively learns the pose prior and accurately predicts the performer's 3D global information and local joint de-

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

tails. The proposed autoregressive Mixture-of-Expert decoder ensures temporal coherence across frames, resulting in natural and realistic motion. Our real-time application showcases MOVIN's robustness to accurately capture diverse motions from subjects with varying body shapes, demonstrating its effectiveness in real-world scenarios.

Acknowledgement

This work was supported by IITP, MSIT, Korea (2022-0-00566) and NRF, Korea (2022R1A4A5033689).

References

[AARS13] AMIN S., ANDRILUKA M., ROHRBACH M., SCHIELE B.: Multi-view pictorial structures for 3d human pose estimation. In *British Machine Vision Conference* (2013). 2

[ACB*22] ALIAKBARIAN S., CAMERON P., BOGO F., FITZGIBBON A., CASHMAN T. J.: Flag: Flow-based 3d avatar generation from sparse observationszhang22motiondiffuse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 13253–13262. 2, 3

[ATS*08] AGUIAR E., THEOBALT C., STOLL C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. *ACM Transactions on Graphics* (02 2008). doi:10.1145/1360612.1360697. 2

[BKL*16] BOGO F., KANAZAWA A., LASSNER C., GEHLER P., ROMERO J., BLACK M. J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image, 2016. arXiv:1607.08128. 2

[BM98] BREGLER C., MALIK J.: Tracking people with twists and exponential maps. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)* (1998), pp. 8–15. doi:10.1109/CVPR.1998.698581. 2

[BMB*11] BAAK A., MÜLLER M., BHARAJ G., SEIDEL H.-P., THEOBALT C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In *2011 International Conference on Computer Vision* (2011), pp. 1092–1099. doi:10.1109/ICCV.2011.6126356. 2

[BSC13] BURENIUS M., SULLIVAN J., CARLSSON S.: 3d pictorial structures for multiple view articulated pose estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3618–3625. doi:10.1109/CVPR.2013.464. 2

[CH05] CHAI J., HODGINS J. K.: Performance animation from low-dimensional control signals. In *ACM SIGGRAPH 2005 Papers*. 2005, pp. 686–696. 3

[CXR*22] CONG P., XU Y., REN Y., ZHANG J., XU L., WANG J., YU J., MA Y.: Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar, 2022. arXiv:2211.16951. 3

[DFJ*22] DONG J., FANG Q., JIANG W., YANG Y., HUANG Q., BAO H., ZHOU X.: Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 10 (2022), 6981–6992. doi:10.1109/TPAMI.2021.3098052. 2

[FNM19] FERSTL Y., NEFF M., MCDONNELL R.: Multi-objective adversarial gesture generation. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (New York, NY, USA, 2019), MIG '19, Association for Computing Machinery. 3

[GMHP04] GROCHOW K., MARTIN S. L., HERTZMANN A., POPOVIĆ Z.: Style-based inverse kinematics. In *ACM SIGGRAPH 2004 Papers*. 2004, pp. 522–531. 3

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), Ghahramani Z., Welling M., Cortes C., Lawrence N., Weinberger K., (Eds.), vol. 27, Curran Associates, Inc. 3

[HAB20] HENTER G. E., ALEXANDERSON S., BESKOW J.: Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG) 39*, 6 (2020), 1–14. 3

[HKA*18] HUANG Y., KAUFMANN M., AKSAN E., BLACK M. J., HILLIGES O., PONS-MOLL G.: Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG) 37*, 6 (2018), 1–15. 2

[HPY*22] HUANG B., PAN L., YANG Y., JU J., WANG Y.: Neural mocon: Neural motion control for physically plausible human motion capture, 2022. arXiv:2203.14065. 2

[HTTM12] HOLTE M. B., TRAN C., TRIVEDI M. M., MOESLUND T. B.: Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of Selected Topics in Signal Processing 6*, 5 (2012), 538–552. doi:10.1109/JSTSP.2012.2196975. 2

[IPOS13] IONESCU C., PAPAVA D., OLARU V., SMINCHISESCU C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence 36*, 7 (2013), 1325–1339. 3

[JKL18] JIN T., KIM M., LEE S.-H.: Aura mesh: Motion retargeting to preserve the spatial relationships between skinned characters. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library, pp. 311–320. 6

[JPL22] JANG D.-K., PARK S., LEE S.-H.: Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG) 41*, 3 (2022), 1–16. 5

[JSQ*22] JIANG J., STRELI P., QIU H., FENDER A., LAICH L., SNAPE P., HOLZ C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V* (2022), Springer, pp. 443–460. 2

[JYG*22] JIANG Y., YE Y., GOPINATH D., WON J., WINKLER A. W., LIU C. K.: Transformer inertial poser: Attention-based real-time human motion reconstruction from sparse imus. *arXiv preprint arXiv:2203.15720* (2022). 2

[KAB20] KOCABAS M., ATHANASIOU N., BLACK M. J.: Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 5253–5263. 2, 6

[KPD19] KOLOTOUROS N., PAVLAKOS G., DANIILIDIS K.: Convolutional mesh regression for single-image human shape reconstruction, 2019. arXiv:1905.03244. 2

[LAZ*22] LI P., ABERMAN K., ZHANG Z., HANOCKA R., SORKINE-HORNUNG O.: Ganimator: Neural motion synthesis from a single sequence. *ACM Trans. Graph. 41*, 4 (jul 2022). 3

[LCC*19] LI M., CHEN S., CHEN X., ZHANG Y., WANG Y., TIAN Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 3595–3603. 5

[LLZ*22] LI Z., LIU J., ZHANG Z., XU S., YAN Y.: Cliff: Carrying location information in full frames into human pose and shape estimation, 2022. arXiv:2208.00571. 2

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG) 34*, 6 (2015), 1–16. 3

[LVC*19] LANG A. H., VORA S., CAESAR H., ZHOU L., YANG J., BEIJBOM O.: Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 12697–12705. 3

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[LWH*12] LEVINE S., WANG J. M., HARAUX A., POPOVIĆ Z., KOLTUN V.: Continuous character control with low-dimensional embeddings. *ACM Transactions on Graphics (TOG) 31*, 4 (2012), 1–10. 3

[LYC*20] LI J., YIN Y., CHU H., ZHOU Y., WANG T., FIDLER S., LI H.: Learning to generate diverse dance motions with transformer. 3

[LYL*19] LEE H.-Y., YANG X., LIU M.-Y., WANG T.-C., LU Y.-D., YANG M.-H., KAUTZ J.: Dancing to music. In *Advances in Neural Information Processing Systems* (2019), Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., (Eds.), vol. 32, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/7ca57a9f85a19a6e4b9a248c1daca185-Paper.pdf. 3

[LZCVDP20a] LING H. Y., ZINNO F., CHENG G., VAN DE PANNE M.: Character controllers using motion vaes. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 40–1. 3

[LZCVDP20b] LING H. Y., ZINNO F., CHENG G., VAN DE PANNE M.: Character controllers using motion vaes. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 40–1. 5

[LZW*22] LI J., ZHANG J., WANG Z., SHEN S., WEN C., MA Y., XU L., YU J., WANG C.: 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* (2022), pp. 20502–20512. 2, 3

[LZWM06] LIU G., ZHANG J., WANG W., MCMILLAN L.: Human motion estimation from a reduced marker set. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games* (2006), pp. 35–42. 3

[MC12] MIN J., CHAI J.: Motion graphs++: A compact generative model for semantic motion analysis and synthesis. *ACM Trans. Graph. 31*, 6 (nov 2012). 3

[MSS*17] MEHTA D., SRIDHAR S., SOTNYCHENKO O., RHODIN H., SHAFIEI M., SEIDEL H.-P., XU W., CASAS D., THEOBALT C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. vol. 36. URL: http://gvv.mpi-inf.mpg.de/projects/VNect/, doi:10.1145/3072959.3073596. 2

[opt09] Optitrack motion capture systems., 2009. URL: https://www.optitrack.com/. 2, 3

[PBV21] PETROVICH M., BLACK M. J., VAROL G.: Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10985–10995. 3

[PZDD17] PAVLAKOS G., ZHOU X., DERPANIS K. G., DANIILIDIS K.: Coarse-to-fine volumetric prediction for single-image 3d human pose, 2017. arXiv:1611.07828. 2

[QYSG17] QI C. R., YI L., SU H., GUIBAS L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems 30* (2017). 4

[RZH*23] REN Y., ZHAO C., HE Y., CONG P., LIANG H., YU J., XU L., MA Y.: Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE Transactions on Visualization and Computer Graphics 29*, 5 (2023), 2337–2347. 3

[SGJ*20] SHI S., GUO C., JIANG L., WANG Z., SHI J., WANG X., LI H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 10529–10538. 3

[SHP04] SAFONOVA A., HODGINS J. K., POLLARD N. S.: Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics (ToG) 23*, 3 (2004), 514–521. 3

[SLY15] SOHN K., LEE H., YAN X.: Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems 28* (2015). 3

[SOS23] SOSLAB: Ml-x model lidar, 2023. [Online; accessed 4-June-2023]. URL: https://www.soslab.co. 4

[SZCL19] SHI L., ZHANG Y., CHENG J., LU H.: Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 7912–7921. 5

[TCL23] TSENG J., CASTELLON R., LIU C. K.: Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023). 3

[TRG*23] TEVET G., RAAB S., GORDON B., SHAFIR Y., BERMANO A. H., COHEN-OR D.: Human motion diffusion model. In *ICLR* (2023). 3

[VAV*07] VLASIC D., ADELSBERGER R., VANNUCCI G., BARNWELL J., GROSS M., MATUSIK W., POPOVIC J.: Practical motion capture in everyday surroundings. *ACM Trans. Graph. 26* (07 2007), 35. doi:10.1145/1276377.1276421. 2

[vic10] Vicon motion capture systems., 2010. URL: https://www.vicon.com/. 2

[viv11] Xsens technologies b.v., 2011. URL: https://www.vive.com/us/accessory/tracker3/. 2

[VMHB*18] VON MARCARD T., HENSCHEL R., BLACK M. J., ROSENHAHN B., PONS-MOLL G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 601–617. 3

[VMRBPM17] VON MARCARD T., ROSENHAHN B., BLACK M. J., PONS-MOLL G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum* (2017), vol. 36, Wiley Online Library, pp. 349–360. 2

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. *Advances in neural information processing systems 30* (2017). 5

[WCX21] WANG Z., CHAI J., XIA S.: Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE Transactions on Visualization and Computer Graphics 27*, 1 (jan 2021), 14–28. 3

[WFH08] WANG J. M., FLEET D. J., HERTZMANN A.: Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 2 (2008), 283–298. doi:10.1109/TPAMI.2007.1167. 3

[WLLL22] WEI W.-L., LIN J.-C., LIU T.-L., LIAO H.-Y. M.: Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video, 2022. arXiv:2203.08534. 2

[WWY22] WINKLER A., WON J., YE Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers* (2022), pp. 1–8. 2

[xse11] Xsens technologies b.v., 2011. URL: https://www.xsens.com/. 2

[YKL21] YANG D., KIM D., LEE S.-H.: Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum* (2021), vol. 40, Wiley Online Library, pp. 265–275. 2

[YLX*19] YAN S., LI Z., XIONG Y., YAN H., LIN D.: Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 4394–4402. 5

[YZ21] YING J., ZHAO X.: Rgb-d fusion for point-cloud-based 3d human pose estimation. In *2021 IEEE International Conference on Image Processing (ICIP)* (2021), pp. 3108–3112. doi:10.1109/ICIP42928.2021.9506588. 2

[YZH*22] YI X., ZHOU Y., HABERMANN M., SHIMADA S., GOLYANIK V., THEOBALT C., XU F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 13167–13178. 2

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[YZX21] YI X., ZHOU Y., XU F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–13. 2

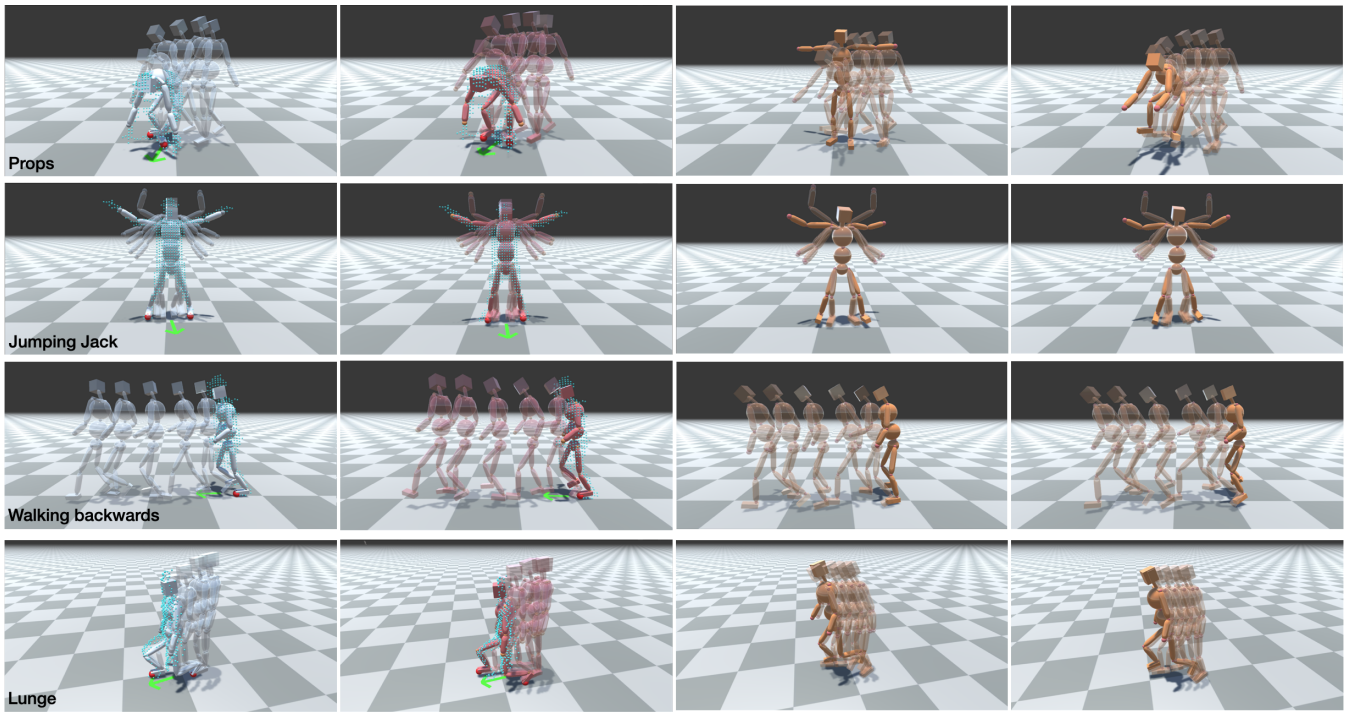
[ZCP*22] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022). 3

[ZML*22] ZHU W., MA X., LIU Z., LIU L., WU W., WANG Y.: Motionbert: Unified pretraining for human motion analysis. *arXiv preprint arXiv:2210.06551* (2022). 3, 6

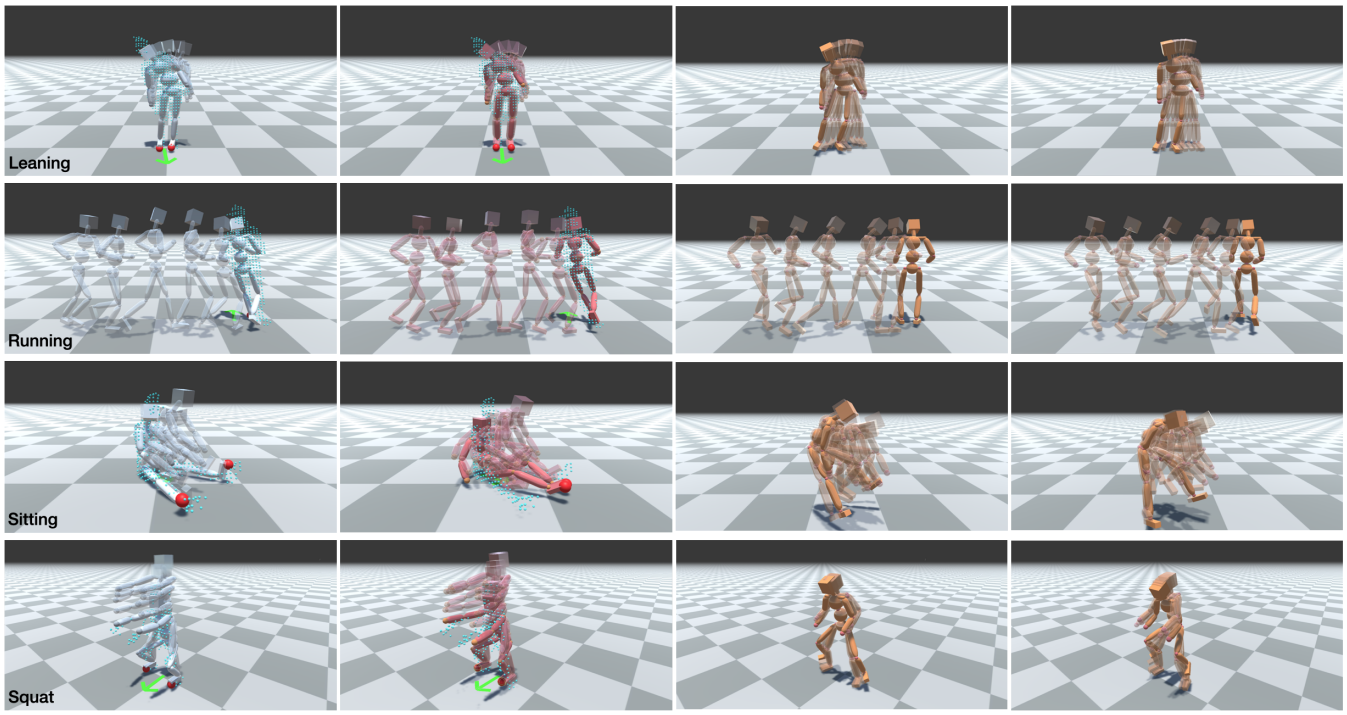
[ZTJF21] ZHENG W., TANG W., JIANG L., FU C.-W.: Se-ssd: Self-ensembling single-stage object detector from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 14494–14503. 2

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



(a) *Male / 170 cm*



(b) *Female / 162 cm*

Figure 8: *Qualitative comparisons of full-body motion outputs: Ground Truth, MOVIN-Ours, VIBE, and MotionBERT (from left to right). Our model, MOVIN, accurately generates output motion that closely resembles the ground truth, with natural joint trajectories. In contrast, baseline methods often suffer from issues such as oversmoothing, inaccurate pose, or temporal discontinuities with noticeable jitter.*

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

14678659, 2023, 7, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.14961 by Cas - Shanghaitech University, Wiley Online Library on [16/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License