CAvatar: Real-time Human Activity Mesh Reconstruction via Tactile Carpets

WENQIANG CHEN, Massachusetts Institute of Technology, USA
YEXIN HU, Carnegie Mellon University, USA
WEI SONG, University of New South Wales, Australia
YINGCHENG LIU, Massachusetts Institute of Technology, USA
ANTONIO TORRALBA, Massachusetts Institute of Technology, USA
WOJCIECH MATUSIK, Massachusetts Institute of Technology, USA

Human mesh reconstruction is essential for various applications, including virtual reality, motion capture, sports performance analysis, and healthcare monitoring. In healthcare contexts such as nursing homes, it is crucial to employ plausible and non-invasive methods for human mesh reconstruction that preserve privacy and dignity. Traditional vision-based techniques encounter challenges related to occlusion, viewpoint limitations, lighting conditions, and privacy concerns. In this research, we present CAvatar, a real-time human mesh reconstruction approach that innovatively utilizes pressure maps recorded by a tactile carpet as input. This advanced, non-intrusive technology obviates the need for cameras during usage, thereby safeguarding privacy. Our approach addresses several challenges, such as the limited spatial resolution of tactile sensors, extracting meaningful information from noisy pressure maps, and accommodating user variations and multiple users. We have developed an attention-based deep learning network, complemented by a discriminator network, to predict 3D human pose and shape from 2D pressure maps with notable accuracy. Our model demonstrates promising results, with a mean per joint position error (MPJPE) of 5.89 cm and a per vertex error (PVE) of 6.88 cm. To the best of our knowledge, we are the first to generate 3D mesh of human activities solely using tactile carpet signals, offering a novel approach that addresses privacy concerns and surpasses the limitations of existing vision-based and wearable solutions. The demonstration of CAvatar is shown at https://youtu.be/ZpO3LEsgV7Y.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: human activity reconstruction, 3D human mesh, pressure and vibrations, tactile sensor

ACM Reference Format:
Wenqiang Chen, Yexin Hu, Wei Song, Yingcheng Liu, Antonio Torralba, and Wojciech Matusik. 2023. CAvatar: Real-time Human Activity Mesh Reconstruction via Tactile Carpets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 151 (December 2023), 24 pages. https://doi.org/10.1145/3631424

1 INTRODUCTION

Human mesh reconstruction, the process of recovering the three-dimensional (3D) structure of the human body from diverse data sources, is essential for a wide range of applications, such as virtual reality, motion capture for animation and film, sports performance analysis, and healthcare monitoring. Specifically, in the context of

Authors' addresses: Wenqiang Chen Massachusetts Institute of Technology, USA, wenqiang@mit.edu; Yexin Hu Carnegie Mellon University, USA; Wei Song University of New South Wales, Australia; Yingcheng Liu Massachusetts Institute of Technology, USA; Antonio Torralba Massachusetts Institute of Technology, USA; Wojciech Matusik Massachusetts Institute of Technology, USA.

2474-9567/2023/12-ART151 $15.00
https://doi.org/10.1145/3631424

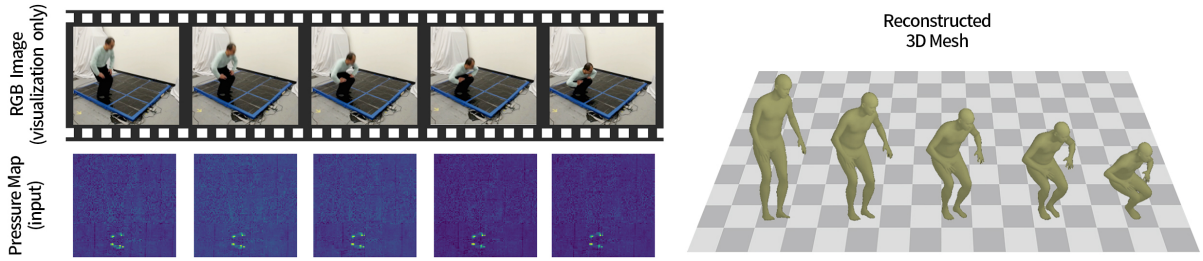

Fig. 1. This study leverages tactile carpet data to create 3D mesh representations of human activities.

healthcare monitoring within nursing homes and assisted living facilities, there is a critical need for plausible and non-invasive methods for human mesh reconstruction that not only ensure the privacy and dignity of patients and residents but also provide valuable insights into their well-being and daily activities. To address these requirements, innovative approaches must be developed that can effectively capture and analyze human motion and interactions while minimizing the impact on individuals' personal space and comfort. In recent years, substantial progress has been achieved in the field of human mesh reconstruction using vision-based techniques, notably with the advent of deep learning-based methods [35]. Nevertheless, these approaches frequently face obstacles stemming from occlusion, viewpoint limitations, and susceptibility to lighting conditions. Furthermore, employing cameras in sensitive settings, such as nursing homes, raises privacy concerns, as patients and residents may experience discomfort with constant video monitoring. Recent RF-based radar methods [78] also present privacy concerns, including the capacity to see through walls into adjacent areas. In addition, motion wearables [3] are hindered by battery constraints and necessitate regular charging, which can be problematic for elderly individuals who may struggle to remember to charge their devices or forget to wear them.

We observe that daily human activities, encompassing locomotion, exercises, and resting, are significantly influenced by tactile interactions between humans and the ground. A tactile carpet, composed of carbon paper and pliable conductive threads, measures floor vibrations and generates a 2D pressure map in real-time. Distinct human activities yield unique 2D pressure maps, rendering carpet pressure sensors a promising alternative that offers a non-invasive and unobtrusive sensing modality capable of addressing the limitations inherent to vision-based techniques while maintaining privacy. By capturing pressure distributions as individuals traverse the surface, these sensors provide invaluable insights into human motion and interactions with the surrounding environment.

In this work, we capitalize on tactile interactions to propose CAvatar, a real-time human mesh reconstruction approach utilizing pressure maps recorded by a tactile carpet as input, as shown in Figure 1. This smart carpet solution represents an innovative and non-intrusive technology that can be installed beneath any carpet, reducing the likelihood of causing discomfort as it is explicitly designed for walking. This method obviates the need for cameras during usage, generating 3D avatars from the carpet without exposing personal details. Our objective is to reconstruct plausible 3D virtual representations, as opposed to real videos, thereby preserving privacy. Harnessing the potential of tactile carpet data enables our approach to overcome the limitations of existing wearable devices and camera-based systems, paving the way for new opportunities in human activity monitoring and mesh reconstruction across various domains. Although prior research [43] has investigated human pose estimation using pressure sensors, the majority of these methods concentrate on simplified representations or skeletal models and lack the granularity necessary for comprehensive 3D human mesh generation.

Inferring 3D body meshes exclusively from carpet signals poses significant challenges. First, unlike cameras and radar, tactile sensors provide only 2D pressure map information, which lacks spatial resolution. Consequently, capturing dynamic 3D body meshes that characterize the human body and its motion becomes arduous. To

address this limitation, we employ a "teacher" model, trained from open-source camera models, to instruct the tactile carpet. By leveraging cross-modality supervision, our approach reconstructs 3D human activities without requiring cameras during the usage process.

Second, the pressure maps produced by the tactile carpet are 2D in nature and may contain noise and ambiguity. Extracting and interpreting meaningful information from these maps to reconstruct a 3D human mesh is a complex endeavor. To overcome this, we design an attention-based deep learning network for generating human mesh, supplemented by a discriminator network to assess the naturalness of the generated pose and shape parameters.

Third, numerous practical challenges arise, such as the variation in user styles while performing the same activity and the presence of multiple users on the carpet simultaneously. Consequently, our model must be adaptable to accommodate unseen users.

We conduct a comprehensive evaluation of the model through both qualitative and quantitative measures. Our experimental results demonstrate the capability of our system to generate plausible 3D human mesh reconstructions. The model exhibits a mean per joint position error (MPJPE) of 5.89 cm and a per vertex error (PVE) of 6.88 cm when compared to the ground truth derived from visual information. Additionally, we perform ablation studies to assess the significance of individual components in our model and evaluate its generalization performance on unseen individuals. Furthermore, our approach can be effectively scaled up to accommodate multi-person 3D mesh estimation scenarios.

The major contributions of this paper are as follows: To the best of our knowledge, we are the first to generate 3D mesh of human activities solely using tactile carpet signals, presenting a novel approach that addresses privacy concerns and overcomes the limitations of existing vision-based and wearable solutions in various domains, particularly healthcare monitoring. We designed a transformer-based neural network, incorporating attention mechanisms and a discriminator network, to effectively predict 3D human pose and shape from 2D pressure maps, enabling the extraction of meaningful information from the tactile data. We evaluated the model through extensive experiments, including qualitative and quantitative assessments, ablation studies, and generalization tests on unseen individuals and multi-person scenarios.

2 RELATED WORK

In this segment, we initially elucidate the current advancements in the field of human activity recognition, followed by a discussion pertaining to tactile sensing. Subsequently, we present an overview of the work conducted in the realm of human mesh reconstruction. As per our comprehensive understanding and research, we are pioneers in generating 3D mesh representations of human activities using solely tactile carpet signals. This innovative approach mitigates privacy concerns and transcends the limitations inherent in existing vision-based and wearable solutions. It is especially applicable in various domains, with a notable emphasis on healthcare monitoring.

2.1 Human Activity Recognition

In recent years, Human Activity Recognition (HAR) employing sensory data has garnered considerable attention due to its growing significance [33, 40]. A diverse array of applications, such as health and fitness monitoring [14, 31, 44, 59], remote patient care [49], and smart homes [25], have adopted HAR technologies. Presently, numerous sensors and wireless methodologies have been implemented for HAR applications, including Radio Frequency Identification (RFID) [48], WiFi [58], radio signals [39], and wearable sensors, such as acoustic sensors [68], accelerometers [23], and gyroscopes [20].

Recently, Peng et al. [53] presented a graph convolutional neural network (GCN) approach for HAR utilizing skeleton data obtained from cameras. Wearable devices, such as accelerometers and gyroscopes, have been extensively employed for capturing motion and orientation data [6–13, 15–19, 29, 71]. Numerous investigations have concentrated on recognizing activities like walking, running, and sitting [2]. In addition to wearable devices,

contemporary research has delved into the application of non-invasive sensors, including pressure sensors, for HAR purposes. For instance, Cheng et al. [21] proposed a deep learning-based framework for activity recognition. Likewise, Pham et al. [56] introduced a smart shoe-based HAR system employing pressure sensors for activity recognition.

Although these recent studies have contributed to advancements in HAR, their primary focus remains on activity recognition rather than 3D human activity mesh reconstruction. This unaddressed research challenge necessitates the development of innovative methods for capturing and processing 3D information from various sensors, and incorporating it with machine learning algorithms capable of reconstructing detailed and accurate human activity meshes.

2.2 Tactile Sensing

In recent years, the field of tactile sensing has garnered considerable attention as an alternative modality for human motion capture and activity recognition. Tactile sensors offer valuable insights into human motion and interaction with the environment through the measurement of pressure distributions. These sensors generally comprise arrays of miniature pressure-sensitive elements, which generate a two-dimensional map of pressure readings. Such sensors have been integrated into a range of wearable devices, including smart gloves [1, 64] and shoes [26], as well as non-wearable solutions like smart beds [50] and smart floors [54, 61, 62, 79].

Capacitive sensing is a predominant approach in tactile sensing technology, where changes in capacitance due to pressure exerted on sensor elements are measured [28, 72]. Capacitive sensors have found applications in numerous fields; however, the intricate manufacturing processes associated with capacitive sensing make it less practical for widespread adoption. Resistive sensing presents an alternative method for tactile sensing, wherein applied pressure to sensor elements results in altered resistance and a measurable voltage drop. Pressure-sensitive carpets [70] and mats [60], for instance, have been suggested for capturing human motion and activity. Similarly, pressure-sensitive insoles have been utilized for identifying walking patterns [51]. Optical tactile sensors, which rely on the deformation of transparent elastomers, have been proposed as an innovative approach to human motion capture. [38] Piezoelectric materials, which generate an electrical charge upon experiencing mechanical stress, represent a recent advancement in tactile sensing. Yu et al. [76] introduced a flexible piezoelectric tactile sensor that can be embedded into wearable devices for human activity recognition.

Nonetheless, the majority of these methodologies primarily focus on activity recognition or pose estimation, rather than three-dimensional human mesh reconstruction. Moreover, most of these studies center on the hardware design of sensors, whereas our research emphasizes the development of algorithms based on sensor readings.

2.3 Human Mesh Reconstruction

The reconstruction of human mesh has been extensively investigated within the fields of computer vision and graphics. Traditional methodologies for human mesh estimation have predominantly relied on model-based techniques, capitalizing on prior knowledge of human anatomy and kinematics [45]. However, these approaches necessitate accurate modeling of the human body, rendering them susceptible to model initialization and fitting issues.

The advent of deep learning-based methodologies has led to significant advancements in this domain [34, 36, 52]. These techniques generally rely on cameras or depth sensors, which can provide high-quality input data [4, 27, 30, 32, 46, 47, 55, 57, 63, 65, 66, 75]. However, the use of cameras raises privacy concerns, particularly in healthcare monitoring contexts. Moreover, cameras may not capture the entire body or may be hindered by occlusions and variable lighting conditions.

Also, researchers have explored alternative sensing modalities for human mesh reconstruction. For example, radar sensors have been employed to recover the 3D human body structure [69, 73, 74, 78]. Despite this, radar-based systems still raise privacy concerns, as they can penetrate walls and reveal adjacent areas. Wearable devices, such as motion sensors or inertial measurement units (IMUs), have also been utilized for human pose estimation and motion tracking [3, 42]. However, these devices may be obtrusive, uncomfortable, or suffer from battery limitations.

Recently, tactile sensing has emerged as a promising alternative modality for human pose and mesh reconstruction [5]. For instance, in-bed pressure sensors have been used for human pose estimation utilizing complete body shape pressure images [22]. However, this approach was solely dedicated to estimating simplified skeletal models of the human body. This approach is limited to predicting the pose of users while they are lying on a bed, as opposed to predicting a diverse range of daily human activities. Similarly, Zhang et al. [77] proposed a method for 3D human mesh estimation using synthetic data, which simulates users lying on a bed sheet and captures the complete shape of the user's body pressure. In contrast, our method estimates the human mesh of activities without requiring a full body image (such as lying on a bed) and builds on this work by utilizing real-world pressure maps of carpets to reconstruct 3D human activities in real-time. The work most closely related to ours is that of Luo et al. [43], who devised a method for estimating human poses using pressure-sensitive carpets. While their approach demonstrated promising results, it was solely dedicated to estimating simplified skeletal models of the human body. In contrast, we are the first to employ pressure maps generated by a tactile carpet to reconstruct a comprehensive 3D human mesh of activities encompassing both human pose and shape.

In summary, our proposed approach represents a significant advancement in human activity mesh reconstruction using a novel sensing modality. It has the potential to revolutionize the field of human activity monitoring, particularly in healthcare settings, where privacy concerns and obtrusive sensing modalities pose significant limitations.

3 PRIMER ON CAVATAR MESH RECONSTRUCTION

In this section, we first expound on the observation that enables the carpet to differentiate between various activities, and potentially render a 3D human mesh. Then, we introduce the tactile carpet dataset.

3.1 Tactile Characteristics for HAR

Leveraging a large-scale, high-resolution tactile sensing platform enables the collection and analysis of individual foot pressure patterns. It further allows for an in-depth documentation of tactile interactions occurring between the human body and the floor surface during complex activities. The existing research indicates that the pressure maps generated by the tactile carpet can effectively classify different human activities. [43, 60, 70]

Figure 2 expounds on the pressure map variations stemming from diverse activities such as lunging, sitting, executing push-ups, sit-ups, twisting the waist, and raising arms. The figure also exemplifies a situation wherein an individual conducts a lateral upper-body rotation and arms raising. Despite the movements being largely confined to the upper body, distinct alterations in the foot pressure map persistently emerge. The upper limb's action, in particular, reveals a noticeable discrepancy in the pressure distribution applied on each foot.

In summary, we have observed that different movements, even those of the upper limbs, produce a different tactile pressure map. This serves as the basis for our use of carpets to identify actions and reconstruct the human mesh of activity.

3.2 Tactile Carpet Dataset

The primary focus of this study is the innovative approach and learning strategies employed to reconstruct human mesh from signals emanating from tactile carpets, as opposed to the creation of a novel sensor. Building

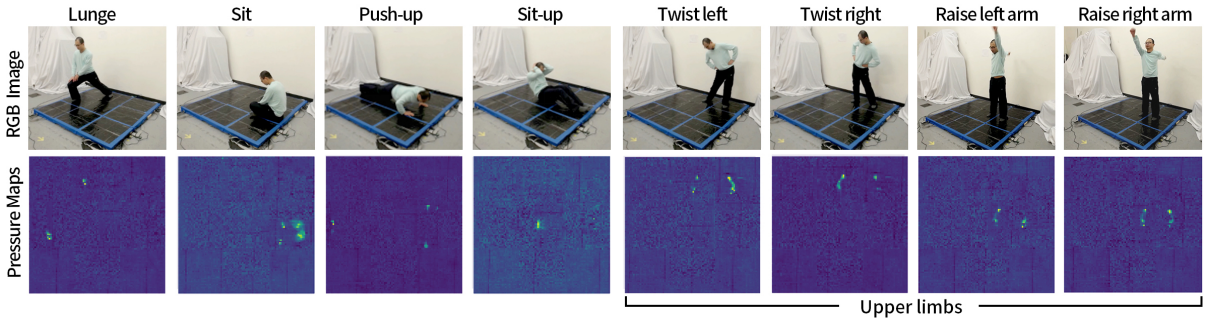

Fig. 2. Pressure map variations are illustrated from different activities. Notably, movements of the upper limbs cause changes in the weight distribution on the foot pressure map.

on the foundation of an existing dataset from open-source tactile carpet sensors [24], we harnessed this resource for our research.

The dataset stems from a large-scale, high-density piezoresistive pressure-sensing carpet encompassing over 36 square feet. This carpet integrates 9,216 sensors, each spaced at intervals of 0.375 inches. The design of the carpet incorporates a piezoresistive pressure-sensing matrix, constructed from a network of orthogonally aligned conductive threads, functioning as electrodes on each side of commercial piezoresistive films. Each sensor is positioned at the intersection of the orthogonal electrodes, boasting the capacity to measure pressure up to 14 kPa with an exceptional sensitivity of 0.3 kPa. A coupled readout circuit is employed to gather tactile frames, resulting in 9,216 unique sensing readouts at a frequency of 14 Hz. All tactile frames sourced from the carpet are complemented by synchronized visual frames procured through camera recordings.

The dataset under analysis includes a spectrum of continuous actions performed by 10 volunteers. To analyze the prediction effects of different actions, we segmented ten distinguishable activities from the video, including stepping, bending over, turning around, deep squatting, twisting the waist, raising arms, lunge, sitting, push-ups, and sit-ups, as shown in Figure 5. From the video in the dataset, each action is continually performed for around one minute when volunteers randomly walk to different locations on the carpet to perform the action. For each action, 80% of the data is designated for training, 10% percent is reserved for validation, and the remainder is allocated for testing. We perform the cross-validation to test all data. We continuously train and predict each frame. To extract more temporal features, we use a sliding window approach, setting the window size to 40 frames, which includes the target frame along with the previous 39 frames. The sliding window advances one frame at a time. For the purposes of our study, 145,500 frames of data in total are utilized for single-person mesh reconstruction, while an additional 8,400 frames are dedicated to multi-person mesh reconstruction.

4 INFERRING THE BODY

The utilization of carpet-based surveillance for human activity recognition confers a certain level of privacy preservation. However, its comparatively lower spatial resolution relative to cameras introduces a unique set of challenges. One particular issue that arises is the conversion of 2D pressure maps into a 3D human mesh.

In addressing this challenge, we propose a novel approach: a learning model that initially leverages camera-based systems during its training phase, but subsequently functions independently in real-world applications. This implies that camera systems would be employed during the training stage to "teach" the carpet-based monitoring system, yet they would be excluded from the system's operational deployment and usage.

In furtherance of this objective, we introduce a novel network, designated as CAvatar, delineated in Figure 3. The specifics of this proposition will be elucidated in the subsequent subsections.

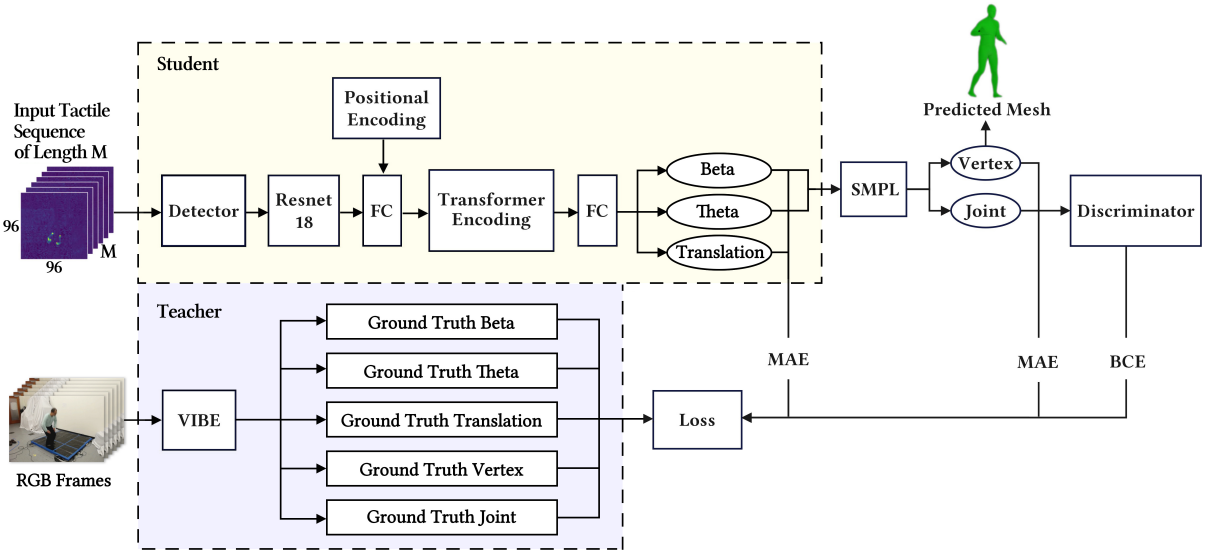

Fig. 3. Architecture of CAvatar. Video from cameras is employed solely during the training phase, serving as a "supervisor" for the tactile signals. However, it is important to note that these camera videos are not utilized during actual practice or application.

4.1 Human Mesh Representation

4.1.1 Skinned Multi-Person Linear Model (SMPL). SMPL is a mathematical model that can only use a few reduced parameters of β ($\dim(\beta) = 10$) and θ ($\dim(\theta) = 24 \times 3 = 72$) to represent a human body mesh of 6890 vertices (3×6890 dimensional space).

The human body representation in SMPL is primarily driven by a template mesh T, which is a triangulated mesh of the human body in a certain rest pose. This mesh has V (V=6890) vertices. The body shape variations are represented as: $S(\beta) = \bar{T} + B_s\beta$, where: $\bar{T}$ is the average template shape. B_s are the shape blend shapes or the shape basis that captures the deformations for different body shapes. β is a vector of shape coefficients ($\dim(\beta) = 10$). Pose-dependent deformations are represented as: $P(\theta) = B_p\theta$, where: B_p are the pose blendshapes or the pose basis that capture the pose-dependent deformations. θ is a vector of pose parameters, where $\dim(\theta) = 24 \times 3 = 72$ (Each joint has 3 angles in axis-angle representation.) The final vertex locations of the deformed template mesh are: $T(\beta, \theta) = S(\beta) + P(\theta)$.

However, simply adding pose and shape deformations isn't enough. SMPL uses a linear blend skinning (LBS) function which also takes into account joint rotations. The LBS function is: $W(T(\beta, \theta), J(\beta), \theta, W) = \sum_{k=1}^{K} w_k T(\beta, \theta) \odot \mathcal{T}_k(\theta)$, where: w_k are the blend weights for each vertex. K is the number of joints. $\mathcal{T}_k(\theta)$ is the transformation matrix for joint k given pose θ.

Directly manipulating the positions of each vertex would require 3 coordinates for each of the 6890 vertices in SMPL, resulting in a 20,670-dimensional space (3×6890) for each single shape. On the other hand, with SMPL's parameterization, only 10 shape coefficients and 72 pose coefficients are required. This drastically reduces the dimensionality and leads to a more compact and efficient representation. Moreover, SMPL's approach of mathematically defining a deformable mesh in terms of pose and shape parameters rather than vertices directly strikes a balance between efficiency, generalization, and semantic understanding, making it an elegant solution to representing the human body in a wide variety of shapes and poses. [41]

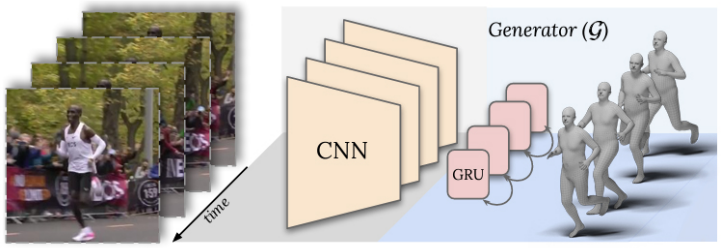

Fig. 4. VIBE Architecture [35].

4.1.2 Video Inference for Human Body Pose and Shape Estimation (VIBE). VIBE's objective is to fit the SMPL model to video data. A key component of this is the reprojection loss, which minimizes the difference between the 2D projections of the 3D keypoints of the model and the 2D keypoints detected in the video.

The loss function can be more precisely represented as:

$$L_{\text{VIBE}}(\beta, \theta) = \alpha L_{2\text{D}}(P(W'M(\beta, \theta)), J_{2\text{D}}) + \lambda L_{\text{SMPL}}(\beta, \theta) + \gamma L_{3\text{D}}(W'M(\beta, \theta), J_{3\text{D}})$$

Where: $L_{2\text{D}}$ is the reprojection loss for 2D keypoints. $J_{2\text{D}}$ are the ground-truth 2D keypoints. L_{SMPL} is the SMPL loss. $L_{3\text{D}}$ is the loss between predicted 3D keypoints and ground-truth 3D keypoints. $J_{3\text{D}}$ are the ground-truth 3D keypoints. α, λ, γ are scalar weights for the respective loss terms. $M(\beta, \theta)$ represents the posed and shaped 3D mesh of the body as estimated by the SMPL model. W' is a pre-trained linear regressor, mapping the vertices of the deformed body mesh to the reduced joint locations. P represents a 2D projection of 3D joints. The parameters within the VIBE model's overall loss function are determined through an iterative process of empirical tuning and domain knowledge. Researchers' experiment with different values to strike a balance that optimizes the model's performance. The choice of hyperparameters is guided by domain-specific insights, adapting them to the relative importance of specific components for the given application. The best hyperparameter values are found by adjusting them iteratively to achieve the highest performance on a validation dataset, with the aim of accurately estimating human pose and shape from video data.

The overall framework of VIBE [35] is summarized in Fig. 4. Given an input video $V = \{I_t\}_{t=1}^{T}$ of length T, of a single person, it extracts the features of each frame I_t using a pretrained CNN. It trains a temporal encoder composed of bidirectional Gated Recurrent Units (GRU) that outputs latent variables containing information incorporated from past and future frames. Then, these features are used to regress the parameters of the SMPL body model at each time instance.

4.2 CAvatar Encoder

This section presents an innovative methodology that incorporates the application of a Temporal Encoder, a fusion of ResNet and Transformer models, in conjunction with the Skinned Multi-Person Linear (SMPL) model for the inference of a 3D human body mesh from tactile carpet data. This methodology is primarily focused on extracting critical body mesh attributes, such as shape parameters β, pose parameters θ, camera translations c, vertices V, and joints J. Please be aware that for human mesh reconstruction, we estimate shape parameters (β), pose parameters (θ), vertices (V), and joints (J). Camera translations (c) are only used for projecting the three-dimensional human mesh onto a designated camera plane. These parameters dictate the position and scale of the human mesh on the camera plane, facilitating its visualization in comparison to RGB images during the evaluation process.

To process our single-channel pressure data, we employ a customized pre-trained ResNet18 model, modified from its initial design for handling RGB images. We feed each frame to the ResNet18 model. This, however, cannot learn the sequential information in our data. Therefore, we set a window of 40 frames (the target frame and the previous 39 frames) as input for a Transformer encoder. The final linear and pooling layers of the ResNet18 have been removed and replaced with a fully connected layer. This ensures that the ResNet18 output dimension matches the Transformer encoder input dimension.

To harness the temporal dependencies within our sequence data, we utilize a Transformer model, renowned for its proficiency in processing such data. The Transformer model recognizes sequential information through the integration of positional encoding, a concept influenced by Vaswani et al. [67]. This form of encoding adds a unique positional signal, allowing the model to acknowledge the order of tokens in the input sequence. The positional encoding for the i^{th} position and d^{th} dimension can be calculated using the following formulas in a sine and cosine alternation [67]:

$$PE_{(i,2d)} = \sin\left(\frac{i}{10000^{2d/d_{\text{model}}}}\right)$$

$$PE_{(i,2d+1)} = \cos\left(\frac{i}{10000^{2d/d_{\text{model}}}}\right)$$

Where $PE_{(i,d)}$ denotes the positional encoding of the i^{th} position in the d^{th} dimension, and d_{model} is the dimension of the input features. This encoding method infuses the concept of absolute positions into the Transformer, which intrinsically does not possess such positional awareness due to its self-attention mechanisms. The approach of directly adding the positionally encoded features to the features obtained from the linear layer is adopted, instead of concatenation. This strategy preserves the integrity of positional information and also maintains the same feature dimensionality. As a result, it does not add to the computational complexity of the model.

Upon processing through the Transformer, we obtain the shape parameters β, pose parameters θ, and camera translations c. Subsequently, we feed the learned shape parameters β, and pose parameters θ into the Skinned Multi-Person Linear (SMPL) model [41]. This model, a differentiable function of body shape and pose parameters, generates a 3D mesh output represented by vertices V and joints J, providing a comprehensive depiction of the human body shape and pose. Directly learning a large number of parameters for vertices V and joints J is computationally intensive and complex. Therefore, we only learn a limited number of parameters, β and θ. Subsequently, we use SMPL to generate vertices V and joints J.

In terms of label generation and loss computation, we employ the Video Inference for Body pose and shape Estimation (VIBE) model [35]. The VIBE model provides the ground truth for shape parameters, pose parameters, camera parameters, vertices, and joints. The loss function, defined as the Mean Absolute Error (MAE), is computed between the predicted parameters and outputs and the ground truth values:

$$L_{CA} = \frac{1}{N_\beta}\sum_{i=1}^{N_\beta}|\beta_i - \hat{\beta}_i| + \frac{1}{N_c}\sum_{i=1}^{N_c}|c_i - \hat{c}_i| + \frac{1}{N_\theta}\sum_{i=1}^{N_\theta}|\theta_i - \hat{\theta}_i| + \frac{1}{N_V}\sum_{i=1}^{N_V}|V_i - \hat{V}_i| + \frac{1}{N_J}\sum_{i=1}^{N_J}|J_i - \hat{J}_i|.$$

It is noteworthy to mention that this loss function effectively captures the deviations between the predicted and ground truth values for all relevant parameters. The MAE is an optimal choice for this purpose, as it is resistant to outliers and is a robust measure of prediction error. It ensures that the model learns to predict parameters that are as close to the ground truth as possible, thereby optimizing the performance of the inference model.

4.3 Self-Attention for Dynamic Human Activities

The Self-Attention Mechanism, an integral component of the proposed methodology, is embedded within the Transformer model. This mechanism is particularly esteemed for its aptitude in handling sequence data efficiently, which is a critical requirement in our context of interpreting tactile carpet data. Given the nature of the carpet data, where each sensor activation is influenced not only by its immediate neighbours but potentially by activations elsewhere in the sequence, the ability to maintain a global view of dynamic human activities becomes indispensable.

In essence, the attention mechanism empowers the model to selectively focus on varying sections of the input sequence during the generation of output for each timestep. It assigns a weight to each input in the sequence, the magnitude of which indicates the relevance or 'attention' that should be given to each tactile carpet signal input when making predictions about a specific human mesh output.

Our proposed methodology initiates with the extraction of tactile features from a carpet at various temporal instances, represented as $H = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_t)$. The self-attention module incorporated within our model leverages a specialized function to selectively focus on different frames over time. Predominantly, we employ the "Scaled Dot-Product Attention" mechanism in our attention module. This mechanism requires inputs in the form of queries (Q), keys (K), and values (V). These Q, K, and V components are obtained by performing a dot product operation between the carpet tactile features $\mathbf{h}_t$ and their corresponding learnable weight matrices. Subsequently, our attention mechanism is computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In the above equation, d_k represents the dimensionality of the keys and queries and serves as a normalization factor. Our model further incorporates the concept of multi-headed attention [67]. This mechanism enables the model to concurrently attend to information from different representational subspaces at various positions. The feature vectors derived from different attention heads are concatenated, culminating in the generation of the final feature set, which is employed for the prediction of the human mesh.

The Self-Attention Mechanism ensures that every discrete piece of data is contemplated within the context of the entire sequence, thereby facilitating an accurate inference of 3D human body mesh of dynamic activities from tactile carpet data.

4.4 Adversarial Mesh Learning

Our encoder, fortified by the application of Mean Absolute Error (MAE) loss, forecasts the 3D human mesh of future instants based on the temporal dependencies associated with the preceding 'n' poses. The inherent challenge lies in the potential minimization of network loss by poses that are naturally implausible. To counteract this, we propose an adversarial mesh learning approach. This approach involves a Generative Adversarial Network (GAN) setup that includes a generator (G) and a discriminator (D). These are trained in an adversarial manner, where the discriminator's function is to penalize poses that diverge from natural human movement patterns.

4.4.1 Generator. Our generator network deviates from traditional GAN approaches. Instead of merely capturing the data distribution to generate deceptive samples, our generator strives to learn the temporal mapping from carpet tactile information to human mesh parameters [37]. Over the course of the training process, the generator progressively enhances its accuracy in pose prediction, thereby increasing its potential to deceive the discriminator. The adversarial loss associated with the generator, denoted by L_{Adv}, is defined as follows:

$$L_{Adv} = \mathbb{E}_{J\sim\hat{P}}[-log(D(J))]$$

Here, J signifies the estimated joints position and $\hat{P}$ signifies the SMPL generated human mesh. This adversarial loss (L_{Adv}) is compounded with the CAvatar encoder loss to facilitate the training of the generator, and the total loss of generator, denoted by L_G, is given by:

$$L_G = L_{Adv} + L_{CA}$$

4.4.2 Discriminator. The discriminator network is designed to differentiate between real data and generated data, thereby ensuring the generation of valid output. It takes as inputs the manifold of ground truth (P) and predicted mesh ($\hat{P}$), and it is entrusted with the task of discerning whether the generated pose aligns with the manifold of natural poses. To this end, the discriminator employs a multi-layer perception, which outputs a value between 0 and 1. This value corresponds to the probability of the pose being within the manifold of naturally plausible poses. The adversarial loss of the discriminator, denoted by L_D, is given by:

$$L_D = \mathbb{E}_{J\sim P}[-log(D(J))] + \mathbb{E}_{J\sim \hat{P}}[-log(1 - D(J))]$$

In the above equation, $\hat{P}$ and P denote the manifolds of generated and ground truth SMPL human mesh, respectively.

4.5 Fine-Tuning for Unseen Individuals

While the general applicability of our adversarial mesh learning approach is crucial, it is equally important to ensure the model can accommodate unseen individuals. While our model exhibits generalizability to unseen users, it encounters challenges when dealing with more complex actions. To overcome this, we employ fine-tuning techniques to effectively adapt to the distinctive characteristics of each new individual. This process only necessitates a short initial training period for a new user.

Fine-tuning involves adjusting the parameters of our pre-trained model to enhance its performance for an unseen user. The primary objective is to capitalize on the previously learned features and apply them to the new unseen user, thereby avoiding the need to train the model from scratch.

4.5.1 Initialization. CAvatar network is first pre-trained on a large dataset comprising a diverse range of individuals performing various activities. This pre-training phase allows the model to learn a broad range of human body shapes and movements, thereby creating a solid foundation for fine-tuning.

4.5.2 Adaptation Phase. In the adaptation phase, the model is fine-tuned on data from the unseen individual. The aim is to adjust the model parameters such that it can accurately predict the 3D human mesh for the new individual. Given the diversity in human body shapes and movements, this fine-tuning process is crucial to ensure the model generalizes well to new individuals.

During fine-tuning, we retain the learned weights from pre-training as initialization and update them using a smaller learning rate. This is done to prevent drastic changes to the model parameters, ensuring the model retains the general knowledge acquired during pre-training while adapting to the specifics of the new individual.

4.6 Simultaneous Multi-User Prediction

To facilitate the concurrent prediction of multiple users on a tactile carpet, we introduce a sophisticated multi-user adaptation mechanism. This mechanism allows for the simultaneous generation of a plurality of human mesh. We achieve this by enhancing our extant ResNet-Transformer Encoder's functionalities, enabling it to classify the count of users on the tactile carpet, and subsequently recalibrate the output parameters commensurately with this classification. As depicted in Figure 3, we initially employ this proposed classifier as a detector to ascertain the number of users on a carpet at a given moment, prior to the operation of the generator.

4.6.1 Modification of Encoder Classification. We introduce a novel modification to the ResNet-Transformer Encoder architecture, equipping it to concurrently function as a classifier. This classifier is responsible for ascertaining the number of users currently on the tactile carpet, symbolized by N. The classifier can be seamlessly integrated into the prevailing encoder architecture as an auxiliary output layer, specifically devised for the estimation of N. For this classification endeavor, we suggest employing the cross-entropy loss function, a well-regarded and efficient tool for classification problems.

4.6.2 Modification of Parameters. Upon the successful classification of the number of users and generation parameters that concurrently represent N users, the shape parameters (β), pose parameters (θ), and transition parameters (c) are multiplied by N and then feed into SMPL model separately. Note that the camera transition parameters (c) have the location information that separates different users in the learning phase.

4.6.3 Adaptation of the SMPL Model. The modified parameters are subsequently fed into the Skinned Multi-Person Linear (SMPL) model orderly. This adaptation permits the concurrent generation of multiple human mesh, each corresponding to one of the N users as determined by the modified encoder. Since people rarely perform actions with one on top of the other, we assume that the pressure maps induced by the actions of different people will not overlap at the given time.

This adaptation of the proposed network enables the concurrent processing of multiple users' tactile data and the simultaneous generation of a unique human mesh for each user. This substantial enhancement to the proposed model's capabilities renders it an indispensable resource for real-world applications, particularly those where the tactile carpet may be occupied by multiple users simultaneously.

4.7 Training Procedure and Implementation Details

The fully-connected layer between ResNet18 and transformer encoder is a single MLP layer with 256 neurons, thereby resulting in the generation of a 256-dimensional feature vector, denoted as $f_i \in \mathbb{R}^{256}$. Our configuration incorporates a sequence length $T = 40$ and a batch size of 32, enabling us to train our model on a single Nvidia RTX4090 laptop GPU. To encode the temporal information inherent in the sequences, we apply positional encoding before feeding them into the transformer encoder. The classic transformer encoder comprises 6 encoder layers, each endowed with 4 heads and a total feature dimension of 256. The feedforward network dimension within the transformer encoder layer is set to 2048, dropout value is set to 0.1 and eps value in layer normalization components is set to 1×10^{-5}. Then we use a single linear layer to receive the output from the transformer encoder layer and generate an output $\Theta \in \mathbb{R}^{86}$, encompassing joint angles, shape, and camera parameters. The SMPL uses these joint angles and shapes as input, and subsequently generates human vertices and joint positions. These joint positions serve as 'fake' samples for the discriminator, while the ground truth joint positions are provided as 'real' samples. The discriminator's network architecture includes a single MLP hidden layer with 256 neurons, utilizing a tanh activation function. The final layer employs a sigmoid activation function and output a single probability for each sample, indicating its classification as 'fake' or 'real'. For the multi-user detector, we maintain an identical architecture as the ResNet-TransformerEncoder, but with a change in the output dimension to be the maximum number of individuals to be detected. In our experiment, the output dimension is 2. The multi-user generator and discriminator architectures are almost the same as those of the single-person model, with the exception of the generator's output dimension and the discriminator's input dimension. These dimensions are multiplied by N, where N represents the number of individuals detected on the carpet. The optimization process leverages the Adam Optimizer, employing a learning rate of 1×10^{-4} (3×10^{-5} for fine-tuning) and a weight decay of 1×10^{-3}, with all other parameters set to their default values, which applies to all of the generators, discriminators and detector during the training process.

During the training phase, it took approximately 20 hours on a single Nvidia RTX4090 laptop GPU. In the testing phase, the acquisition of the SMPL parameters for each frame took about 0.01 seconds per frame, while rendering required about 0.05 seconds per frame.

5 EVALUATION

In this segment of the study, we commence by introducing the evaluation metrics and the ground truth, which encompass both qualitative and quantitative dimensions. Following this, we proceed to evaluate the 3D mesh associated with a variety of human activities, derived from a sample of ten participants. Subsequently, an assessment of body shape is conducted in comparison to RGB images. After this, we delve into an examination of human motion dynamics. Furthermore, we investigate the human mesh of users that have not been previously encountered. Additionally, we carry out an ablation study to obtain quantitative results from different model variants. Lastly, we assess scenarios in which multiple individuals are present on the carpet concurrently.

5.1 Evaluation Metrics

VIBE generates human mesh from the input of camera videos, with mean per joint position error (MPJPE) at 82.9mm, 96.6mm, and 65.9mm on 3DPW, MPI-INF-3DHP, and Human3.6M datasets, respectively. [35] This results in a reliably accurate human mesh representation. Although these three datasets have ground truth of human mesh from real human body scan, they do not have tactile carpet signals. In this study, we employed VIBE [35], to generate ground truth human mesh from the Intelligent Carpet dataset [24] for the qualitative assessment. We reported the widely accepted mean per joint position error (MPJPE) and calculated the per vertex error (PVE) between the predicted mesh and the ground truth. Furthermore, we examined the RMSE of camera parameters derived from the VIBE model and CAvatar, as our study implemented a weak-perspective camera model with scale and translation parameters.

While VIBE has minimal errors, it cannot be considered an absolute ground truth. Hence, we utilized both RGB images and the human mesh derived from VIBE for qualitative evaluation. We've illustrated the human mesh from CAvatar, the human mesh from VIBE, and RGB images to provide a visually intuitive comparison of their human pose and shape.

5.2 Human Mesh of Activities:

Our initial assessment involved a qualitative evaluation of human activity mesh reconstruction. As depicted in Figure 5, our model, CAvatar, effectively reconstructs the human mesh across a range of ten distinct activities, including stepping, bending over, turning around, deep squatting, twisting the waist, raising arms, lunge, sitting, push-ups, and sit-ups, as shown in Figure 5.

When juxtaposed with the RGB images and the ground truth human mesh from VIBE, CAvatar demonstrates a remarkable level of proficiency in the reconstruction of the human mesh.

Notably, CAvatar captures complex activity meshes, unearthing even the most nuanced details, such as the bending of a knee or the rotation of a torso. Furthermore, CAvatar demonstrates its capabilities in reconstructing human meshes with remarkable fidelity during strenuous and complex activities, such as the execution of sit-ups and push-ups. In such instances, CAvatar consistently exhibits its aptitude for accurate activity mesh estimation.

Intriguingly, we have also been successful in accurately predicting arm movements. Our hypothesis for this is that arm movements can cause shifts in the body's weight distribution on the feet, subsequently leading to alterations in the carpet's pressure. Additionally, it's plausible that there's a correlation between arm movements and the positioning of other body parts, such as the legs and torso.

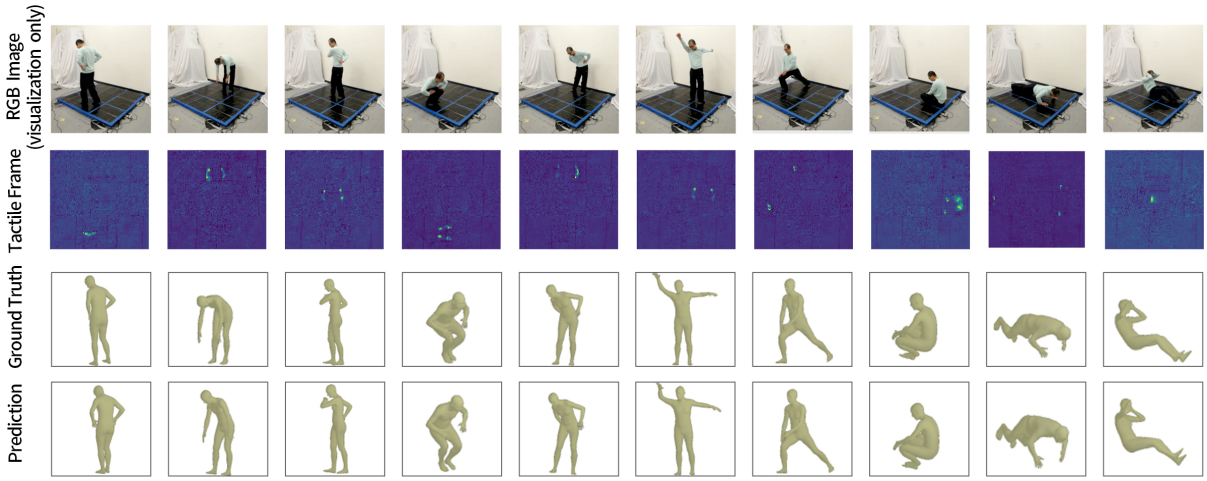

Fig. 5. Human mesh of various activities, derived from tactile frames, are exhibited. For the purpose of comparison, RGB images, tactile frames, and ground truth data are concurrently displayed.

Table 1. Quantitative results for Various Actions

Actions	Step	Bend	Turn	Squat	Twist	Arm	Lunge	Sit	Pushup	Sit-ups
MPJPE (cm)	3.1889	6.8761	5.9748	5.8791	4.4299	7.7079	4.9651	6.9154	8.1743	7.9778
PVE (cm)	3.7586	8.1444	7.2048	6.8088	5.1470	10.5193	5.4998	7.8902	8.9850	9.2185
Translation	2.7610	3.4778	2.7966	4.1136	3.1095	2.8341	3.6669	4.4229	5.2821	4.5469

In our quantitative evaluation, we first test the error of ten activities. The table 1 showcases performance metrics for various human actions. Notably, actions like "twisting the waist" have lower errors, while more complex movements such as "raising arms" exhibit higher errors.

Table 2. Quantitative Results of Human Mesh Reconstruction from Ten Users

#User	1	2	3	4	5	6	7	8	9	10
MPJPE (cm)	5.6757	6.2372	6.6749	5.4763	5.6981	5.8210	5.1544	5.9804	6.3031	5.8452
PVE (cm)	6.5507	7.3266	7.7985	6.4065	6.7911	6.7616	6.0447	6.9960	7.3342	6.8006
Translation	3.7336	3.8555	4.2558	3.4075	3.6303	3.4947	3.3947	3.7308	3.8802	3.7997

Then, the provided table 2 offers a comprehensive overview of the quantitative outcomes obtained from human mesh reconstruction across ten users. The Mean Per Joint Position Error (MPJPE) values show the average deviation between reconstructed and actual joint positions, with variations ranging from 5.4763 to 6.6749 cm. Additionally, the Percentage of Volume Error (PVE) values illustrate the extent to which reconstructed volumes diverge from true values, ranging from 6.0447 to 7.7985 cm. The "Translation" column indicates the translation differences, with values varying between 3.3947 and 4.2558. These errors are key in assessing the model's ability to provide accurate weak-perspective camera translations.

Overall, this table serves as a succinct representation of the accuracy and precision of the human mesh reconstruction process, providing valuable insights into the quality of results achieved across different users.

Table 3. Joint Distances and Standard Deviations

Jts.	Hd.	Nk.	LS.	LE.	LWr.	RS.	RE.	RWr.	LH.	LK.	LA.	RH.	RK.	RA.
ED (cm)	6.03	4.24	4.30	6.82	10.34	4.42	7.08	10.19	1.71	4.89	7.97	1.75	4.68	7.61
STD	5.34	3.92	3.68	5.90	9.67	3.88	5.92	9.44	1.63	4.27	6.40	1.63	4.33	6.46

Furthermore, We proceed to analyze the Euclidean Distance of the 14 joints. (Head, Neck, Left Shoulder, Left Elbow, Left Wrist, Right Shoulder, Right Elbow, Right Wrist, Left Hip, Left Knee, Left Ankle, Right Hip, Right Knee, Right Ankle) Note that we do not analyze each vertex due to the considerable count of 6890 vertices. The table 3 displays joint distances and their corresponding standard deviations. It reveals variations in joint spatial relationships during movements. Notably, joints like "head", "wrist," "elbow," and "ankle" exhibit higher distances and standard deviations, implying dynamic motion and potential complexity. Most importantly, the notable error detected in the wrist measurements indicates an inability to accurately reconstruct hand gestures.

Table 4. Performance Metrics for Leave-One-Action-Out

Actions	Step	Bend	Turn	Squat	Twist	Arm	Lunge	Sit	Pushup	Sit-ups
MPJPE (cm)	11.742	24.567	17.455	16.171	13.363	15.969	13.954	26.726	44.039	25.444
PVE (cm)	13.333	29.976	22.509	18.060	16.680	21.627	15.221	27.267	46.845	28.632
Translation	6.652	6.072	3.748	11.163	4.954	6.071	7.146	18.080	16.237	11.283

At last, we conducted the experiment to present the performance for Leave-One-Action-Out. Analyzing the table, it's evident that "push-up," "sitting," and "sit-ups" exhibit the poorest performance. This could be attributed to the distinctiveness of these three actions compared to the others. While most actions involve pressure maps from only two footprints, "push-up" involves pressure maps from both hands and feet. The majority of other actions share more similarities, resulting in smaller errors. This highlights the importance of collecting a diverse range of actions to improve the performance and practicality of the system.

5.3 Body Shape:

An essential aspect of our evaluation involves the analysis of body shape reconstruction. As depicted in Figure 6, the integration of RGB frames and the VIBE-derived human mesh is achieved by overlaying the estimated 3D human mesh onto the original RGB frames. This juxtaposition allows evaluators to visually assess how accurately the mesh represents the human body in the video, focusing on aspects such as alignment, movement, and overall quality. Overlaying is done by projecting this mesh into 2D image space and then superimposing it onto the original RGB frames using compositing techniques. This combined visualization allows qualitative assessment of the mesh's alignment and accuracy with the human subject in the video. Specifically, human mesh models, which are generated from carpet data and represent a variety of actions performed by ten different individuals, are superimposed onto corresponding RGB images based on the camera translations c, which are denoted by scale and translation parameters $[s, t]$, where $t \in \mathbb{R}^2$. This facilitates the computation of the 2D projection of the 3D vertices $\hat{X}$, as $\hat{x} \in \mathbb{R}^{v\times 2} = s\,\Pi(R\,\hat{X}(\Theta)) + t$, where $R \in \mathbb{R}^3$ denotes the global rotation matrix, and Π signifies orthographic projection.

On close inspection of the images, it is clear that the shapes of the reconstructed human meshes align remarkably well with the actual shapes of the individuals. This alignment is indicative of the model's ability to accurately capture and reproduce the unique body shape of each individual, which is a critical aspect of successful human mesh reconstruction.

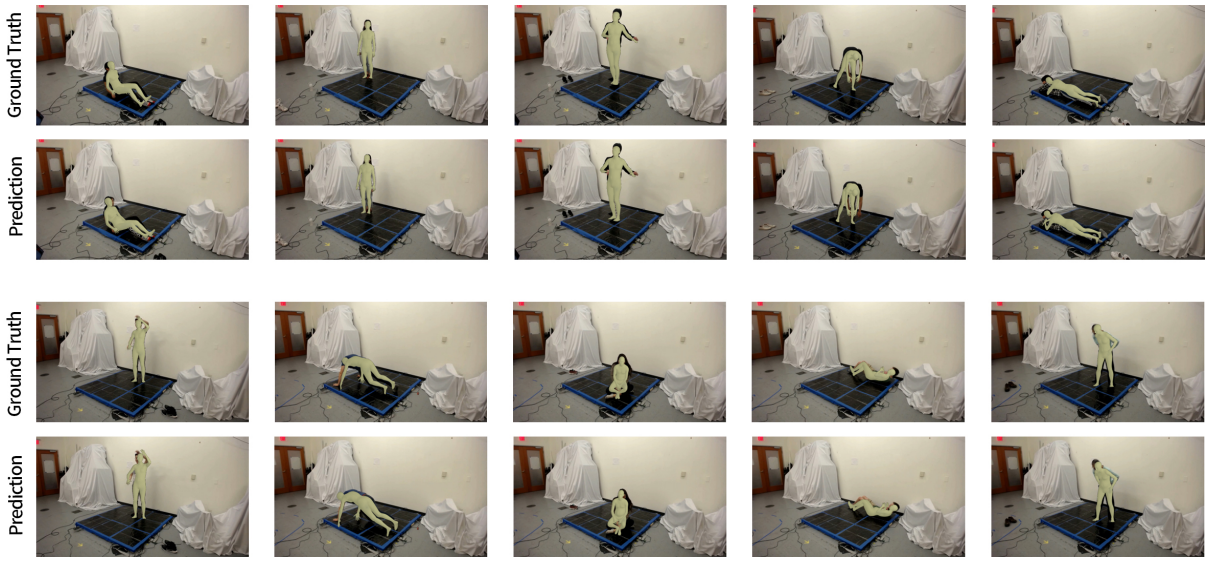

Fig. 6. Human mesh models, representing various actions performed by ten different individuals and generated from carpet data, are superimposed onto RGB images. As can be discerned from the images, the shapes of the human mesh align remarkably well with the actual shapes of the individuals. The resulting shapes display a nearly identical effect to the ground truth.

The shapes of the resulting meshes display a striking resemblance to the ground truth data, almost mirroring the actual body shapes of the individuals. This demonstrates the efficacy of CAvatar in accurately reconstructing the intricate details of individual body shapes from the tactile data.

In conclusion, CAvatar exhibits a high degree of accuracy and precision in body shape reconstruction, highlighting its potential for practical applications in environments where accurate body shape representation is paramount.

5.4 Human Motion Dynamics:

Evaluating human motion dynamics is a crucial part of our study. As shown in Figure 7, our model, CAvatar, successfully generates dynamic 3D meshes that represent a series of activities over a specific period.

A detailed examination of the figure, particularly focusing on the waist-twisting action, demonstrates that the produced meshes throughout the action sequence closely correspond to the actual performed movements. This close correspondence is indicative of the model's strong capability to capture and reconstruct the temporal dynamics of human motion.

The ability of CAvatar to accurately reconstruct human motion dynamics is particularly significant. This ability is not only vital for accurately capturing complex human actions but also instrumental in applications where understanding and predicting human motion over time is essential, such as in sports training or physical rehabilitation.

In summary, CAvatar demonstrates exceptional proficiency in capturing and reconstructing the dynamic aspects of human motion, thereby underlining its potential applicability in a wide range of practical scenarios.

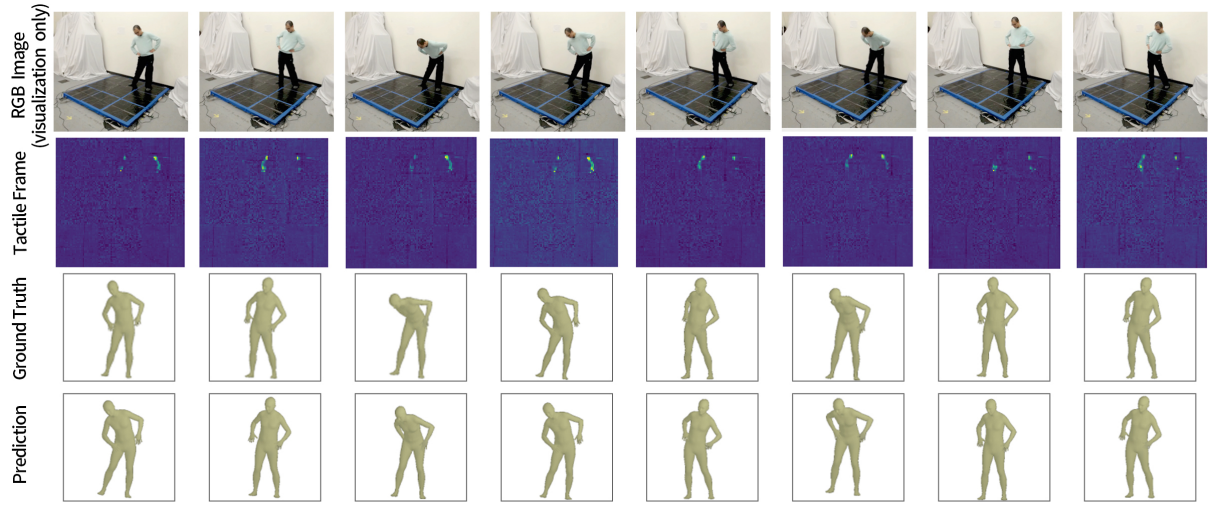

Fig. 7. Our model consistently generates dynamic 3D meshes representing activities over a period of time. As the subject performs waist-twisting actions, the produced meshes over this duration exhibit a close correspondence to the performed actions.

5.5 Unseen Individuals

The capability of our model, CAvatar, to accurately reconstruct human mesh for unseen users is a critical factor in assessing its generalizability. We employed a leave-one-out training strategy to evaluate the model's performance on unseen users.

Qualitatively, as presented in Figure 8, our pre-trained model has shown proficiency in generating convincing human mesh representations across a diverse set of activities performed by unseen users. However, it occasionally grapples with more complex activities, particularly those that involve intricate upper body movements. To address this, we have implemented a specialized fine-tuning method for unseen users. The considerable enhancements in human mesh reconstruction for more challenging activities, as a result of this approach, are well-demonstrated in the accompanying figure.

Quantitatively, we have utilized varying amounts of training data, specifically from 0% to 60%, to fine-tune our model. The outcomes of these measurements are summarized in Table 5.

Examining the Mean Per Joint Position Error (MPJPE) values, it's apparent that as the percentage of samples used for finetuning increases, the errors decrease. This trend suggests that the model benefits from more training data, becoming more adept at reconstructing human meshes for previously unseen users. Notably, the MPJPE drops from 14.4048 cm for 0% samples to 5.5056 cm for 60% samples, showcasing substantial improvement.

Similarly, the Percentage of Volume Error (PVE) values also exhibit a decreasing pattern as the finetuning samples increase. This underscores the model's capacity to better capture the shapes and volumes of the human meshes with a larger pool of training data.

It's worth mentioning that there's minimal difference in performance between our model using 50% and 60% of the training data. Consequently, we selected this proportion of 50% for the final model training. This only equates to a modest initial training requirement (performing actions with cameras) of merely 8.5 minutes per individual.

In conclusion, our model, CAvatar, has demonstrated its ability to adeptly generalize to unseen users through both qualitative and quantitative evaluations, especially when supplemented with a portion of the training data

for fine-tuning. This invaluable capability, essential for practical applications, solidifies the robustness of CAvatar and underscores its potential for extensive real-world utilization.

Table 5. Performance Metrics for Different Proportions of Finetuning

Samples (%)	0	10	20	30	40	50	60
MPJPE (cm)	14.4048	7.3682	7.1082	6.7348	6.4350	5.6679	5.5056
PVE (cm)	16.4528	8.5627	8.2625	7.7750	7.4025	6.5080	6.3784
Translation	2.9952	3.2607	3.6111	3.4918	2.9913	2.6046	2.8460

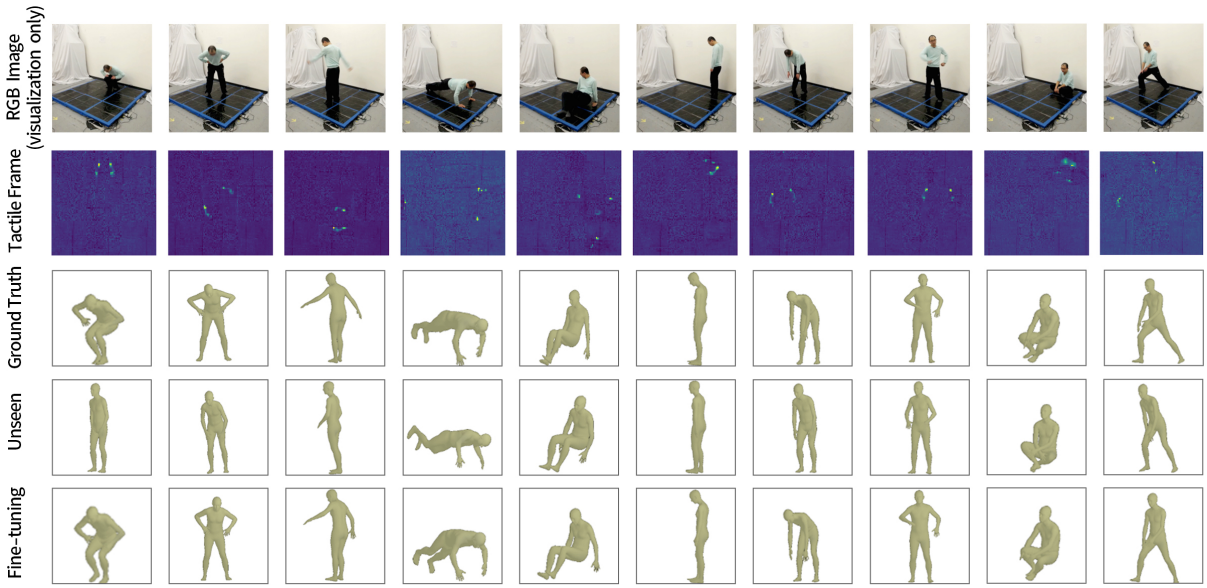

Fig. 8. Our pre-trained model is capable of generating plausible human mesh representations for certain activities performed by unseen users. However, it struggles with more challenging activities, particularly those involving upper body movements. By employing a fine-tuning method tailored for unseen users, we can effectively address this issue, as demonstrated in the accompanying figure.

5.6 Ablation Study

Our ablation study delves into the impact of integral components on our model's overall performance. The performance metrics extracted from the quantitative results, as depicted in Table 6, shed light on this analysis. Our comprehensive model, named *CAvatar*, integrates a loss function that combines shape parameters β, pose parameters θ, camera translations c, vertices V, joints J, and a discriminator.

To discern the significance of each component, we juxtaposed the performance of *CAvatar* against its variants, each lacking a specific component: *w/o Theta*, *w/o Beta*, *w/o Joint*, *w/o Vertex*, *w/o Discriminator*, and *w/o Attention*.

The analysis of the provided metrics highlights several key observations. Firstly, the MPJPE values indicate that the absence of specific components, such as pose parameters (*w/o Theta*) or shape parameters (*w/o Beta*),

Table 6. Quantitative results of different model variants.

	CAvatar	w/o Theta	w/o Beta	w/o J	w/o V	w/o Dis	w/o Att
MPJPE (cm)	5.6679	6.3599	6.2349	6.2833	6.3031	5.8099	6.8718
PVE (cm)	6.5080	7.4436	7.1602	7.1298	7.2692	6.7405	7.9727
Camera	2.6046	3.0630	3.1699	2.7962	2.7986	3.1752	3.3802

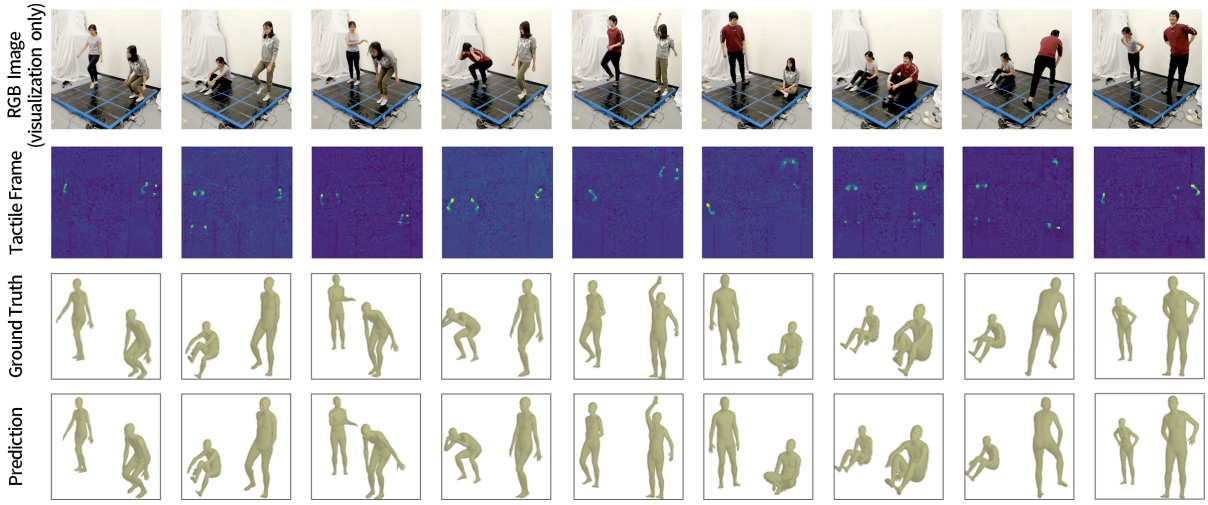

Fig. 9. Our model proficiently generates two concurrent human meshes of different activities.

leads to a marginal increase in errors. Secondly, the removal of joints (*w/o Joint*) or vertices (*w/o Vertex*) results in slightly elevated errors across both MPJPE and PVE metrics.

Furthermore, the impact of the discriminator becomes evident in the differences between *CAvatar* and *w/o Discriminator*, with the latter exhibiting higher errors. Notably, the absence of attention mechanisms (*w/o Attention*) also contributes to increased errors in MPJPE and PVE.

In summary, the quantitative results from the various model variants outlined in Table 6 illustrate the significance of individual components on the overall performance of our model. These insights underscore the importance of holistic integration of shape, pose, vertices, joints, discriminator, and attention mechanisms in achieving optimal accuracy and robustness in human mesh reconstruction.

5.7 Multi-person Scenarios

In scenarios where multiple individuals are present on the carpet at the same time, CAvatar effectively reconstructs the mesh of each person with remarkable precision, as demonstrated in Figure 9. The MPJPE value stands at 3.44 cm, the PVE is 3.94, and the translation measure is 1.95. We observed that the error rate for the dual-person dataset is, in fact, lower than that of the single-person dataset. This same phenomenon was observed in the original paper that contributed to the dataset [43]. We hypothesize that this may be due to the dual-person dataset being smaller in size, while the multi-person dataset spans a wider data distribution. Furthermore, the dual-person dataset is cleaner. We've noticed that the single-person dataset contains significant noise, such as people touching their hair, which are unlearned actions.

5.8 Different Sensing Resolution

The ablation study focusing on carpet sensing resolution (Table 7) highlights a clear relationship between increased resolution and enhanced accuracy in both MPJPE and PVE metrics. Specifically, the finest resolution of 96×96 consistently demonstrates superior results compared to its lower-resolution counterparts. This suggests that higher resolutions are pivotal for achieving precise measurements in such applications. Consequently, when determining the optimal sensing resolution for carpets or similar systems, one should prioritize the clear advantages brought about by finer resolutions in terms of measurement accuracy. Please note that in this context, the original dataset operates at a resolution of 96×96, while the fine-tuning dataset employs resolutions of 48×48 and 32×32.

Table 7. Different sensing resolution.

	96 x 96	**48 x 48**	**32 x 32**
MPJPE (cm)	5.8593	6.5824	6.6179
PVE (cm)	6.8475	7.5911	7.6517

6 LIMITATIONS AND FUTURE WORK

Our model exhibits varying performance when confronted with different types of unseen activities. It demonstrates strong generalization capabilities for poses with pressure maps similar to those encountered during training, but its performance deteriorates when faced with tactile imprints that deviate significantly from the training data. For instance, while the model can generalize well to different directions of the waist-twisting action, it struggles to accurately predict entirely new and different actions. To enhance the reliability of pose estimation in practical real-life deployments, it is crucial to conduct a more systematic data collection procedure that encompasses a broader range of typical human activities, enabling the model to better handle unseen actions and achieve more robust human mesh estimation performance.

Furthermore, our analysis is predominantly focused on scenarios that encompass the simultaneous presence of two individuals on the carpet. In situations where the carpet is occupied by a multitude of users concurrently, it is plausible that the efficacy of our model may be compromised. This introduces the necessity for more in-depth exploration, particularly in the field of human mesh reconstructions, when there is a greater number of individuals present on the carpet.

Although the pressure maps generated by the tactile carpet excel in effectively categorizing diverse human activities, it's crucial to recognize that the task of human mesh reconstruction differs significantly from a mere classification problem. Instead, it presents itself as a regression challenge, inherently possessing greater intricacy. This heightened complexity can result in situations where dissimilar hand movements produce comparable learning outcomes. In consideration of this, gaining insights into the model's confidence in faithfully reconstructing such akin activities in the future would be highly valuable. Also, the pressure distributions captured in the tactile maps may not be substantially influenced by the movements of certain body parts, such as the head, face, wrists, and fingers. This can introduce ambiguity in the resulting reconstructions, as these body parts may not exhibit clear pressure imprints. However, it is important to note that this trade-off in granularity serves a significant purpose in preserving privacy. By limiting the level of detail in the reconstructed data, individuals' privacy is safeguarded, ensuring that personal information related to specific body parts is not exposed.

Additionally, it's important to note that the VIBE model is not infallible and may not serve as a perfect ground truth. In future studies, it would be beneficial to explore alternative ground truths with known and reduced

errors, such as commercial optical or IMU-based motion trackers. Additionally, investigating different frames with larger step sizes should also be considered for a more comprehensive analysis.

Lastly, Various external factors can influence the sensor readings of an intelligent carpet. Temperature fluctuations can affect sensor material properties, while nearby vibrations can introduce noise into tactile readings. Humidity changes and moisture presence can alter electrical properties, especially in capacitive sensors, and electromagnetic interference from nearby devices can distort readings. The system may also suffer from wear and tear, dirt accumulation, static electricity build-up, power supply fluctuations, mechanical stresses, and installation inconsistencies. To address these challenges, it's crucial to incorporate robust sensor design, ensure proper installation, frequently calibrate the system, implement noise filtering algorithms, utilize protective layers against environmental factors, and conduct regular maintenance checks to maintain sensor accuracy and reliability over time. However, in our study, we utilized an existing dataset, so we couldn't directly investigate these noise sources. Our paper's primary contribution is reconstructing 3D human mesh of activities from carpet pressure data. Nonetheless, researching these noise factors will be crucial for future work.

7 CONCLUSION

In conclusion, we have introduced a novel, privacy-preserving approach for real-time human activity mesh reconstruction using tactile carpet signals. Our transformer-based neural network effectively predicts 3D human pose and shape from 2D pressure maps, overcoming limitations of existing vision-based and wearable solutions. Despite certain challenges, such as handling unseen actions and activities with minimal floor contact, the extensive evaluation showcases the model's robustness and generalization capabilities. This pioneering work not only highlights the potential of tactile carpet data for human activity monitoring and mesh reconstruction but also opens up opportunities for a variety of real-world applications, particularly in healthcare monitoring.

ACKNOWLEDGMENTS

With gratitude, we would like to express our appreciation to Yiyue Luo and Yunzhu Li for their invaluable technical guidance. This research was supported in part by the MIT-GIST joint research program and Wistron. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies of the funding agencies.

REFERENCES

[1] AH Akpa, Masashi Fujiwara, Hirohiko Suwa, Yutaka Arakawa, and Keiichi Yasumoto. 2019. A smart glove to track fitness exercises by reading hand palm. *Journal of Sensors* 2019 (2019).

[2] Giuseppe Amato, Davide Bacciu, Stefano Chessa, Mauro Dragone, Claudio Gallicchio, Claudio Gennaro, Hector Lozano, Alessio Micheli, Gregory MP O'Hare, Arantxa Renteria, et al. 2016. A benchmark dataset for human activity recognition and ambient assisted living. In *Ambient Intelligence-Software and Applications–7th International Symposium on Ambient Intelligence (ISAmI 2016)*. Springer, 1–9.

[3] Matteo Bastico, Alberto Belmonte-Hernández, and Federico Álvarez García. 2022. Continuous Person Identification and Tracking in Healthcare by Integrating Accelerometer Data and Deep Learning Filled 3D Skeletons. *IEEE Sensors Journal* 22, 15 (2022), 15402–15409.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[5] Shuo Chen, Yixuan Zhu, Shaogang Gong, and Yongxin Yang. 2019. Tactile sensing for dexterous in-hand manipulation in robotics: A review. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 417–424. https://doi.org/10.1109/IROS40897.2019.8968301

[6] Wenqiang Chen, Daniel Bevan, and John Stankovic. 2021. ViObject: A Smartwatch-based Object Recognition System via Vibrations. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. 97–99.

[7] Wenqiang Chen, Lin Chen, Yandao Huang, Xinyu Zhang, Lu Wang, Rukhsana Ruby, and Kaishun Wu. [n. d.]. Taprint: Secure text input for commodity smart wearables.

[8] Wenqiang Chen, Lin Chen, Yandao Huang, Xinyu Zhang, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2019. Taprint: Secure text input for commodity smart wristbands. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[9] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Shwetak Patel, and John Stankovic. 2021. ViFin: Harness Passive Vibration to Continuous Micro Finger Writing with a Commodity Smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[10] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Patel Shwetak, and John Stankovic. 2020. Continuous micro finger writing recognition with a commodity smartwatch: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 603–604.

[11] Wenqiang Chen, Lin Chen, Kenneth Wan, and John Stankovic. 2020. A smartwatch product provides on-body tapping gestures recognition: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 589–590.

[12] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2018. Vitype: A cost efficient on-body typing system through vibration. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[13] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2019. A Low Latency On-Body Typing System through Single Vibration Sensor. *IEEE Transactions on Mobile Computing* 19, 11 (2019), 2520–2532.

[14] Wenqiang Chen, Maoning Guan, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2017. FLoc: Device-free passive indoor localization in complex environments. In *2017 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[15] Wenqiang Chen, Yanming Lian, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2017. Virtual keyboard for wearable wristbands. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–2.

[16] Wenqiang Chen, Shupei Lin, Elizabeth Thompson, and John Stankovic. 2021. Sensecollect: We need efficient ways to collect on-body sensor-based human activity data! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[17] Wenqiang Chen and John Stankovic. 2022. ViWatch: harness vibrations for finger interactions with commodity smartwatches. In *Proceedings of the 13th ACM Wireless of the Students, by the Students, and for the Students Workshop*. 4–6.

[18] Wenqiang Chen, Ziqi Wang, Pengrui Quan, Zhencan Peng, Shupei Lin, Mani Srivastava, Wojciech Matusik, and John Stankovic. 2023. Robust Finger Interactions with COTS Smartwatches via Unsupervised Siamese Adaptation. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[19] Wenqiang Chen, Ziqi Wang, Pengrui Quan, Zhencan Peng, Shupei Lin, Mani Srivastava, and John Stankovic. 2022. Making Vibration-based On-body Interaction Robust. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 300–301.

[20] Xianda Chen, Yifei Xiao, Yeming Tang, Julio Fernandez-Mendoza, and Guohong Cao. 2021. Apneadetector: Detecting sleep apnea with smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–22.

[21] Jingyuan Cheng, Mathias Sundholm, Bo Zhou, Marco Hirsch, and Paul Lukowicz. 2016. Smart-surface: Large scale textile pressure sensors arrays for activity recognition. *Pervasive and Mobile Computing* 30 (2016), 97–112.

[22] Henry M Clever, Ariel Kapusta, Daehyung Park, Zackory Erickson, Yash Chitalia, and Charles C Kemp. 2018. 3d human pose estimation on a configurable bed from a pressure image. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 54–61.

[23] Diane J Cook, Miranda Strickland, and Maureen Schmitter-Edgecombe. 2022. Detecting Smartwatch-Based Behavior Change in Response to a Multi-Domain Brain Health Intervention. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 3 (2022), 1–18.

[24] Dropbox. 2023. Shared Dropbox Folder. https://www.dropbox.com/sh/5l0lm4po64xf6jd/AACuMt_oGy99Beyz_IMeknQ6a?dl=0 [Online; accessed 12-May-2023].

[25] Yegang Du, Yuto Lim, and Yasuo Tan. 2019. A novel human activity recognition and prediction in smart home based on interaction. *Sensors* 19, 20 (2019), 4474.

[26] Luigi D'Arco, Haiying Wang, and Huiru Zheng. 2022. Assessing impact of sensors and feature selection in smart-insole-based human activity recognition. *Methods and Protocols* 5, 3 (2022), 45.

[27] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2334–2343.

[28] Tobias Grosse-Puppendahl, Sebastian Herber, Raphael Wimmer, Frank Englert, Sebastian Beck, Julian Von Wilmsdorff, Reiner Wichert, and Arjan Kuijper. 2014. Capacitive near-field communication for ubiquitous interaction and perception. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 231–242.

[29] Maoning Guan, Wenqiang Chen, Yandao Huang, Rukhsana Ruby, and Kaishun Wu. 2019. FaceInput: a hand-free and secure text entry system through facial vibration. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[31] Yandao Huang, Wenqiang Chen, Hongjie Chen, Lu Wang, and Kaishun Wu. 2019. G-fall: device-free and training-free fall detection with geophones. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[32] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*. Springer, 34–50.

[33] Jeya Vikranth Jeyakumar, Ankur Sarker, Luis Antonio Garcia, and Mani Srivastava. 2023. X-CHAR: A Concept-based Explainable Complex Human Activity Recognition Model. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–28.
[34] Angjoo Kanazawa, Michael J Zhang, Philip Felsen, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7122–7131.
[35] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5253–5263.
[36] Nikos Kolotouros, Georgios Pavlakos, and Michael J Black. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2252–2261.
[37] Anil Kunchala, Mélanie Bouroche, Lorraine D'Arcy, and Bianca Schoen-Phelan. 2021. Smpl-based 3d pedestrian pose prediction. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–8.
[38] Nathan F Lepora. 2021. Soft biomimetic optical tactile sensing with the TacTip: A review. *IEEE Sensors Journal* 21, 19 (2021), 21131–21143.
[39] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. 2019. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 872–881.
[40] Youpeng Li, Xuyu Wang, and Lingling An. 2023. Hierarchical Clustering-based Personalized Federated Learning for Robust and Fair Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–38.
[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.
[42] Irvin Hussein Lopez-Nava and Angelica Munoz-Melendez. 2016. Wearable inertial sensors for human motion analysis: A review. *IEEE Sensors Journal* 16, 22 (2016), 7821–7834.
[43] Yiyue Luo, Yunzhu Li, Michael Foshey, Wan Shou, Pratyusha Sharma, Tomás Palacios, Antonio Torralba, and Wojciech Matusik. 2021. Intelligent carpet: Inferring 3d human pose from tactile signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11255–11265.
[44] Fen Miao, Yi He, Jinlei Liu, Ye Li, and Idowu Ayoola. 2015. Identifying typical physical activity on smartphone with varying positions and orientations. *Biomedical engineering online* 14 (2015), 1–15.
[45] Thomas B Moeslund and Erik Granum. 2001. A survey of computer vision-based human motion capture. *Computer vision and image understanding* 81, 3 (2001), 231–268.
[46] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in neural information processing systems*. 2277–2287.
[47] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*. Springer, 483–499.
[48] George A Oguntala, Raed A Abd-Alhameed, Nazar T Ali, Yim-Fun Hu, James M Noras, Nnabuike N Eya, Issa Elfergani, and Jonathan Rodriguez. 2019. SmartWall: Novel RFID-enabled ambient human activity recognition using machine learning for unobtrusive health monitoring. *IEEE Access* 7 (2019), 68022–68033.
[49] Venet Osmani, Sasitharan Balasubramaniam, and Dmitri Botvich. 2008. Human activity recognition in pervasive health-care: Supporting efficient remote collaboration. *Journal of network and computer applications* 31, 4 (2008), 628–655.
[50] Sarah Ostadabbas, Maziyar Baran Pouyan, Mehrdad Nourani, and Nasser Kehtarnavaz. 2014. In-bed posture classification and limb identification. In *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*. IEEE, 133–136.
[51] Todd C Pataky, Tingting Mu, Kerstin Bosch, Dieter Rosenbaum, and John Y Goulermas. 2012. Gait recognition: highly unique dynamic plantar pressure patterns among 104 individuals. *Journal of The Royal Society Interface* 9, 69 (2012), 790–800.
[52] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 459–468.
[53] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. 2020. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2669–2676.
[54] Tobias Peter, Simone Bexten, Veit Müller, Viola Hauffe, and Norbert Elkmann. 2020. Object classification on a high-resolution tactile floor for human-robot collaboration. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Vol. 1. IEEE, 1255–1258.
[55] Tomas Pfister, James Charles, and Andrew Zisserman. 2015. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 1913–1921.
[56] Cuong Pham, Nguyen Ngoc Diep, and Tu Minh Phuong. 2017. e-Shoes: Smart shoes for unobtrusive human activity recognition. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 269–274.
[57] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4929–4937.

[58] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2021. Winect: 3D human pose tracking for free-form activity using commodity WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–29.
[59] Christian Seeger, Alejandro Buchmann, and Kristof Van Laerhoven. 2012. myHealthAssistant: a phone-based body sensor network that captures the wearer's exercises throughout the day. In *6th International ICST Conference on Body Area Networks*.
[60] Qiongfeng Shi, Zixuan Zhang, Tianyiyi He, Zhongda Sun, Bingjie Wang, Yuqin Feng, Xuechuan Shan, Budiman Salam, and Chengkuo Lee. 2020. Deep learning enabled smart mats as a scalable floor monitoring system. *Nature communications* 11, 1 (2020), 4609.
[61] Monit Shah Singh, Vinaychandran Pondenkandath, Bo Zhou, Paul Lukowicz, and Marcus Liwickit. 2017. Transforming sensor data to the image domain for deep learning—An application to footstep detection. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2665–2672.
[62] Sarah M Stadig and Anna K Bergh. 2015. Gait and jump analysis in healthy cats using a pressure mat system. *Journal of Feline Medicine and Surgery* 17, 6 (2015), 523–529.
[63] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. 2018. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 529–545.
[64] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. 2019. Learning the signatures of the human grasp using a scalable tactile glove. *Nature* 569, 7758 (2019), 698–702.
[65] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*. 1799–1807.
[66] Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1653–1660.
[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[68] Lei Wang, Xiang Zhang, Yuanshuang Jiang, Yong Zhang, Chenren Xu, Ruiyang Gao, and Daqing Zhang. 2021. Watching your phone's back: Gesture recognition by sensing acoustical structure-borne propagation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–26.
[69] Yichao Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Wi-Mesh: A WiFi Vision-based Approach for 3D Human Mesh Construction. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 362–376.
[70] Irmandy Wicaksono, Don Derek Haddad, and Joseph Paradiso. 2022. Tapis Magique: Machine-knitted Electronic Textile Carpet for Interactive Choreomusical Performance and Immersive Environments. In *Creativity and Cognition*. 262–274.
[71] Kaishun Wu, Yandao Huang, Wenqiang Chen, Lin Chen, Xinyu Zhang, Lu Wang, and Rukhsana Ruby. 2020. Power saving and secure text input for commodity smart watches. *IEEE Transactions on Mobile Computing* 20, 6 (2020), 2281–2296.
[72] Te-Yen Wu, Lu Tan, Yuji Zhang, Teddy Seyed, and Xing-Dong Yang. 2020. Capacitivo: Contact-based object recognition on interactive fabrics using capacitive sensing. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 649–661.
[73] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. 2022. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 391–406.
[74] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 269–282.
[75] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2016. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3073–3082.
[76] Ping Yu, Weiting Liu, Chunxin Gu, Xiaoying Cheng, and Xin Fu. 2016. Flexible piezoelectric tactile sensor array for dynamic three-axis force measurement. *Sensors* 16, 6 (2016), 819.
[77] Shihao Zhang, Yang Lin, Cheng Lv, Yuzhe Guo, Yu-Kun Lai, and Nanning Zheng. 2020. Bodies at rest: 3D human pose and shape estimation from a pressure image using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1459–1469.
[78] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. 2019. Through-wall human mesh recovery using radio signals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10113–10122.
[79] Bo Zhou, Sungho Suh, Vitor Fortes Rey, Carlos Andres Velez Altamirano, and Paul Lukowicz. 2022. Quali-Mat: Evaluating the Quality of Execution in Body-Weight Exercises with a Pressure Sensitive Sports Mat. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–45.