Journal of the Optical Society of America A
OPTICS, IMAGE SCIENCE, AND VISION

Enhancing 3D human pose estimation with NIR single-pixel imaging and time-of-flight technology: a deep learning approach

Carlos Osorio Quero,[1,*] Daniel Durini,[1] Jose Rangel-Magdaleno,[1] Jose Martinez-Carranza,[2] and Ruben Ramos-Garcia[3]

[1]Electronics Department in the Digital Systems Group, Instituto Nacional de Astrofísica Óptica y Electrónica, 72810 Puebla, Mexico
[2]Computer Science Department, Instituto Nacional de Astrofísica Óptica y Electrónica, 72810 Puebla, Mexico
[3]Optics Department, Instituto Nacional de Astrofísica Óptica y Electrónica, 72810 Puebla, Mexico
*caoq@inaoep.mx

Received 6 July 2023; revised 17 December 2023; accepted 9 January 2024; posted 18 January 2024; published 13 February 2024

The extraction of 3D human pose and body shape details from a single monocular image is a significant challenge in computer vision. Traditional methods use RGB images, but these are constrained by varying lighting and occlusions. However, cutting-edge developments in imaging technologies have introduced new techniques such as single-pixel imaging (SPI) that can surmount these hurdles. In the near-infrared spectrum, SPI demonstrates impressive capabilities in capturing a 3D human pose. This wavelength can penetrate clothing and is less influenced by lighting variations than visible light, thus providing a reliable means to accurately capture body shape and pose data, even in difficult settings. In this work, we explore the use of an SPI camera operating in the NIR with time-of-flight (TOF) at bands 850–1550 nm as a solution to detect humans in nighttime environments. The proposed system uses the vision transformers (ViT) model to detect and extract the characteristic features of humans for integration over a 3D body model SMPL-X through 3D body shape regression using deep learning. To evaluate the efficacy of NIR-SPI 3D image reconstruction, we constructed a laboratory scenario that simulates nighttime conditions, enabling us to test the feasibility of employing NIR-SPI as a vision sensor in outdoor environments. By assessing the results obtained from this setup, we aim to demonstrate the potential of NIR-SPI as an effective tool to detect humans in nighttime scenarios and capture their accurate 3D body pose and shape.

https://doi.org/10.1364/JOSAA.499933

1. INTRODUCTION

The process of 3D reconstruction finds extensive applications in various domains, such as human animation [1], human motion recognition [2], augmented reality [3], and virtual reality [4]. However, it presents a formidable challenge to obtain a complete 3D model of the human body from just a single 2D image due to the inherent ill-posed nature of the problem. This is because different 3D locations can have identical projections on the 2D image plane, resulting in ambiguity and difficulty in accurately reconstructing the 3D human body. However, advancements in computational techniques and algorithms are constantly improving the accuracy and robustness of 3D reconstruction methods, paving the way for exciting applications in various fields such as medical [5], computer imaging (CI) [6,7], biomedical [8], games [9], and robotics [10].

Reconstructing an accurate human shape from imperfect input data, accounting for nonrigid deformations and joint articulations, is a challenging task. Recent advances in deep learning techniques have made it possible to achieve end-to-end reconstruction of a human shape [2–11]. However, directly learning a high-dimensional mesh with articulations, such as the 3D human mesh (e.g., with 6890 vertices [12]) remains extremely difficult. Previous approaches using deep neural networks for 3D human reconstruction have produced results that are either rugged [13], blurred [14], or distorted [15]. Fortunately, the skinned multi-person linear model (SMPL) [16] and SMPL eXpressive (SMPL-X) [8] offer compact representations for the 3D human shape and have been integrated with deep neural networks for 3D human reconstruction from RGB images [17]. The typical pipeline involves using deep neural networks to extract powerful image features, followed by direct regression of SMPL shape and pose parameters [17,18].

Currently, various solutions exist for estimating a human pose using different technologies, such as RGB cameras [17], thermal cameras [19], and IR ultra-wideband (UWB) radar [20]. While thermal infrared cameras are commonly used for object detection in low-illumination situations and provide better information for objects with higher temperatures, they have

poor information for objects with lower temperatures [21]. In the case of RGB cameras, they are sensitive to low-illumination scenarios, and a solution for detecting a human pose is to use the camera in the near-infrared (NIR); however, they are expensive. Single-pixel imaging (SPI) systems [22,23] offer a promising solution to the limitations of conventional and thermal cameras in low-light conditions [24]. SPI systems capture images by measuring the light reflected from an object through a single-pixel detector, allowing them to operate in spectral bands such as the infrared spectrum. By exploiting the power of deep learning techniques [25,26], SPI systems can reconstruct high-quality images from sparse measurements, making them an ideal candidate for detecting 3D human poses in nighttime surveillance applications.

A major benefit of SPI systems compared to conventional cameras includes several aspects [22]: (i) Cost-effectiveness: SPI systems are more affordable; (ii) Spectral range flexibility: These devices can operate across a wide range of wavelengths; (iii) Simplicity and robustness: SPI systems offer a straightforward operation and durable design; (iv) High dynamic range: They are capable of capturing images with a vast range of brightness levels; (v) Noise resistance: SPI systems are less prone to image distortion caused by noise; (vi) High-resolution imaging potential: Advanced deep learning techniques enable SPI systems to produce high-resolution images; and (vii) Versatile imaging modalities: an SPI system's capability to capture images in the near-infrared (NIR) spectrum enhances its versatility [27]. NIR imaging provides better visibility in low-light conditions, making it a valuable tool for object detection and tracking in surveillance applications. By combining SPC technology with time-of-flight (TOF) sensing, we can obtain 2D/3D images of the environment, providing additional information about the location and movement of objects [27]. The use of SPI systems in surveillance applications is not limited to nighttime environments. They can also be used to capture images in harsh environments where traditional cameras may fail, such as in dusty or foggy conditions [27,28].

In this work, we propose an SPI vision system with active illumination in the NIR wavelength range of 850–1500 nm, which can be employed using single InGaAs photodetectors [29]. As a strategy for detection, we remove the background of the SPI image by applying a U2Net [30] to identify the object for segmentation of the area of interest containing the element to detect. We then apply the vision transformers (ViT) model [31] to perform silhouette analysis-based gait recognition for human identification [32]. Information will be used to generate a 3D model through the Video Inference for Body Pose and Shape Estimation method (VIBE) [33]. VIBE predicts SMPL-X [8] body model parameters using a convolutional neural network (CNN) pretrained on the AMASS dataset [34] for single-image body pose and shape estimation.

Therefore, in this work, we propose:

- Exploring the capacity of SPI for the generation of a 3D human pose from a 2D low-resolution image;
- 2D human action recognition from silhouette SPI applying the ViT model; and
- addressing a new and challenging task dealing with the prediction of a 3D hand pose from a single 2D binary mask obtained from NIR-SPI imaging.

2. RELATED WORK

A. 2D Human Action Recognition

Extensive research has been conducted in computer vision to study the recognition of human activities [35]. Currently, various techniques for action representation are available, using single-view and multiview recognition methods [36]. These integrate different technologies such as camera mono-stereo [37], radar [38], and lidar [39]. Single-view human action recognition is often studied using three types of features: holistic [40], local features [41], and geometric human body features [41]. Holistic methods use shape or motion-based information [42]; shape-based methods are insensitive to the color, texture, and luminance of a person's clothing, making them ideal for action representation [43]. Motion-based approaches may face challenges such as motion discontinuities, low-quality videos, and background variations [44]. Geometric human body features involve identifying body parts and movements. Local space–time features or interest points describe these features efficiently with a feature descriptor. Single-view approaches require the same or a similar camera view for training and testing [45].

Multicamera view-invariant action recognition has become a popular research topic in the last decade [46]. Multiview approaches are classified into two categories: 3D and 2D multiview methods. In 3D methods, 2D human body silhouettes are joined to obtain a "3D human body pose" representation [47]. These methods typically necessitate a fixed multicamera setup during training and testing. On the other hand, 2D multiview methods propose various types of directions to overcome limitations through the integration of different cameras to compensate points of the scene to determine the direction and cross-view action recognition [48].

Numerous algorithms and systems have been proposed for human action recognition in the literature, proposing two general approaches based on deep learning using CNN [35]. The first approach involves compressing an individual's binary silhouettes of a one gait cycle into a single compact gait representation, called a gait energy image (GEI) [49]. This approach uses a single image as the gait features representation [49]. The second approach considers the gait as a sequence of silhouettes of an individual that are individually used as input for a feature extractor [50]. CNNs have dominated the field of image-based deep learning and have become the standard backbone network used in approaches tackling gait recognition and classification [51] and predicting a body pose from an image [32].

In recent years, the ViT architecture has emerged as a direct competitor to CNNs in the field of image classification [31]. ViT has shown excellent results on many image classification benchmarks, demonstrating their strong generalization capability. Compared to CNNs, the ViT model demands fewer computational resources to train and have a stronger modeling capability, making them ideal for low-memory computing systems.

B. 3D Pose and Shape from a Single Image

Human pose estimation commonly relies on parametric 3D models of human bodies [52], as they can capture human shape statistics and provide a 3D mesh for various tasks [53]. Early work explored different approaches using keypoints and silhouettes as input [54], including "bottom-up" regression [55], "top-down" optimization [56], and multicamera settings [57]. However, these methods were found to be fragile, requiring manual intervention and failing to generalize well to images in natural settings.

The SMPLify model [5] was one of the first end-to-end approaches that fit the SMPL [16] and SMPL-X [8] model to the output of a CNN keypoint detector. Recently, deep neural networks have been trained to directly regress the parameters of the SMPL-X body model from pixels [33]. However, due to the lack of 3D ground-truth labels, these methods use weak supervision signals obtained from a 2D keypoint reprojection loss, body/part segmentation, or human input.

Other models combined regression-based and optimization-based methods by using SMPLify in the training loop [58]. In addition, several nonparametric body mesh reconstruction methods have also been proposed [59], including using voxels as the output body representation [14], directly regressing the vertex locations of a template body mesh using graph convolutional networks, and predicting body shapes using pixel-aligned implicit functions followed by a mesh reconstruction step [60].

C. Bodies, Faces, Hands, and Unified Models SMPL-X

Previous approaches have focused on separate parts of the body and using statistical shape models learned from 3D scans [53]. The FLAME model [61] is unique in that it models the whole head, including 3D head rotations and the neck region, and is critical for connecting the head and the body [13]. However, none of these methods model correlations between the face shape and body shape [62]. Similarly, hand modeling approaches typically ignore the body and rely on non-learned, artist-designed models [63].

The unified model, SMPL-X [8], combines the SMPL + H body model [64] with the FLAME head model. Unlike previous methods that simply graft models together, the authors fit the full model to 3D scans with 6890 vertices and learn the shape and pose-dependent blend shapes [12]. This results in a natural-looking model with a consistent parameterization that is differentiable and easy to integrate into applications that use SMPL [16]. Overall, the SMPL-X model offers a more comprehensive and realistic approach to modeling human bodies, faces, and hands. By modeling correlations between these different parts, the model can better capture natural expressions and movements.

3. SPI RECONSTRUCTION

The SPI technique [24] is used to reconstruct images by measuring the correlated intensity on a detector without spatial resolution. The SPI camera utilizes spatial light modulators (SLMs) such as digital micromirror devices (DMDs) to produce spatially structured light patterns (Hadamard-like patterns) for interrogation of a scene. The SPI camera can operate in two architectures: structured detection and structured illumination (Fig. 1).

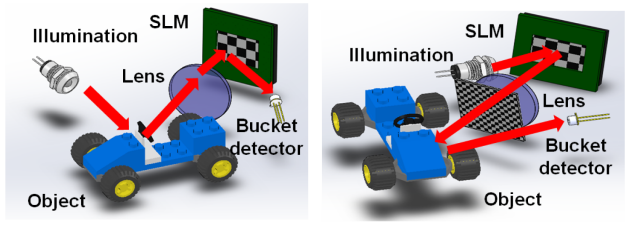

Fig. 1. Two different approaches applied to SPI: (a) structured detection and (b) structured illumination [24].

In structured detection, the object is illuminated by a light source, and the reflected light is projected onto an SLM, followed by detection using a bucket detector. In contrast, in structured illumination Φ, the light source is spatially modulated by the SLM, illuminating the object O, and the reflected light is detected by bucket detector is converted in electrical signal y_i by [24]

$$y_i = \alpha \sum_{i=1}^{M} \sum_{j=1}^{N} O(i, j)\Phi(i, j), \tag{1}$$

where α is a constant factor that depends on the optoelectronic response of the photodetector, the correlation of the light spatial pattern and the reflected light from the object when captured by the photodetector produces an electrical signal. Therefore, projecting a sequence of spatial patterns allows a sequence of electrical signals to be obtained, which can be used to reconstruct the image computationally. In this regard, the image x_i is reconstructed from the captured signal y_i and the corresponding pattern Φ using [24]

$$x_i = \alpha \sum_{i=1}^{M} \sum_{j=1}^{N} y_i \Phi(i, j). \tag{2}$$

To generate Hadamard-like patterns Φ using active illumination, an array of 32×32 NIR-LEDs emitting radiation with a peak wavelength of 1550 nm is used in this work. The choice of wavelength is due to the reduced scattering by water and the reduced absorption coefficient of water. The NIR-LED array is placed perpendicular to the focal length of the lens to project the light pattern to an infinite. However, given the size of the array, the patterns are projected up to a distance of 0.3–3 m. Although the object is not completely illuminated in active illumination, the technique of fast super-resolution CNN (FSRCNN) can be used to reconstruct images with good quality [65]. The active illumination approach offers several advantages, such as operating in different outdoor weather conditions, low-level illumination scenarios, and being less sensitive to background radiation noise. Additionally, the proposed configuration requires fewer optical elements and lower costs, and the modulation rate can be much higher because there are no moving parts involved.

A. SPI Camera

In our research, we propose the utilization of structured illumination to improve the quality of images captured under challenging lighting conditions, including strong backlight and stray light. To achieve this, we employ a time-of-flight (ToF) system with a wavelength of 850 nm and an InGaAs photodiode as the bucket detector operating at a wavelength of 1550 nm.

The architecture we introduce in this study is called NIR-SPI, which consists of two main components. Firstly, we utilize fundamental elements based on the single-pixel principle to generate images. These elements include an InGaAs photodetector (specifically, the Thorlabs FGA015 diode operating at 1550 nm), an array of NIR-LEDs for emission, a ToF system, and an analog-to-digital converter (ADC). This setup is illustrated in Fig. 2(a).

Second, we incorporate a subsystem responsible for processing the electrical output signal obtained from the bucket detector. The signal is digitized using the ADC, and the resulting data is then processed using an embedded system-on-module (SOM) [66], specifically the GPU-Jetson Xavier NX depicted in Fig. 2(a). The SOM performs multiple tasks, including generating Hadamard-like patterns and processing the digitized data from the ADC. The orthogonal matching pursuit GPU (OMP-GPU) algorithm [67] is implemented on the SOM, enabling the generation of 2D images. The processing time for each stage involved in the 2D image reconstruction process is also presented. For further details on the SPI camera, refer to [27].

B. 2D Reconstruction Algorithm

We initiated the process by acquiring and converting the electrical signal y_i using an ADC. This involved applying the Hadamard matrix projection to the signal, resulting in a vector of signals y_i [Eq. (1)]. Subsequently, we utilized the OMP algorithm (Algorithm 1) to extract the image x_i [Eq. (2)]. Our objective was to solve the equation $|y_i - \Phi(i, j)x_i| < \varepsilon$ [24]. To improve the efficiency of reconstructing the 2D SPI image, we employed the Cholesky method for matrix inversion as

Algorithm 1. OMP-GPU algorithm [67], Input: OMP-GPU algorithm input data: Patterns Φ, input signal y_i, target sparsity K, Output: OMP-GPU algorithm output data: sparse representation x_i that fulfills the relation $y_i \approx \Phi x_i$

1: **procedure** OMP-GPU (Φ, y_i, K):
2: set: $L_1 = [1]$, $i = 1$, $p^0 = \Phi^T y_i$
3: set: $\varepsilon = y_i y_i^T$, $G_i = \Phi^T \Phi$, $p = p^0$
4: **while** $\varepsilon_{i-1} > \epsilon$ **do**
5: $k = \arg\max_K |p|$ ▷ Finding the new atom
6: **if** $i > 1$ **then**
7: $w_i = \{L_{i-1} w_i = G_{i-1,K}\}$ ▷ Solver w_i
8: $L_i = \begin{bmatrix} L_{i-1} & 0 \\ w_i^T & \sqrt{1 - w_i^T w_i} \end{bmatrix}$ ▷ Update of Cholesky
10: $x_i = \{L_i L_i^T x_i = p^o\}$ ▷ Solver x_i
11: $\beta = G_i x_i$ ▷ Matrix-sparse-vector product for each path
12: $p = p^o - \beta$
13: $\delta^k = x_i^T \beta$ ▷ Calculate error
14: $\varepsilon^k = \varepsilon^{k-1} - \delta^k + \delta^{k-1}$ ▷ Calculate norm ε
16: $i = i + 1$ ▷ increasing iteration
17: **return** x_i

defined in [68,69]. For this method, it was necessary to precalculate the symmetric and positive Gram matrix, defined as $G_i = \Phi^T \Phi$ [67]. Additionally, we carried out an initial projection $p^0 = \Phi^T y_i$ (Algorithm 1, line 3). This projection was performed to facilitate the implementation of the Cholesky method to get

$$L_{\text{new}} = \begin{bmatrix} L & 0 \\ w^T & \sqrt{1 - w^T w} \end{bmatrix}. \tag{3}$$

The matrix G can be decomposed into two triangular matrices using Cholesky decomposition, represented as $L_i L_i^T$ [Eq. (3)]. Here, L_i is a triangular Cholesky factor [70] (Algorithm 1, line 8). To solve this matrix, we define a system $L_i L_i^T x_i = \Phi^T y_i$. This system can be solved by treating it as a triangular system, where we express the system in the form $L_i u = b$ with $b = \Phi y_i$ and $L_i^T x_i = u$ (Algorithm 1, line 10). The matrix L_i can be calculated using the formulation

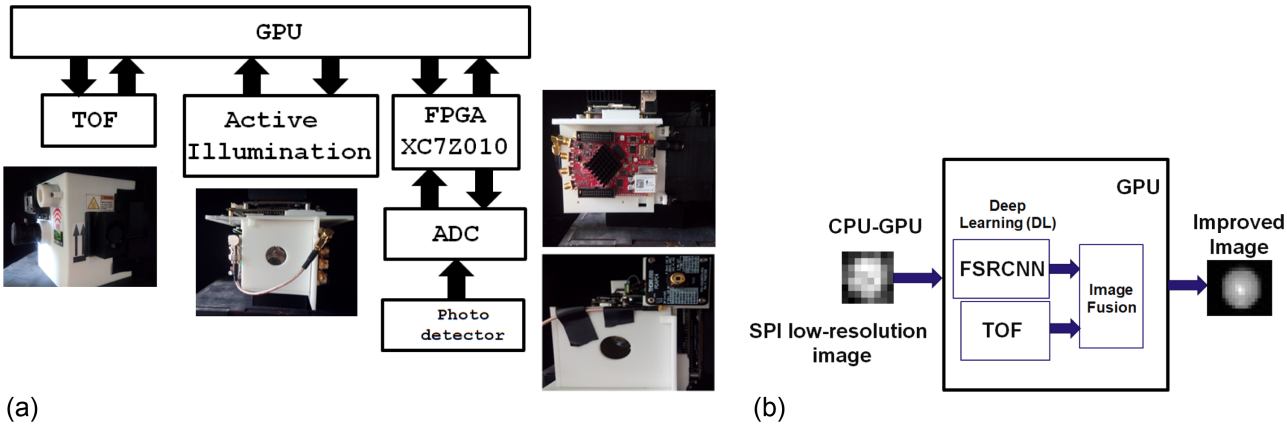

Fig. 2. Proposed vision system's overall block diagram has dimensions of 11 cm × 11 cm × 14 cm. It comprises several components, including a lens with a focal length of 20 cm for projecting active illumination patterns. The system weighs 1.2 kg and consumes 45 W of power. (a) In the first stage module, there are three key elements: a photodiode, an active illumination source, and an InGaAs photodetector diode (FGA015) used for the ToF system, as described in the 3DSPI reference [27]. (b) The second stage incorporates a GPU unit and an ADC. The processing unit utilizes an FSRCNN network to enhance the low-resolution SPI images and combines them with the ToF information captured.

in Eq. (3) [67], where $w_i = L_i^{-1} G_i$ (Algorithm 1, line 7). To obtain the reconstructed signal x_i, which contains the vector image reconstruction and needs to undergo a reshape operation to convert it into an $N \times N$ matrix, we define a stopping criterion to compare the norm of the residual with a threshold ε (Algorithm 1, line 14), eliminating the need to calculate the residual δ (Algorithm 1, lines 11–13). To enhance the efficiency of the algorithm, we propose implementing it on compute unified device architecture (CUDA) to parallelize the reconstruction operation [66,71] (Algorithm 1).

To produce the final 2D image, we combine the SPI image obtained through the Algorithm 1 with post-processed depth information from a ToF system. To enhance the depth data, we utilize a normalization technique. The initial input image is first fused with data from the ToF system using the FSRCNN network method, as described in [27]. This fusion process results in an enhanced image with four times the original resolution. Consequently, we achieve a high-resolution image with dimensions of 64×64 pixels as the final output.

The overall block diagram is shown in Fig. 2, with Fig. 2(a) representing the proposed vision system and Fig. 2(b) showing the processing algorithm used by the proposed NIR-SPI vision system, which takes a low-resolution SPI image, applies an FSRCNN network [27], and fuses it with information captured by ToF system.

4. HUMAN MODELING

The use of parametric human models, such as SMPL-X [8–72], allows for a concise representation of human shapes by utilizing shape and pose parameters to encode variations [6]. The SMPL-X model (Fig. 3) offers various advantages:

- It disentangles the human shape and pose, allowing for independent analysis and control of each human shape [13,14,33–73];
- It avoids modeling rugged and twisted shapes directly, which can pose difficulties for neural network-based methods [13,14,60,61], by utilizing a skinning process to model deformation; and
- It is differentiable and can be easily integrated with neural networks [73]. For this research, we used SMPL-X as the underlying representation for modeling 3D humans.

The SMPL-X model comprises shape parameters β, and pose parameters $\theta \in \mathbb{R}^{3K}$. The body pose is defined by a skeleton rig with $K = 24$ joints including the body root (define the vector positions from 0 to 23 with points of reference over the model SMPL-X human pose [74] 0: Pelvis, 1: L_{Hip}, 2: R_{Hip}, 3: Spine 1, 4: L_{Knee}, 5: R_{Knee}, 6: Spine 2, 7: L_{Ankle}, 8: R_{Ankle}, 9: Spine 3, 10: L_{Foot}, 11: R_{Foot}, 12: Neck, 13: L_{Collar}, 14: R_{Collar}, 15: Head,16: L_{Shoulder}, 17: R_{Shoulder}, 18: L_{Elbow}, 19: R_{Elbow}, 20: L_{Wrist}, 21: R_{Wrist}, 22: L_{Hand}, 23: R_{Hand}), and global translation parameters. Shape parameters ($\beta \in \mathbb{R}^{10}$) are utilized for shape blending and encoding global shape information. Pose parameters are used for pose blending and skinning and encode local information between adjacent joints, with the exception of the root joint's pose parameters, which denote the global rotation of the entire shape. It should be noted that SMPL-X's pose parameters denote the relative rotation from a joint to its parent, which differs from 2D or 3D human pose estimation [75], where the pose refers to joint locations. With β and θ, we can obtain the 3D body mesh $M = f_{\text{SMPL}}(\beta, \theta)$, where $M \in \mathbb{R}^{N \times 3}$ is a triangulated surface with $N = 6890$. We can predict the 3D SMPL-X model locations of the body joints X with the body mesh using a pretrained mapping matrix $W \in \mathbb{R}^{K \times N}$, $X \in \mathbb{R}^{k \times 3} = WM$ [76]. From the 3D human joints and the perspective camera model to project the body joints from 3D to 2D. Assuming the camera parameters are $\delta \in \mathbb{R}^3$, which define the 3D translation of the camera, the 2D keypoints can be defined as $J \in \mathbb{R}^2 = f_{\text{project}(X,\delta)}$ [7–77].

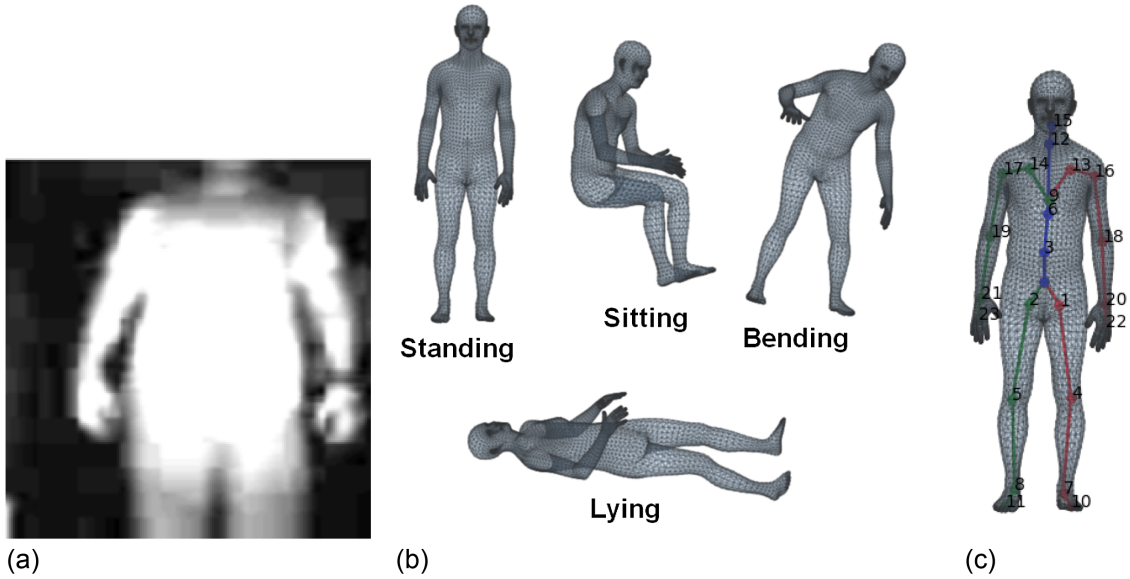

Fig. 3. Human poses, but with same joint positions generated from NIR-SPI imaging: (a) Test NIR-SPI imaging, (b) SMPL-X model generated based on estimation pose (standing, sitting, bending, and lying), (c) SMPL-X model generated with joints.

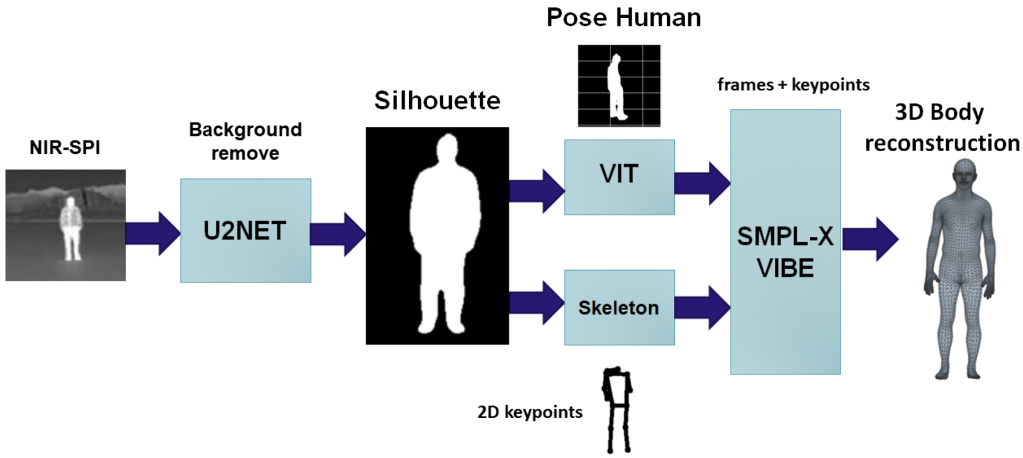

Fig. 4. Overview of the proposed network architecture, which takes NIR single-pixel imaging input and outputs 3D body reconstruction based on SMPL-X shape and pose parameters. The entire network consists of three main modules: (i) NIR-SPI-based image acquisition. (ii) Feature extraction using deep learning: To extract the background, the NIR-SPI image is used to obtain the silhouette. (iii) 3D pose estimation using a regression-based approach: The silhouette image is used to obtain the gait features (shape estimation), which are then used to pose the human using ViT and skeleton joint features. These features are used to pre-define the pose SMPL-X model; from the pre-defined parameters (pose θ, shape β and camera s, R, T), the SMPL-X model is fed to the off-the-shelf SMPL-X model to obtain the reconstructed 3D human mesh.

5. PROPOSED METHOD

The process used to obtain the 3D human model from NIR-SPI is shown in Fig. 4. It involves several steps that use different computer vision techniques to reconstruct a 3D human pose from a single low-resolution image. Here is a detailed explanation of each step:

- Take a single pixel low-resolution image [see Fig. 6(a)]. This step involves capturing an image of a human. The image contrast is adjusted to extract the basic shape of the person, and then the background is removed using U2Net [30], a deep learning model that can accurately segment the foreground and background of an image. Thus, to isolate the person from the background, an image segmentation technique is used (Fig. 4) to obtain the image's silhouette. This image only shows the outline of the person without any details of the surface or texture.
- Applied over the silhouette image, ViT can identify four human poses: lying, bending, sitting, and standing (Fig. 5).

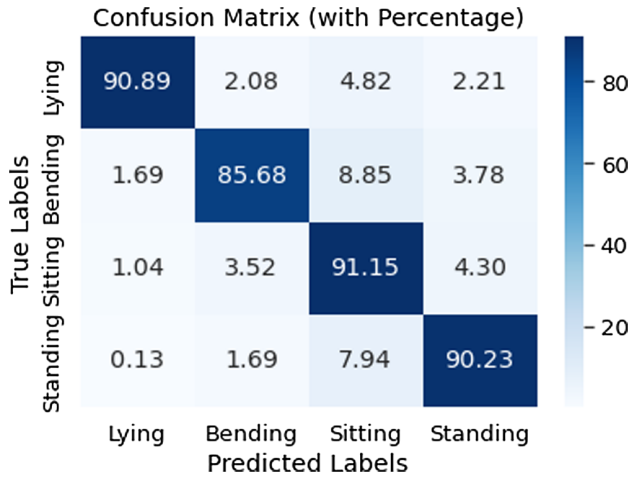

Fig. 5. Confusion matrix ViT identification of human pose: lying, bending, sitting, and standing.

Once the pose is identified, it can be used to generate a 3D human pose using the VIBE method, a deep learning model that can estimate the 3D pose of a human from a single image or video.

- Finally, we can reconstruct the human body shape and pose in 3D space (Fig. 3). This can be done using a tool such as SMPL-X, as discussed above.

6. EXPERIMENTAL RESULTS

The experimental results show the process to obtain a 3D human model from NIR-SPI imaging at a distance of 1 m from an SPI camera and nighttime condition illumination using the proposed method (Fig. 4). The process involves several steps that use different deep learning models such as U2Net, ViT, and VIBE to extract and estimate the 3D pose of a human from a single low-resolution image.

To evaluate the effectiveness of the proposed method, we conducted experiments on a set of datasets. For the ViT transform classification, we used the following datasets: silhouette-based 3D human pose estimation [78], silhouette for human posture recognition [79] the OU-ISIR gait database [80], and the human pose SMPL-X dataset: AMASS dataset [34]) of single-pixel low-resolution images of humans. The results show that the U2Net method can accurately remove the background and extract the silhouette of a person [Fig. 6(b)]. The ViT model can successfully detect four different poses of a person from a silhouette image, and the VIBE model can accurately estimate the 3D pose of the person from the identified pose (Table 1). Using SMPL-X, the researchers were able to reconstruct the 3D shape and pose of the person in space [Fig. 6(c)].

A. Discussion: Proposed Method

For evaluation of the proposed network architecture (Fig. 4), we tested different NIR-SPI-based image reconstruction

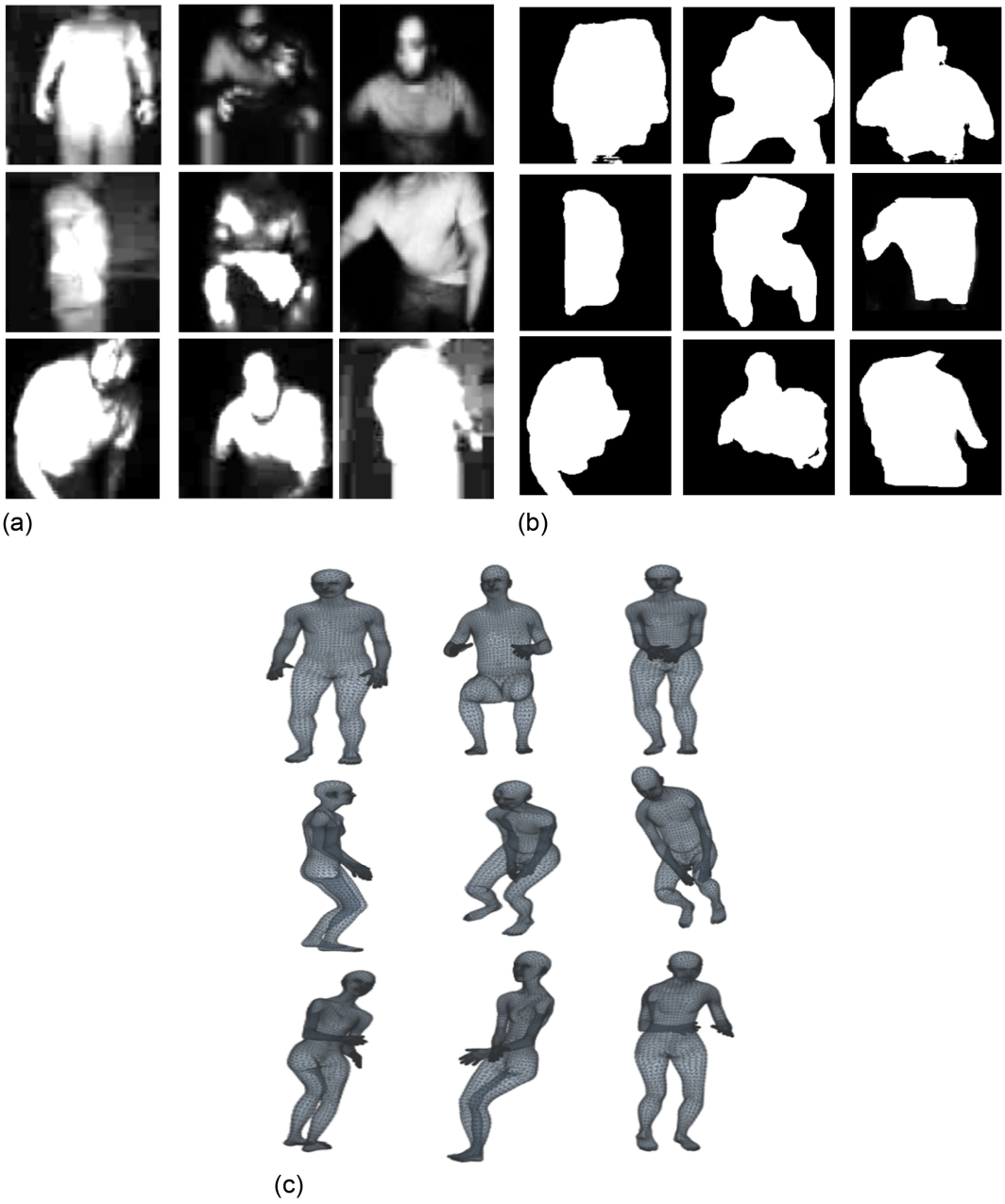

Fig. 6. Capture human poses imaging at a distance of 1 m: (a) Capture NIR-SPI imaging human pose standing, sitting, and bending, (b) silhouette image, and (c) 3D human pose regression based on SMPL-X model (see Visualization 1).

Table 1. Mean Vertex-to-Vertex (V2V) Results [81] and Mean Per-Joint Position Error (MPJPE) [82] Body for Different Human Positions

Human Pose	V2V Error (mm) ↓	MPJPE (mm) Error ↓
Lying	57.29	53.2
Bending	49.86	40.19
Sitting	34.2	33.7
Standing	42	41

approaches using the SMPL-*X* model to reconstruct human positions at nighttime from a distance of 1 m while taking into consideration the limited field of view of the SPC camera, which is 74° × 57°. We captured NIR-SPI images of the human poses, including sitting, standing, bending, and lying. We observed some limitations in the hand and body positions with respect to the reference image, particularly in the bending position, due to a loss of information in the input NIR-SPI image resulting from reflection effects and low resolution in the reconstructed NIR-SPI image. However, for the standing and sitting positions (as shown in Fig. 5), the 3D human reconstruction exhibited better accuracy in terms of vertex and joint positions, as presented in Table 1. Compared to other models of 3D pose estimation (Table 2), this model can achieve acceptable accuracy in 3D pose estimation from low-resolution images, as measured by MPJPE. This is in contrast to other methods that require high-quality images for accurate pose estimation.

Table 2. Various Methods to Estimate 3D Human Poses from Monocular Images[a]

Method	MPJPE (mm)	Complexity	Performance
VIBE [33]	65.6	Uses a combination of DL and optimization techniques	Offers higher accuracy in dynamic scenarios
DenseRaC [83]	76.8	Significantly complex	Provides detailed and accurate 3D reconstructions
HoloPose [84]	60.2	Aims at high-fidelity 3D human pose estimation	High-quality results may not be ideal for real-time
GCMR [59]	71.9	Employs graph convolutional networks for mesh regression	Performance can be computationally intensive
HMR [73]	87.9	Balances between complexity and efficiency	Good trade-off between accuracy and speed
UP [85]	80.7	Complex depending on the implementation	Offers reliable performance
SMPLify [8]	82.3	Optimization-based approach to fit the SMPL body model	Offers reasonable accuracy in controlled environments
Ours	42	Implementation a strategy of ViT and VIBE methods	3D reconstructions from low-resolution images

[a]These include VIBE (Video Inference for Body Pose and Shape Estimation), DenseRaC (Dense Reconstruction of Articulated Characters), HoloPose, GCMR (Graph Convolutional Mesh Regression), HMR (Human Mesh Recovery), and UP (Unite the People). Additionally, the SMPLify algorithm has been considered, alongside the other methods we have proposed.

7. CONCLUSION

The proposed methods to obtain a 3D human model from NIR-SPI imaging, for human poses such as lying, bending, sitting, and standing, were determined by applying ViT transforms classification (Fig. 5). The best accuracy was achieved in the sitting position, with an accuracy of around 91%, as shown in the V2V and MPJPE error table (Table 1). The results demonstrate the effectiveness of the proposed approach, with limitations in hand positioning due to the low contrast of the NIR-SPI image. However, the level position of the core person detection shows an accurate estimation of the 3D pose of the person through qualitative and quantitative evaluations (Table 1). These findings highlight the potential of the proposed approach for 3D human modeling from a single low-resolution image (Fig. 6).

In comparison, the presented SMPL-X model captures the body, face, and hands jointly, and the SMPL-X approach fits the model to a single NIR-SPI image and 2D joint detections. The results of this work demonstrate the expressivity of SMPL-X in capturing bodies, hands, and faces from NIR-SPI images. However, we observed that the bending and lying pose presented the highest level of V2V and MPJPE error, indicating limitations in the pose parameters θ. Therefore, it is recommended to implement a compensation model in future applications. Future work may involve the development of a dataset of in-the-wild SMPL-X fits and the direct regression of SMPL-X parameters from NIR-SPI images.

This work marks a significant advancement in the expressive capture of bodies, hands, and faces using NIR-SPI imaging. Compared to other methods, our proposed approach provides efficient 3D reconstruction of poses, as shown in Table 2, even in low-resolution scenarios. This is particularly advantageous in environments with low illumination, demonstrating the robustness and practicality of our technique.

Funding. National Council for Science and Technology—CONACyT (251992).

Acknowledgment. The first author is thankful to Consejo Nacional de Ciencia y Tecnología (CONACYT) for his scholarship with No. CVU: 661331.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

REFERENCES

1. C. Johnson, J. Seidel, R. Carson, *et al.*, "Evaluation of 3D reconstruction algorithms for a small animal pet camera," in *IEEE Nuclear Science Symposium* (1996), Vol. **3**, pp. 1481–1485.
2. L. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1453–1459 (2000).
3. P. Sudhaman, P. Surendiren, S. Meeran, *et al.*, "Augmented reality in automation using virtual 3D models," in *3rd International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (2012), pp. 1–4.
4. D. Ram, B. Roy, and V. Soni, "A review on virtual reality for 3D virtual trial room," in *IEEE World Conference on Applied Intelligence and Computing (AIC)* (2022), pp. 247–251.
5. C. Prahm, K. Bauer, A. Sturma, *et al.*, "3D body image perception and pain visualization tool for upper limb amputees," in *IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)* (2019), pp. 1–5.
6. C.-H. P. Huang, H. Yi, M. Höschle, *et al.*, "Capturing and inferring dense full-body human-scene contact," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 13274–13285.
7. Z. Zheng, T. Yu, Y. Liu, *et al.*, "Pamir: parametric model-conditioned implicit representation for image-based human reconstruction," IEEE Trans. Pattern Anal. Mach. Intell. **44**, 3170–3184 (2022).
8. G. Pavlakos, V. Choutas, N. Ghorbani, *et al.*, "Expressive body capture: 3D hands, face, and body from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10967–10977.
9. A. Feng, S. Shin, and Y. Yoon, "A tool for extracting 3D avatar-ready gesture animations from monocular videos," in *15th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Association for Computing Machinery, 2022).

10. Y. Qin, H. Su, and X. Wang, "From one hand to multiple hands: imitation learning for dexterous manipulation from single-camera teleoperation," IEEE J. Robot. Autom. Lett. **7**, 10873–10881 (2022).
11. S. S. Jinka, R. Chacko, A. Sharma, *et al.*, "PeeledHuman: robust shape representation for textured 3d human body reconstruction," in *International Conference on 3D Vision (3DV)* (2020), pp. 879–888.
12. G. Pons-Moll, J. Romero, N. Mahmood, *et al.*, "Dyna: a model of dynamic human shape in motion," ACM Trans. Graph. **34**, 120 (2015).
13. O. Litany, A. Bronstein, M. Bronstein, *et al.*, "Deformable shape completion with graph convolutional autoencoders," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 1886–1895.
14. G. Varol, D. Ceylan, B. Russell, *et al.*, "BodyNet: volumetric inference of 3D human body shapes," in *European Conference on Computer Vision*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. (Springer, 2018), pp. 20–38.
15. T. Groueix, M. Fisher, V. G. Kim, *et al.*, "3D-coded: 3D correspondences by deep deformation," in *Computer Vision–ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. (Springer International Publishing, 2018), pp. 235–251.
16. M. Loper, N. Mahmood, J. Romero, *et al.*, "SMPL: a skinned multi-person linear model," ACM Trans. Graph. **34**, 248 (2015).
17. D. Chen, Y. Song, F. Liang, *et al.*, "3D human body reconstruction based on SMPL model," Vis. Comput. **39**, 1893–1906 (2023).
18. S. Zhang and N. Xiao, "Detailed 3D human body reconstruction from a single image based on mesh deformation," IEEE Access **9**, 8595–8603 (2021).
19. H. M. Clever, Z. Erickson, A. Kapusta, *et al.*, "Bodies at rest: 3D human pose and shape estimation from a pressure image using synthetic data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6214–6223.
20. G. W. Kim, S. W. Lee, H. Y. Son, *et al.*, "A study on 3D human pose estimation using through-wall IR-UWB radar and transformer," IEEE Access **11**, 15082–15095 (2023).
21. A. Bañuls, A. Mandow, R. Vázquez-Martín, *et al.*, "Object detection from thermal infrared and visible light cameras in search and rescue scenes," in *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)* (2020), pp. 380–386.
22. C. A. O. Quero, D. Durini, J. d. J. Rangel-Magdaleno, *et al.*, "2D NIR-SPI spatial resolution evaluation under scattering condition," in *19th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)* (2022), pp. 1–6.
23. C. O. Quero, D. Durini, J. Rangel-Magdaleno, *et al.*, "Deep-learning blurring correction of images obtained from NIR single-pixel imaging," J. Opt. Soc. Am. A **40**, 1491–1499 (2023).
24. C. A. Osorio Quero, D. Durini, J. Rangel-Magdaleno, *et al.*, "Single-pixel imaging: an overview of different methods to be used for 3D space reconstruction in harsh environments," Rev. Sci. Instrum. **92**, 111501 (2021).
25. F. Wang, C. Wang, C. Deng, *et al.*, "Single-pixel imaging using physics enhanced deep learning," Photon. Res. **10**, 104–110 (2022).
26. G. M. Gibson, S. D. Johnson, and M. J. Padgett, "Single-pixel imaging 12 years on: a review," Opt. Express **28**, 28190–28208 (2020).
27. C. Osorio Quero, D. Durini, J. Rangel-Magdaleno, *et al.*, "Single-pixel near-infrared 3D image reconstruction in outdoor conditions," Micromachines **13**, 795 (2022).
28. C. O. Quero, D. D. Romero, J. Rangel-Magdaleno, *et al.*, "2D/3D single-pixel NIR image reconstruction method for outdoor applications in presence of rain," Proc. SPIE **11914**, 1191415 (2021).
29. R. Lange, S. Böhmer, and B. Buxbaum, "11 - CMOS-based optical time-of-flight 3D imaging and ranging," in *High Performance Silicon Imaging*, D. Durini, ed., 2nd ed., Woodhead Publishing Series in Electronic and Optical Materials (Woodhead, 2020), pp. 319–375.
30. X. Qin, Z. Zhang, C. Huang, *et al.*, "U^2-Net: going deeper with nested U-structure for salient object detection," Pattern Recogn. **106**, 107404 (2020).
31. H. Wu, B. Xiao, N. Codella, *et al.*, "CVT: introducing convolutions to vision transformers," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 22–31.
32. L. Wang, T. Tan, H. Ning, *et al.*, "Silhouette analysis-based gait recognition for human identification," IEEE Trans. Pattern Anal. Mach. Intell. **25**, 1505–1518 (2003).
33. M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: video inference for human body pose and shape estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 5252–5262.
34. N. Mahmood, N. Ghorbani, N. F. Troje, *et al.*, "AMASS: archive of motion capture as surface shapes," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 5441–5450.
35. M. Gholamrezaii and S. M. Taghi Almodarresi, "Human activity recognition using 2D convolutional neural networks," in *27th Iranian Conference on Electrical Engineering (ICEE)* (2019), pp. 1682–1686.
36. T. Xu, P. Peng, X. Fang, *et al.*, "Single and multiple view detection, tracking and video analysis in crowded environments," in *IEEE 9th International Conference on Advanced Video and Signal-based Surveillance* (2012), pp. 494–499.
37. H. Liu, Y. Cao, and Z. Wang, "A novel algorithm of gait recognition," in *International Conference on Wireless Communications & Signal Processing* (2009), pp. 1–5.
38. Z. Chen, G. Li, F. Fioranelli, *et al.*, "Personnel recognition and gait classification based on multistatic micro-doppler signatures using deep convolutional neural networks," IEEE Geosci. Remote Sens. Lett. **15**, 669–673 (2018).
39. S. Yoon, H.-W. Jung, H. Jung, *et al.*, "Development and validation of 2D-LiDAR-based gait analysis instrument and algorithm," Sensors **21**, 414 (2021).
40. A. Castelli, G. Paolini, A. Cereatti, *et al.*, "A 2D markerless gait analysis methodology: validation on healthy subjects," Comput. Math. Methods Med. **2015**, 186780 (2015).
41. Y.-F. Tsao, W.-T. Liu, and C.-T. Chiu, "Human gait analysis by body segmentation and center of gravity," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (2013), pp. 1–5.
42. H. Su and F.-G. Huang, "Human gait recognition based on motion analysis," in *International Conference on Machine Learning and Cybernetics* (2005), Vol. **7**, pp. 4464–4468.
43. T. Yeoh, H. E. Aguirre, and K. Tanaka, "Clothing-invariant gait recognition using convolutional neural network," in *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)* (2016), pp. 1–5.
44. N. S. Razali and A. A. Manaf, "Gait recognition using motion capture data," in *8th International Conference on Informatics and Systems (INFOS)* (2012), pp. MM–67–MM–71.
45. D. Muramatsu, A. Shiraishi, Y. Makihara, *et al.*, "Gait-based person recognition using arbitrary view transformation model," IEEE Trans. Image Process. **24**, 140–154 (2015).
46. D. Guffanti, A. Brunete, and M. Hernando, "Non-invasive multi-camera gait analysis system and its application to gender classification," IEEE Access **8**, 95734–95746 (2020).
47. G. Zhao, G. Liu, H. Li, *et al.*, "3D gait recognition using multiple cameras," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)* (2006), pp. 529–534.
48. Y. J. Qi, Y. P. Kong, and Q. Zhang, "A cross-view gait recognition method using two-way similarity learning," Math. Probl. Eng. **2022**, 2674425 (2022).
49. L. Yao, W. Kusakunniran, Q. Wu, *et al.*, "Robust CNN-based gait verification and identification using skeleton gait energy image," in *Digital Image Computing: Techniques and Applications (DICTA)* (2018), pp. 1–7.
50. A. M. Saleh and T. Hamoud, "Analysis and best parameters selection for person recognition based on gait model using CNN algorithm and image augmentation," J. Big Data **8**, 1 (2021).
51. P. P. Min, S. Sayeed, and T. S. Ong, "Gait recognition using deep convolutional features," in *7th International Conference on Information and Communication Technology (ICoICT)* (2019), pp. 1–5.
52. Y. Tian, H. Zhang, Y. Liu, *et al.*, "Recovering 3D human mesh from monocular images: a survey," IEEE Trans. Pattern Anal. Mach. Intell. **45**, 15406–15425 (2023).
53. G. Pons-Moll and B. Rosenhahn, "Model-based pose estimation," in *Visual Analysis of Humans* (Springer, 2011), pp. 139–170.

54. D. Anguelov, P. Srinivasan, D. Koller, *et al.*, "SCAPE: shape completion and animation of people," ACM Trans. Graph. **24**, 408–416 (2005).
55. A. Zanfir, E. Marinoiu, M. Zanfir, *et al.*, "Deep network for the integrated 3D sensing of multiple people in natural images," in *32nd International Conference on Neural Information Processing Systems* (Curran Associates, 2018), pp. 8420–8429.
56. W. Jiang, N. Kolotouros, G. Pavlakos, *et al.*, "Coherent reconstruction of multiple humans from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 5578–5587.
57. Y. Zheng, R. Shao, Y. Zhang, *et al.*, "DeepMultiCap: performance capture of multiple characters using sparse multiview cameras," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 6219–6229.
58. V. Choutas, L. Müller, C. P. Huang, *et al.*, "Accurate 3D body shape regression using metric and semantic attributes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 2708–2718.
59. N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4496–4505.
60. S. Saito, Z. Huang, R. Natsume, *et al.*, "PIFu: pixel-aligned implicit function for high-resolution clothed human digitization," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 2304–2314.
61. T. Li, T. Bolkart, M. J. Black, *et al.*, "Learning a model of facial shape and expression from 4D scans," ACM Trans. Graph. **36**, 194 (2017).
62. E. A. Clark, J. Kessinger, S. E. Duncan, *et al.*, "The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review," Front. Psychol. **11**, 920 (2020).
63. C. Cao, Y. Weng, S. Zhou, *et al.*, "FaceWarehouse: a 3D facial expression database for visual computing," IEEE Trans. Vis. Comput. Graph. **20**, 413–425 (2014).
64. J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: modeling and capturing hands and bodies together," ACM Trans. Graph. **36**, 245 (2017).
65. C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. (Springer, 2016), pp. 391–407.
66. P. Kang and S. Lim, "A taste of scientific computing on the GPU-accelerated edge device," IEEE Access **8**, 208337–208347 (2020).
67. C. O. Quero, D. Durini, R. Ramos-Garcia, *et al.*, "Hardware parallel architecture proposed to accelerate the orthogonal matching pursuit compressive sensing reconstruction," Proc. SPIE **11396**, 56–63 (2020).
68. B. L. Sturm and M. G. Christensen, "Comparison of orthogonal matching pursuit implementations," in *20th European Signal Processing Conference (EUSIPCO)* (2012), pp. 220–224.
69. J. Chen and Z. Chen, "Cholesky factorization on heterogeneous CPU and GPU systems," in *9th International Conference on Frontier of Computer Science and Technology* (2015), pp. 19–26.
70. R. Zheng, W. Wang, H. Jin, *et al.*, "GPU-based multifrontal optimizing method in sparse Cholesky factorization," in *IEEE 26th International Conference on Application-specific Systems, Architectures and Processors (ASAP)* (2015), pp. 90–97.
71. H. Choi and J. Lee, "Efficient use of GPU memory for large-scale deep learning model training," Appl. Sci. **11**, 10377 (2021).
72. Y. Feng, V. Choutas, T. Bolkart, *et al.*, "Collaborative regression of expressive bodies using moderation," in *International Conference on 3D Vision (3DV)* (IEEE Computer Society, 2021), pp. 792–804.
73. A. Kanazawa, M. J. Black, D. W. Jacobs, *et al.*, "End-to-end recovery of human shape and pose," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7122–7131.
74. H. Zhang, Y. Tian, X. Zhou, *et al.*, "PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 11426–11436.
75. H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3D human pose estimation," in *European Conference on Computer vision (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. (Springer, 2018), pp. 765–782.
76. X. Xu, H. Chen, F. Moreno-Noguer, *et al.*, "3D human shape and pose from a single low-resolution image with self-supervised learning," in *European Conference on Computer vision (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. (Springer, 2020), pp. 284–300.
77. R. Hartley and A. Zisserman, "Two-view geometry," in *Multiple View Geometry in Computer Vision*, 2nd ed. (Cambridge University, 2004), pp. 237–238.
78. R. Hori, R. Hachiuma, H. Saito, *et al.*, "Silhouette-based synthetic data generation for 3D human pose estimation with a single wrist-mounted 360° camera," in *IEEE International Conference on Image Processing (ICIP)* (2021), pp. 1304–1308.
79. W. Ding, B. Hu, H. Liu, *et al.*, "Human posture recognition based on multiple features and rule learning," Int. J. Mach. Learn. Cybern. **11**, 2529–2540 (2020).
80. C. Xu, Y. Makihara, G. Ogi, *et al.*, "Real-time rendering of aerial perspective effect based on turbidity estimation," IPSJ Trans. Comput. Vis. Appl. **9**, 1 (2017).
81. J. Y. Chang, G. Moon, and K. M. Lee, "V2V-PoseNet: voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 5079–5088.
82. W. Bao, Z. Ma, D. Liang, *et al.*, "Pose ResNet: a 3D human pose estimation network model," in *2nd International Conference on Big Data, Information and Computer Network (BDICN)* (2023), pp. 264–267.
83. Y. Xu, S.-C. Zhu, and T. Tung, "DenseRaC: joint 3D pose and shape estimation by dense render-and-compare," in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 7759–7769.
84. R. A. Güler and I. Kokkinos, "HoloPose: holistic 3D human reconstruction in-the-wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10876–10886.
85. C. Lassner, J. Romero, M. Kiefel, *et al.*, "Unite the people: closing the loop between 3D and 2D human representations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, 2017), pp. 4704–4713.