MiShape: Accurate Human Silhouettes and Body Joints from Commodity Millimeter-Wave Devices

AAKRITI ADHIKARI, University of South Carolina, USA
HEM REGMI, University of South Carolina, USA
SANJIB SUR, University of South Carolina, USA
SRIHARI NELAKUDITI, University of South Carolina, USA

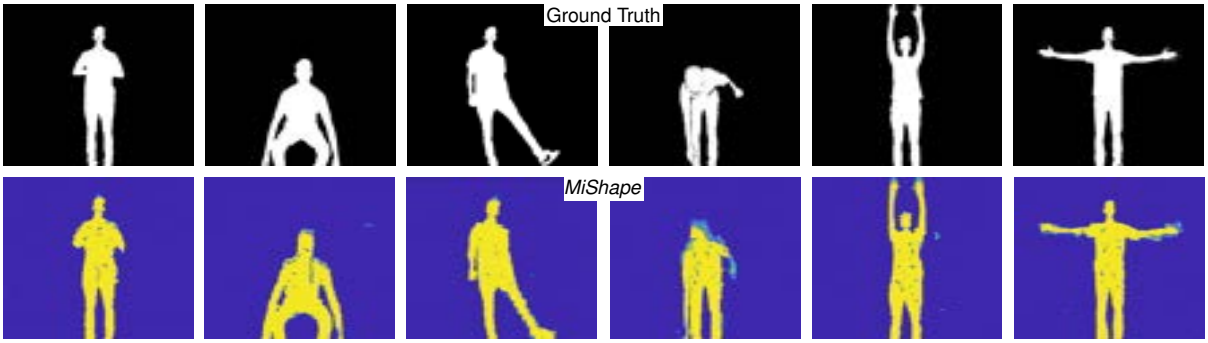


Figure 1. *MiShape* generates high-resolution silhouettes similar to vision-based systems using only millimeter-wave signals.

We propose *MiShape*, a millimeter-wave (mmWave) wireless signal based imaging system that generates high-resolution human silhouettes and predicts 3D locations of body joints. The system can capture human motions in real-time under low light and low-visibility conditions. Unlike existing vision-based motion capture systems, *MiShape* is privacy non-invasive and can generalize to a wide range of motion tracking applications at-home. To overcome the challenges with low-resolution, specularity, and aliasing in images from Commercial-Off-The-Shelf (COTS) mmWave systems, *MiShape* designs deep learning models based on conditional Generative Adversarial Networks and incorporates the rules of human biomechanics. We have customized *MiShape* for gait monitoring, but the model is well adaptive to any tracking applications with limited fine-tuning samples. We experimentally evaluate *MiShape* with real data collected from a COTS mmWave system for 10 volunteers, with diverse ages, gender, height, and somatotype, performing different poses. Our experimental results demonstrate that *MiShape* delivers high-resolution silhouettes and accurate body poses on par with an existing vision-based system, and unlocks the potential of mmWave systems, such as 5G home wireless routers, for privacy-noninvasive healthcare applications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Machine learning approaches**.

Additional Key Words and Phrases: 5G; Millimeter-Wave; Aliasing; Generative Adversarial Networks; Joint Prediction

Authors' addresses: Aakriti Adhikari, aakritit@email.sc.edu, University of South Carolina, USA; Hem Regmi, hregmi@email.sc.edu, University of South Carolina, USA; Sanjib Sur, sur@cse.sc.edu, University of South Carolina, USA; Srihari Nelakuditi, srihari@sc.edu, University of South Carolina, USA.

2474-9567/2022/9-ART96 $15.00
https://doi.org/10.1145/3550300

ACM Reference Format:
Aakriti Adhikari, Hem Regmi, Sanjib Sur, and Srihari Nelakuditi. 2022. MiShape: Accurate Human Silhouettes and Body Joints from Commodity Millimeter-Wave Devices . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 96 (September 2022), 31 pages. https://doi.org/10.1145/3550300

1 INTRODUCTION

The need for understanding and perceiving at-home human activities is critical for numerous applications, such as monitoring behavior of elderly patients in assisted living conditions or detecting falls [1, 2], monitoring recovery of post-surgery or post-stroke patients [3], monitoring infants [4], and monitoring early biomarkers of critical health conditions [5–7]. Optical cameras, IRs, LiDARs, *etc.*, can be used to design such applications [8–12], but they do not perform well in low light and low visibility conditions [13]. Furthermore, the scope of these systems is also limited by additional hardware requirements, such as LiDARs, which are expensive and cumbersome to instrument inside an average home. More importantly, cameras capture the scene through a true-color image representation, making them privacy-invasive and undesirable to the users to implement in their private space, like home or office. Fortunately, ubiquitous networking devices, such as 5G home wireless routers [14], can be augmented with intelligence to enable such at-home monitoring applications without being intrusive. These networking devices have built-in millimeter-wave (mmWave) technology, and they can illuminate the target scene and capture its image using high-frequency mmWave wireless signals [15]. The mmWave imaging systems are robust to low light and low visibility conditions, can enable through-occlusion imaging, and have been widely used in airports and security portals to detect hidden weapons [16, 17]. Furthermore, they provide an advantage over the camera-based systems by capturing only the general body shape or silhouette and preserving users' privacy [15]. So, *the ubiquity of mmWave technology in 5G-and-beyond devices, such as home wireless routers, enables the opportunity for bringing privacy non-invasive human motion capture systems to the masses at-home.*

Wireless signal based motion capture systems traditionally rely on low-frequency signal reflections from Wi-Fi devices to capture at-home human activities [18–21]. These systems deliver information about the human body in the form of skeletons/joints at a coarse-grain scale. Even though skeletal representation is adequate for tracking the movement of a person, they lack meaningful, discriminatory information, such as somatotype, contour of the body, *etc.*, on par with the existing vision-based systems, *e.g.*, RGB cameras, Vicon, or Kinect [22–24]. So, it might result in an identical skeleton for two different somatotypes of the same height, which is undesirable in human monitoring applications. Significant research efforts have been directed towards extracting meaningful information in the form of fine-grained mesh or silhouette from mmWave signals [25, 26]. This is because the smaller wavelength of mmWave signals theoretically allows us to capture a target scene with a higher resolution than Wi-Fi and represent the human body at a fine-grain scale. With such a fine-grained representation, privacy non-invasive information about the human body, comparable to the vision-based system, can be derived. But it is challenging to design mmWave imaging on networking devices for three key reasons.

First, mmWave imaging resolution is still very low compared to optical cameras. Imaging resolution of a system is proportional to the antenna array size and signal bandwidth [27]; so, a practical mmWave networking device, with less than 4 GHz bandwidth and less than 5 cm $\times$ 5 cm of antenna size [28], creates a significant image pixel spread along the horizontal and vertical directions. This results in the elimination of high-frequency components, such as the contour, limbs, and joints, and the final output looks like blobs, making it human or machine imperceptible (see Figure 3[c]). *Second*, mmWave reflections suffer from the challenges of specularity and variable reflectivity. Since human body mainly absorbs mmWave signals [29, 30], most of the signals transmitted do not reach back to the mmWave receiver, and some of the body parts can only create specular reflections [25]. Furthermore, due to the presence of clothing, different body parts would reflect signals differently, creating an imperceptible human shape with many missing parts. *Finally*, existing wireless systems require special antenna arrangements, such as bulky T-shaped or rectilinear, or uniform antenna arrays [25, 31, 32] that are unavailable

in Commercial-Off-The-Shelf (COTS) devices at homes. Deploying traditional mmWave imaging algorithms on a COTS device with non-uniform antenna arrangements is challenging because of the image aliasing effect, where aliasing refers to the presence of spurious/ghost reflections and distortion of the target scene [33–37].

To overcome these challenges, we propose MiShape, a low-barrier system that brings high-resolution mmWave imaging on COTS devices for at-home continuous monitoring of human motions and activities. MiShape relies on the reflected signals from multiple mmWave antennas on a COTS device to estimate human silhouettes accurately. But combining the reflections using traditional imaging algorithms produces low-resolution, distorted human shapes with many missing parts [38, 39]. This is because traditional algorithms reconstruct each pixel on the target scene independently without the knowledge of the well-defined shape and biomechanics of humans. To this end, *MiShape* designs a deep learning framework to learn the representation of mmWave reflections to the human shapes and generate high-resolution silhouette images by identifying patterns from several examples. But instead of trying to learn thousands of pixels in high-resolution silhouettes from only a few points in the reflected signals, which could lead to a network divergence during learning, *MiShape* divides the learning task into three networks. *First*, it designs a customized conditional Generative Adversarial Network (cGAN) to learn very low-resolution silhouettes consisting of hundreds of pixels from the input reflected signals. *Then*, it converts the low-resolution silhouettes to high-resolution using a Super-Resolution Generative Adversarial Network (SRGAN), that is customized for human shapes. *Finally*, from the generated silhouettes, it predicts accurate 3D locations of joints by incorporating well-established rules of human biomechanics.

We design and prototype *MiShape* on a COTS mmWave device and conduct microbenchmark and field-trial experiments for at-home application. Since the current COTS mmWave networking devices cannot switch between Tx and Rx mode within nanoseconds and do not provide user access to the raw signal reflections yet, we built a customized setup using two 77–81 GHz mmWave Radars [40] to collect the reflected signals and a Microsoft Kinect Xbox One [41] to collect the ground truth silhouettes and 3D joint locations. We collect reflected signals and ground truths from 10 volunteers performing 17 different poses, spanning over a period of two months, and our dataset consists of nearly 100 K samples (> 14 GB). The dataset is used to not only train and fine-tune *MiShape* but also benchmark its effectiveness. Our baseline experiments with a single volunteer's data show that *MiShape* can generate human silhouette images with a median Intersection of Union (IoU) and Multi-Scale Structural Similarity Index Measure (MS-SSIM) of 0.72 and 0.96, respectively, where IoU [42] and MS-SSIM [43] measure the shape and quality similarity between generated and ground truth images. In contrast, traditional imaging algorithms can only achieve a median IoU and MS-SSIM of 0.06 and 0.11, respectively. Furthermore, *MiShape* can upsample the silhouettes to generate high-resolution images, which accentuates the fine-grained texture and the frame of the human body, and still maintains a median MS-SSIM of 0.91 for 8× upsampling. *MiShape* also performs well across diverse mmWave antenna configurations with fundamentally different resolutions in their captured samples and consistently outperforms an existing deep learning model. In addition, *MiShape* predicts the 3D locations of joints with an average error of ∼ 10 cm across many critical body joints. We also find that *MiShape* requires little fine-tuning for new volunteers: The model, when fine-tuned with only 2 randomly selected volunteers' data samples, can achieve a median IoU of 0.60 across all 10 subjects. Finally, our field-trial experiments for at-home gait monitoring application show that *MiShape* consistently predicts its key metrics with accuracy similar to the RGB-D camera-based systems.

In summary, we make the following contributions: (1) We design a deep learning based imaging framework for generating high-resolution human silhouettes by overcoming the challenges in COTS devices. To the best of our knowledge, *MiShape* is the first system to address the fundamental aliasing problem, and achieve high-resolution silhouettes on par with the existing vision-based systems with COTS mmWave device. (2) We design a framework based on generative models and rules of human biomechanics to enable at-home healthcare applications, and prototype and evaluate its performance for multiple volunteers with diverse ages, gender, height, and somatotype.

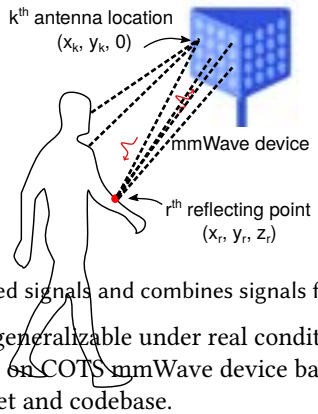


Figure 2. A mmWave device captures reflected signals and combines signals from multiple antennas to generate an image.

Our results demonstrate that *MiShape* is generalizable under real conditions and work consistently with practical COTS devices. To accelerate the research on COTS mmWave device based human imaging and motion tracking, we will open-source our measured dataset and codebase.

2 BACKGROUND AND FUNDAMENTALS

2.1 Millimeter-Wave Imaging

To construct an image, a transceiver illuminates the target scene by radiating mmWave signal from its planar antennas. The signal then bounces off of the target and the background to reach back to the transceiver. By combining the received signals from multiple antennas, the transceiver can generate an image. If the combination is coherent, a human's shape could be identified against the background by estimating the reflection strengths at different spatial points. Let us consider that a human body has R reflecting points, and the location and reflectivity coefficient of r^{th} point is (x_r, y_r, z_r) and σ_r, respectively. The mmWave transceiver sends a wide bandwidth signal from each antenna location $(x_k, y_k, 0)$ and receives reflection from these R points (Figure 2). The received signal of k^{th} antenna can be expressed as a sum of all time-delayed transmitted signals, $p[t]$, and can be written as: $S(x_k, y_k, t) = \sum_{r \in R} \sigma_r \cdot p[t - 2d_{rk}/c]$, where c is the wireless propagation speed (~$3{\times}10^8$ m/s) and $2d_{rk}$ is the round trip distance between the r^{th} point and the k^{th} antenna [44]. To generate the image, *first*, a Fast Fourier Transform (FFT) is applied to the received signal to estimate the intensity and phase at different depths. *Then*, these values are interpolated and compressed *w.r.t.* the mean depth, and a 3D image grid is created along the XYZ dimensions. *Finally*, for each depth bin and for each antenna, the reflected signals are combined and projected onto the image grid to generate the 3D voxel of the target scene [38]. A 2D silhouette image can be generated by identifying a high-intensity cluster inside the voxel and extracting the slice with maximum energy across depth.

2.2 Challenges in Imaging with COTS Millimeter-Wave Devices

Achieving high-resolution images from COTS mmWave networking devices, such as home wireless routers, is challenging because of two reasons.

(1) Small Antenna Array Size: The resolution of mmWave images depends on the antenna array size, and larger antenna arrays, with each antenna placed at the correct location, achieve higher resolution since it can better distinguish close-by reflecting points [27, 44]. However, to be cost-effective, the array size for most mmWave networking devices are very small, typically, 4×4 or 2×4 or 2×8 [1] [45–49]. So, the resolution of the generated images will be extremely poor and the resultant shapes are often imperceptible by both humans and machines. For example, Figure 3 shows the silhouette image of a human posing in front of two mmWave antenna arrays: While a rough shape could be visually perceivable with a 32×32 array, the shape appears like a blob for the 2×8 array. Such a shape also lacks enough discriminating features for simple automation tasks, *e.g.*, recognizing or distinguishing human poses.

[1] N×M represents the number of antennas across vertical and horizontal directions.

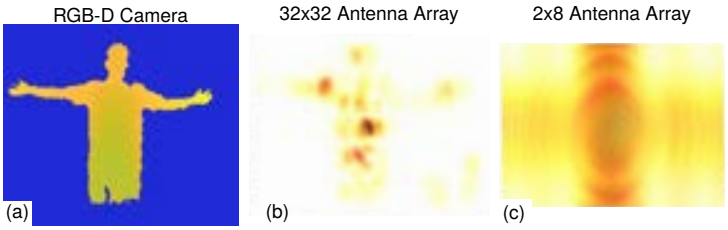


Figure 3. (a) Silhouette image of a human performing a pose in front an RGB-D camera. (b) Silhouette image generated for this pose with rectilinear mmWave antenna array size of 32×32. (c) Silhouette image generated by a more practical and widely available mmWave antenna array size of 2×8.

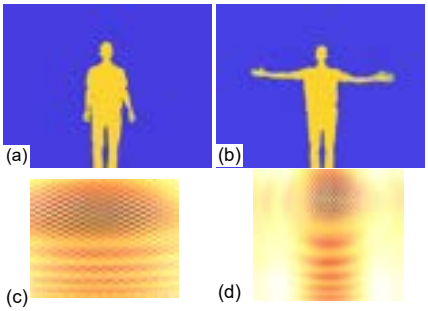


Figure 4. (a–b) Two poses. (c–d) Aliased images generated for these poses.

(2) Non-Uniform Antenna Placements: While the image resolutions can be improved by deploying a larger antenna array, each antenna must be placed by strictly adhering to the Nyquist criterion for alias-free imaging [50] (which states that the critical distance between adjacent antenna elements should be $\sim \lambda/2$, where λ is the signal wavelength). Combining the reflected signals from multiple non-uniformly spaced antennas often leads to image aliasing, distorting the target scene and creating spurious/ghost reflections [34]. This is because the reflected signals are combined incoherently, adding destructively where it should be constructive, and vice versa. While the next-generation of expensive mmWave devices for outdoor networking applications, such as [51, 52], promises to include more antennas, they will distribute the antenna arrays non-uniformly across the device to improve the network coverage. So, they will likely produce aliased images with unrecognizable human shapes. Figures 4(a–d) show that for two different poses, it generates similar-looking indistinguishable silhouette images when two antenna arrays of size 1×4 are placed non-uniformly, which are difficult to perceive or distinguish.

3 *MISHAPE* DESIGN

3.1 Overview

MiShape aims to generate high-resolution human silhouette images and predict accurate 3D joint locations by addressing the practical challenges in COTS mmWave devices. This could enable many at-home healthcare applications, such as physiotherapy and gait monitoring, from ubiquitous mmWave networking devices. Instead of relying on traditional imaging algorithms, *MiShape* trains a set of customized deep learning frameworks with thousands of examples of mmWave signal reflections, ground truth human silhouettes, and 3D joint locations to learn the generalized relationships between them. Then, at run-time, when the model has been trained appropriately, *MiShape* can accurately predict the silhouettes and joints from only the mmWave signal reflections without using the ground truth. Figure 5 shows an overview of the *MiShape* system.

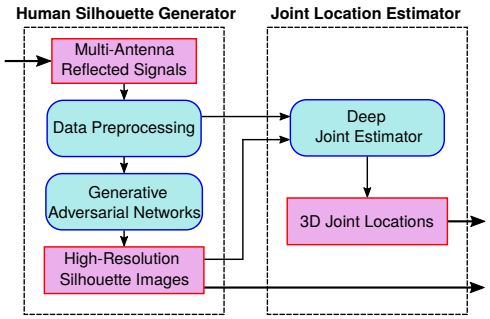


Figure 5. System overview of *MiShape*.

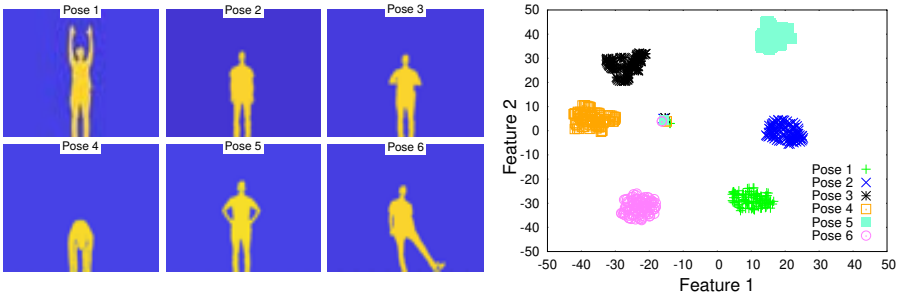
To train *MiShape* with the ground truths, *first*, we collect a diverse dataset with different human poses, activities, *etc.*, which records the signal reflections from mmWave devices and the corresponding ground truth silhouette images and joint locations from a co-located RGB-D camera. The silhouettes and joints are then sanitized, and the reflected signals are synchronized and resampled in time to align with the ground truths. These datasets are then fed to *MiShape* for training, which consists of two components: A human silhouette generator and a joint location estimator. Inside the silhouette generator, the reflected signals are paired with the ground truth silhouette images to train Generative Adversarial Networks (GAN). The GANs, from thousands of data pairs, learn the association between images and signals and can generate the silhouettes from only the reflected signals at run-time. Inside the joint location estimator, the generated silhouette images are paired with the ground truth 3D locations of joints, and the model is trained by incorporating the rules of human biomechanics so that it can predict accurate joint locations at run-time. We now describe these design components in detail.

3.2 Relationship between Human Silhouette and Signal Reflections

MiShape relies on a deep learning model so that mmWave reflected signals can generate accurate human silhouettes. Before building such a model, we first analyze the reflected signals and ground truth silhouettes to understand their relationships. To this end, we have two hypotheses: (1) *The raw reflected signals from various poses of a human should have distinct features* so that a model could extract and learn them, even when the generated mmWave images from these signals are indistinguishable. (2) *The same pose performed by different humans should generate different features in the reflected signals*, so that a model could not only distinguish humans but also learn to generalize itself for multiple humans. To test these hypotheses, we first collect mmWave reflected signals from a single volunteer performing 6 different poses in front of *MiShape* (see Figure 6[a]). Furthermore, we ask 6 volunteers (different heights, somatotypes, see Figure 6[c]) to perform the same pose in front of *MiShape* and collect their reflected signals. Then, we analyze the features of the reflected signals by measuring the t-SNE distribution [53], where similar features in the input space should cluster near each other.

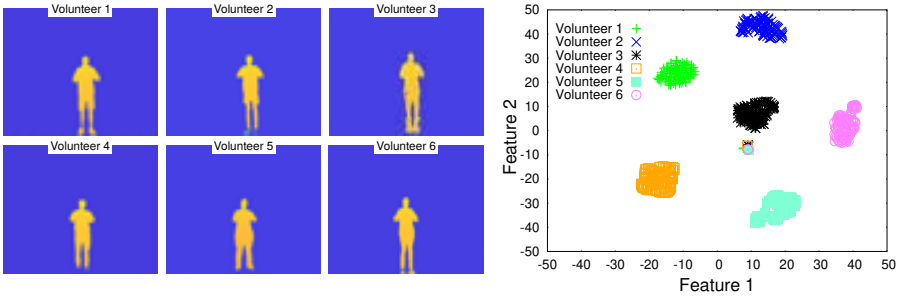
Figures 6(b) and (d) show the t-SNE distributions for these datasets. We observe 6 distinct feature clusters in the reflected signals corresponding to the case when the same volunteer performs distinct poses (Figure 6[b]). Besides, for each pose, its cluster is highly concentrated around the centroid, which shows that the behavior of the reflected signal is consistent across the same pose. We also observe another 6 distinct feature clusters, concentrated around their centroids, corresponding to the case when multiple volunteers perform the same pose (Figure 6[d]). This preliminary analysis validates the hypothesis and suggests that mmWave reflected signals can be used as an input to a learning based system to distinguish between humans and their poses.

However, these results do not showcase whether it is feasible to map these reflected signals to generate silhouette images. To this end, we hypothesize that *visually similar-looking silhouettes likely produce similar mmWave reflections and vice versa*, so that a model could learn the association between the reflected signals and



(a) Single volunteer performing different poses.

(b) t-SNE plot for different poses.



(c) Single pose from different volunteers.

(d) t-SNE plot for different volunteers.

Figure 6. t-SNE analyses of reflected signals from poses and volunteers show dominant sub-clusters in feature space.

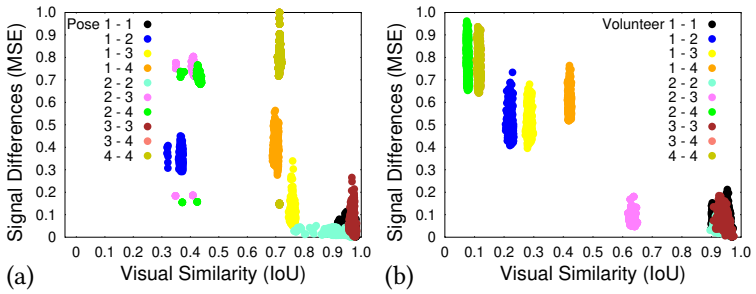


Figure 7. Analyzing one-to-one relational mapping between reflected signals and visual data across: (a) Different poses and (b) Different volunteers.

silhouettes and train itself to generate viable silhouettes given only the reflected signals. To this end, we use our data to identify the visual similarity between silhouette images and the difference in reflected signals across a pair of datasets with one-to-one mapping. The visual similarity is measured using the 2D IoU between the silhouette images [42] (where IoU value 1 means two images are identical), and the signal difference is measured using a normalized Mean Square Error (MSE) between the signals (where MSE value 0 means two signals are identical). We select the datasets for poses 1 through 4 and for volunteers 1 through 4, and find the corresponding IoU and MSE for pairs of poses (*i.e.*, pose 1 vs. 1, pose 1 vs. 2, *etc.*) and pairs of volunteers (*i.e.*, volunteer 1 vs. 1, volunteer 1 vs. 2, *etc.*).

Figures 7(a) and (b) show the relationship between visual similarity and signal differences across pairs of poses and volunteers, respectively. Clearly, the IoU and MSE (in Figure 7[a]) for datasets of the same pose pairs are clustered near 1 and 0, respectively, which indicates that a pose consistently generates similar reflections and silhouettes, irrespective of the pose type. Also, different poses consistently generate different reflections and silhouettes, which should allow mapping the reflections to the silhouettes. Furthermore, the IoU and MSE (in Figure 7[b]) for volunteer pairs also show similar behavior, which should allow not only distinguishing volunteers but also mapping reflected signals to their unique silhouettes.

3.3 Data Preprocessing

To learn only the necessary features from the reflected signals and ground truths, *MiShape* first preprocesses them to remove spurious information. Data preprocessing involves two steps: (1) Silhouette images and joints extraction by subtracting the unwanted background for high-quality ground truths, and (2) Data synchronization and resampling to align the mmWave reflected signals with the ground truths.

3.3.1 Silhouette Images and Joints Extraction. The silhouette images from a typical RGB-D camera often have spurious noise due to its inability to compute correct depth in the presence of background and clutters. Besides, different human body parts have different depths from the device, so thresholding the image with a single depth value to separate the background is infeasible. These noise and variable depths, make human silhouette extraction from the RGB-D cameras non-trivial and error-prone. To this end, we follow [54, 55] to use the body joint locations reported by RGB-D cameras [56] to separate the foreground human body from background noise. These 3D locations provide the seed point of human joints and are essential to locate the region of interest. We first use these seed points in the depth images to grow the region near the joint and cover the body parts that the specific joint has represented, and then, merge all the regions. Extracting the human silhouette from depth image using joint locations not only produces an accurate shape but also works well in diverse settings.

3.3.2 Data Synchronization and Resampling. Since a learning model will rely on the true relationship between the reflected signals, silhouettes, and joints, it is critical that the model is only trained on synchronized samples. Since hardware synchronization is currently unavailable between the mmWave device and the RGB-D camera, we use software synchronization and process the data samples further to remove residual misalignment. During an experiment, we find the local timestamp of subject movement by analyzing the temporal changes in the mmWave signals and then correlate the movement from the RGB-D camera. Since the effect of movement should appear simultaneously on RGB-D camera and mmWave device, we can calibrate their data samples by first identifying the local timestamp of movement start and then offsetting the samples *w.r.t.* the timestamps. Besides, there could be a sampling rate mismatch between the devices (*e.g.*, mmWave device and RGB-D camera in our setup have 25 and 30 fps sampling rates, respectively). So, we resample RGB-D silhouette images in time using a weighted average method, similar to [57]. These preprocessed ground truths aligned with mmWave signal reflections form the output and input pair, respectively, to the human silhouette generator network.

3.4 Human Silhouette Generation

The core purpose of the human silhouette generator network is to convert the mmWave signal reflections to high-resolution human silhouettes and capture diverse human poses. To learn the relationship between reflections to the silhouettes, *MiShape* uses Generative Adversarial Networks (GAN) and trains them with thousands of past examples of reflections and ground truths. *MiShape* uses two GANs in its silhouette generator: A conditional GAN (cGAN) and Super-Resolution GAN (SRGAN), which run in succession. The cGAN first generates low-resolution silhouettes directly from the reflected signals, and then the SRGAN upsamples them to high-resolution. The framework is designed in such a two-step because the size of the input reflected signals is typically very low

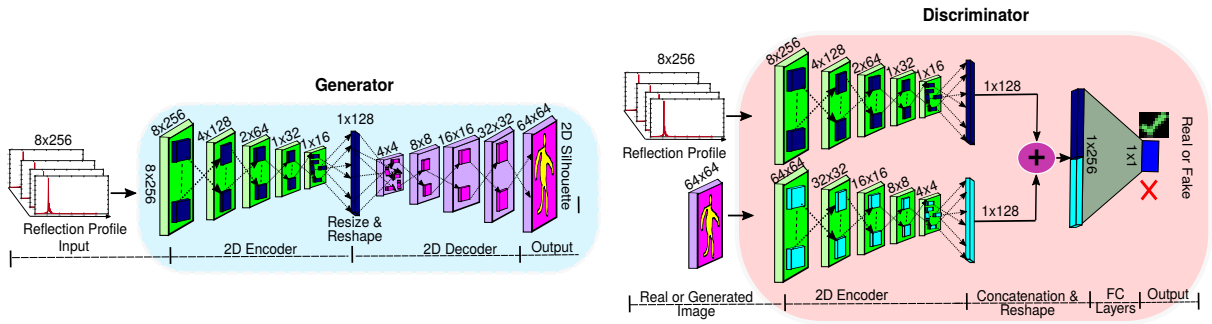


Figure 8. Generator and Discriminator networks of *MiShape*'s cGAN.

compared to the size of output silhouette images, which makes it hard for the network to converge [58–60]. For example, in our design and evaluations with COTS devices, the input size is only 8×256 and the output size is 512×512; so, the network does not converge well when we use only the cGAN to learn the larger output size.

Therefore, *MiShape* is trained in two steps. *First*, the cGAN is trained on low-resolution ground truth images obtained by down-sampling the original silhouettes to learn its association with the reflected signals. This model can be used to generate low-resolution silhouettes directly from the reflected signals. *Then*, the SRGAN is trained to convert those generated low-resolution images to high-resolution. These generated images are the final outputs from *MiShape*'s human silhouette generator network. Both GANs consist of a *Generator G* to learn the association between the input and ground truth, and a *Discriminator D* that guides *G* to learn better association at each iteration. During the run-time, when both networks have been trained properly, *MiShape* can estimate an accurate, high-resolution human silhouette using only the mmWave reflected signals and without using the ground truth. We now discuss the GAN fundamentals and then describe the network components in detail.

3.4.1 GAN Fundamentals. A generative model is an unsupervised learning which automatically discovers and learns the pattern and regularities in the input dataset to generate new datasets with similar distribution as the input. GAN augments the concept of the generative model and improves the quality of output by training the network with supervision via two sub-models: (1) *Generator G* to generate new outputs and (2) *Discriminator D* to classify outputs generated by *G* as either real (*i.e.*, from ground truth datasets) or fake (*i.e.*, generated from *G*) [61]. These two sub-models are trained in an adversarial, zero-sum game, until *G* is able to fool *D* by generating plausible examples [61]. However, traditional GANs cannot control the modes of generated data to a particular domain; therefore, in *MiShape*, we propose to restrict the generated datasets by conditionally training *G* to follow the distribution of the ground truth silhouettes through a customized conditional GAN (cGAN) [62].

3.4.2 MiShape's cGAN Model. Figure 8 shows the network architecture for *MiShape*'s cGAN. The model is conditional on ground truth silhouettes and consists of two network blocks: *Generator* (*G*) and *Discriminator* (*D*).

Generator: The objective of the *Generator G* is to convert the mmWave signal reflections to a human perceivable silhouette with complete pose information. To ensure network convergence with minimal training for any new environment, we need correct feature representation between input and output. With traditional Convolutional Neural Networks, such representations are directly encoded from the dense layers, and the network relies on them to upsample and reconstruct images. However, dense layers perform a linear operation on the layer's 1D feature vectors and are not deep enough to capture the accurate pixel-by-pixel reconstruction. Instead, we leverage an encoder-decoder architecture, where the encoder first extracts abstract features from the input reflections using multiple 2D convolution layers and a flatten layer. Then, the decoder converts those features to a 2D silhouette using multiple deconvolution layers. In our design, we use five 2D convolutional layers and five 2D

deconvolutional layers at the encoder and the decoder, respectively (see Figure 8). We find five convolutions and five deconvolutions to be the minimum number of operations to fit our input and output representation onto a 1D feature vector of 128. We do not want to increase this number to encode onto a feature vector of larger size as it will create sparsity and will increase the network's complexity for the same relevant features encoded by a vector of size 128. Table 1 summarizes the G network parameters.

Table 1. Generator network parameters for *MiShape*'s cGAN. RSC: Reflected Signal Convolutional layer; RSDC: Reflected Signal Deconvolutional layer. Act. Fcn.: Activation Function. LReLU: Leaky ReLU.

	RSC1	RSC2	RSC3	RSC4	RSC5	RSDC1	RSDC2	RSDC3	RSDC4	RSDC5	Output
Filter #	4	8	16	64	128	128	64	16	8	1	
Filter Size	4x3	6x6	6x6	6x6	6x6	4x4	4x4	4x4	4x4	4x4	
Dilation	2x2	2x2	2x2	2x2	2x2	4x4	2x2	2x2	2x2	2x2	
Act. Fcn.	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	Linear

Discriminator: The objective of the *Discriminator* D is to guide G to generate real-looking silhouettes via adversarial training. D classifies the real samples and the generated samples during the training process to assist G in learning a better association between the reflected signals and silhouettes. It takes two inputs in the forms of the reflected signal and the 2D silhouette image, which could either be from real datasets or datasets generated by G. Then, D outputs the probability of whether the input is real or fake. *Discriminator* takes input from two sets of images. The first set of images is the output from the generator and the second set is the ground truth images. Therefore, it needs two networks to represent images into 1D features to classify them as real or fake. To this end, D uses two encoder networks to encode these images: (1) The first encoder with convolution operations that encodes the reflected signals into a 1D feature vector. This encoder follows the similar encoder architecture of G. (2) A second encoder with five 2D convolutions to encode the ground truth into an appropriate 1D feature vector. To distinguish between the inputs, D concatenates the output feature vectors from these two encoders and then feeds them into two fully connected dense layers and one dense layer that outputs the probability. Since the probability values should be between 0 to 1, we use a Sigmoid activation function in D's final dense layer. In this adversarial training, the overall network converges when D consistently outputs a probability of 0.5, *i.e.*, irrespective of the input from real datasets or G, D will assign an equal probability of input being real or fake. Said differently, G is now trained properly to output samples indistinguishable from the ground truth silhouettes. Then, at run-time, when we input the mmWave reflected signals to G, it can output accurate human silhouette images. Table 2 summarizes the D network parameters.

Table 2. Discriminator network parameters for *MiShape*'s cGAN. RSC: Reflected Signal Convolutional layer; 2DC: 2D Convolutional layer; FC: Fully Connected layer; Act. Fcn: Activation Function; LReLU: Leaky ReLU.

	RSC1	RSC2	RSC3	RSC4	RSC5	2DC1	2DC2	2DC3	2DC4	2DC5	FC	Output
Filter #	4	8	16	64	128	4	8	16	64	128	64	1
Filter Size	4x3	6x6	6x6	6x6	6x6	4x3	6x6	6x6	6x6	6x6		
Dilation	2x2	2x2	2x2	2x2	2x2	2x2	2x2	2x2	2x2	2x2		
Act. Fcn.	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	LReLU	Sigmoid

cGAN Loss Function: Loss functions are critical in the learning-based model for training as they allow the network to tune its convolution or deconvolution weights appropriately. Intuitively, *MiShape*'s cGAN loss function should account for not only how well the adversarial learning works but also the pixels and shape quality of generated silhouettes from G. To this end, we use the vanilla GAN loss function $L(G)$ [61] from the output of D and G to tune and maintain the adversarial, zero-sum game. Additionally, we include the $L(2)$ loss, which is the L2-norm between the ground truth and generated images. L2-norm ensures that the network is able to

generate correct human silhouettes by estimating the MSE of pixels between the images. We combine both the loss functions with shape hyperparameters λ_G and λ_M, and the total cGAN loss function is designed as:

$$L_{cGAN} = \lambda_G \cdot L(G) + \lambda_M \cdot L(2); \quad \textit{where,} \quad L(2) = \mathbf{E}\|y - G(x)\|_2 \tag{1}$$

Here $G(x)$ and y are the generated and ground truth images, respectively. During training, we try to minimize our loss function by optimizing the hyperparameters, which aims to find the best weights for the convolutional and deconvolutional layers and ensure the model generalizability on unseen cases. We have discussed more on the choice of these hyperparameters in Section 4.

3.4.3 MiShape's SRGAN Model. The low-resolution human silhouette images generated by the cGAN show the general shape and layout of the body, detect movement of arms, legs for a specific exercise, but fail to reveal the high-resolution and minute details of specific exercises. These visual cues are necessary for machines to predict joints accurately and to classify between various similar exercises, such as Hands on waist, Both arms Up, Namaste, *etc.* (see Figure 6[b]). Besides, the high-resolution enhanced shapes could also bolster human perception for manual activity/exercise recognition. One approach to increase the resolution would be to apply traditional interpolation, such as Bicubic [63] or Nearest Neighbors [64], *etc.* However, interpolation methods only consider the pixels in the local region to determine the intensity of new pixels for high-resolution images; hence it will smoothen the image, and distinct features and details will be lost. Furthermore, interpolation techniques cannot recover missing body parts on low-resolution images, even if the general human shapes are known.

We believe learning based upsampling could be better than interpolation methods since it can learn the general shapes of human body from existing sets of images. The approach is inspired by the existing deep learning enabled high-resolution models [65]. But the challenge here is that the available SRGAN model is trained on millions of RGB color images from the ImageNet database [66], but *MiShape* tries to improve resolution on monochrome depth images. So, the data distribution on which the existing SRGAN model is trained does not match with *MiShape*'s dataset. Moreover, the existing SRGAN also generates the RGB image in the output as resolution improved image while *MiShape* needs a monochrome depth image on its output layer. Therefore, we customize the filter size of the initial convolution layer of the standard SRGAN architecture [65] to input the monochrome depth image, customize the stride size, convolution filter size, and feature size of the convolution blocks of the network to upsample the low-resolution images of size 64×64 to high-resolution images of size 512×512. *MiShape* first learns the mapping from low to high-resolution from thousands of pairs of examples: The ground truth, high-resolution silhouettes are intentionally decimated by a factor of 8 and then paired with the original ground truths to prepare the input-output pair for SRGAN. The generator upsamples images from the low-resolution images into the high-resolution image by passing them through multiple convolution layers, and the Discriminator distinguishes the images generated by the Generator is real or fake. *MiShape*' SRGAN has 16 residual blocks in the generator network G that contains the skip connection between convolution layers with identity transformation, allowing G to preserve high frequency details on high-resolution image. While discriminator has 7 pairs of convolution, batch normalization, and Leaky ReLU activation layers, that encode the high-resolution input image sampled from either generated or ground truth. We then pass these 1D abstract features into the fully connected layer (1024 neurons) and finally to the single neuron output layer to predict a binary decision, *i.e.*, if the input image is real or fake.

SRGAN Loss Function: SRGAN trains independently and doesn't update other networks from its output. However, during the inference, we use a trained SRGAN network to produce high-resolution silhouettes from low-resolution silhouettes that are generated from cGAN. To learn the association between low-resolution and high-resolution images, *MiShape*'s SRGAN uses a different combination of losses to update the network parameters. In loss function, we include content loss (reconstruction loss) $L(C)$ to make sure images produced are perceivable by humans, pixel loss $L(P)$ to learn correct depth values on the reconstructed images, and adversarial loss $L(G)$ to

maintain adversarial learning. We compute the content loss by calculating Mean Squared Error (MSE) between features obtained by passing the generated ($G(x)$) and ground truth (y) shapes through pre-trained VGG-19 [67]. Similarly, we compute pixel loss as MSE between $G(x)$ and y. The total SRGAN loss is then:

$$L_{SRGAN} = L(G) + \lambda_C \cdot L(C) + \lambda_P \cdot L(P);\ \textit{where,}\ L(C) = \mathbf{E}\|VGG(y) - VGG(G(x))\|_2;\ \textit{and}\ L(P) = \mathbf{E}\|y - G(x)\|_2 \quad (2)$$

During training, we optimize hyperparameters λ_C and λ_P to minimize L_{SRGAN}. Optimizing hyperparameters is essential to ensure that the network updates its parameters with proper loss function to learn the association between input and output. Section 4 discusses the choice of these hyperparameters in detail.

3.5 3D Joint Location Estimation

The core purpose of the joint location estimator network is to predict accurate 3D locations of joints from the generated human silhouette images. Since joint locations of major body parts, such as arms, legs, spine, head, *etc.*, determine the general posture of a human, accurately estimating their locations is vital. For example, if a person is performing *Lunges*, we will see that the set of various joints' locations differ than that of a person performing *Squats* (see Figures 6 and 9). One approach could be to directly estimate the joints from the mmWave reflections, but reflected signals lack the spatial features and result in poor joint location estimation. Another approach could be to use the traditional mmWave imaging and then predict the joints, but traditional imaging suffers from low-resolution and aliasing (Section 2.2) and thus will predict incorrect joint locations. To this end, *MiShape* uses the high-resolution silhouette images generated by the silhouette generator network, pair them with the ground truth 3D locations of joints to train a deep learning model, and predict the locations at run-time.

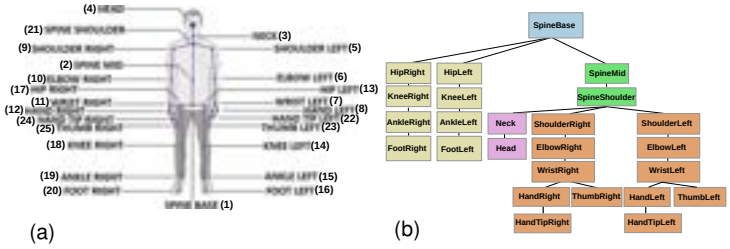


Figure 9. (a) Human joints name-index pair available in our ground truth [68]. (b) Human joints parent-child hierarchy.

MiShape designs a customized deep neural network, Joint Estimator (JE), for this purpose. Instead of only learning the relationship between silhouette images to the absolute joint locations, *MiShape*'s JE network also learns the arrangements of one joint *w.r.t.* another. Human joints follow a well-defined kinematic chain with a hierarchical structure ([18], Figure 9[b]). For example, the *Spine Base* (in Figure 9[a]) is connected to the *Hip Right* or *Hip Left* and limits their movements; so, the *Spine Base* is considered a parent of the *Hips* in the hierarchical structure (Figure 9[b]). Similarly, *Knee Right* will be a child of *Hip Right*, and so on. So, if we can somehow incorporate this structure into the learning model, the network can predict better joint locations.

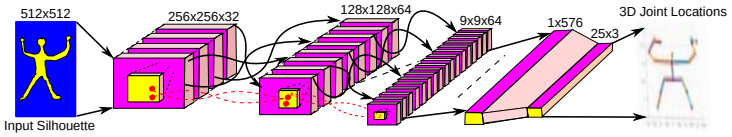


Figure 10. Joint estimator network architecture of *MiShape*.

Table 3. Network parameters for *MiShape*'s JE. 2DC: 2D Convolutional layer; FC: Fully Connected layer; Act. Fcn: Activation Function; PReLU: Parametric ReLU.

	2DC1	2DC2	2DC3	2DC4	2DC5	2DC6	FC1	FC2	Output
Filter #	32	64	64	64	64	64	576	75	75
Filter Size	3x3	3x3	3x3	3x3	3x3	3x3			
Dilation	2x2	2x2	2x2	3x3	3x3	3x3			
Act. Fcn	PReLU	PReLU	PReLU	PReLU	PReLU	PReLU	Relu	Relu	Linear

3.5.1 MiShape's Joint Estimator Model. The model takes high-resolution silhouettes of size 512×512 generated from SRGAN to predict 25 three-dimensional joints of human pose/skeleton. To this end, JE first extracts the abstract features from the images supplied at the input by using multiple convolutional layers (2DC1 - 2DC6, see Table 3) with different filter sizes, feature sizes, dilation, and pooling. We use convolutional and pooling layers with increasing feature size until the spatial dimension converges to multi-channel 1×1 features; this is needed to extract not only the local spatial relationship but also the global relationships between the image pixels. Finally, we supply the abstract features to the fully connected layers (FC1 and FC2), which predicts the X, Y, Z coordinates of the 25 joint locations (*i.e.*, 75 real values). We observe that the JE network with only MSE loss of absolute joint locations is insufficient; thus, we exploit the hierarchical structure of joints and enforce them in the model through loss function and hyperparameters. Figure 10 shows the JE network architecture, and Table 3 shows the network parameters in more detail.

JE Loss Function: The loss function for JE combines both the MSE loss, L_{MSE}, between the predicted and ground truth joint locations and the parent-child distance loss, L_{PC}, following the joint hierarchy. The parent-child distance loss keeps track of the parent's 3D joint location while simultaneously predicting the child's 3D joint location, and the total loss function is defined as:

$$L_{JE} = \lambda_J \cdot L_{MSE} + \lambda_D \cdot L_{PC};\ \textit{where, } L_{MSE} = \sum_{i=1}^{N} MSE(Z_p^i, Z_g^i);\ \textit{and } L_{PC} = \sum_{i=1}^{N} |d_p^i - d_g^i| \tag{3}$$

Where N is the total number of joints, Z_p^i and Z_g^i represent the predicted and ground truth i^{th} joint locations, respectively, and d_p^i and d_g^i are the predicted and ground truth distances between i^{th} joint and its parent joint. Note that for the joint with no parent (*i.e.*, *Spine Base*), both d_p and d_g are zero, and we enforce this explicitly in the model. We will discuss the choice of the hyperparameters λ_J and λ_D in more detail in Section 4.

In summary, MiShape first generates a low-resolution human silhouette from the mmWave reflected signals with cGAN, improves its quality through SRGAN, and finally predicts the 3D joint locations with the JE network.

3.6 Gait Monitoring Application of *MiShape*

Both the high-resolution silhouettes and 3D joint locations can facilitate continuous, privacy non-invasive monitoring applications at-home. In this work, we design and evaluate a Gait monitoring application.

Approximately 90% of patients with chronic stroke ambulate with impaired coordination: Their gait is slow, endurance is poor, and walking pattern has diminished quality and adaptability [69, 70]. But they exhibit improvements with physical therapy interventions [71, 72]. Thus, knowing a stroke survivor's gait in their home, and more importantly, knowing when it gets worse, could be a strong indicator to intervene with more therapy or physical activity. The benefit of monitoring walking steps has also been linked with the early detection of strokes and Alzheimer's [73, 74]. To assist the user in monitoring their gait, *MiShape* generates a real-time silhouette of the user during a regular walk, predicts their body joints, and records their gait. *MiShape* can also facilitate experts remotely tracking the gait of a patient and providing feedback. Gait is comprised of two phases: A *Stance* phase that begins when the foot first touches the ground and ends when the same foot leaves the ground, and a *Swing* phase that begins when the foot first leaves the ground and ends when the same foot touches the ground.

A complete gait cycle for healthy individuals consists of ~60% of *Stance* and ~40% of *Swing* [75], and they can be quantified using the following standard gait parameters ([76–79], see Figure 11): (a) *Step length* is the distance between different feet at their *Stance*; (b) *Stride length* is the distance covered by the same foot after one *Swing*, and ; (c) *Cadence* is the measure of the total number of steps taken within a minute. We customize *MiShape* to enable gait monitoring application and present the prediction results of these parameters in Section 5.2.

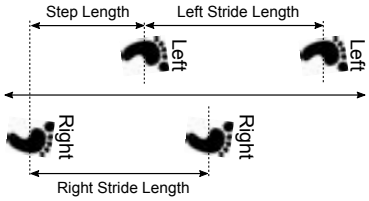


Figure 11. Gait parameters.

4 IMPLEMENTATION

Hardware Platform: Due to the unavailability of open-source datasets of mmWave signal reflections, silhouette images, and joint locations, we train and test *MiShape* with real data collected from a customized hardware platform. Since existing 5G mmWave routers do not provide raw signal reflections yet, we design a setup integrating COTS mmWave transceivers and an RGB-D camera for data collection. The mmWave transceiver is a 77–81 GHz device, TI IWR1443BOOST [40], and the RGB-D camera is a Microsoft Kinect Xbox One [41] (see Figure 12). The mmWave device has one transmit and four receive antennas on a linear axis, and in its horizontal orientation, it can resolve reflection points in azimuth and depth only. To resolve the points in elevation as well, we use two of these devices with one rotated 90° counter-clockwise *w.r.t.* another, so the resultant setup has two antenna arrays arranged in 1×4 and 4×1 configurations. With this approach, we can resolve the reflections in depth, azimuth, and elevation with a resolution of 9.4 cm, 28°, 56°, respectively. We apply traditional Frequency

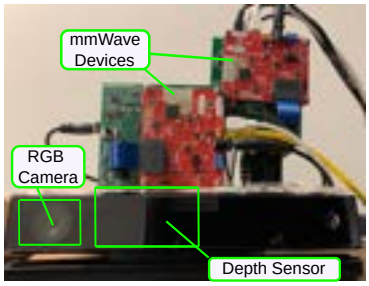


Figure 12. Ground truth data collection setup with two 77-81 GHz devices and a Kinect Xbox One.

Modulated Continuous Wave (FMCW) processing on the received signals, with the following signal parameters: Ramp start frequency – 77 GHz; frequency ramp slope – 70.3 MHz/μS; baseband sampling rate – 5 Msps; number of ADC samples – 256; chirp sweep duration – 60 μS; pulse repetition rate – 1 kHz; and maximum antenna gain – 10.5 dBi. We also attach a data capture module, TI DCA1000EVM [80], to each IWR1443BOOST device to capture the signals in real-time. The DCA1000EVM module can temporarily store 2 GB of data in its FPGA buffer and then transfer it via an Ethernet cable to a host laptop. The Kinect is co-located with our mmWave devices and can collect the silhouette images and joint locations at 30 fps, whereas mmWave device collects signals at 25 fps. Recall that the data collected from the mmWave device and Kinect cannot be fed directly into *MiShape*'s learning

model without sanitizing the input-output pairs due to the differences in the sampling rate, background, spurious noise, and software delays. To this end, we follow the preprocessing method discussed in Section 3.3 to calibrate the recorded datasets. We implement *MiShape* in Matlab and Python environments running on a host PC, which uses the mmWave reflected signals as input and generates human silhouette and 3D joint locations as output.

Table 4. Information about poses.

Pose #	Exercise name	Pose #	Exercise name	Pose #	Exercise name
1	Lunges	2	Zoom in-out	3	Arm stretch
4	Arm extension	5	Alternate toe touch	6	Alternate arms up
7	Both arms up	8	Arms up and down	9	Hands on waist
10	Leg extension	11	Namaste	12	Squat
13	Standing	14	Stretch	15	Walk fast
16	Walk normal	17	Walk slow		

Table 5. Information about volunteers.

Characteristics	Number
Total number of volunteers	10
Age range (years)	12 - 52
Male/Female (%)	80/20
Height range (cm)	152 - 183

Real Data Collection: For our microbenchmark evaluations, we collect datasets from a single subject performing 17 diverse poses. Table 4 shows the high-level information about these poses. We have selected these 17 poses that are found to be common in physiotherapy for patients and elderly [81, 82], post orthopedic-surgery [83], and home exercises [84], and involve various ranges of motion of arms, legs, and body. For example, we have included Pose 7, *"Both arms up,"* where the user only moves arms keeping both leg and body at rest. Different from Pose 7, Pose 10, *"Leg extension,"* involves only legs' motion keeping body and arms at rest. In addition to these categories involving separate body parts, we have also included several poses, such as Pose 5, *"Alternate toe touch,"*, that involve complex simultaneous motions of arms, legs, and body. Including such variations of motions in individual and multiple body parts in our training dataset makes *MiShape* easily generalizable to different exercises that may not have been included in the training process. For each pose, the subject stands at approximately 2 m distance (except for poses 15, 16, and 17, which involve walking around) from the setup during the data collection. The subject wears a similar outfit during the experiments, and the background is drywall without any clutters. A single experiment takes approximately 12 seconds to complete and generates a dataset for a given pose, and a trial involves all 17 poses. We collect data for over 14 days, where the subject is asked to repeat each trial 10 times. After preprocessing the dataset, pruning for noise, *etc.*, we have 26,796 input-output pairs of mmWave reflections, human silhouette images, and 3D joint locations.

To study the performance variations and understand the generalizability of *MiShape*, we also conduct experiments involving 9 additional volunteers and collect the data following similar processes as above. Table 5 shows the information about the volunteers in more detail. We ask each volunteer to perform diverse exercises (poses 2, 4, 6, 7, and 11 in Table 4). All experiments involving the additional volunteers are conducted within a single trial for about an hour, and the total duration for data collection was 9 hours spanning one week. After preprocessing the datasets, we have a new set of 65,502 input-output pairs. Furthermore, to evaluate *MiShape*'s performance in facilitating applications, such as at-home gait monitoring, we collect data from an individual volunteer for three different walking speeds (see poses 15, 16, and 17 in Table 4). So, in total, we have collected and analyzed nearly 100 K input-output pairs with a data size of over 14 GB involving different activities from 10 volunteers with diverse ages, gender, height, and somatotype. *This data diversity and scale allow us to evaluate MiShape's robustness as well as verify its generalizability across multiple, diverse conditions.*

Network Training: *MiShape* requires training of three independent models: (1) cGAN, (2) SRGAN, and (3) JE network. *First*, to train the cGAN, *MiShape* explores different network parameter settings to ensure a near-optimal model convergence. We set the initial value of our total epochs to be 1200, and then, monitor the loss function from cGAN and stop its training when the model shows little to no improvement for consecutive 30 epochs. We also explore different optimizers, such as, *Adam*, *Rmsprop*, *etc.*, and observe that *Rmsprop* performs the best with a learning rate of $5{\times}10^{-4}$. We also explore different combination of the hyperparameters λ_G and λ_M

(Equation 1) and found that the network performs the best when λ_M is approximately 10× λ_G, *e.g.*, (λ_G, λ_M) = (0.1, 1). This ensures cGAN pays extra attention towards reconstructing the actual shape of humans along with adversarial learning. To train the SRGAN, we follow a similar strategy as above, and find that the *Adam* optimizer performs the best with a learning rate of 1×10^{-4}. We add pixel and content loss to our adversarial loss with hyperparameters of λ_P and λ_C, respectively, and we observe that the model performs better when the ratio between λ_P and λ_C is ∼100, *e.g.*, (λ_C, λ_P) = (0.01, 1). Said differently, the model converges better when the network gives more weightage to shape reconstruction. To train our JE network, we use the loss function in Equation 3 and optimize the hyperparameters λ_J and λ_D that balances the weights between MSE loss and parent-child distance loss, respectively. We observe that the setting with equal weights on λ_J and λ_D provides the minimum loss in JE network. This is intuitive since it is equally important to optimize for absolute joint locations as well as maintain an accurate parent-child joint relationship in the prediction. All our networks are implemented in Python with TensorFlow 2.4 [85] using Spyder IDE [86] and Anaconda version 4.10.3 distribution [87] in a PC with Intel Xeon CPU @3.5 GHz, 32 GB RAM, and NVIDIA's GeForce GTX 1070 GPU. Our training time varies across models and requires between 24 hours to 72 hours for completion, but it can be reduced significantly by using powerful GPUs or cloud TPUs [88, 89].

5 PERFORMANCE EVALUATION

We evaluate *MiShape*'s silhouette generator network performance using 2 metrics commonly used to compare the quality of images and another metric to evaluate the joint location estimation and application performance.
Intersection of Union (IoU): IoU measures the extent of overlapping between two images [42], and in *MiShape*, we use it to measure the overlapping between generated and ground truth images. IoU ranges from 0 to 1, and 1 indicates the generated image is a perfect replica of the ground truth.
Multi-Scale Structural Similarity Index Measure (MS-SSIM): MS-SSIM is a perceptual metric that quantifies image degradation in terms of shape, resolution, pixel intensity, and orientation *w.r.t.* ground truth [43]. It ranges from 0 to 1, and a value near 1 indicates that the generated image quality is near reference image quality.
Mean Square Error (MSE): MSE of joint locations measures the Euclidean distance between the ground truth joint location and predicted joint location. Lower MSE indicates that the prediction is close to the ground truth.

Evaluation Summary: (1) *MiShape*'s cGAN model generates human silhouette images of size 64×64 from the mmWave reflected signals with a median and 90^{th} percentile IoU of 0.72 and 0.84, respectively, and median and 90^{th} percentile MS-SSIM of 0.96 and 0.98, respectively. Furthermore, *MiShape*'s SRGAN upsamples the cGAN output by 8×, but preserves the finer texture in the silhouettes with image sharpness almost 2× as compared to the traditional interpolation methods. We observe that SRGAN generates high-resolution silhouettes with median and 90^{th} percentile MS-SSIM of 0.91 and 0.93, respectively. (2) *MiShape*'s joint location estimator can localize joints in 3D with a mean error of ∼10 cm across most critical joints, and the accuracy is consistent across all poses. (3) *MiShape* also performs well across diverse mmWave antenna configurations and consistently achieves more than 0.70 in median IoU, outperforming traditional imaging and an existing deep learning technique. Moreover, *MiShape*'s base model requires little fine-tuning, and with only 2 randomly selected volunteers' data samples, it can achieve a median IoU of 0.60 across all 10 volunteers. (4) Finally, for at-home gait monitoring, *MiShape* can estimate the *Step length* and *Stride length* with median errors of 0.19 m and 0.20 m, respectively, and nearly accurate *Cadence* across different types of walking.

5.1 Microbenchmark Results

5.1.1 Silhouette Generation from cGAN. To evaluate the effectiveness of *MiShape*'s cGAN, we use the single volunteer's dataset collected for 17 different poses. *First*, we preprocess the dataset to generate the input-output pairs of mmWave reflected signals and silhouette images, and *then*, randomly select the training and testing

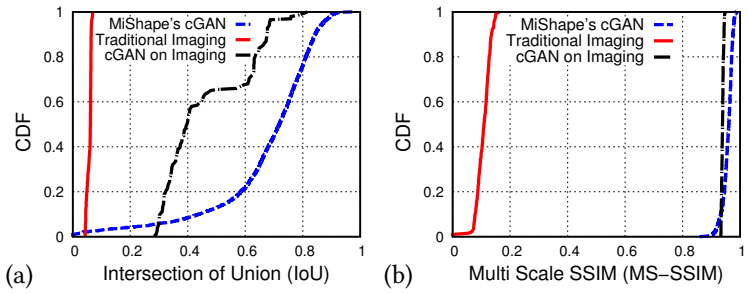


Figure 13. *MiShape*'s performance distribution comparison with traditional imaging and cGAN on imaging on two metrics: (a) IoU and (b) MS-SSIM.

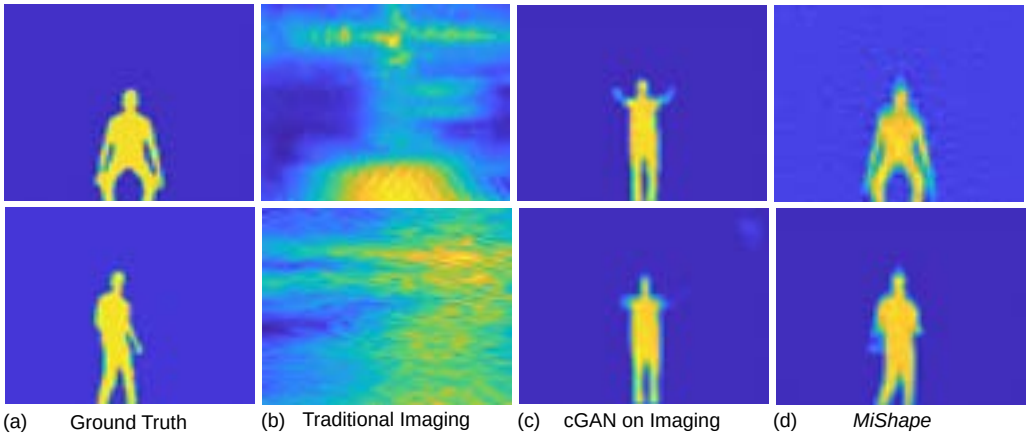


Figure 14. (a) Ground truth silhouette of a person performing a *Squat* (top row) and standing tilted to right (bottom row). Silhouette images using mmWave reflected signals: (b) Traditional imaging. (c) cGAN on traditional imaging. (d) *MiShape*.

samples. For training, we select ~3300 samples, and for testing, we select another ~3300 samples; both sets of samples are distributed evenly among all poses. After training, we feed the mmWave reflected signals from the test samples to the *Generator* of cGAN to generate the silhouette images. To find the efficacy of *MiShape*, we also compare and contrast it with two other approaches. In the first approach, we use the reflected signals to generate the silhouettes directly using a traditional imaging algorithm (Section 2.1). Since these images could suffer from fundamental aliasing and low-resolution issues, we seek to improve their quality using existing deep learning models. So, in the second approach, we use the 3D voxel generated by the traditional imaging, and then feed it through the existing cGAN model proposed in [13] to generate the silhouette images. The learning model is trained and tested with the same datasets used for training and testing *MiShape*, and we tune its hyperparameters for best network convergence. Finally, we estimate both the IoU and MS-SSIM between the ground truth and the generated images for traditional imaging, cGAN on traditional imaging, and *MiShape*.

Figure 14 shows a visual example of silhouette images generated by the three approaches and compares them with the ground truth. Clearly, traditional imaging cannot generate any shapes due to the low-resolution and aliasing issues (Section 2.2). While applying an existing deep learning model makes the silhouette more human-like, the resultant pose is far from the ground truth. This is because, the learning model when trained on aliased images couldn't distinguish between samples, and thus, produces random poses from its training dataset. In contrast, *MiShape* is trained with raw mmWave reflected signals and learns to overcome the aliasing effects to

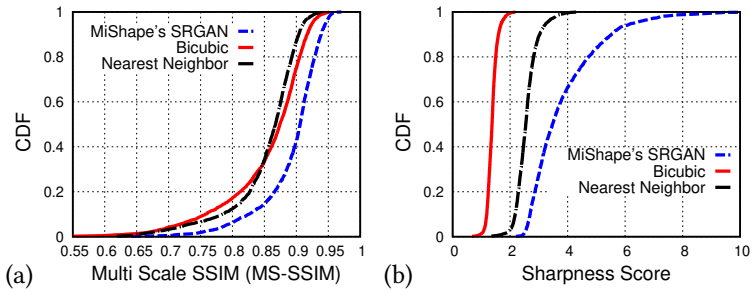


Figure 15. Comparison of the image quality upsampled by *MiShape* with two traditional interpolation methods: Bicubic and Nearest neighbor using two metrics: (a) MS-SSIM and (b) Sharpness Score.

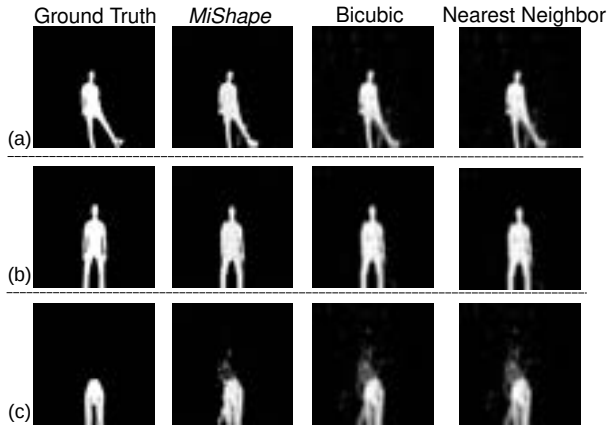


Figure 16. Silhouette comparison of *MiShape*, Bicubic, and Nearest Neighbor: (a) Leg extension; (b) Squat; and (c) Stretch.

generate a similar pose to the ground truth. Figures 13(a–b) further show the Cumulative Distributive Function (CDF) of the IoU and MS-SSIM across all test samples for the three approaches. Under traditional imaging, the median and 90^{th} percentile IoU is only 0.06 and 0.07, respectively, and median and 90^{th} MS-SSIM is 0.11 and 0.14, respectively, which are extremely low. Applying existing deep learning over these images improves their quality, and the median IoU and MS-SSIM increase to 0.4 and 0.94, respectively. However, these improvements still do not help generating accurate silhouettes (Figure 14[c]). *MiShape* outperforms both these approaches significantly: The median and 90^{th} percentile for IoU are 0.72 and 0.84, respectively, and for MS-SSIM, they are of 0.96 and 0.98, respectively, and the generated poses are consistent with the ground truth obtained from the RGB-D camera based system. *These results show that MiShape's silhouette generator network is well generalizable for multiple poses for a single subject, and it generates human silhouettes similar to the ground truth consistently.*

5.1.2 Resolution Improvement by SRGAN. While cGAN produces the general shape and layout of the body, its resolution is far from the high-quality silhouettes produced by RGB-D cameras. *MiShape* aims to improve the silhouette resolutions by employing a customized deep learning model, SRGAN. To this end, we evaluate its performance using the same volunteer dataset as in cGAN. For each low-resolution silhouette image (64×64) generated by cGAN, we pair it with the ground truth high-resolution image (512×512) and train the SRGAN. We also use the same training and testing samples as in cGAN, and after training, we feed ~3300 low-resolution test samples to SRGAN. To evaluate the efficacy of SRGAN, we implement two interpolation based methods: Bicubic [63] and Nearest Neighbor [64], that interpolate the silhouette pixel-by-pixel without the knowledge of human

shapes, and evaluate them on the same set of test samples. We then find the MS-SSIM between the high-resolution ground truth and the generated images. Finally, to understand the overall image resolution improvement in addition to the shape, we also measure the Sharpness Score [90, 91] for the generated images.

Figure 15 shows the CDF of both MS-SSIM and Sharpness Score for the three approaches. While both the interpolation methods can produce reasonable shapes with median MS-SSIM around 0.86, the final results are blurry (see Figure 16 for a visual example). This is expected since traditional interpolation methods apply weighted filters to local pixels on the neighborhood and do not take into account the global relationship among pixels. This is also visible in the Sharpness Score, where the 90^{th} scores are only, 1.37 and 2.54, respectively. In contrast, *MiShape* improves the quality of the images, even when they are upsampled by 8×, the median and 90^{th} percentile for Sharpness Score are 3.49 and 5.56, respectively, and the median and 90^{th} percentile for MS-SSIM values are of 0.91 and 0.93, respectively. Figures 28 and 29 in the Appendix A further show multiple examples of static and dynamic poses generated by *MiShape*. *In sum, MiShape's SRGAN maintains the silhouette quality even when the images are upsampled by 8×, and the generated shapes are similar to the ground truth.*

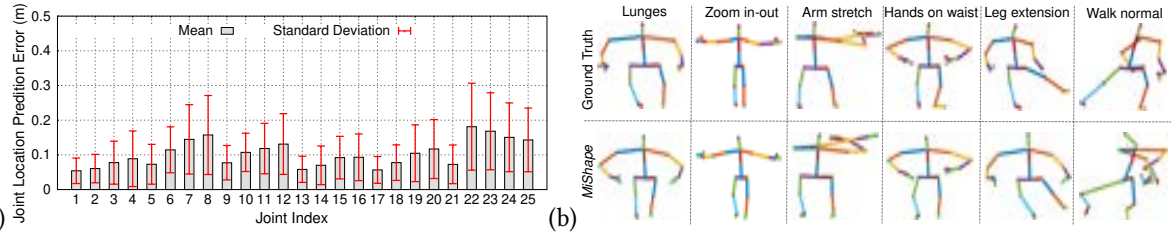


Figure 17. (a) *MiShape*'s prediction errors for 3D locations of 25 joints across 17 poses. (b) Ground truth and predicted joint examples of an individual performing different poses.

5.1.3 Effectiveness of 3D Joint Estimator. Apart from silhouette images, accurate joint location estimation is also vital for many tracking applications. To evaluate the effectiveness of the *MiShape*'s JE network, we use the same training and testing dataset as before, and pair the high-resolution images predicted by SRGAN with the ground truth 3D locations for 25 joints to train JE. Then, we test ~3300 samples distributed across the 17 different poses and find the MSE between the ground truth and predicted joint locations.

Figure 17(a) shows the result, where we observe that most of the joint locations could be predicted with a mean error of less than 10 cm. However, for joint numbers 7, 8, 22, 23, 24, and 25 (wrist left, hand left, hand tip left, thumb left, hand tip right, thumb right, Figure 9[b]), the prediction error is high, around 15 cm on average. These joints are the limb edges of the upper body, and are very difficult to recognize at a distance even with RGB-D cameras [92], and thus could be erroneous for *MiShape* too. However, a majority of the tracking applications involving critical joints in legs, feet, arms, spine, *etc.*, can still be accurately enabled by *MiShape*. Figure 17(b) shows different skeletons of human poses predicted by *MiShape* in comparison to the ground truth. In all examples, we see that leg joints are predicted more accurately than arm joints, and we believe this is due to very weak reflections from the human arms. *The result shows that MiShape is generalizable across multiple poses, and it can predict 3D locations of joints with low errors for most of the critical joint locations.*

5.1.4 Effect of Human Motion. We now demonstrate *MiShape*'s joint prediction error under different human walking speeds. We test *MiShape* with ~600 additional samples collected with the same volunteer walking at 3 different speeds: slow walking (< 1 mph), normal walking (2-4 mph), and fast walking (5-10 mph). For each case, we use the output from *MiShape* joint predictor and find the Euclidean distance of each predicted joint from its ground truth. Figure 18 shows that *MiShape* can predict all joint locations with a median error of 10 cm for static case, and the error increases to 17 cm when user starts walking with less than 1 mph speed. Even though

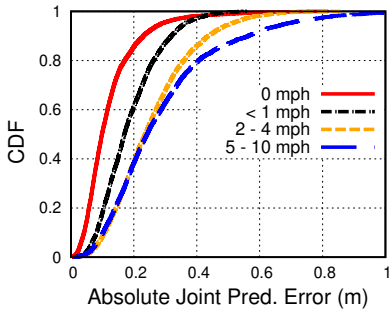


Figure 18. Error for mobile human.

median error reaches to 23.5 cm when user starts to walk normally but it doesn't exceed 24 cm even when user is walking faster (*i.e.* 10 mph). Ability of *MiShape* to predict joints accurately at higher speeds suggest that we can predict different exercises without sacrificing accuracy.

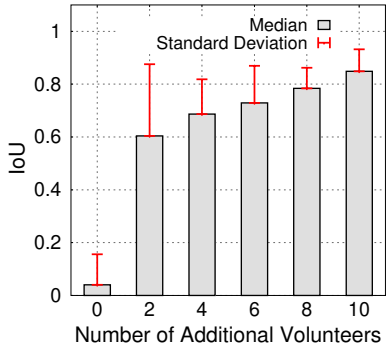


Figure 19. *MiShape* generates significantly better silhouettes with only a little fine-tuning.

5.1.5 Effect of Model Fine-Tuning. So far, we have evaluated *MiShape*'s performance with a single volunteer performing diverse poses. We now evaluate the generalizability of *MiShape*'s cGAN to generate silhouettes for a diverse set of volunteers. This evaluation is useful in understanding the amount of fine-tuning required to deploy *MiShape* in a new environment. To this end, we randomly select a set of 1300 data samples from the volunteer datasets, with each volunteer contributing 130 samples. *MiShape*'s baseline cGAN has never been trained on these samples before. We then create five sets of samples, each created by randomly selecting two volunteers and adding their samples. We first test *MiShape*'s base cGAN model without training any samples from these sets, and then we progressively add volunteers data to fine-tune the base model and test samples from all 10 volunteers.

Figure 19 shows the performance with different levels of fine-tuning. Without fine-tuning, the generated images from *MiShape* has very poor quality with a median IoU of only 0.04. This is expected since 90% of the datasets are from completely unseen volunteers, with varying heights and somatotypes, and the base cGAN model has been trained with one volunteer only. However, fine-tuning *MiShape* with 2 additional volunteers for only 10 epochs improves the generalization of network [93] and median IoU increases to 0.60. This is because by adding additional volunteers with potentially different physical features, the fine-tuned cGAN is able to capture the underlying correlation between physical features and mmWave reflections. By further increasing the number of volunteers for fine-tuning, we see image quality improves consistently, but the improvements show diminishing

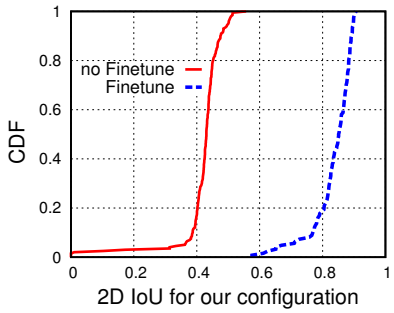


Figure 20. *MiShape*'s cGAN prediction on indoor scenario with multiple objects shows that it works well even with the presence of multiple object in surrounding.

return. *These results show that MiShape adapts very well for a large set of volunteers with minimal fine-tuning samples from a small set.*

5.1.6 Effect of Different Objects in the Surroundings. To validate the performance of *MiShape* in the presence of different objects of non-interest, we conduct additional experiments in two different environments. In the Environment 1 as shown in Figure 21(a), we collect data from a volunteer in presence of different objects in indoor setting. We collect total data of 842 samples that are measured for ∼60 seconds. We ask a volunteer to stay ∼ 2.2 m from the experimental setup and perform pose 3 (see Table 4). This environment includes objects like chair, ladder, cluttered boxes, and a bed similar to an at-home setting. In the Environment 2 (Figure 21[e]), we ask the same volunteer to perform same pose and collect data in a similar way. This environment setting is similar to our old setting where there are no objects in the surrounding. We first test *MiShape*'s base cGAN model without finetuning on Environment 1, we observe that *MiShape* can generate a rough silhouette as in Figure 21(c). We see that *MiShape* gives reasonable silhouette even when we do not do any finetuning. Futhermore, with little finetuning *i.e.* for 10 epochs using 100 samples, prediction improves as evident in Figure 21(d). This confirms that *MiShape* can infer signatures associated with a pose in a new environment but little finetuning can help improve the prediction in generating accurate human silhouette.

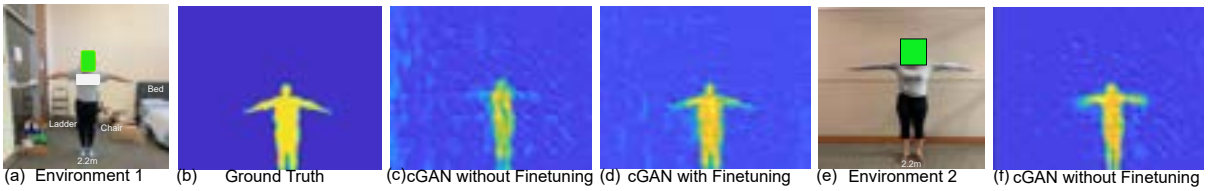


Figure 21. (a) A volunteer in Environment 1 doing pose 3 in presence of different objects of non interest. (b) Ground truth obtained from Kinect. (c) cGAN's output without finetuning. (d) cGAN's output with finetuning for 10 epochs. (e) Same volunteer in Environment 2 doing pose 3. (f) cGAN's output without finetuning.

Further, we test *MiShape*'s prediction on Environment 2 without finetuning, and observe that *MiShape* generates accurate human silhouette in Figure 21(e). This suggests that *MiShape* needs no finetuning in a similar environment for accurate prediction but, for change in environment, it can make a prediction but with very little finetuning, we can obtain highly accurate silhouette. Figure 20 shows the performance of *MiShape* in the presence of different objects of non-interest with finetuning and without finetuning. We observe that *MiShape* can predict human silhouette with median IoU of 0.43 and 90^{th} percentile IoU of 0.47. But with finetuning for 10 epochs that takes ∼ 10 minutes , we see improvement in median and 90^{th} percentile of 0.85 and 0.89, respectively. *These results show*

Figure 22. (a) A volunteer performing pose 3 in normal environment. (b) cGAN's output without finetuning. (c) A volunteer performing pose 3 in occluded environment. (d) cGAN's output with finetuning for 10 epochs. (e) Signal reflection during occlusion show that highest reflection comes from human standing 2.2m behind the occlusion. (f) cGAN's output without finetuning for occluded environment.

that MiShape *can generate silhouettes similar to ground truth consistently even in a new environment with many objects of non-interest.*

5.1.7 Effect of Occlusion and Low Light. We now evaluate the effect of occlusion and low-light conditions on *MiShape*'s performance. To experiment with occlusion, we put a thin sheet of clothing (bed sheet) at ~1.1 m, in front of our experimental scene in Figure 22(a), in line-of-sight with our experimental setting (Figure 22[c]). A volunteer is present behind the occlusion at ~2.2 m from our experimental setup doing a similar pose as in Figure 22(a). Figure 22(e) shows the reflection signal strength from the occluded scene. We see that human present at ~2.2 m reflects the signal with a higher strength compared to other reflections. We also see a little bit of reflection coming from the occlusion at ~1.1 m. Even with occlusion around human, we observe a distinct strong reflections coming from human. Thus, we can generate human silhouette by feeding these reflections to *MiShape*'s cGAN without finetuning. We observe in Figure 22(f) that *MiShape* can predict human silhouette even in the presence of occlusion. We couldn't make a comparison against ground truth as we rely on Kinect for our ground truth and it cannot penetrate through the occlusion.

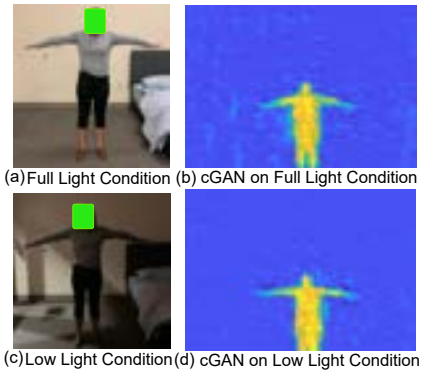


Figure 23. (a) A volunteer performing pose 3 in full light condition. (b) cGAN's output on full light condition. (c) Same volunteer performing same pose in low light condition. (d) cGAN's output.

To evaluate *MiShape*'s performance in low light conditions, we conduct two experiments at night when the source of the light is fluorescent tube lights present in indoor setting. *First*, we turn on all fluorescent lights and the setting looks like Figure 23(a). Then, we collect data from a volunteer present at ~2.2 m performing a pose. *Second*, we turn off all lights to replicate low light conditions typical to an indoor setting, causing the scene to look like Figure 23(c). As before, we collect data from a volunteer present at ~2.2 m performing the same pose. To make a comparison between full light condition and low light condition, we use our pre-trained model from this

experimental setting and generate human silhouette from *MiShape*'s cGAN without finetuning. Ideally, *MiShape* should perform same even with no finetuning as low light conditions do not have any impact on reflection signals coming from a human. Figure 23(b) and 23(d) show that *MiShape* produces highly accurate human silhouette and has no impact in low light conditions. *These results confirm that* MiShape*'s cGAN can process mmWave reflection signal to generate human silhouettes even in the presence of occlusion and in low light conditions at night.*

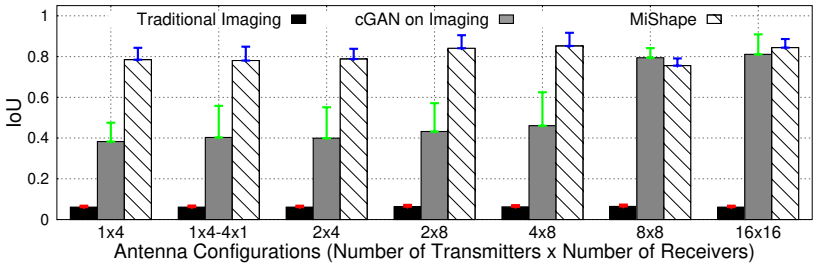


Figure 24. *MiShape* generates high-quality images consistently across diverse antenna configurations, outperforming traditional imaging and an existing deep learning model.

5.1.8 Effect of Different mmWave Antenna Configurations. In all our analyses thus far, we have used a COTS mmWave device with a specific antenna configuration (two arrays of size 1×4 and 4×1). Recall that, future COTS devices could be equipped with larger antenna arrays, and the array size directly influences the performance of an imaging system (Section 2.2). So, we now evaluate the effect of different antenna configurations on *MiShape*. Due to the unavailability of real devices with different configurations, we evaluate the performance of *MiShape* on different configurations through emulation. To this end, we emulate six additional antenna configurations with the following sizes: 1×4, 2×4, 2×8, 4×8, 8×8, and 16×16. For each configuration, we use 1556 ground truth silhouette images and follow a ray-tracing method similar to [13, 94] to generate synthetic mmWave reflections from the human body. Then, we divide the dataset into 1356 samples for training and 200 samples for testing. We implement a traditional imaging algorithm to generate the 3D voxel and train the existing deep learning model [13], and use the reflected signals directly to train *MiShape*. We customize the encoder in *MiShape*'s cGAN architecture based on the size of the reflection profile for each antenna configuration to map input representation onto 1D feature vector of 128 as shown in Figure 8. Finally, we use the test samples and generate the silhouettes from traditional imaging, cGAN on traditional imaging, and *MiShape*. For each antenna configuration, and for each of the three approaches, we estimate the IoU of the generated image *w.r.t.* the ground truth.

Figure 24 shows the performance of the three approaches for different configurations. As expected, the traditional imaging performs poorly, and its generated silhouettes show little to no improvement even with a larger array of size 16×16: The median and 90^{th} percentile IoU are 0.06 and of 0.07, respectively, across all configurations. Employing existing deep learning techniques on these images can improve their quality, but the generated silhouettes are similar to the ground truth only in case of larger array sizes, such as 8×8 and 16 ×16. In contrast, *MiShape*'s median IoU is consistently near 0.8, irrespective of the antenna configuration. Thus, it outperforms both these approaches for smaller array sizes (up to 4×8), and achieves a similar performance to the existing deep learning method for larger arrays (8×8 and 16×16). More importantly, *MiShape* can match the imaging performance of the existing deep learning based method with a larger array of size 16×16 (median IoU = 0.81) using only a small array of size 1×4 (median IoU = 0.78). *In summary, MiShape is generalizable to different antenna configurations, achieves consistently high imaging quality, and can approximate the performance of expensive hardware with cheaper hardware and intelligent software.*

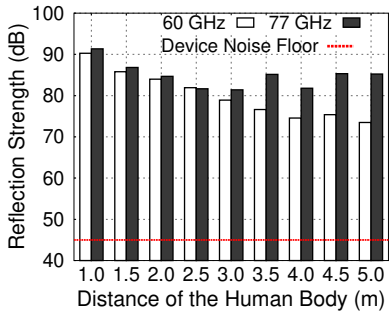


Figure 25. Strength comparison between 60 GHz and 77 GHz reflections from the human body. Median strength of 60 GHz is lower across all distances because it is absorbed more by the body. Still, the strength is much higher than the device noise floor, allowing for accurate silhouette generation and joint location predictions.

5.1.9 Human Reflection Strength Comparison between 77 GHz and 60 GHz. *MiShape* has been evaluated using a pair of 77 GHz mmWave transceivers; however, next-generation mmWave networking devices, such as the home wireless routers, could operate at various frequencies, including the most popular 60 GHz mmWave band following the IEEE 802.11ad standard [95]. Since 60 GHz signals could be absorbed more by the human body compared to the 77 GHz signals, the body silhouette generation and joint location prediction could be challenging for 60 GHz mmWave devices. Therefore, a comparative study on the effects of reflections from human body for 60 GHz *w.r.t.* 77 GHz is critical. To this end, we use a 60 GHz transceiver, TI IWR6843ISK [96] and measure the reflection strengths from human body at various distances from the device. We have a volunteer in front of our set up at different distances, from 1 m to 5 m with an increment of 0.5 m, and for each distance, we locate the range bin corresponding to the location of the human in the signal reflection profile, and then find the reflected signal strength. We then calculate the median strength across 3000 frames for each distance. Since the 60 GHz device uses a 3.07 GHz bandwidth, it can resolve reflections in range with a resolution of 4.84 cm. To minimize the effect of multipath, we ensure that the line-of-sight between the device and the body is open, and the other objects and walls are at least 5 m away from the human body. We also repeat the experiments for the 77 GHz mmWave transceiver under identical conditions. Finally, we calibrate the 77 GHz reflection strengths to account for the known transmit power difference between the 60 and 77 GHz transceivers. Figure 25 shows the comparative results of the reflection strengths for 60 and 77 GHz, and the strength is consistently lower at 60 GHz across all distances. This is because 60 GHz signals are absorbed more by the human body compared the 77 GHz signals. At a higher distance, such as 5 m, this difference could be even higher due to the effect of the multipath. However, we observe that the absolute signal strength is still very high, irrespective of the distance: Even at 5 m, 60 GHz reflection strength from the human body is 28.48 dB higher than the average device noise floor. *These results demonstrate that, although 60 GHz signals could have higher absorption by the human body compared to the 77 GHz signals, the absolute signal strength is strong enough to produce good quality silhouettes and joint locations.*

5.2 Application Results with *MiShape*

We now evaluate *MiShape*'s ability to enable gait monitoring applications with COTS mmWave devices. To this end, we collect datasets from a young, healthy volunteer for three walking speeds: fast, normal, and slow (poses 15, 16, and 17 in Table 4). We ask the volunteer to walk around the room and capture the reflected signals and the ground truth body joint locations at 25 fps. Then, we feed the reflected signals to *MiShape*, which outputs silhouettes and 25 joint locations frame-by-frame. Finally, we measure the three gait parameters, *Step length*, *Stride length*, and *Cadence* (Section 3.6) from the ground truth and *MiShape*'s predicted joint locations. Figures 26(a–c) show the results for fast, normal, and slow walking trajectories, respectively. We see that for all walking

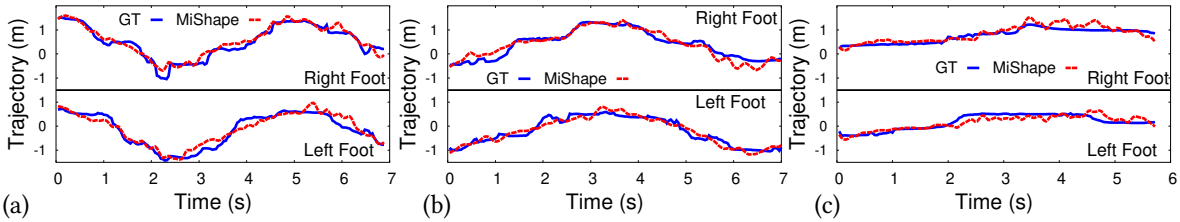


Figure 26. *MiShape* follows ground truth walking trajectories well for three walking types: (a) Fast; (b) Normal; (c) Slow.

Table 6. *MiShape*'s *Cadence* prediction (steps/minute).

Type	Ground Truth	*MiShape*	Acceptable range
Fast walk	148.2	156.9	> 135
Normal walk	129.7	129.7	120-135
Slow walk	115	114	<120

trajectories, *MiShape* can follow the ground truth reasoably well. Further, Figures 27(a–b) show the absolute *Step length* and *Stride length* prediction errors in *MiShape*. It predicts the *Step length* with a median absolute error of 0.19 m, 0.22 m, and 0.13 m for fast walk, normal walk, and slow walk, respectively. Furthermore, it can predict the *Stride length* with a median absolute error of 0.2 m, 0.25 m, and 0.1 m, respectively, which are tolerable in practice. By improving *MiShape*'s JE network performance with fine-tuning, we can minimize potentially

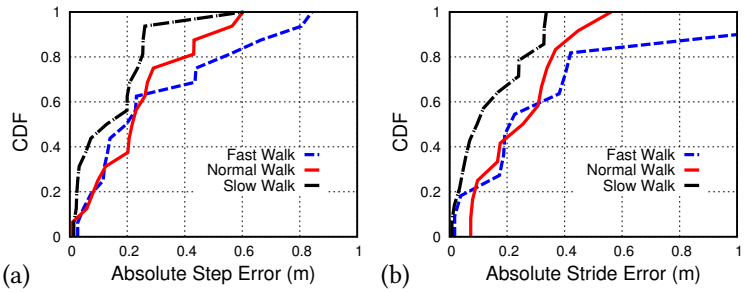


Figure 27. *MiShape*'s gait parameters' prediction errors for 3 walking speeds: (a) *Step length* error. (b) *Stride length* error.

minimize these errors. Table 6 also shows the predicted *Cadence* results from *MiShape*. For both normal and slow walk, the predicted *Cadence* values match with the ground truth. The predicted *Cadence* for fast walk is higher than the ground truth; still, *MiShape* can classify the type of walk accurately since the predicted value is greater than the acceptable range (>135 for a fast walk). *In summary, MiShape can potentially enable gait monitoring applications from COTS mmWave devices and works well for different walking types.*

6 RELATED WORK

Human Pose Estimation: Vision-based systems achieve excellent performance in human motion estimation using optical cameras, depth sensors, and LiDARs. The approaches in these systems can be broadly classified into top-down and bottom-up: In top-down, these systems first identify each person in an image and then performs aggregation to obtain key points, and in the bottom-up, they first detect all key points in an image and then map these points to the same person. Both these approaches are popular in industry and research [97, 98]. Apart from RGB-D cameras, LiDARs have been used to improve accuracy in human pose estimation [12, 99], and

state-of-the-art systems have achieved almost 99% accuracy in correctly estimating pose for real-time tracking [100]. But these systems are limited by poor lighting and occlusion, and they are privacy-invasive.

RF imaging can overcome these limitations and has achieved reasonable accuracy in estimating human poses. They mostly use Wi-Fi signals either in the form of spatial heatmap or CSI information from multiple receivers [18–20, 31] and generate key points for humans in a 2D or 3D space, but their image quality is far from existing vision-based approaches. To match the quality of vision-based systems, RF imaging systems need to learn features in the RF signal and correlate with the visual images. To this end, [101] formulated the problem of 3D pose estimation as identifying the locations of 14 anatomical key points on the body from RF signals. It follows a similar approach as in [19] but decomposes RF signals into two 3D heatmaps and incorporates temporal information. Furthermore, [18] predicts 3D location from CSI by implementing RNNs. Such tracking is useful in enabling healthcare applications, such as gait or exercise monitoring. RFID or IMU based tracking can also enable these applications [102, 103], but these systems require extra hardware to be attached to the body, which could be cumbersome. [25, 26, 104, 105] have also explored predicting key points of humans from mmWave signals. For example, [26] takes the 3D point clouds generated from mmWave signals, similar to [106], learns the key points from those points, and then passes the points through an open-source human-mesh generator to create general body shapes [32]. However, due to variable reflectivity and specularity, the performance could be poor. *MiShape* aims to improve the quality of the mmWave images by employing deep learning models, that not only produce correct silhouettes on par with existing vision-based systems but also work consistently across multiple subjects.

Enhancing Resolution with Learning: To improve the resolution, traditional approaches, such as Bicubic, Nearest Neighbor, Sparse Coding, *etc.*, use interpolation and fixed models to upsample the images [63, 64, 107]. These methods mostly generate blurry images, which lack finer details, and they are unable to remove image artifacts. Recently, super-resolution deep learning models promise to improve the quality by training with thousands of examples of low-resolution and high-resolution images as input-output pairs [108, 109]. Unfortunately, such methods could not be applied to mmWave images directly since the images are either aliased with spurious information or have poor resolutions with many missing body parts. Existing cGAN models have been able to successfully enhance the mmWave image resolution through adversarial learning [13, 26, 94]. However, they work with a specific mmWave antenna configuration and do not address the challenges with aliasing problems in COTS mmWave devices. In contrast, *MiShape* is designed to generate high-resolution human silhouettes and 3D joint locations using mmWave reflected signals from commodity devices, and it works well across many practical antenna configurations under real environmental conditions.

7 DISCUSSION AND FUTURE WORKS

In this work, we have designed and evaluated *MiShape* on a COTS mmWave Radar; however, for at-home monitoring, we need to deploy *MiShape* on a typical mmWave networking device, such as routers/access points. The standard Commercial-Off-The-Shelf (COTS) mmWave networking devices, such as those following the IEEE 802.11ad standard [95], already have the capability to resolve the reflections from multiple directions and operate on a 2 GHz bandwidth; thus, they can achieve range resolutions on the order 7.5 cm. In contrast, the mmWave devices used in *MiShape* evaluation operate on a lower bandwidth, 1.62 GHz, and can achieve a range resolution of 9.26 cm only. Higher range resolution means a device can distinguish reflections from two different points with more confidence. So, we expect the performance of *MiShape* would be even better on COTS mmWave communication devices. Furthermore, the networking devices can measure the channel response, which carries the amplitude and phase information of the reflected signals at different range, from each standard packet using the Channel Estimation (CE) header field [95]. They will not require any modifications to the communications protocols or packet formats. However, to capture the reflected signals from a transmitted packet, a COTS mmWave networking device needs to switch between Tx and Rx mode within nanoseconds. Such capability might be too

stringent for low-cost COTS mmWave devices [95, 110], and may require device firmware or antenna hardware modifications. Besides, the COTS mmWave networking devices do not provide raw signal reflections yet. So, we have designed a custom setup integrating COTS mmWave Radars for data collection, test, and evaluation. In the future, we propose to extend *MiShape* to work with COTS networking devices, such as [45]. One option could be to find avenues for switching between Tx and Rx mode quickly by modifying the device firmware. Another option could be to use networking device with multiple phased-array antennas [51], and use one as Tx and another as Rx. We will thoroughly investigate the issues that arise and benefits that accrue from designing and implementing *MiShape* on a COTS mmWave networking device.

8 CONCLUSION

In this work, we present *MiShape*, an imaging system that can generate high-quality human silhouettes and predicts 3D locations of body joints on par with existing vision-based systems. The system employs customized deep learning models to overcome the challenges of poor image resolution, specularity of signal reflections, and image aliasing in the COTS mmWave system. The experimental results demonstrate that *MiShape* generalizes to multiple subjects with little fine-tuning and works well for systems with different mmWave hardware capabilities. We have customized *MiShape* for gait monitoring applications, but the system can be adapted to facilitate other tracking and monitoring applications. We believe *MiShape* can unlock the potential of 5G mmWave systems, such as home wireless routers, in facilitating privacy-noninvasive, high-quality imaging.

ACKNOWLEDGMENTS

We sincerely thank the reviewers for their comments and feedback. This work is partially supported by the NSF under grants CNS-1910853, MRI-2018966, and CAREER-2144505 and by the UofSC ASPIRE II award.

REFERENCES

[1] Yao, Lina and Sheng, Quan Z. and Ruan, Wenjie and Gu, Tao and Li, Xue and Falkner, Nick and Yang, Zhi, "RF-Care: Device-Free Posture Recognition for Elderly People Using A Passive RFID Tag Array," in *Proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services on 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2015.

[2] Wang, Xueyi and Ellul, Joshua and Azzopardi, George, "Elderly Fall Detection Systems: A Literature Survey," *Frontiers in Robotics and AI*, vol. 7, 2020.

[3] Dawes, A J and Lin, A Y and Varghese, C and Russell, M M and Lin, A Y, "Mobile Health Technology for Remote Home Monitoring After Surgery: A Meta-Analysis," *British Journal of Surgery*, vol. 21, 10 2021.

[4] Ghazi, Mustafa A. and Ding, Lei and Fagg, Andrew H. and Kolobe, Thubi H.A. and Miller, David P., "Vision-Based Motion Capture System for Tracking Crawling Motions of Infants," in *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2017.

[5] Anh L. Bui and Gregg C. Fonarow, "Home Monitoring for Heart Failure Management," *Journal of the American College of Cardiology*, vol. 59, no. 2, 2012.

[6] Sica, Marco et al , "Continuous Home Monitoring of Parkinson's Disease using Inertial Sensors," *PLOS ONE*, no. 2, 2021.

[7] Tiersen, Federico et al, "Smart Home Sensing and Monitoring in Households With Dementia," *JMIR Aging*, no. 3, 2021.

[8] Buzzelli, Marco and Albe, Alessio and Ciocca, Gianluigi, "A Vision-Based System for Monitoring Elderly People at Home," *Applied Sciences*, vol. 10, no. 1, 2020.

[9] Gutierrez J, Rodriguez V and Martin S, "Comprehensive Review of Vision-Based Fall Detection Systems," *Sensors*, vol. 21, no. 3, 2021.

[10] Y. Feng and L. Max, "Accuracy and Precision of a Custom Camera-Based System for 2-D and 3-D Motion Tracking During Speech and Nonspeech Motor Tasks," *Journal of Speech, Language, and Hearing Research*, vol. 57, 2014.

[11] Sun Guanghao, et al., *Noncontact Monitoring of Vital Signs with RGB and Infrared Camera and Its Application to Screening of Potential Infection*, 1st ed. IntechOpen, 2018.

[12] Michael Furst and Shriya T. P. Gupta and Rene Schuster and Oliver Wasenmuller and Didier Stricker, "HPERL: 3D Human Pose Estimation from RGB and LiDAR," 2020. [Online]. Available: https://arxiv.org/abs/2010.08221

[13] Guan, Junfeng and Madani, Sohrab and Jog, Suraj and Gupta, Saurabh and Hassanieh, Haitham, "Through Fog High-Resolution Imaging Using Millimeter Wave Radar," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Verizon, "Verizon 5G Home Router," 2022. [Online]. Available: https://www.verizon.com/support/knowledge-base-220089/

[15] Roger Appleby, Duncan A. Robertson and David Wikner, "Millimeter Wave Imaging: A Historical Review," in *Proc. SPIE*, 2017.
[16] "ProVision Automatic Target Detection," 2015. [Online]. Available: http://www.sds.l-3com.com/advancedimaging/provision-at.htm
[17] K. Mowery, E. Wustrow, T. Wypych, C. Singleton, C. Comfort, E. Rescorla, S. C. J. A. Halderman, and H. Shacham, "Security Analysis of a Full-Body Scanner," in *USENIX Security Symposium*, 2014.
[18] Jiang, Wenjun and Xue, Hongfei and Miao, Chenglin and Wang, Shiyang and Lin, Sen and Tian, Chong and Murali, Srinivasan and Hu, Haochen and Sun, Zhi and Su, Lu, "Towards 3D Human Pose Construction Using WiFi," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.
[19] Zhao, Mingmin and Li, Tianhong and Alsheikh, Mohammad Abu and Tian, Yonglong and Zhao, Hang and Torralba, Antonio and Katabi, Dina, "Through-Wall Human Pose Estimation Using Radio Signals," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
[20] Fadel Adib and Zach Kabelac and Dina Katabi and Robert C. Miller, "3D Tracking via Body Radio Reflections," in *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*, 2014.
[21] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the Human Figure Through a Wall," in *ACM SIGGRAPH Asia*, 2015.
[22] C. S. Pros, "CCTV." [Online]. Available: https://www.cctvsecuritypros.com
[23] "RoboRealm. Microsoft Kinect, 2013." [Online]. Available: http://www.roborealm.com/help/MicrosoftKinect.php
[24] "Vicon." [Online]. Available: https://www.vicon.com/applications/life-sciences/gait-analysis-neuroscience-and-motor-control/
[25] Zhang, Feng and Wu, Chenshu and Wang, Beibei and Liu, K. J. Ray, "mmEye: Super-Resolution Millimeter Wave Imaging," *IEEE Internet of Things Journal*, vol. 8, no. 8, 2021.
[26] Xue, Hongfei and Ju, Yan and Miao, Chenglin and Wang, Yijiang and Wang, Shiyang and Zhang, Aidong and Su, Lu, "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*, 2021.
[27] M. Skolnik, *Introduction to Radar Systems*. McGraw-Hill Book Co., 1962.
[28] IEEE Standards Association, "IEEE Standards 802.11ad-2012, Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band," goo.gl/r2JeYd, 2012.
[29] Wu, Ting and Rappaport, Theodore S. and Collins, Christopher M., "The Human Body and Millimeter-Wave Wireless Communication Systems: Interactions and Implications," in *2015 IEEE International Conference on Communications (ICC)*, 2015.
[30] Sanjib Sur and Vignesh Venkateswaran and Xinyu Zhang and Parmesh Ramanathan, "60 GHz Indoor Networking through Flexible Beams: A Link-Level Profiling," in *Proc. of ACM SIGMETRICS*, 2015.
[31] Adib, Fadel and Katabi, Dina, "See through Walls with WiFi!" in *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, 2013.
[32] Zhao, Mingmin and Liu, Yingcheng and Raghu, Aniruddh and Zhao, Hang and Li, Tianhong and Torralba, Antonio and Katabi, Dina, "Through-Wall Human Mesh Recovery Using Radio Signals," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
[33] Wang, Weiwei and Yang, Kehu, "A Method for Millimeter-Wave Imaging of Concealed Objects Via De-Aliasing," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[34] Kazemi, Mahmoud and Kavehvash, Zahra and Shabany, Mahdi, "K-Space Analysis of Aliasing in Millimeter-Wave Imaging Systems," *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 3, 2021.
[35] Gao, Xiangyu and Roy, Sumit and Xing, Guanbin, "MIMO-SAR: A Hierarchical High-Resolution Imaging Algorithm for mmWave FMCW Radar in Autonomous Driving," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, 2021.
[36] Jing, Handan and Li, Shiyong and Miao, Ke and Wang, Shuoguang and Cui, Xiaoxi and Zhao, Guoqiang and Sun, Houjun, "Enhanced Millimeter-Wave 3-D Imaging via Complex-Valued Fully Convolutional Neural Network," *Electronics*, vol. 11, no. 1, 2022.
[37] Lam H. Nguyen, "Millimeter-wave forward-looking 3-D SAR imaging challenges," in *Passive and Active Millimeter-Wave Imaging*, 2019.
[38] Evan C. Zaugg and David G. Long, "Generalized Frequency Scaling and Backprojection for LFM-CW SAR Processing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, 2015.
[39] Watts, Claire M. and Lancaster, Patrick and Pedross-Engel, Andreas and Smith, Joshua R. and Reynolds, Matthew S., "2D and 3D Millimeter-Wave Synthetic Aperture Radar Imaging on a PR2 Platform," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.
[40] Texas Instruments, "IWR1443 Single-Chip 76-GHz to 81-GHz MmWave Sensor Evaluation Module," 2020. [Online]. Available: https://www.ti.com/tool/IWR1443BOOST
[41] Gamesto, "Kinect," 2022. [Online]. Available: https://www.gamestop.com/gaming-accessories/controllers/xbox-one/products/microsoft-kinect-for-xbox-one/10115985.html
[42] Hao Zhu and Xinxin Zuo and Sen Wang and Xun Cao and Ruigang Yang, "Detailed Human Shape Estimation from a Single Image by Hierarchical Mesh Deformation," 2019. [Online]. Available: https://arxiv.org/abs/1904.10506
[43] Wang, Z. and Simoncelli, E.P. and Bovik, A.C., "Multiscale Structural Similarity for Image Quality Assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, 2003.
[44] Mehrdad Soumekh, *Synthetic Aperture Radar Signal Processing*, 1st ed. John Wiley & Sons, Inc., 1999.
[45] NETGEAR, Inc., "Nighthawk X10 Smart WiFi Router," 2022. [Online]. Available: https://www.netgear.com/landings/ad7200/

[46] TP-Link Corporation Limited, "Talon AD7200 Multi-Band Wi-Fi Router," 2022. [Online]. Available: https://www.tp-link.com/us/home-networking/wifi-router/ad7200/
[47] IgniteNet, "MetroLinq 10G Tri-Band Omni," 2022. [Online]. Available: https://www.ignitenet.com/wireless-backhaul/ml-10g-omni/
[48] MikroTik, "wAP 60G," 2022. [Online]. Available: https://mikrotik.com/product/wap_60g
[49] Sivers Semiconductors AB, "60 GHz Evaluation Kits (EVK)," 2022. [Online]. Available: https://www.sivers-semiconductors.com/sivers-wireless/evaluation-kits/
[50] Rojhani, Neda and Passafiume, Marco and Lucarelli, Matteo and Collodi, Giovanni and Cidronali, Alessandro, "Assessment of Compressive Sensing 2x2 MIMO Antenna Design for Millimeter-Wave Radar Image Enhancement," *Electronics*, vol. 9, no. 4, 2020.
[51] Airfide Networks, "Airfide Brings High Performance Home and Enterprise 5G-NR Wireless," 2022. [Online]. Available: https://airfidenet.com/
[52] Xinyu Zhang, "M-Cube: An Open-Source Programmable Millimeter-Wave Experimental Platform," 2022. [Online]. Available: http://m3.ucsd.edu/
[53] tsne, "tsne," 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
[54] Brekel, "Face V1," 2022. [Online]. Available: https://brekel.com/
[55] Cleveland, Jerika and Lewis, Jacob and Mitra, Dipankar and Braaten, Benjamin D and Allen, Jeffery and Allen, Monica, "On the Image Analysis of Conducting Magneto-Responsive Micro-Particles for Applications in Leaky Wave Antenna Beam Steering," in *2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting*. IEEE, 2020.
[56] Lisa Jamhoury, "Understanding Kinect V2 Joints and Coordinate System," 2022. [Online]. Available: https://lisajamhoury.medium.com/understanding-kinect-v2-joints-and-coordinate-system-4f4b90b9df16
[57] Schellberg, Jacqueline M and Sur, Sanjib, *ViSAR: A Mobile Platform for Vision-Integrated Millimeter-Wave Synthetic Aperture Radar*. Association for Computing Machinery, 2021.
[58] Xun Huang and Yixuan Li and Omid Poursaeed and John Hopcroft and Serge Belongie, "Stacked Generative Adversarial Networks," 2017. [Online]. Available: https://arxiv.org/abs/1612.04357
[59] Denton, Emily L and Chintala, Soumith and szlam, arthur and Fergus, Rob, "Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks," in *Advances in Neural Information Processing Systems*, 2015.
[60] Xiaolong Wang and Abhinav Gupta, "Generative Image Modeling using Style and Structure Adversarial Networks," 2016. [Online]. Available: https://arxiv.org/abs/1603.05631
[61] Goodfellow, Ian, et al., "Generative Adversarial Networks," *Commun. ACM*, vol. 63, no. 11, 2020.
[62] Mehdi Mirza and Simon Osindero, "Conditional Generative Adversarial Nets," 2014. [Online]. Available: https://arxiv.org/abs/1411.1784
[63] R. E. Carlson and F. N. Fritsch, "Monotone Piecewise Bicubic Interpolation," *SIAM Journal on Numerical Analysis*, vol. 22, no. 2, 1985.
[64] Song, Yibing and Gong, Lijun, "Analysis and Improvement of Joint Bilateral Upsampling for Depth Image Super-Resolution," in *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*. IEEE, 2016.
[65] Christian Ledig and Lucas Theis and Ferenc Huszar and Jose Caballero and Andrew Cunningham and Alejandro Acosta and Andrew Aitken and Alykhan Tejani and Johannes Totz and Zehan Wang and Wenzhe Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," 2017. [Online]. Available: https://arxiv.org/abs/1609.04802
[66] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Kai Li and Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[67] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556
[68] Roegiers, Sanne et al. , "Human Action Recognition using Hierarchic Body Related Occupancy Maps," *Integrated Computer-Aided Engineering*, vol. 26, no. 3, 2019.
[69] G. Chen, C. Patten, D. H. Kothari, and F. E. Zajac, "Gait Deviations Associated with Post-Stroke Hemiparesis: Improvement During Treadmill Walking using Weight Support, Speed, Support Stiffness, and Handrail Hold," *Gait & Posture*, vol. 22, no. 1, 2005.
[70] Anouk Lamontagne and Joyce Fung, "Faster is Better: Implications for Speed-Intensive Gait Training After Stroke," *Stroke*, vol. 35, no. 11, 2004.
[71] Neil F. Gordon and Meg Gulanick and Fernando Costa and Gerald Fletcher and Barry A. Franklin and Elliot J. Roth and Tim Shephard, "Physical Activity and Exercise Recommendations for Stroke Survivors: An American Heart Association Scientific Statement from the Council on Clinical Cardiology, Subcommittee on Exercise, Cardiac Rehabilitation, and Prevention; the Council on Cardiovascular Nursing; the Council on Nutrition, Physical Activity, and Metabolism; and the Stroke Council," *Circulation*, vol. 109, no. 16, 2004.
[72] Stacy L. Fritz and Ashlee L. Pittman and Anna C. Robinson and Skylar C. Orton and Erin D. Rivers, "An Intense Intervention for Improving Gait, Balance, and Mobility for Individuals with Chronic Stroke: A Pilot Study," *Journal of Neurologic Physical Therapy*, vol. 31, no. 2, 2007.
[73] Chen-Yu Hsu, et al., "Enabling Identification and Behavioral Sensing in Homes using Radio Reflections," in *ACM CHI*, 2019.
[74] Wang, Wei and Liu, Alex X. and Shahzad, Muhammad, "Gait Recognition Using Wifi Signals," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.

[75] Gerald F. Harris and Jacqueline J. Wertsch, "Procedures for gait analysis," *Archives of Physical Medicine and Rehabilitation*, vol. 75, no. 2, 1994.

[76] John H. Hollman and Eric M. McDade and Ronald C. Petersen, "Normative Spatiotemporal Gait Parameters in Older Adults," *Gait and Posture*, vol. 34, no. 1, 2011.

[77] Amanda E. Chisholm and Shelley Makepeace and Elizabeth L. Inness and Stephen D. Perry and William E. McIlroy and Avril Mansfield, "Spatial-Temporal Gait Variability Poststroke: Variations in Measurement and Implications for Measuring Change," *Archives of Physical Medicine and Rehabilitation*, vol. 95, no. 7, 2014.

[78] Ion Martinikorena and Alicia MartÃınez-Ramirez and Marisol Gomez and Pablo Lecumberri and Alvaro Casas-Herrero and Eduardo L. Cadore and Nora Millor and Fabricio Zambom-Ferraresi and Fernando Idoate and Mikel Izquierdo, "Gait Variability Related to Muscle Quality and Muscle Power Output in Frail Nonagenarian Older Adults," *Journal of the American Medical Directors Association*, vol. 17, no. 2, 2016.

[79] Renata Noce Kirkwood and Bruno de Souza Moreira and Marcia L.D.C. Vallone and Sueli Aparecida Mingoti and Rosangela Correa Dias and Rosana Ferreira Sampaio, "Step Length Appears to be a Strong Discriminant Gait Parameter for Elderly Females Highly Concerned about Falls: A Cross-Sectional Observational Study," *Physiotherapy*, vol. 97, no. 2, 2011.

[80] Texas Instruments, "DCA1000EVM: Real-time Data-Capture Adapter for Radar Sensing Evaluation Module," 2020. [Online]. Available: https://www.ti.com/tool/DCA1000EVM

[81] Palermo Physiotherapy, "Palermo Physiotherapy and Wellness Center," 2022. [Online]. Available: https://palermophysio.ca/yoga-basics-part-2-most-common-poses/

[82] Henry, Kristin D and Rosemond, Cherie and Eckert, Lynn B, "Effect of number of home exercises on compliance and performance in adults over 65 years of age," *Physical Therapy*, vol. 79, no. 3, 1999.

[83] Moraes, Alberto da Rocha and Sanches, Monique Lalue and Ribeiro, Eduardo Cotecchia and Guimarães, Antonio Sérgio, "Therapeutic exercises for the control of temporomandibular disorders," *Dental press journal of orthodontics*, vol. 18, no. 5, 2013.

[84] Ann Pizer, "VeryWellFit," 2022. [Online]. Available: https://www.verywellfit.com/essential-yoga-poses-for-beginners-3566747

[85] Open-Source, "TensorFlow," 2022. [Online]. Available: https://www.tensorflow.org/

[86] ——, "Spyder IDE," 2022. [Online]. Available: https://www.spyder-ide.org/

[87] ——, "ANACONDA," 2022. [Online]. Available: https://www.anaconda.com/

[88] NVIDIA, "GEFORCE," 2022. [Online]. Available: https://www.nvidia.com/en-us/geforce/

[89] Google, "Cloud TPU," 2022. [Online]. Available: https://cloud.google.com/tpu

[90] Ferzli, Rony and Karam, Lina J, "A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB)," *IEEE transactions on image processing*, vol. 18, no. 4, 2009.

[91] Tolga Bridal, "Sharpness Estimation From Image Gradients," 2022. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/32397-sharpness-estimation-from-image-gradients

[92] Mortazavi, Fatemeh and Nadian, Ali, "Stability of Kinect for Range of Motion Analysis in Static Stretching Exercises," *PLOS ONE*, vol. 13, p. e0200992, 07 2018.

[93] Zhu, Weiqiang and Mousavi, S Mostafa and Beroza, Gregory C, "Seismic signal augmentation to improve generalization of deep neural networks," in *Advances in geophysics*. Elsevier, 2020, vol. 61, pp. 151–177.

[94] Regmi, Hem and Saadat, Moh Sabbir and Sur, Sanjib and Nelakuditi, Srihari, "SquiggleMilli: Approximating SAR Imaging on Mobile Millimeter-Wave Devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 3, 2021.

[95] IEEE Standards Association, "IEEE Standards 802.11ad-2012: Enhancements for Very High Throughput in the 60 GHz Band," 2012.

[96] Texas Instruments, "IWR6843 Single-Chip 60-GHz MmWave Sensor Evaluation Module," 2020. [Online]. Available: https://www.ti.com/tool/IWR6843ISK

[97] George Papandreou and Tyler Zhu and Nori Kanazawa and Alexander Toshev and Jonathan Tompson and Chris Bregler and Kevin Murphy, "Towards Accurate Multi-person Pose Estimation in the Wild," 2017. [Online]. Available: https://arxiv.org/abs/1701.01779

[98] Zhe Cao and Gines Hidalgo and Tomas Simon and Shih-En Wei and Yaser Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," 2019. [Online]. Available: https://arxiv.org/abs/1812.08008

[99] Khalife, Joe and Ragothaman, Sonya and Kassas, Zaher M., "Pose Estimation with LiDAR Odometry and Cellular Pseudoranges," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[100] Tasnim, Nusrat and Islam, Mohammad Khairul and Baek, Joong-Hwan, "Deep Learning Based Human Activity Recognition Using Spatio-Temporal Image Formation of Skeleton Joints," *Applied Sciences*, vol. 11, no. 6, 2021.

[101] Zhao, Mingmin and Tian, Yonglong and Zhao, Hang and Alsheikh, Mohammad Abu and Li, Tianhong and Hristov, Rumen and Kabelac, Zachary and Katabi, Dina and Torralba, Antonio, "RF-Based 3D Skeletons," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018.

[102] Jin, Haojian and Yang, Zhijian and Kumar, Swarun and Hong, Jason I., "Towards Wearable Everyday Body-Frame Tracking Using Passive RFIDs," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, 2018.

[103] Anderson, Boyd and Shi, Mingqian and Tan, Vincent Y. F. and Wang, Ye, "Mobile Gait Analysis Using Foot-Mounted UWB Sensors," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, 2019.
[104] Sengupta, Arindam and Jin, Feng and Zhang, Renyuan and Cao, Siyang, "mm-Pose: Real-Time Human Skeletal Posture Estimation Using mmWave Radars and CNNs," *IEEE Sensors Journal*, vol. 20, no. 17, 2020.
[105] Arindam Sengupta and Siyang Cao, "mmPose-NLP: A Natural Language Processing Approach to Precise Skeletal Pose Estimation using mmWave Radars," 2021. [Online]. Available: https://arxiv.org/abs/2107.10327
[106] Qian, Kun and He, Zhaoyuan and Zhang, Xinyu, "3D Point Cloud Generation with Millimeter-Wave Radar," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, 2020.
[107] Egiazarian, Karen and Katkovnik, Vladimir, "Single Image Super-Resolution via BM3D Sparse Coding," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015.
[108] Chao Dong and Chen Change Loy and Kaiming He and Xiaoou Tang, "Image Super-Resolution Using Deep Convolutional Networks," 2015. [Online]. Available: https://arxiv.org/abs/1501.00092
[109] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep Learning for Single Image Super-Resolution: A Brief Review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, 2019.
[110] Sabbir Saadat and Sanjib Sur and Srihari Nelakuditi and Parmesh Ramanathan, "MilliCam: Hand-held Millimeter-Wave Imaging," in *IEEE International Conference on Computer Communications and Networks (ICCCN)*, 2020.

A MULTIPLE SILHOUETTE RECONSTRUCTION RESULTS

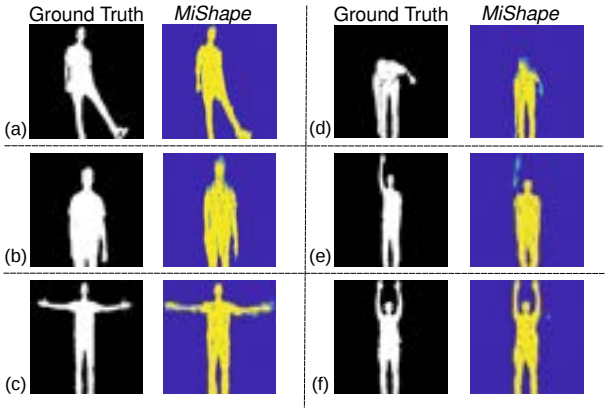


Figure 28. Examples of static human silhouettes generated by *MiShape*: (a) Leg extension; (b) Lunges; (c) Arm extension; (d) Toe touchdown; (e) One arm up; (f) Both arms up.

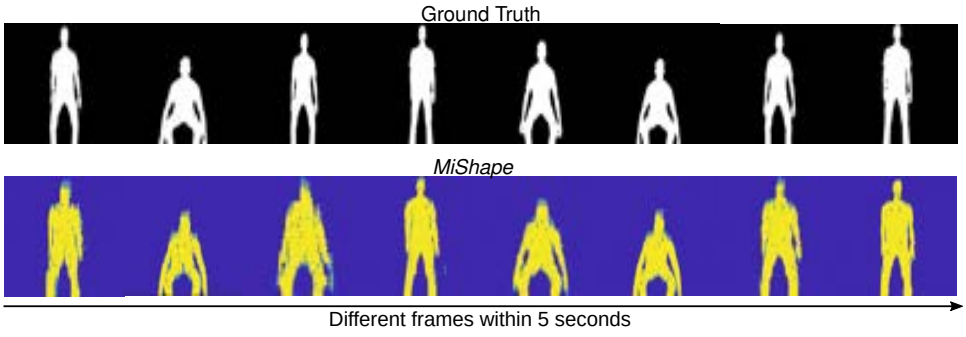


Figure 29. Examples of dynamic human silhouettes generated by *MiShape* for a video frame of 5 seconds.