Construct 3D Hand Skeleton with Commercial WiFi

Sijie Ji*, Xuanye Zhang*, Yuanqing Zheng+, Mo Li†*
†The Hong Kong University of Science and Technology,
*Nanyang Technological University, +The Hong Kong Polytechnic University
Email:{sijie001, c200212}@ntu.edu.sg, yqzheng@polyu.edu.hk, lim@cse.ust.hk

ABSTRACT

This paper presents HandFi, which constructs hand skeletons with practical WiFi devices. Unlike previous WiFi hand sensing systems that primarily employ predefined gestures for pattern matching, by constructing the hand skeleton, HandFi can enable a variety of downstream WiFi-based hand sensing applications in gaming, healthcare, and smart homes. Deriving the skeleton from WiFi signals is challenging, especially because the palm is a dominant reflector compared with fingers. HandFi develops a novel multi-task learning neural network with a series of customized loss functions to capture the low-level hand information from WiFi signals. During offline training, HandFi takes raw WiFi signals as input and uses the leap motion to provide supervision. During online use, only with commercial WiFi, HandFi is capable of producing 2D hand masks as well as 3D hand poses. We demonstrate that HandFi can serve as a foundation model to enable developers to build various applications such as finger tracking and sign language recognition, and outperform existing WiFi-based solutions. Artifacts can be found: https://github.com/SIJIEJI/HandFi

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Computing methodologies** → *Neural networks.*

KEYWORDS

Wireless Sensing, Multi-task Learning, 3D Hand Pose, Gesture Recognition

ACM Reference Format:
Sijie Ji, Xuanye Zhang, Yuanqing Zheng, Mo Li. 2023. Construct 3D Hand Skeleton with Commercial WiFi. In *ACM Conference on Embedded Networked Sensor Systems (SenSys '23), November 12–17, 2023, Istanbul, Turkiye.* ACM, Istanbul, Turkiye, 13 pages. https://doi.org/10.1145/3625687.3625812

1 INTRODUCTION

Hand sensing plays a crucial role in human-computer interaction, enabling a broad range of applications in video games, education, and healthcare. These applications further create new opportunities for people with communication disabilities to interact with devices and others. Numerous hand-sensing systems have been developed, wherein wearable-based solutions offer high accuracy but are limited by their cumbersome nature [47]. On the other hand, contact-free solutions offer greater flexibility and versatility. However, acoustic-based solutions are susceptible to environmental noise and have a limited sensing range [42, 49], while computer vision-based solutions may raise privacy concerns. In contrast, RF-based solutions offer a wider sensing range, preserve privacy and are robust to illumination variation. Among these technologies, WiFi-based solutions are promising to repurpose ubiquitous low-cost WiFi devices for contact-free hand sensing.

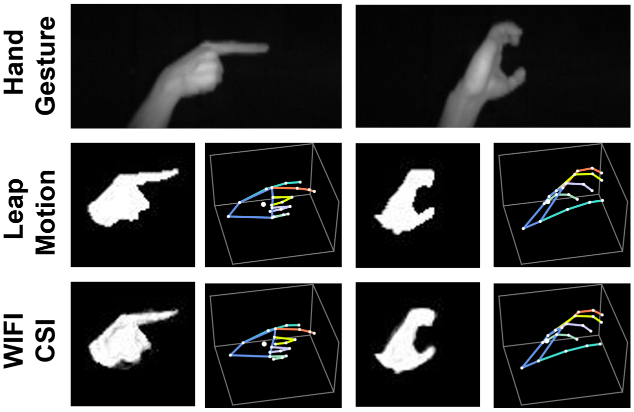


Figure 1: Leap Motion Output VS HandFi Output.

However, due to the low spatial resolution of WiFi signals and the small scale of a hand, directly modeling the radio reflection from a hand is challenging. Therefore, existing WiFi-based solutions rely on WiFi signal patterns caused by hand motions to their corresponding hand gestures [17, 21, 30, 36, 46, 54], or employ geometric constraints to track variations in signal propagation for hand tracking [33, 37, 43, 52, 53]. None of the existing WiFi-based hand-sensing systems directly models the relationship between the hand skeleton and the reflected signals of interests. As a result, specific patterns or models need to be identified and built for different downstream applications. Thus, these gesture recognition systems are typically limited to a small pre-defined set of hand gestures, while signal propagation modeling-based hand tracking solutions are vulnerable to noise and environmental changes. Nonetheless, the existing WiFi-based hand sensing systems show that WiFi signals do carry information about human hands and hand movements. In addition, WiFi has the potential to achieve millimeter-level sensing resolution [14, 41]. The challenge lies in how to separate the signal of interest of the target reflector from the environment-related multi-path and hardware imperfection-induced noise, which are

SenSys '23, November 12–17, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0414-7/23/11…$15.00
https://doi.org/10.1145/3625687.3625812

often nonlinearly superimposed. On the other hand, deep learning is particularly well-suited for extracting the signal of interest from the nonlinear superimposed high-dimensional data. A deep learning model could potentially learn the intricate relationships of the superimposed WiFi CSI and extract the reflected signal from the hand through proper design of neural networks. The question remains whether it is possible to extract rich hand semantic information from WiFi signals and achieve vision-like results such as 3D hand pose construction with the novel design of neural networks. If such a result can be achievable, we will be able to build various downstream applications directly and integrally without being limited by pre-defined hand gestures.

This paper explores the possibility of obtaining vision-like results solely using commercial WiFi by presenting HandFi. HandFi is capable of constructing the shape and skeleton of a hand simultaneously. Figure 1 presents the output of HandFi compared with the output of a commercial depth camera (leap motion). The hand shape is represented by a binary matrix-based hand mask, while the hand skeleton is represented by a set of vectors indicating the 3D coordinates of 21 key joints of a hand. The core of HandFi is a novel deep neural network called HandNet, which is trained in an end-to-end manner with cross-modality supervision using labels obtained from the depth camera. Once a model is trained, HandFi can directly infer hand shape and skeleton, which can be used to support a variety of downstream applications with flexible extension. For example, finger tracking can be directly enabled by tracking the coordinate of an index finger in 3D hand skeleton. Based on the 3D hand skeleton, a new hand gesture can be defined and added in a more efficient rule-based manner (e.g., relative position of key joints) rather than a data-driven manner which would otherwise require cumbersome new data collection and model retraining.

Developing such a hand-sensing foundation model capable of constructing vision-like shape and skeleton information solely using commercial WiFi entails two major technical challenges, even with the assistance of a depth camera during model training. First, the palm is a dominant reflector compared to the fingers. It is challenging to separate different scales of reflectors from the received signal and model the relationship between the dominant reflector (the palm) and other reflectors (the fingers) to further understand the structure of the hand. Second, the hand's varying positions and changes in the ambient environment can lead to distinct CSI multi-path profiles. As such, the developed model must be capable of accurately extracting the signals of interest from the hand while remaining robust to any other changes.

To address the aforementioned challenges, we develop HandNet, a novel symbolically-constrained multi-task learning framework. HandNet first adds one more modality (hand mask) to learn together with the hand pose by sharing the same encoder. This additional modality provides dense supervision, constraining the learning process and balancing the information between the dominant reflector (palm) and other reflectors (fingers). Simultaneously, the encoder is designed to preserve the phase information (distance information) of the CSI plus multi-scale feature extractors to exploit reflectors of different sizes that are of interest. Further, a set of symbolically constrained loss functions based on the defined parameterized hand model are adopted to ensure the reconstruction of anatomically plausible skeletons. In addition, a domain generalization method is introduced during HandNet training and we collect a comprehensive training data set to train the model. The model is encouraged to learn only the features representing the hands of interest that are common across different hand positions and environments. These techniques put together allow us to construct 3D hand skeleton with practical WiFi devices.

HandFi is evaluated comprehensively in various usage scenarios and we summarize the results and contributions as follows :

- We develop HandFi which is capable of constructing 2D hand mask and 3D hand pose with commercial WiFi devices. HandFi achieves a 91% overlap with ground truth for 2D hand masks and a 2.07 cm joint error for 3D hand skeletons, sufficiently accurate for various HCI applications.
- HandFi can serve as a foundation model for the development of a variety of downstream hand-sensing applications. For example, finger-level sensing applications developed based on HandFi outperform existing WiFi-based sensing systems.
- HandFi has been prototyped and evaluated under various conditions, including different sensing ranges, hand positions, different users, and occlusion scenarios. HandFi can work under occlusion and offers an extended field-of-view and sensing range.

2 OVERVIEW

HandFi takes as input the channel state information (CSI) that measures the environment by multiple antennas and then goes through HandNet to construct 2D hand mask and 3D hand pose as illustrated in Figure 2. The depth camera is only used in the training stage to provide ground truth labels. The core of HandFi is the design of a multi-task learning network, HandNet (Figure 4), including four main components. The first component is the RF signal embedding layer to deal with complex-valued CSI from different antenna streams, aiming to preserve both amplitude and phase information of CSI. The second component consists of a shared deep multi-scale encoder, which is designed to squeeze and exploit the signals of interest from the palm and finger reflectors contained within the high-dimensional noisy CSI and then transform relevant deep hand semantic features into a latent space. Two task-specific decoders then reconstruct the 2D hand mask and 3D hand pose from the latent deep hand semantic features. This decoding process provides complementary information if a WiFi frame is unable to encode all parts of the hand. To make the system deployable in unseen hand positions and environments, HandFi further adopts a domain generalization technology. In the end, a wide range of applications can be built on top of the inferred hand mask and hand pose. For example, one can encode different gestures as a one-hot vector and add one fully connected layer as a classification head to classify the gestures in a flexible way. New gestures then can be added by updating the one-hot vector. Likewise, one can focus on one of the fingers and enable finger tracking.

3 METHODOLOGY

This section outlines the methodology of HandFi and its core multi-task learning-based framework, HandNet. Although the range resolution of WiFi is low (with a range resolution of only 15 meters for a 20MHz bandwidth), the measurement granularity of CSI is high

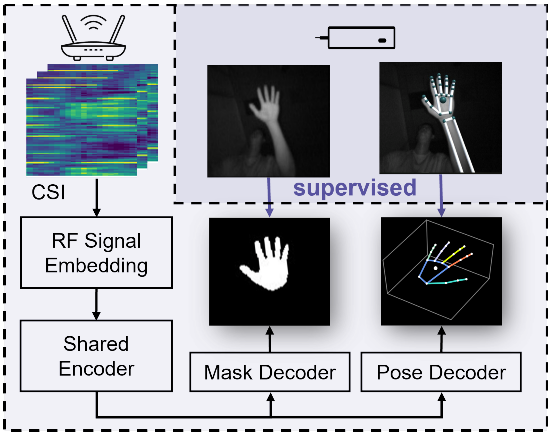


Figure 2: System overview.

and sufficient to distinguish subtle multipath profile variation over a short distance.

For exmaple, the CSI data obtained at 5GHz with a wavelength $\lambda = 5.7cm$, and $\pi/3$ phase measurement granularity is capability of distinguishing 5mm distance resolution [41]. However, the imperfections in WiFi hardware can induce mixture phase rotations to the CSI, making it difficult to separate the phase rotation caused by the target reflector. Therefore, we resort to the deep learning technology, which is good at extracting features from a high-dimensional data to model the relationship between hand pose and CSI data. To establish a one-to-one relationship with ground truth label and incorporate symbolic prior knowledge, we first introduce the proposed hand model. Next, we explicitly embed the complex-valued CSI through point-wise group convolution, which prevents information loss and preserves the physical information of CSI. The embedded CSI is then input into a multi-scale perception encoder that extracts features from different domains with different scales and transforms them into a deep hand semantic space. This allows us to focus on signals of interest and filter out irrelevant noise. Two task-specific decoders are connected to the same encoder, which works in parallel to reconstruct the 2D hand mask by 0-1 classification and 3D hand pose by regression, respectively. A set of customized loss functions is utilized for both tasks. The back-propagation of the 2D hand mask reconstruction task regularizes the learned latent space of the shared encoder, which then benefits the 3D hand pose task. Further, the domain generalization technique is applied during training to learn position-agnostic latent features for practical concerns.

3.1 Hand Modeling

In order to establish one-to-one correspondences with ground truth labels for supervised learning, HandFi first symbolically models the hand. The depth camera returns 24 key joints of the hand, with 20 joints representing the fingers, one joint indicating the center of the palm and three joints denoting the wrist (as shown in the top right corner of Figure 2). The depth camera provides sufficiently accurate results that we take as ground truth. We keep the 20 finger joints along with the palm joint, while discarding the remaining joints to build our hand model as shown in Figure 3. Our model differs from the well-recognized 21-joint hand model in the computer vision (CV) domain [23], where the wrist point is used as the root point. Instead, we choose the palm point as the root point. This is not only because the palm is a dominant reflector, but also because the depth camera uses it as the base to estimate fingers and the wrist, providing a more accurate ground truth for locating the center of the palm. The 3D hand pose is defined by a set of coordinates $\mathbf{j_i} = (x_i, y_i, z_i)$, which describe the locations of J keypoints in 3D space, i.e., $i \in [1, ..., J]$ where $J = 21$ in our case. The joint matrix $\boldsymbol{J} \in \mathbb{R}^{21\times3}$. The depth camera does not directly provide the hand mask. However, HandFi extracts the hand mask from the raw image frame captured by the depth camera. The hand mask is represented as a 0-1 matrix $\boldsymbol{M} \in \mathbb{R}^{114\times114}$, where the areas with and without the hand are annotated respectively. We model the 2D hand mask and 3D hand pose to direct HandNet to learn the gesture-independent hand shape and hand skeleton.

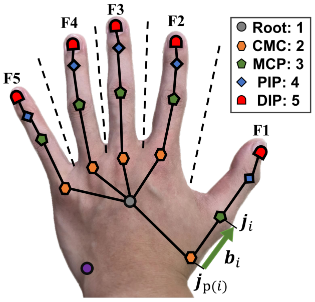


Figure 3: The hand model.

Since the 21 joints follow the hand's anatomy and the hand is a complex articulated object, one can further formulate them to facilitate the incorporation of prior knowledge in subsequent learning procedures. Specifically, the order of the joints is placed elaborately. The hand joints are grouped by fingers. Each finger consists of sequential joints that provide constrained motion, referred to as a kinematic chain - Carpometacarpal Joint (CMC) , Metacarpophalangeal Joint (MCP), Proximal Interphalangeal Joint (PIP) and Distal Interphalangeal Joint(DIP), denoted as the respective set $\boldsymbol{F}_q$, $q \in [1, ..., 5]$ and $\boldsymbol{F} \in \mathbb{R}^{4\times3}$. Except for the root joint $\mathbf{j_1}$, each joint has a parent $p(i)$ and we define a bone as a vector pointing from the parent joint to its child joint, $\mathbf{b_i} = \mathbf{j}_{i+1} - \mathbf{j}_{p(i+1)}$, so $[\mathbf{b_1}, ..., \mathbf{b_{20}}] = \boldsymbol{B} \in \mathbb{R}^{20\times3}$. The bones are named according to the child joint. For example, the bone connecting CMC to MCP is called MCP bone. Intuitively, the CMC bones share one root joint $\mathbf{j_1}$ are those that lie within and define the palm. We define the CMC bones $\mathbf{b_1}, ..., \mathbf{b_5}$ to correspond to the fingers $\boldsymbol{F}_1, ..., \boldsymbol{F}_5$. Figure 3 visualizes the hand model (The purple joint is excluded from the model and learning process, and is solely utilized for plotting the 3D hand pose for illustration purposes) .

3.2 RF Signal Embedding

Specular reflections in RF signals make it difficult to determine if each frame contains all the necessary components of the hand. We aim to retain as much information as possible during signal processing, trusting in the power of deep neural networks to filter out irrelevant features. To achieve this goal, we design an RF signal embedding layer that preserves the physical information carried by CSI, serving as an adapter between complex-valued signals and real-valued deep learning building blocks. Notably, it is important to preserve the phase information, particularly as it represents the corresponding distance of hand reflectors. Unlike images where all pixels have the same magnitude (0-255), CSI values are much more dynamic. For example, the pathloss grows exponentially with distance, and the CSI, $h = a + bi$, is complex-valued, containing both magnitude and phase information of the channel coefficient. Traditional normalization methods, such as scaling the values to a certain range [0,1], are not appropriate for CSI. Therefore, we first normalize CSI by the average power of each packet:

$$\hat{\boldsymbol{h}} = \boldsymbol{h}/(\sum_{i=1}^{F} \|h_i\| /F) \tag{1}$$

where $\boldsymbol{h}$ is a vector of CSI of one packet, the length of the vector depends on the number of subcarrier F. Then, we separate the real and imaginary parts of the normalized CSI to form a real CSI matrix and an imaginary CSI matrix accordingly. We then stack them together to form a tensor denoted as $\boldsymbol{H} \in \mathbb{R}^{F \times T \times 2}$, where F is the number of subcarriers and T is the number of packets. After pre-processing, CSI data meets the requirements of the operators of deep learning.

Although we separate the real and imaginary parts to adapt to existing deep learning frameworks, we do not want to lose the physical information contained in the complex-valued CSI. For example, the absolute value of a and b is the amplitude of the CSI and the ratio between a and b is the phase of the CSI. Inspired by the CLNet [12], we explicitly embed the CSI by making the first layer of HandNet as 1×1 point-wise convolution without bias term such that:

$$E_i = w_i a + w_i b \tag{2}$$

where w_i is the weight of the convolution filter and i is the number of filters. Such embedding preserve the phase information by maintaining the ratio between a and b, and the amplitude information is simply scaled by the w_i parameter. Unlike CLNet which only considers one CSI matrix, a wireless sensing system typically contains multiple antennas that render multiple CSI matrices. Hence, we stack different CSI matrices across the channel dimension, $\boldsymbol{H} \in \mathbb{R}^{F \times T \times (2 \times Ant)}$, where Ant denotes the number of spatial streams, and apply group convolution that embeds each pair of CSI matrices from the same antenna with learnable weights to preserve the physical information of CSI as the CLNet does. Figure 4 depicts the RF signal embedding operation.

3.3 Multi-task Learning

The RF signals are now organized into a time-frequency-spatial tensor with the physical information preserved to feed in the shared encoder with two task-specific decoders as shown in Figure 4. Multi-task learning improves the performance of tasks by utilizing the limited training samples to learn generic features that benefit from the effect of regularization brought by parameter sharing, which has been successfully adopted in many domains [51], including hand pose estimation in CV domain [55]. In HandFi, the ground truth 2D hand mask and 3D hand pose are acquired from the same egocentric camera. Although they may have different distributions and unequal information, they are inherently paired and possess intrinsic correspondence, enabling them to be learned jointly. By harnessing the dense supervision provided by the 2D hand mask, the information regarding the overall hand structure can be propagated to the task of reconstructing the 3D hand pose. This process aids in disambiguating the hand pose amidst the complexity of the WiFi signals.

3.3.1 Shared Multi-scale Perception Encoder. The information pertaining to the hand is intricately encoded within high-dimensional and noisy WiFi signals, where the palm assumes the role of the dominant reflector, while the fingers exhibit comparatively smaller reflections. To capture deep semantic information of the hand across diverse scales, HandNet has devised an encoder equipped with varying receptive fields, allowing HandNet to extract and integrate information from different feature dimensions. The job of encoder is to exploit and extract deep semantic information of the hand regardless of noise and other irrelevant CSI patterns. To address this problem, we equip the network with modified InceptionV3 [35] blocks to enhance feature extraction across multiple scales. Specifically, HandNet creates four learning pathways (from top to bottom of block 1 and block 2), as shown in Figure 4. After RF signal embedding layer, $\boldsymbol{H} \in \mathbb{R}^{F \times T \times Ant}$, where Ant represents the number of antenna streams ($Ant = 3$ in our case). The tensor's height, width, and depth respectively represent information in the frequency subchannels, time domain, and spatial domain. The four pathways aim to extract features of different dimensions and scales. In particular, the combination of average pooling (AP) and 1x1 convolution filters are specifically designed to capture spatial features by condensing all time-frequency features into a single numerical value through AP and then filtering responses throughout the depth of the tensor. The second pathway is intended to extract fused global time-frequency information using a combination of filters of different scales in a large receptive field. The 1x7 filter focuses solely on the time domain, the 7x1 filter focuses solely on the frequency domain, and the 1x1 filter is used to increase the tensor dimension for improved learning. The third pathway is designed to focus on local time-frequency features that are complementary to those extracted by the second pathway. Similarly, the fourth pathway is designed to focus on local spatial features that complement those extracted by the first pathway. All four pathways are concatenated to fuse features with different scales of the time-frequency-spatial domains. At the end, the multi-scale encoder will obtain a deep representation of $\boldsymbol{H}$, CSI, denoted as $\mathbf{r}$.

3.3.2 2D Hand Mask Generation. A task-specific decoder is devised to generate the 2D mask $\hat{\mathbf{M}}$ from the deep hand representation $\mathbf{r}$ using ground truth 2D hand mask supervision, as illustrated in the top right of Figure 4. The decoder is composed of 14 residual blocks [10]. These residual blocks are designed to be more sensitive to gradient changes and to prevent feature loss in the mask decoder. A transport convolution layer with five 5x5 convolution blocks is then used to rescale the feature map to a fixed hand mask size for the supervision training. Instead of directly regressing a hand mask from CSI, we consider the hand mask generation as a pixel-level classification problem and thus adopt binary cross entropy (BCE)

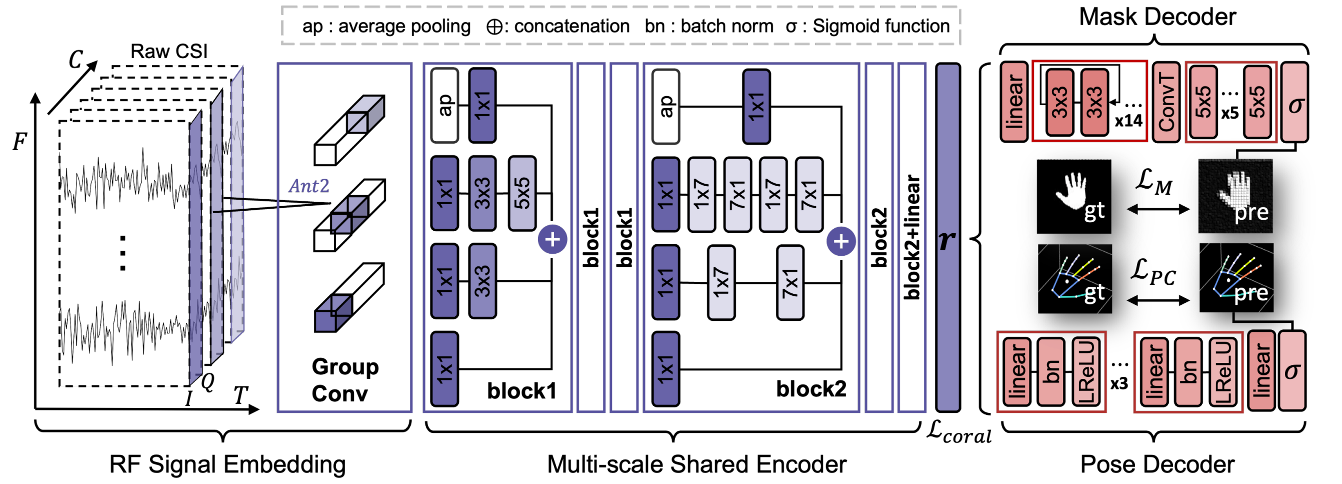


Figure 4: Architecture of HandNet.

loss rather than mean square error (MSE) loss such that:

$$\mathcal{L}_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}[\mathbf{m}\log(\widehat{\mathbf{m}}) + (1-\mathbf{m})\log(1-\widehat{\mathbf{m}})] \tag{3}$$

where n is the number of training samples of the hand mask. Compared to small value gradient of MSE loss, the BCE loss return gradient is proportional to the difference between the prediction and the truth and improve the network ability to distinguish the hand class. Inspired by Focal Loss [18], we add a factor $(1-\mathbf{m}_t)^\gamma$ to $\mathcal{L}_{BCE}$ for the purpose of focusing on hard, misclassified elements, which in our case is the boundary part of the hand and the fingers areas. For notational convenience, we define $\mathbf{m}_t$:

$$\mathbf{m}_\mathrm{t} = \begin{cases} \widehat{\mathbf{m}} & \text{if m=1} \\ 1-\widehat{\mathbf{m}} & \text{otherwise} \end{cases} \tag{4}$$

So, our loss for 2D hand mask generation is defined as:

$$\mathcal{L}_M = -(1-\mathbf{m}_t)^\gamma \log(\mathbf{m}_t) \tag{5}$$

where $\gamma \geq 0$ is a tunable focusing parameter. Intuitively, this scaling factor can automatically down-weight the contribution of easy elements (the background and the palm areas) during training and rapidly focus the model on hard elements. In particular, if the element is misclassified and $\mathbf{m}_t$ is small, the factor is near 1 and the loss is unaffected. As $\mathbf{m}_t \longrightarrow 1$, the factor goes to 0 and the loss for well-classified elements is down-weighted. The focusing parameter γ is increased the effect of the factor is likewise increased. When $\gamma = 0$, $\mathcal{L}_M$ is equivalent to $\mathcal{L}_{BCE}$. γ is set to 2 in our case.

3.3.3 *3D Hand Pose Estimation.*

Another pose regression decoder is used to infer the 3D hand pose, $\hat{\boldsymbol{J}} \in \mathbb{R}^{21\times3}$, from the deep representation $\mathbf{r}$, with the supervision of ground truth as illustrated in the bottom right of Figure 4 such that:

$$\mathcal{L}_J = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{J} - \hat{\boldsymbol{J}}) \tag{6}$$

where n is the number of training samples of the hand pose. When solely doing pose estimate task using the $\mathcal{L}_J$, we note that the network easily gets stuck at a local optimum, resulting in 21 joints clustering around a specific location and therefore the network hard to regress a meaningful hand pose. When learning concurrently with the hand mask generation task, this phenomenon disappears, implying that the hand's structure aids in regularizing hand pose reconstruction. Inspired by this observation, additional prior knowledge about the hand can be imported into the pose decoder to further constrain the estimated hand pose. First, HandNet adopts bone length loss [19, 34] that is utilized in human body estimation task to constrain the fingers by penalizing invalid bone lengths. As mentioned in Section 3.1, hand is an articulated object comprising the palm and five independent fingers, and each of the finger consists of three types of bones arranged in a hierarchical manner. The bone length regularization loss is defined such that:

$$\mathcal{L}_{BL} = \frac{1}{15}\sum_{i=1}^{15}\mathcal{R}\left(\|\mathbf{b}_i\|_2\,; b_i^{\min}, b_i^{\max}\right) \tag{7}$$

Here, $\mathcal{R}$ denotes a range-constrained function that restricts the variable x to fall within a specific range $[a, b]$ such that:

$$\mathcal{R}(x; a, b) = \max(a - x, 0) + \max(x - b, 0) \tag{8}$$

Each of the bones i has a range $\left[b_i^{\min}, b_i^{\max}\right]$ of valid length that can be found in [6].

In addition, the five CMC bones exhibit a lower degree of freedom compared to other hand bones. The five CMC bones, along with the root joint and five CMC joints, collectively span a palmar structure [31]. The four angular distance ϕ of five CMC bones is within a certain degree from each other in the plane they span [29]. The angle is calculated by:

$$\phi_i = \arccos\left(\frac{\mathbf{b}_i^T\mathbf{b}_{i+1}}{\|\mathbf{b}_i\|_2\,\|\mathbf{b}_{i+1}\|_2}\right) \tag{9}$$

At the same time, the curvature c between two CMC joints is limited in a range as well. The curvature between two bones can be calculated approximately by following [28]:

$$c_i = \frac{(\mathbf{e}_{i+1} - \mathbf{e}_i)^T (\mathbf{b}_{i+1} - \mathbf{b}_i)}{\|\mathbf{b}_{i+1} - \mathbf{b}_i\|^2}, \text{ for } i \in \{1, 2, 3, 4\} \tag{10}$$

where $\mathbf{e}_i$ is the edge norm at bone $\mathbf{b}_i$ such that:

$$\mathbf{n}_i = \text{norm}(\mathbf{b}_{i+1} \times \mathbf{b}_i), \text{ for } i \in \{1, 2, 3, 4\}$$
$$\mathbf{e}_i = \begin{cases} \mathbf{n}_1, & \text{if } i = 1 \\ \text{norm}(\mathbf{n}_i + \mathbf{n}_{i-1}), & \text{if } i \in \{2, 3, 4\} \\ \mathbf{n}_4, & \text{if } i = 5 \end{cases} \tag{11}$$

where $\text{norm}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. To ensure biomechanical valid palm structures, both angular distance ϕ and curvature c are taken into consideration and define the palmar structure constraints loss as follows:

$$\mathcal{L}_P = \frac{1}{4} \sum_{i=1}^{4} \left(\mathcal{R}\left(c_i; c_i^{\min}, c_i^{\max}\right) + \mathcal{R}\left(\phi_i; \phi_i^{\min}, \phi_i^{\max}\right) \right) \tag{12}$$

In summary, the pose constraints loss is constructed as follows:

$$\mathcal{L}_{PC} = \beta \mathcal{L}_J + \gamma \mathcal{L}_{BL} + \lambda \mathcal{L}_P \tag{13}$$

where β,γ and λ are weights to balance the individual loss terms.

3.4 Domain Generalization

Different hand placement positions can result in distinct multipath profiles with different CSI data distributions, which leads the learned model to perform poorly. In fact, this issue is known as *domain shift*, which refers to the change in the data distribution. In this paper, we refer to a pair of relative positions of hand and routers in a specific environment as *domain*. The domain where the model is trained as source domain and the domain where the model is applied as target domain. Literature has proposed numerous methods to cope with the domain shift issue and tried to improve the generalization of models. Traditional Empirical Risk Minimization (ERM) [39] approaches aim to collect sufficient data from diverse domains to provide the model with strong generalization capabilities. Alternatively, domain adaptation [4] approaches adjust network parameters to adapt the learned model to different target domains. However, collecting data from all domains is a costly and impractical endeavor. Domain adaptation assumes that the model has access to some information about the new domain, enabling it to perform meta-learning or few-shot learning, which is not the case for HandFi scenario. When deploying HandFi in real-world scenarios, one of the challenges is that we may not know the exact relative location of the present hand or the WiFi routers. As such, the model must distinguish the causal signal feature of interest rather than spurious signal feature specific to a target domain. An opportunity for HandFi is that regardless the different domain, our label's distribution are always the same. Therefore, HandFi resorts to the advanced domain generalization (DG) technique.

In particular, suppose we have K similar but distinct source domains $\mathcal{S} = \left\{S_k = \left\{\left(x^{(k)}, y^{(k)}\right)\right\}\right\}_{k=1}^{K}$, where x is the CSI data, y is the label, each associated with a joint distribution $P_{XY}^{(k)}$. Note that $P_{XY}^{(k)} \neq P_{XY}^{(k')}$ with $K \neq K'$ and $k, k' \in \{1, ..., K\}$. Note that the marginal distribution of label space remains the same, $P_Y^{(\mathcal{T})} = P_Y^{(\mathcal{S})}$. DG tries to learn a model $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ from several different but related source domains data such that the error of the model on *unseen* target domain $\mathcal{T} = \left\{x^{\mathcal{T}}\right\}$ is minimized [5, 59]. The assumption behind this is that even if we do not have information from all domains, it is possible to have a subset of domains that are sufficient for identifying true causal features and distinguishing them from statistically associated features that may vary across different domains [3].

To learn the position-agnostic features, we adopt deepCORAL [32] into HandNet. Concretely, we collect the same set of data but in different domain and obtain their deep representation $S_1 = \left\{\boldsymbol{r}_i^1\right\}$ and $S_2 = \left\{\boldsymbol{r}_i^2\right\}$, where we choose two domains for clarification. In practice, there can be many domains. A CORAL loss is applied to minimize the distance between the second-order statistics (covariances) of the different domain latent features:

$$\mathcal{L}_{CORAL} = \frac{1}{4d^2} \left\|\mathrm{C}_{S_1} - \mathrm{C}_{S_2}\right\|_F^2 \tag{14}$$

where d is the dimension of $\mathbf{r}$ and $\|\cdot\|_F^2$ denotes the squared matrix Frobenius norm. The covariance matrices are given by:

$$\begin{aligned} \mathrm{C}_{S_1} &= \frac{1}{bs-1} \left(S_1^\top S_1 - \frac{1}{bs} \left(\mathbf{1}^\top S_1\right)^\top \left(\mathbf{1}^\top S_1\right) \right) \\ \mathrm{C}_{S_2} &= \frac{1}{bs-1} \left(S_2^\top S_2 - \frac{1}{bs} \left(\mathbf{1}^\top S_2\right)^\top \left(\mathbf{1}^\top S_2\right) \right) \end{aligned} \tag{15}$$

where bs is the batch size and $\mathbf{1}$ is a column vector with all elements equal to 1. Note that $\mathcal{L}_{CORAL}$ is differentiable so it can incorporate to end-to-end deep neural network and the gradient can be back-propagated. Unlike the original CORAL method, which aims to align the feature representations of different networks before the classification head, HandNet uses DeepCORAL by alternately feeding data from different domains on a batch basis to compel HandNet to capture domain-invariant causal feature during the extraction of deep hand semantic information. The whole HandNet is trained end-to-end with the summation of all the losses such that Eq 16 will be minimized:

$$\mathcal{L} = \alpha \mathcal{L}_M + \beta \mathcal{L}_J + \gamma \mathcal{L}_{BL} + \lambda \mathcal{L}_P + \zeta \mathcal{L}_{CORAL} \tag{16}$$

where $\alpha, \beta, \gamma, \lambda$ and ζ are weights to balance the individual loss terms.

4 EVALUATION

4.1 Implementation

Commercial Wi-Fi: We collect the CSI data using two commercial routers (TP-LINK WDR4310) equipped with Atheros SoC AR9344. The transmitting router has a single antenna, while the receiving router has three antennas. The raw CSI data is collected using the Atheros-CSI-Tool [45] and sent to a server through a TCP link using the general socket API. The WiFi signals are transmitted on the 5GHz channel with a 40MHz bandwidth, and the packet rate is set at 200 packets per second.

Server: The server employed for both CSI data collection and model inference is a Linux desktop computer equipped with an Intel Core i9-9820X CPU and one Nvidia 2080Ti GPU card.

Leap Motion: The ground truth 2D mask and 3D joint data are obtained using a commercial egocentric depth infrared stereo camera, Leap Motion. The depth camera provides sub-millimeter accuracy of the hand joints within a field of view of approximately 135°, with an effective sensing range above the device of approximately 25 to 600 millimeters [16].

Ground truths: The Leap Motion sensor is positioned at 30cm directly below the hand. Using the official python API, we collect the 3D coordinates of 21 hand joints as defined by our hand model as our ground truth. We then normalize the coordinates from 0 to 1. A 2D hand mask is generated from the leap image using the GrabCut [27] algorithm by applying OpenCV for foreground segmentation to extract the hand from the background. The resulting mask uses a binary representation, with 1 representing the hand area and 0 representing the background. To avoid the class imbalance problem, we crop the hand mask with the reference of the center of the palm. We use a local network time protocal (NTP) server to synchronize the routers and the depth camera.

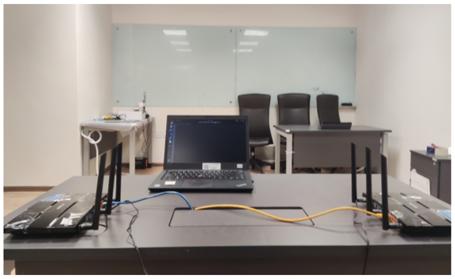


Figure 5: Testbed.

Data Collection: We invited 6 subjects and collected data in 5 different environments with 5 different hand positions relative to the WiFi transceivers. [1] In order to validate the effectiveness of our system on domain shifts (e.g., unseen subjects, gestures, environments, and hand positions), we construct a training dataset (main dataset) from two subjects for model training. The training dataset includes 27 hand postures consisting of 26 American Sign Language (ASL) alphabets and one fully open hand posture. As suggested by RFPose [57], the specularity of the human body causes unequal reflection, and therefore we collect data as a sequence to enhance the learning process. Each posture is performed within a 2-second window with a slight up-and-down motion, except for the dynamic alphabets J and Z. This approach is also consistent with the principles for super-resolution, which involves utilizing complementary information from adjacent frames to perform super-resolution reconstruction [38]. Deep neural networks are a powerful tool for achieving super-resolution by learning patterns and features from a large amount of data [50]. The training dataset is collected in a testbed as shown in Figure 5, where the leap motion is placed at the center of the two WiFi routers. Volunteers are invited to perform 27 hand gestures between the routers and above the leap motion. Two volunteers (#P1: 21-year old male and #P2: 27-year old female) perform each gesture for about 4.5 minutes respectively. As the router outputs 200 CSI per second and we consider every 20 CSI measurements as one sample, we generate 10 samples per second. Thus, each sample can be represented as a vector with the size of 3x114x20, representing 3 spatial streams, 114 WiFi subcarriers, and 20 CSI samples, respectively. The training dataset is randomly split into 9:1 for training and validation of the HandNet model.

Training details: The HandNet is trained using the ADAM optimizer [15] with Cosine Annealing Learning Rate scheduler [20]. The initial learning rate is 0.001. The batch size is 24 and HandNet is implemented by PyTorch.

4.2 Evaluation Metric

Figure 1 shows the qualitative results of HandFi, which demonstrates that the WiFi-based HandFi is capable of achieving competitive perceptual accuracy to that of the CV-based system. To further analyze it, we adopt several metrics that are widely measured in the CV domain. The mean Pixel Accuracy (mPA) and Intersection-over-Union (IoU) are used for 2D mask evaluation. The Mean Per Joint Position Error (MPJPE) and the Percentage of Correct Keypoints (PCK) are used for 3D joint evaluation. The metrics are detailed below:

mPA: Pixel accuracy is a classic semantic segmentation metric, defined as the ratio between the amount of accurately classified pixels and the total number of pixels in the image. Accordingly, mPA represents the average percentage of accurately classified pixels in the entire image.:

$$mPA = \frac{1}{k}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{17}$$

where p_{ii} is the total number of true positive pixels for class i, p_{ij} stands for the pixel that belongs to class i but mistakenly as class j and k is the total number of class, where $k = 2$ in our case. A drawback of mPA is that it favors the dominant class. For example, if the hand is very far away from the camera and just occupies a small portion of the mask, the background class is dominant. In an extreme case, even if a predicted mask is without a hand at all, the accuracy will still be almost 100%. Thus, we introduce IoU as well.

IoU: IoU is the area of overlap between the predicted hand and the ground truth hand in the mask, divided by the area of union between the predicted hand and the ground truth hand:

$$IoU = \frac{p_{ii}}{p_{ii} + p_{ji} + p_{ij}} \tag{18}$$

MPJPE: MPJPE calculates the average Euclidean distance between the predicted joint coordinates and the ground truth joint coordinates:

$$MPJPE = \frac{\sum_{i=1}^{J}\left(\left\|\hat{\mathbf{j}}_i - \mathbf{j}_i\right\|_2\right)}{J} \tag{19}$$

PCK: PCK describes the percentage of correct predicted joints, where the predicted joint is considered correct if the distance between it and the ground truth joint is within a certain threshold:

$$PCK@a = \frac{\sum_{i=1}^{J}\delta\left(\left\|\hat{\mathbf{j}}_i - \mathbf{j}_i\right\|_2 \leq a\right)}{J} \tag{20}$$

where δ is a logical operation that outputs 1 if Ture and output 0 if False. We set $a = 2cm$.

[1] All experiments that involve humans have been approved by IRB.

Table 1: Ablation study of HandFi.

	mPA	IoU	MPJPE(cm)	PCK@2cm
BL.A-single task	/	/	20.41	0.19
BL.B-ResNet50	/	/	25.27	0.10
BL.C-UNet	/	/	22.65	0.14
BL.D-embedding	/	/	13.98	0.29
BL.E-multi-task	0.90	0.78	*6.43*	*0.71*
BL.F-BCE loss	0.92	0.80	6.41	0.72
BL.G-mask loss	0.94	0.91	5.72	0.76
BL.H-HandNet	**0.94**	**0.91**	***2.07***	***0.93***

4.3 Ablation Study

Since there is currently no model available that can extract hand masks and poses from RF signals, we conduct an ablation study by designing several baseline (BL) models to understand the role of different components in the proposed HandNet architecture. Specifically, these baseline models are designed by removing or replacing specific modules of HandNet as follows:

Baseline A - single task: This baseline only contains the multi-scale encoder of HandNet and the pose decoder with joint regression loss $\mathcal{L}_J$.

Baseline B - ResNet50 encoder: Baseline B changes the proposed multi-scale encoder of baseline A to ResNet50 [10] as the encoder.

Baseline C - UNet encoder: This baseline uses UNet [26] with joint regression loss $\mathcal{L}_J$.

Baseline D - signal embedding: RF signal embedding layer + Baseline A.

Baseline E - multi-task: Baseline E adds one mask decoder with MSE loss to the baseline D to make it a multi-task encoder-decoder structure.

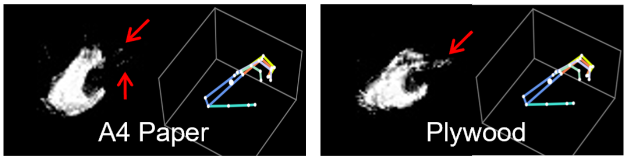
Baseline F - BCE loss: Use BCE loss to replace baseline E's MSE loss.

Baseline G - mask loss: Use $\mathcal{L}_M$ to replace baseline F's BCE loss.

Baseline H - HandNet: $\mathcal{L}_{PC}$ + baseline G, which is the whole HandNet.

Table 1 presents the average quantitative results for 5 runs of each baseline on the main training dataset. The aim of Baseline A-C is to evaluate different choices of backbone networks. ResNet50 [10] and UNet [26] are popular backbones with good pose estimation performance in CV domain. While UNet is better than ResNet in terms of lower average joint error (MPJPE) and higher correct joint rate (PCK), our proposed multi-scale backbone outperforms them both. The reason for this is that our multi-scale backbone focuses on extracting features from different scales, instead of solely focusing on identifying useful features.

Baselines D and E evaluate the necessity of the RF signal embedding layer and the multi-task structure. The results show that organizing the complex-valued RF signal properly boosts the accuracy by about 42.07%, and the added multi-task structure takes the hand pose accuracy to another level. Baselines F-H evaluates different loss functions, with the BCE loss generally providing better results than the MSE loss. The $\mathcal{L}_M$ significantly increases the IoU result by improving the finger boundary while also benefiting the pose estimation task. The $\mathcal{L}_{PC}$ loss provides an additional lift for the hand pose accuracy. The added mask generation task and the designed constraints for hand joints are key to the success of HandFi.



Scenarios	A4 Paper	Plywood
Leap Motion	X	X
HandFi	IoU:0.78 PCK:0.90	IoU:0.67 PCK:0.86

Figure 6: Examples of masks and poses under occluded scenarios and HandFi accuracy under occlusion.

4.4 Evaluation of Robustness

We test HandFi in various conditions to evaluate the robustness and the characteristics of the system.

4.4.1 *Under Occlusion.*

Another advantage of RF sensing is that it can sense under occlusion. We use an A4 paper and a plywood to occlude the hand and play the ASL letter c 10 times independently and let the HandFi infers the hand information. We note that because of the occlusion, the Leap Motion cannot work in this experiment. The quantitative result[2] is reported in the table of Figure 6 along with examples of the hand mask and hand pose. As expected, HandFi works well under occlusion, but interestingly, we saw some extra components on the masks due to the reflection of the occlusion.

4.4.2 *Effective Sensing Range.*

WiFi offers omnidirectional (360°) sensing as demonstrated in § 4.5, while Leap Motion has the field of view of a cone roughly 135°. The sensing distance of Leap Motion is limited by its FoV and can be roughly calculated by $P * tan(135/2)$, where P indicates the perpendicular distance of the hand to Leap Motion. For example, if the hand is placed 30cm above the Leap Motion, the sensing range for Leap Motion is +/- 23cm. In contrast, WiFi has a room-scale sensing range. In the following, we conduct the experiment to understand the effective sensing range of HandFi.

In particular, starting from the midpoint between two WiFi APs, we place an open hand at every 0.2-meter interval. The hand holds roughly at the same plane level, and at each range, we collect 1s with 10 samples for one-time hand placement. We compute the IoU of the mask output by the HandFi system and the PCK@2cm value of the pose at each range. Figure 7 reports the results. The results are presented in Figure 10, where we showcase the inferred 2D hand masks at distances of 0.2m, 0.8m, 1.6m, 2.4m, and 2.6m. It can be observed from the masks that their clarity decreases with increasing distance. Some hand areas are mistakenly identified as background areas, resulting in holes in the masks, as can be seen in mask (b). Moreover, misclassifications of finger areas become increasingly frequent, as evident from masks (c) and (d). Finally, the model becomes more and more confused with other gestures, as demonstrated by mask (e). As for the pose, although the accuracy of

[2] The ground truth is obtained by removing the occlusion immediately.

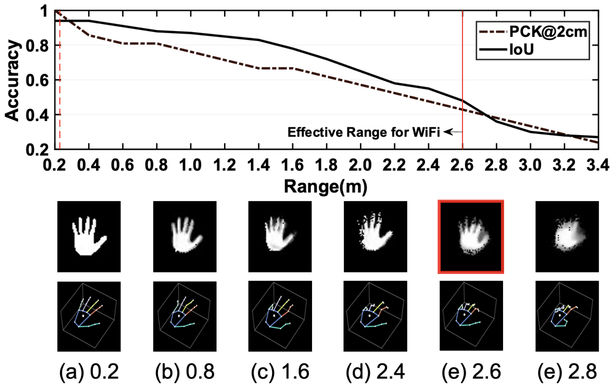
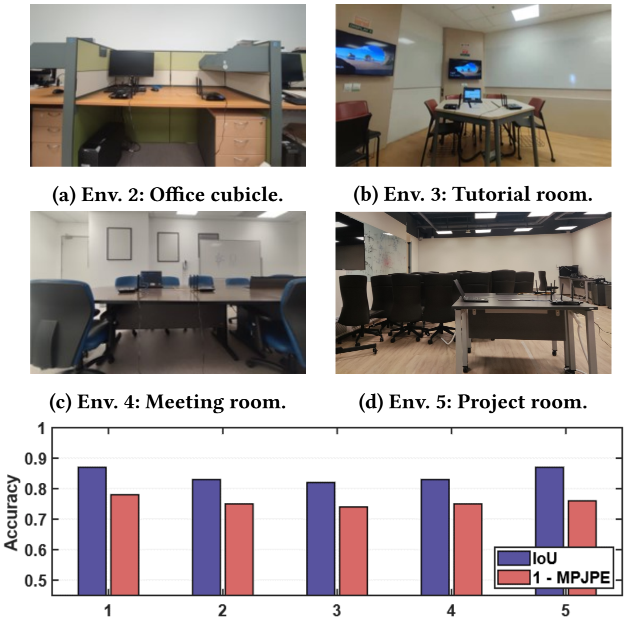


Figure 7: Quantitative results of HandFi at different sensing distances and the inferred hand mask at the corresponding sensing distances.


(a) Env. 2: Office cubicle.

(b) Env. 3: Tutorial room.

(c) Env. 4: Meeting room.

(d) Env. 5: Project room.

Figure 8: Performance of HandFi under different environments.

finger joints is getting worse, the palm structure still aligns with the open palm. Since at 2.6m, we are still able to tell that the mask/pose is an open hand and thus the effective sensing range of WiFi is +/- 2.6m, x11 larger than Leap Motion.

4.4.3 Different Environments. We directly migrate HandFi that was trained in the testbed (Fig 5) to several typical indoor environments as illustrated in Figure 8, and ask the subject #P1 to perform the rock-scissors-paper gesture, where the rock and scissors gestures were not seen in the training. Each gesture is performed by ten times. The average IoU and 1-MPJPE are reported in Figure 8, and the mean variances of the mask and pose results are 0.058% and 0.023%. Therefore it is safe to say HandFi is robust to different environments.

Table 2: Performance of HandFi under different users (gender and age).

	M,21(#P1)	F,27(#P2)	M,22	M,20	M,21	F,20
IoU	0.91	0.88	0.84	0.85	0.85	0.81
MPJPE	1.95	2.07	2.24	2.19	2.21	2.71

4.4.4 Different Users. We conducted a study involving 6 subjects of varying ages and genders, who were instructed to perform rock-scissors-paper gestures. Each gesture was repeated ten times for each subject in our Testbed 1. The results are reported in Table 2. Note that the first and the second users contribute to the training dataset of HandFi. The results suggest that the accuracy of HandFi decreases slightly for unseen users, but still outperforms the baselines. In addition, we observed that the system performs less accurately for female users compared to male users. This could be because female hands are generally smaller and have less reflection information. We leave the investigation of this observation for our future work.

4.5 Different Location of the Hand

To evaluate the effectiveness of the domain generalization technique, we collected datasets at different hand positions and conducted evaluations. Specifically, we collected data using Testbed 1 (Figure 5) in Env. 5 (Figure 8d). The volunteers performed hand gestures at 5 different positions (domains) as illustrated in Figure 9 The data collection process at each hand position is the same as that of the main dataset (Section 4.1). Data from the same position are aggregated into a dataset, and we collected 5 datasets at different positions in total. We report the quantitative results in Figure 11 when HandFi is trained and tested in the same dataset as well as different datasets. As can be seen from Figure 11, when HandFi is tested on the same domain (i.e., the diagonal line of the table), its performance is consistently good regardless of the position of the hand with 0.97 IoU of the hand mask and 2.00cm joint errors. However, when testing HandFi on different domains (i.e., the non-diagonal part of the table), a significant performance drop is observed, with a range of 16.5-56.1% for the mask task and 20.3-103.0% for the pose task, respectively.

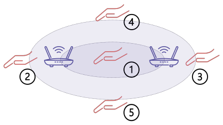


Figure 9: Hand positions .

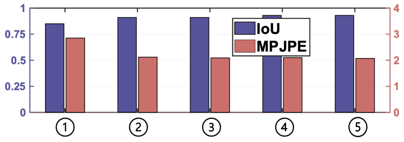
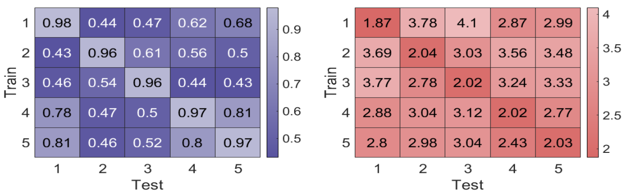


Figure 10: Performance of HandFi in different hand positions with domain generalization employed (trained on other 4 positions).

By adopting the domain generalization training strategy introduced in Section 3.4, we train the model with different dataset from 4 different positions and evaluate HandFi in the unseen hand position and report the quantitative results in Figure 10. Compared to



(a) IoU (b) MPJPE(cm)

Figure 11: Performance of HandFi in different hand positions without domain generalization employed.

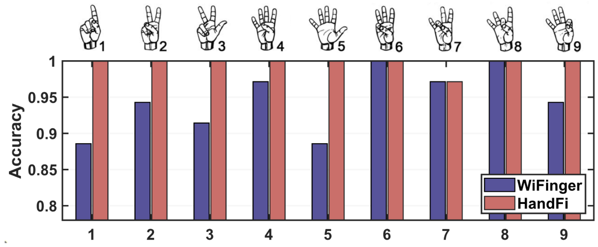


Figure 12: Break down recognition accuracy of american sign language (ASL) digits.

Figure 11a and Figure 11b, although neither IoU nor MPJPE results can achieve the reconstructed mask and pose as well as the diagonal result in Figure 11, the performance drop in the unseen domain is effectively mitigated, and comparable results are achieved by using the domain generalization training strategy.

4.6 Discussion

The current version of HandFi operates with a fixed transceiver setting around one meter apart to ensure the quality of WiFi signals. In our experiments, we observe that when the distance between transceivers increases, our system performance could be affected. We are investigating whether the performance degradation is mainly due to signal attenuation or the limitation of our model generalizability. In practice, we believe it can be acceptable for users to configure transceivers so as to ensure signal strength and high sensing accuracy.

HandFi is currently designed to work with one hand of a user at a time. This restricts HandFi from supporting tasks that require the coordination of both hands or involve more complex multi-user interactions. We are investigating if data augmentation techniques can be applied to generalize the model to both hands without collecting an excessive amount of new data. Multi-user scenario is challenging and we plan to explore if multi-antenna could help separate reflections from multiple users. We leave the above research problems for our future work.

5 APPLICATION CASES

We developed two downstream applications on top of HandFi to show that the obtained 2D hand mask and 3D hand pose can directly boost the hand-related applications.

5.1 Sign Language Recognition

Sign language recognition is the most challenging task because of the inter-similarity between different signs and the fine-grained finger motion. WiFinger [17] is the first solution to use commercial WiFi to recognize 9-digit finger-level signs from American Sign Language (ASL), the most widely used sign language. WiFinger only uses one pair of the transmitter as HandFi does and it proposes a series of signal processing techniques with the k-Nearest Neighbor (KNN) and Dynamic Time Wrapping (DTW) to classify 9 sign digits. We conducted the same experiment as WiFinger, collecting the same amount of data from a single user, with each of the 9 sign digits having 35 samples. The single fully connected layer with softmax function serves as the classification head plug after the HandNet to classify the 9 digits. The results are reported in Figure 12. We see that except for number 7, all of the sign digits are classified with 100% accuracy with the benefit of clear semantic hand information from the generated mask. Compared to the 94.60% accuracy achieved by the WiFinger, HandFi achieves 99.68% accuracy with 5.37% performance increase. It is worth noting that these 9 gestures are unseen gestures for HandFi. HandFi is able to reconstruct the mask with IoU = 0.81 and the pose with MPJPE = 4.37cm and PCK@2cm = 0.78. It can be envisioned that once the low-level features of the hand are extracted, gesture recognition becomes easier and more stable.

5.2 Finger Tracking

Finger tracking, or finger drawing, offers natural and fine-grained interaction with hardware devices. Although it is well-studied in the literature, WiFi-based finger tracking is extremely challenging because of the low resolution of CSI. Since HandFi is able to obtain 3D hand pose, finger tracking becomes a relatively easy task, which focuses on the specific finger (index finger) in a clear way without worrying about self-occlusion. We conduct a finger tracking experiment following the most recent and the first sub-wavelength level finger tracking system, FingerDraw [43], and compare results with them. In particular, we print three templates (a triangle, the letter 'Z', and a closed half-circle-like letter 'D') on cardboard and use the index finger to follow the template trajectories, the same as FingerDraw. We collect 150 finger drawings in total. Each template has 50 drawings. The drawing speed is roughly 5 cm per second.

Table 3: Finger tracking accuracy.

	50% error (cm)			90% error (cm)		
Methods	△	Z	D	△	Z	D
FingerDraw	1.12	1.46	1.29	2.98	3.38	3.52
HandFi	**0.98**	**0.81**	**1.07**	**1.10**	**1.01**	**1.22**
HandFi-3D	2.20	1.05	1.50	2.24	1.15	1.50

Table 3 reports the results, with HandFi achieving an overall median error of 0.95 cm and a 90th percentile error of 1.11 cm. These values are 24.69% and 66.18% higher, respectively, than those achieved by FingerDraw. Additionally, we take the cardboard off and draw the same 150 samples freely in the air. The corresponding results are reported in the last row of Table 3. Figure 14 reveals the

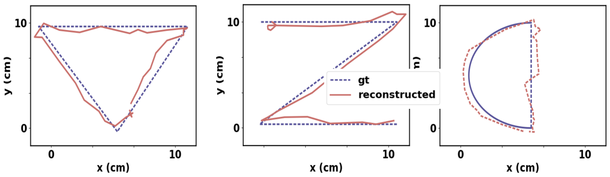


Figure 13: 2D finger tracking with constraint.

trajectory in a free space without the constraint of cardboard, where natural jitters are observed in our finger compared to Figure 13. This indicates HandFi's ability to capture hand tremors, which can enable health monitoring applications such as indicative of Parkinson's disease or anxiety disorders.

In addition to HandFi's higher accuracy and tracking in 3D space, it is worth noting that FingerDraw has accumulative tracking errors, while HandFi does not have such an issue. In addition, FingerDraw requires multiple routers and at least two receivers placed orthogonality, while HandFi only needs one pair of transmitters.

5.3 Computation Overhead

Our system is evaluated in an offline manner to facilitate a more comprehensive evaluation process, but it can support online end-to-end execution. To demonstrate HandFi's capabilities, real-time finger tracking is performed using the trained model. The current version of HandFi does not solve the continuous segmentation problem. Instead, we use a sliding window plus a stop sign to conduct hard segmentation. We average the computation cost from all samples and report the computation overhead in Table 4. The inference task can be executed on a server without a high-end GPU. The latency of CPU execution is approximately five times longer compared to GPU execution.

Table 4: Computation overhead of HandFi.

Transmission Latency (ms)	**0.314**
Collecting Latency (s)	**0.1**
Inference Time (GPU) (ms)	**11.32**
Inference Time (CPU) (ms)	**50.28**

6 RELATED WORK

CV-based Hand Pose Estimation. In the following, we review recent advanced deep learning-based hand sensing methods [8]. Generally, existing CV-based methods are sensitive to changes in illumination and background. The closest works to our 2D hand mask and 3D hand pose tasks are Hand3D [60] and HIU [55], respectively. However, due to the inherent differences between WiFi signals and image signals, existing CV-based methods cannot be directly applied. Specifically, image signals inherently contain clear semantic information, with the task focusing on accurately locating the key points of hands in the image. In contrast, WiFi sensing focuses on effectively extracting features from abstract reflected signals and modeling them according to hand models.

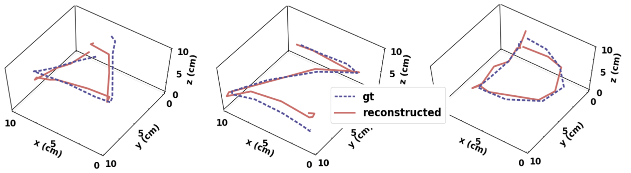


Figure 14: 3D finger tracking in free space.

RF-based Body Pose Estimation. Previous RF-based human pose estimation works either require specialized FMCW devices like RF-Pose [57] and RF-Pose3D [58], or they are not robust to environmental variations [40] and require specific multi-device and multi-antenna placement [14]. These works all rely on body model so cannot be directly applied to hand pose estimation. Hand model has higher degrees of freedom and the reflection scale of hand is much smaller [24, 25]. HandFi is the first to use commercial WiFi to obtain vision-like hand shape and hand pose.

WiFi-based Hand Sensing. Current WiFi-based hand sensing applications are built towards specific applications such as gesture recognition [1, 22, 36, 56], sign language recognition [21, 30, 46, 54], finger tracking [37, 43, 52, 53], and keystroke detection [2]. In contrast, HandFi has the capability to support diverse applications. Furthermore, existing systems mainly rely on mapping hand motion patterns (which requires capturing a time sequence of sensing data and learning the timed pattern) to a limited set of specific gestures, rather than recognizing the basic elements of hand motion as HandFi does. Additionally, HandFi outperforms existing methods in both finger tracking [17] and sign language recognition [17] tasks. HandFi is the first achieving 3D finger tracking.

Generalization in RF Sensing. Over-reliance on deep learning methods during training can result in performance degradation of machine learning systems when tested on data outside the training domain, commonly known as the domain shift. The heterogeneous and imperfect nature of RF data, owing to hardware imperfections and multi-path effects, exacerbates this problem. Prior research in RF sensing has sought to address this issue [7, 9, 11, 13, 44, 48], with many studies resorting to meta-learning or few-shot learning. However, these methods assume access to samples in the target domain, which may not be practical. In contrast, HandFi assumes the target domain is inaccessible and uses DG techniques to train a robust model with slightly lower accuracy in the source domain.

7 CONCLUSION

This paper presents HandFi, a WiFi hand sensing system that can construct hand shape and hand skeleton from commercial WiFi. To this end, we propose a novel symbolically constrained multi-task learning framework, HandNet, and incorporate the domain generalization technique to enhance the sensing performance in unseen environments. We build a prototype system and conduct comprehensive evaluation in various experiment settings. We further develop two finger-level downstream applications to demonstrate the effectiveness of HandFi as a foundation model. Based on HandFi, we believe new downstream applications can be developed to improve accessibility and convenience for individuals with disabilities such as sign language recognition.

ACKNOWLEDGMENTS

We thank the anonymous shepherd and reviewers for their helpful comments. We thank Yupeng Huang for the preliminary validation experiments and all the volunteers involved in this study. This research is supported by HKUST grant R9899, Ministry of Education Singapore MOE AcRF Tier 2 MOE-T2EP20220-0004, and Hong Kong GRF Grant 15206123. Mo Li is the corresponding author.

REFERENCES

[1] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. 2015. Wigest: A ubiquitous wifi-based gesture recognition system. In *2015 IEEE conference on computer communications (INFOCOM)*. IEEE, 1472–1480.

[2] Kamran Ali, Alex X Liu, Wei Wang, and Muhammad Shahzad. 2015. Keystroke recognition using wifi signals. In *Proceedings of the 21st annual international conference on mobile computing and networking*. 90–102.

[3] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant Risk Minimization. *ArXiv* abs/1907.02893 (2019).

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79 (2010), 151–175.

[5] Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems* 24 (2011).

[6] Fai Chen Chen, Silvia Appendino, Alessandro Battezzato, Alain Favetto, Mehdi Mousavi, and Francesco Pescarmona. 2013. Constraint study for a hand exoskeleton: human hand kinematics and dynamics. *Journal of Robotics* 2013 (2013).

[7] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 517–530.

[8] Bardia Doosti. 2019. Hand pose estimation: A survey. *arXiv preprint arXiv:1903.01013* (2019).

[9] Mingda Han, Huanqi Yang, Tao Ni, Di Duan, Mengzhe Ruan, Yongliang Chen, Jia Zhang, and Weitao Xu. 2023. mmSign: mmWave-based Few-Shot Online Handwritten Signature Verification. *ACM Transactions on Sensor Networks* (2023).

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Mingda Jia Zehua Sun Pengfei Hu Yu Zhang Tao Gu Huanqi Yang, Mingda Han and Weitao Xu. 2023. Cross-Modal Translation via Deep Generative Sensing for RF-based Gait Recognition. In *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*.

[12] Sijie Ji and Mo Li. 2021. CLNet: Complex input lightweight neural network designed for massive MIMO CSI feedback. *IEEE Wireless Communications Letters* 10, 10 (2021), 2318–2322.

[13] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th annual international conference on mobile computing and networking*. 289–304.

[14] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using WiFi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] Inc. Leap Motion. [n.d.]. Leap Motion Overview. https://developer-archive.leapmotion.com/documentation/csharp/devguide/Leap_Overview.html.

[17] Hong Li, Wei Yang, Jianxin Wang, Yang Xu, and Liusheng Huang. 2016. WiFinger: Talk to your smart devices with finger-grained gesture. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 250–261.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[19] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time arm skeleton tracking and gesture inference tolerant to missing wearable sensors. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 287–299.

[20] Ilya Loshchilov and Frank Hutter. [n.d.]. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*.

[21] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. Signfi: Sign language recognition using wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–21.

[22] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. 27–38.

[23] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. 2014. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1106–1113.

[24] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. 2014. Realtime and Robust Hand Tracking from Depth. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1106–1113. https://doi.org/10.1109/CVPR.2014.145

[25] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[27] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. In *ACM SIGGRAPH 2004 Papers (SIGGRAPH '04)*. ACM, New York, NY, USA, 309–314.

[28] Szymon Rusinkiewicz. 2004. Estimating curvatures and their derivatives on triangle meshes. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004*. IEEE, 486–493.

[29] Chr Ryf and A Weymann. 1995. The neutral zero method—a principle of measuring joint function. *Injury* 26 (1995), 1–11.

[30] Jiacheng Shang and Jie Wu. 2017. A robust sign language recognition system with multiple Wi-Fi devices. In *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*. 19–24.

[31] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. 2020. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision*. Springer, 211–228.

[32] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 443–450.

[33] Li Sun, Souvik Sen, Dimitrios Koutsonikolas, and Kyu-Han Kim. 2015. Widraw: Enabling hands-free drawing in the air on commodity wifi devices. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 77–89.

[34] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional human pose regression. In *Proceedings of the IEEE international conference on computer vision*. 2602–2611.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[36] Sheng Tan and Jie Yang. 2016. WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition. In *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*. 201–210.

[37] Sheng Tan, Jie Yang, and Yingying Chen. 2020. Enabling fine-grained finger gesture recognition on commodity wifi devices. *IEEE Transactions on Mobile Computing* 21, 8 (2020), 2789–2802.

[38] Jing Tian and Kai-Kuang Ma. 2011. A survey on super-resolution imaging. *Signal, Image and Video Processing* 5 (2011), 329–342.

[39] Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* 10, 5 (1999), 988–999.

[40] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained person perception using WiFi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5452–5461.

[41] Hao Wang, Daqing Zhang, Junyi Ma, Yasha Wang, Yuxiang Wang, Dan Wu, Tao Gu, and Bing Xie. 2016. Human respiration detection with commodity WiFi devices: Do user location and body orientation matter?. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 25–36.

[42] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the limit of acoustic gesture recognition. *IEEE Transactions on Mobile Computing* 21, 5 (2020), 1798–1811.

[43] Dan Wu, Ruiyang Gao, Youwei Zeng, Jinyi Liu, Leye Wang, Tao Gu, and Daqing Zhang. 2020. FingerDraw: Sub-wavelength level finger motion tracking with WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–27.

[44] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. Onefi: One-shot recognition for unseen gesture via cots wifi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 206–219.

[45] Yaxiong Xie, Zhenjiang Li, and Mo Li. 2015. Precise power delay profiling with commodity WiFi. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 53–64.

[46] Tianzhang Xing, Qing Yang, Zhiping Jiang, Xinhua Fu, Junfeng Wang, Chase Q Wu, and Xiaojiang Chen. 2022. WiFine: Real-time Gesture Recognition Using Wi-Fi with Edge Intelligence. *ACM Transactions on Sensor Networks* 19, 1 (2022),

1–24.

[47] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[48] Chao Yang, Lingxiao Wang, Xuyu Wang, and Shiwen Mao. 2022. Environment Adaptive RFID-Based 3D Human Pose Tracking With a Meta-Learning Approach. *IEEE Journal of Radio Frequency Identification* 6 (2022), 413–425.

[49] Qiang Yang, Kaiyan Cui, and Yuanqing Zheng. 2023. VoShield: Voice Liveness Detection with Sound Field Dynamics. In *Proceedings of IEEE INFOCOM*.

[50] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. 2019. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia* 21, 12 (2019), 3106–3121.

[51] Yongxin Yang and Timothy Hospedales. 2016. Deep multi-task representation learning: A tensor factorisation approach. *arXiv preprint arXiv:1605.06391* (2016).

[52] Nan Yu, Wei Wang, Alex X Liu, and Lingtao Kong. 2018. QGesture: Quantifying gesture distance and direction with WiFi signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–23.

[53] Liming Zhang, Jie Wang, Qinghua Gao, Xuanheng Li, Miao Pan, and Yuguang Fang. 2018. LetFi: Letter recognition in the air using CSI. In *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.

[54] Lei Zhang, Yixiang Zhang, and Xiaolong Zheng. 2020. Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–24.

[55] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. 2021. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11281–11292.

[56] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2021. Widar3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8671–8688.

[57] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.

[58] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 267–281.

[59] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), 1–20. https://doi.org/10.1109/TPAMI.2022.3195549

[60] Christian Zimmermann and Thomas Brox. 2017. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*. 4903–4911.