OCHID-Fi: Occlusion-Robust Hand Pose Estimation in 3D via RF-Vision

Shujie Zhang[1]*, Tianyue Zheng[1]*, Zhe Chen[2], Jingzhi Hu[1],
Abdelwahed Khamis[3], Jiajun Liu[3], Jun Luo[1]
[1] Nanyang Technological University   [2] Fudan University   [3] CSIRO
{shujie002,tianyue002,jingzhi.hu,junluo}@ntu.edu.sg
zhechen@fudan.edu.cn   abdelwahed.khamis@data61.csiro.au   jiajun.liu@csiro.au

Abstract

Hand Pose Estimation (HPE) is crucial to many applications, but conventional cameras-based CM-HPE methods are completely subject to Line-of-Sight (LoS), as cameras cannot capture occluded objects. In this paper, we propose to exploit Radio-Frequency-Vision (RF-vision) capable of bypassing obstacles for achieving occluded HPE, and we introduce OCHID-Fi as the first RF-HPE method with 3D pose estimation capability. OCHID-Fi employs wideband RF sensors widely available on smart devices (e.g., iPhones) to probe 3D human hand pose and extract their skeletons behind obstacles. To overcome the challenge in labeling RF imaging given its human incomprehensible nature, OCHID-Fi employs a cross-modality and cross-domain training process. It uses a pre-trained CM-HPE network and a synchronized CM/RF dataset, to guide the training of its complex-valued RF-HPE network under LoS conditions. It further transfers knowledge learned from labeled LoS domain to unlabeled occluded domain via adversarial learning, enabling OCHID-Fi to generalize to unseen occluded scenarios. Experimental results demonstrate the superiority of OCHID-Fi: it achieves comparable accuracy to CM-HPE under normal conditions while maintaining such accuracy even in occluded scenarios, with empirical evidence for its generalizability to new domains.

1. Introduction

We have witnessed tremendous efforts put into Computer Vision (CV) research in the past decade, driven by applications such as facial recognition [37, 45], hand pose estimation [24, 38], object detection [14, 57], and augmented/virtual reality [16, 53]. Among various sensing technologies behind CV, Optical Vision (OV) has so far been the dominant path, fuelled by the widely available OV devices (i.e., cameras, lidars) and large-scale datasets [10, 18]. However, OV often suffers from a few major limiting

(a) Normal   (b) Wood
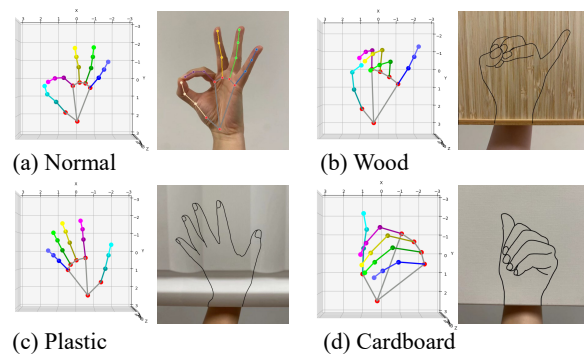
(c) Plastic   (d) Cardboard

Figure 1. Unlike CM-HPE models, OCHID-Fi can extract 3D hand keypoints behind occlusions (b-d). Note that the wrist and the position of the hand are shown here only as an example. In the actual experiment, the hand can be fully occluded and its position can vary. Drawn outlines here are for illustration purposes only.

factors: it requires Line-of-Sight (LoS) [10, 18] and certain lighting conditions [4, 46], it is prone to background clutter [17, 40], and is challenged by privacy concerns [1, 3].

Motivated by these shortcomings, non-optical vision technologies have also been explored in the CV community [7, 11, 28–30, 43, 47, 48, 52, 54, 55], among which *Radio-Frequency*-vision (*RF-vision*) stands out in many aspects including low complexity (thus real-time responsiveness), energy efficiency, and ready deployability [8, 29]. These strengths have motivated the use of RF-vision for problems previously tackled by OV [30, 43, 52]. However, one of the biggest strengths of RF-vision, i.e., *occlusion robustness*, has only been lightly touched by RF-Pose [52], for coarse-grained human pose estimation trained only in LoS domain and used directly in obstructed scenarios without accounting for the impact of the obstacles that create occlusion.

In this paper, we focus on utilizing RF-vision to perform fine-grained 3D *Hand Pose Estimation* (HPE) for occluded scenes. It is important to distinguish RF-HPE from RF-enabled *Hand Gesture Recognition* (RF-HGR) [42, 56]: While the former requires a more detailed understanding of hand keypoints [24, 39, 49], the latter only performs basic classification tasks. As a result, RF-HPE is highly non-

---

*Equal contribution. https://github.com/DeepWiSe888/OCHID-Fi

trivial and we summarize three major challenges:

- *Non-Euclidean Mapping of Keypoints:* RF data does not enjoy a direct Euclidean mapping from signal space to keypoint locations. It is hence difficult for a deep learning model to uncover the intrinsic relations.

- *Model Design for Low-resolution, Complex-valued RF data:* CM-HPE models cannot process low-resolution complex-valued RF data, while RF-HGR models are designed exclusively for classification while RF-HPE is inherently a regression task.

- *Model Training for Occluded Scenes:* RF data in occluded scenes are significantly affected by reflection and refraction caused by the obstacles. It is highly non-trivial to design a training mechanism for an RF-HPE model in such scenarios.

To address these challenges, we propose OCHID-Fi, the first 3D RF-HPE model capable of extracting 3D hand keypoints behind full occlusion. To provide a taste of what OCHID-Fi can achieve, we plot the outputs of OCHID-Fi and a state-of-the-art CM-HPE solution (Google MediaPipe Hands [49]) for a clear comparison in Figure 1. Trained in *cross-modality* and *cross-domain* manner, OCHID-Fi exploits RF-vision to tackle the occlusion issue where CM-HPE fails. OCHID-Fi translates RF signals to hand keypoints through cross-modality training aided by a pre-trained CM-HPE model. Specifically, OCHID-Fi employs a synchronized pair of camera and RF sensor during data collection in LoS scenarios. The CM-HPE network is first trained with pseudo ground truth collected by the Opti-Track [25] system, and then we transfer its learned knowledge to OCH-Net by supervising OCH-Net together with the RF ground truth data. To handle the complex-valued RF data, a deep complex-valued network is specifically designed to perform feature extraction. While completing the first training stage allows the deep complex-valued network OCH-Net (OCcluded Hand-Net) to work independently under LoS situation, the second stage training OCH-AL (OC-cluded Hand-Adversarial Learning) is performed to further transfer knowledge across domains (from LoS to occluded), for which we leverage adversarial learning in an unsupervised manner. In summary, our major contributions are:

- To the best of our knowledge, OCHID-Fi is the first occlusion-robust 3D RF-HPE model.
- OCHID-Fi transfers the knowledge from the OV to the RF, effectively bridging the gap between complex RF-vision data and hand keypoints.
- OCH-Net is proposed as the complex-valued RF-HPE model to fit RF signals, making it possible to fully exploit the intrinsic RF features.
- OCH-AL leverages adversarial learning in an unsupervised manner, so as to further transfer knowledge from the LoS domain into the occluded one.

- We perform extensive experiments to validate that, in occluded scenes where OV fails completely, OCHID-Fi achieves similar accuracy to that of CM-HPE in LoS conditions. Our empirical results also demonstrate that OCHID-Fi generalizes to unseen occluded scenes.

2. Related Work

Several OV methods exist for HGR [6, 23, 26, 50] using photos and videos. However, their functions cannot meet the need from the HPE problem, as HGR aims to only classify hand gestures rather than estimate locations of hand keypoints (such as knuckles) accurately. To this end, novel solutions for addressing the HPE problem have been devised based on visual inputs [9, 12, 13, 35, 38, 49]. Specifically, HandFoldingNet [9] uses depth image as input, Open-Pose [35] employs a multi-camera system for fine-grained hand keypoint detection, and others [12, 13, 38, 49] rely on neural networks for extracting hand keypoint out of single RGB images. Unfortunately, all these methods fail to accomplish the HPE tasks in the presence of occlusion where hands hidden behind, for example, cardboard or sleeves.

Recent research has demonstrated that it is possible to distinguish RF signals reflected by different parts of the human body [43, 52]. The shape and action amplitudes of body parts impact the intensity of reflected RF signals, thus enabling the reconstruction of human poses through deep analytics. However, these approaches are not directly applicable to tackle HPE, because they typically involve signal accumulation over time to detect body pose at a larger scale [43, 52]. Although RF has succeeded in addressing HGR [5, 36], these solutions are incompatible with what HPE demands for the same reason as explained earlier for CM-HGR. Again, none of these proposals is capable of handling HPE under occlusion.

3. OCHID-Fi Design

In this section, we present three key components of OCHID-Fi, namely a deep cross-modality framework, a deep complex-valued network OCH-Net, as well as a deep adversarial learning algorithm OCH-AL.

3.1. Overview

The overview of OCHID-Fi is illustrated in Figure 2. The camera and RF sensor are calibrated extrinsically according to their different positions [20], and synchronized using the network time protocol [22]. Before feeding RF signals to the neural models, an RF preprocessing module is employed to suppress the noise and improve the quality of signals. The RF preprocessing module employs a smooth filter followed by a band-pass filter to process RF signals [7]. After data collection and preparation, OCHID-Fi consists of the following three major components:
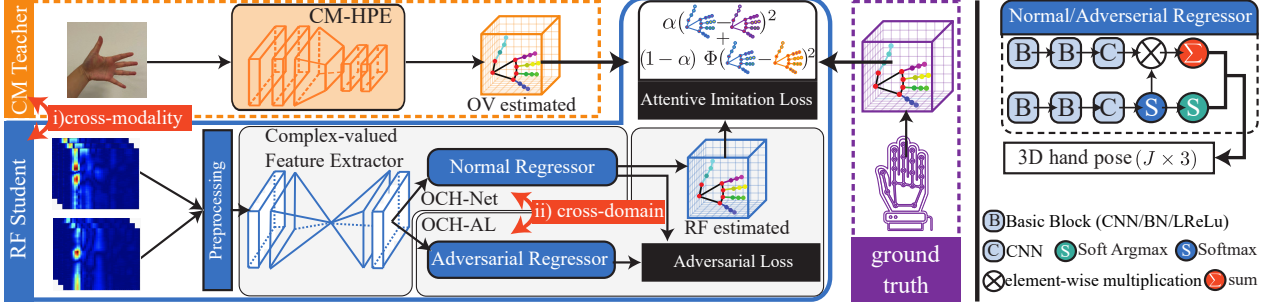
Figure 2. Architecture of OCHID-Fi: i) cross modality from OV to RF and ii) cross-domain from LoS to occluded.

C1: We pre-train a CM-HPE network using ground truth obtained from OptiTrack and transfer knowledge cross-modality from CM-HPE to RF-HPE under LoS situation, by minimizing an attentive imitation loss (Section 3.2), while the detailed construction of RF-HPE (feature extractor and regressor) is postponed to later sections.

C2: To handle complexed-valued RF signals, we construct our RF-HPE network OCH-Net as a complex-valued feature extractor and a 2D-3D regressor (Section 3.3).

C3: For generalizing towards occluded domain with unseen obstacles, we further employ adversarial learning to transfer knowledge from the LoS domain to the occluded one. Specifically, an adversarial regressor is added to form a minimax game with the OCH-Net regressor, allowing for an unsupervised training (Section 3.4).

Overall, the two-stage training of OCHID-Fi realizes both cross-modality from OV to RF and cross-domain from LoS to occluded, resulting in an occlusion-robust RF-HPE model. This is in stark contrast to the troublesome alternative that employs a synchronized camera at a vantage point (only if possible) for every occluded scenario to provide direct OV annotations and then trains an RF-vison model in a supervised manner. Essentially, this alternative still involves cross-modality and cross-domain training that are not always feasible for all scenarios, hence it is hardly generalizable to unseen domains.

3.2. Deep Cross-Modality Framework

In this section, we draw inspiration from recent research works [34, 44, 52] and propose a teacher-student framework for cross-modality training, aiming to tackle the challenge of mapping RF signals to hand keypoints. For brevity, we use the notations T for the OV teacher network, S for the RF student network, and $\boldsymbol{y}^{\mathsf{T}}$, $\boldsymbol{y}^{\mathsf{S}}$, $\boldsymbol{y}$ for the predictions of the CM-HPE network, the RF-HPE network, and the ground truth, respectively. Unlike a classification task, we cannot utilize the "dark knowledge" of the soft output in the logit space for knowledge distillation of HPE regression. Therefore, to effectively distill knowledge from the CM-HPE network we employ an attentive imitation loss [34] as follows:

$$L_y = (1/n)\sum_{i=1}^{n} \alpha(\boldsymbol{y}_i^{\mathsf{S}} - \boldsymbol{y}_i)^2 + (1-\alpha)\Phi_i(\boldsymbol{y}_i^{\mathsf{S}} - \boldsymbol{y}_i^{\mathsf{T}})^2,$$

where α is a scale factor, $\Phi_i = \left(1 - \frac{(\boldsymbol{y}_i^{\mathsf{T}} - \boldsymbol{y}_i)^2}{\eta}\right)$ is the normalized teacher loss for the i-th sample, $\eta = \max\left(e^{\mathsf{T}}\right) - \min\left(e^{\mathsf{T}}\right)$ is a normalization parameter obtained as the difference between the maximum and minimum of e^{T}, where $e^{\mathsf{T}} = \left\{(\boldsymbol{y}_j^{\mathsf{T}} - \boldsymbol{y}_j)^2 : j = 1, \ldots, N\right\}$ is a set of teacher loss from entire training data. In the loss function, Φ_i assigns different relative importance to each component, making the loss attentive. Since Φ_i is computed differently for each sample, it down-weights unreliable predictions made by the teacher T. This means that, during knowledge transfer from T to S, the cross-modality training relies more on the example data that T is good at predicting.

3.3. OCH-Net: A Deep Complex Perspective

Despite the guidance of the CM-HPE model, the RF-HPE model still faces challenges in understanding RF signals due to their complex-valued and high-dimensional tensor nature. Essentially, the *RF tensor* consists of data from multiple antennas acquiring RF pulses reflected by targets in the environment. Since a single RF pulse is insufficient for vision, multiple pulses are transmitted both in time and space (via different antennas) and the received signals are combined to form the RF tensor $\boldsymbol{X}$, whose every element is a complex number with I/Q components $\boldsymbol{x}(n, t, d) = \boldsymbol{x}_{\mathsf{I}} + j\boldsymbol{x}_{\mathsf{Q}}$, where n, t and d respectively indicate the transmitter-receiver pair in the antenna array, the number of RF pulses, and the number of discretized range bins [8]. In particular, the phase of each $\boldsymbol{x}(n, t, d)$ carries important information about the relative displacement of the targets. To fully exploit the potential in the I/Q components of RF data, we propose OCH-Net that specializes in handling the complex-valued RF data tensor.

Building Blocks for Complex Before diving into the full architecture, we first introduce our updated version of 2D convolution building blocks for handling complex-valued RF tensor $\boldsymbol{X} = [\boldsymbol{x}(n, t, d)]$, in which n is treated as the number of channels of 2D input, t and d are treated as height and width. Conventional 2D CNN for real-valued data at the ℓ-th layer could be described as $\boldsymbol{X}^{\ell} = \boldsymbol{W}^{\ell} * \boldsymbol{X}^{\ell-1}$, where $\boldsymbol{W}^{\ell}$ is the convolution kernel and the symbol $*$ represents the convolution operator. Given $\boldsymbol{X} = \boldsymbol{X}_{\mathsf{I}} + j\boldsymbol{X}_{\mathsf{Q}}$

as the RF tensor, we perform the complex-valued convolution using two new real-valued kernels $\boldsymbol{W}_\mathrm{I}^\ell$ and $\boldsymbol{W}_\mathrm{Q}^\ell$ [41]. Instead of simply stacking the two parts, we make $\boldsymbol{X}^\ell = (\boldsymbol{W}_\mathrm{I}^\ell * \boldsymbol{X}_\mathrm{I}^{\ell-1} - \boldsymbol{W}_\mathrm{Q}^\ell * \boldsymbol{X}_\mathrm{Q}^{\ell-1}) + j(\boldsymbol{W}_\mathrm{I}^\ell * \boldsymbol{X}_\mathrm{Q}^{\ell-1} + \boldsymbol{W}_\mathrm{Q}^\ell * \boldsymbol{X}_\mathrm{I}^{\ell-1})$. We also redefine the complex-valued nonlinear activated function as $\sigma_\mathrm{C}(\boldsymbol{X}) = \sigma(\boldsymbol{X}_\mathrm{I}) \oplus \sigma(\boldsymbol{X}_\mathrm{Q})$, where σ is the original activation function, and $\oplus$ represents the operation of concatenation. The downsampling, upsampling, and batch normalization layers are similarly redefined by processing the real and imaginary branches separately with their real-valued counterparts and then concatenating the results. This whole procedure is visualized in Figure 3.

OCH-Net Architecture Leveraging the complex-valued building blocks, we design OCH-Net with a feature extractor followed by a regressor, as illustrated in Figure 2. The feature extractor is based on a popular encoder-decoder architecture with skip connections [32]. The encoder and decoder blocks are paired and connected via a skip connection to facilitate information flow. In this way, OCH-Net utilizes fine-grained details learned in the encoder part to estimate hand poses in the decoder part.

In particular, each encoder block in the system consists of a complex-valued convolutional layer, a batch normalization layer, and a nonlinear activation layer with leaky ReLU. After every three blocks, the number of channels increases, and a max-pooling layer is applied to enhance the most prominent feature and reduce the dimension of the hidden layers, thus reducing complexity. Once the bottleneck block extracts a representation in the latent space, an upsampling layer is utilized to reverse the compression for the decoding process. Similarly, each decoder block contains the same components as the encoder block. After every three blocks, the number of channels decreases, and an upsampling layer is inserted to maintain the shape of the feature map. Finally, the decoded features are fed to a regressor to map to 3D hand keypoints. As shown in the upper-right Figure 2, the lower branch of the regressor recovers the 2D hand keypoints, while the upper branch recovers the depth of the keypoints. Finally, the 3D keypoints are reconstructed by combining the 2D and depth estimations.

To demonstrate the necessity of using OCH-Net to handle complex-valued RF signals, we visualize the feature
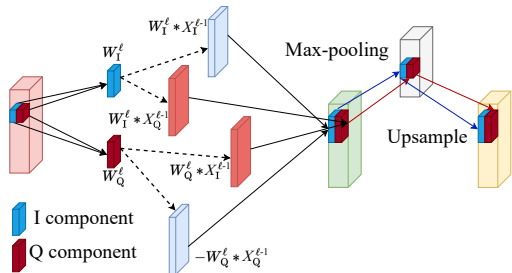

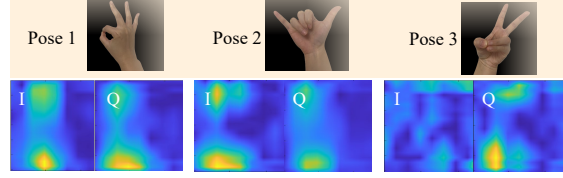Figure 3. Network operations for processing RF tensor.

Figure 4. Activated feature maps of different poses.

maps in one of OCH-Net's hidden layer for various hand pose RF data. The results in Figure 4 clearly indicate that I and Q data generates feature maps with distinctive differences, with bright yellow regions representing activated neurons in the feature maps. These findings highlight that relying solely on either I or Q component of the RF data as input can be insufficient for capturing all intrinsic features of RF data for estimating hand poses, potentially resulting in missing crucial information.

3.4. Deep Adversarial Learning for Occlusion

Since OCH-Net only works for the normal domain, we need to transfer the learned knowledge to the occluded domain. We denote the n-th sample of the RF data as $\boldsymbol{x}_n$, and in unsupervised deep adversarial learning, we denote the labeled normal domain as $D^\mathrm{NO} = \{(\boldsymbol{x}_n^\mathrm{NO}, \boldsymbol{y}_n^\mathrm{NO})\}_{n=1}^N$, and the unlabeled occluded domain as $D^\mathrm{OC} = \{\boldsymbol{x}_m^\mathrm{OC}\}_{m=1}^M$. The objective of this training algorithm is to maximize the regressive accuracy in the occluded domain.

Minimax Game Due to the lack of OV annotations in $D^\mathrm{OC} = \{\boldsymbol{x}_m^\mathrm{OC}\}_{m=1}^M$, we need an unsupervised training approach. We introduce an adversarial regressor g' to form a minimax game with the OCH-Net regressor g, so as to minimize the expected loss of g on the occluded domain while maintaining good performance on the normal domain. To this end, we leverage the disparity discrepancy theory [15, 51] that allows for a proper alignment between two domains via disparity reduction.

We define the *disparity* between two regressors g and g' as the expected loss over a domain D, denoted by $\mathrm{disp}_D(g', g) = \mathbb{E}_D L(g', g)$, and the *disparity discrepancy* induced by g' as the supremum of the difference between the disparities of D^NO and D^OC over a hypothesis space $\mathcal{G}$, denoted by $d_{g,\mathcal{G}}(D^\mathrm{NO}, D^\mathrm{OC}) = \sup_{g' \in \mathcal{G}}(\mathrm{disp}_{D^\mathrm{OC}}(g', g) - \mathrm{disp}_{D^\mathrm{NO}}(g', g))$. It is proven in [51] that we can strictly bound the expected error of g on the occluded domain by the following minimization objective:

$$\min_{g \in \mathcal{G}} \epsilon_{D^\mathrm{NO}}(g) + d_{g,\mathcal{G}}(D^\mathrm{NO}, D^\mathrm{OC}),$$

where $\epsilon_{D^\mathrm{NO}}(g) = \mathbb{E}_{(\boldsymbol{x}^\mathrm{NO}, \boldsymbol{y}^\mathrm{NO}) \in D^\mathrm{NO}} L_a(g(\boldsymbol{x}^\mathrm{NO}), \boldsymbol{y}^\mathrm{NO})$ is the expected regression loss L_a of g in the normal domain. Since $\epsilon_{D^\mathrm{NO}}(g)$ is determined by g during the pre-training stage, only the second term, i.e., the disparity discrepancy between the two domains, needs to be minimized. This term

can be approximated by maximizing over g' as a deep learning model instead of taking supremum in the space $\mathcal{G}$, with ϕ as the parameter-fixed feature extractor:

$$d_{g,\mathcal{G}}(D^{\mathrm{NO}}, D^{\mathrm{OC}}) \approx \max_{g'} \left(\mathrm{disp}_{D^{\mathrm{OC}}}(g', g) - \mathrm{disp}_{D^{\mathrm{NO}}}(g', g) \right)$$

$$= \max_{g'} \mathbb{E}_{\boldsymbol{x}^{\mathrm{OC}} \in D^{\mathrm{OC}}} L_a \left((g' \circ \phi)(\boldsymbol{x}^{\mathrm{OC}}), (g \circ \phi)(\boldsymbol{x}^{\mathrm{OC}}) \right)$$

$$- \mathbb{E}_{\boldsymbol{x}^{\mathrm{NO}} \in D^{\mathrm{NO}}} L_a \left((g' \circ \phi)(\boldsymbol{x}^{\mathrm{NO}}), (g \circ \phi)(\boldsymbol{x}^{\mathrm{NO}}) \right). \quad (1)$$

After training g' to approximate the disparity discrepancy $d_{g,\mathcal{G}}(*)$, minimizing the following equation will decrease the error of g in the occluded domain effectively:

$$\min_{\phi, g} \mathbb{E}_{(\boldsymbol{x}^{\mathrm{NO}}, \boldsymbol{y}^{\mathrm{NO}}) \in D^{\mathrm{NO}}} L_a \left((g \circ \phi)(\boldsymbol{x}^{\mathrm{NO}}), \boldsymbol{y}^{\mathrm{NO}} \right) + d_{g,\mathcal{G}}(*).$$

To implement this step, we fix the parameters of g', and update the parameters of g and ϕ by backpropagation. This process essentially creates a minimax game between two regressors, each working towards achieving opposing goals. However, their collaboration enables the adaptation from the normal domain to the occluded domain.

Bounding Output Space. Compared with a classification task, HPE involves regression with a much larger output space. Specifically, if we consider a target hand pose existing in a 3D voxel space with dimensions of height H, width W, and depth R, and treat each output voxel as a class, we would have $H \times W \times R$ classes. Consequently, the large number of output classes will increase the bound of occluded domain error. Therefore, it is necessary to reduce the output space for the HPE network. As pointed out by [15], there is an intrinsic sparsity in the keypoint positions: when inferring hand poses under occlusion using $g \circ \phi$, even if the estimated hand pose is incorrect, the keypoint positions are still on the hand and likely overlap with other keypoint positions. For instance, the incorrect little finger keypoint position may appear on the index finger's position, but not in the background (more examples are illustrated in Section 4.2.2). This indicates that the output space size can be bounded to a limited set.

Therefore, during the training of OCH-AL, it only needs to pay more attention to the limited set. To achieve this, we can accumulate all incorrectly predicted hand poses while training in the normal domain and compute the distribution of each false hand pose heatmap $m_{\mathrm{GF}}(\hat{\boldsymbol{y}}_i^{\mathrm{NO}}) = \sum_{q \neq i} \mathcal{N}(\hat{\boldsymbol{y}}_q^{\mathrm{NO}})$ where $\mathcal{N}$ is the 3D truncated Gaussian function, and $\boldsymbol{y}_i^{\hat{\mathrm{NO}}}$ is the i-th keypoint heatmap of a hand pose predicted by the normal domain regressor g. Furthermore, in Eqn. (1), we observe that only the occluded predictions $(g' \circ \phi)(\boldsymbol{x}^{\mathrm{OC}})$ are used to maximize the disparity $\mathbb{E}_{\boldsymbol{x}^{\mathrm{OC}} \in D^{\mathrm{OC}}} L_a((g' \circ \phi)(\boldsymbol{x}^{\mathrm{OC}}), (g \circ \phi)(\boldsymbol{x}^{\mathrm{OC}}))$ by updating the adversarial regressor g'. Therefore, since the estimated supremum of the disparity discrepancy likely occurs on a false prediction, we can further leverage the false heatmap distribution $m_{\mathrm{GF}}(\hat{\boldsymbol{y}}_i^{\mathrm{NO}})$ to get $\tilde{\mathcal{L}}_a(\boldsymbol{x}^{\mathrm{OC}}) =$

$\mathbb{E}_{\boldsymbol{x}^{\mathrm{OC}} \in D^{\mathrm{OC}}} L_a((m_{\mathrm{GF}} \circ g' \circ \phi)(\boldsymbol{x}^{\mathrm{OC}})), (g \circ \phi)(\boldsymbol{x}^{\mathrm{OC}}))$, and revise Eqn. (1) with the updated first term to obtain:

$$\max_{g'} d_{g,\mathcal{G}}(D^{\mathrm{NO}}, D^{\mathrm{OC}}) = \tilde{\mathcal{L}}_a(\boldsymbol{x}^{\mathrm{OC}}) - \mathcal{L}_a(\boldsymbol{x}^{\mathrm{NO}}),$$

where $\mathcal{L}_a(\boldsymbol{x}^{\mathrm{NO}}) = \mathbb{E}_{\boldsymbol{x}^{\mathrm{NO}} \in A} L_a((g' \circ \phi)(\boldsymbol{x}^{\mathrm{NO}}), (g \circ \phi)(\boldsymbol{x}^{\mathrm{NO}}))$.

4. Evaluation

4.1. Experiment Setup

Data Collection For OV dataset, we use a camera with 1080×1920 pixels and $30\,\mathrm{Hz}$ frame rate. For RF dataset, since accessing raw signals from commodity devices (e.g., iPhone) is impossible, we emulate such an RF sensor by an IR-UWB radar [27] with 10 antennas [8]. The frame rate of the RF sensor is set to $150\,\mathrm{Hz}$, and we use a Rockchip PX30 [31] to control the sensor. Both the camera and RF sensor are connected to a PC to be synchronized. Opti-Track [2] is used for obtaining 3D hand pose ground truth.

We collect data with 30 volunteers in 5 different environments including a classroom, a living room, a bedroom, a lab cubicle, and a conference room. We use hand poses from American Sign Language [19] (including transition hand poses between two signs) along with randomly moving wrists and fingers, covering nearly all feasible variations of the hand's degrees of freedom. The dataset includes 20 hours of normal condition data and 20 hours of occluded data. In the occluded scenarios, we utilize a variety of obstacles including wood, plastic, (frosted) glass, and cardboard sheets. These sheets are available in different areas, ranging from $0.5\mathrm{m}^2$ to $1\mathrm{m}^2$, with various widths between $1\,\mathrm{cm}$ and $10\,\mathrm{cm}$, and placed at $10\,\mathrm{cm}$ in front of the RF sensor. The distance from the hand to the RF sensor varies from $20\,\mathrm{cm}$ to $80\,\mathrm{cm}$. To annotate the OV dataset, we attach motion capture markers to the keypoints on a hand and retrieve the corresponding 3D coordinates from OptiTrack. Our evaluations use 10 hours of normal condition data (120,000 samples) to train OCH-Net, then these data with additional 2 hours of occluded data (24,000 samples) to train OCH-AL, and finally all the remaining data are used for testing.

Teacher Network We use MediaPipe Hands [49], a widely adopted CM-HPE network in many applications as our teacher network. We specifically set two key parameters of MediaPipe Hands, i.e., the maximum number of hands and minimum detection confidence as 1 and 0.5, respectively. The network is first pre-trained with OV data and ground truth collected by OptiTrack. Subsequently, we leverage MediaPipe's output and the ground truth to transfer knowledge to OCH-Net.

Training Details We train and evaluate our method on a server with NVIDIA RTX 1080 GPU. We implement the OCH-Net and OCH-AL on Python 3.8.16 and Pytorch 1.10.0. The input RF tensors have the size of $10 \times 40 \times 40$,

with the scale factor α set to 0.5. In the feature extractor, the number of channels is set to 10, 64, 128, 256, and 512 in the encoder, and 512, 256, 128, 64, and 32 in the decoder for each convolutional layer. The regressor, as mentioned in Section 3.3, takes in concatenated real and imaginary features of 32 channels each to have in total 64 channels. These channels are then used to predict 21 keypoint heatmaps. We set the batch size to 8, and adopt an Adam optimizer with a learning rate of 0.001, β_1 of 0.9, and β_2 of 0.999.

Evaluation Metric The performance of HPE is evaluated using the percentage of correct keypoint (PCK) metric [24, 38, 39] defined as follows:

$$\text{PCK@}a = \frac{1}{N} \sum_{n=1}^{N} \Xi \left(\frac{\|\boldsymbol{y}_n^{\text{pred}} - \boldsymbol{y}_n^{\text{gt}}\|_2^2}{\sqrt{w_n^2 + h_n^2 + d_n^2}} \leq a \right),$$

where N is the number of test samples, Ξ is a logical operation that outputs 0 if the expression is false and 1 if true, $\boldsymbol{y}_n^{\text{pred}}$ denotes the predicted keypoint position, $\boldsymbol{y}_n^{\text{gt}}$ denotes the ground truth keypoint position, and $\sqrt{w_n^2 + h_n^2 + d_n^2}$ is the bounding box size of the hand. The PCK score ranges from 0 to 1, with higher scores indicating better performance. Typically, a normalized distance error of $a = 0.2$ is used as the threshold for successful HPE [24, 38, 39]. To gain a clearer understanding of the performance of different parts of the hand, we also calculate PCKs at metacarpophalangeal (MCP), proximal interphalangeal (PIP), distal interphalangeal (DIP), and fingertip joints, as defined in [38].

4.2. Performance Evaluations

4.2.1 Performance of OCH-Net

We study the overall performance of OCH-Net using normal condition data with two existing RF vision methods as baselines, namely *Person-in-WiFi* [43] and *RF-Pose* [52]. To the best of our knowledge, there is no directly related research work designed for hand pose with RF vision. Therefore, we have to modify the current human body skeleton neural network models [43, 52] to make them comparable to OCH-Net. To be specific, we maintain the main part of their network architecture but replace their input and output to match our HPE task. Moreover, OCH-Net can adapt to a different number of RF data streams as the network considers it as the number of input channels. To validate this, we create another baseline *OCH-Net-Slim* by extracting 2 data streams out of the 10 data streams collected from the RF antennas. In this experiment, we test all approaches under both normal and occluded scenarios.

The results of the experiment, as shown in Figure 5, demonstrate that both OCH-Net and OCH-Net-Slim outperform Person-in-WiFi and RF-Pose in terms of PCK@0.2 for all parts of the hand. In particular, OCH-Net outperforms the two baselines by more than 10%. Furthermore, since OCH-Net utilizes more RF data streams, it also outperforms

(a) MCP. (b) PIP.
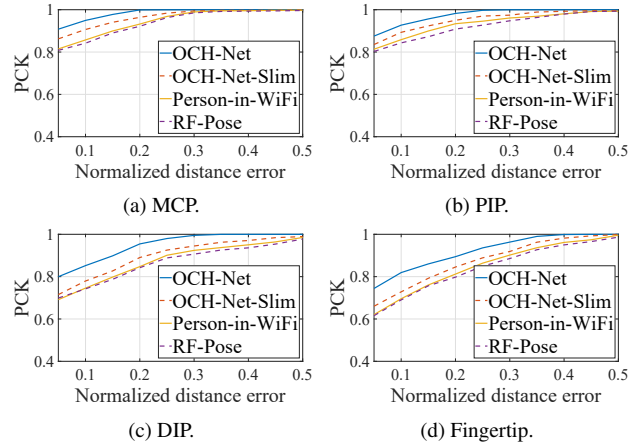
(c) DIP. (d) Fingertip.

Figure 5. The PCKs of OCH-Net and different baselines.

OCH-Net-Slim by a small margin of approximately 5% for all parts of the hand. There are two main advantages of using OCH-Net over the two baselines. On one hand, OCH-Net gains its knowledge from both the ground truth and the teacher network, effectively bridging the gap between complex RF-vision data and hand keypoints. This enables us to achieve non-Euclidean mapping. On the other hand, the deep complex-valued building blocks in OCH-Net are better suited for interpreting and encoding RF data, as explained in Section 3.3.

4.2.2 Performance of OCH-AL under Occlusion

As described in Section 3.4, we employ the OCH-AL framework to adapt from the normal domain into the occluded one. Data from all 4 types of obstacles, namely wood, plastic, glass, and cardboard sheets are used for adaptation. To demonstrate the performance of OCH-AL, we present examples of recovered poses under various oc-
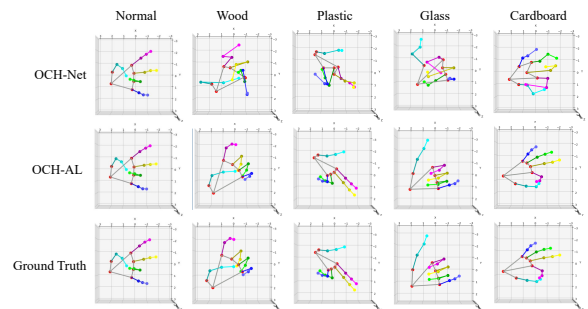
(a) 3D HPE before and after OCH-AL under LoS and occlusion.
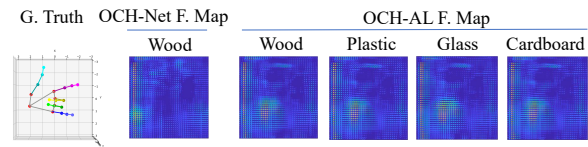
(b) Feature maps of before and after OCH-AL under occlusion.

Figure 6. Qualitative results before and after OCH-AL.
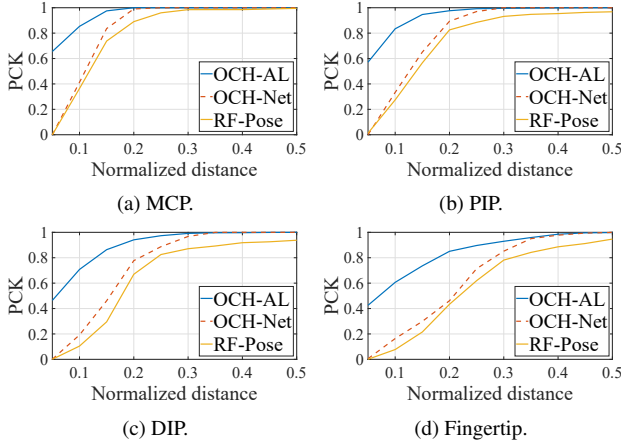
(a) MCP. (b) PIP.

(c) DIP. (d) Fingertip.

Figure 7. PCKs of OCH-Net and OCH-AL under occlusion.

cluded scenarios in Figure 6a: while OCH-Net can partially recover hand poses under occlusion, most of the recovered poses deviate from the ground truth, with a few joints overlapping with each other. In comparison, OCH-AL successfully adapts to the occluded domain and learns to handle the "twisted" RF signals caused by occlusion. We also show the feature maps at the bottleneck of the feature extractor in Figure 6b; it further confirms that OCH-Net achieves almost invariant feature maps for the same gesture despite occluded by distinct obstacles.

We further plot the performance of OCH-AL for different keypoints in Figure 7. The results show OCH-AL achieves PCKs@0.2 of 0.9998, 0.9763, 0.9410, and 0.8506 for MCP, PIP, DIP, and fingertip, respectively. In comparison, the PCKs@0.2 of OCH-Net for the same keypoints are 0.9904, 0.8934, 0.7772, and 0.4615, respectively. These findings demonstrate OCH-AL successfully adapts OCH-Net from the normal domain to the occluded one, resulting in significant performance improvements for all hand parts. Moreover, OCH-AL outperforms RF-Pose without domain adaptation, in terms of PCKs@0.2, by 0.1094, 0.1510, 0.2713, and 0.4169 for MCP, PIP, DIP, and fingertip respectively. These improvements emphasize the need for OCH-AL: although RF-vision can bypass obstacles, its signals may be substantially altered by obstacle materials. Such signal variations result in different data distributions, rendering domain adaptation a necessary step.

4.2.3 Generalization to Unseen Occluded Scenarios

To validate the generalizability of OCH-AL, we conduct additional experiments to assess its performance on unseen obstacles during adaptation. Specifically, we select three types of obstacles from wood, plastic, glass, and cardboard sheets for adaptation, and leave one out for testing. The results presented in Figure 8 show the PCKs@0.2 values achieved by OCH-AL for various keypoints. Our findings suggest that OCH-AL yields remarkable PCKs@0.2 scores

far exceeding those of the average PCKs of OCH-Net. Furthermore, our experiments reveal that OCH-AL performs slightly better when the obstacle is plastic or cardboard than when it is glass or wood, possibly due to the lower dielectric and loss tangent of these materials. These results strongly demonstrate the generalizability of OCH-AL to unseen domains, which is a significant advantage over methods that require time-consuming and cumbersome retraining.

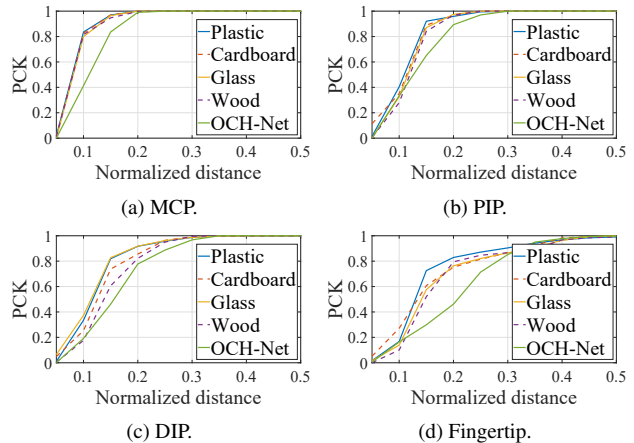(a) MCP. (b) PIP.

(c) DIP. (d) Fingertip.

Figure 8. Comparing PCKs of OCH-AL and the average PCKs of OCH-Net under unseen occlusion.

4.2.4 Adversarial Learning Algorithm Comparison

We compare our OCH-AL with two other adversarial learning algorithms DANN [21] and MCD [33]. To ensure a fair comparison, we use the same feature extractor in all three algorithms. We calculate the average PCKs@0.2 for each algorithm and show them in Figure 9. All three algorithms improve the original OCH-Net 's performance, but our OCH-AL achieves the highest PCK@0.2. Moreover, we observe that while MCD and DANN help OCH-Net adapt to occlusion, their average PCKs@0.2 in the normal domain decrease by 0.10 and 0.13, respectively. In contrast, OCH-AL maintains consistent performance in both normal and occluded domains. As described in Section 3.4, OCH-AL is specifically designed for the HPE regression task by bounding the size of the output space, so it achieves successful domain adaptation while avoiding unnecessary parameter updates, thus preserving OCH-Net's performance in the normal domain.
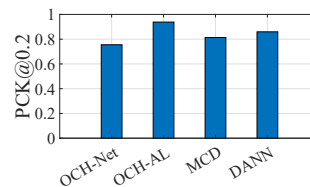

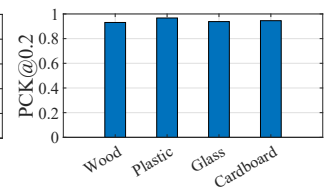Figure 9. Different adversarial learning algorithms.

Figure 10. Different obstacle materials.

15118

4.2.5 Ablation Study

We hereby evaluate the different components of OCH-Net, and study their impact on the performance. There are three baselines called *Real-Net*, *I-Net*, and *Q-Net*. The Real-Net is implemented by removing all the complex-valued CNN operations designed in Section 3.3, and simply concatenating the I and Q branches. The I-Net and Q-Net use the same structure as Real-Net, but are trained only on the I or Q RF data, respectively. In this experiment, we consider only data from the normal domain.

(a) MCP.
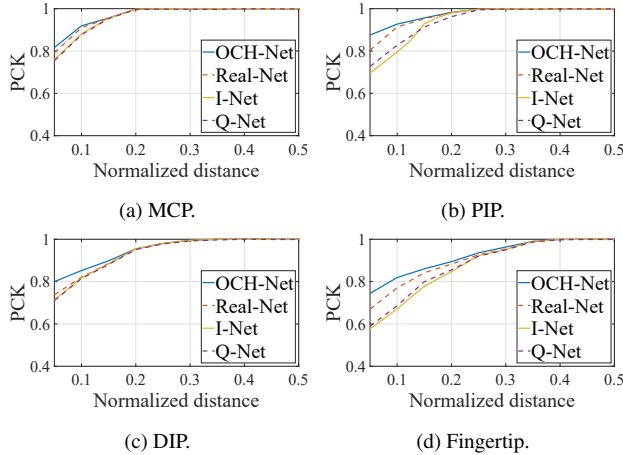
(b) PIP.

(c) DIP.

(d) Fingertip.

Figure 11. PCKs of different neural network structures.

Apparently, OCH-Net outperforms all three baselines for all parts of the hand. The PCKs@0.2 of OCH-Net are 0.9988, 0.9823, 0.9546, and 0.8943 for MCP, PIP, DIP, and fingertip, respectively. Due to the power of complex-valued CNN operations, OCH-Net obtains noticeable improvement over Real-Net, I-Net, and Q-Net. Moreover, I-Net and Q-Net achieve similar PCKs@0.2 for all parts of the hand, and are consistently the worst among all baselines. The reason is that single I or Q neural network structure cannot represent the whole RF data intrinsically. Moreover, it can be observed that PCKs decrease slightly from Figure 11a to Figure 11d, probably because the motion-induced errors from MCP to fingertip grows larger progressively.

4.2.6 Impact of Different Obstacle Materials

We investigate the impact of obstacle materials including wood, plastic, glass, and cardboard on the performance of our system. To achieve this, we place these materials in front of the RF sensor to block all LoS RF signals. As shown in Figure 10, the average PCKs@0.2 are found to be 0.9309, 0.9672, 0.9381, and 0.9464 for wood, plastic, glass, and cardboard, respectively. Notably, the worst performance is observed with the wood block, which causes the largest interference to RF signals among the four obstacles, thereby altering the RF signal distribution reflected by a hand to the most extent.
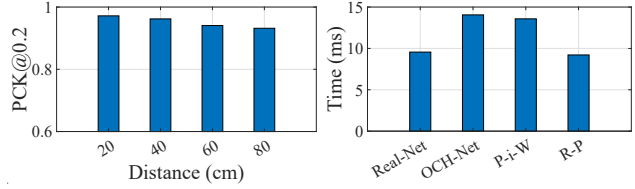

Figure 12. Different distances.   Figure 13. Inference time.

4.2.7 Impact of Different Distances

To evaluate the impact of different distances, we test the trained OCHID-Fi with data containing hands at different distances from the RF sensor. The results, depicted in Figure 12, reveal that the average PCKs@0.2 decreases by 4% as the distance increases from 20cm to 80cm. This degradation can be explained by the fact that the RF sensor captures more interference as the distance increases. Moreover, the receiving power of RF signals decreases with distance, resulting in lower SNRs in reflected RF signals. Despite the slight performance degradation, OCHID-Fi can still provide sufficient HPE accuracy at a further distance.

4.2.8 Inference Time of OCHID-Fi

We further evaluate the inference time of OCHID-Fi. In the inference stage, we only keep the feature extractor ϕ and the normal regressor g for assessment. We compare the inference time of OCH-Net with three baselines, and show the average inference time in Figure 13. The average inference time of OCH-Net is 14ms, which is only slightly higher than those of the baselines by up to 5ms. The main reason for the extra computing overhead is the use of complex-valued neural networks. If we replace the OCH-Net with Real-Net, the inference time is decreased to 10ms. However, we believe that the modest overhead of only 4ms is worth the superior performance that OCH-Net provides for our HPE task.

5. Conclusion

HPE in occluded scenarios is a crucial yet challenging problem pertinent to human-computer interaction. In this paper, to overcome the LoS limitations of CM-HPE, we resort to RF-vision, and propose OCHID-Fi as a cross-modality, cross-domain method for occlusion-robust HPE. Employing the carefully designed cross-modality framework, we have demonstrated OCHID-Fi's ability to map RF signals to hand keypoints in a non-Euclidean manner. Furthermore, OCHID-Fi has successfully adapted its neural model OCH-Net to deal with diversified obstacles by leveraging the power of adversarial learning. Extensive experiments have been conducted to demonstrate that OCHID-Fi achieves high accuracy in HPE, even in occluded scenarios. The results strongly support the effectiveness of this method, showing its potential for practical applications in fields such as human-computer interaction (HCI), smart-home controls, and medical rehabilitation.

References

[1] Are Security Cameras Legal? https://www.security.org/security-cameras/legality/. Accessed: 2022-05-19. 1

[2] OptiTrack - Motion Capture Systems. https://optitrack.com/. Accessed: 2023-02-19. 5

[3] Security Camera Laws, Rights, and Rules. https://www.safewise.com/security-camera-laws/. Accessed: 2022-05-19. 1

[4] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. A Histogram Modification Framework and Its Application for Image Contrast Enhancement. *IEEE Transactions on Image Processing*, pages 1921–1935, 2009. 1

[5] Xiaodong Cai, Jingyi Ma, Wei Liu, Hemin Han, and Lili Ma. Efficient Convolutional Neural Network for FMCW Radar based Hand Gesture Recognition. In *Adjunct Proc. of the UbiComp/ISWC*, pages 17–20, 2019. 2

[6] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks With Spatiotemporal Transformer Modules. In *Proc. of the CVF/IEEE ICCV*, pages 3763–3771, 2017. 2

[7] Zhe Chen, Chao Cai, Tianyue Zheng, Jun Luo, Jie Xiong, and Xin Wang. RF-based Human Activity Recognition using Signal Adapted Convolutional Neural Network. *IEEE Transactions on Mobile Computing*, 2021. 1, 2

[8] Zhe Chen, Tianyue Zheng, and Jun Luo. Octopus: A Practical and Versatile Wideband MIMO Sensing Platform. In *Proc. of the 27th ACM MobiCom*, pages 601–614, 2021. 1, 3, 5

[9] Wencan Cheng, Jae Hyun Park, and Jong Hwan Ko. Handfoldingnet: A 3d hand pose estimation network using multiscale-feature guided folding of a 2d hand skeleton. In *Proc. of the IEEE/CVF CVPR*, pages 11260–11269, 2021. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *Proc. of the IEEE/CVF CVPR*, pages 248–255, 2009. 1

[11] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. RF-Net: A Unified Meta-learning Framework for RF-enabled One-shot Human Activity Recognition. In *Proc. of the 18th ACM SenSys*, pages 517–530, 2020. 1

[12] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D Hand Shape and Pose Estimation From a Single RGB Image. In *2019 IEEE/CVF CVPR*, pages 10825–10834, 2019. 2

[13] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In *Proc. of ECCV*, pages 118–134, 2018. 2

[14] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proc. of the ECCV*, pages 784–799, 2018. 1

[15] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive Domain Adaptation for Unsupervised Keypoint Detection. In *Proc. of the IEEE/CVF CVPR*, pages 6780–6789, 2021. 4, 5

[16] Jinwoo Kim, Woojae Kim, Heeseok Oh, Seongmin Lee, and Sanghoon Lee. A Deep Cybersickness Predictor Based on Brain Signal Analysis for Virtual Reality Contents. In *Proc. of the CVF/IEEE ICCV*, pages 10580–10589, 2019. 1

[17] Eirini Konstantinou, Joan Lasenby, and Ioannis Brilakis. Adaptive Computer Vision-based 2D Tracking of Workers in Complex Environments. *Automation in Construction*, 103:168–184, 2019. 1

[18] Akio Kosaka, Akito Saito, Yukihito Furuhashi, and Takao Shibasaki. Augmented Reality System for Surgical Navigation using Robust Target Vision. In *Proc. of the IEEE/CVF CVPR*, volume 2, pages 187–194, 2000. 1

[19] Scott K. Liddell and Robert E. Johnson. American Sign Language: The Phonological Base. *Sign Language Studies*, 64(1):195–277, 1989. 5

[20] Teck-Yian Lim, Spencer A Markowitz, and Minh N Do. RaDICaL: A Synchronized FMCW Radar, Depth, IMU and RGB Camera Data Dataset With Low-Level FMCW Radar Signals. *IEEE Journal of Selected Topics in Signal Processing*, 15(4):941–953, 2021. 2

[21] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional Adversarial Domain Adaptation. *Proc. of NeurIPS*, 31, 2018. 7

[22] David L. Mills. Internet Time Synchronization: the Network Time Protocol. *IEEE Transactions on Communications*, 39(10):1482–1493, 1991. 2

[23] Yuecong Min, Yanxiao Zhang, Xiujuan Chai, and Xilin Chen. An Efficient PointLSTM for Point Clouds Based Gesture Recognition. In *Proc. of the IEEE/CVF CVPR*, pages 5761–5770, 2020. 2

[24] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single rgb Image. In *Proc. of ECCV*, pages 548–564, 2020. 1, 6

[25] Gergely Nagymáté and Rita M. Kiss. Application of Opti-Track Motion Capture Systems in Human Movement Analysis: A Systematic Literature Review. *Recent Innovations in Mechatronics*, 5(1):1–9, 2018. 2

[26] Xuan Son Nguyen, Luc Brun, Olivier Lézoray, and Sébastien Bougleux. A Neural Network Based on SPD Manifold Learning for Skeleton-Based Hand Gesture Recognition. In *Proc. of the IEEE/CVF CVPR*, pages 12036–12045, 2019. 2

[27] Novelda AS. The World Leader in Ultra Wideband (UWB) Sensing. https://novelda.com/technology/, 2017. Accessed: 2022-07-23. 5

[28] Omer Oralkan, A. Sanli Ergun, Jeremy A. Johnson, Mustafa Karaman, Utkan Demirci, Kambiz Kaviani, Thomas H. Lee, and Butrus T. Khuri-Yakub. Capacitive Micromachined Ultrasonic Transducers: Next-generation Arrays for Acoustic Imaging? *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 49(11):1596–1610, 2002. 1

[29] Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-View Radar Semantic Segmentation. In *Proc. of the CVF/IEEE ICCV*, pages 15671–15680, 2021. 1

[30] Kun Qian, Shilin Zhu, Xinyu Zhang, and Li Erran Li. Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar and Radar Signals. In *Proc. of the IEEE/CVF CVPR*, pages 444–453, 2021. 1

[31] Ltd. Rockchip Electronics Co. Rockchip PX30 Datasheet. https://rockchip.fr/PX30%20datasheet%20V1.1.pdf, 2018. Accessed: 2022-07-23. 5

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 4

[33] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In *Proc. of the IEEE/CVF CVPR*, pages 3723–3732, 2018. 7

[34] Muhamad Risqi U. Saputra, Pedro P.B. De Gusmao, Yasin Almalioglu, Andrew Markham, and Niki Trigoni. Distilling Knowledge from a Deep Pose Regressor Network. In *Proc. of the CVF/IEEE ICCV*, pages 263–272, 2019. 3

[35] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *Proc. of the IEEE/CVF CVPR*, pages 1145–1153, 2017. 2

[36] Sruthy Skaria, Akram Al-Hourani, Margaret Lech, and Robin J Evans. Hand-Gesture Recognition Using Two-Antenna Doppler Radar With Deep Convolutional Neural Networks. *Sensors Journal*, 19(8):3041–3048, 2019. 2

[37] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Differential Siamese Network. In *Proc. of the CVF/IEEE ICCV*, pages 773–782, 2019. 1

[38] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-Supervised 3D Hand Pose Estimation from Monocular RGB via Contrastive Learning. In *Proc. of the CVF/IEEE ICCV*, pages 11230–11239, 2021. 1, 2, 6

[39] Spurr, Adrian and Song, Jie and Park, Seonwook and Hilliges, Otmar. Cross-Modal Deep Variational Hand Pose Estimation. In *Proc. of the IEEE/CVF CVPR*, pages 89–98, 2018. 1, 6

[40] Jian Sun, Weiwei Zhang, Xiaoou Tang, and Heung-Yeung Shum. Background Cut. In *Proc. of ECCV*, pages 628–641, 2006. 1

[41] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. Deep Complex Networks. In *Proc. of ICLR*, pages 1–19, 2018. 4

[42] Raghav H Venkatnarayan, Griffin Page, and Muhammad Shahzad. Multi-user Gesture Recognition using WiFi. In *Proc. of the 16th ACM MobiSys*, pages 401–413, 2018. 1

[43] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-WiFi: Fine-grained Person Perception using WiFi. In *Proc. of the CVF/IEEE ICCV*, pages 5452–5461, 2019. 1, 2, 6

[44] Lin Wang and Kuk-Jin Yoon. Knowledge Distillation and Student-teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[45] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang. Orthogonal Deep Features Decomposition for Age-Invariant Face Recognition. In *Proc. of ECCV*, pages 738–753, 2018. 1

[46] Ke Xu, Xin Yang, Baocai Yin, and Rynson W.H. Lau. Learning to Restore Low-Light Images via Decomposition-and-Enhancement. In *Proc. of the IEEE/CVF CVPR*, June 2020. 1

[47] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. RadarNet: Exploiting Radar for Robust Perception of Dynamic Objects. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proc. of ECCV*, pages 496–512, 2020. 1

[48] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel Augmented Joint Learning for Visible-infrared Recognition. In *Proc. of IEEE/CVF CVPR*, pages 13567–13576, 2021. 1

[49] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. MediaPipe Hands: On-device Real-time Hand Tracking. *arXiv preprint arXiv:2006.10214*, 2020. 1, 2, 5

[50] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning Spatiotemporal Features Using 3DCNN and Convolutional LSTM for Gesture Recognition. In *Workshops at the CVF/IEEE ICCV*, pages 3120–3128, 2017. 2

[51] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging Theory and Algorithm for Domain Adaptation. In *Proc. of ICML*, pages 7404–7413, 2019. 4

[52] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall Human Pose Estimation using Radio Signals. In *Proc. of the IEEE/CVF CVPR*, pages 7356–7365, 2018. 1, 2, 3, 6

[53] Yiqin Zhao and Tian Guo. Pointar: Efficient lighting estimation for mobile augmented reality. In *Proc. of ECCV*, pages 678–693, 2020. 1

[54] Tianyue Zheng, Chao Cai, Zhe Chen, and Jun Luo. Sound of Motion: Real-time Wrist Tracking with A Smart Watch-Phone Pair. In *Proc. of the 41st IEEE INFOCOM*, pages 110–119, 2022. 1

[55] Tianyue Zheng, Zhe Chen, Jun Luo, Lin Ke, Chaoyang Zhao, and Yaowen Yang. SiWa: See into Walls via Deep UWB Radar. In *Proc. of the 27th ACM MobiCom*, page 323–336, 2021. 1

[56] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort Cross-domain Gesture Recognition with Wi-Fi. In *Proc. of the 17th ACM MobiSys*, pages 313–325, 2019. 1

[57] Yousong Zhu, Chaoyang Zhao, Jinqiao Wang, Xu Zhao, Yi Wu, and Hanqing Lu. Couplenet: Coupling Global Structure with Local Parts for Object Detection. In *Proc. of the CVF/IEEE ICCV*, pages 4126–4134, 2017. 1