Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

RFPose-OT: RF-based 3D human pose estimation via optimal transport theory*

Cong YU†1, Dongheng ZHANG2, Zhi WU2, Zhi LU2, Chunyang XIE1, Yang HU3, Yan CHEN†‡2

[1]*School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*

[2]*School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China*

[3]*School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China*

†E-mail: congyu@std.uestc.edu.cn; eecyan@ustc.edu.cn

Received Nov. 7, 2022; Revision accepted June 15, 2023; Crosschecked Sept. 12, 2023

Abstract: This paper introduces a novel framework, i.e., RFPose-OT, to enable three-dimensional (3D) human pose estimation from radio frequency (RF) signals. Different from existing methods that predict human poses from RF signals at the signal level directly, we consider the structure difference between the RF signals and the human poses, propose a transformation of the RF signals to the pose domain at the feature level based on the optimal transport (OT) theory, and generate human poses from the transformed features. To evaluate RFPose-OT, we build a radio system and a multi-view camera system to acquire the RF signal data and the ground-truth human poses. The experimental results in a basic indoor environment, an occlusion indoor environment, and an outdoor environment demonstrate that RFPose-OT can predict 3D human poses with higher precision than state-of-the-art methods.

Key words: Radio frequency sensing; Human pose estimation; Optimal transport; Deep learning
https://doi.org/10.1631/FITEE.2200550 **CLC number:** TP391.4

1 Introduction

Due to the non-contact and privacy-preserving characteristics of radio signals, radio frequency (RF) based human sensing tasks have drawn increasing attention in recent years. Existing signal processing based wireless sensing works include mainly human vital sign monitoring (Conte et al., 2010; Yue et al., 2018; Zhang DH et al., 2019, 2020), gesture recognition (Niu et al., 2022), human gait authentication (Ji et al., 2021), human position tracking (Kotaru et al., 2015; Rampa et al., 2015; Zhang DH et al., 2018, 2021; Chen Y et al., 2020; Ito and Godsill, 2020), and human speed estimation (Qian et al., 2018; Zhang F et al., 2018). With the development of deep learning, some learning-based methods (Kim and Park, 2018; Chen Y et al., 2021; Li YD et al., 2021; Zhang BB et al., 2021; Qiu et al., 2022) have been proposed to handle wireless sensing tasks.

In addition to the above classic wireless sensing tasks, some researchers (Zhao et al., 2018a, 2018b; Li TH et al., 2019; Wang F et al., 2019; Jiang et al., 2020; Song et al., 2022; Wu et al., 2022; Yu et al., 2022) have explored using RF signals to perceive human movements finely based on deep learning methods, e.g., designing deep learning models to construct fine-grained human poses from radio signals. Specifically, Zhao et al. (2018b) proposed a teacher–student network model to estimate two-dimensional (2D) human poses from frequency

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 62201542 and 62172381), the National Key R&D Programmes of China (Nos. 2022YFC2503405 and 2022YFC0869800), the Fellowship of China Postdoctoral Science Foundation (No. 2022M723069), and the Fundamental Research Funds for the Central Universities, China

ORCID: Cong YU, https://orcid.org/0000-0001-6744-021X; Yan CHEN, https://orcid.org/0000-0002-3227-4562

modulated continuous wave (FMCW) signals. Wang F et al. (2019) used a U-Net model to generate 2D human pose heatmaps from Wi-Fi signals. Taking this one step further, a three-dimensional (3D) human pose estimation model based on RF signals was proposed in Zhao et al. (2018a), where the pose estimation task was regarded as a keypoint classification problem in the 3D space. Wi-Fi-based 3D pose reconstruction has been explored in Jiang et al. (2020).

Although achieving promising performance, existing RF-based human pose estimation methods transform RF signals to human poses at the signal level directly, ignoring the structure difference between RF signals and human poses; i.e., the RF signals record human activities based on the signal reflections that are processed as signal projection heatmaps, while target human poses are represented as skeletons in the real physical space based on the human visual system. Therefore, in our work, as shown in Fig. 1, we pay more attention to the feature level, propose RFPose-OT to transform RF signals to the target pose feature domain based on the optimal transport (OT) theory, and then generate pose keypoints from the transformed features. Specifically, three phases are designed to train RFPose-OT: (1) We first train a pose encoder and a keypoint predictor to obtain target human pose representations in the feature space with the supervision of ground-truth keypoints; (2) Then, an RF encoder is trained to transform RF signals to the target pose feature domain using the OT distance (defined through the OT theory) as the training loss; (3) Finally, we fine-tune the RF encoder and the keypoint predictor using ground-truth keypoints for fine-grained estimation. Note that during RFPose-OT inference, the pose encoder can be removed, and only the RF encoder and the keypoint predictor are used for 3D human pose estimation.

To evaluate RFPose-OT, we build a radio system to capture the horizontal and vertical RF signals, which are preprocessed to the RF heatmaps. The ground-truth human pose keypoints are obtained using a multi-view camera system. In both basic and occlusion indoor environments, we compare our proposed RFPose-OT with state-of-the-art methods and also conduct ablation studies. We further test RFPose-OT in an outdoor environment. The experimental results demonstrate that RFPose-OT can predict 3D human poses with high precision and outperform the alternative methods.

2 Related works

With the popularity of radio devices, recent years have witnessed more and more interest in using radio signals to deal with sensing tasks (He Y et al., 2020).

2.1 Classical wireless human sensing

Based on signal processing algorithms, some researchers (Patwari et al., 2014; Zhang DH et al., 2019) have tried to monitor human vital signs such as breathing by analyzing the heaving chest through radar or Wi-Fi signals. In an indoor environment, some researchers focused on localization (Majeed et al., 2016) and tracking issues (Zhang DH et al., 2018, 2021; Wang L et al., 2021) and others attempted to estimate human speed from radio signals (Qian et al., 2018; Zhang F et al., 2018). Furthermore, some works tried to identify humans (Zeng et al., 2016; Hsu et al., 2019) or recognize gestures (Niu et al., 2022) from radio signals based on signal processing technologies.

Meanwhile, deep learning methods have made remarkable achievements in many areas (LeCun et al., 2015; Li J, 2018; Zhang QS and Zhu, 2018;

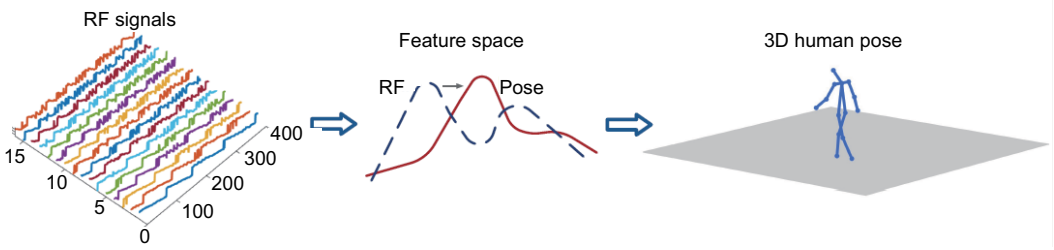

Fig. 1 RFPose-OT transforms the radio frequency (RF) signals to the pose domain to enable fine-grained three-dimensional (3D) human pose estimation

Ma et al., 2021; Yang et al., 2021; Liu et al., 2022). Therefore, in addition to using signal processing methods, more and more researchers have explored designing deep learning frameworks to push the limit of radio sensing tasks. For example, Zhao et al. (2017) constructed a conditional adversarial architecture to monitor sleep stages from radio signals via convolutional and recurrent neural networks. Xu et al. (2022) enabled human breath detection from acoustic signals in noisy driving environments by training a deep learning model. Taking this a step further, deep neural networks can learn electrocardiograms from millimeter wave (mmWave) signals (Chen JB et al., 2022). In addition to the above coarse-grained human perception, many finer-grained sensing tasks, such as human pose estimation, have been explored recently.

2.2 Human pose estimation

Before using radio signals to infer human poses, human pose estimation was a well-studied problem in computer vision literature (Wei et al., 2016; Cao et al., 2017; Fang et al., 2017; He KM et al., 2017; Martinez et al., 2017; Zheng et al., 2021). However, vision-based human pose estimation methods often suffer from occlusion or bad illumination, whereas radio signals can traverse occlusion and do not rely on lights. Hence, radio-based human pose estimation has a wider range of application scenarios and has drawn increasing attention. For example, Zhao et al. (2018b) designed a teacher–student network model to estimate 2D human poses from FMCW signals with the supervision of a vision-based human pose estimation model, and further extended the 2D version to the 3D version to achieve 3D human pose construction in Zhao et al. (2018a). After that, Li TH et al. (2019) used FMCW signals to recognize human actions based on the human pose estimation results. Efforts have been made to use Wi-Fi signals to predict human poses, including 2D (Wang F et al., 2019) and 3D (Jiang et al., 2020) human pose estimation. RF-based human pose segmentation (Wu et al., 2022) and visual synthesis (Yu et al., 2022) have been explored recently.

However, the above methods usually follow the technologies in computer vision literature to predict human poses from radio signals directly, ignoring the structural difference between radio signals and human poses. Thus, we propose to first transform radio signals to the pose feature domain based on the OT theory (Monge, 1781; Kantorovich, 1942), and then predict human poses.

The OT-based pose estimation method has been explored in Zhou et al. (2020), as an image-based human pose estimation work, where the human pose in the optical image has the same structure as the target human pose skeleton. In contrast, human pose information embedded in the RF signals is obscure and is divided into horizontal and vertical planes, with a totally different structure from the target human pose skeleton. Thus, RF-based 3D human pose estimation is a cross-domain problem, and thus is a much more challenging task than the image-based one, and the OT-based method is much more suitable for this cross-domain task.

3 Primer of optimal transport theory

In this study, we use the OT distance to train our model, which is defined through the OT theory. OT theory discusses the problem of how to transport one distribution to another with the lowest cost, where the transport map is defined as follows:

Definition 1 Map $T:\Omega\to\Psi$ transports measure $\mu\in\mathcal{P}(\Omega)$ to measure $\nu\in\mathcal{P}(\Psi)$, and we call T a transport map, if for all ν-measurable sets B, Eq. (1) holds:

$$\nu(B)=\mu(T^{-1}(B)). \tag{1}$$

Fig. 2 visualizes the transport map, where $A=\{z|z\in\Omega,T(z)\in B\}$, so $\mu(A)=\nu(B)$. Further, we define a cost function $C:\Omega\times\Psi\to[0,+\infty]$ to indicate the transportation cost, and the overall transportation cost from μ to ν can be expressed using the Monge formulation (Monge, 1781) as

$$\mathcal{M}(T)=\int_{\Omega}C(z,T(z))\mathrm{d}\mu(z). \tag{2}$$

OT theory finds a transport map $T^{\dagger}$ to determine the minimum transportation cost $\mathcal{M}(T)$. Assume that there exists a transport map $T^{\dagger}$. Then

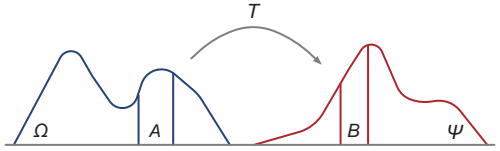

Fig. 2 Transport map

the minimum cost $\mathcal{M}(T^\dagger)$ can be defined as the OT distance between μ and ν in the geometric space. If $\mathcal{M}(T^\dagger)=0$, we think that Ω and Ψ have the same distribution.

The generalization of the Monge formulation is the Kantorovich formulation (Kantorovich, 1942):

$$\mathcal{K}(\gamma)=\int_{\Omega\times\Psi}C(z,\hat{z})\mathrm{d}\gamma(z,\hat{z}), \tag{3}$$

where $z\in\Omega$ and $\hat{z}\in\Psi$, and γ denotes the transport map from Ω to Ψ, which is subject to

$$\begin{cases}\gamma(A\times\Psi)=\mu(A),\\ \gamma(\Omega\times B)=\nu(B),\end{cases} \tag{4}$$

for all measurable sets $A\subseteq\Omega$ and $B\subseteq\Psi$. Then, assuming that the optimal transport map $\gamma^\dagger$ exists, $\mathcal{K}(\gamma^\dagger)$ is the OT distance.

4 RF signal collection and preprocessing

We use mmWave radar with multiple-input multiple-out (MIMO) antenna arrays to transceive RF signals. The received signals can be expressed as

$$\boldsymbol{s}_{m,n}(t)=\sum_l a_l(t)\boldsymbol{\psi}_{l,m,n}(t)\boldsymbol{\phi}_{l,m,n}(t), \tag{5}$$

where m, n, l, and t denote the indices of the receiver antenna, frequency point, signal propagation path, and time, respectively, $a_l(t)$ is the complex attenuation coefficient, and $\boldsymbol{\psi}_{l,m,n}(t)$ and $\boldsymbol{\phi}_{l,m,n}(t)$ are the phase shifts, which can be expressed as

$$\boldsymbol{\psi}_{l,m,n}(t)=\mathrm{e}^{-\mathrm{j}2\pi f_n\frac{(m-1)d\cos\boldsymbol{\theta}_{l,m}(t)}{c}}, \tag{6}$$

$$\boldsymbol{\phi}_{l,m,n}(t)=\mathrm{e}^{-\mathrm{j}2\pi f_n\boldsymbol{\tau}_{l,m}(t)}, \tag{7}$$

where $\boldsymbol{\theta}_{l,m}(t)$ denotes the angle of arrival (AoA), $\boldsymbol{\tau}_{l,m}(t)$ denotes the time of flight (ToF), d is the inter-element distance of the antenna array, and c is the signal propagation speed.

Human poses are often defined in the rectangular coordinate system. Thus, as shown in Fig. 3, we transform the received signals to the spatial domain:

$$\boldsymbol{S}_{x,y}(t)=\sum_m\sum_n\boldsymbol{s}_{m,n}(t)\mathrm{e}^{\mathrm{j}2\pi\varphi(x,y,m)}, \tag{8}$$

where $\mathrm{e}^{\mathrm{j}2\pi\varphi(x,y,m)}$ is the phase shift determined by the spatial positions of the signal and the antenna. $\boldsymbol{S}$ is the RF heatmap that represents the signals in the rectangular coordinate system. According to the analysis in Zhang DH et al. (2021), there exists a static multipath in the raw RF heatmap. Therefore, we further apply the differential operation on $\boldsymbol{S}_{x,y}(t)$ along time t to remove the static multipath.

In our work, we use two mmWave radars to obtain RF heatmaps on the horizontal and vertical planes, separately, which are denoted as $\boldsymbol{S}_\mathrm{H}$ and $\boldsymbol{S}_\mathrm{V}$ in the following sections.

5 RFPose-OT

We propose a novel framework, i.e., RFPose-OT, to enable 3D human pose estimation based on RF heatmaps $\boldsymbol{S}_\mathrm{H}$ and $\boldsymbol{S}_\mathrm{V}$. In the following, we first introduce the problem setup and explain the motivation for RFPose-OT. Then we discuss the network structures and the model training of RFPose-OT.

5.1 Problem setup

RFPose-OT aims to predict 3D human poses from the horizontal and vertical RF heatmaps. Because human poses can be constructed using some

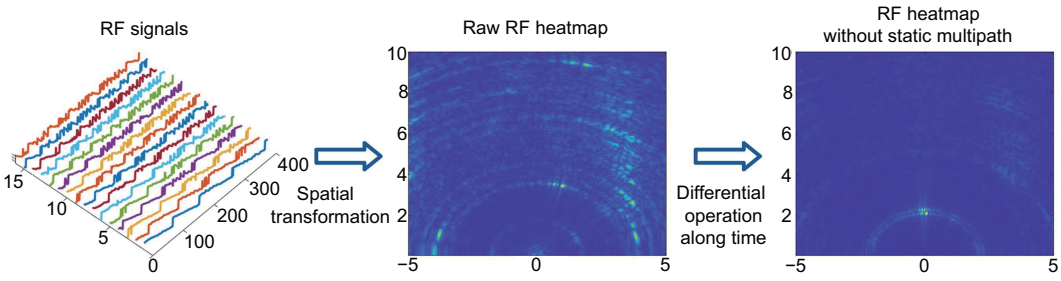

Fig. 3 Radio frequency (RF) signal preprocessing

We first transform the collected RF signals to the spatial domain to obtain RF heatmaps, and then apply the differential operation along time to remove the static multipath

body keypoints, the objective of RFPose-OT is to generate the 3D coordinates of the keypoints:

$$\hat{\boldsymbol{p}}_{K\times 3} = \text{RFPose-OT}(\boldsymbol{S}_{\text{H}}, \boldsymbol{S}_{\text{V}}), \tag{9}$$

where $\hat{\boldsymbol{p}}_{K\times 3}$ denotes the estimated keypoint coordinates, K means the number of body keypoints, and 3 means the dimension of the 3D space.

Although RF signals have been transformed to the spatial domain, from Fig. 4, we can see that $\boldsymbol{S}_{\text{H}}$ and $\boldsymbol{S}_{\text{V}}$ still have a totally different representation of the human pose compared with the human pose skeleton which is based on the human visual system; i.e., the RF heatmaps and the human poses belong to two different feature domains, and predicting keypoints from RF heatmaps directly may be difficult.

To tackle the above limitation, RFPose-OT tries to encode the RF heatmaps to the target pose domain. Specifically, as shown in Fig. 4, we first learn the human pose embedding based on ground-truth human poses, and then train an RF encoder to transform the RF heatmaps to the human pose embedded space using the OT distance as the training loss. Finally, the RF encoder and the keypoint predictor are fine-tuned to generate target human pose keypoints. Note that once trained, the pose encoder can be removed and only the RF encoder and the keypoint predictor are needed in the inference phase.

5.2 Pose embedding

To obtain a representation of the human pose in the pose domain, as shown in Fig. 4, we draw a 3D pose heatmap based on the ground-truth pose keypoint coordinates. Specifically, for each keypoint, we synthesize a 3D point heatmap using the Gaussian kernel function as follows:

$$P_{x,y,z}^{(k)} = \text{e}^{-\frac{(x-x_{\text{p}k})^2+(y-y_{\text{p}k})^2+(z-z_{\text{p}k})^2}{2\sigma^2}}, \tag{10}$$

where k denotes the index of the keypoint, and $x_{\text{p}k}$, $y_{\text{p}k}$, and $z_{\text{p}k}$ denote the coordinates of the k^{th} keypoint of the pose on the x, y, and z axes, respectively. After that, we combine all 3D point heatmaps to obtain the 3D pose heatmap:

$$P_{x,y,z} = \sum_{k=1}^{K} P_{x,y,z}^{(k)}. \tag{11}$$

$\boldsymbol{P}$ represents the human pose in the 3D physical space. Then, a pose encoder is followed to map the 3D pose heatmap $\boldsymbol{P}$ to the pose feature vector $\boldsymbol{Z}_{\text{p}}$ in the pose domain:

$$\boldsymbol{Z}_{\text{p}} = E_{\text{P}}(\boldsymbol{P}), \tag{12}$$

where E_{P} denotes the pose encoder, and $\boldsymbol{Z}_{\text{p}}$ is a manifold embedding code that contains the keypoint

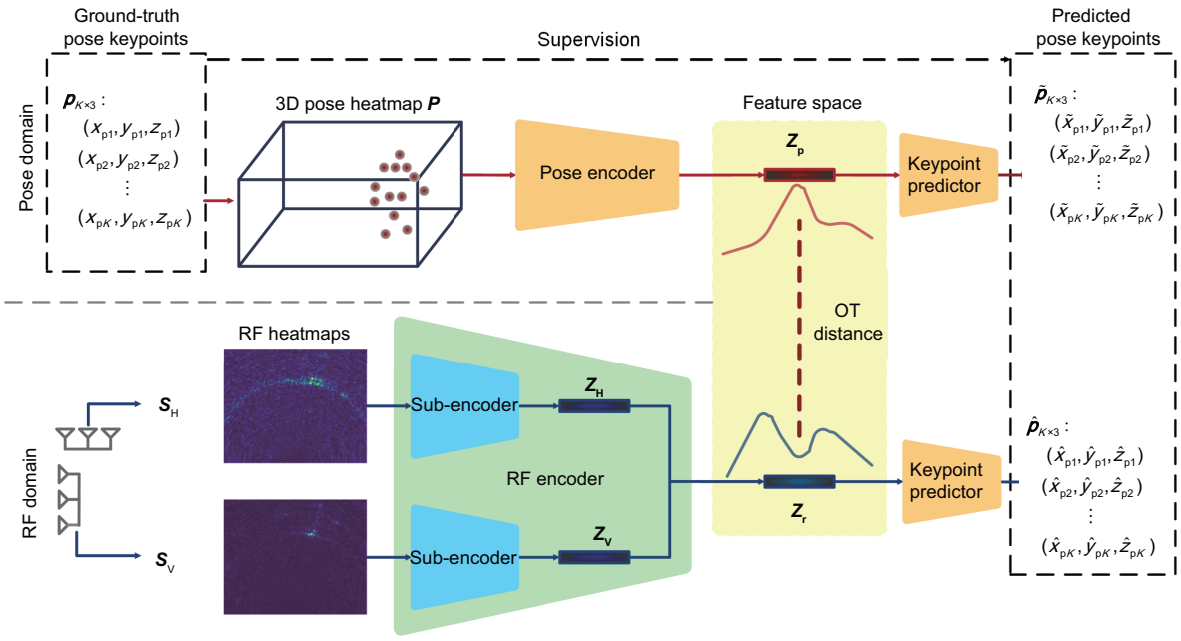

Fig. 4 The RFPose-OT architecture which consists of a pose encoder, a radio frequency (RF) encoder, and a keypoint predictor. Once trained, only the RF encoder and the keypoint predictor are retained to predict 3D human poses from RF heatmaps

location information and the spatial relationship between the keypoints. Finally, a keypoint predictor is designed to transform the pose feature vector $\boldsymbol{Z}_{\mathrm{p}}$ to the pose keypoint coordinates:

$$\tilde{\boldsymbol{p}}_{K\times 3}=F(\boldsymbol{Z}_{\mathrm{p}}). \tag{13}$$

Obviously, the predicted keypoint coordinates $\tilde{\boldsymbol{p}}_{K\times 3}$ from $\boldsymbol{P}$ should be the same as the ground-truth pose keypoint coordinates. Hence, we train the pose encoder and the keypoint predictor using the following two measurements:

$$\mathcal{L}_{\mathrm{P}}=\left\|\tilde{\boldsymbol{p}}_{K\times 3}-\boldsymbol{p}_{K\times 3}\right\|_2, \tag{14}$$

$$\mathcal{L}_{\mathrm{PO}}=\left\|\left(\tilde{\boldsymbol{p}}_{K\times 3}-\frac{1}{K}\sum_{k}^{K}\tilde{\boldsymbol{p}}_k\right)-\left(\boldsymbol{p}_{K\times 3}-\frac{1}{K}\sum_{k}^{K}\boldsymbol{p}_k\right)\right\|_2, \tag{15}$$

where $\tilde{\boldsymbol{p}}_k$ denotes the k^{th} predicted keypoint coordinates and $\boldsymbol{p}_k$ denotes the k^{th} ground-truth keypoint coordinates. Thus, $\frac{1}{K}\sum_k^K\tilde{\boldsymbol{p}}_k$ and $\frac{1}{K}\sum_k^K\boldsymbol{p}_k$ are the center point coordinates of the predicted pose and the ground-truth pose, respectively. Therefore, $\mathcal{L}_{\mathrm{P}}$ measures the absolute location error of pose keypoints, and $\mathcal{L}_{\mathrm{PO}}$ pays more attention to the relative human poses regardless of the absolute locations.

5.3 Transporting radio frequency to pose

After pose embedding, with the parameters of the pose encoder and the keypoint predictor being fixed, we design an RF encoder to map the RF heatmaps to the RF feature vector. Specifically, as shown in Fig. 4, the RF encoder contains two sub-encoders to encode the horizontal and vertical RF heatmaps separately. Then the extracted RF representations $\boldsymbol{Z}_{\mathrm{H}}$ and $\boldsymbol{Z}_{\mathrm{V}}$ are fused and mapped to the RF feature vector $\boldsymbol{Z}_{\mathrm{r}}$:

$$\boldsymbol{Z}_{\mathrm{r}}=E_{\mathrm{R}}(\boldsymbol{S}_{\mathrm{H}},\boldsymbol{S}_{\mathrm{V}}), \tag{16}$$

where E_{R} denotes the RF encoder, and $\boldsymbol{Z}_{\mathrm{r}}$ has the same dimension as the pose feature vector $\boldsymbol{Z}_{\mathrm{p}}$.

Considering the geometry of the feature spaces, i.e., the RF feature space $\mathbb{Z}_{\mathrm{r}}$ and the pose feature space $\mathbb{Z}_{\mathrm{p}}$, we use the OT distance (recall the primer in Section 3) to assess the divergence between them. Specifically, assume that μ_{r} and μ_{p} are probability measures on spaces $\mathbb{Z}_{\mathrm{r}}$ and $\mathbb{Z}_{\mathrm{p}}$ respectively, and let $C:\mathbb{Z}_{\mathrm{r}}\times\mathbb{Z}_{\mathrm{p}}\to[0,+\infty]$ be a cost function where $C(\boldsymbol{Z}_{\mathrm{r}},\boldsymbol{Z}_{\mathrm{p}})$ measures the cost of transporting one unit of mass from $\boldsymbol{Z}_{\mathrm{r}}\in\mathbb{Z}_{\mathrm{r}}$ to $\boldsymbol{Z}_{\mathrm{p}}\in\mathbb{Z}_{\mathrm{p}}$, based on Kantorovich's OT theory. The OT distance can be expressed as

$$\mathcal{L}_{\mathrm{OT}}=\int_{\mathbb{Z}_{\mathrm{r}}\times\mathbb{Z}_{\mathrm{p}}}C(\boldsymbol{Z}_{\mathrm{r}},\boldsymbol{Z}_{\mathrm{p}})\mathrm{d}\gamma(\boldsymbol{Z}_{\mathrm{r}},\boldsymbol{Z}_{\mathrm{p}}), \tag{17}$$

where $\gamma\in\mathcal{P}(\mathbb{Z}_{\mathrm{r}},\mathbb{Z}_{\mathrm{p}})$ denotes the optimal transport map, indicating the amount of mass transported from $\boldsymbol{Z}_{\mathrm{r}}$ to $\boldsymbol{Z}_{\mathrm{p}}$ and satisfying the following marginal constraints:

$$\begin{cases}\gamma(\mathbb{Z}_{\mathrm{r}}^i\times\mathbb{Z}_{\mathrm{p}})=\mu_{\mathrm{r}}(\mathbb{Z}_{\mathrm{r}}^i),\\ \gamma(\mathbb{Z}_{\mathrm{r}}\times\mathbb{Z}_{\mathrm{p}}^i)=\mu_p(\mathbb{Z}_{\mathrm{p}}^i),\end{cases} \tag{18}$$

for all measurable sets $\mathbb{Z}_{\mathrm{r}}^i\subseteq\mathbb{Z}_{\mathrm{r}}$ and $\mathbb{Z}_{\mathrm{p}}^i\subseteq\mathbb{Z}_{\mathrm{p}}$. The cost function $C(\boldsymbol{Z}_{\mathrm{r}},\boldsymbol{Z}_{\mathrm{p}})$ is set as $\|\boldsymbol{Z}_{\mathrm{r}}-\boldsymbol{Z}_{\mathrm{p}}\|_1$, so the optimal $\gamma(\boldsymbol{Z}_{\mathrm{r}},\boldsymbol{Z}_{\mathrm{p}})$ can be computed using the method in Bonneel et al. (2011). Ideally, we hope that the RF feature space $\mathbb{Z}_{\mathrm{r}}$ has the same distribution as the pose feature space $\mathbb{Z}_{\mathrm{p}}$ after RF encoder training; i.e., the OT distance is supposed to be 0. Hence, we use $\mathcal{L}_{\mathrm{OT}}$ as the objective function to train the RF encoder.

5.4 Fine-tuning

After transporting the RF domain to the pose domain, we can estimate the pose keypoint coordinates from $\boldsymbol{Z}_{\mathrm{r}}$ using the keypoint predictor:

$$\hat{\boldsymbol{p}}_{K\times 3}=F(\boldsymbol{Z}_{\mathrm{r}}). \tag{19}$$

For better human pose estimation, we further fine-tune the RF encoder and the keypoint predictor. The objective functions are similar to Eqs. (14) and (15), where $\tilde{\boldsymbol{p}}_{K\times 3}$ is replaced by $\hat{\boldsymbol{p}}_{K\times 3}$:

$$\mathcal{L}_{\mathrm{R}}=\left\|\hat{\boldsymbol{p}}_{K\times 3}-\boldsymbol{p}_{K\times 3}\right\|_2, \tag{20}$$

$$\mathcal{L}_{\mathrm{RO}}=\left\|\left(\hat{\boldsymbol{p}}_{K\times 3}-\frac{1}{K}\sum_{k}^{K}\hat{\boldsymbol{p}}_k\right)-\left(\boldsymbol{p}_{K\times 3}-\frac{1}{K}\sum_{k}^{K}\boldsymbol{p}_k\right)\right\|_2. \tag{21}$$

The whole training and fine-tuning procedure is described in Algorithm 1.

5.5 Inference setting

After training, the 3D pose heatmap and its embedding $\boldsymbol{Z}_{\mathrm{p}}$, as well as the corresponding pose encoder, can be removed, and only the trained RF encoder is needed to estimate $\boldsymbol{Z}_{\mathrm{r}}$ from the RF heatmaps. Because the RF encoder has been trained,

Algorithm 1 Training and fine-tuning algorithm of RFPose-OT

Set: batch size b and learning rate η

Initialize: pose encoder parameters Φ_{E_P}, RF encoder parameters Φ_{E_R}, and keypoint predictor parameters Φ_F

while Φ_{E_P} and Φ_F have not converged **do**
  Sample a batch of $\{\boldsymbol{p}_{K\times 3}\}$
  Update Φ_{E_P} and Φ_F using Adam with
  $\Phi_{E_\mathrm{P}} \leftarrow \Phi_{E_\mathrm{P}} - \eta\frac{1}{b}\nabla_{\Phi_{E_\mathrm{P}}}\sum_{i=1}^{b}(\mathcal{L}_\mathrm{P}+\mathcal{L}_\mathrm{PO})$
  $\Phi_\mathrm{F} \leftarrow \Phi_\mathrm{F} - \eta\frac{1}{b}\nabla_{\Phi_\mathrm{F}}\sum_{i=1}^{b}(\mathcal{L}_\mathrm{P}+\mathcal{L}_\mathrm{PO})$
end while
while Φ_{E_R} has not converged **do**
  Sample a batch of $\{\boldsymbol{S}_\mathrm{H}, \boldsymbol{S}_\mathrm{V}, \boldsymbol{Z}_\mathrm{p}\}$
  Update Φ_{E_R} using Adam with
  $\Phi_{E_\mathrm{R}} \leftarrow \Phi_{E_\mathrm{R}} - \eta\nabla_{\Phi_{E_\mathrm{R}}}\mathcal{L}_\mathrm{OT}$
end while
while Φ_{E_R} and Φ_F have not converged **do**
  Sample a batch of $\{\boldsymbol{S}_\mathrm{H}, \boldsymbol{S}_\mathrm{V}, \boldsymbol{p}_{K\times 3}\}$
  Update Φ_{E_R} and Φ_F using Adam with
  $\Phi_{E_\mathrm{R}} \leftarrow \Phi_{E_\mathrm{R}} - \eta\frac{1}{b}\nabla_{\Phi_{E_\mathrm{R}}}\sum_{i=1}^{b}(\mathcal{L}_\mathrm{R}+\mathcal{L}_\mathrm{RO})$
  $\Phi_\mathrm{F} \leftarrow \Phi_\mathrm{F} - \eta\frac{1}{b}\nabla_{\Phi_\mathrm{F}}\sum_{i=1}^{b}(\mathcal{L}_\mathrm{R}+\mathcal{L}_\mathrm{RO})$
end while

it can directly map the RF heatmaps to the pose domain, and then the output $\boldsymbol{Z}_\mathrm{r}$ is inputted into the keypoint predictor to estimate the pose keypoint coordinates.

6 Experiments

In this section, we describe first the implementation details and then the experiments conducted to evaluate the performance of the proposed RFPose-OT.

6.1 Dataset

To collect RF signal data, we built a radio system using two mmWave radars (horizontal and vertical), where each radar was equipped with 12 transmitters and 16 receivers with a MIMO antenna array. To avoid mutual interference, one radar worked at 77 GHz while the other worked at 79 GHz, both with a 1.23 GHz bandwidth. To obtain the ground-truth human poses, we built a multi-view camera system with 13 camera nodes, and a calibration method (Zhang Z, 2000) was applied to obtain the ground-truth 3D pose keypoint coordinates.

During data collection, we captured the RF signal reflections at 20 Hz, and the camera system recorded videos at 10 frames per second. The radio system and the camera system were synchronized using the network time protocol (NTP) through transmission control protocol (TCP) connection, which achieved a millisecond-level synchronization error. We collected the data for 10 indoor scenes. In total, we collected 89 090 RF signal samples and obtained the corresponding 3D pose keypoint coordinates, 80% of which were used for training and the rest for testing.

Because RF signals can traverse occlusions, we collected more data in the occlusion environment as the additional testing set, where the radio system was occluded by baffles. Then 1180 RF signal samples were collected and the corresponding ground-truth poses were obtained by the camera system.

6.2 Network structure

RFPose-OT includes a pose encoder, an RF encoder, and a keypoint predictor.

The pose encoder consists of five convolutional layers and one linear layer. The RF encoder contains two sub-encoders with the same network structure, and each sub-encoder consists of six convolutional layers. The keypoint predictor consists of four linear layers. Details of these network structures are shown in Tables 1–3, where conv denotes the convolution layer, linear denotes the linear layer, BN denotes batch normalization, LN denotes layer normalization, ReLU denotes the rectified linear unit, ks denotes the kernel size, cs denotes the number of channels, and st denotes the stride.

6.3 Training details

RFPose-OT was trained using the Adam solver. The numbers of epochs for pose embedding training and RF transporting training were both 100, and the number of epochs for fine-tuning was set to 50. The initial learning rate was set to 0.002 and decayed by half every 10 epochs for all training phases. The batch size during model training was always 64. We implemented our proposed RFPose-OT using PyTorch and all experiments can be run on a commodity workstation with a single GTX-1080 graphics card.

Table 1 Pose encoder

Architecture	Parameter(s)
Conv-BN-ReLU	ks = 5 × 5, cs = 32, st = 2 ks = 5 × 5, cs = 64, st = 2 ks = 5 × 5, cs = 64, st = 1 ks = 5 × 5, cs = 128, st = 2 ks = 5 × 5, cs = 128, st = 1
Linear	256

Table 2 Sub-encoder in the RF encoder

Architecture	Parameters
Conv-BN-ReLU	ks = 3 × 3, cs = 1, st = 1 ks = 5 × 5, cs = 8, st = 2 ks = 5 × 5, cs = 32, st = 2 ks = 5 × 5, cs = 128, st = 2 ks = 5 × 5, cs = 128, st = 1
Conv-BN	ks = 5 × 5, cs = 512, st = 2

Table 3 Keypoint predictor

Architecture	Parameter
Linear-LN-ReLU	cs =256 cs =256 cs =256
Linear	cs =42

6.4 Metric

To assess the precision of human pose estimation, we calculated the spatial location error (SLE) between the predicted keypoints and the corresponding ground-truth keypoints using the Euclidean distance:

$$\mathrm{SLE}_k = \frac{1}{U}\sum_{u=1}^{U}\left\|\hat{\boldsymbol{p}}_k^{(u)} - \boldsymbol{p}_k^{(u)}\right\|_2, \tag{22}$$

where k denotes the keypoint index, and U is the number of test samples.

6.5 Comparison with baselines

We compared our proposed RFPose-OT with RF-Pose3D (Zhao et al., 2018a) and mm-Pose (Sengupta et al., 2020).

1. RF-Pose3D: RF-Pose3D is a classic deep learning model for 3D human pose estimation from RF signals. It regards pose estimation as a keypoint classification problem and is constructed using 18 convolution layers.

2. mm-Pose: mm-Pose is a real-time 3D human pose estimation model based on mmWave radars. It consists of six convolution layers and four linear layers, and predicts the coordinates of human pose keypoints from the horizontal and vertical RF signals directly.

The quantitative comparison results are summarized in Table 4 in the basic environment, from which we can find that RFPose-OT outperformed baseline methods in almost all pose keypoints, and achieved much higher estimation precision in small body parts, e.g., wrists and ankles, which means that the pose perception of our proposed RFPose-OT is much more fine-grained. We also show the qualitative results in Fig. 5a, from which we can see that although RF-Pose3D and mm-Pose can predict correct human locations, inaccurate poses were generated, whereas RFPose-OT can generate target 3D human poses that were consistent with ground truth.

Compared with the vision-based pose estimation method, the RF-based pose estimation model can work in the occlusion environment. Thus, we further tested RFPose-OT and the baseline methods using an additional testing set that was collected in the occlusion environment. The quantitative and qualitative results are shown in Table 4 and Fig. 5b, from which we can see that RFPose-OT can still estimate 3D human poses from RF signals with high precision and outperformed alternative methods. The above experimental results demonstrated the effectiveness of our proposed RFPose-OT model in the occlusion environment.

6.6 Ablation studies

In this subsection, we report our ablation experiments to discuss the effects of some components in the RFPose-OT model.

1. RFPose: In our full RFPose-OT model, an OT loss $\mathcal{L}_{\mathrm{OT}}$ was designed for training the RF encoder to enable the transformation from the RF domain to the pose domain. The ablation native model RFPose removes this component; i.e., RFPose predicts pose keypoints from the RF heatmaps directly.

2. RFPose-L2: RFPose-L2 uses the L2 distance as the loss function to train the RF encoder to minimize the difference between the RF feature vector and the pose feature vector.

3. RFPose-OT without $\mathcal{L}_{\mathrm{PO+RO}}$: In the pose embedding and fine-tuning phases, $\mathcal{L}_{\mathrm{PO}}$ and $\mathcal{L}_{\mathrm{RO}}$ are proposed to pay more attention to the relative human poses. In RFPose-OT without $\mathcal{L}_{\mathrm{PO+RO}}$, we discuss the effects of $\mathcal{L}_{\mathrm{PO}}$ and $\mathcal{L}_{\mathrm{RO}}$.

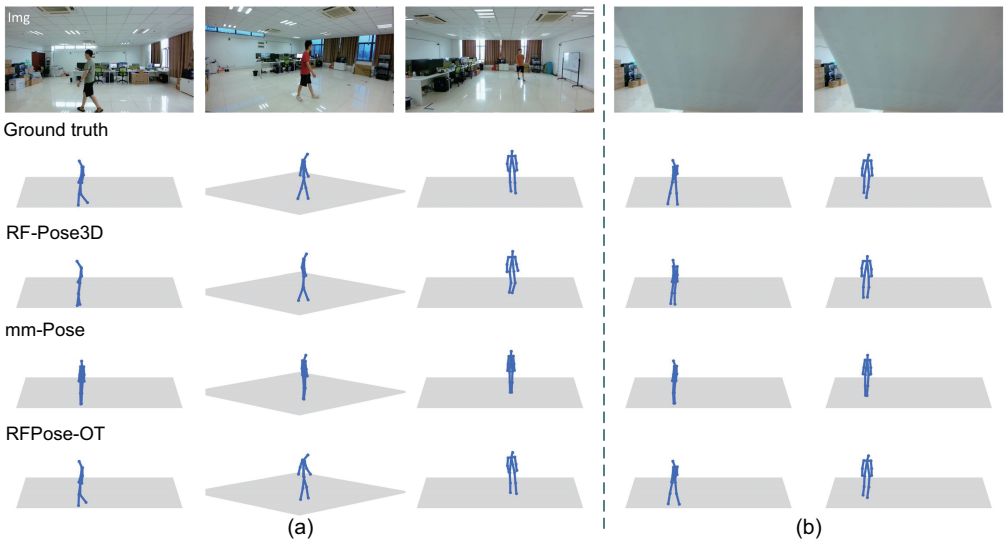

Fig. 5 Qualitative results of different methods in the basic (a) and occlusion (b) indoor environments

The 1st row shows the images captured by a camera that is attached to the radio system, the 2nd row shows the ground-truth 3D human poses, and the 3rd, 4th, and 5th rows show the 3D human poses estimated by RF-Pose3D (Zhao et al., 2018a), mm-Pose (Sengupta et al., 2020), and our proposed RFPose-OT, respectively

Table 4 Quantitative evaluation results of different methods in the basic *a* and occlusion *b* indoor environments

Env	Method	SLE (cm)								
		Nose	Neck	Shoulders	Elbows	Wrists	Hips	Knees	Ankles	Overall
a	RF-Pose3D (Zhao et al., 2018a)	8.11	**5.21**	7.57	9.92	15.74	6.64	11.31	21.10	11.27
	mm-Pose (Sengupta et al., 2020)	8.19	5.30	7.23	9.67	15.29	**6.20**	10.83	19.04	10.72
	RFPose-OT	**7.90**	6.14	**6.76**	**7.99**	**11.67**	6.39	**8.34**	**12.60**	**8.68**
b	RF-Pose3D (Zhao et al., 2018a)	**6.53**	4.86	6.65	8.75	14.05	6.95	11.26	21.52	10.70
	mm-Pose (Sengupta et al., 2020)	6.64	**3.88**	**6.34**	9.16	14.84	6.98	11.28	19.28	10.45
	RFPose-OT	7.85	6.42	6.78	**7.90**	**11.41**	**6.82**	**9.35**	**14.05**	**9.07**

Bold numbers represent the minimum spatial location error (SLE) values. Env: environment

Table 5 Quantitative evaluation results of different ablation models and the full model in the basic *a* and occlusion *b* indoor environments

Env	Method	SLE (cm)								
		Nose	Neck	Shoulders	Elbows	Wrists	Hips	Knees	Ankles	Overall
a	RFPose	8.69	6.82	7.58	8.97	12.93	7.19	9.31	14.11	9.69
	RFPose-L2	8.02	**5.93**	6.88	8.12	12.02	6.36	8.57	12.96	8.84
	RFPose-OT w/o $\mathcal{L}_{\text{PO+RO}}$	8.32	6.04	7.01	8.52	12.66	**6.30**	8.57	13.94	9.17
	RFPose-OT (full)	**7.90**	6.14	**6.76**	**7.99**	**11.67**	6.39	**8.34**	**12.60**	**8.68**
b	RFPose	7.87	6.64	7.35	8.69	12.10	7.57	10.11	15.27	9.76
	RFPose-L2	7.88	6.59	7.23	8.62	12.47	7.33	10.08	15.20	9.74
	RFPose-OT w/o $\mathcal{L}_{\text{PO+RO}}$	**7.83**	**6.11**	7.01	8.36	12.04	**6.61**	9.51	15.55	9.44
	RFPose-OT (full)	7.85	6.42	**6.78**	**7.90**	**11.41**	6.82	**9.35**	**14.05**	**9.07**

Bold numbers represent the minimum spatial location error (SLE) values. Env: environment; w/o: without

The quantitative evaluation results in the basic and occlusion environments are shown in Table 5, from which we can see that the OT-based model RFPose-OT and the L2-based model RFPose-L2 both outperformed the native version RFPose; i.e., transforming the RF domain to the pose domain

first and then predicting pose keypoints can improve the pose estimation precision. RFPose-OT performed better than RFPose-L2, because it can achieve higher estimation precision than RFPose-L2 in all pose keypoints in the occlusion environment. The above experimental results demonstrated that mapping the RF signals to the pose feature space can improve the performance, and that using the OT distance outperforms the use of the L2 distance, which validates the advantage of the proposed OT-based method. Furthermore, compared with RFPose-OT without $\mathcal{L}_{\mathrm{PO+RO}}$, the full RFPose-OT model performed better, which means that $\mathcal{L}_{\mathrm{PO}}$ and $\mathcal{L}_{\mathrm{RO}}$ contribute to the pose estimation.

6.7 Performance in the outdoor scene

Even though the RFPose-OT model was trained in the indoor environment, we directly applied it to a new outdoor environment to evaluate its generalization. Because it is difficult to move the multi-view camera system to the new scenario to provide ground-truth poses, here we qualitatively evaluated only performance. Fig. 6 shows the 3D poses generated by our model and the corresponding snapshots, from which we can see that our model can correctly estimate 3D human poses from RF signals in the outdoor environment. This confirms the cross-environment generalization ability of RFPose-OT.

7 Discussion

7.1 Model complexity

In this subsection, we calculated the number of parameters and the number of multiply-accumulate operations (MACs) to evaluate the complexity of our proposed RFPose-OT, and further tested RFPose-OT on a single GTX-1080 graphics card to calculate the average runtime for predicting one frame of human pose from RF signals. The results are shown in Table 6, from which we can see that RFPose-OT can support real-time processing.

Table 6 Model complexity and runtime

Number of parameters ($\times 10^3$)	Number of multiply-accumulate operations ($\times 10^6$)	Runtime (s/frame)
7.024	1.302	0.029

7.2 Trajectory tracking

RFPose-OT is a fine-grained human activity sensing framework. Obviously, it can be used for handling some classic wireless sensing tasks, e.g., tracking a moving person. In this subsection, we used the trained RFPose-OT to track a subject who was asked to walk randomly in the basic and occlusion scenes. The ground-truth trajectories were obtained by the multi-view camera system. RF signals were inputted into RFPose-OT, and we calculated the average value of the horizontal coordinates of the output keypoints as the predicted trajectories. The quantitative errors are shown in Table 7, and the qualitative results are shown in Fig. 7, from which we can see that RFPose-OT can recover the moving trajectories in both basic and occlusion indoor environments.

Table 7 Quantitative trajectory tracking errors

Environment	Tracking error (cm)	
	x axis	y axis
Basic	3.36	3.58
Occlusion	3.61	3.94

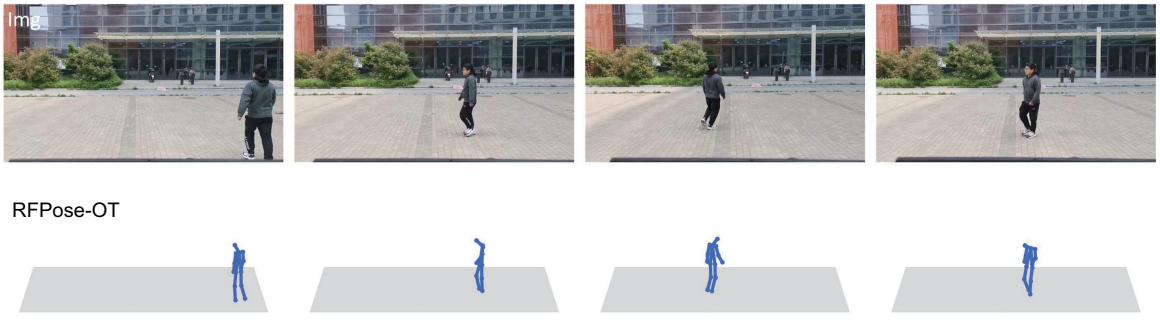

Fig. 6 Qualitative results by RFPose-OT in an outdoor environment

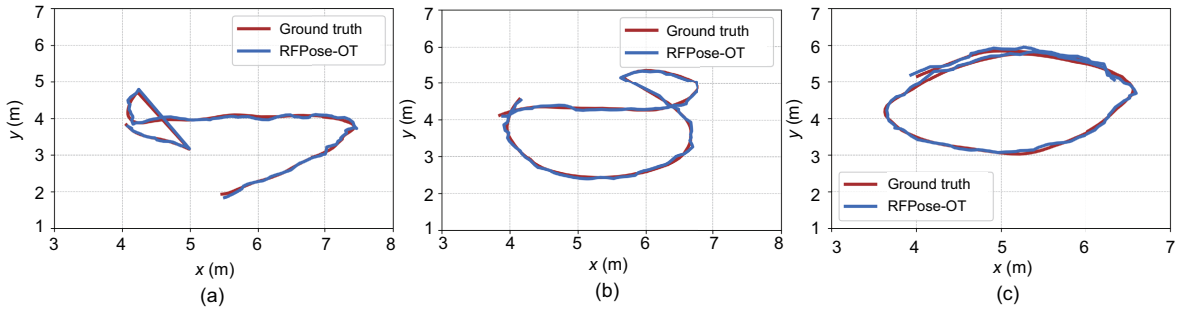

Fig. 7 Trajectories of a moving person in the basic environment (a, b) and occlusion environment (c)
Red line is the ground-truth trajectory and blue line is the predicted trajectory. References to color refer to the online version of this figure

7.3 Scope and limitations

Experimental results demonstrated the effectiveness of RFPose-OT in basic and occlusion environments. However, RFPose-OT has some limitations. On one hand, the operating distance of our radio system was limited to 20 m. Extra transmission power would be needed to cover a larger space. On the other hand, micro hand motions may be missed by RFPose-OT due to few signal power reflections.

8 Conclusions

In this paper, we proposed a novel RF-based 3D human pose estimation model, RFPose-OT, which first transports the RF domain to the target pose domain based on the OT theory, and then estimates human pose keypoints from the transported RF features. To assess RFPose-OT, we conducted experiments in both basic and occlusion indoor environments and an outdoor environment, and the experimental results demonstrated that our proposed RFPose-OT can estimate 3D human poses with higher precision. We believe that this work provides a new and valid framework to tackle RF-based human sensing tasks.

Contributors

Cong YU and Yan CHEN conceived the method. Cong YU designed and implemented the RFPose-OT model. Cong YU and Zhi LU performed the experiments and conducted the comparisons. Dongheng ZHANG, Zhi WU, and Chunyang XIE collected and processed the data. Cong YU drafted the paper. Yang HU and Yan CHEN helped organize the paper. Yan CHEN supervised all aspects of the project. All authors contributed to designing the experiments and revising and finalizing the paper.

Compliance with ethics guidelines

Cong YU, Dongheng ZHANG, Zhi WU, Zhi LU, Chunyang XIE, Yang HU, and Yan CHEN declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bonneel N, van de Panne M, Paris S, et al., 2011. Displacement interpolation using Lagrangian mass transport. Proc SIGGRAPH Asia Conf, p.1-12. https://doi.org/10.1145/2024156.2024192

Cao Z, Simon T, Wei SE, et al., 2017. Realtime multi-person 2D pose estimation using part affinity fields. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.7291-7299. https://doi.org/10.1109/CVPR.2017.143

Chen JB, Zhang DH, Wu Z, et al., 2022. Contactless electrocardiogram monitoring with millimeter wave radar. *IEEE Trans Mob Comput*, early access. https://doi.org/10.1109/TMC.2022.3214721

Chen Y, Su X, Hu Y, et al., 2020. Residual carrier frequency offset estimation and compensation for commodity WiFi. *IEEE Trans Mob Comput*, 19(12):2891-2902. https://doi.org/10.1109/TMC.2019.2934106

Chen Y, Deng HY, Zhang DH, et al., 2021. SpeedNet: indoor speed estimation with radio signals. *IEEE Int Things J*, 8(4):2762-2774. https://doi.org/10.1109/JIOT.2020.3022071

Conte E, Filippi A, Tomasin S, 2010. Ml period estimation with application to vital sign monitoring. *IEEE Signal Process Lett*, 17(11):905-908. https://doi.org/10.1109/LSP.2010.2071382

Fang HS, Xie SQ, Tai YW, et al., 2017. RMPE: regional multi-person pose estimation. Proc IEEE Int Conf on Computer Vision, p.2334-2343. https://doi.org/10.1109/ICCV.2017.256

He KM, Gkioxari G, Dollár P, et al., 2017. Mask R-CNN. Proc IEEE Int Conf on Computer Vision, p.2961-2969. https://doi.org/10.1109/ICCV.2017.322

He Y, Chen Y, Hu Y, et al., 2020. WiFi vision: sensing, recognition, and detection with commodity MIMO-OFDM WiFi. *IEEE Int Things J*, 7(9):8296-8317. https://doi.org/10.1109/JIOT.2020.2989426

Hsu CY, Hristov R, Lee GH, et al., 2019. Enabling identification and behavioral sensing in homes using radio reflections. Proc CHI Conf on Human Factors in Computing Systems, p.1-13. https://doi.org/10.1145/3290605.3300778

Ito N, Godsill S, 2020. A multi-target track-before-detect particle filter using superpositional data in non-Gaussian noise. *IEEE Signal Process Lett*, 27:1075-1079. https://doi.org/10.1109/LSP.2020.3002704

Ji HR, Hou CP, Yang Y, et al., 2021. A one-class classification method for human gait authentication using micro-Doppler signatures. *IEEE Signal Process Lett*, 28:2182-2186. https://doi.org/10.1109/LSP.2021.3122344

Jiang WJ, Xue HF, Miao CL, et al., 2020. Towards 3D human pose construction using WiFi. Proc 26th Annual Int Conf on Mobile Computing and Networking, p.1-14. https://doi.org/10.1145/3372224.3380900

Kantorovich LV, 1942. On the translocation of masses. *Dokl Akad Nauk USSR*, 37:199-201 (in Russian).

Kim HI, Park RH, 2018. Residual LSTM attention network for object tracking. *IEEE Signal Process Lett*, 25(7):1029-1033. https://doi.org/10.1109/LSP.2018.2835768

Kotaru M, Joshi K, Bharadia D, et al., 2015. SpotFi: decimeter level localization using WiFi. Proc ACM Conf on Special Interest Group on Data Communication, p.269-282. https://doi.org/10.1145/2785956.2787487

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. *Nature*, 521(7553):436-444. https://doi.org/10.1038/nature14539

Li J, 2018. Cyber security meets artificial intelligence: a survey. *Front Inform Technol Electron Eng*, 19(12):1462-1474. https://doi.org/10.1631/FITEE.1800573

Li TH, Fan LJ, Zhao MM, et al., 2019. Making the invisible visible: action recognition through walls and occlusions. Proc IEEE/CVF Int Conf on Computer Vision, p.872-881. https://doi.org/10.1109/ICCV.2019.00096

Li YD, Zhang DH, Chen JB, et al., 2021. Towards domain-independent and real-time gesture recognition using mmWave signal. *IEEE Trans Mob Comput*, early access. https://doi.org/10.1109/TMC.2022.3207570

Liu SP, Tian GH, Cui YC, et al., 2022. A deep Q-learning network based active object detection model with a novel training algorithm for service robots. *Front Inform Technol Electron Eng*, 23(11):1673-1683. https://doi.org/10.1631/FITEE.2200109

Ma L, Zhong QY, Zhang YY, et al., 2021. Associative affinity network learning for multi-object tracking. *Front Inform Technol Electron Eng*, 22(9):1194-1206. https://doi.org/10.1631/FITEE.2000272

Majeed K, Sorour S, Al-Naffouri TY, et al., 2016. Indoor localization and radio map estimation using unsupervised manifold alignment with geometry perturbation. *IEEE Trans Mob Comput*, 15(11):2794-2808. https://doi.org/10.1109/TMC.2015.2510631

Martinez J, Hossain R, Romero J, et al., 2017. A simple yet effective baseline for 3D human pose estimation. Proc IEEE Int Conf on Computer Vision, p.2640-2649. https://doi.org/10.1109/ICCV.2017.288

Monge G, 1781. Mémoire sur la théorie des déblais et des remblais. Mémoires de Mathématique et de Physique, Presentés à l'Académie Royale des Sciences, p.666-704 (in French).

Niu K, Zhang FS, Wang XZ, et al., 2022. Understanding WiFi signal frequency features for position-independent gesture sensing. *IEEE Trans Mob Comput*, 21(11):4156-4171. https://doi.org/10.1109/TMC.2021.3063135

Patwari N, Wilson J, Ananthanarayanan S, et al., 2014. Monitoring breathing via signal strength in wireless networks. *IEEE Trans Mob Comput*, 13(8):1774-1786. https://doi.org/10.1109/TMC.2013.117

Qian K, Wu CS, Yang Z, et al., 2018. Enabling contactless detection of moving humans with dynamic speeds using CSI. *ACM Trans Embed Comput Syst*, 17(2):1-18. https://doi.org/10.1145/3157677

Qiu CR, Zhang DH, Hu Y, et al., 2022. Radio-assisted human detection. *IEEE Trans Multim*, 25:2613-2623. https://doi.org/10.1109/TMM.2022.3149129

Rampa V, Savazzi S, Nicoli M, et al., 2015. Physical modeling and performance bounds for device-free localization systems. *IEEE Signal Process Lett*, 22(11):1864-1868. https://doi.org/10.1109/LSP.2015.2438176

Sengupta A, Jin F, Zhang RY, et al., 2020. mm-Pose: real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sens J*, 20(17):10032-10044. https://doi.org/10.1109/JSEN.2020.2991741

Song RY, Zhang DH, Wu Z, et al., 2022. RF-URL: unsupervised representation learning for RF sensing. Proc 28th Annual Int Conf on Mobile Computing and Networking, p.282-295. https://doi.org/10.1145/3495243.3560529

Wang F, Zhou S, Panev S, et al., 2019. Person-in-WiFi: fine-grained person perception using WiFi. Proc IEEE/CVF Int Conf on Computer Vision, p.5452-5461. https://doi.org/10.1109/ICCV.2019.00555

Wang L, Sun K, Dai HP, et al., 2021. WiTrace: centimeter-level passive gesture tracking using OFDM signals. *IEEE Trans Mob Comput*, 20(4):1730-1745. https://doi.org/10.1109/TMC.2019.2961885

Wei SE, Ramakrishna V, Kanade T, et al., 2016. Convolutional pose machines. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.4724-4732. https://doi.org/10.1109/CVPR.2016.511

Wu Z, Zhang DH, Xie CY, et al., 2022. RFMask: a simple baseline for human silhouette segmentation with radio signals. *IEEE Trans Multim*, early access. https://doi.org/10.1109/TMM.2022.3181455

Xu XY, Yu JD, Chen YY, 2022. Leveraging acoustic signals for fine-grained breathing monitoring in driving environments. *IEEE Trans Mob Comput*, 21(3):1018-1033. https://doi.org/10.1109/TMC.2020.3015828

Yang Y, Zhuang YT, Pan YH, 2021. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Front Inform Technol Electron Eng*, 22(12):1551-1558. https://doi.org/10.1631/FITEE.2100463

Yu C, Wu Z, Zhang DH, et al., 2022. RFGAN: RF-based human synthesis. *IEEE Trans Multim*, 25:2926-2938. https://doi.org/10.1109/TMM.2022.3153136

Yue SC, He H, Wang H, et al., 2018. Extracting multi-person respiration from entangled RF signals. *Proc ACM Interact Mob Wearab Ubiq Technol*, 2(2):1-22. https://doi.org/10.1145/3214289

Zeng YZ, Pathak PH, Mohapatra P, 2016. WiWho: WiFi-based person identification in smart spaces. Proc 15th ACM/IEEE Int Conf on Information Processing in Sensor Networks, p.1-12. https://doi.org/10.1109/IPSN.2016.7460727

Zhang BB, Zhang DH, Li YD, et al., 2021. Unsupervised domain adaptation for device-free gesture recognition. https://arxiv.org/abs/2111.10602v1

Zhang DH, He Y, Gong XY, et al., 2018. Multitarget AOA estimation using wideband LFMCW signal and two receiver antennas. *IEEE Trans Veh Technol*, 67(8):7101-7112. https://doi.org/10.1109/TVT.2018.2827408

Zhang DH, Hu Y, Chen Y, et al., 2019. BreathTrack: tracking indoor human breath status via commodity WiFi. *IEEE Int Things J*, 6(2):3899-3911. https://doi.org/10.1109/JIOT.2019.2893330

Zhang DH, Hu Y, Chen Y, et al., 2020. Calibrating phase offsets for commodity WiFi. *IEEE Syst J*, 14(1):661-664. https://doi.org/10.1109/JSYST.2019.2904714

Zhang DH, Hu Y, Chen Y, 2021. MTrack: tracking multi-person moving trajectories and vital signs with radio signals. *IEEE Int Things J*, 8(5):3904-3914. https://doi.org/10.1109/JIOT.2020.3025820

Zhang F, Chen C, Wang BB, et al., 2018. WiSpeed: a statistical electromagnetic approach for device-free indoor speed estimation. *IEEE Int Things J*, 5(3):2163-2177. https://doi.org/10.1109/JIOT.2018.2826227

Zhang QS, Zhu SC, 2018. Visual interpretability for deep learning: a survey. *Front Inform Technol Electron Eng*, 19(1):27-39. https://doi.org/10.1631/FITEE.1700808

Zhang Z, 2000. A flexible new technique for camera calibration. *IEEE Trans Patt Anal Mach Intell*, 22(11):1330-1334. https://doi.org/10.1109/34.888718

Zhao MM, Yue SC, Katabi D, et al., 2017. Learning sleep stages from radio signals: a conditional adversarial architecture. Proc 34th Int Conf on Machine Learning, p.4100-4109.

Zhao MM, Tian YL, Zhao H, et al., 2018a. RF-based 3D skeletons. Proc Conf of the ACM Special Interest Group on Data Communication, p.267-281. https://doi.org/10.1145/3230543.3230579

Zhao MM, Li TH, Abu Alsheikh M, et al., 2018b. Through-wall human pose estimation using radio signals. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7356-7365. https://doi.org/10.1109/CVPR.2018.00768

Zheng C, Zhu SJ, Mendieta M, et al., 2021. 3D human pose estimation with spatial and temporal transformers. Proc IEEE/CVF Int Conf on Computer Vision, p.11656-11665. https://doi.org/10.1109/ICCV48922.2021.01145

Zhou L, Chen YY, Gao YZ, et al., 2020. Occlusion-aware Siamese network for human pose estimation. Proc 16th European Conf on Computer Vision, p.396-412. https://doi.org/10.1007/978-3-030-58565-5_24