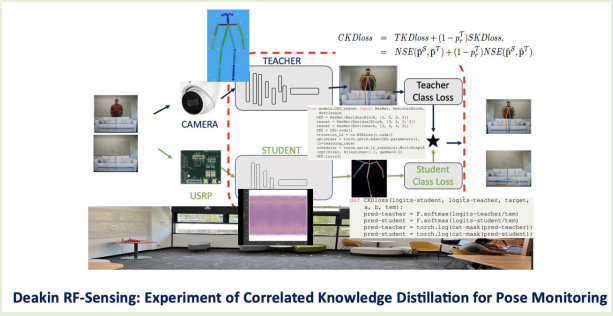
Deakin RF-Sensing: Experiments on Correlated Knowledge Distillation for Monitoring Human Postures With Radios

Shiva Raj Pokhrel, *Senior Member, IEEE*, Jonathan Kua, *Member, IEEE*, Deol Satish, Philip Williams, Arkady Zaslavsky, *Senior Member, IEEE*, Seng W. Loke, *Member, IEEE*, and Jinho Choi, *Fellow, IEEE*

***Abstract*—In this article, we study the feasibility of a novel idea by coupling radio frequency (RF) sensing technology with correlated knowledge distillation (CKD) theory toward designing lightweight, near real-time, and precise human pose monitoring systems through an in-house experimental testbed development at Deakin University. The datasets collected from the developed testbed are fed to the CKD framework, which transfers and fuses pose knowledge from a robust "Teacher" model to a parameterized "Student" model and can be a promising technique for obtaining accurate yet lightweight pose estimates. As a result, it becomes possible to identify the current pose of a human body (e.g., an elderly individual in a care home) solely using RF signals (e.g., WiFi signals in a home) without relying on visual signals or video information, meaning that this approach effectively addresses privacy concerns by allowing pose identification without compromising personal privacy. To ensure its efficacy, we implement CKD for distilling logits in our integrated software-defined radio (SDR)-based experimental setup and investigate the RF-visual signal correlation. Our CKD-RF sensing technique is characterized by two modes—a camera-fed teacher class network (e.g., images and videos) with an SDR-fed student class network (e.g., RF signals). Specifically, our CKD model trains a dual multibranch teacher–student network by distilling and fusing knowledge bases. The resulting CKD models are then subsequently used to identify the multimodal correlation and teach the student branch in reverse. Instead of simply aggregating their learnings, CKD training comprises multiple parallel transformations with the two domains, i.e., visual images and RF signals. Once trained, our CKD model efficiently preserves privacy and utilizes the multimodal correlated logits from the two different neural networks (NNs) to estimate poses solely using RF signals.**



Deakin RF-Sensing: Experiment of Correlated Knowledge Distillation for Pose Monitoring

***Index Terms*— Body radio reflections, correlation, feature extraction, fusion, knowledge distillation, machine learning (ML), motion detection, pose monitoring, radio frequency (RF), RF-based vision, RF localization, RF sensing, RF signals, sensor phenomena and characterization, software-defined radio (SDR), wireless sensor networks.**

I. Introduction

THE advent of the Internet of Things (IoT), artificial intelligence (AI)/machine learning (ML), and computer vision have significantly advanced the field of human pose estimation and have resulted in many new application domains, such as near real-time fall detection in patient/aged care homes [1], activity sensing [2], and sign language recognition [3]. Most state-of-the-art techniques rely on cameras and, more recently, AI-driven computer vision technologies to detect, identify, and react to human pose changes in public/private spaces. These systems can be reliable and provide accurate human pose estimates, but require expensive equipment and incur high computational costs. In addition, there are data privacy concerns when deploying high-definition visual spectrum cameras where the activity of human subjects is being monitored.

The use of radio frequency (RF)-sensing technologies for human pose estimation over the past decade has gained traction and has resulted in significant benefits across a wide range of settings (see [4] and references). For example, people in assisted living facilities can benefit from remote monitoring to ensure their health and safety while maintaining their independence, but privacy is invaded by the use of

Manuscript received 17 August 2023; accepted 21 September 2023. Date of publication 3 October 2023; date of current version 14 November 2023. This work was supported by the School of IT, Deakin University, through the research program, CITECORE-RPI-2023 "Next-Generation IoT Communications Infrastructure, Analytics and Intelligence." The associate editor coordinating the review of this article and approving it for publication was Dr. Avik Santra. *(Corresponding author: Shiva Raj Pokhrel.)*

The authors are with the IoT Research Laboratory, School of Information Technology, Deakin University, Geelong, VIC 3220, Australia (e-mail: shiva.pokhrel@deakin.edu.au).

Digital Object Identifier 10.1109/JSEN.2023.3320131

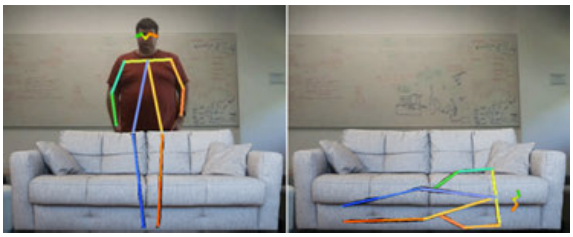


Fig. 1. Motivating example of RF-based pose estimation and tracking through occlusions. Human subjects could be hidden from plain view, as RF signals can penetrate objects.

cameras. RF-sensing could not only monitor diseases of the elderly and disabled but also has great potential for military applications (e.g., detecting sitting, standing, and lying face up or down) and emergency response such as fires and floods, among others.

Some relevant works directly apply [5], [6], [7], [8], [9], [10] ML methods for RF-sensing, which have been reported to be strongly biased due to background RF noises and fail to learn helpful information for pose estimates [2], [11]. In a different context, [3] uses ML to demonstrate the recognition of sign language with RF sensing. However, researchers in this field have poorly studied the potential of multimodal learning with knowledge fusion/distillation to improve RF detection, which is the primary focus of this article.

Fig. 1 shows a motivating example of using RF signals to detect and track human poses (the subject could be hidden from plain view, as RF signals can penetrate objects). We aim to develop a system to accurately detect and predict human activity without compromising privacy (movement tracking through occlusions) and alleviating the impacts of RF noise. With computational and wireless network resources becoming increasingly cheaper and available, we design an experimental testbed and propose a new technique by coupling RF sensing with knowledge fusion and distillation.

The proposed **Deakin RF-Sensing** is based on the correlated knowledge distillation (CKD) theory, which aims to drive lightweight, near real-time, for precise human pose estimates. Our CKD-based technique uses "Teacher" and "Student" deep network models to train and implement two modes—a camera-fed teacher class network (e.g., images and videos) with an RF-fed student class network (e.g., RF signals). Our CKD model trains a dual multibranch teacher–student network and acquires a joint CKD model [12], [13], and slightly differently in the decoupled version [14]. The resulting CKD models are subsequently used to identify the multimodal correlation and teach the Student branch in reverse [15], [16]. Instead of simply aggregating their learnings, CKD training always seeks to comprise multiple parallel transformations with the two domains, i.e., visual images and RF signals, as feature mapping obtained by feature fusion seldom serves as a good teacher to guide their training [17]. Once trained, using radio signals generated from a software-defined radio (SDR), our CKD model can efficiently preserve privacy and utilize the multimodal correlated logits from the Teacher to the Student network without using visual signals/video frames.

A. Main Novelties and Contributions

We have explored and exploited the SDR mechanism to realize a hybrid RF and visual camera architecture for recognizing human actions nonintrusively. To this end, we established an in-house custom experimental testbed to investigate and research the performance of our suggested strategies. We present a novel CKD-based methodology to find and infer the complex relationships between RF and visual camera signals. The RF sensing can be driven separately from the camera's vision system once a CKD model is trained on the data and knowledge and insights about the correlation between the two signals are captured.

We used two Omni-directional VERT2450 antennas and a USRP N200 with a UBX 40 USRP Daughterboard. Each of the two antennas handles reception, while the other handles transmission. We utilized the SDR data for our estimation of posture annotations. Our CKD model is trained (supervised learning, applying proper pose annotations) on the acquired data (approximate stick figures). Combining an RGB camera with Alphapose enables such an accomplishment. Our testing results demonstrate the efficacy of the proposed integrated CKD-RF sensing framework.

Our main contributions to this article are explained below.

1) We develop a purpose-built experimental testbed framework that implements a joint SDR and visual camera-based system for near real-time human pose detection and analytics.

1) Purpose-Built Experimental Testbed: Our testbed network faces unique difficulties in its design and training compared to visual methods of posture examination such as Alphapose [18]. In our case, especially with RF signals, there is a lack of labeled data—one cannot realistically add keypoint annotations to RF broadcasts [19]. Therefore, we employ correlated monitoring to solve this issue, simply by pairing the camera with our RF-sensing, thus synchronizing the RF and Video streams during training. In training, we utilize RF stream supervision based on poses extracted from the video frames. Once trained, the system takes the RF as its only input and outputs the results. The end result of the proposed approach is that it can estimate a person's position only from RF signals, without any help from the camera. Moreover, the RF-based framework learns to quantify poses even when humans are entirely obscured, which is an exciting capability, and it achieves this capability despite never being exposed to such situations throughout its training.

1) We design and propose a novel CKD theory to rapidly track the changes in the physical environment by exploiting the multimodal correlation with the fusion between RF and visual signals.

2) Major Rethink on KD: We dig into the details and process of how information gets distilled in the classic [12] and decoupled [14] KD. Using ideas from [14], we recast the KD loss as a weighted sum of class-independent and class-dependent components. In particular, in the proposed CKD, the KD classification prediction problem is divided into two levels: 1) a binary prediction for the camera-based teacher class and all SDR-based student classes (RF-sensing) and 2) a multicategory prediction for each Student class. We use mathematics as a tool to reason the complexity of the developed CKD theory in Section II-B. Furthermore, we investigated the impact of each element of the CKD framework and exposed

the shortcomings of traditional KD. Using these results as inspiration, we present the new insights of CKD that perform very well in the context of the **Deakin RF-Sensing** proposed in this article.

1) We implement a tractable CKD model for private identification and predictions of human poses without feeding camera images.

3) Privacy Compliance: The ability of our framework to monitor the whereabouts of people without invading their privacy is a major selling point of this research project. All that can be done to identify people using RF-sensing is to gather their silhouettes, as no information about their faces or other attributes can be transmitted. Moving forward, we are creating a **Deakin RF-Sensing spinoff** product (a commercial application that is secure, efficient, and up to date with the current privacy norm) as privacy becomes more and more stringent.

B. Roadmap

The remainder of the article is organized as follows. Section I-C provides background information and reviews related work in the area. Section II presents our proposed joint RF and camera-based CKD approach, and Section III presents the technical setup of our experimental testbed. Section IV presents our experimental evaluation details and provides insights from our findings. Section V identifies key challenges and describes future research directions. Section VI concludes this article.

C. Literature

In this section, we present the background information and related works pertaining to our findings presented in this article.

1) Software-Defined Radio (SDR): SDR has recently gained attention in the sensing and communication community due to its unique ability to unify two fundamental technologies, i.e., software and digital radio [20]. The work in [21] shows SDR as a radio communication system enabling the modeling and control of complex RF sensing tasks using a localizing environment to tune the waveforms. There are currently limited research outcomes documented in the literature on using SDR to monitor and identify human activities and/or pose changes.

2) Camera-Based Motion Sensing: The visual camera produces RGB images, and any human pose estimation from these images generally falls under top–down and bottom–up approaches. In a top–down approach, the camera first detects the person from the image, then applies a single-person pose estimator to each person to extract key points [18], [22]. In a bottom–up approach, the camera first detects all the key points in the image first, then does some postprocessing to associate the key points [23], [24]. There are several approaches to using multiperson pose estimation from RGB images, for example, Alphapose [18], Yolo [25], and FPN [26]. We exploit Aphapose (which uses a two-stage scheme) as our primary pose supervision resource which can "teach" the mapping from WiFi signals to a person's pose.

3) WiFi-Based Human Activity Recognition: There has been much interest in exploiting wireless signals for locating subjects and monitoring human activities. They can be classified into two categories. The first category utilizes very high frequencies, e.g., millimeter wave (mmWave) or terahertz (THz) [27]. It can realistically image the surface of the human body (such as airport security scanners), but it cannot penetrate objects such as walls and furniture. The second category utilizes lower frequencies around a few GHz, thus allowing it to scan through walls and occlusions.

Important related works include crowd counting [5], [6], human identification through physical constraints [7], through-the-wall human detection [8], [28] or through-the-wall activity human activity or gesture recognition [4], [9], [10]. In particular, the work in [19] proposes a state-of-the-art vision model to provide cross-modal supervision. The authors extracted pose information from the visual stream input and used it to guide the training process of the network model. Once the model is sufficiently trained, it will only use wireless signals for pose estimation. Another work in [9] proposed a fully convolutional network to estimate multiperson pose from the collected data and analytics by training with the help of a visual stream from a camera.

The application use cases of the proposed mechanisms and techniques in this field are broad. The ability to accurately understand, correlate, and fuse data, before performing (near) real-time prediction is invaluable. The large body of work in this field has seen new and emerging use cases. For example, the notable "RF-Sleep" work presented in [29] and [30] learns to predict sleep stages from radio measurements without any attached sensors on subjects. A predictive model combining convolutional and recurrent neural networks (NNs) are developed to extract sleep-specific subject features from RF signals and progressively capture the temporal progression of sleep. An important aspect of the proposed method is the modified adversarial training regime that discards individual-specific information, which helps preserve the privacy of the subjects monitored. A recent work in [31] proposed an innovative contactless approach to monitor patients' blood oxygen remotely by analyzing the radio signals in the room where the patients are located. This is performed without any sensors or wearables attached to the patients. The patient's respiration is being monitored from the radio signals that are reflected on their bodies. This mechanism uses a novel NN that infers a patient's oxygen level from their breathing signal, with a gated transformer model. Advancements in coupling RF sensing and ML techniques have found important applications in healthcare, smart transportation and logistics, onset detection of diseases, agriculture, wildlife tracking, where privacy preservation is paramount [32], [33], [34], [35], [36], [37]. The use of CKD is limited in the literature, which forms the primary focus and novel contributions of our work presented in this article.

4) Knowledge Distillation (KD): CKD proposed in this article is an enhancement to KD by exploiting a knowledge fusion that extends a "teacher–student" approach to transfer knowledge from data/instances derived and collected from separate sources and transfer both instance-level and also

Fig. 2. Observations without using proposed CKD framework. Left: teacher class over original video frame with Alphapose detection. Right: student class CSI-trained model and pose detection using classic knowledge distillation.

the information of correlation between instances [14], [15], [16]. To further understand this, we have tested standard KD [12] teacher–student by plugging Alphapose [18] into our experimental testbed. Our observations are shown in Fig. 2. We find that the inaccuracy in the pose estimates that we can see in the right of Fig. 2 is due to RF noise and lack of knowledge fusion. In a different context, Zhao et al. [38], Peng et al. [15], Li et al. [16], and Chen et al. [14] proposed knowledge distillation (and with fusion [17]), which describes a learning style in which a more extensive teacher NN guides a student's NN learning process for numerous assignments. This understanding is passed on to students by such teachers using hidden labels [13] and fused multiscale distillation [17].

To the best of our knowledge, there are currently no existing works in the literature that study the use of CKD by employing knowledge fusion distillation theory to correlate RF and visual signals (using both SDR and visual camera) and build an integrated deep model to rapidly identify and predict human pose changes. Our work builds on existing literature in RF/WiFi sensing of human activities but extends to using SDR for RF signal generation and using CKD models for data correlation.

II. CKD Design

In this section, we present the design and technical details of our proposed joint CKD-RF sensing framework and approach for monitoring human pose.

A. Proposed CKD Framework

With relevant insights from cutting-edge KD works [14], [15], [16], [17], [38], in our proposed CKD approach, the classification prediction problem in the underlying KD is separated into two correlated levels: 1) a binary prediction for the camera-based teacher class and all SDR-based student classes (RF-Sensing) and 2) a multicategory prediction for each Student class. As shown in Fig. 3, using ideas from [14], we decompose the KD loss into two components.

1) Teacher Class Loss: In camera-based teacher class training, the prediction of the teacher class, denoted by $\mathcal{T}$, is provided, but the training of each student class is based only on information transfer through binary logit distillation. One plausible explanation is that the KD teacher class conveys information on the difficulty of training models with human poses; that is, the information specifies how challenging it is to recognize each pose.

2) Student Class Loss: In student class, we consider only the knowledge among student class logits that solely use the outcomes that are equivalent to or better than those of the conventional KD, demonstrating the critical role of knowledge included in student logits, which may provide insightful hidden information.

We have the following new information to develop the proposed CKD approach. The potential for logit distillation is constrained for some reasons, leading to uncorrelated results. Furthermore, straightforward early feature fusion with logits does not make a competent teacher direct their own training. We develop a module for correlated feature fusing and distillation, a novel technique for self-distillation. Logits are at more of a semantic and contextual level than deep network features; therefore, intuitively, logit distillation can achieve better efficiency than feature distillation. Therefore, we propose decomposing the celebrated KD mechanism toward correlated KD for multimodal learning. Such redesign of the KD approach enables us to investigate the effects of each component separately.

Algorithm 1 Details of the Pseudocode of CKD Algorithm

```
# temp: the temperature for KD and CKD
# a, b: hyper-parameters
.................................
Creating class  CKD(Distiller)
        self.lossweight = cfg.DKD.CE-WEIGHT
        self.a = cfg.CKD.A
        self.b = cfg.CKD.B
        self.tem = cfg.CKD.T

Procedure Forward Train
        logitsStudent= self.Student(csi)
        logitsTeacher = self.Teacher(frame)
End Procedure
.................................
Procedure LOSS
 lossce = lossweight*CrossEntropy(logits, target)
 lossCKD(epoch/warmup,CKDloss(logits,target,self)
EndProcedure

CKDloss(logitsStudent,logitsTeacher,target,a,b,tem)
    Teacher = F.softmax(logitsTeacher/tem)
    Student = F.softmax(logitsStudent/tem)
    Compute torch.log of (Teacher, Student)
   TKDloss =F.NSELoss(Student,Teacher)

    Teacher1 = F.log-softmax(Teacher, mask(logits,
    target))
    Student1 = F.log-softmax(Student, mask(logits,
    target))
    SKDloss =F.NSELoss(Student1,Teacher1)
return (TKDloss + b * SKDloss)
..............................
cat-mask(t): compute mask1, mask2
    rt = torch.cat([., .])
return rt
```

B. Details of CKD

The details of the proposed CKD approach are shown as pseudocode in Algorithm 1. For tractability [14], we use the following notations to quantify the correlated features classification relevant and irrelevant to the student class, $\mathcal{S}$. The classification probability of the logit sample, with l_i representing the logits of the ith classification, relevant to the Student class, is given by $\mathbf{p} = [p_1, p_2, \ldots, p_i, \ldots, p_N] \in \mathbb{R}^{1\times N}$, where N is the number of classifications and p_i is

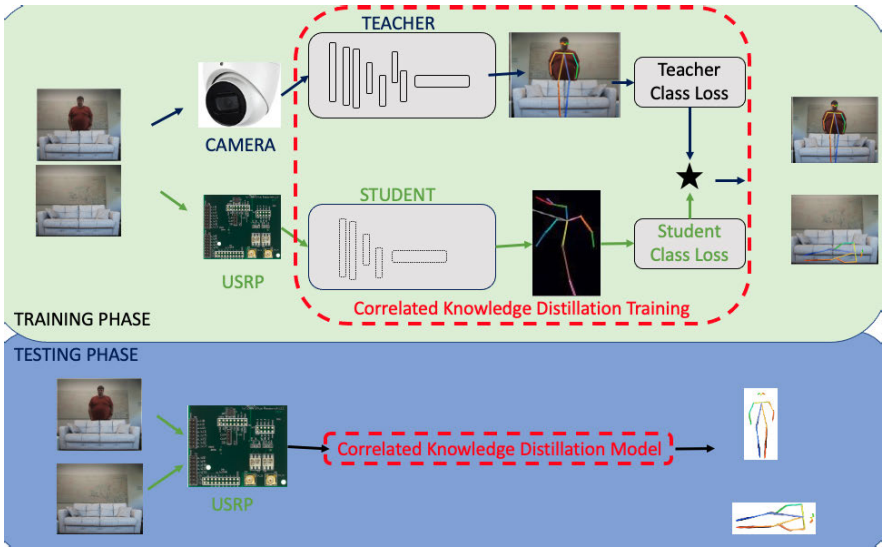


Fig. 3. Abstract view of the proposed CKD framework. Observe that the classification prediction problem of the standard KD is separated into two levels: 1) a binary prediction for the camera-based teacher class and all RF-sensing-based student classes and 2) a multicategory prediction for each student class. More importantly, we decompose the KD loss into two separate loss components.

the probability of the ith classification obtained using *softmax function*

$$p_i = \frac{\exp(l_i)}{\sum_i^N \exp(l_i)}. \tag{1}$$

Consider $\bar{\mathbf{p}} = [p_r, p_{\bar{r}}] \in \mathbb{R}^{1\times 2}$ as the binary probabilities of the relevant classification (p_r) and nonrelevant classification ($p_{\bar{r}}$), which can be computed as

$$p_r = \frac{\exp(l_r)}{\sum_i^N \exp(l_i)} \tag{2}$$

$$p_{\bar{r}} = \frac{\sum_{i\neq r}^N \exp(l_i)}{\sum_i^N \exp(l_i)}. \tag{3}$$

In addition, without considering the relevant rth classification, we quantify the probability $\hat{p}_i$ as

$$\hat{p}_i = \frac{\exp(l_i)}{\sum_{i\neq r}^N \exp(l_i)} \tag{4}$$

for independently determining the model probabilities

$$\hat{\mathbf{p}} = \left[\hat{p}_1, \hat{p}_2, \ldots, p_{r-1}, p_{r+1}, \ldots, p_N\right] \in \mathbb{R}^{1\times(N-1)}.$$

In general, the novelty of the proposed CKD can be seen in the form of a tractable classification loss process as follows. To enable the student model to learn the logit of the teacher model directly, we compute the loss by applying a Kullback–Leibler (KL) divergence-like measure or the normalized squared error (NSE) between the logit vectors as

$$\mathrm{NSE}(\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2) = \Delta(\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2) = \frac{\left\|\log \bar{\mathbf{p}}_1 - \log \bar{\mathbf{p}}_2\right\|^2}{N}. \tag{5}$$

Observe in (5) that the normalization of the squared error of the CKD estimator is the average of the squared differences between the logarithmic estimated classification probabilities of the teacher and student class. It is a risk function representing the predicted cost of squared deviation, which is never negative, and a value closer to 0 is preferable. Considering that the NSE is the second moment of the difference, $\Delta(\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2) = (\|\log \hat{\mathbf{p}}_1 - \log \hat{\mathbf{p}}_2\|^2)/N$, we include both the estimator's variance and its bias, to reflect the importance of correlated multimodal learning in the CKD framework.

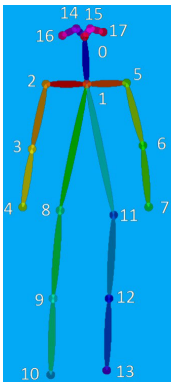
Using a similar intuition to that of KL-divergence (in the standard KD) for tractability, and using the aforementioned analyses [14], we determine the CKD loss (a joint NSE loss of two classes) as

$$\begin{aligned}\mathrm{CKDloss} &= \mathrm{TKDloss} + \left(1 - p_r^{\mathcal{T}}\right)\mathrm{SKDloss} \\ &= \mathrm{NSE}\left(\bar{\mathbf{p}}^{\mathcal{S}}, \bar{\mathbf{p}}^{\mathcal{T}}\right) + \left(1 - p_r^{\mathcal{T}}\right)\mathrm{NSE}\left(\hat{\mathbf{p}}^{\mathcal{S}}, \hat{\mathbf{p}}^{\mathcal{T}}\right).\end{aligned} \tag{6}$$

Intuitively, TKDloss is the feature classification loss, SKDloss is the model probabilities loss, and $p_r^{\mathcal{T}}$ is the relevant binary classification probability of the Teacher class network. To this end, the standard KD [12] with knowledge fusion allows us to look at the distinct effects of TKD and SKD, highlighting the inadequacy of existing approaches [14] in exploiting the underlying correlation and multimodal feature fusion.

C. Technical Setup and Design Details

To implement the proposed CKD, we extend AlphaPose [18] to process the video and annotate the poses with confidence levels and coordinates for each frame accurately. This extension produces a JSON output file that contains person pose coordinates. The number of posing coordinates may differ depending on the dataset and model; nevertheless, we simply employ FastPose with the Yolov3 detector and Resnet50 as the backbone for tractability. Unlike the AlphaPose [18] generating Teacher class model, we extract pose annotations and perform pose estimation from the WiFi channel state information (CSI) with learning to correlate with the student class model. However, we find that we cannot accurately determine the pose coordinates of a person simply by analyzing the CSI. See Fig. 3 for details. To ameliorate the shortcomings of such Student class training, we have implemented the proposed CKD approach, shown in Algorithm 1, with a camera-oriented teacher model training alongside the




```
//Result for COCO (18 body parts)
    {0,  "Nose"},
    {1,  "Neck"},
    {2,  "RShoulder"},
    {3,  "RElbow"},
    {4,  "RWrist"},
    {5,  "LShoulder"},
    {6,  "LElbow"},
    {7,  "LWrist"},
    {8,  "RHip"},
    {9,  "RKnee"},
    {10, "RAnkle"},
    {11, "LHip"},
    {12, "LKnee"},
    {13, "LAnkle"},
    {14, "REye"},
    {15, "LEye"},
    {16, "REar"},
    {17, "LEar"},
```

Fig. 4. Eighteen pose keypoint coordinates.

WiFi USRP to train the correlated model with videos and CSI.[1] For tractability [18], the underlying Alphapose of the teacher model in our configuration returns 18 key points as shown in Fig. 4 or pose coordinates and contributes to determining the corresponding teacher class loss. It is worth noting that these 18 key points are computed in the form of a JSON file format with a set of $\{x_1, y_1, c_1, \ldots, x_k, y_k, c_k\}$.

Our setup and design fully correlate the camera video with WiFi CSI which demonstrates the need for the proposed CKD approach. The WiFi CSI recording system consists of two ends, a sender with three antennas and a USRP with three antennas. When receiving WiFi transmissions, the USRP parses CSI [39] from WiFi signals. The parsed CSI is a tensor of the size of $(n \times 53 \times 3 \times 3)$, where n is the total number of signals received, 53 is the subcarrier and the two $3's$ are the number of sender and receiver antennas, respectively. AlphaPose produces n three-element predictions in the format of $(x_i, y_i; c_i)$, where n is the number of critical points to be estimated, x_i and y_i are the coordinates of the ith keypoint, and c_i is the confidence of the aforementioned coordinates. In experiments, we set $n = 18$, and use the keypoint configuration as shown in Fig. 4.

Assuming that I_t and C_t are a pair of synchronized image frames and RF CSI series, respectively, t denotes the sampling time, the training dataset we generate is denoted by $D = (I_t, C_t), t \in [1, n]$. By developing CKD as described in Algorithm 1, we trained D and learned their correlation. This is the learning mapping rules from the CSI series to the frame pose points. The Student class network $\mathcal{S}(\cdot)$ fed with CSI, and the teacher class network fed with AlphaPose, $\mathcal{T}(\cdot)$, is the entire CKD framework. $\mathcal{T}(\cdot)$ accepts I_t as input, and using the person detector and posture regressor, returns the body keypoint coordinates and confidence, $(x_t, y_t; c_t)$, for each pair of (I_t, C_t). The outputs are converted into an adjacent matrix to the body pose, or PAM_t. 1. Using $\mathcal{T}(I_t) \rightarrow \text{PAM}_t$, we model how the Teacher network teaches $\mathcal{S}(\cdot)$, where PAM_t stands for cross-modality supervision to teach $\mathcal{S}(\cdot)$.

In addition to the CKD, we have an encoder, feature generator, and decoder. Encoder upsamples C_t to proper width and height, suitable for the mainstream convolutional backbone networks and ResNet. Our CSI samples are $30 \times 3 \times 3$ and the RGB frame is $3 \times 244 \times 244$. Therefore, for CSI, by using eight gradually stacked transposed convolutional layers, we upsampled C_t. The feature generator learns effective features to estimate the pose coordinates of a person using the upsampled C_t. To this end, we require a robust feature extractor to generate spatial information from the upsampled C_t, compared to an image frame, as CSI lacks spatial information of the key points of the body. We stack four basic ResNet blocks (16 convolutional layers) as feature generators to extract insights (a size of $300 \times 18 \times 18$), which transforms the upsampled C_t to F_t.

Our decoder is for the adaptation and correlation of the shape between the learned characteristic, F_t, and the supervised output of $\mathcal{T}(\cdot)$, PAM_t. A body keypoint can be localized in the posture estimation job as two coordinates, namely, the x- and y-axes. Therefore, the decoder is designed to take F_t as input and predict the adjacency matrix within (x, y) dimensions, resulting in a *predicted pose adjacent matrix* $\overline{\text{PAM}_t}$. To accomplish this, we have two convolutional layers where the first layer is to release channel-wise information, and the second is to reorganize further the spatial information using kernels.

In fact, the student class deep network predicts a pose adjacent matrix on each CSI input, C_t, using the encoder, feature extractor, and decoder. This process is designated as $S(C_t) \in \overline{\text{PAM}_t}$. Furthermore, with the CKD approach, the PAM_t resulting from the teacher network is feedback to supervise every predicted $\overline{\text{PAM}_t}$ during the student training.

III. Experimental Methodology

In this section, we present the setup of our SDR-based testbed and the implementation details of our proposed CKD framework.

Our testbed implements a USRP N200 with a UBX 40 USRP Daughterboard and two Omni-directional VERT2450 Antennas. Using this configuration, we first implemented a frequency-modulated continuous-wave (FMCW) radar utilizing the GNU radio, which is a free and open-source radio platform for generating and interpreting radio signals based on C++ and Python. Some of the images of our testbed setup and data collection are shown at the top of Fig. 5. Fig. 6 shows our GNU radio configuration, which is extended with essential information from PiRadar. The SDR operates at 5.75 GHz and is fed with a saw tooth wave to modulate the carrier. We transmit the modulated carrier and record the reflected signal to detect state variations and predict pose estimates. This preliminary setup provides significant findings; however, we are unable to capture body movements from the binary data, which reflects very little of the pose changes. Also, the generated dataset is insufficient to feed deep NN training and potential knowledge distillation.

With relevant insights from our first experiment, we extend the testbed to capture CSI using the 802.11ac WiFi standard. Images of our extended testbed setup and data collection are also shown in the two lower of Fig. 5. WiFi access points transmit orthogonally modulated beacons containing information about available WiFi, such as SSID, encryption type, and supported rates. Using the same SDR of our earlier

[1]The RGB video is captured using a laptop webcam at 20 frames/s and a 720p resolution, which saves the timestamps for each frame and synchronizes with the corresponding CSI.

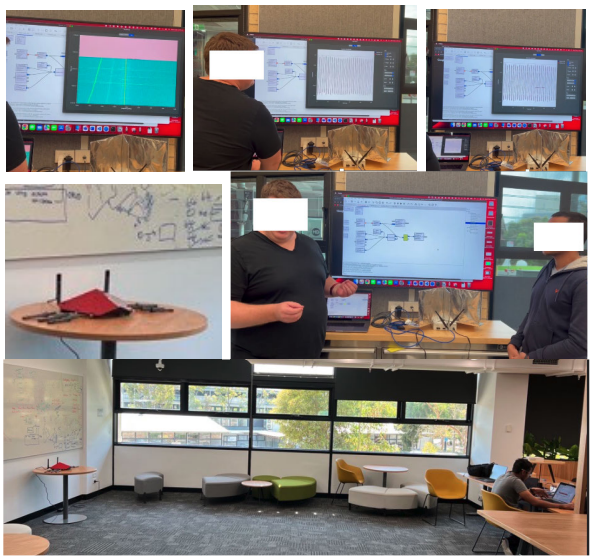


Fig. 5. Images of the testbed setup, videos, and RF data collection. The images at the top demonstrate our preliminary approach. The two bottoms demonstrate our extended testbed setup and data collection.

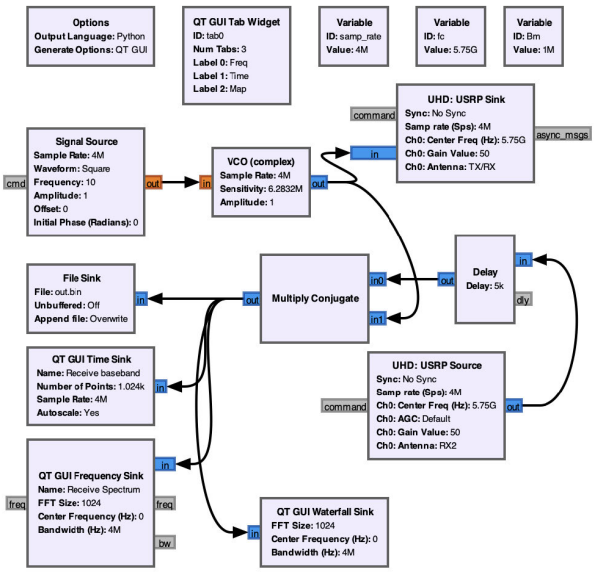


Fig. 6. FMCW radar design in GNU radio.

mentioned setup, we first capture the beacons generated from a different access point in the environment to record the CSI values. We employ PicoScenes [40],[2] to capture our CSI information. In this improved testbed, the USRP N200 is connected to a MacBook Pro-15-in 2017 with Ubuntu 20.04 to facilitate PicoScenes capturing network traffic over frequency 5220 MHz (5 GHz) and channel 44 driven by a Dlink DIR-895L access point. The video frames captured by the MacBook Pro are produced with a timestamp file for synchronization with the CSI.

We position the WiFi access points in our laboratory to have a clear line of sight from the USRP; 4 and 8 m is the typical separation between the WiFi access points and the USRP. The project team has followed the Deakin University policies and processes for carrying out human subjects' experimentation and data collection. In the two Deakin Laboratory rooms on the Burwood campus, two authors (Satish and Williams) performed everyday pose tasks for five weeks. Fig. 5 shows floor plans and data collection locations. We use the setup in Fig. 5 to simultaneously record CSI samples and videos of 5–7 s while the pose actions occur (see Fig. 3). Each person's 85% of videos and CSI recordings are used to train the networks, while 15% of CSI are used to test the networks. Training and testing have data sizes of 72 250, and 12 750 videos and CSI samples, respectively. We have received written consent from all participants.

[2]PicoScenes [40] is a middleware tool to capture the CSI and is compatible with our USRP-based testbed.

Algorithm 2 Pseudocode of Residual Block Developed for Ameliorating the Vanishing Gradients Impacts in the Proposed CKD Approach

```
Creating class ResidualBlock(Neural Network Module)
    super(ResidualBlock,self)\ldots
    self.bn1 = nn.BatchNorm2d(out_channels)
    self.relu = nn.ReLU(inplace=True)
    self.conv1 = conv3x3(in_channels,out_channels,.)
    self.conv2 = conv3x3(out_channels,out_channels)
    self.bn2 = nn.BatchNorm2d(out_channels)
forward(self, x)
    out = self.conv1(x)
    out = self.bn2(self.conv2(..)))
    out = self.relu(out)
return out
```

A. Implementation Challenges and Training Details

All convolution and deep networks mentioned above are implemented with PyTorch 1.0. They are trained using the Adam optimizer for 20 epochs with an initial learning rate of 0.001. In the teacher–student training of the proposed CKD, we noted that the learning rate decreases by approximately 0.3 near the seventh, tenth, and 18th epochs. After training, we determine the diagonal elements in $\overline{\text{PAM}}_t$ as the body keypoint prediction, e.g., $x_k^* = \overline{\text{PAM}}_{1,k}$ and $y_k^* = \overline{\text{PAM}}_{2,k}$, $k \in [1, 18]$.

1) Vanishing Gradient Problem: We have found that the straightforward implementation and training of the proposed CKD approach suffer from vanishing gradient problems. Therefore, we develop a residual block module in PyTorch, which helps to overcome the vanishing gradient problem prevalent in deep NN training. In general, the gradients keep decreasing as they are backpropagated through multiple layers leading to the vanishing problem, which creates difficulty for the optimizer to update the weights in the deeper network layers. The developed residual block module is used to establish a bypass connection that enables gradients to skip levels and directly propagate to deeper layers, thus alleviating the adverse impacts of such vanishing gradients.

The details of the developed residual block are shown in Algorithm 2. It takes as input the number of input channels, the number of output channels, the stride, and an optional downsample layer for scaling the tensor dimensions. As outlined in Algorithm 2, it has two convolutional layers with a batch normalization layer, a ReLU activation function, and two

convolutional layers between each. The residual connection is either the input tensor or the output of the downsample layer.

Algorithm 3 Pseudocode of the Extended ResNet Module for the CKD Approach

```
Extending class ResNet(nn.Module):
        super(ResNet, self)
        self.conv = nn.Sequential(Conv2,BatchNorm,.)
        self.layer1 = self.make_layer(.)
        \ldots\ldots\ldots
        self.layer4 = self.make_layer(.)
        self.decode = nn.Sequential(.),
forward(self, x):
        x = F.interpolate(.)
        x = self.layer1(x)
        \ldots\ldots\ldots
        x = self.decode(x)
return x
```

2) Revising ResNet for CKD: We revise and extend the standard ResNet module along the lines of the developed residual block discussed above. The revision of the ResNet module is to take the developed residual block module as input with a list of the number of layers in each block. As shown in Algorithm 3, our ResNet starts with a convolutional neural layer with 150 input and 150 output channels, followed by a batch normalization layer and ReLU activation. Adopting the *make_layer* function, it creates four residual blocks, each with increasing input and output channels. Observe in Algorithm 3 that the *make_layer* is fed with the residual block, the number of input channels, the number of blocks, and the stride value, which produces a list of residual block modules.[3]

Algorithm 4 Interaction of the Components of CKD Approach: How Algorithm 1 Works With Algorithms 2 and 3

```
from models.CKD_resnet
import ResNet, ResidualBlock, Bottleneck
    CKD = ResNet(ResidualBlock, [2, 2, 2, 2])
    resnet = ResNet(ResidualBlock, [3, 4, 6, 3])
    resnet = ResNet(Bottleneck, [3, 4, 6, 3])
    CKD = CKD.cuda()
criterion_L2 = nn.NSELoss().cuda()
optimizer = Optimize(Adam,CKDparam,learningrate)
scheduler = Optimize(MultiStepLR(optimizer),gamma)
    CKD.train()
```

The decoding layer at the end of the developed ResNet has a convolutional layer with 300 input channels and 64 output channels. It is followed by a batch normalization layer and a ReLU activation function. An input tensor is upsampled using the *F.interpolate* of the ResNet while the output tensor is created by first passing the tensor through the decoding layer.

3) Interaction of Algorithm 1 With Algorithms 2 and 3: The interactions of CKD with Algorithms 2 and 3 are shown in Algorithm 4. In our implementation, all other variables, parameters, and constraints such as *NSE*, and adam optimizer

[3]The first output block of the revised ResNet has the provided number of input channels and output channels, and succeeding blocks have output channels equal to the previous block's input channels.

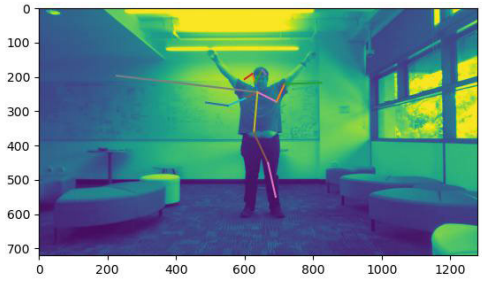


Fig. 7. Testing the student class model using standard KD [13] with test CSI data, pose estimates plotted over the original frame. It is erroneous and demonstrates that only CSI with the KD model cannot quantify the poses accurately.

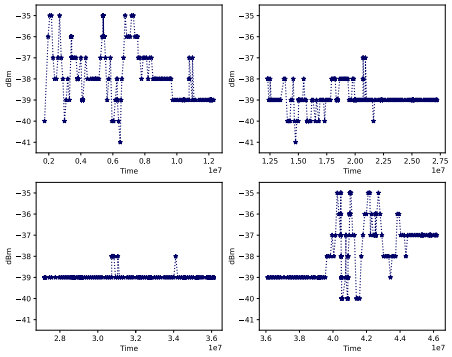


Fig. 8. Signal strength of CSI captured at USRP for the same pose in the same environment at different timestamps.

are inherited for the standard ResNet class constructed with a total of 8 Residual Blocks.

IV. Performance Evaluation and Analysis

We have implemented the CKD framework and used the dataset collected from the testbed. Small video sequences of 5–7 s are captured with the timestamps alongside PicoScenes, producing a CSI file containing all WiFi frames. Each frame has several features. We extracted the system time from the raw CSI data. These data are then converted/stored as a NumPy array and fed into the CKD along with the matching video frames.

1) Benchmarking With Standard KD: Using a standard KD [12], we performed the training and testing with the collected dataset for benchmarking. In Fig. 7, we can see that the pose estimation using standard KD is entirely inaccurate as the KD Loss framework is unable to capture the distinct difference in the classification loss of the video frame with that of CSI via the student and the teacher network model. In Fig. 8, the corresponding variation in the signal strength of the CSI captured by the USRP at four different timestamps under the same posture of the same person is shown in Fig. 7. This is due to the adverse impact of RF noise and WiFi traffic fluctuations and demonstrates why the pose estimation directly from RF signals, such as using CSI shown in Fig. 7, is erroneous and is highly challenging (without knowledge of fusion and correlated distillation from multimodal aspects).

2) Training Efficiency of the Proposed CKD Framework: To demonstrate the excellent training efficiency of the proposed CKD, we compare training costs with other cutting-edge distillation techniques [12], [38]. Our CKD produces optimal results

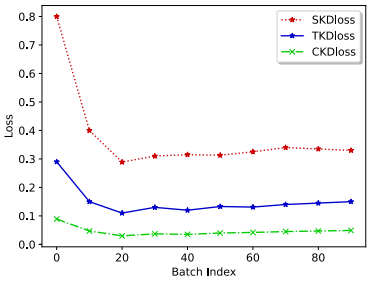


Fig. 9. Evolution of the teacher (TKDloss), student (SKDloss), and CKD (CKDloss) loss over several epochs. The notable improvement (see SKDloss → CKDloss) considerably increases the accuracy in pose detection, and most pose estimation testing obtained with the trained CKD model is much more accurate.

TABLE I
TRAINING TIME AND ACCURACY OF THE CKD COMPARED TO OTHERS [12], [38] WITH THE COLLECTED DATASET IN OUR SETUP AS EXPLAINED IN SECTION III

Approach	Top-1 Acc.	Parameters	Time in ms
KD [12]	70.23%	Same parameters	10 milliseconds
R-KD [38]	72.6%	1.8 Million +	22 milliseconds
CKD [38]	76.46%	Same parameters	11 milliseconds

by optimizing the tradeoff between model performance and training costs such as training time and additional parameters. The training time, top-1 accuracy, and additional parameter requirements of the proposed CKD framework with those of others [12], [38] are tabulated in Table I. Observe that, as a reformed version of the classic KD [12], CKD requires virtually the same computing complexity as KD and, indeed, the same set of parameters. Almost all celebrated feature-based distillation algorithms ([38] and its enhancements) need additional training time to distill intermediate layer features, a huge number of additional parameters, and comparatively high GPU memory expenses, which are impractical in the context of the Deakin RF-Sensing studied in this work.

With the proposed CKD approach, we have successfully minimized the training loss considerably over the epoch index, as shown in Fig. 9. We can see the differences in the evolution of the teacher (TKDloss), student (SKDloss), and CKD (CKDloss) loss over several epochs. This notable improvement (see SKDloss → CKDloss) considerably increases the accuracy in pose detection, and most pose estimation tests obtained with the trained CKD model are much more accurate. We have shown some examples of our test images with the accuracy of the estimation of the CKD in Fig. 10.

To further understand this, we compute the MSE for each batch in different training epochs with the proposed CKD. As shown in Fig. 11, we found that CKD can exploit correlation to minimize loss by considerably increasing the number of batches. Sample training and testing data can be found in the corresponding GitHub repositories. We have seen practically how the CKD method accelerates learning and improves accuracy. Our experiment shows that the CKD model successfully learns a person's posture when the person is relatively stationary and moving slowly but struggles when the person's posture is rapidly changing (as shown in Fig. 10).

Remark 1 (Monitoring Rapidly changing Poses): In our testbed, the RF data are captured differently (as complex numbers) and are noticed from different angles

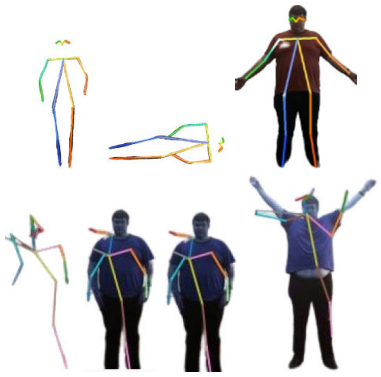


Fig. 10. Testing accuracy of the proposed CKD framework. Top: highly accurate pose estimates; bottom: relatively low accurate pose estimates; both using the test samples over the same experimental testbed.

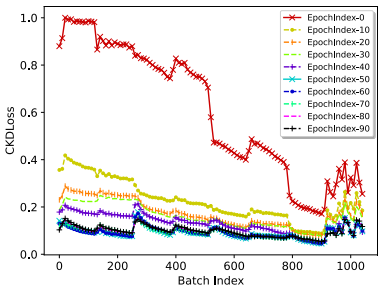


Fig. 11. Comparison of the CKDloss over each batch index of data in different epochs as observed in our testing setup.

(horizontal and vertical projections) than the video frames, as the spatial resolution of RF signals is substantially lower than the video signals (e.g., wall-penetrating frequencies). It is known that when the wavelength is longer than the surface roughness, a physics phenomenon known as RF-specularity takes place. Therefore, our body is specular in frequencies that may even go through walls. Intuitively, in the context of this research and our testbed setup, our body behaves more like a reflector (a mirror) than a scatterer. With a wavelength of around 3–5 cm, humans serve as mirrors. The signal's reflection direction depends on whether or not each limb's surface is toward our USRP. It is worth noting that RF differs from a camera system in that a video frame only contains information on a subset of our limbs. We have observed that, because of a single USRP sensor in our testbed, when a person is running, we sometimes miss limbs and body parts whose orientation at that moment echoes the signal away from the USRP. We anticipate that tracking such rapidly changing aspects requires sophisticated testbed development and appears plausible only with multiple USRPs at spatially dispersed locations to fill such voids/misses.

V. APPLICATIONS, CHALLENGES, AND FUTURE DIRECTIONS

Initial findings and experimental insights of our proposed CKD-RF framework demonstrated the feasibility and practicality of correlating two (or more) independent data sources. Our proposed CKD-RF framework has broad application domains, spanning healthcare, smart transportation/logistics, mission-critical applications, education, and so forth.

In addition to our motivating scenario of privacy-preserving patient monitoring, the use of our framework can be directly applicable to those similar to previous pioneering work in the field, including sleep monitoring [29], [30], onset detection of diseases [31], pose estimation [41] with fall detection use cases in homes/care/educational centers [32], [33], [34], [35], [36]. With advanced enhancements and improvements, our framework could be extended to other avenues such as vehicular traction vehicular traction control (with multiple sensors fed into the CKD model), mission-critical applications for subject localization, contact tracking, and movement control during a disease outbreak, accurate positioning and localization unmanned aerial vehicles (UAVs) for goods delivery, and wildlife tracking [37].

Although our findings improved with CKD, the CSI data we collect contain only one dimension because we only employ a single WiFi transmitter and USRP receiver. Therefore, the proposed CKD framework can only accurately extract crucial spatial information by reshaping and upsampling the poor 1-D CSI. Future work should expand this study to include scenarios with multiple access points and multiple USRP to test the feasibility of our proposed framework.

There are also challenges surrounding the unwanted impacts of RF noise in our experiments. There are some well-known and documented alternative approaches for ameliorating the adverse impacts of RF noise that could be considered in our future experiments. For example, it is feasible to use ultra-wideband (UWB) technology [42], [43] to generate extremely narrow pulses or very short-duration signals which will mitigate the influence of radio interference and enable the signals to coexist with signals generated by other wireless systems. Our USRP SDR (and similar SDR and cognitive radio-based technology) [44] allows us to adjust the transmitting/receiving frequencies, modulation schemes, and other associated parameters. Incorporating the ability to dynamically adjust and adapt these factors and parameters makes our system more resilient to RF noises.

The selection of frequency hopping and spread spectrum techniques (such as direct sequence spread spectrum and frequency hopping spread spectrum) within the generated RF signals in the SDR [45], [46] will intrinsically reduce the impact of RF noises. Our CKD-RF framework in conjunction with multiple directional antennas (such as those used in antenna arrays) [47], [48], [49] for data transmission/reception will significantly improve the accuracy by concentrating the signal toward a particular receiver, thus improving the signal-to-noise ratio. In addition, new SDR devices can enable dynamic spectrum access [50], [51] where the transmission frequency and power levels are intelligently adapted to the environment to avoid areas with high levels of noise and congestion. Unused frequencies can be quickly identified and opportunistically utilized.

We understand that the RF noise in the proposed sensing system restricts how stable the RF field can become. We have found that the RF controller can dampen noise, but still affects the cavity fields. The stability of the RF field is affected by feedback from the SDR detector noise, which in turn, impacts the accuracy and precision of the field measurements. A thorough investigation of the noise's origins and effects on the RF field is necessary to establish the RF system's performance requirements. For such a kind of quantification and development of tightly coupled approaches with CKD for noise amelioration, one needs to develop noise transfer relationships in RF control loops. In addition, noise models of several essential RF components (like amplifiers, mixers, and analog-digital converters) are to be developed, which indicates an important research direction to be considered in the future.

There is a great deal of potential in this area and in the ways it may be used. Still, we encountered some challenges (as discussed earlier) during implementation that are now resolved, and we can confidently use the proposed CKD approach. We have identified the following three important directions.

1) Our preliminary experiments and studies (along the lines of Figs. 8 and 10) with two access points suggest multiple access points and USRP receivers can potentially generate higher accuracy and capture a spatial offset to quantify the dynamic change in CSI efficaciously, which requires extensive experiments in the future.
2) We have observed that the volume of WiFi traffic fluctuates over time and cannot generate enough RF signals for effective passive listening. Therefore, we need to look for an alternative approach that can generate more RF steadily with more success and ameliorate the adverse impacts of RF noise. Such an approach can extend the proposed CKD along the lines of the captured signals and by capturing seamless datasets.
3) It appears possible to generate a precise RF signal and continuously monitor it from the access points. An additional module on the top of the CKD can help understand and predict the received RF signal, which can be more easily detected for deviations.

VI. Conclusion

In this work, we developed a CKD system to teach our inhouse wireless network, comprising WiFi and USRP, to infer human postures and movements. The CKD system aimed to achieve a hybrid RF and camera framework for privacy-preserving human activity detection using only RF signals. CKD training involved two parallel classification training approaches with images and radios, allowing the fusion and distillation of knowledge. Once trained, the CKD model effectively preserved privacy and leveraged the correlated multimodal logits from the Teacher to the Student network, eliminating the need for images and video information. Through testing and validation using our SDR-based experimental testbed, we demonstrated the feasibility of the developed CKD framework, highlighting its potential for significant impact on the field. Finally, we also note that the method of training using correlated twin NNs, where one is a "Teacher network," is useful for leveraging pretrained networks in the absence of (or limited) data.

Acknowledgment

The authors appreciate the time and efforts of the two authors (Satish and Williams) who helped build their datasets

and perform coding as human subjects. They also appreciate the constructive criticism provided by the anonymous reviewers.

REFERENCES

[1] A. Esteva et al., "Deep learning-enabled medical computer vision," *npj Digit. Med.*, vol. 4, no. 1, p. 5, Jan. 2021.

[2] W. Taylor, M. Z. Khan, A. Tahir, A. Taha, Q. H. Abbasi, and M. A. Imran, "An implementation of real-time activity sensing using Wi-Fi: Identifying optimal machine-learning techniques for performance evaluation," *IEEE Sensors J.*, vol. 22, no. 21, pp. 21127–21134, Nov. 2022.

[3] S. Z. Gurbuz et al., "American sign language recognition using RF sensing," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3763–3775, Feb. 2021.

[4] S. Yue, Y. Yang, H. Wang, H. Rahul, and D. Katabi, "BodyCompass: Monitoring sleep posture with wireless signals," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 2, pp. 1–25, Jun. 2020.

[5] S. Depatla and Y. Mostofi, "Crowd counting through walls using WiFi," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2018, pp. 1–10.

[6] A. Hanif, M. Iqbal, and F. Munir, "WiSpy: Through-wall movement sensing and person counting using commodity WiFi signals," in *Proc. IEEE SENSORS*, Oct. 2018, pp. 1–4.

[7] L. Cheng and J. Wang, "Walls have no ears: A non-intrusive WiFi-based user identification system for mobile devices," *IEEE/ACM Trans. Netw.*, vol. 27, no. 1, pp. 245–257, Feb. 2019.

[8] H. Sun, L. G. Chia, and S. G. Razul, "Through-wall human sensing with WiFi passive radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 4, pp. 2135–2148, Aug. 2021.

[9] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with WiFi fingerprints," *IEEE Access*, vol. 7, pp. 80058–80068, 2019.

[10] H. Abdelnasser, K. Harras, and M. Youssef, "A ubiquitous WiFi-based fine-grained gesture recognition system," *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2474–2487, Nov. 2019.

[11] T. Li, L. Fan, Y. Yuan, and D. Katabi, "Unsupervised learning for human sensing using radio signals," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1091–1100.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Stat*, vol. 1050, p. 9, Mar. 2015.

[13] Z. Li, J. Ye, M. Song, Y. Huang, and Z. Pan, "Online knowledge distillation for efficient pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11720–11730.

[14] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11943–11952.

[15] B. Peng et al., "Correlation congruence for knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5006–5015.

[16] X. Li, J. Wu, H. Fang, Y. Liao, F. Wang, and C. Qian, "Local correlation consistency for knowledge distillation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K.: Springer, Aug. 2020, pp. 18–33.

[17] L. Li et al., "Knowledge fusion distillation: Improving distillation with multi-scale attention mechanisms," *Neural Process. Lett.*, vol. 55, pp. 6165–6180, 2023, doi: 10.1007/s11063-022-11132-w.

[18] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[19] M. Zhao et al., "Through-wall human pose estimation using radio signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7356–7365.

[20] W. H. W. Tuttlebee, *Software Defined Radio*. Hoboken, NJ, USA: Wiley, 2002.

[21] M. Z. Khan, A. Taha, W. Taylor, M. A. Imran, and Q. H. Abbasi, "Non-invasive localization using software-defined radios," *IEEE Sensors J.*, vol. 22, no. 9, pp. 9018–9026, May 2022.

[22] G. Papandreou et al., "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3711–3719.

[23] L. Pishchulin et al., "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4929–4937.

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[27] Y. Zhu, Y. Zhu, B. Y. Zhao, and H. Zheng, "Reusing 60 GHz radios for mobile radar imaging," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2015, pp. 103–116.

[28] M. Zhao et al., "Through-wall human mesh recovery using radio signals," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10112–10121.

[29] F. Zhang et al., "SMARS: Sleep monitoring via ambient radio signals," *IEEE Trans. Mobile Comput.*, vol. 20, no. 1, pp. 217–231, Jan. 2021.

[30] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 4100–4109.

[31] H. He, Y. Yuan, Y.-C. Chen, P. Cao, and D. Katabi, "Contactless oxygen monitoring with gated transformer," 2022, *arXiv:2212.03357*.

[32] D. Vasisht, A. Jain, C.-Y. Hsu, Z. Kabelac, and D. Katabi, "Duet: Estimating user position and identity in smart homes using intermittent and incomplete RF-data," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 2, pp. 1–21, Jul. 2018.

[33] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "RF-based fall monitoring using convolutional neural networks," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–24, Sep. 2018.

[34] C.-Y. Hsu, R. Hristov, G.-H. Lee, M. Zhao, and D. Katabi, "Enabling identification and behavioral sensing in homes using radio reflections," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–13.

[35] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, "Making the invisible visible: Action recognition through walls and occlusions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 872–881.

[36] T. Li, H. Chang, S. K. Mishra, H. Zhang, D. Katabi, and D. Krishnan, "MAGE: Masked generative encoder to unify representation learning and image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2142–2152.

[37] K. VonEhr, S. Hilaski, B. E. Dunne, and J. Ward, "Software defined radio for direction-finding in UAV wildlife tracking," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, May 2016, pp. 0464–0469.

[38] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5006–5015.

[39] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, Jan. 2011.

[40] Z. Jiang et al., "Eliminating the barriers: Demystifying Wi-Fi baseband design and introducing the PicoScenes Wi-Fi sensing platform," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4476–4496, Mar. 2022.

[41] M. Zhao et al., "RF-based 3D skeletons," in *Proc. Conf. ACM Special Interest Group Data Commun.*, Aug. 2018, pp. 267–281.

[42] G. R. Aiello and G. D. Rogerson, "Ultra-wideband wireless systems," *IEEE Microw. Mag.*, vol. 4, no. 2, pp. 36–47, Jun. 2003.

[43] Z. Sahinoglu, S. Gezici, and I. Guvenc, *Ultra-Wideband Positioning Systems*. New York, NY, USA: Cambridge, 2008.

[44] T. Ulversoy, "Software defined radio: Challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 4, pp. 531–550, 4th Quart., 2010.

[45] T. J. Rouphael, *RF and Digital Signal Processing for Software-Defined Radio: A Multi-Standard Multi-Mode Approach*. Oxford, U.K.: Newnes, 2009.

[46] R. Kohno, "Signal processing and ASIC's for ITS telecommunications—Spread spectrum, array antenna and software defined radio for ITS," *IEICE Trans. Fundamentals Electron. Commun. Comput. Sci.*, vol. 85, no. 3, pp. 566–572, 2002.

[47] M. M. Rahman, H. E. Baidoo-Williams, R. Mudumbai, and S. Dasgupta, "Fully wireless implementation of distributed beamforming on a software-defined radio platform," in *Proc. ACM/IEEE 11th Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2012, pp. 305–315.

[48] A. M. Ashleibta, Q. H. Abbasi, S. A. Shah, M. A. Khalid, N. A. AbuAli, and M. A. Imran, "Non-invasive RF sensing for detecting breathing abnormalities using software defined radios," *IEEE Sensors J.*, vol. 21, no. 4, pp. 5111–5118, Feb. 2021.

[49] V. Goverdovsky, D. C. Yates, M. Willerton, C. Papavassiliou, and E. Yeatman, "Modular software-defined radio testbed for rapid prototyping of localization algorithms," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 7, pp. 1577–1584, Jul. 2016.

[50] R. A. Rashid, M. A. Sarijari, N. Fisal, S. K. S. Yusof, and S. H. S. Ariffin, "Enabling dynamic spectrum access for cognitive radio using software defined radio platform," in *Proc. IEEE Symp. Wireless Technol. Appl. (ISWTA)*, Sep. 2011, pp. 180–185.

[51] D. A. Clendenen, "A software defined radio testbed for research in dynamic spectrum access," Ph.D. dissertation, Dept. Elect. Comput. Eng., Purdue Univ., West Lafayette, IN, USA, 2012.

Shiva Raj Pokhrel (Senior Member, IEEE) was born in Damak, Nepal. He received the B.E. and M.E. degrees from NCIT, Pokhara University, Lekhnath, Nepal, in 2007 and 2013, respectively, and the Ph.D. degree from the Swinburne University of Technology, Hawthorn, VIC, Australia, in 2017.

He was a Research Fellow at the University of Melbourne, Melbourne, VIC, Australia, and a Network Engineer at Nepal Telecom, Kathmandu, Nepal, from 2007 to 2014. He has been a Deakin RF-Sensing Research Program Leader at the IoT Research Laboratory/Deakin University, Geelong, VIC, Australia, since 2022, where he is a Senior Lecturer of Mobile Computing. His research interests include multiconnectivity, federated learning, industry 4.0 automation, blockchain modeling, optimization, smart grid, recommender systems, 6G, cloud computing, dynamics control, the Internet of Things, and cyber-physical systems and their applications in smart manufacturing, autonomous vehicles, and cities.

Dr. Pokhrel received the prestigious Marie Curie Fellowship from 2017 to 2020 and the Comcast Innovation 2021 and was the finalist of the IEEE Future Networks' *Connecting The Unconnected Challenge* (CTUC) in 2021. He serves/served as the Workshop Chair/Publicity Co-Chair for several IEEE/ACM conferences, including IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, and ACM MobiCom. He is an Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and *MDPI Drones*.

Jonathan Kua (Member, IEEE) received the B.Eng. (Hons.) degree in telecommunications and network engineering and the Ph.D. degree in telecommunications engineering from the Swinburne University of Technology, Melbourne, VIC, Australia, in 2014 and 2019, respectively.

He is currently a Lecturer in the Internet of Things with the School of Information Technology, Deakin University, Melbourne. His research interests include data transport protocols and active queue management, adaptive streaming and content delivery, the Internet of Things, and industry 4.0, with an overarching focus on improving their network performance and quality of service.

Deol Satish is a Research Assistant at the IoT Research Laboratory, Geelong, VIC, Australia, and a recent Deakin University Software Engineering (Hons.) Graduate with exceptional technical and coding abilities. His expertise in working on RF-sensing projects involving various technologies, including kernel development, cloud computing, and machine learning is phenomenal.

Philip Williams is pursuing the master's degree in information technology with the IoT Research Laboratory, Geelong, VIC, Australia.

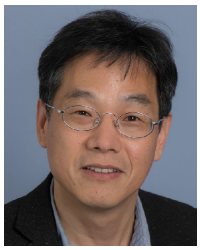
He is a Research Assistant of Deakin RF-Sensing at the IoT Research Laboratory. He has been an Implementation Analyst with 20 years in the retail industry in Melbourne, VIC, Australia.

Arkady Zaslavsky (Senior Member, IEEE) is currently a Distributed Systems and Security Professor at Deakin University, Melbourne, VIC, Australia. He is also leading and participating in research and development projects in the Internet of Things, mobile analytics, and context-awareness science areas. He is also a Technical Leader of the EU Horizon-2020 Project bIoTope—Building IoT Open Innovation Ecosystem for connected smart objects. He holds adjunct professorship appointments with several Australian and international universities, including UNSW, Sydney, NSW, Australia; La Trobe University, Melbourne, VIC, Australia; University of Luxembourg, Esch-sur-Alzette, Luxembourg; and ITMO University, St. Petersburg, Russia. He has published over 400 research publications throughout his professional career and supervised completing more than 40 Ph.D. students.

Mr. Zaslavsky is a Senior Member of ACM and the IEEE Computer and Communication Societies.

Seng W. Loke (Member, IEEE) received the Ph.D. degree in computer science from the University of Melbourne, Melbourne, VIC, Australia, in 1998.

He is currently a Professor of Computer Science at the School of Information Technology, Deakin University, Melbourne. He has authored or coauthored over 300 publications. He also authored *Context-Aware Pervasive Systems: Architectures for a New Breed of Applications by Auerbach* (CRC Press, December 2006), *Crowd-Powered Mobile Computing and Smart Things* (Springer, 2017), and *The Automated City: Internet of Things and Ubiquitous Artificial Intelligence* (Springer, 2021). His research interests include the Internet of Things (IoT), pervasive and mobile computing, and intelligent transport systems, with a current focus on complex trusted cooperation among things (including smart cooperative vehicles), and the social impact of technology innovation.



Jinho Choi (Fellow, IEEE) was born in Seoul, South Korea. He received the B.E. (magna cum laude) degree in electronics engineering from Sogang University, Seoul, in 1989, and the M.S.E. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1991 and 1994, respectively.

He is a Professor at the School of Information Technology, Burwood, Deakin University, Melbourne, VIC, Australia. Prior to joining Deakin University, in 2018, he was with Swansea University, Swansea, U.K., as a Professor/Chair in Wireless, and the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, as a Professor. He has authored two books published by Cambridge University Press in 2006 and 2010 and one book by Wiley-IEEE in 2022. His research interests include the Internet of Things (IoT), wireless communications, and statistical signal processing.

Prof. Choi received a number of best paper awards including the 1999 Best Paper Award for Signal Processing from EURASIP. He has been on the list of the World's Top 2% Scientists by Stanford University since 2020. Currently, he is a Senior Editor of IEEE WIRELESS COMMUNICATIONS LETTERS and an Associate Editor of IEEE TRANSACTIONS ON MOBILE COMPUTING. He has also served as a Division Editor for *Journal of Communications and Networks* (JCN), an Associate Editor or Editor for other journals including IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, *JCN*, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and *ETRI journal*.