Cross-Dual Path Attention for Concurrent CSI-based Applications in Indoor Environments

Shervin Mehryar
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
shervin.mehryar@utoronto.ca

***Abstract*—Due to the omnipresence of radio frequency signals, the Channel State Information (CSI) can offer an alternate source to image, video, and other high-dimensional streams in a great many Internet-of-Things (IoT) applications. As a result, an ever increasing number of researchers are advocating for the use of passive CSI data for ranging, tracking, perception and automation across many domains such as robotics, healthcare, and surveillance. Specifically in indoor environments where movements cause classifiable effects, the CSI can be leveraged to provide a high-dimensional signal source for a broad set of applications including activity, gesture, pose, location, and orientation recognition. This task however remains a challenge on two accounts. On the one hand, the radio frequency channel is highly susceptible to environment changes and artifacts. On the other hand, there is a lack of robust models that cover the full range of applications for practical deployment. In this work, we focus on tackling these issues by proposing a novel cross-dual-path-attention architecture that is robust against environment variations and achieves high accuracy across multiple tasks in practical settings. Our experiments on multiple datasets verify that the proposed architecture consistently outperforms the state-of-the-art methods when tested for concurrent application.**

***Index Terms*—Cross-dual path attention, channel state information, activity recognition, pose estimation, gesture recognition, finger printing.**

I. Introduction

Radio frequency signals have enabled innovative solutions across various indoor IoT applications by leveraging their ability to capture environment-dependent information. This non-intrusive approach is increasingly used for applications such as location, gesture, pose, and human activity recognition in domains such as healthcare, robotics, and surveillance. There are several approaches to radio frequency-based recognition, including device-based and device-free. In device-based approaches, input radio signals are collected directly from sensors attached onto subjects. Equipping subjects with digital sensors is inherently cumbersome, and in residential and most commercial applications impractical. Alternatively, passive sensing from radio frequency signals tranceived between edge-devices in indoor settings is remarkably non-intrusive (i.e. no wearables) and of high resolution [1].

Considering the advantages of device-free solutions, recently the channel state information (CSI) from network interface cards has gained a lot of attention [2]. WiFi enabled devices readily exist in many indoor environments (e.g. laptops, tablets, etc) at no additional cost. In essence, for passive sensing then all that is required is a wireless channel (for example at 2.4GHz or 5Ghz in the 802.11ax standard) established between various WiFi enabled devices. Because the movement of human subjects in relation to objects changes the multipath characteristics of the channel, CSI will have a different amplitude/phase at any given time. As opposed to received signal strength (RSS) which provides coarse information (MAC layer), the CSI carries fine-grained information (PHY layer) measured through orthogonal frequency division multiplexing (OFDM) on multiple sub-carriers [3]. Other sources such as ultra-wide band also exist often at an additional per device cost [4]. The CSI which contains sufficient features, is nevertheless accessible, economical, and therefore highly lucrative as a high dimensional resource.

Building fully ubiquitous passive recognition systems from a technological standpoint remains a challenge due to the fact that small changes in environment setting greatly impact the received patterns. The predominant factors here are: 1) equipment setting (e.g. distance and antennas/sub-carrier specifications); 2) environment artefacts (e.g. furniture layout); 3) location of devices and subjects; 4) pose/orientation of subjects with respect to transceivers. These effects can vary within a given environment, for instance when furniture are rearranged. As a result, a recognition model that is trained for a specific setting may not work well in another one. Device-based approaches can straight forwardly filter such effects out as background noise inside the sensory hardware. In passive sensing however, there is no apparent solution in practical settings and for multiple concurrent applications.

In this work, we propose a novel architecture that addresses the above challenges related to environment and task variability through feature learning from CSI input frames using a cross-dual path attention (CDPA) mechanism. In particular, the CSI inputs are fed through a Spatial Transformation (ST) block in parallel with a Temporal Transformation (TT) block with learnable parameters to extract application dependent features. The CDPA subsequently fuses the respective transformations into a common embedded representation for concurrent prediction in multi application settings. We introduce the CDPA model along with ST and TT components in Section II. We train and test the model on a variety of applications related to activity, gesture, pose, orientation, and location, including 10 different datasets. The performance of the proposed method in each case is evaluated and compared to the state-of-the-art approaches which we report in Section III. To the best of our knowledge, this is the first design of its kind to cover a diverse range of indoor applications.

II. Proposed Method

In this section, we provide an in depth explanation of the proposed CSI-based architecture for multiple applications. The components of the system are as follows. The input data are first transformed into a spatial representation to capture subtleties related to fine motions (e.g. swipe, etc), as explained in Section II-B. The spatial representation is obtained through a Convolution Neural Network (CNN). A temporal representation of the input is also obtained through a Transformer architecture to further capture the inter-subcarrier corelations, as explained in Section II-C. The output representations are fused through a cross-attention block, termed Cross-Dual-Path-Attention (CDPA), as explained in Section II-D, in order to determine interactions across representations using attention mapping. These components are fully and end-to-end trainable given labeled data.

A. Model Inputs

Given a pair of receiver and transmitter WiFi antennas, each with N_{rx} and N_{tx} antennas and K sub-carriers, the channel information are collected at a given time as input data. The vector of channel state information is constructed as $\vec{\mathbf{h}} = [h_1, \cdots, h_K] \in \mathbb{C}^K$. Consider N (discrete) channel measurements, the collected CSI in compact matrix form can be constructed by $\mathbf{H} = [\vec{\mathbf{h}}_1, \cdots, \vec{\mathbf{h}}_N] \in \mathbb{C}^{K \times N}$, where N is the sequence length corresponding to the time index and K the total number of sub-carriers over a multi-antenna, multi-channel communication link (the number of antennas times the number of sub-carriers). The CSI tensor $\mathbf{H}$ consisting of the raw magnitude/phases forms the input to our system. We employ two different neural network architectures to extract features from the raw input representations. First, we employ a CNN architecture which takes the raw input to extract rich features across the spatial dimension. Next we employ a Transformer architecture which is equipped with a self-attention mechanism ideal for capturing time index dependencies. The input transformations are elaborated in the following subsection.

B. Spatial Transformation

In the spatial transformation modality (ST), the representations of the raw CSI data are fed into a CNN in order to extract time and frequency related features. The tensor of raw data, i.e. $\mathbf{H}$, is transformed through the CNN blocks followed by a single layer perceptron (SLP) to produce its final spatial representations $\hat{\mathbf{H}}_{\mathcal{S}}$. The relation between the matrix of original raw and the transformed matrix of CSI through spatial transformation $\mathcal{S}$, denoted by $\hat{\mathbf{H}}_{\mathcal{S}}$, is denoted by:

$$\hat{\mathbf{H}}_{\mathcal{S}} = f_{\mathcal{S}}(\mathbf{H}), \tag{1}$$

where $f_{\mathcal{S}}$ is the spatial transformation function, parameterized by $\Theta_{\mathcal{S}}$ as shown in Spatial Transformation block in Figure 1.

C. Temporal Transformation

In the temporal transformation modality, we aim to capture the long term temporal dependencies in the input representations. Let $\mathcal{T}$ be the self-attention map/matrix, where the higher the value of j'th element in the i'th column, the higher the corelation between the j'th and i'th sub-carriers at a given time. Given the input representation $\mathbf{H}$, the attention map/matrix is given by $\mathcal{T} = \textit{softmax}(\mathbf{W}^{\mathcal{T}} \circ \mathbf{H})$, where $\mathbf{W}^{\mathcal{T}}$ are learnable parameters. The new representation is thereby obtained by $\mathbf{H}^T\mathcal{T}$, followed by an addition and normalization operation, and a dense layer (SLP for dimension matching). We also apply a mask in the self-attention map so that $\mathbf{W}^{\mathcal{T}}$ is lower triangular, to ensure forward-in-time causality. The relation between the matrix of original raw CSI representations and the transformed matrix of CSI through attention map $\mathcal{T}$, denoted by $\hat{\mathbf{H}}_{\mathcal{T}}$, becomes:

$$\hat{\mathbf{H}}_{\mathcal{T}} = f_{\mathcal{T}}(\mathbf{H}), \tag{2}$$

where $f_{\mathcal{T}}$ is the temporal transformation function, parameterized by $\Theta_{\mathcal{T}}$ as shown in Temporal Transformation block in Figure 1.

D. Feature Fusion

In the following, we introduce Cross-Dual-Path Attention (CDPA) for achieving high accuracy and scalable classification and feature learning across disparate target domains and applications. Since the spatial and temporal representations from the previous stages are learned in completely separate parameter spaces, naive combining of such representations could make the learning task near impossible as each module tends to update the gradient independently. The CDPA block is added therefore to properly blend representations from temporal and spatial transformation blocks. We show empirically in Section III the importance of this effect and explain its implementation details in the following.

Given two representations denoted by $\mathcal{S}$ and $\mathcal{T}$ as input, the CDPA applies parameterized transformations as $\mathbf{Q} = \mathbf{W}_q^{\mathcal{ST}}\hat{\mathbf{H}}_{\mathcal{T}}$, $\mathbf{K} = \mathbf{W}_k^{\mathcal{ST}}\hat{\mathbf{H}}_{\mathcal{S}}$, and $\mathbf{V} = \mathbf{W}_v^{\mathcal{ST}}\hat{\mathbf{H}}_{\mathcal{S}}$. Formally, let $\Psi_{\mathcal{ST}} = [\mathbf{W}_q^{\mathcal{ST}}, \mathbf{W}_k^{\mathcal{ST}}, \mathbf{W}_v^{\mathcal{ST}}]$ be a set of distinct parameters, the normalized coefficients ϕ_i across K dimensions between the two representations are computed by:

$$(\phi)_{ij} = \frac{1}{\sqrt{N}} \frac{(\mathbf{K}\mathbf{Q}^T)_{ij}}{\sum_{j'=1}^{K}(\mathbf{K}\mathbf{Q}^T)_{ij'}}, \quad \forall i, j \in \{1, \cdots, K\} \tag{3}$$

using which the transformed input vectors are computed as $\hat{\mathbf{h}}_i = \vec{\phi}_i\mathbf{V}$. The relation between input representations $\hat{\mathbf{H}}_{\mathcal{S}}$ and $\hat{\mathbf{H}}_{\mathcal{T}}$ through the CDPA operations, denoted by $\hat{\mathbf{H}}_{\mathcal{ST}}$, becomes:

$$\hat{\mathbf{H}}_{\mathcal{ST}} = f_{\Phi}^{\mathcal{ST}}(\hat{\mathbf{H}}_{\mathcal{S}}, \hat{\mathbf{H}}_{\mathcal{T}}), \tag{4}$$

where $f_{\Phi}^{\mathcal{ST}}$ is the cross-dual-path transformation function, parameterized by Φ.

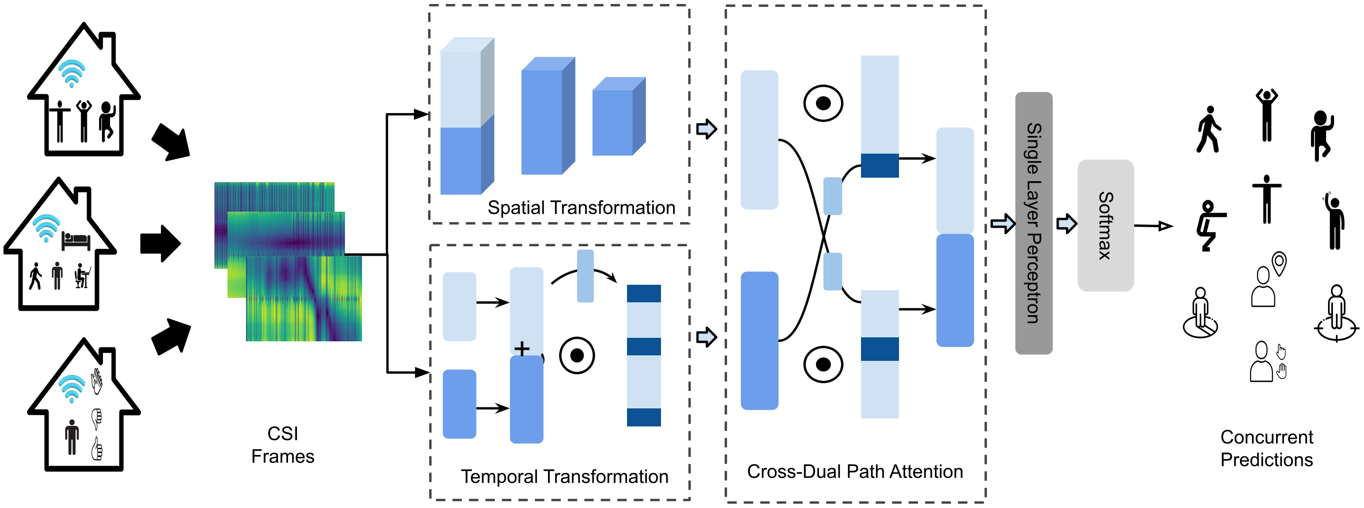

Fig. 1. The CDPA architecture for concurrent CSI-based applications, namely activity, gesture, pose, orientation, and location prediction. It consists of a cross-dual path attention block, trained on CSI datasets across different environments and tasks concurrently using the loss function in Relation (6).

Given $\hat{\mathbf{H}}_{\mathcal{ST}} \in \mathbb{C}^{K\times K}$ as a matrix of cross dual path features, in a single application scenario the class prediction task is performed as:

$$p(y=c|\hat{\mathbf{H}}_{\mathcal{ST}};\Theta') = \frac{\exp(\theta_c'^T\hat{\mathbf{H}}_{\mathcal{ST}})}{\sum_{j=1}^{C}\exp(\theta_j'^T\hat{\mathbf{H}}_{\mathcal{ST}})}, \qquad (5)$$

where y is the model prediction among C classes, θ_c' denotes the weight corresponding to the feature set of class c in the model, and $\Theta' = [\theta_1', \cdots, \theta_C']$ in compact notation.

E. Multi-Application Recognition

Thus far in the discussion, we have provided a complete account of all the components needed in order to implement a full architecture for a single application, such as activity recognition, from raw input CSI to predictive labels. Ultimately, we are interested in supporting multiple CSI-based applications concurrently. In what follows the terms Application and Task are used interchangeably. In multi-task learning, multiple tasks are solved jointly, sharing inductive bias between them [5].

Consider a multi-task learning problem given a large dataset of size N' with input and per-task labels $\{\mathbf{H}_i, y_i^1, \cdots, y_i^T\}_{i=1}^{N'}$, where T is the number of tasks and y_i^t is the label of the t^{th} task for input $\mathbf{H}_i$. For the CDPA model $f(\mathbf{H};\Theta_{\mathcal{S}},\Theta_{\mathcal{T}},\Phi)$ and given task-specific loss function $\mathcal{L}^t(.,.)$, the empirical risk minimization formulation can be written as follows:

$$\min_{\Theta_{\mathcal{S}},\Theta_{\mathcal{T}},\Phi} \sum_{i=1}^{N'}\sum_{t=1}^{T} c_{i,t}\mathcal{L}^t(f(\mathbf{H}_i;\Theta_{\mathcal{S}},\Theta_{\mathcal{T}},\Phi), y_i^t), \qquad (6)$$

where coefficients $c_{i,t}$ weigh contributions from each input to the t^{th} loss component, with no contribution when $c_{i,t} = 0$. Optimizing the above function results in a single network sharing parameters across multiple tasks, which can be performed with GradNorm algorithm [6].

III. Experimentation & Results

In this section, we provide experimental results and evaluate the performance of the proposed algorithm. We train a CDPA model with the following components. We employ a ResNet18 stem for the ST block and a Vision Transformer stem for the TT block with *emb_dim* $=$ 120 , *depth* $=$ 1, and *num_heads* $=$ 1. The model is trained for 100 epoches, with 20% of data held out for test on an Apple M2 computer using the loss function in Relation (6). In particular, we focus on a set of experiments related to $T = 10$ combined tasks with different CSI characteristics for activity, gesture, pose, orientation, and location based on the following datasets. Each dataset comprises varying equipment settings, subjects, environment configurations, and task complexities.

1) Activity Recognition: the datasets used for the task of activity recognition are:

- **StanWiFi** [16] with seven activities including " lie down", "fall", "walk", "run", "sit down", "stand up", and "pick up". These activities were performed twenty times by 6 subjects. Each data frame is 500 (the number of samples) by 90 (the number of subcarriers), by 90 (the number of timestamps).
- **Apartment** [8] with 4 activities including "pickup", "sitdown", "standup", and "walk" performed by a single subject, and a null class where no subject is present. Each data frame is collected over 156 sub-carriers with the dimensions as above.
- **E-EYE** [17] with five different activities, "falling", "standing", "walking", "sitting down" and "standing up" from a chair, and "picking a pen" from the ground, each repeated 20 times. In total, 3,000 samples were collected corresponding to 30 subjects, 5 experiments per subject performed for 20 times each.
- **MultiE** [4], including 6 different activities ("wiping", "walking" ,"moving" ,"rotating" , "sitting", and "stand-

TABLE I
PERFORMANCE ON CONCURRENT TASKS BY THE STATE-OF-THE-ART (SOTA) AND BY THE PROPOSED METHOD WITH - ST: SPATIAL TRANSFORMATION, TT: TEMPORAL TRANSFORMATION, SLP: SINGLE-LAYER PERCEPTRON, CDPA: CROSS DUAL-PATH ATTENTION. METRICS REPORTED ARE ACCURACY (ACC) IN PERCENTAGE, AS WELL AS PRECISION (P), RECALL (R), AND F1-SCORE (F1) BETWEEN 0 AND 1.

Dataset	SOTA	ST+SLP				TT+SLP				ST+TT+SLP				ST+TT+CDPA			
	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	ACC	P	R	F1	Acc
StanWiFi [7]	98	0.84	0.80	0.82	84	0.85	0.81	0.82	85	0.92	0.93	0.92	96	0.94	0.96	0.95	97
Apartment [8]	98	0.98	0.98	0.98	98	1.00	1.00	1.00	100	1.00	1.00	1.00	100	1	1	1	100
E-Eye [9]	94	0.77	0.74	0.74	77	0.83	0.81	0.77	83	0.89	0.90	0.89	90	0.93	0.93	0.93	94
MultiE [10]	71	0.48	0.46	0.45	48	0.89	0.88	0.88	89	0.87	0.87	0.87	87	0.90	0.90	0.90	90
NTU-Fi [11]	99	0.96	0.95	0.95	96	0.91	0.88	0.88	91	0.94	0.93	0.93	93	0.96	0.96	0.96	96
Widar [12]	92	0.95	0.95	0.95	95	0.45	0.27	0.20	45	0.98	0.97	0.97	97	0.96	0.96	0.96	96
SignFi [13]	98	0.99	0.99	0.99	99	0.99	0.99	0.99	99	0.99	0.99	0.99	99	1	1	1	100
mmWPose [14]	95	0.68	0.73	0.66	68	0.87	0.85	0.84	87	0.99	0.99	0.99	99	0.98	0.98	0.98	99
DirWiFi [15]	92	1.00	1.00	1.00	100	1.00	1.00	1.00	100	1.00	1.00	1.00	100	1	1	1	100
FP-Loc	-	0.96	0.95	0.95	94	0.91	0.86	0.87	91	1	1	1	100	1	1	1	100
Average:	93	0.85	0.84	0.84	85	0.86	0.83	0.82	86	0.95	0.95	0.95	96	0.96	0.96	0.96	97

ing up"), performed by 6 different subjects across 100 different environment settings (e.g. furniture and location variations).

2) Gesture Recognition: the datasets used for the task of gesture recognition are:

- **NTU-Fi** [13], comprising classes "running", "walking", "falling", "boxing", "circling arms", and "cleaning" the floor—performed by 20 volunteers repeated 20 times. Each data frame spans a duration of 1 second of size 3 by 114 by 500.
- **Widar** [12] - CSI dataset designed for human gesture recognition, participants performed 22 distinct gestures, including drawing numbers from 0 to 9 in the horizontal plane. The dataset includes an extensive collection of 43,000 samples, providing a robust foundation for studying gesture recognition in diverse scenarios.
- **SignFi** [18] - designed for CSI-based sign language gesture recognition focusing on 150 different gestures by five participants in two environments. Each data frame has dimensions of 3 × 30 × 200.

3) Pose Estimation: the datasets used for the task of pose and orientation recognition are:

- **mmWPose** [14] - the dataset for CSI-based pose estimation with three participants (two males and one female, with varying body shapes and heights) performing a set of eight poses: "Arms up", "Left hand up", "Right lean", "Right hand up", "Left lean", "Empty", and "Arms wide". Each pose was held for 15 seconds, and each participant performed 20 rounds of the pose set, yielding approximately five minutes of data per pose.
- **DirWiFi** [1]- the dataset for CSI-based direction estimation with five distinct hand gestures: "drawing a circle, "crossing hands", "clapping", "raising hands", and "lowering hands", corresponding to actions like starting, stopping, switching context, increasing, and decreasing. Gestures were performed in 24 directions sampled at 15° intervals, with a natural deviation of approximately ±5° in gesture direction. Three volunteers performed the five gestures 40 times in each of the 24 directions.

4) Fingerprinting / Localization: the dataset used for the task of localization based on CSI finger printing is:

- **FP-Loc** [19] - the datasets for CSI-based localization collected in two distinct indoor environments: a laboratory and a meeting room, to represent both Non-Line-of-Sight (NLOS) and Line-of-Sight (LOS) scenarios, consisting of 317 locations spaced 50 cm apart and of 176 locations spaced 60 cm apart, respectively.

For each dataset we report the accuracy for the task achieved by the best state-of-the-art (SOTA) method. In addition to the proposed CDPA model with a ResNet stem (ST) and Transformer stem (TT) referred to as ST+TT+CDPA, we evaluate each task using the ST stem only, the TT stem only, and using a single-layer perceptron (SLP) in place of the CDPA block (ST+TT+SLP). Along with the accuracy of the predictions (ACC), we also report precision (P), recal (R), and f1-score (F1) for the models.

Table I summarizes the above experimental results. For the activity recognition tasks, the CDPA with ST and TT stems (i.e. ST+TT+CDPA) achieves 97, 100, 94, and 90 percent accuracies on StanWiFi, Apartment, E-Eye, and MultiE datasets on par with or improving the state of the art results. Without the CDPA where the output representations from ST and TT are concatenated and passed through a SLP (i.e. ST+TT+SLP), the performance degrades indicating the importance of cross-dual path component. Using either ST or TT individually, corresponding to a single ResNet18 or a single Vision Transformer, results in relatively lower performance specifically on MultiE dataset due to varying environment configurations. The addition of CDPA on the other hand proves robust to these variations.

For the gesture recognition tasks, the full CDPA model outperforms the state of the art methods in precision, recall, and f1-scores consistently on NTU-Fi, Widar, and SignFi datasets. Across pose and orientation tasks with accuracies 99% and 100%, improvements by 4 and 8 percent are achieved on mmWPose and DirWiFi respectively. In terms of CSI-

[1] https://gitlab.com/yuxiqin/direction-independent

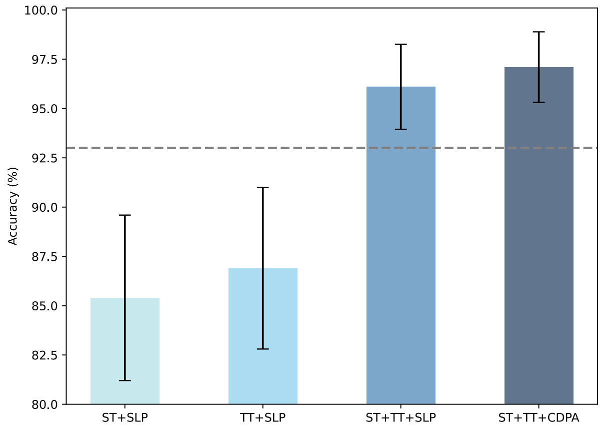

Fig. 2. Improvements in concurrent accuracy (mean and standard deviation) through different model combinations, including Spatial Transformer (ST), Temporal Transformation (TT), Single Layer Percept (SLP), and Cross-Dual Path Attention (CDPA). The average state-of-the-art accuracy across all tasks is at 93% shown by the dashed line.

based finger printing, the accuracy of ST+TT+CDPA is at the highest. The accuracy of 96% by SS+TT+CDPA is lower than the best reported accuracy of 99% for NTU-Fi, the only task where the concurrent accuracy is not improved. Across all tasks, the proposed model however outperforms the state-of-the-art with 97% accuracy and 0.97 f1-score on average. Figure 2 shows concurrent performance across all tasks as compared to baseline average of 93% by the state of the art methods.

IV. Conclusions

The presence of WiFi infrastructure in indoor environments presents an opportunity for non-intrusive applications based on the channel state information. In this work, we present a novel architecture which leverages the spatial and temporal features in the CSI for concurrent processing and performance across tasks related to activity, gesture, pose, orientation and location estimation in smart environments using passive devices. The model utilizes a cross-dual path attention mechanism to fuse features and is trained on a diverse set of CSI-based tasks resulting in improved performance over the state of the art methods. In particular, the model achieves 97% performance accuracy when trained and tested on a set of 10 concurrent CSI-based tasks.

References

[1] J. A. Armenta-Garcia, F. F. Gonzalez-Navarro, and J. Caro-Gutierrez, "Wireless sensing applications with wi-fi channel state information, preprocessing techniques, and detection algorithms: A survey," *Computer Communications*, vol. 224, pp. 254–274, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366424002214

[2] J. Ma, H. Wang, D. Zhang, Y. Wang, and Y. Wang, "A survey on wi-fi based contactless activity recognition," in *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications*, 2016, pp. 1086–1091.

[3] N. Damodaran, E. Haruni, M. Kokhkharova, and J. Schäfer, "Device free human activity and fall recognition using wifi channel state information (csi)," *CCF Transactions on Pervasive Computing and Interaction*, vol. 2, no. 1, pp. 1–17, 2020.

[4] S. Ding, Z. Chen, T. Zheng, and J. Luo, "Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 517–530.

[5] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul. 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[6] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks," Jun. 2018, arXiv:1711.02257 [cs]. [Online]. Available: http://arxiv.org/abs/1711.02257

[7] S. Mekruksavanich, W. Phaphan, N. Hnoohom, and A. Jitpattanakul, "Attention-based hybrid deep learning network for human activity recognition using wifi channel state information," *Applied Sciences*, vol. 13, no. 15, p. 8884, 2023.

[8] S. Mehryar, "Location-aided activity recognition from channel state information with deep cross-modal learning," in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2022, pp. 1–5.

[9] M. S. Islam, M. K. A. Jannat, M. N. Hossain, W.-S. Kim, S.-W. Lee, and S.-H. Yang, "Stc-nlstmnet: An improved human activity recognition method using convolutional neural network with nlstm from wifi csi," *Sensors*, vol. 23, no. 1, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/1/356

[10] S. Ding, Z. Chen, T. Zheng, and J. Luo, "Rf-net: a unified meta-learning framework for rf-enabled one-shot human activity recognition," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, ser. SenSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 517–530. [Online]. Available: https://doi.org/10.1145/3384419.3430735

[11] J. Yang, X. Chen, D. Wang, H. Zou, C. X. Lu, S. Sun, and L. Xie, "Sensefi: A library and benchmark on deep-learning-empowered wifi human sensing," *Patterns*, vol. 4, no. 3, 2023.

[12] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3.0: Zero-effort cross-domain gesture recognition with wi-fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8671–8688, 2022.

[13] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, "Efficientfi: Toward large-scale lightweight wifi sensing via csi compression," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 13 086–13 095, 2022.

[14] N. N. Bhat, J. Sameri, J. Struye, M. T. Vega, R. Berkvens, and J. Famaey, "Multi-modal pose estimation in xr applications leveraging integrated sensing and communication," in *Proceedings of the 1st ACM Workshop on Mobile Immersive Computing, Networking, and Systems*, ser. ImmerCom '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 261–267. [Online]. Available: https://doi.org/10.1145/3615452.3617944

[15] Y. Qin, S. Sigg, S. Pan, and Z. Li, "Direction-agnostic gesture recognition system using commercial wifi devices," *Computer Communications*, vol. 216, pp. 34–44, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366423004747

[16] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, 2017.

[17] A. Baha'A, M. M. Almazari, R. Alazrai, and M. I. Daoud, "A dataset for wi-fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments," *Data in Brief*, vol. 33, p. 106534, 2020.

[18] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using wifi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018. [Online]. Available: https://doi.org/10.1145/3191755

[19] X. Zhu, T. Qiu, W. Qu, X. Zhou, M. Atiquzzaman, and D. Wu, "BLS-Location: A Wireless Fingerprint Localization Algorithm Based on Broad Learning," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 115–128, 2023.