Department of Mathematics and Computer Science
Interconnected Resource-aware Intelligent Systems Research Group

Towards domain agnostic Wi-Fi CSI gesture classification

Master's Thesis

Yvan Putra Satyawan

Supervisors:
Prof. Dr. Ir. Nirvana Meratnia
Bram R.D. van Berlo, M.Sc
Dr. Maryam Tavakol

Second Draft

Eindhoven, August 2023

Contents

Chapter 1

Introduction

In this ever more connected climate that we find ourselves in, IoT devices everywhere are adding little conveniences to our every day lives. With IoT devices becoming ever more common and reaching a forecasted 27 billion devices by 2025 [12], the dream of ubiquitous computing and sensing is transitioning from a mere dream to the reality of our every day lives. Additionally, approximately 19% of new devices bought in 2020 also utilize some form of Wi-Fi radio for communications with a forecasted increase to 24% by 2025.

It is clear that this trend towards integrating computing technology into everyday objects will only accelerate in the future. The increased convenience and efficiency may be the biggest boon of such technologies. For example, smart thermostats can predict heating requirements and adjust accordingly, leading to lower heating costs in a house while maintaining the convenience of having a well heated space.

With all these connected devices becoming and edge computing ability comes ubiquitous sensing, enabling new modalities of interaction and improving data collection and ana-

Figure 1.1: A WiZ A60 Tunable White light bulb, one example of an IoT smart device with an integrated Wi-Fi Radio [29].

lytics. It is now possible to envision households with complete presence detection coverage, for example, through the use of smart motion sensors, enabling increased efficiency by intelligently identifying which rooms require heating and lighting and which do not. Always-on voice control systems, such as Amazon Alexa and Google Assistant speakers are also increasingly common, making a connected AI-assistant only one call away. The always-connected nature of these sensors also enable the gathering and analysis of vast amounts of user data, potentially providing valuable insights into various aspects of our lives.

One challenge that has continued to plague ubiquitous devices is the lack of a ubiquitous user interface which does not require input devices. It is not too common to use the end device itself as the input. Even in the case of voice control systems, a dedicated smart speaker is still required and must be placed in every room from which interaction is desired. For example, in the case of smart lights or smart thermostats, the end device would be the light bulbs and heating system, respectively. In both cases, a separate control unit, the light switch or thermostat, respectively, is still required.

With these challenges in mind, Wi-Fi-based sensing provides one potential solution. Many IoT devices already contain some sort of Wi-Fi radio, such as the Phillips Hue Light bulb in Figure 1.1, and it would be a rather safe assumption to make that spaces with IoT devices would also have some sort of Wi-Fi infrastructure in place as well. With this in mind, the idea of a gesture-based interface based around Wi-Fi becomes rather appealing. The use of Channel-State Information (CSI) also enables of finer-grain signals to be extracted from consumer-grade Wi-Fi radios, enough to enable reliable gesture recognition [1]. This method does, though, suffer from the domain-shift problem, achieving the best accuracy only in cases with a prediction model fine-tuned to a specific person and environment is used.

With this thesis, we aim to explore the use of CNN architectures and domain-shift mitigation methods to improve the state-of-the-art in Wi-Fi CSI-based gesture classification. Specifically, we will look at using preprocessing methods to transform the input signal from CSI into an image using table-to-image and signal-to-image transformations, the use of traditional signal processing algorithms to process the incoming CSI signal, and the use of Reinforcement Learning (RL) to perform domain auto-labeling and provide the CNN classifier with additional information.

1.1 Context and Background

IoT devices are without a doubt increasingly prevalent in everyday life. The Atlas building at the Technical University of Eindhoven (TU/e), for example, uses centrally networked lights for all of its lighting fixtures powered through Power over Ethernet (PoE) and this is at least partially credited as a reason the building has the best efficiency of any academic building in Europe when it was constructed [10].

All over the building, presence detection, in the form of motion detection sensors, is also used to automatically set appropriate lighting levels for each room. This is just one example of how ubiquitous computing and sensing has now entered the mainstream and is no longer a dream of a few enthusiasts and IoT evangelists. Developments in AI and big data processing has also made the usefulness of ubiquitous computing much more evident, legitimizing its use in everyday objects.

Finally, the deployment of 5G networks in densely populated areas is working towards enabling faster speeds and lower latency, essential in many ubiquitous computing and sensing applications.

With these advances, the question has shifted towards what sort of interface we should utilize to provide an always-available non-intrusive experience for the users. One possible solution is gesture-based interfaces. With any ubiquitous computing and sensing product, especially in the consumer space, minimal setup on the user's part is desired; otherwise the product will not become something which is widely accepted and used.

Wi-Fi gesture recognition can solve these issues, providing a gesture-based interface requiring potentially zero additional setup requirements. As a bonus, this would also be a low-cost solution which many IoT devices already having the necessary hardware for regardless.

There also exists a task group for Wireless Local Area Network (WLAN) sensing, called 802.11bf, within the IEEE 802.11 working group, the group which sets the standards for WLAN, with members from large companies including Huawei, Qualcomm, and Meta [9]. This shows there is genuine interest in the industry to utilize WLAN for these purposes. With an approval date set for September 2024, it is clear that WLAN sensing is not just some theoretical possibility confined to a lab, but rather a very real technology that may soon become widespread.

To make such an approach possible, we utilize machine learning (ML) to process the incoming CSI data and classify user gestures. Wi-Fi technology, while being very different to radar technology, are both forms of radio technology. Although topologically very different, it naturally, or not so naturally, follows then, can Wi-Fi be used for remote sensing? The answer to this question, according to [1] and [7], the answer is a resounding yes!

There are even commercial products which have become available on the commercial market during the writing of this very thesis that incorporates parts of this technology. Signify's line of WiZ smart lights uses *SpaceSense* technology, which is a proprietary algorithm that detects motion through "disturbances [in the Wi-Fi signal] created by a person's presence" and is available through a software update [28]. It should be noted, though, that this is only capable of presence detection and not of gesture recognition.

However, ML suffers from degraded performance when faced with domain-shifts. When dealing with new, unseen users and environments, gesture classification accuracy degrades significantly [7, 1, 22, 2, 13, 15, 39, 14, 18]. As such, factors to mitigate this domain-shift problem are required in any implementation outside of a pristine laboratory setting.

For the purposes of our thesis, the Wi-Fi information we utilize for gesture classification is known as the Channel State Information (CSI). This comes in the form of a set of complex numbers denoting signal differences for each receiver access point (AP) from each transmitting AP. This complex number can then be represented as a set of amplitudes, or radii, and phase shifts, or angles, when considered as polar coordinates on the complex plane. CSI itself is a description of the multipath effects of a signal traveling from the transmitting AP to the receiving AP. In the realm of WLAN, the estimated CSI of incoming signals is used to correct for these multipath effects, making it possible for the system to adapt to current environmental conditions. Importantly, human activity in an environment also affects the CSI, making it possible to infer activity through CSI.

1.2 Motivation

There are significant potential commercial and practical benefits to enabling domain-agnostic Wi-Fi gesture classification.

The upcoming release of the 802.11bf standard, which standardizes the requisite hardware technology for Wi-Fi sensing, marks a pivotal moment from a commercial standpoint. To fully exploit these emerging technologies, it is of utmost importance to investigate their practical applications.

Wi-Fi sensing is also unique in that the hardware for it is already commercially available and widespread, though not necessarily standardized. As such, it is in a unique place where it is an almost completely software-based solution for preexisting, mass-adopted hardware enabling completely new modes of interaction. This makes it unique and, ultimately, much more commercially desirable as a technology to be adopted.

With respect to domain-agnosticism, the motivating factor is that the largest hurdle for Wi-Fi gesture classification is its loss of performance in new, unseen domains. A model that cannot adapt to such issues will inevitably be nonviable for commercial adaptation lest the end-user be required to perform the same gesture hundreds of times in various positions after installing every single smart IoT device.

We also wish to further the state-of-the-art in domain-agnostic models. The results of this thesis is not only applicable to the field of Wi-Fi gesture classification. Domain-agnostic systems are required for many industry applications, such as for patient sensing in healthcare, for consumer behavior tracking in retail, and for hands-free control of smart IoT devices in consumer homes. The conclusions of this thesis will hopefully be able to advance the field and provide new insight into what future avenues of research may end up being fruitful.

Finally, a Sci-Fi future where all your devices are controlled through your thoughts might seem like a dream, but in reality, we are not too far from this future. The use of Wi-Fi sensing to detect gestures is one step closer to this futuristic dream and in addition to all the potential commercial and practical benefits this technology could bring, the sheer "coolness" of the technology should not be underestimated as a motivation to perform research.

1.3 Problem Statement

Models already exist to classify gestures through Wi-Fi CSI data. Ideally models are solely influenced by those factors which contribute to the correct classification of the gesture, this is, in reality, not the case. These models are susceptible to the influence of "domain factors", encompassing variables such as the identity of the gesturing subject and the surrounding environment in which the gesture is performed.

These domain factors cause feature domain shifts between the training data and the data encountered during actual use or inference despite both domains containing the phenomenon of interest, for example the gesture we wish to classify. Due to these shifts, performance of the model is degraded [7, 1, 22, 2, 13, 15, 39, 14, 18]. It is thus interesting to be able to create a domain-agnostic model whose output is independent of the aforementioned domain factors.

There exist proposals to mitigate this degradation including using very large datasets. This is one method especially championed by large companies such as Tesla and OpenAI which have vast resources at their disposal. However, the gathering of such large datasets is only feasible for specific scenarios. In the case of Tesla, for example, having a large fleet of vehicles capable of recording what is essentially supervised training data makes it possible to collect the vast amounts of data required to build a dataset usable for the training of self-driving vehicles [32]. Various OpenAI projects, on the other hand, simply scrape large amounts of websites and collect these as part of their dataset [6]. Within the scope of our research, due to its nature, such widespread data collection methods are non-viable, and we must resort to more novel approaches towards data-agnosticism.

Most alternative methods to simply using very large datasets rely on a ground truth domain label being provided [15, 36]. In these approaches, a "discriminator" network is used to predict domain labels from the latent embedding of the data and an adversarial training procedure is then used. The "generator" which produces these latent embeddings must thus generate embeddings which contain no domain information while still maintaining enough information that a classifier model can use its output to correctly classify target features.

In [18], a method is presented which does not require manually labeled domain labels to be provided. Instead, a CNN encoder and state machine neural network are used and their output is fed into a recurrent neural network (RNN) to provide a classification. The RNN is trained through reinforcement learning (RL) to predict features correctly and independently of domain factors.

We hypothesize that we can extend the RL component of [18] to facilitate domain auto-labeling and eliminate the use of a state machine neural network by using signal-to-image preprocessing methods. Towards these goals, we specify our research questions as follows:

1. To what extent can a reinforcement learning agent utilize the latent signal representation produced by the CNN to accurately produce a latent representation of the domain space, measured by performance metric difference in a domain factor leave-out cross validation setting between a classifier provided the domain space representation and one which has not?

2. To what extent would changing the latent representation of the domain space from a one-hot encoding to a probability measure affect the performance of the classifier, measured by the performance metric difference in a domain factor leave-out cross validation between both domain space representation types?

3. To what extent can signal-to-image transformation replace the state machine neural network presented in [18], measured by comparing the performance metric difference in a domain factor leave-out cross validation setting between the model presented in [18] and our approach provided no self-label?

The rest of this thesis will present a literature review, background on required knowledge, the proposed methodology, experimental results, and discussions and conclusions of those results.

Chapter 2

Literature Review

In this chapter we review important works in the literature which form the foundation of this thesis. We first discuss the initial set of works which cover Wi-Fi activity detection as well as other related works which do not use CSI data specificially before discussing those works which do utilize CSI data and publicly available datasets for this purpose. We then look at various signal-to-image transformation methods which may be applied to time-series signal data, enabling the use of techniques from the image processing domain. Finally, we look into domain shift mitigation methods and specifically reinforcement learning for domain shift mitigation.

2.1 Wi-Fi for Activity Detection

Past works have investigated the use of Wi-Fi signals generally for the purposes of activity detection.

The first work regarding the use of Wi-Fi signals for the detection of humans subjects we could find is the work of Chetty, Smith, and Woodbridge in 2012 [7]. Their work utilized passive Wi-Fi signals propagating through a building with receivers placed outside the building for presence detection. This method achieved reasonable results and proved that Wi-Fi signals could be used to detect human presence in buildings, although it required the indoor and outdoor APs to be synchronized through wires and was unable to detect precise activities of the human subjects.

The first work we could find discussing the use of Wi-Fi for activity detection is the work from Fadel Adib and Dina Katabi, published in 2013 [1]. This work shows the potential of using signals which could be produced by Wi-Fi APs to detect human activity from through a wall. The most important idea in this work is the elimination of the radio “flash” which comes with the signal hitting a wall and bouncing back towards the transceiver. Their work focused more on the radar technology implications and not on the use of consumer Wi-Fi APs for gesture detection. They did, though, show that using matched filters was enough to perform rudimentary gesture recognition, given coarse enough gestures.

In the same year, a different group published a paper showing how to use signals in the 2.4 GHz range, i.e., compatible with Wi-Fi transceivers, for simple gesture detection using Doppler shift identification [22]. This paper proposes the use of a narrowband pulse with a very narrow bandwidth of only a few Hertz and detecting the Doppler shift from the returned signal. Using this method, the researchers were able to identify 9 different gestures with a claimed 94% accuracy.

The same group as [1] also published a separate paper in 2014 detailing the use of a custom-built Wi-Fi based device which could detect course body motions by leveraging the geometric position of its antennas and measuring values through a Time of Flight (ToF) approach [2].

Finally, it is also important to note that IEEE has a task group 802.11bf assigned specifically to standardize Wi-Fi sensing technologies [9]. This group is focused on standardizing the hardware requirements, specifically enabling CSI accessibility and specific measurement procedures which future devices can implement. Their target is to standardize these requirements for future devices both in the sub-7 GHz range and in the 60 GHz range. They additionally provide suggestions for what methods can be then be used to interpret the data provided, including the use of Fast Fourier Transform (FFT) algorithms to calculate a Channel Impulse Response from CSI and a Doppler FFT, which may be directly used for gesture recognition. The standards for 802.11bf is set to be ratified and published by September 2024.

2.2 Wi-Fi CSI for Gesture Recognition

A selection of works which utilize Wi-Fi CSI data specifically for the task of gesture recognition are discussed in this section.

To the best of our knowledge, the first work discussing the use of CSI for gesture recognition was published in 2015 by He et al. [13]. This work looks into the use of CSI and outlier detection to detect gestures, achieving 92% gesture recognition accuracy on four gestures in a line-of-sight experiment and 88% accuracy in a non-line-of-sight experiment.

The 2019 work titled Person-in-WiFi from Wang et al. proposes the use of an array of three transmitter and three receiver antennas to directly predict body segmentation and pose estimation of persons located in between the aforementioned antennas [34]. In this work, they used an RGB camera to provide ground truth annotations. The ground truth body segmentation masks were generated using Mask-RCNN while the Body-25 model of OpenPose was used for pose detection. This work, shows that body segmentation and pose estimation is possible with only CSI data, achieving an mAP of 0.38 for body segmentation and around 0.1 meter error for join estimation. Qualitatively, the results are quite impressive and it is clear that at the very least, the model performs well given that its input data is one-dimensional.

Using the same dataset, Geng, Huang, and De La Torre published DensePose in 2022 performs similarly, but instead produces UV coordinates of the subjects. This work also provides some interesting preprocessing steps on the raw CSI data to improve prediction performance.

WiGAN, published in 2020, proposes the use of a Generative Adversarial Network (GAN) to augment training data using the generator as well as using the discriminator as a feature fusion and extraction module [14]. The output of the discriminator is then used fed into a Support Vector Machine (SVM) which classifies then gesture seen in the input data. The authors claim that by using the discriminator, which fuses together multiple layers of a CNN module through yet another convolution, their method is able to “learn and recognize the importance of different depth features by itself”. On publicly available datasets, WiGAN indeed achieves better results than the competing methods at the time, achieving 98% accuracy on Widar 3.0 with known subjects and ≈8–10% lower performance on new domains.

The same group, in the same year, published DeepMV which proposes the use of multiple APs and audio sources, in the form of ultrasound signals, with a domain discriminator and embedding generator [36]. This work is based on the intuition that the fusion of sensor data from multiple APs and audio sources placed around the room, which intuitively should provide more data. The embedding generator produces a latent representation of the action which can then be used by a fully connected module to classify the action being performed while the domain discriminator uses the latent representation to predict the domain of the action. A minimax game is played between the embedding generator and domain discriminator to minimize the maximum accuracy of the domain discriminator while maximizing the minimum accuracy of the action classifier. On their self-collected dataset, they were able to achieve 83.7% mean accuracy, outperforming all other benchmarked methods.

The 2021 paper by Zhuravchak et al. proposes the use of an LSTM as a classifier [40]. In this method approach, CSI data is provided as an input with a fixed length and the LSTM provides a single output representing the action detected within the input window. Their method achieves 87.8% accuracy on a self-collected dataset.

Ma et al., published in 2021, proposes the use of a CNN and neural network state machine encoder and a LSTM trained using RL to eliminate the need of domain specific information [18]. This work also contains a curated benchmark of many previous works in this field. Their proposal is the use of a neural network state machine relies on the assumption that there is a temporal dependency within and across CSI segments. For example, it is unlikely that a person who is currently standing will immediately be sitting within the next CSI segment without a sit-down transition segment in between. Although unlikely, the probability is non-zero, due to incontinuities in the data and mislabeling in the ground-truth labels, and thus a neural network state machine is used. The results show >97% mean accuracy on in-domain test samples with a drop of 14–17% on out-of-domain samples.

Additionally, a few bachelors thesis' fromt he past years at the IRIS research cluster at TU/e have focused on this problem as well. The 2022 thesis by van den Biggelaar proposes the use of reinforcement learning with Deep Q-Networks as the gesture classifier [5]. Their result shows $\sim$88% mean accuracy on Widar 3.0, dropping by $\sim$4% on out-of-domain test samples.

The 2022 thesis by Oerlemans compares how different preprocessing methods appear to affect gesture recognition performance [21]. Specifically, signal filtering through a finite impulse response filter and phase unwrapping, transformation to a Doppler frequency spectrum (DFS), and transformation using Gramian Angular Fields (GAF) was explored. On the SignFi dataset, they show that each of the above steps do indeed significantly improve model performance with the GAF transformation coupled with signal filtering resulting in the best performance.

2.3 Wi-Fi CSI datasets for Gesture Recognition

Three Wi-Fi CSI datasets for gesture recognition were considered for this thesis and they are each explained in this section.

The Widar 3.0 dataset, forthwith referred to as the Widar dataset for brevity, presents a dataset with developed specifically for "cross-domain learning solutions" [39]. The solution provided in this dataset is two-fold: 1) the (relatively) high number of domains that the data was collected with, and 2) the proposal of Body-coordinate Velocity Profile (BVP) with is a theoretically domain-independent representation of the data. More details of BVP is discussed in Section . Finally, this paper also provides a baseline model to compare against.

Desecribe in background section

SignFi is a dataset of Wi-Fi CSI data specifically for sign language recognition [19]. This dataset contains over 276 sign gestures in a lab and home environment with five different users. A baseline CNN model is also presented in this paper, capable of achieving a $\sim$87%

mean accuracy over 150 sign gestures.

Person-in-WiFi is the dataset used in the paper by Wang et al. [34]. This dataset was made public and includes both Wi-Fi CSI data and RGB camera data of the activity from a fixed position. This dataset was collected specifically for pose estimation solutions and is not meant specifically for activity recognition, as no activity labels are provided in the dataset.

2.4 Signal-to-Image Transformations

Different preprocessing methods have been investigated to transform raw tabular data into images for deep learning. Four state-of-the-art approaches are DeepInsight [27], REFINED [3], GAF, and MTF [35]. A search of the current body of literature did not yield any research into a direct comparison of these techniques on a common dataset. Instead, a previous unpublished work by the author of this thesis for the Seminar course at the TU/e has shown that these four methods performed best among state-of-the-art signal-to-image transformations [24]. A more thorough description of each of these methods are described in Section

Describe in background section

2.5 Domain Shift Mitigation Methods

There are a number of papers focusing on domain shift mitigation methods. A selection of these papers, specifically those which are highly related to our problem domain, are discussed in this section.

The 2021 paper by Zinys et al. focuses on the use of GANs for domain shift mitigation, called Adversarial Domain Adaptation (ADA) [41]. In this work, the discriminator attempts to predict gesture and domain while the generator produces sample data. The discriminator is provided a loss function based on ground truth data and accurately identifying generated data while the generator produces sample data which is in the same domain and gesture as its provided input. Their results, tested on Widar show significantly better performance than the baseline Widar model on unseen domains.

Zhang et al., in 2022, proposes the use of federated learning for domain shift mitigation [37]. The concept is to allow for each user to train their own neural networks and using matched average federated learning to combine all user models together. Tested on the Widar dataset, they show performance competitive with state-of-the-art techniques.

Van Berlo et al. discusses attempts at using mini-batch alignment to generate domain factor independent latent representations of the data [4]. They showed that unfortunately, the proposed mini-batch alignment pipeline did not lead to better performance across domains. The authors believe this may be due to a lack of sufficient domain factor information, leading to poor mini-batch alignment. Alternatively, the assumptions of the underlying probability distributions may be incorrect. In any case, it seems that this method, given current publicly available datasets, does not improve the SOTA.

Finally, the 2022 bachelors thesis by Sips investigates the use of network pruning and quantization for domain shift mitigation [30]. The basic concept is to improve model robustness by enforcing sparsity and utilizing mixed precision training. Tested against a baseline where sparsity and mixed precision training was not used, the modified networks performed slightly worse on in-domain test samples while they had mixed results on out-of-domain samples.

2.6 Domain Shift Mitigation using Reinforcement Learning

The following section contains some selected works in domain shift mitigation, mainly those related directly with RL of self-supervised techniques.

Zhang et al. in 2021 published research into using RL based feature selection for domain shift mitigation [38]. The proposed method would be able to select the most relevant features across two domains by employing Q-learning to learn policies for feature selection, utilizing the performance of a domain discriminator as its reward function. By doing so, they attempt to align the feature manifolds between both domains and produce domain invariant features in the vision field through the use of reinforcement learning. This also has led to our own hypothesis that reinforcement learning can be used for domain auto-labeling. Benchmarked on publicly available datasets, this method achieves the best mean accuracy among SOTA methods.

Martini et al. published a technique called "Domain-Adversarial Neural Networks" in 2021 [20]. This technique uses a feature extractor, a domain classifier, and a label predictor. The feature extractor maps the input to a latent representation, the domain classifier predicts the domain given the latent representation, and the label predictor provides a class prediction given the extracted features. The loss function of the model balances the label predictor loss, using Cross Entropy Loss, and the domain classifier loss, also using Cross Entropy Loss, with the goal of providing a latent representation from the feature extractor which is domain-invariant, yet still discriminative such that the domain classifier is still capable of accurately classifying the domain. Additionally, the paper proposes the use of Maximum Mean Discrepancy (MMD) to measure the difference between two domains. The MMD measures the kernel-based difference between feature means of each domain. This work proposes the use of MMD with the goal of minimizing this distance while maintaining the distinctness of each domain, allowing for the domain classifier to still work.

inline]Add the discrete wavelet transform paper ”A WiFi-based Smart Home Fall Detection System using Recurrent Neural Network.” [8] [

Chapter 3

Background Knowledge

In this chapter, we will explore some of the background knowledge required to understand the methodology chosen and experiments undertaken throughout this thesis. We will first explore Variable Autoencoders before delving into some background about Reinforcement Learning and finally concluding with a discussion of Wi-Fi Channel State Information (CSI) and Body-coordinate Velocity Profiles (BVP).

3.1 Variational Autoencoders

Variational autoencoders (VAEs) are a type of generative model which combines elements of both autoencoders and probabilistic modeling and are suited for unsupervised learning and representation learning and was first introduced by Kingma and Weling in [16]. Standard autoencoders are a network architecture in which data is encoded into a lower-dimensional latent space by an encoder. A decoder then reconstructs the original input from the latent representation. VAEs introduce probabilistic modeling to autoencoders by using a probability distribution over the latent space instead of using the latent space directly.

Its original design aimed at using a generative model as an implicit form of regularization. By forcing the model to learn a representation which is also useful for data generation, the representation learned must have some sort of statistically independent but meaningful representation of the variations in the input data, leading to better performance at both the auxiliary task of data generation as well as the main task of discrimination [17].

Generally, VAEs are described as two coupled but independent models: the encoder and decoder. The encoder is a Bayesian network of the form $q(\boldsymbol{z}|\boldsymbol{x})$ where $\boldsymbol{z}|\boldsymbol{x}$ may be a (deep) neural network. Similarly, the decoder is also a Bayesian network of the form $p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ where $\boldsymbol{x}|\boldsymbol{z}$ may also be a (deep) neural network. The input signal $\boldsymbol{x}$ is thus represented by $\boldsymbol{z}$.

The purpose of the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ with parameters ϕ is to produce an approximation of the true, but intractable, posterior. The encoder neural network is then used to produce the set of parameters for latent variables such that

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}) = \text{EncoderNeuralNet}_\phi(\boldsymbol{x}) \tag{3.1}$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma})) \tag{3.2}$$

where $\mathcal{N}$ is the normal distribution.

The purpose of the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is to produce a mapping between the latent space $p_\theta(\boldsymbol{z})$ and the original, observed distribution through learning a joint distribution $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})$.

The VAE is optimized through the evidence lower bound (ELBO), which incorporates Kullback-Leibler divergence, the details of which are thoroughly explained in [17].

By using the reparameterization trick introduced in [16], the ELBO can then be differentiated with respect to both parameters ϕ and θ at the same time through stochastic gradient descent.

VAEs are used in various areas including image generation, data compression, denoising, and image recognition. By using a latent representation of the data which is statistically meaningful, the generated data is much more likely to come from the same underlying distribution as the original data.

3.2 Reinforcement Learning

Reinforcement learning (RL) is a machine learning paradigm which teaches agents how to make decision in complex environments [31]. The general paradigm is that of an agent which takes actions in an environment as a response to its observations of the current environment state. The environment then provides a reward for the agent and transitions into a new state as a response to the action. The agent is given the goal of maximizing the cumulative reward over time. The main use cases of RL are in robotics, gaming, finance, and healthcare where performing complex tasks or optimizing control systems is necessary.

Throughout this work, we focus on model-free RL. Model-free RL is an approach to RL which focuses on learning directly from interactions with the environment without necessarily modeling the dynamics of the environment explicitly. With this approach, the agent learns policies or value functions which it uses to make decisions. Some well-known model-free RL algorithms include Q-Learning, State-Action-Reward-State-Action (SARSA), Deep Q-

Networks (DQN), Proximal Policy Optimization (PPO), and Asynchronous Advantage Actor-Critic (A3C).

Furthermore, there are two main learning paradigms to RL, namely on- and off-policy learning.

On-policy RL learns directly from data collected by the current policy. This entails updating the policy directly using data, i.e., the reward gathered, of the same policy. This makes it especially suitable for scenarios where data collection is cheap, for whatever measure of cheap is appropriate in the given scenario. By updating the policy directly, the learning trajectory may be more stable due to consistency. On the other hand, this same consistency can lead to reduced exploration of the action space as well as reduced sample efficiency due to the slow exploration. An example of this learning paradigm is PPO, which is one of the two methods used in this thesis.

Off-policy RL, conversely, learns from a *replay buffer* of past experiences. A replay buffer is a store of past experiences that an agent has experienced. During each step taken where the agent interacts with its environment, the action and resulting observation and reward is stored. During the update stage, a random sampling from the replay buffer is then taken and used to update the agent's policy or value function. By doing so, this enables *off-policy* learning, where the agent learns from previous experiences, including those performed under a different policy. This also allows for updates to be performed in batches, increasing throughput. The increased data diversity and ability to use samples from previous policies leads to better exploration and better sample efficiency. One example of off-policy RL is Deep Deterministic Policy Gradient (DDPG).

3.2.1 Proximal Policy Optimization

PPO is a model-free on-policy RL algorithm proposed in Schulman et al. [25]. It is a further development based on Trust Region Policy Optimization (TRPO) [26] and works by iteratively sampling data through interactions with the environment and optimizing a "surrogate" objective function using stochastic gradient ascent. The paper explores two approaches to this "surrogate" function, one using a penalty on KL divergence and one using a clipped objective function.

In the KL divergence approach, for each update of a given policy π_θ, the KL-penalized objective L^{KLPEN} is optimized with

$$L^{KLPEN}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_t heta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t - \beta KL\left(\pi_{\theta_{old}}(\cdot|s_t), \pi_t heta(\cdot|s_t)\right) \right] \tag{3.3}$$

where t is the current timestep, a the action, s the environment state, $\hat{A}_t$ an estimator of the advantage function at timestep t, and $\hat{\mathbb{E}}_t$ the empirical average over a finite batch of samples. With every policy update, we also update β by first computing $d = \hat{\mathbb{E}}_t\left(\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right)$. If $d < d_{targ}/1.5$, then we half β. If $d > d_{targ} \times 1.5$, then we double β. β is a hyperparemter that can be chosen, but is “not important in practice because the algorithm quickly adjusts it” [25].

In the clipped surrogate objective approach, for each update of the policy, the clipped surrogate objective L^{CLIP} for a given policy π_θ is updated, where

$$L^{CLIP}(\theta) = \min\left[\frac{\pi_t heta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t, g(\epsilon,\hat{A}_t)\right] \text{where } g(\epsilon,\hat{A}_t) \quad = \begin{cases}(1+\epsilon)\cdot\hat{A}_t & \hat{A}_t \geq 0\\(1-\epsilon)\cdot\hat{A}_t & \hat{A}_t < 0\end{cases} \tag{3.4}$$

The clipped surrogate object approach shows better empirical performance in [25] and is the one that we use in this thesis.

The algorithm itself is relatively simple and is best understood by directly reading the pseudocode seen in Algorithm 1.

Algorithm 1: Proximal Policy Optimization (PPO) Algorithm

Input: Initialize policy parameters θ_0 and value function parameters ϕ_0

for $k = 0, 1, 2, \ldots$ **do**

- Collect trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi_{\theta_k}$ in the environment;
- Compute rewards $\hat{R}_t$ for each action;
- Calculate advantage estimates $\hat{A}_t$ using the value function V_{ϕ_k};
- Update the policy by maximizing L^{CLIP}:

$$\theta_{k+1} = \arg\min_\theta \frac{1}{|\mathcal{D}_k|T} \sum_{\tau\in\mathcal{D}_k} \sum_{t=0}^{T} L^{CLIP} \tag{3.5}$$

  using Adam;
- Fit the value function by regression on mean-squared error over the collected trajectories $\mathcal{D}_k$;

end

3.2.2 Deep Deterministic Policy Gradient

Chapter 4

Methodology

The general intuition behind our approach is the lack of ability for a model to perform on unseen domains without any additional information. The general architecture of our approach can be seen in Figure 4.1. We first propose a signal preprocessing and signal-to-image transformation pipeline, described in Section 4.1 and 4.1.2. This transforms the input signal into an image that is then passed through a CNN encoder, as described in Section sec:methodology-signal-encoding, encoding the signal into a latent representation. This latent representation is then used in two different ways: As an input for the multi-task learning heads, described in Section 4.4, and as the state observation for our RL agent, described in Section 4.3.

To mitigate domain-shift, our RL agent is tasked with producing the best possible embedding of the domain D_r as its action given the signal latent representation as its state observation.

To provide the RL agent with a reward function, inspired by triplet loss, we use two different multi-task learning modules. One of these modules is provided D_0, representing a vector of zeros, and the other D_r.

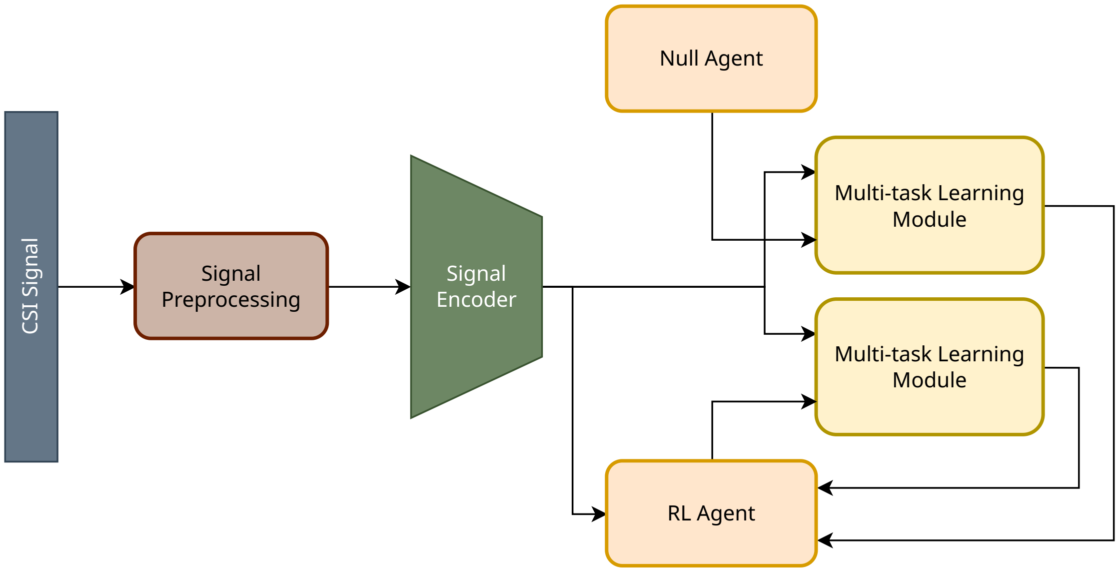

Figure 4.1: An abstracted diagram of the proposed architecture in this thesis.

4.1 Signal Preprocessing

As the old adage goes, garbage in, garbage out. As such, we propose a signal preprocessing module, visualized in Figure 4.2. This module first cleans the signal using traditional signal preprocessing and then transforms it into an image for use by the signal encoder.

4.1.1 Signal Processing

We propose the following steps from traditional signal processing to clean up the input signal for our ML algorithm.

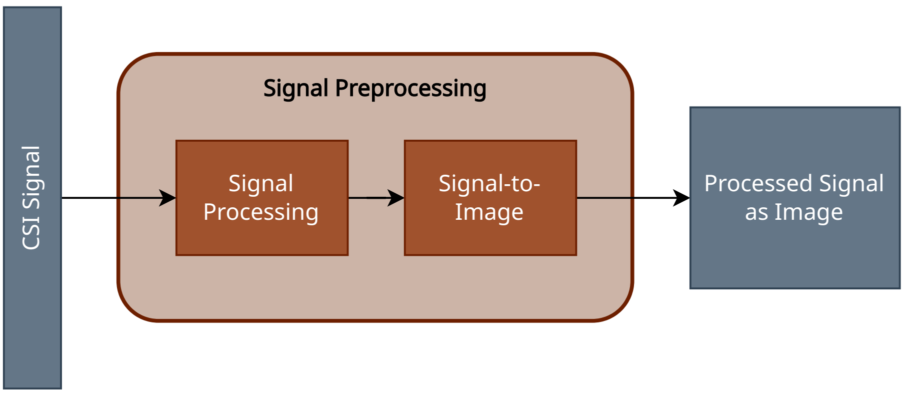

Figure 4.2: Details of the signal preprocessing module. The module is comprised of traditional signal preprocessing and a signal-to-image transformation.

- Noise filtering with a low-pass filter, since high-freq noise is most likely not human movement
- CSI Phase unwrapping and linear fitting as suggested by [11]
- CSI Phase derivative, to keep values from changing magnitude as in the case of unwrapping
- DWT seems to have been attempted in some papers, so let's see if that helps.
- We'll do an ablation study to see if these actually help

The signal can now be considered clean, or at least clean enough that we can continue with further steps.

4.1.2 Signal-to-Image Transformation

The next stage in our method is transform the signal into an image, leveraging advances from computer vision. There are many works from the literature which show that there are certainly improvements that can be gained from performing image processing on temporal data. We will experiment with the following four methods for signal-to-image transformation: DeepInsight [27], REFINED [3], GAF/FGAF (Feature-wise GAF) [35, 24], and MTF [35].

DeepInsight

DeepInsight [27] applies t-SNE on the columns of V, resulting in each AP k being mapped to a point in 2D space. The entire space is then rotated to fit within the minimum rectangular bounding box. These rotated points then become the pixel coordinates for a given AP. The value of the RSSI of a given AP k is then directly mapped to the value of its corresponding pixel. If multiple features share a pixel location, the average value of these features is then used as the value of the pixel.

REFINED

REpresentation of Features as Images with NEighborhood Dependencies [3] uses a technique the authors call "Bayesian Multidimensional Scaling". A 2D embedding of the data V_{embed} is calculated using Multidimensional Scaling (MDS) with a Euclidean distance metric. A pixel grid P of p^2 pixels, where $p = \lceil \sqrt{K} \rceil$ is then produced with dimensions $p \times p$. A mapping of features to pixels is calculated by considering all permutations which minimize the Euclidean distance of the pixel mapping to the feature location in V_{embed} while keeping at most one feature per pixel iteratively. This final mapping is calculated using a hill-climb algorithm.

GAF

Gramian Angular Fields [35] first transform the incoming signal for each AP to polar coordinates with

$$\vec{w_{i,k}} = \begin{cases} \phi_{i,k} = \arccos(V_{n,k}) \\ r_{i,k} = i/I \end{cases} \tag{4.1}$$

where $\vec{w_{i,k}}$ is the resulting vector of a signal, $V_{i,k}$ the value of the signal from AP K at sample i, i the sample number of the fingerprint, and I the total number of timesteps at the given sample location. For the case of our dataset, there are 6 timesteps for each location.

The Gramian is than calculated as

$$G_k = \begin{bmatrix} w_{1,k} \cdot w_{1,k} & \cdots & w_{1,k} \cdot w_{I,k} \\ \vdots & \ddots & \vdots \\ w_{I,k} \cdot w_{1,k} & \cdots & w_{I,k} \cdot w_{I,k} \end{bmatrix} \tag{4.2}$$

for each AP k creating an image with k channels.

We discovered in a previous project that calculating vector $\vec{w}$ across features, i.e., channels, instead of across timesteps can also be useful. This results in

$$\vec{w_{n,k}} = \begin{cases} \phi_{n,k} = \arccos(V_{n,k}) \\ r_{n,k} = k/K \end{cases} \tag{4.3}$$

where n is the sample index and k is the feature index. This leads to the Gramian for each sample

$$G_n = \begin{bmatrix} w_{n,1} \cdot w_{n,1} & \cdots & w_{n,1} \cdot w_{n,K} \\ \vdots & \ddots & \vdots \\ w_{n,K} \cdot w_{n,1} & \cdots & w_{n,K} \cdot w_{n,K} \end{bmatrix} \tag{4.4}$$

We call this transformation Feature-wise GAF (FGAF).

In either case, G is finally normalized.

MTF

Markovian Transition Fields [35] first quantizes the signal into q bins, each bin representing an interval of RSSI strength. A Markovian transition matrix M_t is then constructed where each row represents a fingerprint and each column a bin. The values in M_t are the size of each bin for a given sample. M_t is then normalized and aligned along the temporal axis such that in this new matrix M, $M_{i,j}$ represents the probability of a transition from bin i to bin j. M is then an image of size $q \times q$.

Regardless of the chosen method, the signal has now been transformed into an image and we can proceed with the encoding of this image into a latent space.

4.2 Signal encoding

The encoding of the signal, now an image, into a latent space is performed using CNN-based image processing methods.

- The signal encoding backbone is a standard CNN
- This will probably be something simple, like ResNet since the signal shouldn't be too complex that it requires something very SOTA
- Alternatively, mobilenet may be used as well
- This may actually be the least important aspect of this thesis

The latent representation of this signal is then passed to both the reinforcement learning agent as its state observation as well as to the multi-task learning modules. How the RL agent uses this state observation is described in Section 4.3 while its use by the multi-task learning module is described in Section 4.4.

4.3 Unsupervised Domain Representations through Reinforcement Learning

To mitigate domain shift, we implement a novel method for unsupervised domain representations through reinforcement learning.

- Base terminology: Agent, action, state observation (state), and reward
- The encoder-decoder architecture described in Section 4.2 and 4.4 is the environment.
- RL loop:
  1. The agent receives the image embedding produced by the encoder/backbone as its state
  2. The reinforcement learning agent produces an embedding of the domain as its action

3. The environment is trained with one pair of heads receiving the action of the agent while the other pair receives senseless values (either all 0s if the representation is one-hot or a uniform distribution if the representation is a probability measure)
4. After training of the agent, the RL agent is provided a reward from the environment which is used to improve the agent

- Let a performance metric function M denote the gesture classification performance of a multi-task learning module given the gesture classification prediction $\hat{y}_c$. Then, the reward function R of the agent is $R\left(M\left(\hat{y}_{c,rl}, y_c\right), M\left(\hat{y}_{c,0}, y_c\right)\right)$ and calculates the gesture prediction performance difference between the module given the domain embedding and the module given the null embedding.
  - The intuition is, the reward is based on how much better the head pair with the action should perform than the head pair without the action
  - This is based on the idea behind triplet loss, where in an unsupervised scenario, model performance can be determined by the metric performance differential between two inputs. Maximizing this difference implies better total model performance.
  - We will also experiment with L1 vs L2 regularization on the reward, representated by the function R.
- To speed up training, we take a page from AutoML competitions and the environment is trained within 1 minute and be focused on improving performance as fast as possible instead of achieving the best performance, at least during hyperparam optimization
- After hyperparameters are chosen, then we train properly and fully.
- A reward is then calculated using the function $R(M(\hat{y}_{rl}, y) - M(\hat{y}_0, y))$, where M is the metric used to calculate the performance of the gesture classifier in each multi-task learning module and R is some reward function.
- We do not use the absolute difference, as we are interested in the having negative values representing the module provided D_0 having better performance.
- We will be looking at PPO [25] and SARSA [23]
  - We've used both these techniques before in a previous project, so implementation should be relatively straightforwards
- *Motivation*

- We use the RL agent inspired by [18] and hypothesize that the RL architecture there can be extended to domain auto-labeling
- We use the triplet-loss approach since triplet loss is a proven way to provide a numerical value showing the difference between two input
  * In other words, this is analogous to testing the null hypothesis, where H_0 is no RL agent produced domain embedding and H_1 is a RL agent produced domain embedding
  * Our hypothesis is that the domain embedding will produce better results
  * If the RL agent can be trained, then the difference $(M(\hat{y}_{rl}, y) - M(\hat{y}_0, y)$ will approach $M(\hat{y}_{rl}, y)$ producing the maximum reward for the RL agent

4.4 Multi-task Learning

The actual gesture classification module is built using a multi-task learning module, a visualization of which can be seen in Figure 4.3.

- The idea is to enforce some sort of representation that *can* be used to get a domain-independent representation of the data
- We utilize multi-task learning for this, where one head is used to predict BVP, which is theoretically domain-independent [39]
  - We are not actually interested in the BVP, but having it as part of our training will enforce the need for a representation which is theoretically domain-independent.
  - [20] suggests the use of MMD and an adversarial approach to training the model to ensure that while the representation is domain-invariant, it still retains enough discriminatory powers such that the domain classifier can still perform.
  - Instead of this approach, we use the BVP prediction to enforce a representation which is domain-invariant and we use an RL agent instead of the domain classifier suggested in [20].
- The other head is a classifier head and classifies gestures
- It's been shown that doing multi-task learning like this leads to good results with latent representations which are more robust [33]
- We duplicate this pair of heads to enable a reward function for the RL agent inspired by triplet loss

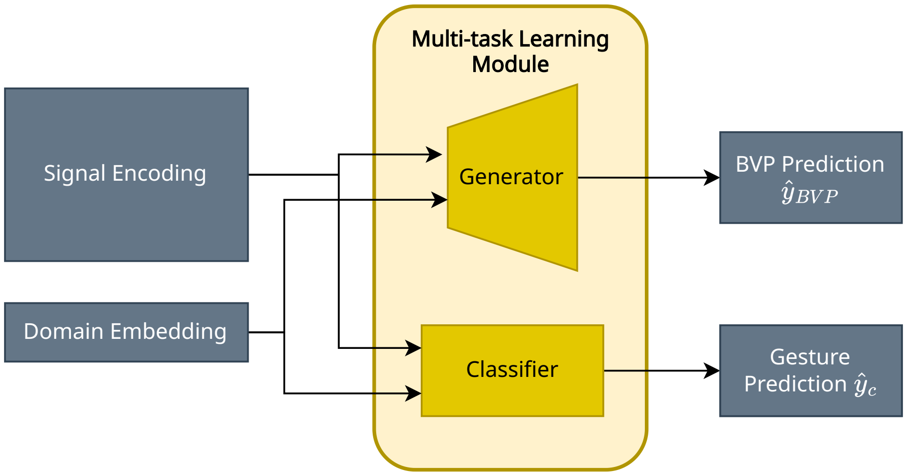

Figure 4.3: Details of the multi-task learning module. The generator is made up of a series of deconvolutional layers and produces the BVP prediction while the classifier is a fully connected network and produces a probability distribution over each gesture class.

- Loss function for the gesture classifier is cross entropy loss
- Loss function for the BVP generator is cross entropy loss on the produced image.
- *Motivation*
  - We use multi-task learning since its been shown that multitask learning can produce better/more powerful/more robust/more efficient latent representations due to the latent representation being used for multiple tasks
  - For this multi-task objective, we use BVP as the target of the generator, since we have ground-truth BVP data and [39] makes a point to show that BVP is domain-independent
    - ⋆ Therfore, if the generator can produce a BVP close to the ground-truth, then we have a latent representation that can be used to produce domain-independent features
  - We want try to get a latent representation capable of producing domain-independent features but we don't want to enforce it directly on the latent representation, since this may not work well [4] and [20] suggests giving the representation some discriminatory abilities w.r.t. the domain.
    - ⋆ Therefore, we believe that enforcing domain independence on the output will result in a representation that is domain-independent enough yet still retain discriminatory power.

- Cross Entropy loss on the produced image is used since it is the appropriate loss function for image generation (i.e., pixel-wise feature value prediction)
- We use gesture as the target of the classifier, since that's the goal of this thesis
- Cross Entropy loss on the prediction is used since it is the appropriate loss function for a multi-class prediction task.

Bibliography

[1] Fadel Adib and Dina Katabi. "See through walls with WiFI". In: *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. 2013, pp. 75–86.

[2] Fadel Adib et al. "3D tracking via body radio reflections". In: *11th* {*USENIX*} *Symposium on Networked Systems Design and Implementation (*{*NSDI*} *14)*. 2014, pp. 317–329.

[3] Omid Bazgir et al. "Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks". In: *Nature communications* 11.1 (2020), pp. 1–13.

[4] Bram van Berlo, Tanir Ozcelebi, and Nirvana Meratnia. "Insights on mini-batch alignment for wifi-csi data domain factor independent feature extraction". In: *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE. 2022, pp. 527–532.

[5] Lieke A.J. van den Biggelaar. "Gesture recognition using Deep Q-Neworks on WiFi-CSI data". MA thesis. Eindhoven University of Technology, 2022.

[6] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *CoRR* abs/2005.14165 (2020). arXiv: `2005.14165`. URL: `https://arxiv.org/abs/2005.14165`.

[7] Kevin Chetty, Graeme E Smith, and Karl Woodbridge. "Through-the-wall sensing of personnel using passive bistatic wifi radar at standoff distances". In: *IEEE Transactions on Geoscience and Remote Sensing* 50.4 (2011), pp. 1218–1226.

[8] Jianyang Ding and Yong Wang. "A WiFi-based smart home fall detection system using recurrent neural network". In: *IEEE Transactions on Consumer Electronics* 66.4 (2020), pp. 308–317.

[9] Rui Du et al. "An overview on IEEE 802.11 bf: WLAN sensing". In: *arXiv preprint arXiv:2207.04859* (2022).

[10] Technical University of Eindhoven. *Atlas most sustainable education building in the world Atlas most sustainable education building in the world Atlas most sustainable education building in the world*. Apr. 2019. URL: `https://www.tue.nl/en/our-university/tue-campus/buildings/atlas/atlas-most-sustainable-education-building-in-the-world/`.

[11] Jiaqi Geng, Dong Huang, and Fernando De la Torre. “DensePose From WiFi”. In: *arXiv preprint arXiv:2301.00250* (2022).

[12] Mohammad Hasan. *State of IoT 2022 Number of connected IoT devices growing 18% to 14.4 billion globally*. URL: `https://iot-analytics.com/number-connected-iot-devices/` (visited on 05/18/2022).

[13] Wenfeng He et al. “WiG: WiFi-based gesture recognition system”. In: *2015 24th International Conference on Computer Communication and Networks (ICCCN)*. IEEE. 2015, pp. 1–7.

[14] Dehao Jiang, Mingqi Li, and Chunling Xu. “Wigan: A wifi based gesture recognition system with gans”. In: *Sensors* 20.17 (2020), p. 4757.

[15] Wenjun Jiang et al. “Towards environment independent device free human activity recognition”. In: *Proceedings of the 24th annual international conference on mobile computing and networking*. 2018, pp. 289–304.

[16] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[17] Diederik P. Kingma and Max Welling. “An Introduction to Variational Autoencoders”. In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392. DOI: `10.1561/2200000056`. URL: `https://doi.org/10.1561%2F2200000056`.

[18] Yongsen Ma et al. “Location-and person-independent activity recognition with WiFi, deep neural networks, and reinforcement learning”. In: *ACM Transactions on Internet of Things* 2.1 (2021), pp. 1–25.

[19] Yongsen Ma et al. “Signfi: Sign language recognition using wifi”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018), pp. 1–21.

[20] Mauro Martini et al. “Domain-Adversarial Training of Self-Attention-Based Networks for Land Cover Classification Using Multi-Temporal Sentinel-2 Satellite Imagery”. In: *Remote Sensing* 13.13 (2021). ISSN: 2072-4292. DOI: `10.3390/rs13132564`. URL: `https://www.mdpi.com/2072-4292/13/13/2564`.

[21] C.H.J.M Oerlemans. "The Effect of Data Preprocessing On the Performance of Few-shot Learning for Wi-Fi CSI- based Gesture Recognition The Effect of Data Preprocessing on the Performance of Few-shot Learning for Wi-Fi CSI-based Gesture Recognition". MA thesis. Tilburg University, 2022.

[22] Qifan Pu et al. "Whole-home gesture recognition using wireless signals". In: *Proceedings of the 19th annual international conference on Mobile computing & networking*. 2013, pp. 27–38.

[23] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. Vol. 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[24] Yvan Putra Satyawan. "CNNs and Preprocessing: The Dynamic Duo of Wi-Fi Localization". Jan. 2023.

[25] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[26] John Schulman et al. *Trust Region Policy Optimization*. 2017. arXiv: `1502.05477 [cs.LG]`.

[27] Alok Sharma et al. "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture". In: *Scientific reports* 9.1 (2019), pp. 1–7.

[28] Signify B.V. *Groundbreaking motion detection turns your lights on and off without the need for sensors*. Press Release. Sept. 2022. URL: `https://www.signify.com/global/our-company/news/press-releases/2022/20220916-signify-delivers-new-way-to-automate-your-lights-using-wi-fi-sensing-technology` (visited on 08/31/2023).

[29] Signify B.V. *Smart lighting for your daily living: WiZ launches a brand new generation of products in Europe*. Aug. 2020. (Visited on 08/31/2023).

[30] Suze E. Sips. "Impact analysis of network pruning and quantization on the domain shift problem in a recognition context, using WiFi CSI data". MA thesis. Eindhoven University of Technology, 2022.

[31] Maryam Tavakol. *Data Intelligence Challenge*. Lecture. Apr. 2022.

[32] Tesla Inc. *Data Sharing End User License Agreement*. EULA. May 2017.

[33] Lukas Tuggener et al. "The DeepScoresV2 dataset and benchmark for music object detection". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 9188–9195.

[34] Fei Wang et al. "Person-in-WiFi: Fine-grained person perception using WiFi". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5452–5461.

[35] Zhiguang Wang and Tim Oates. "Imaging time-series to improve classification and imputation". In: *24th International Joint Conference on Artificial Intelligence*. 2015.

[36] Hongfei Xue et al. "DeepMV: Multi-view deep learning for device-free human activity recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.1 (2020), pp. 1–26.

[37] Weidong Zhang, Zexing Wang, and Xuangou Wu. "WiFi signal-based gesture recognition using federated parameter-matched aggregation". In: *Sensors* 22.6 (2022), p. 2349.

[38] Youshan Zhang, Hui Ye, and Brian D Davison. "Adversarial reinforcement learning for unsupervised domain adaptation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 635–644.

[39] Yue Zheng et al. "Zero-effort cross-domain gesture recognition with Wi-Fi". In: *Proceedings of the 17th annual international conference on mobile systems, applications, and services*. 2019, pp. 313–325.

[40] Andrii Zhuravchak, Oleg Kapshii, and Evangelos Pournaras. "Human Activity Recognition based on Wi-Fi CSI Data-A Deep Neural Network Approach". In: *Procedia Computer Science* 198 (2022), pp. 59–66.

[41] Augustinas Zinys, Bram van Berlo, and Nirvana Meratnia. "A domain-independent generative adversarial network for activity recognition using wifi csi data". In: *Sensors* 21.23 (2021), p. 7852.