



Koupai, A. K., Bocus, M. J., Santos-rodriguez, R., Piechocki, R. J., & Mcconville, R. (2022). Self-Supervised Multimodal Fusion Transformer for Passive Activity Recognition. *IET Wireless Sensor Systems*, 12(5-6), 149-160. <https://doi.org/10.1049/wss2.12044>

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
[10.1049/wss2.12044](https://doi.org/10.1049/wss2.12044)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the final published version of the article (version of record). It first appeared online via Wiley at <https://doi.org/10.1049/wss2.12044>. Please refer to any applicable terms of use of the publisher.


University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

ORIGINAL RESEARCH

Self-supervised multimodal fusion transformer for passive activity recognition

Armand K. Koupai | Muhammad J. Bocus  | Raul Santos-Rodriguez |
Robert J. Piechocki | Ryan McConville

School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths, University of Bristol, Bristol, UK

Correspondence

Mohammad J. Bocus, Digital Health (DH) Lab, University of Bristol, First Floor, 1 Cathedral Square, Trinity Street, Bristol BS1 5DD, UK.
Email: mb13530@bristol.ac.uk

Funding information

Engineering and Physical Sciences Research Council, Grant/Award Number: EP/R018677/1

Abstract

The pervasiveness of Wi-Fi signals provides significant opportunities for human sensing and activity recognition in fields such as healthcare. The sensors most commonly used for passive Wi-Fi sensing are based on passive Wi-Fi radar (PWR) and channel state information (CSI) data, however current systems do not effectively exploit the information acquired through multiple sensors to recognise the different activities. In this study, new properties of the Transformer architecture for multimodal sensor fusion are explored. Different signal processing techniques are used to extract multiple image-based features from PWR and CSI data such as spectrograms, scalograms and Markov transition field (MTF). The Fusion Transformer, an attention-based model for multimodal and multi-sensor fusion is first proposed. Experimental results show that the Fusion Transformer approach can achieve competitive results compared to a ResNet architecture but with much fewer resources. To further improve the model, a simple and effective framework for multimodal and multi-sensor self-supervised learning (SSL) is proposed. The self-supervised Fusion Transformer outperforms the baselines, achieving a macro F1-score of 95.9%. Finally, this study shows how this approach significantly outperforms the others when trained with as little as 1% (2 min) of labelled training data to 20% (40 min) of labelled training data.

KEYWORDS

deep learning, multimodal/sensor fusion, passive WiFi-based HAR, self-supervised learning, vision transformer (ViT)

1 | INTRODUCTION

In recent years, there has been growing research interest in healthcare applications to diagnose and prevent mental and physical diseases, often within the home, and often with the objective to relieve the burden on healthcare services. Many systems have been developed to collect and provide information about a person's health condition in this way [1–7]. A wide array of sensors have been deployed, from wearables, to cameras, to more recently passive sensing systems using radio frequency (RF) signals. Sensors such as Wi-Fi are particularly promising for in-home healthcare applications, as they: (1) perform sensing passively, (2) avoid any discomfort for the user (as no sensors need to be worn), (3) are ubiquitous,

(4) and they are more privacy-friendly than alternatives such as cameras. Wi-Fi-based sensing systems have been studied for tasks such as gesture recognition [8–10], sign language recognition [11] and fall detection [12, 13]. These systems can also be used for human activity recognition (HAR) [14–18] as human activities cause changes in the wireless signal transmitted by the passive Wi-Fi sensors in terms of frequency shifts, multipath propagation and signal attenuation [19]. Two Wi-Fi sensors are commonly used in HAR, namely passive Wi-Fi radar (PWR) and channel state information (CSI). CSI represents how a wireless signal propagates from the transmitter to its receiver at particular carrier frequencies along multiple paths. The CSI data, which can be extracted from specific network interface cards (NICs) such as Intel 5300

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *IET Wireless Sensor Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

[20] or Atheros [21], can be viewed as a 3D time-series matrix of complex values representing both the amplitude attenuation and the phase shift of multiple propagation paths. It captures how wireless signals travel through surrounding objects or humans in time, frequency and spatial domains. Despite that both CSI and PWR sensors use the same signal source and have a similar function, PWR works differently. A PWR system correlates the transmitted signal from a Wi-Fi access point and the reflected signal from the surveillance area and calculates the distance between the antenna and the object or human [17].

Research in radio-based human sensing and activity recognition has moved towards deep learning, principally because deep learning models can learn complex representations from the data, even if the latter is noisy in nature. Deep neural network architectures that are commonly used for Wi-Fi CSI-based sensing include convolutional neural networks (CNN) where the raw CSI data is transformed into image-like representations such as spectrograms (e.g. [16–18, 22]) or recurrent neural networks (RNN) which work directly on the raw Wi-Fi data (e.g. [23, 24]). In this work, we study and propose to use multiple synchronised sensors, or views, in an indoor environment to improve the performance of a passive HAR system. More specifically, we propose to build a system that collects raw data from synchronised RF sensors, and after some signal processing, all modalities can be fused effectively; in this case using Transformers.

Recent work has demonstrated that the Vision Transformer (ViT) [25] architecture is capable of competitive or superior performance on image classification tasks at a large scale. Instead of convolutions, it uses a self-attention mechanism to aggregate information across locations. Thus, we investigate the potential of the ViT for sensor and feature fusion. For this purpose, we need to address the possible challenges: firstly, how the self-attention mechanism present in the ViT can be used for sensor feature fusion? Secondly, does ViT benefit from sensor fusion and lead to better predictions for HAR? In this paper, we study these questions and compare our findings with a traditional ResNet model. The main contributions of this work are the following:

- We propose a model, the Sensor Fusion Vision Transformer (SF-ViT), based on the Vision Transformer (ViT) architecture [25], that can fuse multiple image features and views.
- We extend it to a more general framework which we called the Fusion Transformer, that can effectively fuse multiple features from different types of sensors and we assess the effectiveness of our transformer-based model for multimodal and multi-sensor fusion.
- We evaluate the effectiveness of the Fusion Transformer on a HAR dataset (collected using Wi-Fi sensors) in a purely supervised fashion, and compare its performance against ResNet.
- We also propose a new method for multimodal and multi-sensor self-supervised learning (SSL) that outperforms the baselines using multiple image features and views for passive HAR.

This paper is organised as follows: Related works on multimodal sensor fusion are presented in Section 2. The methodology and system design of our multimodal sensor fusion transformer models are described in Section 3, including details on the signal processing of Wi-Fi-based signals. Section 4 provides detailed information on the experimental setup. Section 5 presents the results obtained using our fully supervised Fusion Transformer model on a HAR dataset. Section 6 describes our self-supervised multimodal sensor fusion transformer architecture, along with details on the experiment setup and results. Finally, conclusions are drawn in Section 7.

2 | RELATED WORK

Unimodal and multimodal sensing based on vision sensors (e.g. RGB-D cameras, infrared, thermal cameras etc.) and inertial wearable sensors (e.g. accelerometers, gyroscopes, magnetometers, audio-signals etc.) have previously been studied for HAR. The interested reader is kindly referred to [28] for a comprehensive review on different multimodal HAR methods and fusion techniques (non-radio based). Most works on multimodal or multi-sensor fusion for human action recognition using RF, inertial and/or vision sensors, have considered either decision-level fusion or feature-level fusion. For example, the authors of ref. [29] perform multimodal fusion at the decision level to combine the advantages of Wi-Fi and vision-based sensors using a hybrid deep neural network (DNN) model to achieve a 97.5% cross-validation accuracy on average for three activities: sitting, standing and walking. The model essentially consists of a Wi-Fi sensing module (CNN architecture) and a vision sensing module (based on the convolutional 3D model) for processing Wi-Fi and video frames for unimodal inference, followed by a multimodal fusion module. Multimodal fusion is performed at the decision level (after both the Wi-Fi and vision modules have made a classification) because this framework is stated to be more flexible and robust to unimodal failure compared to feature level fusion. The authors of ref. [30] present a method for HAR, which leverages four sensor modalities, namely, skeleton sequences, inertial and motion capture measurements and Wi-Fi fingerprints. The fusion of signals is formulated as a matrix concatenation. The individual signals of different sensor modalities are transformed and represented as an image. The resulting images are then fed into a 2D CNN (EfficientNet B2) for classification. The authors evaluated their approach on four different datasets; the NTU RGB + D 120 dataset for skeleton data, the UTD-MHAD dataset for skeleton and inertial data, the ARIEL dataset for Wi-Fi data and the Simitate dataset for motion capture data. Good experimental results were achieved across the different sensor modalities. The authors of ref. [31] proposed a multimodal HAR system that leverages Wi-Fi and wearable sensor modalities to jointly infer human activities. They collected CSI data from a standard Wi-Fi NIC, alongside the user's local body movements via a wearable inertial measurement unit (IMU) consisting of an accelerometer, gyroscope, and magnetometer sensors. They

calculated the time-variant mean Doppler shift (MDS) from the processed CSI data and magnitude from the inertial data for each sensor of the IMU. Then, various time and frequency domain features were separately extracted from the magnitude data and the MDS. The authors applied a feature-level fusion method which sequentially concatenates feature vectors that belong to the same activity sample. Finally, supervised machine learning techniques were used to classify four activities, such as walking, falling, sitting, and picking up an object from the floor. The authors of ref. [17] conducted a comprehensive study on the comparison of two RF sensing devices for the purpose of HAR, namely, CSI and PWR systems. Two pipelines were proposed for filtering and processing the raw signals from the two sensors into Doppler spectrograms, which were then used to train a simple supervised CNN to evaluate the HAR performance. They considered the combined activity data from three different layouts. In the first layout, the transmitter and receiver were facing each other (in a line-of-sight configuration) while in the second layout, the transmitter and receiver were at 90° to each other. Finally, in the third layout, the transmitter and receiver were co-located (placed next to each other). The CSI system achieved an overall accuracy of 67,3% while the PWR system had an accuracy of 66,7%. Although this work presents a simple system which combines CSI and PWR spectrograms by merging probabilities from two networks (decision-level fusion), current state-of-the-art models are not specifically designed for the fusion of multiple passive Wi-Fi devices.

While CNN architecture was the de-facto standard for computer vision tasks, gradually, ViT showed very promising results when pre-trained on large amounts of data and then fine-tuned on mid-sized or small-sized image recognition benchmarks while requiring fewer learnable parameters for training [32, 33]. However, most ViT models are trained on natural images of very large sizes, together with pre-training and very strong data augmentation techniques. A similar work which also trained a ViT with spectrograms is the audio spectrogram transformer (AST) [34], which presents a new method for audio classification with a ViT using spectrogram

data. Recent works showed that ViTs could outperform ResNets without pre-training or strong data augmentations [35], notably by using sharpness-aware minimisation technique [36], which simultaneously minimises the loss value and loss sharpness by seeking parameters that lie in neighbourhoods and having uniformly low loss. However, this technique requires the computation of two forward-backward propagations to estimate the ‘sharpness-aware’ gradient, and thus, leads to an increased training time.

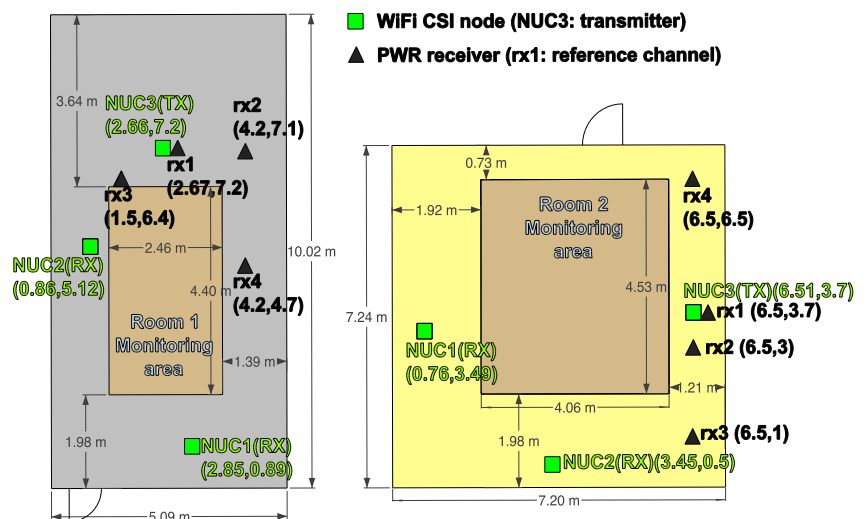
In this paper, we evaluate the performance of our Transformer-based sensor fusion model for HAR using image data generated from multiple sensors. We evaluate its potential for sensor fusion and propose a method for multimodal and multi-sensor self-supervised learning (SSL).

3 | METHODOLOGY AND SYSTEM DESIGN

3.1 | Signal processing of RF sensors

Inspired by another work in this area [17], which has explored two pipelines for extracting image features from RF sensors using signal processing techniques, we apply the same principles. In this work, we use the OPERAnet dataset [27], which includes publicly available data from both CSI and PWR systems (as well as Kinect and ultra-wideband systems). The dataset was collected with the intention to evaluate HAR and localisation techniques with measurements obtained from synchronised RF devices and vision-based sensors. The experimental setup established to collect both CSI and PWR data is shown in Figure 1. Figure 2a,b show some examples of the generated spectrograms with these two pipelines, for each of the six activities, namely, sitting down on a chair (*‘sit’*), standing from chair (*‘stand’*), lying down on the floor (*‘liedown’*), standing from floor (*‘standff’*), body rotation (*‘bodyrotate’*), and walking (*‘walk’*). The pipelines are as follows:

FIGURE 1 The CSI system and PWR system deployment [26]. The CSI system consisted of 2 receivers (denoted as NUC1 and NUC2) while the PWR system consisted of 3 receivers/surveillance channel (denoted as rx2, rx3 and rx4). For more details on the dataset, the interested reader is kindly referred to [26, 27].



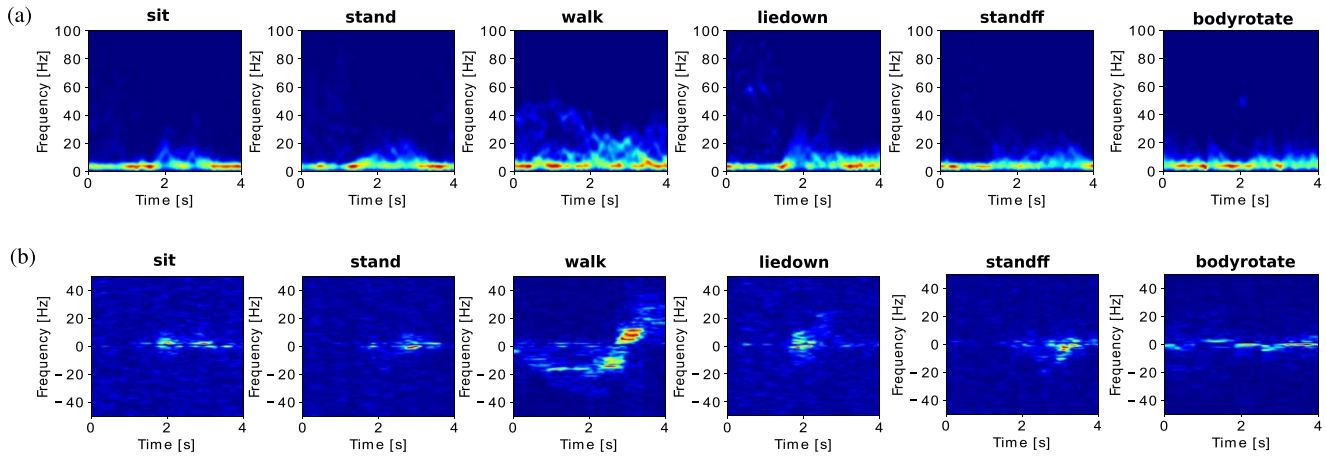


FIGURE 2 Visualisation of CSI and PWR spectrograms for each activity: (a) Amplitude CSI spectrograms from view 1 (NUC1) and (b) PWR spectrograms from first surveillance channel (rx2).

- In Figure 2a: we denoise the CSI signal using discrete wavelet transform (DWT) and median filtering, then reduce the dimensionality using principal component analysis (PCA) and generate a spectrogram using short-time Fourier transform (STFT).
- In Figure 2b: we apply cross ambiguity function to the raw PWR data, use the CLEAN algorithm and constant false alarm rate (CFAR) for direct signal cancellation and noise reduction, generating as output a Doppler spectrogram [17].

These two pipelines are necessary to extract informative data from CSI and PWR sensors. The raw CSI data is very noisy in nature, and thus the DWT technique helps to filter out high frequency components and remove noises, while preserving most of the information and avoiding the distortion of the signal [16]. Afterwards, we perform median filtering to remove any undesired transients in the CSI measurements which have not been caused by human motion. Despite that the CSI data is highly informative, it consists of a lot of complex values per second, depending on the number of transmit and receive antennas, orthogonal frequency-division multiplexing (OFDM) subcarriers and packet rate (for example, the Intel 5300 chipset captures complex CSI data over 3 transmit antennas, 3 receive antennas and 30 subcarriers). Therefore, we use PCA to reduce the computational complexity of such data, while preserving as much information as possible. Finally, we convert the PCA signal into spectrograms using STFT [17].

For the PWR signal, we first apply the cross ambiguity function (CAF) to extract target range and Doppler information. However, we also capture an interference source which is the strong direct signal emitted from the Wi-Fi access point and which is captured by the PWR surveillance channels. Thus, to remove this signal, we employ the CLEAN algorithm [37]. The last step consists of reducing the noise on the CAF surface. We use CFAR to estimate the background noise and apply it to the CAF surface. PWR's Doppler spectrogram is

generated by selecting the maximum Doppler pulse from each Doppler bin within the CAF surface [17].

The interested reader is kindly referred to our previous works [16–18] for more details on the signal processing pipelines for Wi-Fi CSI and PWR data. In this paper, we focus mainly on the design of models that can fuse data from multiple modalities/sensors effectively for the purpose of HAR. It should be noted that all the devices were synchronised to the same network time protocol (NTP) server and were labelled in sync. Thus, the raw data could be segmented as per the ground truth activity labels and processed accordingly.

Using the CSI data, we also generate other features such as scalograms and Markov transition fields (MTF). Each feature captures particular information about the activity. Given all of these different features, we aim to build a network that can fuse all these images together effectively to improve the overall system performance. In this work, we have extracted 15 different features (see Figure 3):

- PWR spectrogram data collected from the three receiver surveillance channels, rx2, rx3, and rx4 in Figure 1 (denoted as ‘PWR channel 1’, ‘PWR channel 2’, ‘PWR channel 3’, respectively, in Figure 3);
- Spectrograms generated using STFT on amplitude CSI data from the two receivers, NUC1 (‘Amp. spec. N1’) and NUC2 (‘Amp. spec. N2’);
- Spectrograms generated using STFT on phase difference CSI data for each of the two receivers (‘Ph. diff. N1’, ‘Ph. diff. N2’);
- Markov transition field (MTF) [38] features generated from phase difference CSI data acquired from two receivers (‘MTF ph. diff. N1’, ‘MTF ph. diff. N2’);
- MTF features generated from amplitude CSI data acquired from two receivers (‘MTF amp. N1’, ‘MTF amp. N2’);
- Scalograms generated by applying continuous wavelet transform (CWT) on the amplitude CSI data from NUC1 (‘Amp. scal. N1’) and NUC2 (‘Amp. scal. N2’) receivers;

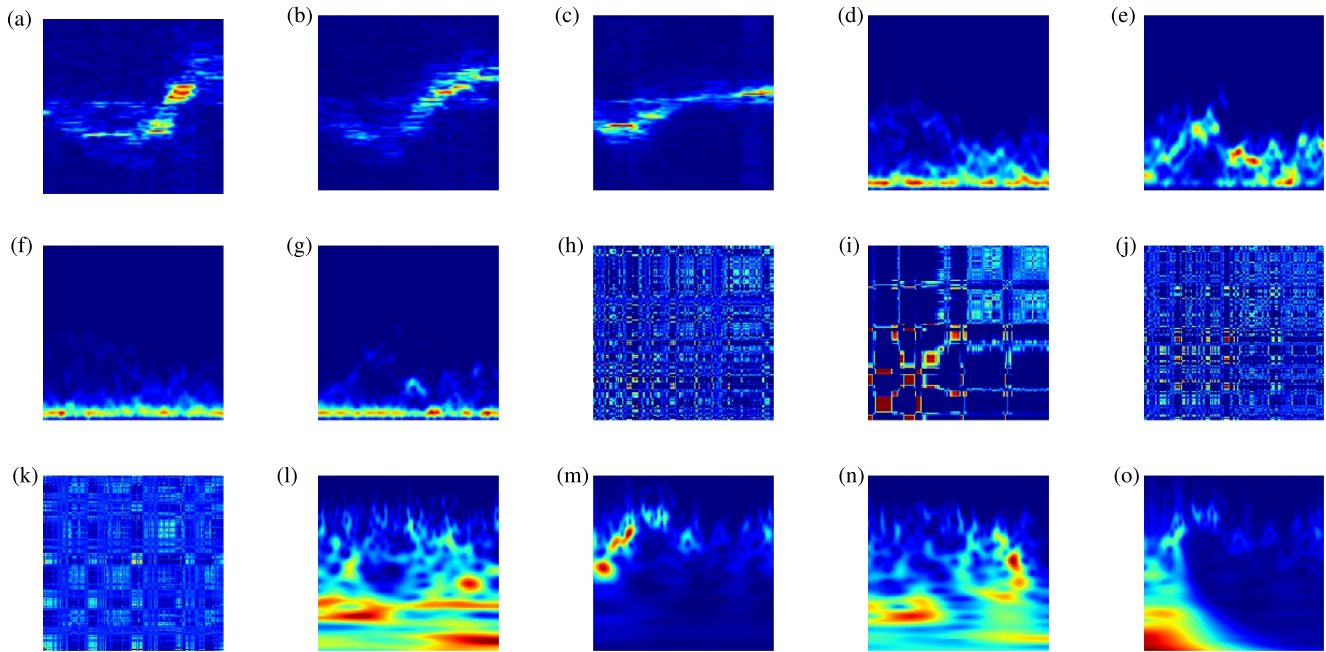


FIGURE 3 Fifteen image features extracted via signal processing techniques representing a person walking in the monitoring area for a duration of 4 seconds: (a) PWR channel 1, (b) PWR channel 2, (c) PWR channel 3, (d) Amp. spec. N1, (e) Amp. spec. N2, (f) Ph. diff. N1, (g) Ph. diff. N2, (h) MTF ph. diff. N1, (i) MTF ph. diff. N2, (j) MTF amp. N1, (k) MTF amp. N2, (l) Amp. scal. N1, (m) Amp. scal. N2, (n) Ph. diff. scal. N1, and (o) Ph. diff. scal. N2.

- Scalograms generated using CWT on the phase difference CSI data from the two CSI receivers (*Ph. diff. scal. N1*, *Ph. diff. scal. N2*).

Each channel and receiver can be seen as another view of the human activity performed in the room. Previously, we presented the spectrograms of CSI and PWR data, which give a visual representation of the spectrum of frequencies of a signal varying through time. Spectrograms are generated through STFT by applying a sliding window to obtain equally sized segments of the signal and then FFT is performed on the samples in each segment, which converts the signal from the time domain to the frequency domain. Similar to STFT, the scalogram is a time-frequency representation of a signal and it is obtained from the absolute value of the CWT of a signal. Finally, we also introduce another type of representation called the Markov transition field (MTF), which is an image generated from time series data, representing a field of transition probabilities for a discretised time series.

3.2 | Multimodal sensor fusion transformer

3.2.1 | A first approach: Sensor fusion Vision Transformer

We will first present the Sensor-Fusion Vision Transformer (SF-ViT), which uses a similar architecture to the conventional Vision Transformer (ViT). Nevertheless, in most applications

where ViT is used, the model is trained with ‘natural’ images of size $224 \times 224 \times 3$ (height, width, channels) that are divided into small patches of size 16×16 or 32×32 . Here instead, we concatenate all image features and obtain an image of size $224 \times (224 \times N) \times 1$ where N is the number of different image features concatenated. Instead of dividing our image into small patches of size 16×16 , we patch the image so that each patch represents a different image-based feature. Figure 4 illustrates an overview of the SF-ViT, where the shape of the input image has been changed for convenience.

The SF-ViT trains a transformer to recognise human activities by assigning a high attention weight to relevant features (i.e. our patches of different features), and a low attention weight to less pertinent image features. The SF-ViT's inspiration is that the more unique the image features that are used with the ViT are, the more effective is the model for recognising human activities, as each image feature will represent or capture different information about the activity, and thus combining them effectively should lead to a better performance.

3.2.2 | The Fusion Transformer

One potential issue with the SF-ViT approach is that we do a linear projection of patches of size $224 \times 224 \times 1$ into a feature space of size 512, which is computationally expensive and results in a very large number of trainable parameters and potential over-fitting, as each input pixel is connected to the linear layer. To remedy this, we instead first encode each image-

based feature using a CNN encoder, which transforms the raw image feature of size $224 \times 224 \times 1$ into an image of size $16 \times 16 \times 64$. This new architecture can be considered as a multimodal model, where each modality is first passed into an encoder that transforms the raw modality into a smaller feature space. The corresponding model architecture, which we call Fusion Transformer, is presented in Figure 5.

4 | EXPERIMENTAL SETUP

We evaluate the capabilities of our Fusion Transformer on the HAR task and compare its performance with ResNet and show that the Fusion Transformer is successful in achieving competitive results while requiring less trainable parameters than ResNet. In this section, the experimental setup used

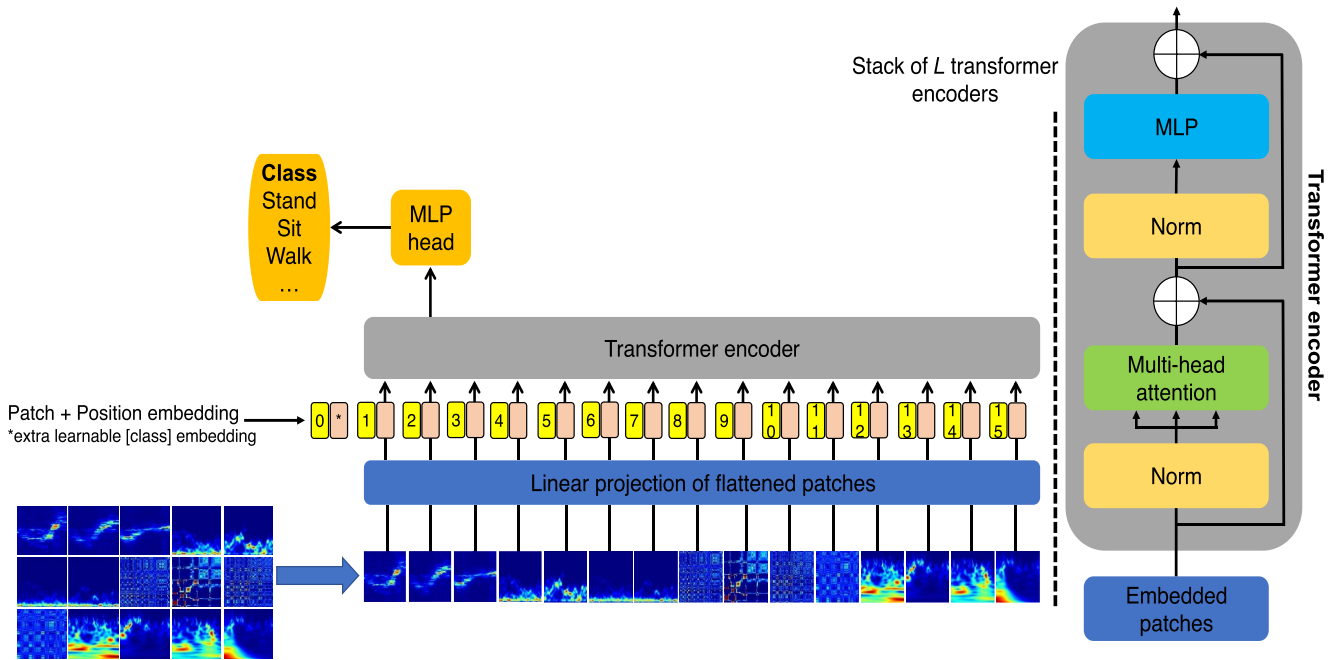


FIGURE 4 Sensor-Fusion Vision Transformer (SF-ViT) model overview. We split our concatenated image into patches of fixed size, 224×224 , where each patch corresponds to one of the image features. We linearly embed each of them, add position embeddings, and feed the output embeddings to the transformer encoder. In order to perform classification, we add an extra learnable ‘classification token’. The network model is inspired from the original ViT architecture [25].

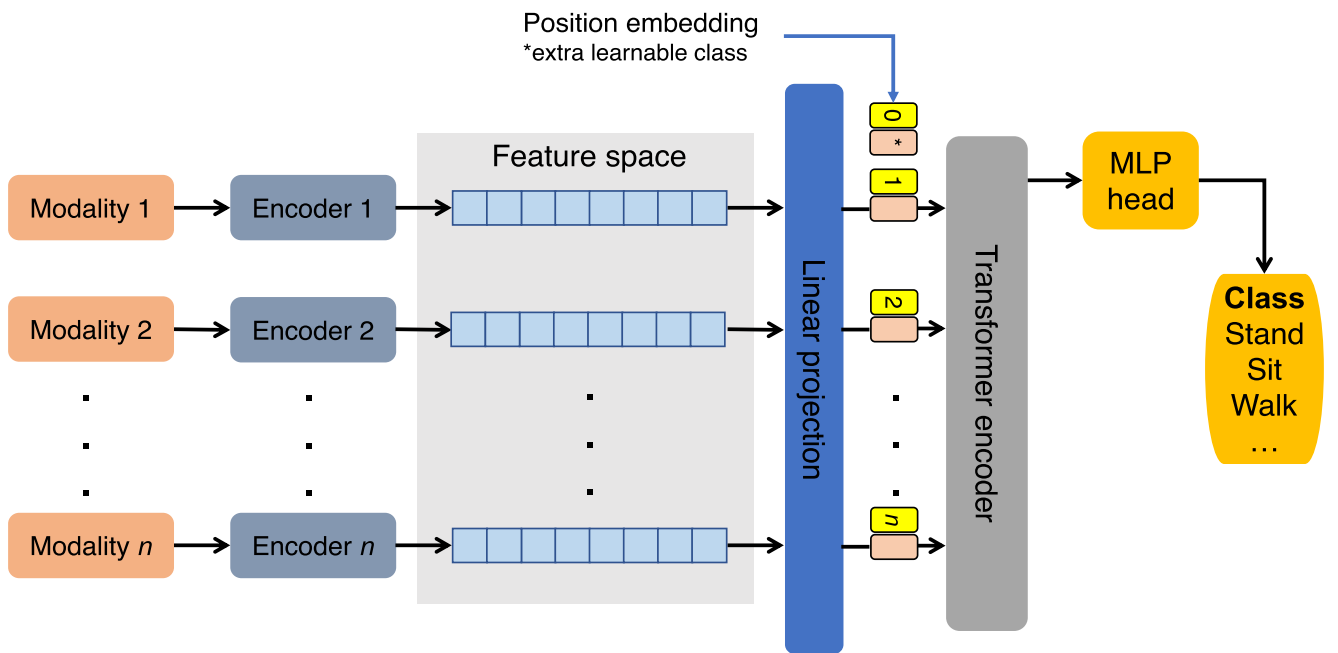


FIGURE 5 The Fusion Transformer model overview. Each modality is encoded into a new feature space and then linearly embedded. We add position embeddings and feed the output embeddings to the transformer encoder. We add an extra learnable classification token.

throughout the findings of the paper is presented. The system was developed in PyTorch and all models have been trained on a single GPU (Nvidia 2080Ti).

4.1 | Dataset and metrics

As mentioned previously, for our experimentation, we used the OPERAnet dataset [27], which includes publicly available data from both CSI and PWR systems. The RF sensors captured the changes in the wireless signals while six daily activities were being performed by six participants, namely, sitting down on a chair (*'sit'*), standing from chair (*'stand'*), lying down on the floor (*'liedown'*), standing from floor (*'standff'*), body rotation (*'bodyrotate'*), and walking (*'walk'*). It should be noted that the six activities were performed in two different rooms and in each room the participants performed the activities at different locations. The distribution of the six activities performed by the six participants in the two rooms is reported in [26]. Applying the signal processing pipelines outlined in Section 3.1, led to a dataset composed of 2897 data samples (non-overlapping windows each representing 4 s of an activity) for the six activities. Worth noting however, as is the case in reality, the distribution of the different activities a human engages in, is highly imbalanced. In this case, we have an imbalanced dataset where the two most represented classes are 'body rotating' and 'walking', representing respectively 30% and 33% of the total observations. The two classes which are less represented are 'standing from floor' and 'lying down', each representing 7% of the dataset. For training and validation purposes, we randomly split the dataset into a train set and a validation set, respectively composed of 80% and 20% of the total dataset samples. These two sets consist of activity samples from the two experimental rooms as well as samples at different locations within a given room (across all participants). The main

objective of the HAR algorithm/model is that, irrespective of the physical environment or participant's location within the environment or participant's demographics, it should be able to generalise well and classify the human activities with high accuracy. For this dataset, we use the accuracy and macro-averaged F1-score as our main metrics.

4.2 | Models

In this section, we will initially focus on the Fusion Transformer. All experiments have been performed with the configuration presented in Table 1. The width of the image is $N \times 224$, where N is the number of different image features generated. We compare and train the model with different number of features to analyse how the model's performance scales.

In the Fusion Transformer shown in Figure 5, we add a CNN encoder for our images to extract more relevant features from the raw images and to reduce the size of the images. The CNN encoder is composed of 4 blocks, where each block consists of a convolution, ReLU and pooling layers. Each image feature of size $224 \times 224 \times 1$ is embedded in a new image representation of dimension $16 \times 16 \times 64$.

To compare the performance of the Fusion Transformer with a baseline, we also train a ResNet model to evaluate whether it achieves better performance when trained with multiple image features, by considering each feature as a new channel. We trained two models: ResNet18 and ResNet34.

4.3 | Training

All models, including ResNet, have been trained using the AdamW optimiser, with $\beta_1 = 0.90$ and $\beta_2 = 0.999$, with a weight decay settled at 0.01 and a batch size of 64. The learning

TABLE 1 ViT parameters

Image size	Patch size	Channels	emb. dim.	Depth	qkv bias	Drop out	MLP ratio
224, $N \times 224$	224, 224	1	512	3	False	0.1	1.0

TABLE 2 Performance comparison of the Fusion Transformer with ResNet

	SF-ViT		ResNet18		ResNet34		Fusion Transformer	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
CSI amplitude spectrogram (view 1 = 1 feature)	80.3%	71.7%	92.8%	89.7%	73.3%	71.0%	88.3%	83.5%
CSI amplitude spectrogram (2 views = 2 features)	85.9%	78.6%	93.1%	90.0%	35.6%	43.4%	92.1%	87.0%
CSI ph. diff + amp. spectrograms (4 features)	84.3%	77.4%	94.5%	91.4%	95.7%	93.9%	92.2%	88.4%
CSI (amp. + ph. diff.) & PWR spectrograms (7 features)	91.6%	88.2%	95.0%	92.4%	96.6%	94.9%	95.9%	94.3%
CSI (amp. + ph. diff.) spectrograms + PWR spectrograms + CSI (amp. + ph. diff.) MTF (11 features)	91.9%	88.6%	93.8%	91.1%	91.9%	88.2%	94.3%	91.9%
All 15 image features	92.8%	89.5%	91.0%	86.5%	93.3%	90.4%	93.6%	91.1%

Note: Bold font indicates the best result achieved with each model.

rate has been initialised at $1e-4$, which is reduced during training using a learning step scheduler with a unitary step size and $\gamma = 0.5$. The loss function used for these experiments is cross-entropy.

Despite that recent works train ViTs using pre-training or transfer learning on large datasets, we decided to train our model from scratch, to more closely study the benefit of our sensor fusion model for activity recognition. Furthermore, when training on smaller datasets, ViT-based models have a weaker inductive bias compared to CNNs and this leads to an increased reliance on model regularisation or data augmentation [39]. In the case of CSI and PWR data, using similar data augmentation techniques as those used on natural images is not possible. Thus, throughout all our experiments, we did not use data augmentation.

5 | RESULTS

5.1 | Fully supervised fusion transformer results

Our experiments showed that when training both our Fusion Transformer and ResNet from scratch, the Fusion Transformer obtained competitive results without any pre-training on a small amount of images. In Table 2, we present the results of SF-ViT, ResNet and Fusion Transformer performance on the validation set when varying the number of image-based features used for training. With our Fusion Transformer architecture, we obtained our best results when using only PWR and CSI spectrograms, reaching a macro F1-score of 94.3%. With ResNet34, we also obtained the best results when using PWR and CSI spectrograms, reaching a macro F1-score of 94.9%. The two confusion matrices are shown in Figure 6.

Thus, ResNet34 seems to achieve slightly better performance for HAR. However, the Fusion Transformer can achieve competitive performance with less parameters when trained from scratch, without pre-training. The Fusion Transformer has 11.7 M trainable parameters against 12.4 M for the ResNet34. One benefit of the Fusion Transformer is that the number of trainable parameters is invariant to the addition of new image-based representations. Unlike the Fusion Transformer, the number of trainable parameters increases with ResNet when doing so.

We also assess the generalisation capability of the supervised Fusion Transformer model across different environments and participants (of different demographics). We use the PWR and CSI spectrograms (7 features) as input to the Fusion Transformer as these achieved the highest performance in Table 2. The same model parameters as in Section 4.3 are used and the model is trained for 100 epochs. In the first case, the activity samples across all six participants for Room 1 are used as training set (1886 samples) and the activity samples across all six participants in Room 2 are used as validation set (1011 samples). In the second case, 5 participants' (persons 1, 2, 4, 5

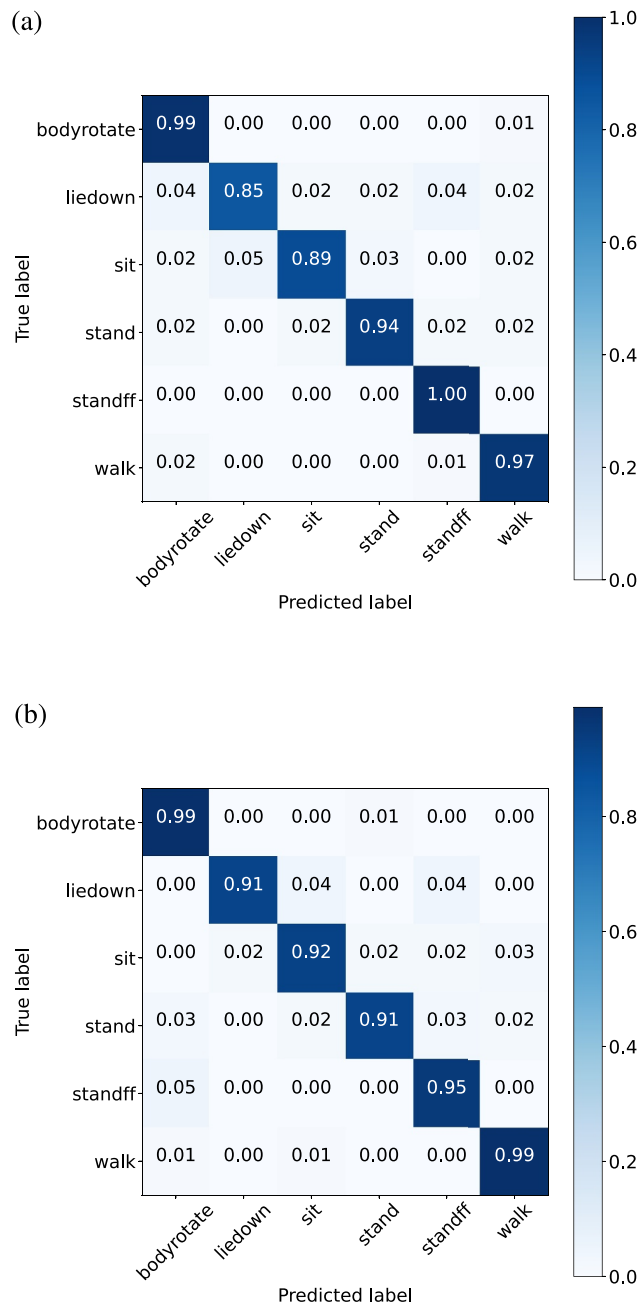


FIGURE 6 Visualisation of the confusion matrices of the models with highest scores: (a) Confusion matrix for the Fusion Transformer model and (b) Confusion matrix for the ResNet34 model.

and 6) activity samples are used as training set (2227 samples) while the validation set (670 samples) consists of the unseen activity samples from only one participant (person 3). Finally, in the third case, 3 participants' (persons 1, 2 and 3) activity samples are used as training set (1776 samples) and the validation set (1121 samples) consists of the activity samples from the remaining participants (i.e. persons 4, 5 and 6). The last two cases assess the model generalisation capability for HAR across people of different demographics. The results are shown in Figure 7 where an accuracy above 90% is observed in each

case, demonstrating that the supervised Fusion Transformer model generalises well across different environments and people.

6 | TOWARDS SELF-SUPERVISION

Transformers outperform many state-of-the-art models when trained on large scale datasets. In this work, we succeeded in achieving competitive results compared to ResNet while training our model from scratch. However, we believe that with self-supervision, the Fusion Transformer can outperform ResNet34 for HAR. Instead of training a model from scratch with weights initialised arbitrarily, the model can be pre-trained via different self-supervised learning (SSL) methods.

We propose a self-supervised method based on image masking as in [40]. However, instead of masking some parts of a natural image, in our approach, we mask multiple image features and we pre-train our model to predict the masked image features. The architecture used during the pre-training phase is presented in Figure 8, where a lightweight one-layer head (e.g. a linear layer) is used to predict the raw pixel values of the masked image features and it performs learning using a simple L1 loss function.

6.1 | Pre-training phase: Experimental setup

We pre-train our model with both PWR and CSI spectrograms, using all different views and image-based features. We masked

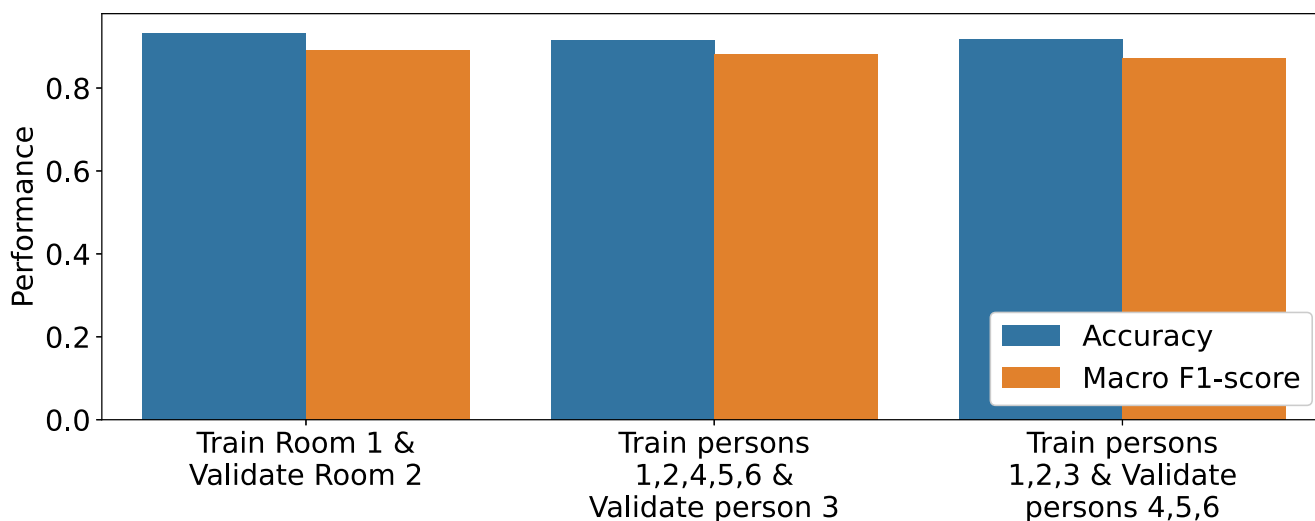


FIGURE 7 Assessing the generalisation capability of the Fusion Transformer model across different environments and participants (of different demographics) for the HAR task. HAR, human activity recognition

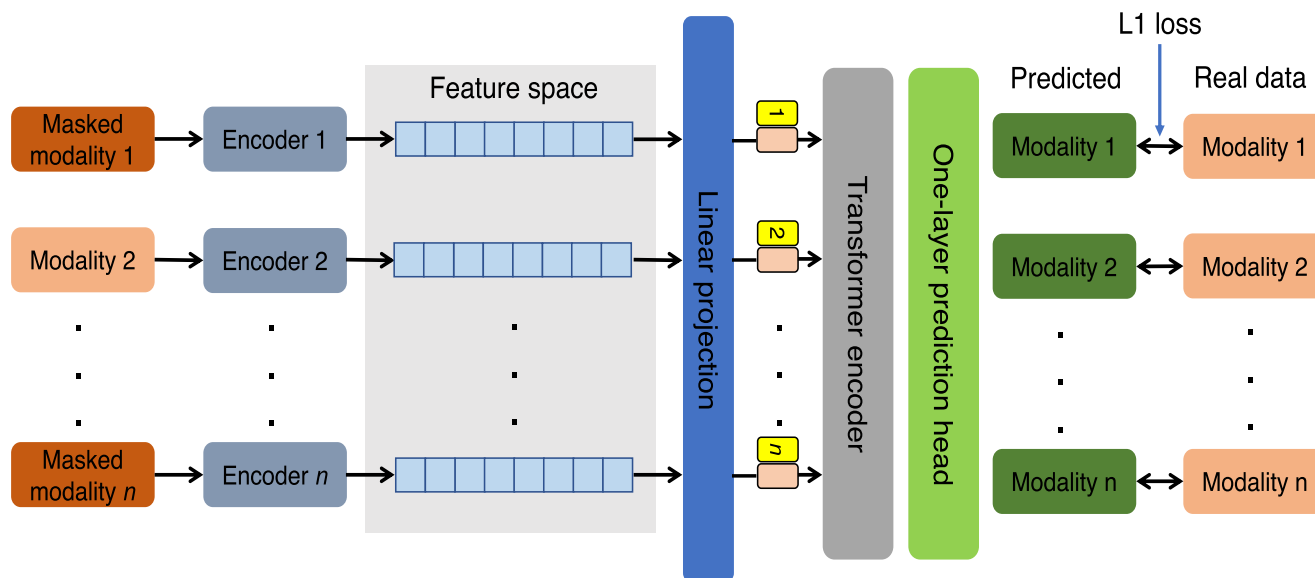


FIGURE 8 Self-supervised Fusion Transformer method. We randomly mask 60% of our modalities and we train a model to predict the masked modalities.

	Macro F1-score with different training set size D						
	1 sample per class (%)	2.5% of D (%)	5% of D (%)	10% of D (%)	15% of D (%)	20% of D (%)	Full D (%)
Fusion Transformer (with SSL)	56.30	77.00	84.50	89.70	90.40	91.20	95.90
Fusion Transformer (no SSL)	32.80	60.00	67	83.10	84.40	84.40	94.30
ResNet34	32.60	43.40	56.90	62.70	62.20	73.80	94.90

Note: The best results are shown in bold font.

60% of the image-based features and pre-train our model for 500 epochs. We use an AdamW optimiser and a multi-step learning rate scheduler. The batch size is fixed as 64, the base learning rate as $5e-4$, weight decay as 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a warm-up for 10 epochs.

6.2 | Fine-tuning phase: Experimental setup and results

Next, we fine-tune the pre-trained model in a supervised way. The strength of self-supervised learning is that we can fine-tune the pre-trained model on a smaller training set. This is particularly useful when labelling the data is time consuming and expensive. We train the model on different number of training samples: 1 sample per class, 5% (10 min), 10% (20 min), 15% (30 min), 20% (40 min) of the train set and also the full train set. We fine-tuned our model using the following hyper-parameters settings: an AdamW optimiser, a base learning rate fixed at $1e-3$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a warm-up for 10 epochs. We added multiple regularisation methods: a weight decay of 0.05 and a stochastic depth [41] ratio of 0.1. We report the results of our self-supervised method in Table 3.

ResNet as an architecture is not well-defined for SSL when having multimodal and multi-sensor data. Existing approaches involve contrastive learning methods [42, 43], which require multiple views for each modality, data augmentation techniques and pairs of negative/positive samples. A similar work has proposed a self-supervised contrastive pre-training method for passive Wi-Fi based activity recognition [44]. Although we can simply pre-train a model using a contrastive method with two views on many different CNN benchmark models, this framework is not well defined for multi-view and multimodal pre-training and led to worse results than those presented by our non pre-trained ResNet34. Thus, in the context of multi-sensor fusion with multi-views, we cannot rely on a ResNet architecture.

In this work, we have proposed a simple but yet very effective method for multimodal and multi-sensor self-supervised learning with a Fusion Transformer which outperforms the results obtained with a non pre-trained ResNet34 and a non pre-trained Fusion Transformer, regardless of the training set size. The strength of the Fusion Transformer is that it can be easily pre-trained with multiple views, sensors and modalities, thanks to the transformer architecture.

TABLE 3 Comparison between pre-trained and supervised Fusion Transformer for various sizes of training set D

7 | CONCLUSION

We proposed a new architecture for multimodal, multi-sensor passive Wi-Fi based HAR. Using signal processing, we extracted 15 image-based features from multiple sensors. With our Fusion Transformer architecture, we first embed each modality via an encoder and then pass it into our transformer network. The Fusion Transformer can fuse multiple image-based features and train a classifier to predict six daily activities performed by six participants. The best results of this model were achieved with PWR and CSI spectrograms, achieving competitive performance with ResNet34, but with less trainable parameters. We next demonstrated that with our proposed self-supervision technique, our pre-trained model outperformed non pre-trained ResNet34, achieving a macro F1-score of 95.9% when fine-tuned on the full training set. Furthermore, it outperformed the other models when fine-tuned with as little as 1% (2 min) of labelled training data with a macro F1-score of 56.3%, while the macro F1-score achieved with 20% (40 min) of training data was 91.2%. These results are promising given the need to collect training data for each new indoor environment.

AUTHOR CONTRIBUTION

Armand K. Koupai: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing – original draft, Writing – review & editing. **Mohammad J. Bocus:** Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualisation, Writing – review & editing. **Raul Santos-Rodriguez:** Funding acquisition, Project administration, Resources, Supervision. **Robert J. Piechocki:** Conceptualisation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualisation, Writing – review & editing. **Ryan McConville:** Conceptualisation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualisation, Writing – review & editing.

ACKNOWLEDGEMENT

This work was performed as a part of the OPERA Project, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/R018677/1.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in figshare at <https://doi.org/10.6084/m9.figshare.c.5551209.v1>, reference number [27].

ORCID

Mohammad J. Bocus  <https://orcid.org/0000-0001-7843-3445>

REFERENCES

- McConville, R., et al.: Vesta: a digital health analytics platform for a smart home in a box. *Future Generat. Comput. Syst.* 114, 106–119 (2021). <https://doi.org/10.1016/j.future.2020.07.046>
- Tan, B., et al.: Wi-Fi based passive human motion sensing for in-home healthcare applications. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), pp. 609–614. (2015)
- Serpush, F., et al.: Wearable sensor-based human activity recognition in the smart healthcare system. *Comput. Intell. Neurosci.* 2022, 1–31 (2022). [Online]. Available. <https://doi.org/10.1155/2022/1391906>
- Li, W., Tan, B., Piechocki, R.: Passive radar for opportunistic monitoring in e-health applications. *IEEE J. Transl. Eng. Health Med.* 6, 1–10 (2018). <https://doi.org/10.1109/jtehm.2018.2791609>. <https://ieeexplore.ieee.org/document/8269297/>
- Gavrilova, M.L., et al.: Kinect sensor gesture and activity recognition: new applications for consumer cognitive systems. *IEEE Consum. Electron. Mag.* 7(1), 88–94 (2018). <https://doi.org/10.1109/mce.2017.2755498>
- Hämäläinen, M., et al.: Ultra-wideband radar-based indoor activity monitoring for elderly care. *Sensors* 21(9), 3158 (2021). <https://doi.org/10.3390/s21093158>. <https://www.mdpi.com/1424-8220/21/9/3158>
- Bregar, K., Hrovat, A., Mohorčič, M.: UWB radio-based motion detection system for assisted living. *Sensors* 21(11), 3631 (2021). <https://doi.org/10.3390/s21113631>. <https://www.mdpi.com/1424-8220/21/11/3631>
- Thariq Ahmed, H.F., Ahmad, H.A.C.V.: Device free human gesture recognition using Wi-Fi CSI: a survey. *Eng. Appl. Artif. Intell.* 87, 103281 (2020). <https://doi.org/10.1016/j.engappai.2019.103281>. <https://www.sciencedirect.com/science/article/pii/S0952197619302441>
- Tian, Z., et al.: WiCatch: a Wi-Fi based hand gesture recognition system. *IEEE Access* 6, 16911–16923 (2018). <https://doi.org/10.1109/access.2018.2814575>
- Tan, S., Yang, J.: WiFinger: leveraging commodity WiFi for fine-grained finger gesture recognition. In: Proceedings Of the 17th ACM International Symposium On Mobile Ad Hoc Networking and Computing, Ser. MobiHoc 16, pp. 201–210. Association for Computing Machinery, New York (2016). <https://doi.org/10.1145/2942358.2942393>
- Ma, Y., et al.: SignFi: sign language recognition using WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2(1), 1–21 (2018). <https://doi.org/10.1145/3191755>
- Palipana, S., et al.: FallDeFi: ubiquitous fall detection using commodity Wi-Fi devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1(4), 1–25 (2018). <https://doi.org/10.1145/3161183>
- Damodaran, N., et al.: Device free human activity and fall recognition using WiFi channel state information (CSI). *CCF Trans. Pervasive Comput. Interact.* 2(1), 1–17 (2020). <https://doi.org/10.1007/s42486-020-00027-1>
- Alazrai, R., et al.: A dataset for Wi-Fi-based human-to-human interaction recognition. *Data Brief* 31, 105668 (2020). <https://doi.org/10.1016/j.dib.2020.105668>
- Alsaify, B.A., et al.: A dataset for Wi-Fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments. *Data Brief* 33, 106534 (2020). <https://doi.org/10.1016/j.dib.2020.106534>
- Bocus, M.J., et al.: Translation resilient opportunistic WiFi sensing. In: 2020 25th International Conference on Pattern Recognition, pp. 5627–5633. ICPR (2021)
- Li, W., et al.: On CSI and passive Wi-Fi radar for opportunistic physical activity recognition. *IEEE Trans. Wireless Commun.* 21(1), 607–620 (2022). <https://doi.org/10.1109/twc.2021.3098526>
- Li, W., et al.: A taxonomy of WiFi sensing: CSI vs passive WiFi radar. In: 2020 IEEE Globecom Workshops (GC Wkshps), pp. 1–6. (2020)
- Ma, Y., Zhou, G., Wang, S.: WiFi sensing with channel state information: a survey. *ACM Comput. Surv.* 52(3), 1–36 (2019). <https://doi.org/10.1145/3310194>
- Halperin, D., et al.: Tool release: gathering 802.11n traces with channel state information. *SIGCOMM Comput. Commun. Rev.* 41(1), 53 (2011). <https://doi.org/10.1145/1925861.1925870>
- Xie, Y., Li, Z., Li, M.: Precise power delay profiling with commodity WiFi. In: Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, Ser. MobiCom 15, pp. 53–64. ACM, New York (2015). <http://doi.acm.org/10.1145/2789168.2790124>
- Shi, F., Chetty, K., Julier, S.: Passive activity classification using just WiFi probe response signals. In: 2019 IEEE Radar Conference (RadarConf), pp. 1–6. (2019)
- Chen, Z., et al.: WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Trans. Mobile Comput.* 18(11), 2714–2724 (2019). <https://doi.org/10.1109/tmc.2018.2878233>
- Yousefi, S., et al.: A Survey of Human Activity Recognition Using WiFi CSI (2017). <https://arxiv.org/abs/1708.07129>
- Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria (2021)
- Bocus, M.J., et al.: OPERAnet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Sci. Data* 9(1), 474 (2022). <https://doi.org/10.1038/s41597-022-01573-2>
- Bocus, M.J., et al.: A comprehensive multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Figshare* (2022). <https://doi.org/10.6084/m9.figshare.c.5551209.v1>
- Yadav, S.K., et al.: A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowl. Base Syst.* 223, 106970 (2021). <https://doi.org/10.1016/j.knosys.2021.106970>. <https://www.sciencedirect.com/science/article/pii/S0950705121002331>
- Zou, H., et al.: WiFi and vision multimodal learning for accurate and robust device-free human activity recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 426–433. (2019)
- Memmesheimer, R., Theisen, N., Paulus, D.: Gimme signals: discriminative signal encoding for multimodal activity recognition. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10394–10401. (2020)
- Muaaz, M., et al.: WiWeHAR: multimodal human activity recognition using Wi-Fi and wearable sensing modalities. *IEEE Access* 8, 164453–164470 (2020). <https://doi.org/10.1109/access.2020.3022287>
- Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: ICCV 2021—International Conference on Computer Vision, Virtual, pp. 1–21. France (2021). <https://hal.archives-ouvertes.fr/hal-03323359>
- Bao, H., et al.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022). <https://openreview.net/forum?id=p-BhZSsZ59o4>
- Gong, Y., Chung, Y.-A., Glass, J.: AST: audio spectrogram transformer. *Proc. Interspeech* 2021, 571–575 (2021)
- Chen, X., Hsieh, C.-J., Gong, B.: When vision transformers outperform ResNets without pre-training or strong data augmentations. In: International Conference on Learning Representations (2022). <https://openreview.net/forum?id=LtKcMgGOeLt>

36. Foret, P., et al.: Sharpness-aware minimization for efficiently improving generalization. In: International Conference on Learning Representations (2021). <https://openreview.net/forum?id=6Tm1mposlrM>
37. Chetty, K., Smith, G.E., Woodbridge, K.: Through-the-wall sensing of personnel using passive bistatic WiFi radar at standoff distances. *IEEE Trans. Geosci. Rem. Sens.* 50(4), 1218–1226 (2012). <https://doi.org/10.1109/tgrs.2011.2164411>
38. Gamboa, J.C.B.: Deep Learning for Time-Series Analysis. (2017). <https://arxiv.org/abs/1701.01887>
39. Steiner, A.P., et al.: How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers. *Transactions on Machine Learning Research* (2022). <https://openreview.net/forum?id=4nPswr1KcP>
40. Xie, Z., et al.: SimMIM: A Simple Framework for Masked Image Modeling (2021)
41. Huang, G., et al.: Deep networks with stochastic depth. In: Leibe, B., et al. (eds.) *Computer Vision—ECCV 2016*, pp. 646–661. Springer International Publishing, Cham (2016)
42. Dao, S.D., et al.: Multi-label image classification with contrastive learning. (2021). <https://arxiv.org/abs/2107.11626>
43. Jaiswal, A., et al.: A survey on contrastive self-supervised learning. *Technologies* 9(1), 2 (2021). <https://doi.org/10.3390/technologies9010002>
44. Lau, H.-S., et al.: Self-supervised wifi-based activity recognition. (2021). <https://arxiv.org/abs/2104.09072>

How to cite this article: Koupai, A.K., et al.: Self-supervised multimodal fusion transformer for passive activity recognition. *IET Wirel. Sens. Syst.* 12(5-6), 149–160 (2022). <https://doi.org/10.1049/wss2.12044>