UNIVERSITY OF BRISTOL

Piechocki, R. J., Wang, X., & Bocus, M. J. (2022). Multimodal sensor fusion in the latent representation space. Manuscript submitted for publication. https://doi.org/10.48550/arXiv.2208.02183

Early version, also known as pre-print

Link to published version (if available):
10.48550/arXiv.2208.02183

Link to publication record in Explore Bristol Research
PDF-document

Multimodal sensor fusion in the latent representation space

Robert J. Piechocki[1,*]**, Xiaoyang Wang**[1]**, and Mohammud J. Bocus**[1]

[1]School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths, University of Bristol, Bristol, BS8 1UB, UK.
[*]r.j.piechocki@bristol.ac.uk

ABSTRACT

A new method for multimodal sensor fusion is introduced. The technique relies on a two-stage process. In the first stage, a multimodal generative model is constructed from unlabelled training data. In the second stage, the generative model serves as a reconstruction prior and the search manifold for the sensor fusion tasks. The method also handles cases where observations are accessed only via subsampling i.e. compressed sensing. We demonstrate the effectiveness and excellent performance on a range of multimodal fusion experiments such as multisensory classification, denoising, and recovery from subsampled observations.

Introduction

Controlled hallucination[1] is an evocative term referring to the Bayesian brain hypothesis[2]. It posits that perception is not merely a function of sensory information processing capturing the world as is. Instead, the brain is a predictive machine - it attempts to infer the causes of sensory inputs. To achieve this, the brain builds and continually refines its world model. The world model serves as a prior and when combined with the sensory signals will produce the best guess for its causes. Hallucination (uncontrolled) occurs when the sensory inputs cannot be reconciled with, or contradict the prior world model. This might occur in our model, and when it does, it manifests itself at the fusion stage with the stochastic gradient descent procedure getting trapped in a local minimum. The method presented in this paper is somewhat inspired by the Bayesian brain hypothesis, but it also builds upon multimodal generative modelling and deep compressed sensing.

Multimodal data fusion attracts academic and industrial interests alike[3] and plays a vital role in several applications. Automated driving is arguably the most challenging industrial domain[4]. Automated vehicles use a plethora of sensors: Lidar, mmWave radar, video and ultrasonic, and attempt to perform some form of sensor fusion for environmental perception and precise localization. A high-quality of final fusion estimate is a prerequisite for safe driving. Amongst other application areas, a notable mention deserves eHealth and Ambient Assisted Living (AAL). These new paradigms are contingent on gathering information from various sensors around the home to monitor and track the movement signatures of people. The aim is to build long-term behavioral sensing machine which also affords privacy. Such platforms rely on an array of environmental and wearable sensors, with sensor fusion being one of the key challenges.

In this contribution, we focus on a one-time snapshot problem (i.e. we are not building temporal structures). However, we try to explore the problem of multimodal sensor fusion from a new perspective, essentially, from a Bayesian viewpoint.The concept is depicted in Fig. 1, alongside two main groups of approaches to sensor fusion. Traditionally, sensor fusion for classification tasks has been performed at the decision level as in Fig. 1(a). Assuming that conditional independence holds, a pointwise product of final pmf (probability mass function) across all modalities is taken. Feature fusion, as depicted in Fig. 1(b), has become very popular with the advent of deep neural networks[3], and can produce very good results. Fig. 1(c) shows our technique during the fusion stage (Stage 2). Blue arrows indicate the direction of backpropagation gradient flow during fusion.

Contributions:

- A novel method for multimodal sensor fusion is presented. The method attempts to find the best estimate (*maximum a posteriori*) for the causes of observed data. The estimate is then used to perform specific downstream fusion tasks.

- The method can fuse the modalities under lossy data conditions i.e. when the data is subsampled, lost and/or noisy. Such phenomena occur in real-world situations such as the transmission of information wirelessly, or intentional subsampling to expedite the measurement (rapid MRI imaging and radar) etc.
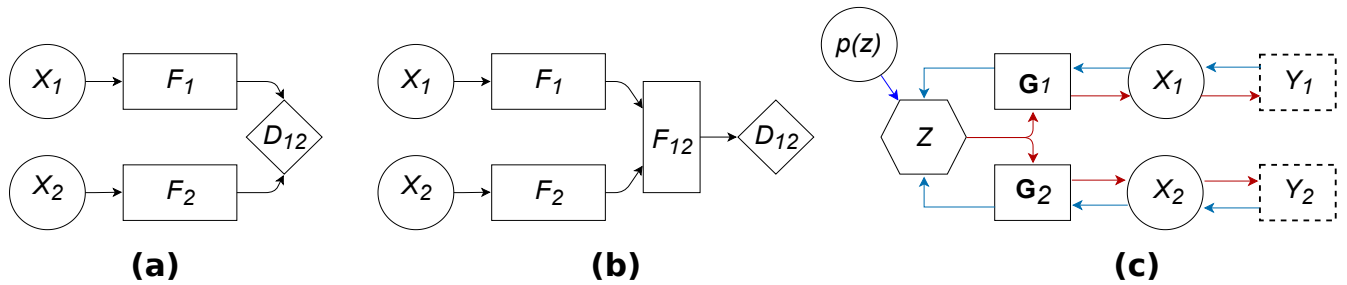
Figure 1. Multimodal Sensor Fusion: (**a**) Decision fusion, (**b**) Feature fusion, (**c**) Our technique: fusion in the latent representation with optional compressed sensing measurements; F features, $p(z)$ prior model, $\mathbf{G}$ generators, X complete data, Y subsampled data. For clarity $M = 2$ modalities are shown, the concept generalises to any M.

- It can leverage between modalities. A strong modality can be used to aid the recovery of another modality that is lossy or less informative (weak modality). This is referred to as asymmetric Compressed Sensing.

Related Work

In this section, we review the state-of-the-art in three areas directly relevant to our contribution: multimodal generative modeling, sensor fusion, and compressed sensing. One of the main aims of Multimodal Variational Autoencoders (MVAEs) is to learn shared representation across different data types in a fully self-supervised manner, thus avoiding the need to label a huge amount of data, which is time-consuming and expensive[5]. It is indeed a challenge to infer the low-dimensional joint representation from multiple modalities, which can ultimately be used in downstream tasks such as self-supervised clustering or classification. This is because the modalities may vastly differ in characteristics, including dimensionality, data distribution, and sparsity[6]. Recently, several methods have been proposed to combine multimodal data using generative models such as Variational Autoencoders (VAEs)[5,7–11]. These methods aim to learn a joint distribution in the latent space via inference networks and try to reconstruct modality-specific data, even when one modality is missing. In these works, a modality can refer to natural images, text, captions, labels or visual and non-visual attributes of a person. JMVAE (Joint Multimodal Variational Autoencoder)[9] makes use of a joint inference network to learn the interaction between two modalities and they address the issue of missing modality by training an individual (unimodal) inference network for each modality as well as a bimodal inference network to learn the joint posterior, based on the product-of-experts (PoE). They consequently minimize the distance between unimodal and multimodal latent distribution. On the other hand, MVAE[7], which is also based on PoE, considers only a partial combination of observed modalities, thereby reducing the number of parameters and improving the computational efficiency. Reference[8] uses the Mixture-of-Experts (MoE) approach to learn the shared representation across multiple modalities. The latter two models essentially differ in their choices of joint posterior approximation functions. MoPoE (Mixture-of-Products-of-Experts)-VAE[5] aims to combine the advantages of both approaches, MoE and PoE, without incurring significant trade-offs. DMVAE (Disentangled Multimodal VAE)[10] uses a disentangled VAE approach to split up the private and shared (using PoE) latent spaces of multiple modalities, where the latent factor may be of both continuous and discrete nature. CADA (Cross- and Distribution Aligned)-VAE[11] uses a cross-modal embedding framework to learn a latent representation from image features and classes (labels) using aligned VAEs optimized with cross- and distribution- alignment objectives.

In terms of multimodal/sensor fusion for human activity sensing using Radio-Frequency (RF), inertial and/or vision sensors, most works have considered either decision-level fusion or feature-level fusion. For instance, the work in[12] performs multimodal fusion at the decision level to combine the benefits of WiFi and vision-based sensors using a hybrid deep neural network (DNN) model to achieve good activity recognition accuracy for 3 activities. The model essentially consists of a WiFi sensing module (dedicated Convolutional Neural Network (CNN) architecture) and a vision sensing module (based on the Convolutional 3D model) for processing WiFi and video frames for unimodal inference, followed by a multimodal fusion module. Multimodal fusion is performed at the decision level (after both WiFi 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. Reference[13] presents a method for activity recognition, which leverages four sensor modalities, namely, skeleton sequences, inertial and motion capture measurements and WiFi 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 to a two-dimensional CNN (EfficientNet B2) for classification. The authors of[14] proposed a multimodal HAR system that leverages WiFi and wearable sensor modalities to jointly infer human activities. They collect Channel Sate Information (CSI) data from a standard WiFi Network Interface Card (NIC), alongside the user's local body movements via a wearable

Inertial Measurement Unit (IMU) consisting of an accelerometer, gyroscope, and magnetometer sensors. They compute 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 are separately extracted from the magnitude data and the MDS. The authors apply a feature-level fusion method which sequentially concatenates feature vectors that belong to the same activity sample. Finally supervised machine learning techniques are used to classify four activities, such as walking, falling, sitting, and picking up an object from the floor.

Compared to the aforementioned works[12–14] which consider supervised models with feature-level fusion or decision-level fusion, our technique, in contrast, performs multimodal sensor fusion in the latent representation space leveraging a self-supervised generative model. Our method is different from current multimodal generative models such as those proposed in[5,7–9] in the sense that it can handle cases where observations are accessed only via subsampling (i.e. compressed sensing with significant loss of data and no data imputation). And crucially our technique attempts to directly compute the MAP (*maximum a posteriori*) estimate.

The presented method is related to and builds upon Deep Compressed Sensing (DCS) techniques[15,16]. DCS, in turn, is inspired by Compressed Sensing (CS)[17,18]. In CS, we attempt to solve what appears to be an underdetermined linear system, yet the solution is possible with the additional prior sparsity constraint on the signal: $\min L0$. Since $L0$ is non-convex, $L1$ is used instead to provide a convex relaxation, which also promotes sparsity and allows for computationally efficient solvers. DCS, in essence, replaces the $L0$ prior with a low dimensional manifold, which is learnable from the data using generative models. Concurrently to DCS, Deep Image Prior[19] was proposed. It used un-trained CNNs to solve a range of inverse problems in computer vision (image inpainting, super-resolution, denoising).

Methods

Assume the data generative process so that latent and common cause Z gives rise to X_m, which in turn produces observed Y_m, i.e. $Z \rightarrow X_m \rightarrow Y_m$ forms a Markov chain. Here, X_m is the full data pertaining to m^{th} modality, $m \in \{1, \ldots, M\}$. Crucially, the modalities collect data simultaneously "observing" the same scene. As an example, in this work, we consider the different views (obtained via multiple receivers) from the opportunistic CSI WiFi radar as different modalities. The variable Z encodes the semantic content of the scene and is typically of central interest. Furthermore, X_m is not accessed directly, but is observed via a *subsampled* Y_m. This is a compressed sensing setup: $Y_m = \chi_m(X_m)$: χ_m is a deterministic and known (typically many-to-one) function. The only condition we impose on χ_m is to be Lipschitz continuous. With the above, the conditional independence between modalities holds (conditioned on Z). Therefore, the joint density factors as:

$$p(z, x_{1:M}, y_{1:M}) = p(z) \prod_{m=1}^{M} p(y_m|x_m) p(x_m|z). \tag{1}$$

The main task in this context is to produce the best guess for latent Z, and possibly, to recover the full signal(s) X_m, given subsampled data $Y_{1:M}$. We approach the problem in two stages. First we build a joint model which approximates equation (1), and will be instantiated as a Multimodal Variatational Autoencoer (M-VAE). More specifically, the M-VAE will provide an approximation to $p_{\phi_{1:M}, \psi_{1:M}}(z, x_{1:M})$, parameterized by deep neural networks $\{\phi_1, \ldots, \phi_M\}$, $\{\psi_1, \ldots, \psi_M\}$, referred to as *encoders* and *decoders*, respectively. The trained M-VAE will then be appended with $p_{\chi_m}(y_m|x_m)$ for each modality m: $\{\chi_1, \ldots, \chi_M\}$ referred to as *samplers*. In the second stage, we use the trained M-VAE and $\chi_{1:M}$ to facilitate the fusion and reconstruction tasks. Specifically, our sensor fusion problem amounts to finding the maximum a posteriori (MAP) $\hat{z}_{MAP}$ estimate of the latent cause for a given (i^{th}) data point $Y_{1:M} = y_{1:M}^{(i)}$:

$$\hat{z}_{MAP} = \arg\max_z p\left(z|Y_{1:M} = y_{1:M}^{(i)}\right), \tag{2}$$

where,

$$p\left(z|Y_{1:M} = y_{1:M}^{(i)}\right) \propto p(z) \prod_{m=1}^{M} \int_{X_m} p(Y_m = y_m^{(i)}|x_m) p(x_m|z) \, dx_m. \tag{3}$$

The above MAP estimation problem is hard, and we will resort to approximations detailed in the sections below.

Multimodal VAE

The first task is to build a model of equation (1). As aforementioned, this will be accomplished in two steps. Firstly, during the training stage we assume access to full data $X_{1:M}$, therefore training an approximation to $p_{\phi_{1:M}, \psi_{1:M}}(z, x_{1:M})$ is a feasible task. The marginal data log-likelihood for the multimodal case is:

$$\log p(x_{1:M}) = D_{KL}(q(z|x_{1:M})||p(z|x_{1:M})) \tag{4}$$

$$+ \left[\sum_{X_m} \mathbb{E}_{z \sim q(z|x_{1:M})} \log p(x_m|z) - \mathbb{E}_{z \sim q(z|x_{1:M})} \log \frac{q(z|x_{1:M})}{p(z)} \right], \tag{5}$$

where D_{KL} is the Kullback–Leibler (KL) divergence. The first summand in equation (5), i.e. the sum over modalities follows directly from the conditional independence. And since KL is non-negative, equation (5) represents the lower bound (also known as Evidence Lower Bound - ELBO) on the log probability of the data (and its negative is used as the loss for the M-VAE). There exist a body of work on M-VAEs, the interested reader is referred to[5,7–9] for details and derivation. The key challenge in training M-VAEs is the construction of variational posterior $q(z|x_{1:M})$. We dedicate a section in the Supplementary Information document S1 to the discussion on choices and implications for the approximation of variational posterior. Briefly, we consider two main cases: a missing data case – i.e. where particular modality data might be missing ($X_m = x_m^{(i)} = \emptyset$); and the full data case. The latter is straightforward and is tackled by enforcing a particular structure of the encoders. For the former case variational Product-of-Experts (PoE) is used:

$$q_\Phi(z|x_{1:M}) = p(z) \prod_{m=1}^{M} q_{\phi_m}(z|x_m). \tag{6}$$

Should the data be missing for any particular modality, $q_{\phi_m}(z|x_m) = 1$ is assumed. Derivation of equation (6) can be found in the Supplementary Information document S1.

Fusion on the M-VAE prior

Recall the sensor fusion problem as stated in equation (2). The $p(z)$ is forced to be isotropic Gaussian by M-VAE, and the remaining densities are assumed to be Gaussian. Furthermore, we assume that $p(x_m|z) = \delta(x_m - \psi_m(z))$. Therefore equation (2) becomes:

$$\hat{z}_{MAP} = \arg\max_z p\left(z|Y_{1:M} = y_{1:M}^{(i)}\right) \propto \exp\left(-\|z\|^2\right) \prod_{m=1}^{M} \exp\left(-\frac{1}{2\sigma_m^2}\|y_m^{(i)} - \chi_m(\psi_m(z))\|^2\right). \tag{7}$$

Hence, the objective to minimize becomes:

$$\mathscr{L}(z) = \lambda_0\|z\|^2 + \sum_{m=1}^{M} \lambda_m\|y_m^{(i)} - \chi_m(\psi_m(z))\|^2. \tag{8}$$

Recall that the output of the first stage is $p(z)$ and the decoders $\prod_m p_{\psi_n}(x|z)$ are parametrized by $\{\psi_{1:M}\}$, $\{\lambda_{0:M}\}$ are constants. The MAP estimation procedure consists of backpropagating through the sampler χ_m and decoder ψ_m using Stochastic Gradient Descent (SGD). In this step $\{\psi_{1:M}\}$ are non-learnable, i.e. jointly with χ_m are some non-linear known (but differentiable) functions.

$$z \leftarrow z - \eta_0 \nabla_z(\|z\|^2) - \sum_{m=1}^{M} \eta_m \nabla_z(\|y_m^{(i)} - \chi_m(\psi_m(z))\|^2). \tag{9}$$

The iterative fusion procedure is initialized by taking a sample from the prior $z^0 \sim p(z)$, $\{\eta_{0:M}\}$ are learning rates. One or several SGD steps are taken for each modality in turn. The procedure terminates with convergence - see Algorithm 1. In general, the optimization problem as set out in equation (8) is non-convex. Therefore, there are no guarantees of convergence to the optimal point ($\hat{z}_{MAP}$). We deploy several strategies to minimize the risk of getting stuck in a local minimum. We consider multiple initialization points (a number of points sampled from the prior with Stage 2 replicated for all points). In some cases it might be possible to sample from: $z^0 \sim p(z) \prod p\left(z|X = \check{x}_m^{(j)}\right)$. Depending on modality, this might be possible with data imputation ($\check{x}_m$ are imputed data). The final stage will depend on a particular task (multisensory classification/reconstruction), but in all cases it will take $\hat{z}_{MAP}$ as an input. In our experiments, we observe that the success of Stage 2 is crucially dependent on the quality of M-VAE.

Experiments

In this work, we investigate the performance of the proposed method on two datasets for multimodal sensor fusion and recovery tasks: i) a synthetic "toy protein" dataset and ii) a passive WiFi radar dataset intended for Human Activity Recognition (HAR).

Algorithm 1 Multimodal Sensor Fusion in the Latent Representation Space (SFLR)

1: Training data: $\mathscr{D}_{\mathscr{T}} \equiv \{X_{1:M}^{(1:I)}\}$, Test data $\mathscr{D}_{\mathscr{P}} \equiv \{X_{1:M}^{(1:J)}\}$, Samplers $\{\chi_{1:M}\}$
2: **Stage 1**: Train M-VAE using $\mathscr{D}_{\mathscr{T}}$
3: Output: $p(z)$, Encoders $\{\phi_{1:M}\}$, Decoders $\{\psi_{1:M}\}$
4: **Stage 2**: Fusion
5: $y_{1:M}^{(i)} \sim \mathscr{D}_{\mathscr{P}}$
6: Sample the initial point $z^0 \sim p(z)$
7: **while** not converged **do**
8: $\quad z \leftarrow z - \eta_0 \nabla_z(\|z\|^2) - \eta_1 \nabla_z(\|y_1^{(i)} - \chi_1(\psi_1(z))\|^2)$
9: $\quad z \leftarrow z - \eta_0 \nabla_z(\|z\|^2) - \eta_2 \nabla_z(\|y_2^{(i)} - \chi_2(\psi_2(z))\|^2)$
10: $\quad \vdots$
11: $\quad z \leftarrow z - \eta_0 \nabla_z(\|z\|^2) - \eta_M \nabla_z(\|y_M^{(i)} - \chi_M(\psi_M(z))\|^2)$
12: **end while**
13: $\hat{z}_{MAP} \leftarrow z$
14: Downstream tasks: $\hat{x}_m = \psi_m(\hat{z}_{MAP})$, classification tasks K-NN$(\hat{z}_{MAP})$

Synthetic toy protein dataset

A synthetic dataset containing two-dimensional (2D) protein-like data samples with two modalities is generated. The latent distribution $p(z), z \in \mathbb{R}^4$ is a Gaussian mixture model with 10 components, simulating 10 different "classes" for samples. For each modality, the data generative model $p(x_m|z), x_m \in \mathbb{R}^N$ is a one-layer multilayer perceptron (MLP) with random weights. Here $m = 1, 2$ represents two modalities. 10,000 pairs of samples are generated using the generative model, with the protein size $N = 32$. Fig. 2(a) shows an instance of the 2D protein data with $N = 64$.
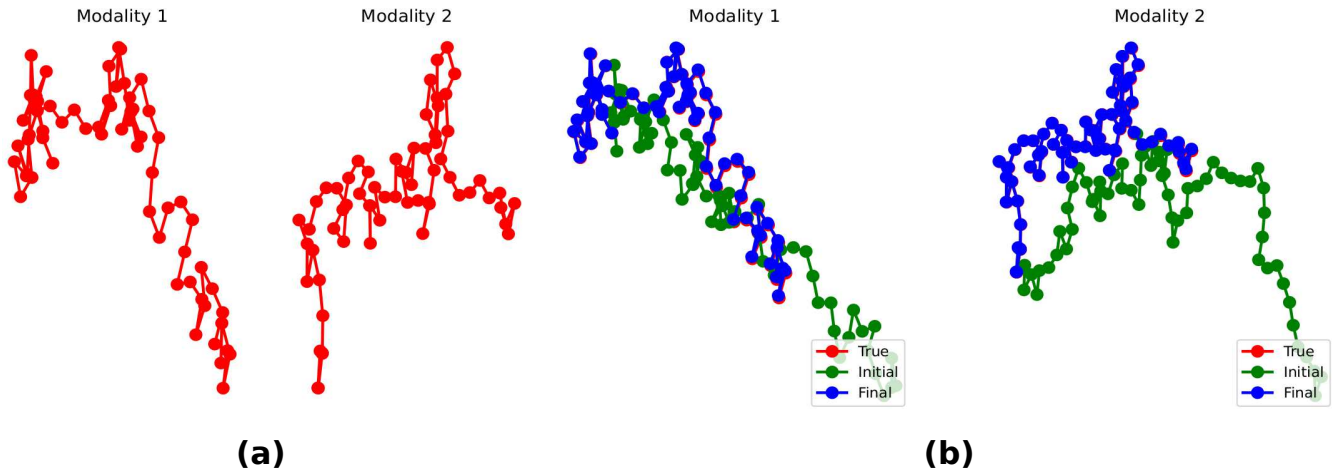

Figure 2. (**a**) Generated toy proteins examples ($N = 64$) and (**b**) reconstruction from compressed sensing observations. With 2 out of 64 measurements (3.125%), near perfect reconstruction is possible even though the modalities are individually subsampled.

Passive WiFi radar dataset

We use the OPERAnet[20] dataset which was collected with the aim to evaluate human activity recognition (HAR) and localization techniques with measurements obtained from synchronized Radio-Frequency (RF) devices and vision-based sensors. 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 the chair ("stand"), laying down on the floor ("laydown"), standing from the floor ("standff"), upper body rotation ("bodyrotate"), and walking ("walk"). We convert the raw time-series CSI data from the WiFi sensors into the image-like format, namely, spectrograms using signal processing techniques. More details are available in Section S2 of the Supplementary Information document. 2,906 spectrogram samples (each of 4s duration window) were generated for the 6 human activities and 80% of these were used as training data while the remaining 20% as testing data
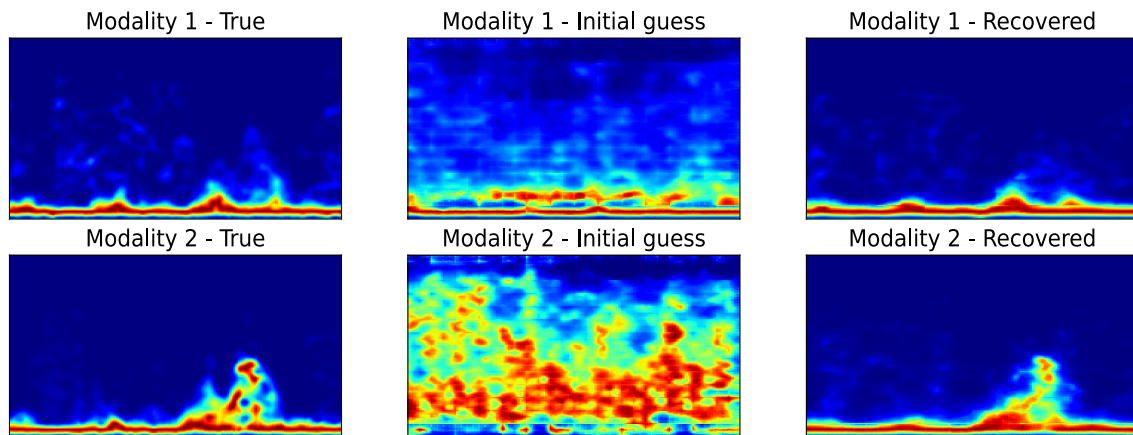
Figure 3. Illustration of spectrogram recovery (for sitting down activity) using compressed sensing with measurements as low as 784 out of 50,176 (1.56%). No additive white Gaussian noise is considered. The left column shows the true spectrogram sample, the middle column shows reconstruction with an initial guess (no optimization) while the right column shows reconstruction with $\hat{Z}_{MAP}$.

(random train-test split).

Results and Discussion

Classification results of WiFi CSI spectrograms for HAR

In this section, we evaluate the HAR sensor fusion classification performance under a few-shot learning scenario, with 1, 5 and 10 labelled examples per class. These correspond to 0.05%, 0.26% and 0.51% of labelled training samples, respectively. We randomly select 80% of the samples in the dataset as the training set and the remaining 20% is used for validation. The average F_1-macro scores for the HAR performance are shown in Table 1 for different models. To allow for a fair comparison, the same random seed was used in all experiments with only two modalities (processed spectrograms data obtained from two different receivers).

Prior to training our model (see Supplementary Fig. S1), the spectrograms were reshaped to typical image dimensions of size $(1 \times 224 \times 224)$. Our model was trained for 1,000 epochs using the training data with a fixed KL scaling factor of $\beta = 0.02$. The encoders comprised of the ResNet-18 backbone with the last fully-connected layer dimension having a value of 512. For the decoders, corresponding CNN deconvolutional layers were used to reconstruct the spectrograms from each modality with the same input dimension. The latent dimension, batch size, and learning rate are set at 64, 64, and 0.001, respectively. In the second stage, the generative model serves as a reconstruction prior and the search manifold for the sensor fusion tasks. Essentially, in this stage, we obtain the maximum a posteriori estimate of $\hat{z}_{MAP}$ through the process described in Algorithm 1. The final estimate of the class is produced by K-NN in the latent representation space, with labelled examples sampled from the training set.

To benchmark our technique we investigate the performance of other state-of-the-art sensor fusion techniques. The feature-fusion is represented by CNN models (1-channel CNN, 2-channel CNN, dual-branch CNN). All are trained in a conventional supervised fashion from scratch using the ResNet-18 backbone and a linear classification head is appended on top of it consisting of a hidden linear layer of 128 units and a linear output layer of 6 nodes (for classifying 6 human activities). The dual-input CNN refers to the case where the embeddings from the two modalities' CNNs are concatenated, and a classification head is then added (as illustrated in Fig. 1(b)). The "Probability Fusion" (decision fusion) model refers to a score-level fusion method where the classification probabilities (P_1 and P_2) from each modality are computed independently (using an output SoftMax layer) and then combined using the product rule (this is optimal given conditional independence). These models are fine-tuned with labelled samples over 200 epochs, with a batch size of 64 and the Adam optimizer was used with learning rate of 0.0001, weight decay of 0.001 and $\beta_1 = 0.95$, $\beta_2 = 0.999$.

It can be observed from Table 1 that our method significantly outperforms all other conventional feature and decision fusion methods. The confusion matrix for HAR classification using our SFLR (Sensor Fusion in the Latent Representation space) model is shown in Fig. S8 in the Supplementary Information document for the case when only ten labelled examples are used at the (classification) fusion stage.

Sensor fusion from subsampled observations

Next, we evaluate the recovery performance of spectrograms under different numbers of compressed sensing measurements. The measurement function χ_m is a matrix initialized randomly and we assume that there is no additive Gaussian noise. The Adam optimizer is used to optimize $\hat{z}_{MAP}$ with a learning rate of 0.01. The algorithm is run for 10,000 iterations. After the loss in equation (8) has converged during the optimization process, the samples are decoded/recovered for modality 1 and modality 2 using their respective decoders $\hat{x}_m = \psi_m(\hat{z}_{MAP})$. Table 2 shows the compressed sensing results when a batch of 50 images is taken from the testing dataset and evaluated under different number of measurements (without noise). It can be observed that the samples can be recovered with very low reconstruction error when the number of measurements is as low as 196 (0.39%). An illustration is also shown in Fig. 3 where very good reconstruction is observed for the case when the number of measurements is equal to 784. More illustrations are shown in Fig. S7 in the Supplementary Information document, with further experimental results in Sections S4, S5, S6.

Toy protein classification

Similarly to the experiments on the OPERAnet dataset, we perform two tasks, classification and sensor fusion from compressed sensing observations, on the synthetic toy protein dataset.

As mentioned previously, the toy protein dataset contains 10 classes. The dataset is split into a training set and a test set, containing 80% and 20% of samples, respectively. We evaluate the classification performance under a few-shot learning setting, using 1, 5 or 10 labelled samples per class. The few-shot classification via the SFLR model consists of two stages. In the first stage, the M-VAE is trained in an unsupervised manner using the training set. Using the maximum a posterior $\hat{z}_{MAP}$ and a few labels, the K-NN classifier is applied to the latent representation space. Here the *encoder* and *decoder* in M-VAE are two-layer MLPs, with 16 neurons in the hidden layer.

We compare the SFLR method with 4 baseline models. The single modality model only considers one modality without sensor fusion. The probability fusion model independently computes the classification probability for each modality, which is a representative model for decision-fusion (Fig. 1(a)). The dual-branch feature fusion model concatenates the embedding of two modalities before the classification layer, which is a feature fusion method (Fig. 1(b)). All baseline models are trained in a supervised manner, with the same neural network structure as the *encoder*. Table 3 shows the $F1$-macro scores for different methods on the test set. On the 10-class protein dataset, SFLR outperforms other sensor fusion models using limited labelled samples.

Sensor fusion from subsampled toy proteins

Another advantage of the proposed SFLR model is that it can fuse modalities in subsampled cases. We use a set of *samplers* $\chi_{1:M}$ to simulate the subsampled observations. The measurement function χ_m is a matrix initialized randomly. Here we use 10 initialization points to reduce the risk of getting trapped in a local minimum (points sampled from the prior with Stage 2 replicated for all of them). Fig. 2(b) shows the recovered protein from subsampled observations, with only 2 measurements for each modality. Both modalities are successfully recovered from the latent representation space, even though the initial guess z^0 is far from the true modality. Note that the proteins in Fig. 2 have a higher dimension than in the dataset, showing the robustness of the SFLR method. Table 4 shows the average reconstruction error of the synthetic protein dataset using different subsamplers. The reconstruction error reduced significantly when having 2 measurements for each modality, showing superior sensor fusion abilities.

The Supplementary Information document (see Section S7) contains additional experiments, including tasks showcasing the ability to leverage between modalities, where a strong modality can be used to aid the recovery of a weak modality. It also presents the performance under subsampled and noisy conditions.

Conclusions and Broader Impacts

The paper presents a new method for sensor fusion. Specifically, we demonstrate the effectiveness of classification and reconstruction tasks from radar signals. The intended application area is human activity recognition, which serves a vital role in the E-Health paradigm. New healthcare technologies are the key ingredient to battling spiralling costs of provisioning health services that beset a vast majority of countries. Such technologies in a residential setting are seen as a key requirement in empowering patients and imbuing a greater responsibility for own health outcomes. However, we acknowledge that radar and sensor technologies also find applications in a military context. Modern warfare technologies (principally defensive) could potentially become more apt if they were to benefit from much-improved sensor fusion. We firmly believe that, on balance, it is of benefit to the society to continue the research in this area in the public eye.

Data availability

The raw dataset used in the Passive WiFi radar experiments is available from[20]. The toy protein dataset is not publicly available at this time but can be made available from the authors upon reasonable request.

References

1. Seth, A. *Being You: A New Science of Consciousness (The Sunday Times Bestseller)* (Faber & Faber, 2021).

2. Doya, K., Ishii, S., Pouget, A. & Rao, R. P. N. *Bayesian Brain: Probabilistic Approaches to Neural Coding* (The MIT Press, 2007).

3. Gao, J., Li, P., Chen, Z. & Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Comput.* **32**, 829–864, DOI: 10.1162/neco_a_01273 (2020).

4. Wang, Z., Wu, Y. & Niu, Q. Multi-sensor fusion in automated driving: A survey. *IEEE Access* **8**, 2847–2868, DOI: 10.1109/ACCESS.2019.2962554 (2020).

5. Sutter, T. M., Daunhawer, I. & Vogt, J. E. Generalized multimodal ELBO. Preprint at https://arxiv.org/abs/2105.02470 (2021).

6. Minoura, K., Abe, K., Nam, H., Nishikawa, H. & Shimamura, T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods* **1**, 100071, DOI: https://doi.org/10.1016/j.crmeth.2021.100071 (2021).

7. Wu, M. & Goodman, N. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 5580–5590 (Curran Associates Inc., Red Hook, NY, USA, 2018).

8. Shi, Y., N, S., Paige, B. & Torr, P. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 32 (Curran Associates, Inc., 2019).

9. Suzuki, M., Nakayama, K. & Matsuo, Y. Joint multimodal learning with deep generative models. Preprint at https://arxiv.org/abs/1611.01891 (2016).

10. Lee, M. & Pavlovic, V. Private-shared disentangled multimodal VAE for learning of latent representations. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, 1692–1700, DOI: 10.1109/CVPRW53098.2021.00185 (Computer Vision Foundation / IEEE, 2021).

11. Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T. & Akata, Z. Generalized zero- and few-shot learning via aligned variational autoencoders. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8239–8247, DOI: 10.1109/CVPR.2019.00844 (2019).

12. 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)*, 426–433, DOI: 10.1109/CVPRW.2019.00056 (2019).

13. 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)*, 10394–10401, DOI: 10.1109/IROS45743.2020.9341699 (2020).

14. Muaaz, M., Chelli, A., Abdelgawwad, A. A., Mallofré, A. C. & Pätzold, M. WiWeHAR: Multimodal human activity recognition using Wi-Fi and wearable sensing modalities. *IEEE Access* **8**, 164453–164470, DOI: 10.1109/ACCESS.2020.3022287 (2020).

15. Bora, A., Jalal, A., Price, E. & Dimakis, A. G. Compressed sensing using generative models. In Precup, D. & Teh, Y. W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 537–546 (PMLR, 2017).

16. Wu, Y., Rosca, M. & Lillicrap, T. Deep compressed sensing. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 (PMLR, 2019).

17. Candès, E. J., Romberg, J. K. & Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. on Pure Appl. Math.* **59**, 1207–1223, DOI: https://doi.org/10.1002/cpa.20124 (2006).

18. Donoho, D. Compressed sensing. *IEEE Transactions on Inf. Theory* **52**, 1289–1306, DOI: 10.1109/TIT.2006.871582 (2006).

19. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

20. Bocus, M. J. *et al.* OPERAnet: A multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. Preprint at https://arxiv.org/abs/2110.04239 (2021).

21. 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*, MobiCom '15, 53–64, DOI: 10.1145/2789168.2790124 (ACM, New York, NY, USA, 2015).

22. Halperin, D., Hu, W., Sheth, A. & Wetherall, D. Tool release: Gathering 802.11n traces with channel state information. *SIGCOMM Comput. Commun. Rev.* **41**, 53, DOI: 10.1145/1925861.1925870 (2011).

23. Bocus, M. J. *et al.* Translation resilient opportunistic WiFi sensing. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 5627–5633, DOI: 10.1109/ICPR48806.2021.9412263 (2021).

Acknowledgements

Author contributions statement

All authors, R.P, X.W and M.B, contributed equally to this work. The main tasks involved conceiving and conducting the experiments, algorithm implementation, analysis, validation and interpretation of results, and finally preparing and reviewing the manuscript.

Additional information

Competing interests
The authors declare no competing interests.

Figure legends

1. Multimodal Sensor Fusion: (**a**) Decision fusion, (**b**) Feature fusion, (**c**) Our technique: fusion in the latent representation with optional compressed sensing measurements; F features, $p(z)$ prior model, **G** generators, X complete data, Y subsampled data. For clarity $M = 2$ modalities are shown, the concept generalises to any M.

2. (**a**) Generated toy proteins examples ($N = 64$) and (**b**) reconstruction from compressed sensing observations. With 2 out of 64 measurements (3.125%), near perfect reconstruction is possible even though the modalities are individually subsampled.

3. Illustration of spectrogram recovery (for sitting down activity) using compressed sensing with measurements as low as 784 out of 50,176 (1.56%). No additive white Gaussian noise is considered. The left column shows the true spectrogram sample, the middle column shows reconstruction with an initial guess (no optimization) while the right column shows reconstruction with $\hat{Z}_{MAP}$.

Table 1. Few-shot learning sensor fusion classification results (F_1 macro) for Human Activity Recognition.

Model	1 example	5 examples	10 examples
2-channel CNN	0.427272	0.570888	0.618501
1-channel CNN (Modality 1)	0.349084	0.451328	0.504462
1-channel CNN (Modality 2)	0.446554	0.600084	0.605678
Probability fusion (product rule)	0.440414	0.584726	0.641922
Dual-branch CNN	0.508243	0.568795	0.575914
SFLR (ours)	**0.652699**	**0.718180**	**0.737507**

Table 2. Compressed sensing mean reconstruction error over a batch of 50 WiFi spectrogram data samples (No additive Gaussian noise). An illustration is shown in Fig. 3.

No. of measurements	Modality 1	Modality 2
1 (0.002%)	0.03118854	0.15024841
10 (0.02%)	0.00938917	0.02824161
196 (0.39%)	0.00348606	0.00613665
784 (1.56%)	0.00305005	0.00505758
1,568 (3.125%)	0.00284343	0.00489433

Table 3. Few-shot learning sensor fusion classification results (F_1 macro) for synthetic proteins.

Model	1 example	5 examples	10 examples
Single modality (Modality 1)	0.3188	0.4342	0.5843
Single modality (Modality 2)	0.3221	0.4849	0.5555
Probability fusion (product rule)	0.2256	0.3736	0.3836
Dual-branch feature fusion	0.3769	0.4973	0.5953
SFLR (ours)	**0.4183**	**0.5501**	**0.6120**

Table 4. Compressed sensing mean reconstruction error over a batch of 100 protein samples.

No. of Measurements	Modality 1 (10^{-5})	Modality 2 (10^{-5})
1 (3.125%)	4,622.4	4,923.5
2 (6.250%)	22.5	27.9
4 (12.500%)	7.1	7.4
8 (25.000%)	2.3	2.7

Appendix

S1 Approximations to variational posterior

Given the objective, the variational joint posterior $q_\phi(z|x_{1:M})$ can be learned by training one single encoder network that takes all modalities $X_{1:M}$ as input to explicitly parametrize the joint posterior. This is our baseline model and an example for $M = 2$ modalities is given in Fig. S1. However, this approach requires all modalities to be present at all times, thus making cross-modal generation difficult. Alternatively, the joint variational posterior can be modelled using the following approaches:

Variational Product of Experts

In this section we reproduce the arguments from[7]. The first option is to approximate the joint variational posterior as a product:

$$q_\phi(z|x_{1:M}) \equiv p(z) \prod_{m=1}^{M} q_{\phi_m}(z|x_m). \tag{S10}$$

In case of a missing expert, we assume: $q_{\phi_m}(z|x_m) = 1$.

For a system of N modalities, 2^N inference networks need to be specified, $q(z|X)$ for each subset of modalities $X \subseteq \{X_1, X_2, \ldots, X_M\}$. The optimal inference network $q(z|x_1, \ldots, x_N)$ would be the true posterior $p(z|x_1, \ldots, x_N)$. The conditional independence assumptions in the generative model imply a relation among joint- and single-modality posteriors[7]:

$$\begin{aligned}
p(z|x_1, \ldots, x_N) &= \frac{p(x_1, \ldots, x_N|z)p(z)}{p(x_1, \ldots, x_N)} \\
&= \frac{p(z)}{p(x_1, \ldots, x_N)} \prod_{i=1}^{N} p(x_i|z) \\
&= \frac{p(z)}{p(x_1, \ldots, x_N)} \prod_{i=1}^{N} \frac{p(z|x_i)p(x_i)}{p(z)} \\
&= \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)} \cdot \frac{\prod_{i=1}^{N} p(x_i)}{p(x_1, \ldots, x_N)} \\
&\propto \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)}.
\end{aligned} \tag{S11}$$

If we approximate $p(z|x_i)$ with $q(z|x_i) \equiv \tilde{q}(z|x_i)p(z)$, where $\tilde{q}(z|x_i)$ is the underlying inference network, the quotient term can be omitted[7]:

$$\begin{aligned}
p(z|x_1, \ldots, x_N) &\propto \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)} \\
&\approx \frac{\prod_{i=1}^{N} [\tilde{q}(z|x_i)p(z)]}{\prod_{i=1}^{N-1} p(z)} \\
&= p(z) \prod_{i=1}^{N} \tilde{q}(z|x_i).
\end{aligned} \tag{S12}$$

Equation (S12) implies that we can use a Product-of-Experts (PoE), including a "prior expert" (e.g., spherical Gaussian), as the approximating distribution for the joint-posterior. This derivation is easily extended to any subset of modalities yielding $q(z|X) \propto p(z) \prod_{x_i \in X} \tilde{q}(z|x_i)$.

Variational Mixture of Experts

$$q_\phi(z|x_{1:M}) \equiv \sum_{m=1}^{M} \frac{1}{M} q_{\phi_m}(z|x_m), \tag{S13}$$

where the above assumes an equitable distribution of power among experts. Non-uniform weights can also be used. Missing expert: $q_{\phi_m}(z|x_m) = 0$.
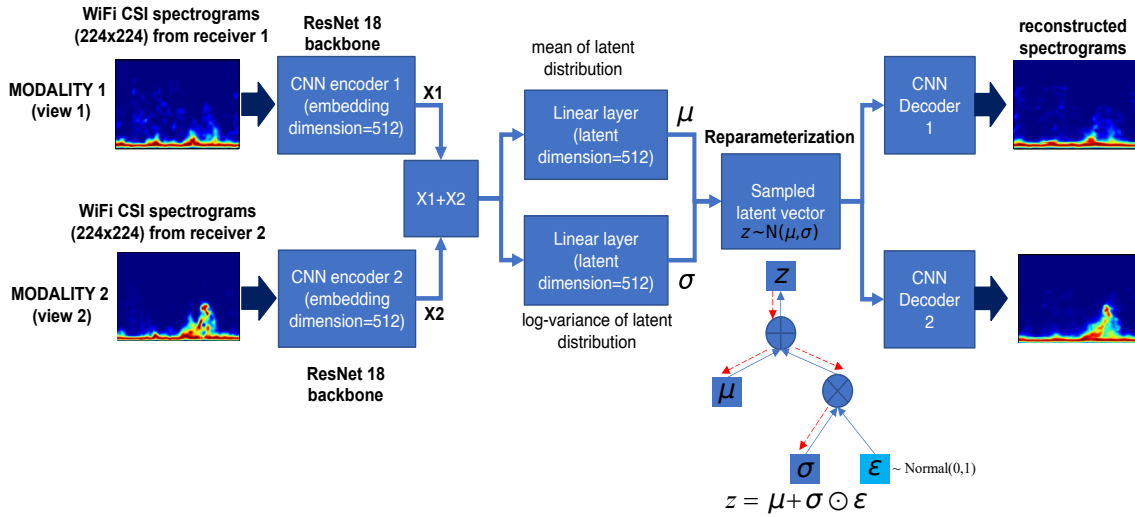
Figure S1. M-VAE for a full data case: Single encoder network takes all modalities.

S2 Signal processing pipelines for passive WiFi radar

A typical scenario for opportunistic passive WiFi Radar is depicted in Figure S2. This is an extremely challenging problem since the WiFi waveform was not specifically designed to lend itself to Radio-Frequency (RF) imaging. In addition, commercial WiFi chipsets have noisy RF chains and tend to suffer from phase drifts. The WiFi backscatter does contain information about the dynamic changes in the radio channel which is incapsulated in the Channel State Information (CSI). Dedicated tools need to be used to extract the CSI from WiFi network interface cards such as Atheros[21] or Intel 5300 (IWL5300)[22]. The raw CSI data is obtained as a 3-dimensional (3D) matrix per transmitted packet, with $n_t \times n_r \times N_{sc}$ complex values, where n_t is the number of transmit antennas, n_r is the number of receive antennas and N_{sc} is the number of subcarriers. Since the raw CSI data is very noisy in nature, the Discrete Wavelet Transform (DWT) technique can be used to filter out in-band noise and preserve the high frequency components, thus avoiding the distortion of the signal[23]. Afterwards, median filtering can be used to remove any undesired transients in the CSI measurements which are not due to human motion. The Intel 5300 chipset has a 3×3 antenna configuration and only 30 subcarriers are reported by this chipset. Thus the number of complex values per packet is equal to $3 \times 3 \times 30 = 270$. Considering a packet rate as high as 1.6 kHz, this results in a significant amount of data that needs to be processed. Therefore, we also apply Principal Component Analysis (PCA) to reduce the computational complexity of such high dimensional data. PCA identifies the time-varying correlations between the CSI streams which are optimally combined to extract only a few components that represent the variations caused by human activities. Finally, we convert the resultant data into spectrograms (time-frequency domain) using Short Time Fourier Transform (STFT), which are similar to those generated by Doppler radars. The CSI is highly sensitive to the surrounding environment and signal reflections from the human body result in different frequencies when performing different activities. The Doppler spectrogram generated from STFT helps to identify the change of frequencies over time. The generated spectrograms can be directly fed to CNNs to automatically identify a set of features, which can ultimately be used in downstream tasks. The CSI system consisted of two receivers. For more details on the experimental setup of the data collection, the interested reader is kindly referred to[20]. Each receiver can be seen as another view of the human activity being performed in the environment.

S3 Latent representation of WiFi spectrogram data

Our base model is shown in Fig. S1. The trained latent space for different latent dimensions are shown in Fig. S3. The trained latent space in Fig. S3 shows distinct clusters using UMAP (Uniform Manifold Approximation and Projection) visualization. The model was trained in a self-supervised fashion, the six clusters representing the six different human activities can be seen.

S4 Sensor fusion under noisy conditions (WiFi spectrogram data)

In this experiment, we analyze the sensor fusion performance when the data samples from the test dataset are affected by different amount of additive Gaussian noise. In this case, the measurement matrices are initialized as identity matrices with dimensions 50,176×50,176. No noise was injected to the input data from two modalities during training. The SFLR (Sensor Fusion in the Latent Representation space) algorithm (see Algorithm 1 in manuscript) is run for 1,000 iterations and the
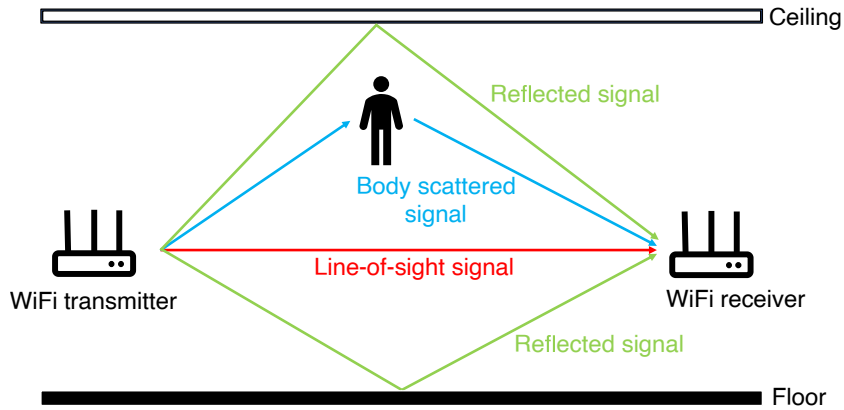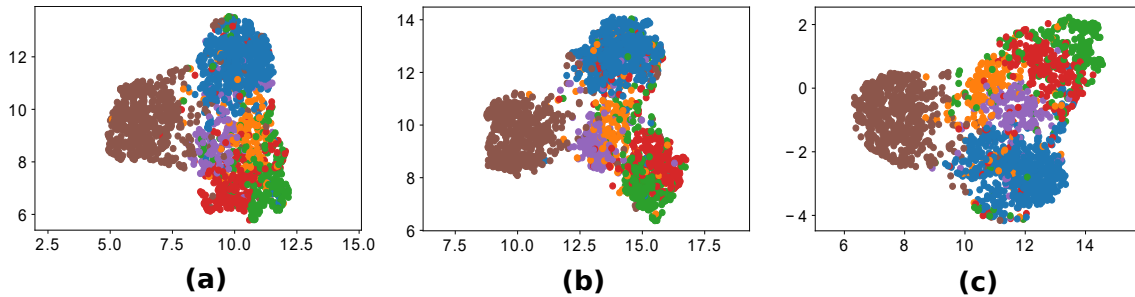
Figure S2. Opportunistic Passive WiFi Radar.

(a)   **(b)**   **(c)**

Figure S3. UMAP projection of trained latent space using our model on real WiFi CSI spectrogram data: (**a**) latent dimension=16, (**b**) latent dimension=64, and (**c**) latent dimension=128.

corresponding results are shown in Fig. S4, showing one sample in the test dataset. It can be observed that even under extreme noisy conditions, the noisy samples are denoised efficiently. These results are further validated in Table S1 where it can be seen that the fusion error remains essentially constant for different noise standard deviation values considering a batch of 50 images from the test dataset.

Table S1. Noisy measurements mean reconstruction error over a batch of 50 WiFi spectrogram data samples (full measurements considered).

Noise standard deviation	Modality 1	Modality 2
0.01	0.00347822	0.00573636
0.05	0.00348499	0.00574255
0.20	0.00349233	0.00569815
0.40	0.00354153	0.0058348
0.60	0.00357407	0.00594901
0.80	0.00367526	0.00608692

S5 Sensor fusion performance with missing pixels (WiFi spectrogram data)

In this experiment, we evaluate the fusion performance of the samples under different ratios of missing pixels. The results are shown in Fig. S5 when a true data sample is randomly chosen from the test dataset and a randomly generated (binary) mask is applied to it to simulate different missing pixel ratios. Each measurement corresponds to an observed pixel. Therefore, in this case the measurement matrices will be diagonal matrices with their diagonal entries corresponding to the mask elements (1's and 0's). From Fig. S5, it can be observed that the recovered samples are very close to the true ones, even when the missing pixel ratio for both modalities is as high as 0.8. In fact, as shown in Table S2, the reconstruction error remains essentially constant with increasing missing pixel ratio for both modalities.

Table S2. Missing pixel mean reconstruction error over a batch of 50 WiFi spectrogram data samples. Illustrations of spectrogram fusion under different missing pixel ratios are shown in Fig. S5.

Missing pixel ratio	Modality 1	Modality 2
0.1	0.00351316	0.00573596
0.2	0.00345235	0.00575154
0.4	0.00347854	0.00575382
0.6	0.003477	0.00571794
0.8	0.00349462	0.00579462

S6 Sensor fusion from asymmetric Compressed Sensing (WiFi spectrogram data)

In this experiment, we analyse the reconstruction of the WiFi spectrogram samples under two different scenarios, where we want to demonstrate the benefits of having multiple modalities. We are interested in recovery for one modality that is subsampled (loss of data) and noisy. This can be referred to as the weak data (or weak modality). Using the SFLR method, we leverage the second modality data, which has no loss of information or does not suffer from noise (strong modality), to improve the recovery for the modality of interest i.e., the weak modality. In the first case, only modality 1 (subsampled and noisy) is considered in the reconstruction process. In the second case, the good modality 2 is added in the iterative fusion process to improve the reconstruction quality of modality 1.

The results are tabulated in Table S3, where additive Gaussian noise with a standard deviation of 0.1 is considered. The results show the mean reconstruction errors (over 50 WiFi spectrogram samples) when modality 1 is subsampled to different extents. We see that reconstruction error has a general tendency to decrease with increasing number of measurements. It can be observed that the samples can be recovered with very low reconstruction error when the number of measurements is as low as 196 (0.39%). Furthermore, from Table S3, we observe that when only modality 1 is considered in the reconstruction process, the reconstruction errors are high when the number of measurements is equal to 1 (0.002%) and 10 (0.02%). However, by leveraging the good modality 2, the reconstruction quality is greatly improved for the same number of measurements, demonstrating the clear benefit of having multiple modalities. An illustration of the reconstruction quality is depicted in Fig. S6, where it can be observed that the unimodal reconstruction of modality 1 is far from the true sample. On the other hand, the reconstruction quality of modality 1 is improved by leveraging the good modality data.

Table S3. Mean reconstruction error over 50 WiFi spectrogram data samples. Noise standard deviation: 0.1

	No. of Measurements	Modality 1	Modality 2
	1 (0.002%)	0.0246185	-
	10 (0.02%)	0.01075371	-
Modality 1 with compressed sensing	196 (0.39%)	0.00258467	-
	784 (1.56%)	0.00195997	-
	1,568 (3.125%)	0.00184247	-
	1 (0.002%)	0.00892453	0.00380795
Modality 1 with compressed sensing	10 (0.02%)	0.00798366	0.00420512
	196 (0.39%)	0.0034269	0.00460956
	784 (1.56%)	0.0030373	0.00466936
Modality 2 with full information	1,568 (3.125%)	0.0028537	0.00469946

S7 Toy protein dataset: additional results

Sensor fusion from subsampled and noisy toy proteins

In this section, we present the sensor fusion results for toy protein reconstruction under subsampled and noisy observations, as an extension to Section "Sensor fusion from subsampled toy proteins" in the main document. Table S4 shows the mean reconstruction error of subsampled toy protein samples, with different levels of additive Gaussian noise. The proposed SFLR method recovers both modalities from as low as 4 subsampled and noisy observations.

Sensor fusion from asymmetric Compressed Sensing of toy proteins

We show the results of sensor fusion from asymmetric compressed sensing, regarding the third contribution of this paper. We claim that a strong modality can be used to aid the recovery of another modality that is lossy or less informative (weak

Table S4. Compressed sensing mean reconstruction error over a batch of 100 protein samples, with different noise levels.

Noise standard deviation	No. of Measurements	Modality 1 (10^{-3})	Modality 2 (10^{-3})
0.05	1 (3.125%)	52.301	55.382
	2 (6.250%)	11.678	9.200
	4 (12.500%)	0.834	0.715
	8 (25.000%)	0.387	0.450
0.1	1 (3.125%)	36.611	50.118
	2 (6.250%)	20.267	14.638
	4 (12.500%)	3.413	2.769
	8 (25.000%)	2.386	2.411
0.2	1 (3.125%)	43.466	48.271
	2 (6.250%)	17.864	19.435
	4 (12.500%)	1.528	1.435
	8 (25.000%)	5.005	5.063

modality). Table S5 shows the recovery results in two cases. In the first case, the subsampled modality 1 with additive Gaussian noise is observed and recovered. In the second case, the noise-free modality 2 with full observation is used to help the sensor fusion. We can see that modality 2 significantly helps with the recovery of modality 1, especially when the number of observations are relatively small.

Table S5. Mean reconstruction error over 100 samples with asymmetric compressed sensing. Noise standard deviation: 0.1.

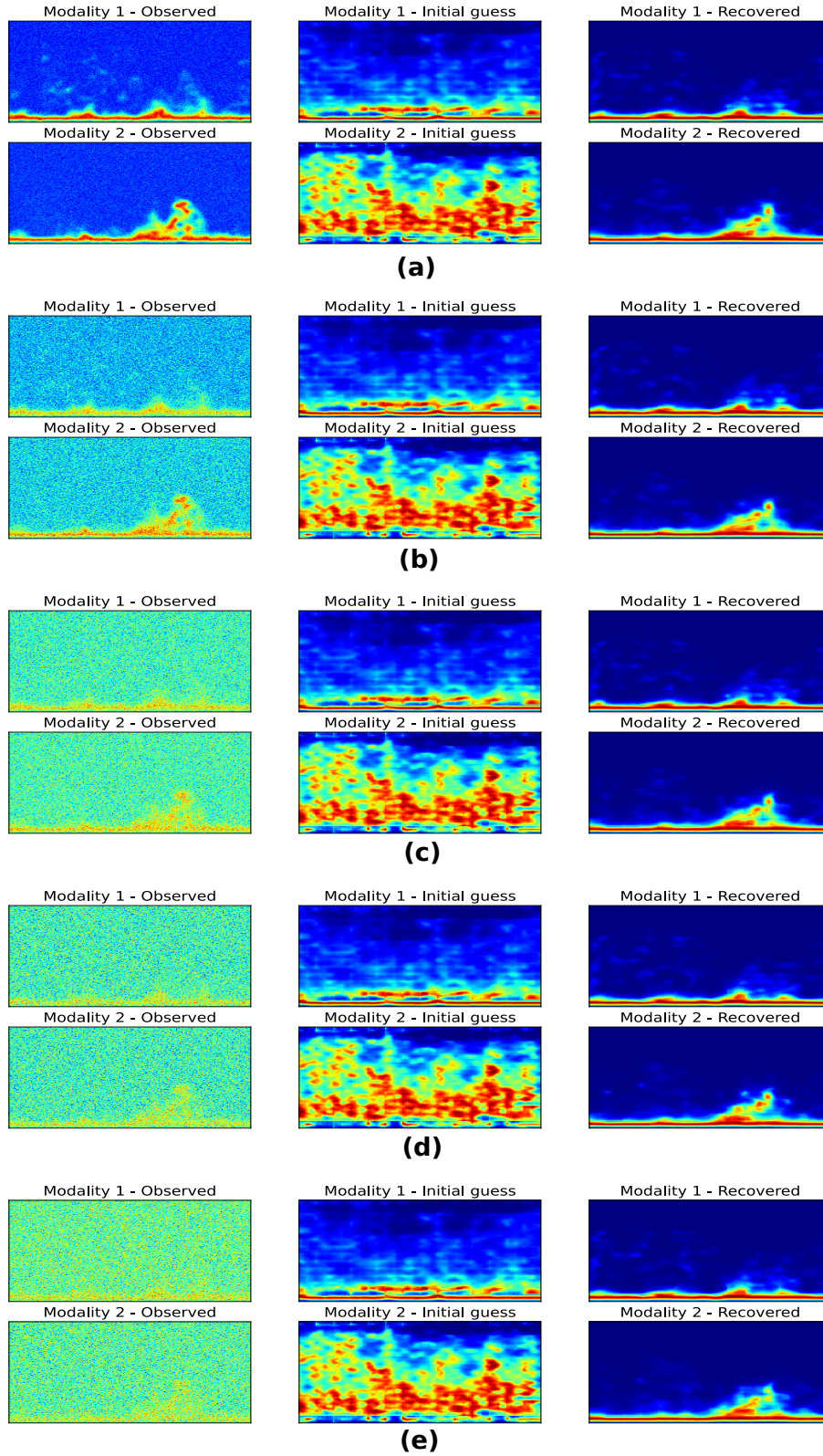|	No. of Measurements	Modality 1	Modality 2
Modality 1 with compressed sensing	1 (3.125%)	0.0542	-
	2 (6.250%)	0.0366	-
	4 (12.500%)	0.0205	-
	8 (25.000%)	0.0021	-
Modality 1 with compressed sensing	1 (3.125%)	0.0076	0.0073
	2 (6.250%)	0.0067	0.0062
Modality 2 with full information	4 (12.500%)	0.0023	0.0024
	8 (25.000%)	0.0033	0.0031

Figure S4. Impact of additive Gaussian noise on sensor fusion from the two modalities: (**a**) Std Dev=0.05, (**b**) Std Dev=0.2, (**c**) Std Dev=0.4, (**d**) Std Dev=0.6, and (**e**) Std Dev=0.8. Left column shows noisy spectrogram sample, middle column shows fusion with initial guess (no optimization) while right column shows fusion with $\hat{z}_{MAP}$.
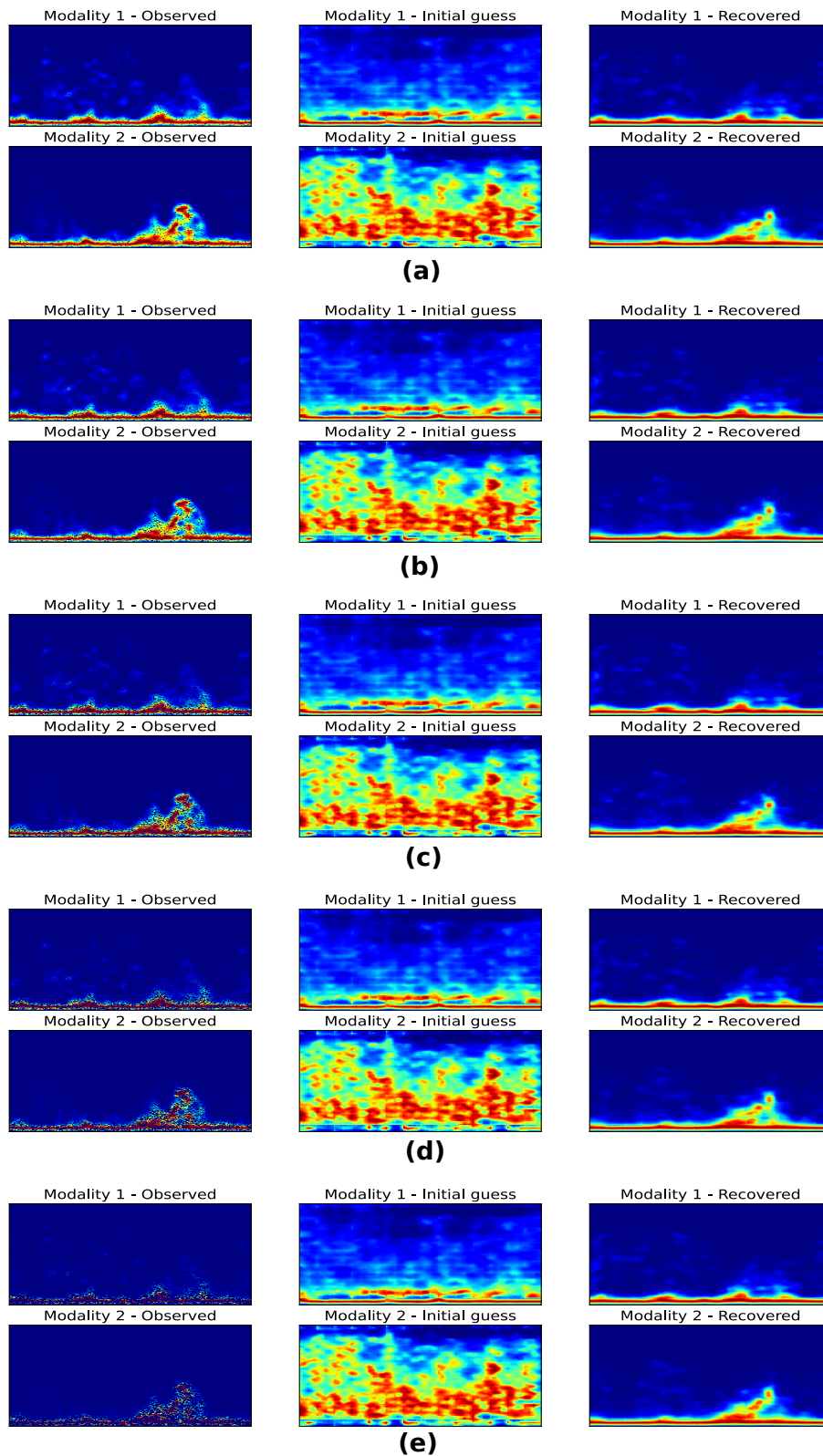
Figure S5. Impact of missing pixels on spectrogram recovery from the two modalities (no additive Gaussian noise): (**a**) missing pixel ratio=0.1, (**b**) missing pixel ratio=0.2, (**c**) missing pixel ratio=0.4, (**d**) missing pixel ratio=0.6, and (**e**) missing pixel ratio=0.8. Left column shows spectrogram sample with missing pixels, middle column shows reconstruction with initial guess (no optimization) while right column shows reconstruction with $\hat{z}_{MAP}$. Very good recovery performance is observed in all cases.
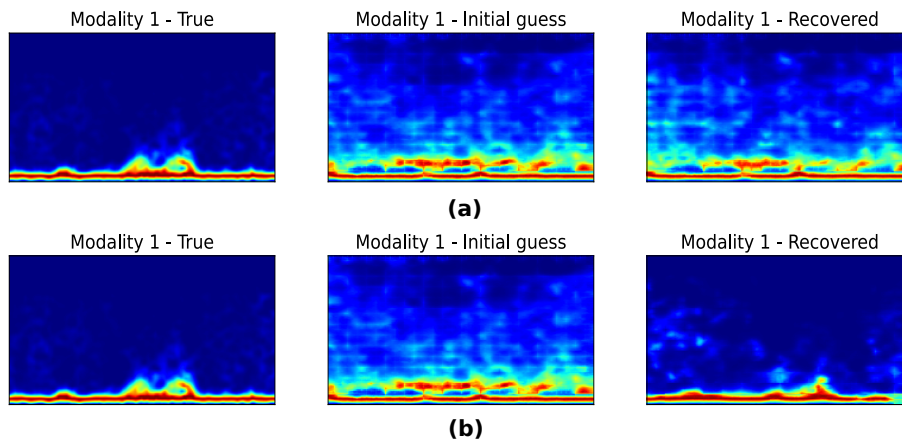
Figure S6. Reconstruction examples showing the benefit of multimodal system compared to a unimodal system. Modality 1 is subsampled data with 1 single measurement while modality 2 has full information (no noise and no loss of data). Additive Gaussian noise with a standard deviation of 0.1 is considered in this example: (**a**) reconstruction with modality 1 only, (**b**) reconstruction with both modalities 1 and 2. Left column shows true spectrogram sample, middle column shows reconstruction with initial guess (no optimization) while right column shows reconstruction with $\hat{z}_{MAP}$. Adding modality 2 during reconstruction stage helps in the sample recovery of modality 1.
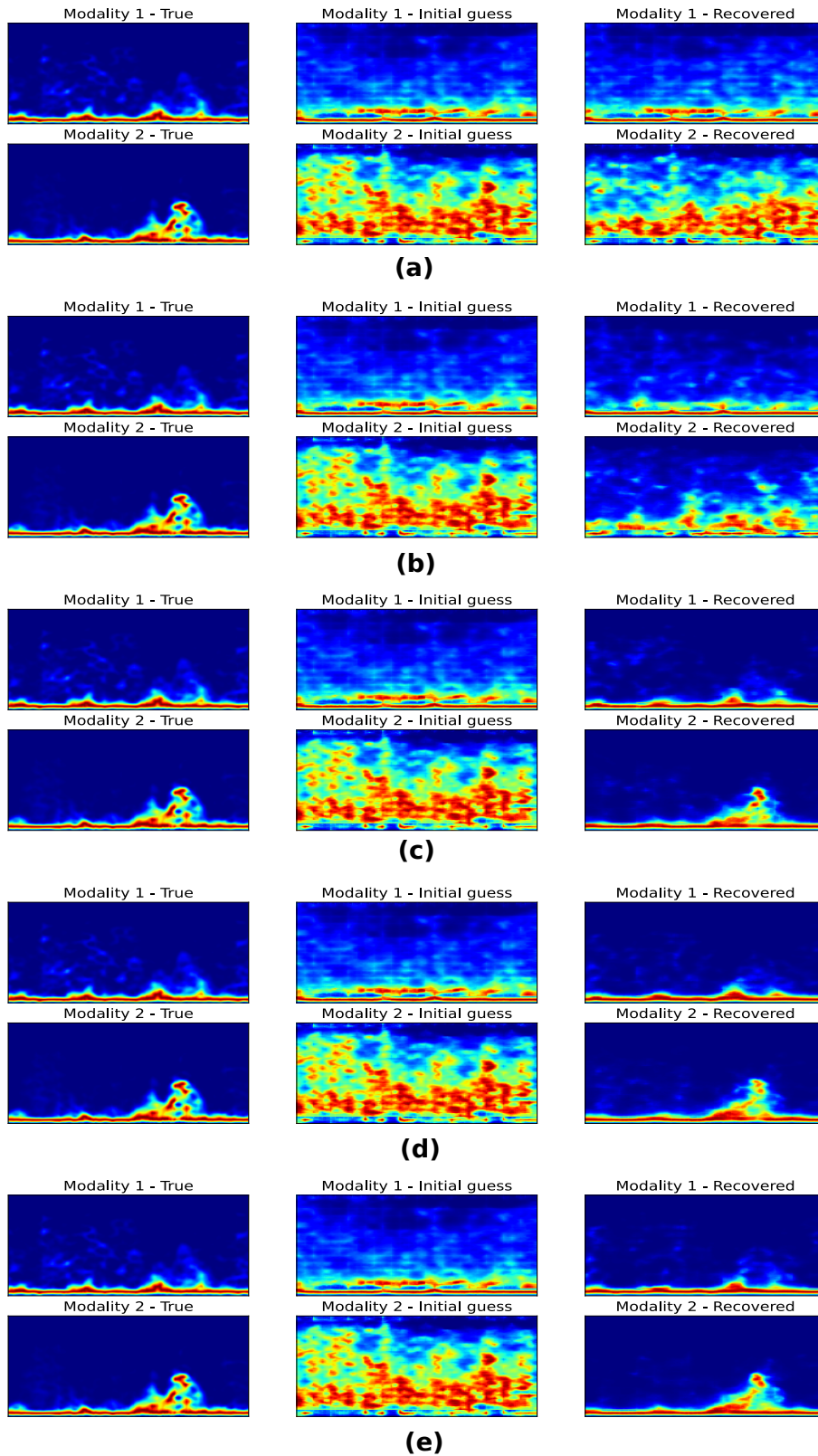
Figure S7. Compressed sensing performance on different number of measurements without additive Gaussian noise: (**a**) 1 measurement out of 50,176 (0.002%), (**b**) 10 measurements out of 50,176 (0.02%), (**c**) 196 measurements out of 50,176 (0.39%), (**d**) 784 measurements out of 50,176 (1.56%), and (**e**) 1,568 measurements out of 50,176 (3.125%). Left column shows true spectrogram sample, middle column shows reconstruction with initial guess (no optimization) while right column shows reconstruction with $\hat{z}_{MAP}$.
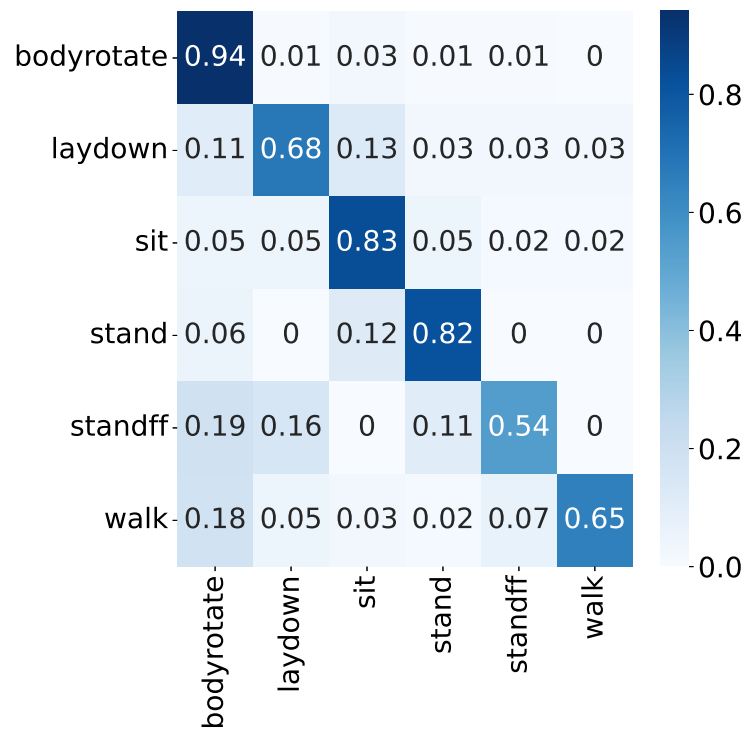
Figure S8. Confusion matrix of Human Activity Recognition (HAR) classification using our SFLR model (with compressed sensing). Ten labelled examples per class are considered (refer to Table 1 in main manuscript for classification results in terms of macro F_1 score).