Self-Supervised Multimodal Sensor Fusion for Human Actvity Recognition

Robert Piechocki[1], Kevin Chetty[2], N.D. Lane[3], Mohammud J. Bocus[1]

[1] School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths, University of Bristol

[2] Department of Security and Crime Science, University College London

[3] Department of Computer Science & Technology, University of Cambridge

OPERA

University of BRISTOL

UNIVERSITY OF CAMBRIDGE

UCL

1. Introduction

- Sensors such as Wi-Fi are particularly promising for in-home healthcare applications.
- Most passive Human Activity Recognition (HAR) deep learning-based systems are unimodal, i.e., they use information from only one type of sensor.
- Propose to use multiple synchronised sensors, or views, to improve the performance of a passive HAR system.
- 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.

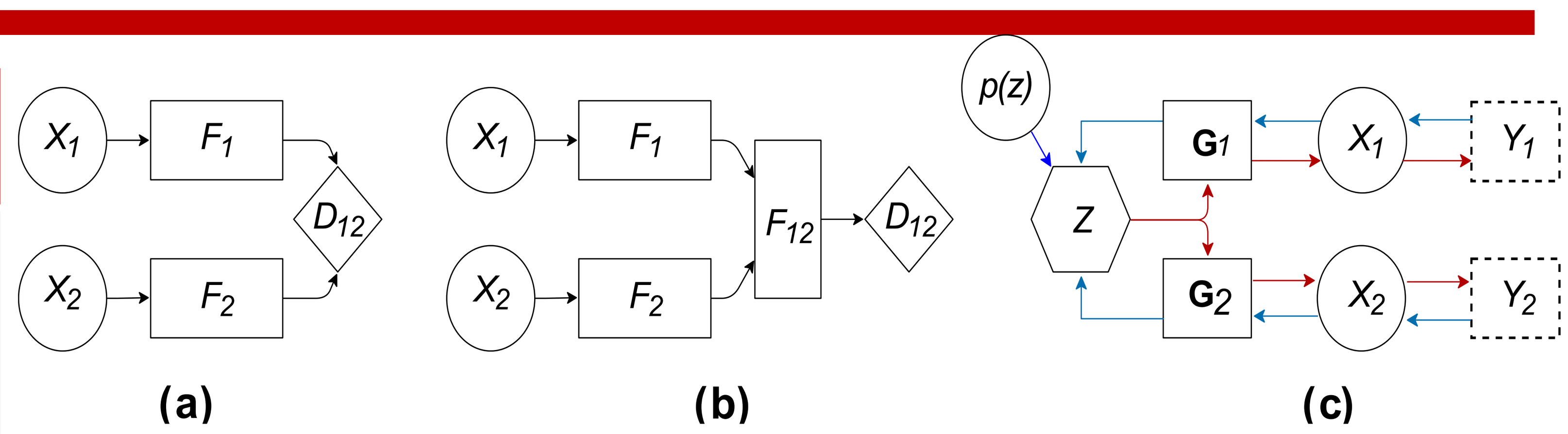


Figure 1. Multimodal Sensor Fusion: (a) Decision fusion, (b) Feature fusion, (c) Our technique: fusion in the latent representation with optional compressed sensing measurements; *F* features, *p(z)* prior model, *G* generators, *X* complete data, *Y* subsampled data.

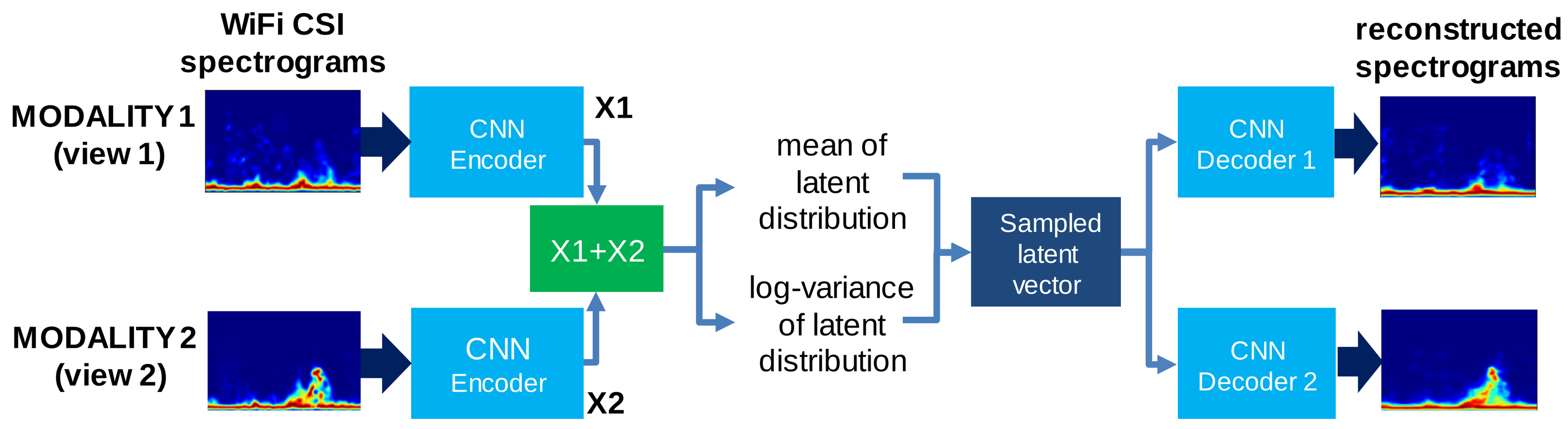


Figure 2. Multimodal Variational Autoencoder (M-VAE)

2. Results - Multimodal Sensor Fusion in the Latent Representation Space

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

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 $\{\chi_{1:M}\}$
Stage 1: Train M-VAE using $\mathscr{D}_{\mathscr{T}}$ Full data in first stage
Output: $p(z)$, Encoders $\{\phi_{1:M}\}$, Decoders $\{\psi_{1:M}\}$ Non-learnable in 2nd stage
Stage 2: Fusion
$y_{1:M}^{(i)} \sim \mathscr{D}_{\mathscr{P}}$ subsampled data
Sample the initial point $z^0 \sim p(z)$ Sample from prior
while not converged **do**
    $z \leftarrow z - \eta_0 \nabla_z(\|z\|^2) - \eta_1 \nabla_z(\|y_1^{(i)} - \chi_1(\psi_1(z))\|^2)$
    $z \leftarrow z - \eta_0 \nabla_z(\|z\|^2) - \eta_2 \nabla_z(\|y_2^{(i)} - \chi_2(\psi_2(z))\|^2)$
    $\vdots$
    $z \leftarrow z - \eta_0 \nabla_z(\|z\|^2) - \eta_M \nabla_z(\|y_M^{(i)} - \chi_M(\psi_M(z))\|^2)$
    One or several SGD steps are taken for each modality in turn.
end while
$\hat{z}_{MAP} \leftarrow z$
Downstream tasks: $\hat{x}_m = \psi_m(\hat{z}_{MAP})$, classification tasks K-NN$(\hat{z}_{MAP})$

MAP estimation procedure consists of backpropagating through the sampler and decoder using Stochastic Gradient Descent (SGD).

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

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

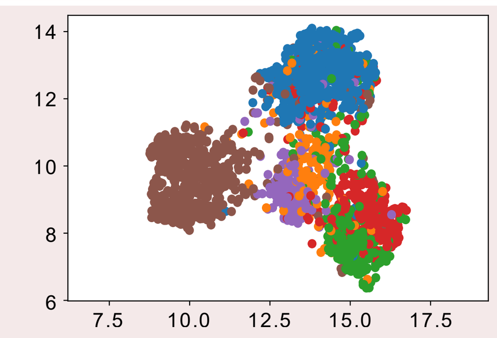


Figure 3. Trained Latent Space (6 clusters = 6 activities)

Table 1. Few-shot learning classification results (F1 macro) for HAR

	1 example per class	5 examples per class	10 examples per class
2-channel CNN	0.4273	0.5709	0.6185
1-channel CNN (Modality 1)	0.3491	0.4513	0.5045
1-channel CNN (Modality 2)	0.4466	0.6000	0.6057
Probability fusion (product rule)	0.4404	0.5847	0.6419
Dual-branch CNN	0.5082	0.5688	0.5759
SFLR (ours)	**0.6527**	**0.7182**	**0.7375**

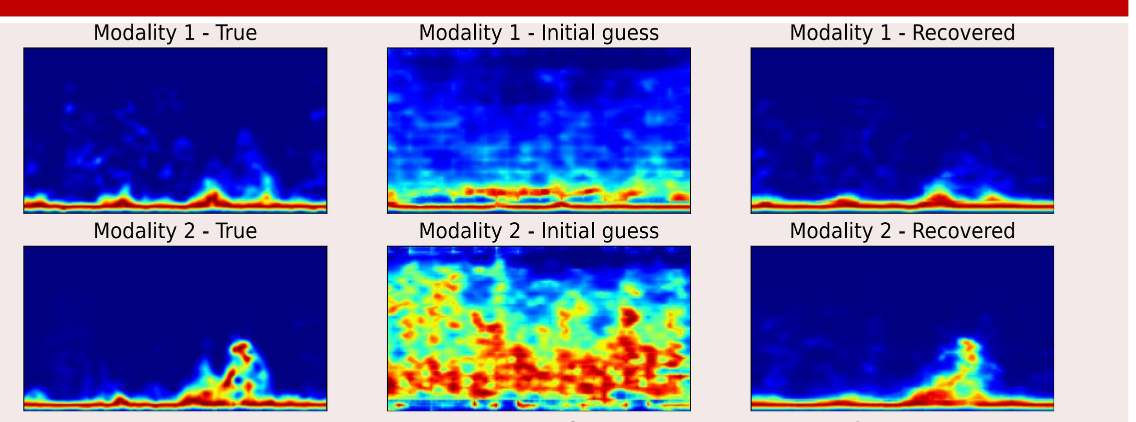


Figure 4. Recovery with compressed sensing measurements as low as 784 out of 50,176 (1.56%).

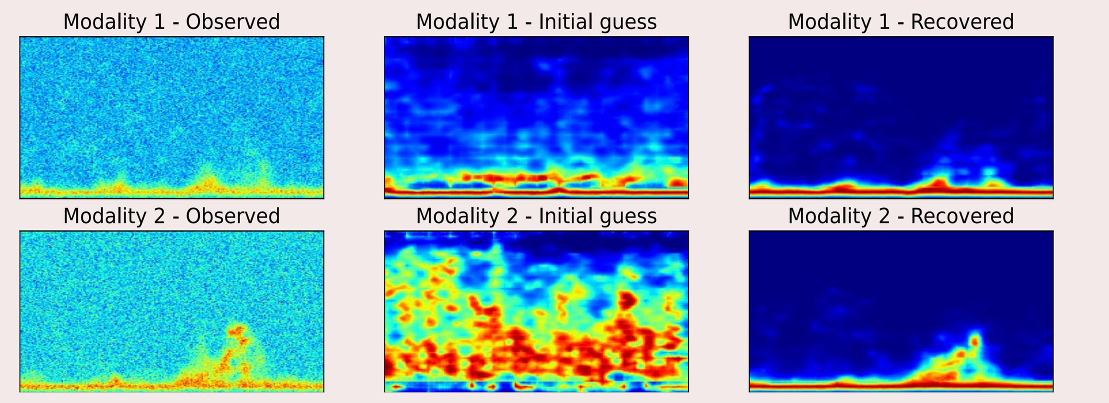
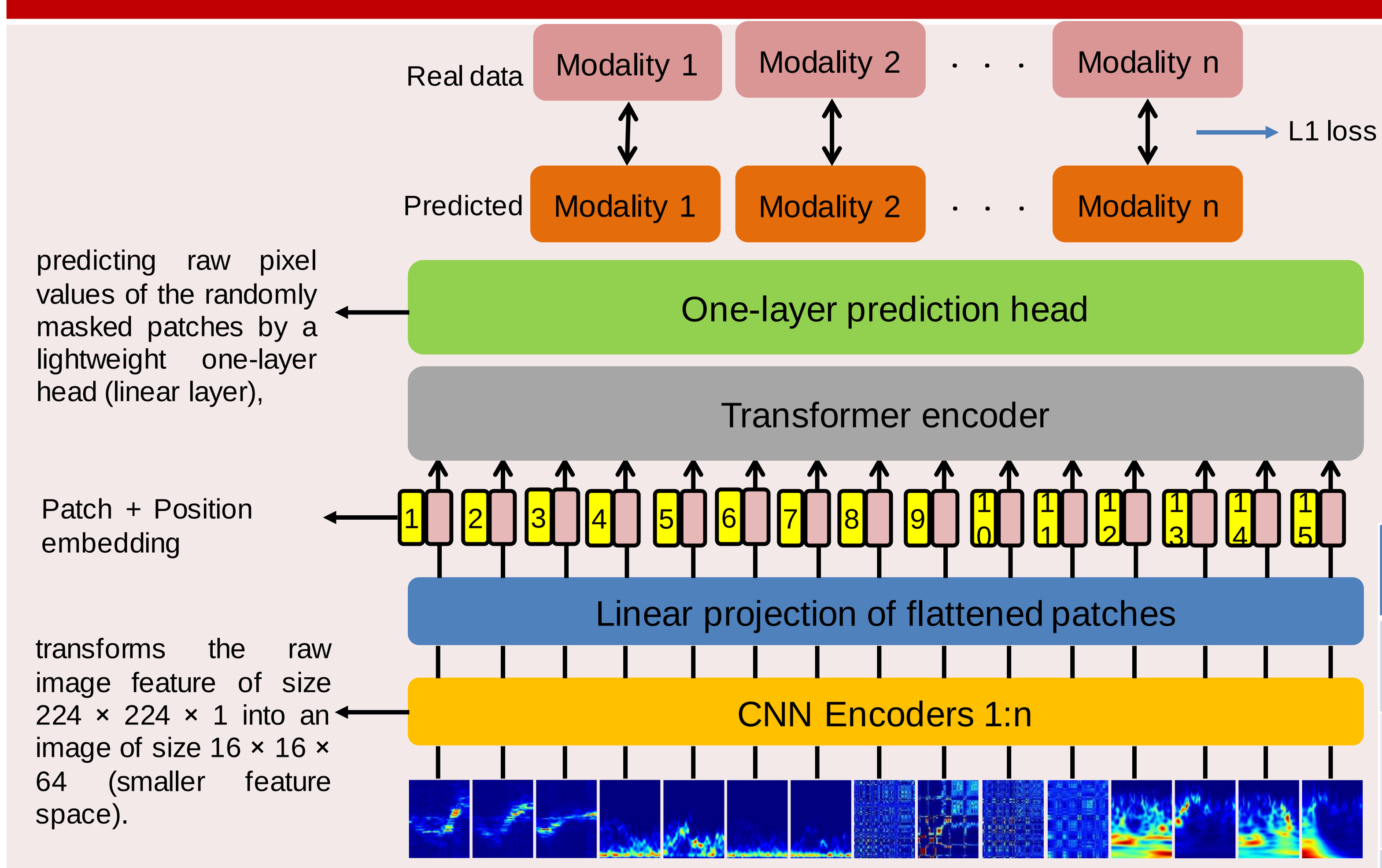


Figure 5 Recovery under additive Gaussian noise (std dev. = 0.4).

3. Results - Multimodal Fusion Transformer for Passive HAR



15 images features (modalities) extracted from **WiFi CSI** and **PWR** data (spectrograms, scalograms and Markov Transition Field). Each channel and receiver can be seen as another view of the human activity being performed. **60% of image features are masked during pretraining.**

1. mask 60% of image features and pre-train model to predict the masked image features.
2. pretrain model with both PWR and CSI spectrograms, using all different views and image-based features
3. fine-tune the pretrained model in a supervised way using few labelled examples.

	F1-score for different amount of labelled samples in training set				
	1 sample	5%	10%	20%	Full
Fusion Transformer with SSL	**56.3%**	**84.5%**	**89.7%**	**91.2%**	**95.9%**
ResNet-34 (CNN)	32.6%	56.9%	62.7%	73.8%	94.9%

4. Conclusions

- Multimodal sensor fusion brings improved performance for downstream tasks such as Human Activity Recognition (HAR), which serves a vital role in the E-Health paradigm.
- Pretraining models such as Multimodal Variational Autoencoder (MVAE) and multimodal Vision Transformer (ViT) in a self-supervised fashion outperforms non-pretrained models under few-shot learning (i.e. under the condition that few labelled samples are available).

References

[1] Bocus, M.J., Li, W., Vishwakarma, S. *et al. OPERAnet*, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Sci Data* **9**, 474 (2022).

[2] Piechocki, R. J., Wang, X. and Bocus, M. J., "Multimodal sensor fusion in the latent representation space", 2022, https://arxiv.org/abs/2208.02183.

[3] Koupai, A. K., Bocus, M. J., Santos-Rodriguez, R., Piechocki, R. J. and McConville, R., "Self-Supervised Multimodal Fusion Transformer for Passive Activity Recognition ", 2022, https://arxiv.org/abs/2209.03765.

UKRI Engineering and Physical Sciences Research Council

Digital Health and Care

sphere

IRC Next Steps Plus: OPERA – Opportunistic Passive Radar for Non-Cooperative Contextual Sensing (2019-2022, £1.36M, EP/R018677/1)