ICAM-reg: Interpretable Classification and Regression with Feature Attribution for Mapping Neurological Phenotypes in Individual Scans

Cher Bass[14], Mariana da Silva[1], Carole H. Sudre[1], Logan Z.J. Williams[1], Petru-Daniel Tudosiu[1], Fidel Alfaro-Almagro[2], Sean P. Fitzgibbon[1], Matthew Glasser[3], Stephen M. Smith[2], Emma C. Robinson[1]

1, King's College London; 2, University of Oxford; 3, Washington University in St Louis; 4, Panakeia Technologies

- Task - interpretable classification and regression with feature attribution (FA)
- Goal - prediction with a feature map for explanation
- Approach - VAE-GAN network with a shared attribute latent space and classification and regression layers to **disentangle class-relevant from class-irrelevant features**
- Dataset - UK Biobank for age prediction (3D MRI)
- Results - we show that ICAM can be used to analyse ageing by examining the latent space
- Extended arXiv paper (with code) - https://arxiv.org/abs/2103.02561

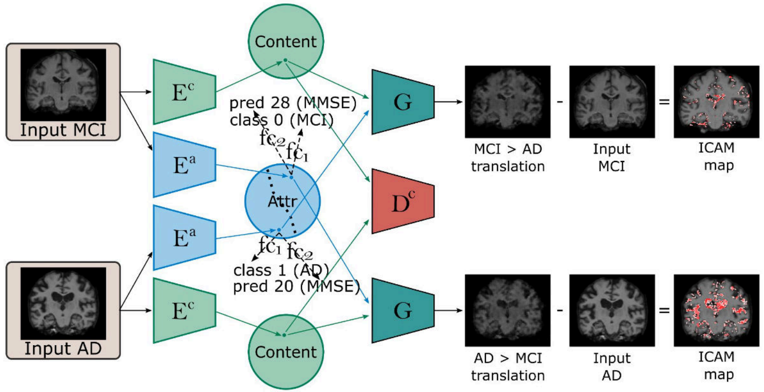
ICAM: training and inference

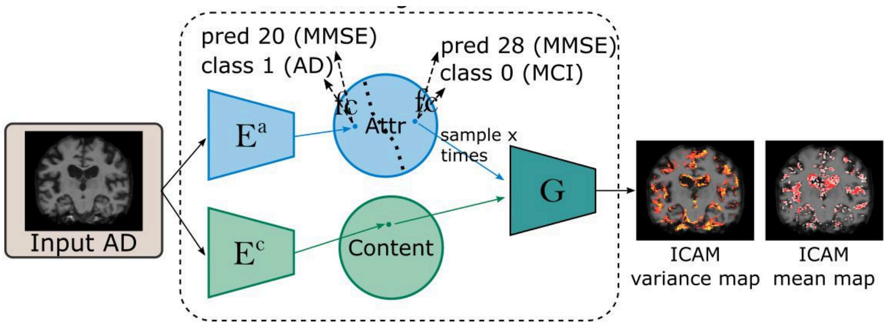
ICAM during training:

ICAM uses unpaired data during training, taking 2 input images of different classes, and uses Attribute and Content Encoders to disentangle class and non-class relevant features. The attribute space is swapped, to generate a translated image, and then the original image is subtracted to compute the FA map. The attribute space is used for prediction, using 2 linear layers, 1 for classification and 1 for regression.

ICAM during inference:

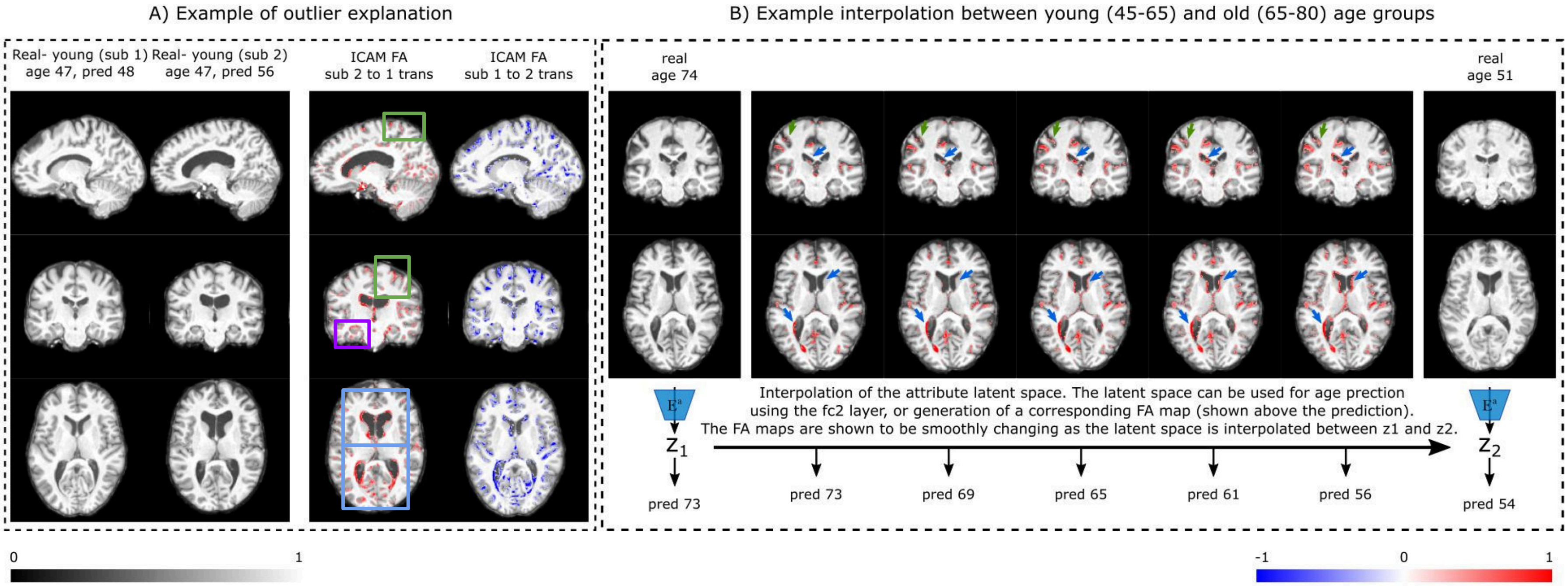
During inference, using ICAM, translation can be achieved using a single input image, in addition to translating between 2 images. An input image is encoded into a content space. The attribute space is then randomly sampled until a random vector of the required class is sampled, by checking its class using the classification layer. The newly sampled vector is passed to the generator along with the encoded content space to achieve translation.





Explaining Brain Ageing

A, translating between images of *same true age* but *different brain age*; B, interpolation between brain ages



Interpolation of the attribute latent space. The latent space can be used for age prection using the fc2 layer, or generation of a corresponding FA map (shown above the prediction). The FA maps are shown to be smoothly changing as the latent space is interpolated between z1 and z2.

Hippocampal atrophy

Ventricle atrophy

Cortical atrophy