In recent years, human pose estimation has greatly benefited from deep learning and huge gains in performance have been achieved on popular benchmarks [1, 3, 4]. The trend to maximise the accuracy on benchmarks, however, resulted in computationally expensive deep network architectures that require expensive hardware and pre-training on large datasets. In this work, we propose an efficient deep network architecture that can be efficiently trained on mid-range GPUs without the need of any pre-training and that is on par with much more complex models on the benchmarks [1, 3, 4].

Our proposed Fully Convolutional GoogLeNet (FCGN) network (see Figure 1) is based on the network architecture from [2]. We take the first 17 layers of [2] and add a deconvolution layer to make it fully convolutional. In addition, we introduce a skip layer and combine two FCGNs with shared weights to obtain a multi-resolution network. Belief maps for each joint are then obtained by a deconvolution layer with large kernel size in combination with a sigmoid function for normalisation and spatial drop out for regularisation.

We compare the performance of the proposed architecture against convolutional pose machines [5] on the well-known FLIC, LSP, and MPII benchmarks [1, 3, 4]. Our proposed network outperforms most previous approaches and achieves competitive performance to the more complex model of [5], while requiring only 3GB of memory and far less training time.