Implementation details

All the images are scaled to $[0,1]$. In all cases, the training is done using the ADAM optimizer with learning rate $10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The Keras default Xavier initialization is used. Shared layers are denoted in red and connected by a bidirectional arrow: $\leftrightarrow$. $\text{Conv} \times (k \times k)$ stands for a convolution layer with $n$ filters of size $k \times k$. $\text{ReLU}$ stands for a rectified linear unit, i.e. the function $\max(x, 0)$. $\text{MP} k \times k \text{strides}$ stands for max-pooling of size $k \times k$ with stride $s$. $\text{FC} k$ stands for a fully-connected layer of size $k$. The symbol $\oplus$ stands for the merge operation. For instance, if it appears after fully-connected layers of size 500 each, it denotes the resulting merged layer of size 1000. Outputs are processed by a SoftMax.

Unsupervised learning - MNIST

For the MNIST reconstruction experiments, we utilize a CNN-based version of the autoencoder and JAE presented in Figure 1 in the body. Mini-batch size is set to 256, with 10 epochs. The JAE losses are weighed equally.

Unsupervised learning - CIFAR-10

Mini-batch size is set to 128, with 10 epochs. The JAE losses are weighed equally. “$\text{Deconv} \times k \times k$” stands for a deconvolution layer with $n$ filters of size $k \times k$ with $2 \times 2$ upsampling.
Unsupervised learning - celebA

The images are rescaled to $64 \times 64 \times 3$. Mini-batch size is set to 64, with 30 epochs. The JAE losses are weighed equally. We omit the ReLU activations for brevity, and use “Conv $n, k, s$” and “Deconv $n, k, s$” to denote convolutions and deconvolutions with $n$ filters of size $k$ with strides $s$.

Transfer learning - MNIST $\leftrightarrow$ USPS

Mini-batch size is set to 64, with 10 epochs. The reconstruction losses are weighed 4 times higher than the classification losses.

Transfer learning - SVHN $\rightarrow$ MNIST

Mini-batch size is set to 64, with 10 epochs. The reconstruction losses are weighed 4 times lower than the classification losses. In this case, as opposed to the previous one, the classification task is challenging enough to avoid early overfitting.
Transfer learning - SVHN→MNIST+USPS

Mini-batch size is set to 64, with 20 epochs. The reconstruction losses are weighed 4 times lower than the classification losses.