Supplementary Materials

A Detailed derivation of the learning objective

We here provide the details for deriving equation (12), the lower bound of our learning objective $\mathcal{L}'$. The derivation is similar to that of the original DVIB literature. Remark that the objective is:

$$\text{maximize} \quad \mathcal{L}' = I(Z'; Y) - \beta \cdot I(Z; X) \tag{1}$$

Here, as in DVIB, we make the assumption that the joint distribution $p(\mathbf{x}, y, \mathbf{z})$ is factorized as

$$p(\mathbf{x}, y, \mathbf{z}) = p(\mathbf{x})p(y|\mathbf{x})p(\mathbf{z}|\mathbf{x}) \tag{2}$$

which means that the corresponding directed graph is $Z \leftarrow X \rightarrow Y$.

The lower bound for the first term $I(Z'; Y)$ is:

$$\begin{aligned} I(Z'; Y) &= \iint p(y, \mathbf{z}') \log \frac{p(y, \mathbf{z}')}{p(y)p(\mathbf{z})} dy d\mathbf{z}' \\ &= \iint p(y, \mathbf{z}') \log p(y|\mathbf{z}') dy d\mathbf{z} - H[Y] \\ &= \int p(\mathbf{z}') \Big[\int p(y|\mathbf{z}') \log p(y|\mathbf{z}') dy \Big] d\mathbf{z}' - H[Y] \\ &\geq \int p(\mathbf{z}') \Big[\int p(y|\mathbf{z}') \log q(y|\mathbf{z}') dy \Big] d\mathbf{z}' - H[Y] \\ &= \iint p(y, \mathbf{z}') \log q(y|\mathbf{z}') dy d\mathbf{z}' - H[Y] \\ &= \iiint p(\mathbf{x}, y) p(\mathbf{z}'|\mathbf{x}, y) \log q(y|\mathbf{z}') dy d\mathbf{z}' d\mathbf{x} - H[Y] \\ &= \iint p(\mathbf{x}, y) \Big[\int p(\mathbf{z}'|\mathbf{x}) \log q(y|\mathbf{z}') d\mathbf{z}' \Big] d\mathbf{x} dy - H[Y] \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{z}'|\mathbf{x}_i)} \Big[\log q(y_i|\mathbf{z}') \Big] - H[Y]. \end{aligned} \tag{3}$$

The inequality is due to $\text{KL}[p(y|\mathbf{z}')||q(y|\mathbf{z}')] \geq 0$.

The upper bound for the second term $I(X; Y)$ is:

$$\begin{aligned} I(X; Z) &= \iint p(\mathbf{x}, \mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z})p(\mathbf{x})} d\mathbf{x} d\mathbf{z} \\ &= \iint p(\mathbf{z}, \mathbf{x}) \log p(\mathbf{z}|\mathbf{x}) d\mathbf{x} d\mathbf{z} - \int p(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z} \\ &\leq \iint p(\mathbf{z}, \mathbf{x}) \log p(\mathbf{z}|\mathbf{x}) d\mathbf{x} d\mathbf{z} - \int p(\mathbf{z}) \log q(\mathbf{z}) d\, \mathbf{z} \\ &= \iint p(\mathbf{x}) p(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} d\mathbf{x} d\mathbf{z} \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \Big[\text{KL}[p(\mathbf{z}|\mathbf{x}_i)||q(\mathbf{z})\Big]. \end{aligned} \tag{4}$$

The inequality is due to $\text{KL}[p(\mathbf{z})||q(\mathbf{z})] \geq 0$. Putting all together yields

$$\mathcal{L}' \geq \frac{1}{n} \sum_{i=1}^{n} \Big[\mathbb{E}_{p(\mathbf{z}'|\mathbf{x}_i)} \Big[\log q(y_i|\mathbf{z}') \Big] - \beta \cdot \text{KL}[p(\mathbf{z}|\mathbf{x}_i)||q(\mathbf{z})] \Big] - H[Y] \tag{5}$$

and since $H[Y]$ is a constant, we are safe to drop it from the objective for optimization.

B Details of the network architecture

The convolutional neural networks (CNN) employed in the experiments contain 20 layers that are grouped into 5 stages, as summarized in Figure 1.

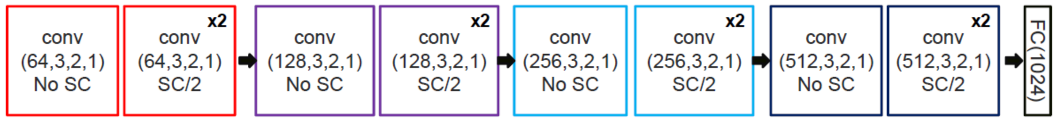


Figure 1: The detailed architecture of the CNNs employed in the experiments.

in which:

- *Conv* means the convolutional layer, the figures (n, s, p, r) mean that there are n filters with $s \times s$ size in the layer, the stride is p, and the padding s r;
- *No SC* means that there is no short cut connection and *SC/2* means that there is a short cut connection between every two layers;
- *FC* indicates the fully connected layer. There are 1024 units in the FC layer.
- Parametric Rectified Linear Unit (pReLU) is adopted as the non-linearity in the network. The activation function of pReLU is:

$$pReLU(x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{if } x \leq 0 \end{cases} \tag{6}$$

  where a is a learnable parameter. The initial value of a is set to be $a = 0.25$.

The weights of the CNN would be jointly trained with that in the subsequent network through BP.

C Details of the modified Carlini-Wanger algorithm

Here we provide the details of the modified Carlini-Wanger attack for constructing adversarial biometrics in our experiment. Remark that to find the adversarial biometric $\tilde{\mathbf{x}}_1$ we need to optimize the following objective:

$$J'_{\text{adv}}(\tilde{\mathbf{x}}_1) = \|\mathbf{x}_1 - \tilde{\mathbf{x}}_1\|_2 + \lambda \cdot \cos(f(\tilde{\mathbf{x}}_1), f(\mathbf{x}_2)) \tag{7}$$

which is subject to the constraint $x_k \in [0, 1]$. To remove this constraint we reparameterize each $\mathbf{x}$ as

$$\mathbf{x} = h(\mathbf{v}) = \frac{1}{2} \tanh \mathbf{v} + 1 \tag{8}$$

with which we can rewrite (7) as:

$$J'_{\text{adv}}(\tilde{\mathbf{z}}_1)) = \|h(\mathbf{z}_1) - h(\tilde{\mathbf{z}}_1)\|_2 + \lambda \cdot \cos(f(h(\tilde{\mathbf{z}}_1)), f(h(\mathbf{z}_2))) \tag{9}$$

and we can now learn $\tilde{\mathbf{z}_1}$ by gradient descent.

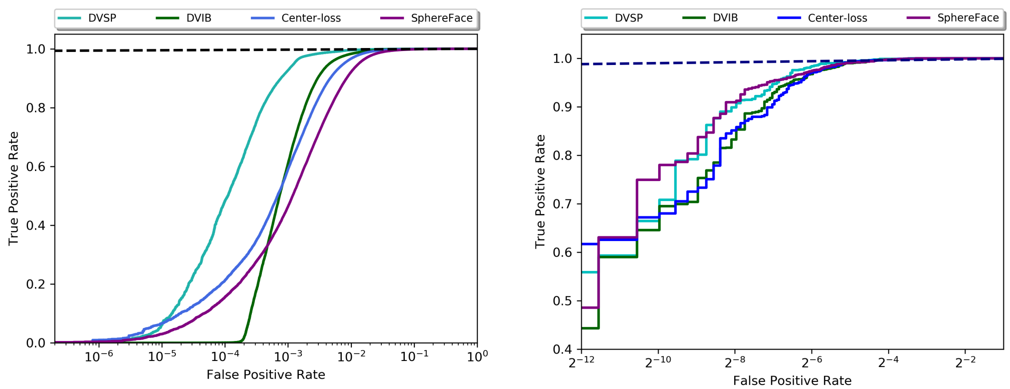
For the selection of λ, we find the optimal value of λ by an iterative procedure. Starting from $\lambda = 1$, we will update the value of λ as follows:

$$\lambda = \begin{cases} 10\lambda & \text{if the solved } \tilde{\mathbf{z}}_1 \text{ in (9) satisfies: } \cos(f(h(\tilde{\mathbf{z}}_1)), f(h(\mathbf{z}_2))) \leq T \\ \lambda/2 & \text{if the solved } \tilde{\mathbf{z}}_1 \text{ in (9) satisfies: } \cos(f(h(\tilde{\mathbf{z}}_1)), f(h(\mathbf{z}_2))) \geq T \end{cases} \tag{10}$$

This procedure is repeated until convergence. T is selected as the threshold at which the equal error rate (EER) is attained. All optimization is done by Adam with its default settings.

D More experimental results

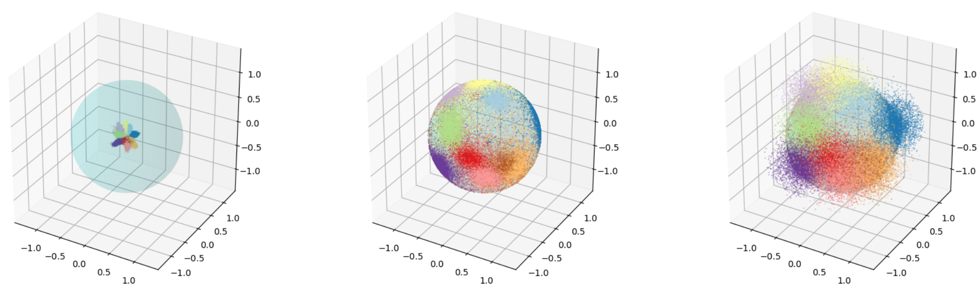
D.1 ROC curves



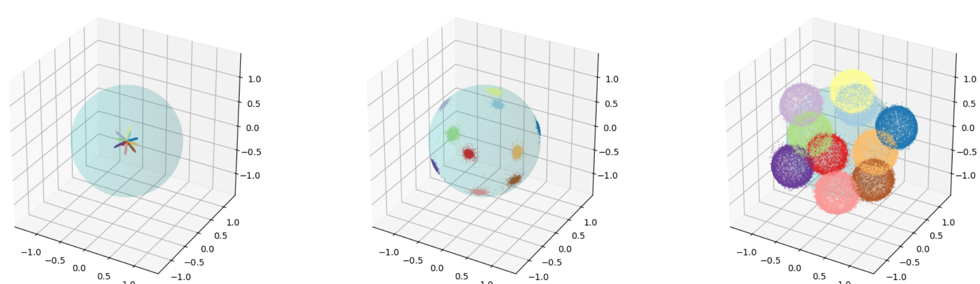
(a) MNIST (b) CASIA/LFW

Figure 2: The ROC curves on MNIST and CASIA/LFW datasets.

D.2 Visualization of learned features



(a) epoch 1: original features (b) epoch 1: projected features (c) epoch 1: sampled features



(d) epoch 50: original features (e) epoch 50: projected features (f) epoch 50: sampled features

Figure 3: Visualizing the feature learning process in DVSP. First columns: the original features $\mathbf{z}$ in DVIB. Second columns: the features $\mathbf{z}'$ after sphere projection. Note that these features are the final features used in recognition. Third columns: the features randomly sampled from $\mathbf{z}' \sim p(\mathbf{z}'|\mathbf{x})$. It can be seen that after 50 epochs the (angular) margin between the features $\mathbf{z}'$ are visually very large.