A. Proofs of Theorems

Proof of Theorem 3.2. We calculate the gap between $\mathrm{I}_{\mathrm{vCLUB}}$ and $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$:

$$\begin{aligned}
\tilde{\Delta} :=& \mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x};\boldsymbol{y}) - \mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \\
=& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{y})\right] \\
=& \left[\mathbb{E}_{p(\boldsymbol{y})}[\log p(\boldsymbol{y})] - \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})]\right] - \left[\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})]\right] \\
=& \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}[\log \frac{p(\boldsymbol{y})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})}] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log \frac{p(\boldsymbol{y}|\boldsymbol{x})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})}] \\
=& \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}[\log \frac{p(\boldsymbol{x})p(\boldsymbol{y})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}] \\
=& \mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) - \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})).
\end{aligned}$$

Therefore, $\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x};\boldsymbol{y})$ is an upper bound of $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$ if and only if $\mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) \geq \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$.

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, $p(\boldsymbol{x})p(\boldsymbol{y}) = p(\boldsymbol{x},\boldsymbol{y})$. Then, $\mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) = \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$ and $\tilde{\Delta} = 0$. Therefore, $\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{I}(\boldsymbol{x};\boldsymbol{y})$, the equality holds. □

Proof of Corollary 3.3. If $\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) \leq \epsilon$, then

$$\mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log \frac{p(\boldsymbol{x},\boldsymbol{y})}{q_\theta(\boldsymbol{x},\boldsymbol{y})}] = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log \frac{p(\boldsymbol{y}|\boldsymbol{x})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})}] = \mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) \leq \epsilon.$$

By the condition $\mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}) > \mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$, we have $\mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) < \varepsilon$.

Note that the KL-divergence is always non-negative. From the proof of Theorem 3.2,

$$\begin{aligned}
|\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x};\boldsymbol{y}) - \mathrm{I}(\boldsymbol{x};\boldsymbol{y})| =& |\mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) - \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))| \\
<& \max\left\{\mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})), \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))\right\} \leq \varepsilon,
\end{aligned}$$

which supports the claim. □

B. Network Expressiveness in Variational Inference

In Section 3.2, when analyze the properties of the vCLUB estimator, we claim a reasonable assumption that with high expressiveness of the neural network $q_\theta(\boldsymbol{y}|\boldsymbol{x})$, we can achieve $\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) < \varepsilon$. Here we provide a analysis under the scenario that the conditional distribution is a Gaussian distribution, $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}^*(\boldsymbol{x}), \mathbf{I})$. The variational approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ is parameterized by $q_\theta(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}), \mathbf{I})$.

Then training samples pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ can be treated as $(\boldsymbol{x}_i, \boldsymbol{\mu}^*(\boldsymbol{x}_i) + \boldsymbol{\xi}_i)$, where $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then

$$\begin{aligned}
\log p(\boldsymbol{y}|\boldsymbol{x}) =& \log \prod_{d=1}^{D}[\frac{1}{\sqrt{2\pi}} e^{(y^{(d)} - \mu*^{(d)}(\boldsymbol{x}))^2/2}] = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}^*(\boldsymbol{x})\|^2, \\
\log q_\theta(\boldsymbol{y}|\boldsymbol{x}) =& \log \prod_{d=1}^{D}[\frac{1}{\sqrt{2\pi}} e^{(y^{(d)} - \mu_\theta^{(d)}(\boldsymbol{x}))^2/2}] = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\mu}_\theta(\boldsymbol{x})\|^2.
\end{aligned}$$

The log-ratio between $p(\boldsymbol{y}_i|\boldsymbol{x}_i)$ and $q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is

$$\log \frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i)}{q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)} = \log p(\boldsymbol{y}_i|\boldsymbol{x}_i) - \log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i) = [\boldsymbol{\mu}^*(\boldsymbol{x}_i) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_i)]^T[\boldsymbol{y}_i - \boldsymbol{\mu}_\theta(\boldsymbol{x}_i) + \boldsymbol{\xi}_i].$$

We further assume $\|\boldsymbol{\mu}^*(\boldsymbol{x}) - \boldsymbol{\mu}_\theta(\boldsymbol{x})\| < A$ is bounded. Then $|\log p(\boldsymbol{y}_i|\boldsymbol{x}_i) - \log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)| < A\|\boldsymbol{y}_i - \boldsymbol{\mu}_\theta(\boldsymbol{x}_i) + \boldsymbol{\xi}_i\|$.

Denote a loss function $l(\boldsymbol{\mu}_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i) = \|\boldsymbol{y}_i - \boldsymbol{\mu}_\theta(\boldsymbol{x}_i) + \boldsymbol{\xi}_i\|$. With all reasonable assumptions in Hu et al. (2019), and applying the Theorem 5.1 in Hu et al. (2019), we know that when the number of samples $n \to \infty$, the expected error $\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[l(\boldsymbol{\mu}_\theta(\boldsymbol{x}), \boldsymbol{y})] \to \infty$ with probability $1 - \delta$.

$$\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log q_\theta(\boldsymbol{y}|\boldsymbol{x})] < A \cdot \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[l(\boldsymbol{\mu}_\theta(\boldsymbol{x}), \boldsymbol{y})].$$

Therefore, when given a small number $\varepsilon > 0$, having the sample size n large enough, we can guarantee that $\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x}))$ is smaller than ε.

C. Properties of Variational Upper Bounds

In the Section 2, we introduce two variational MI upper bounds with neural network approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ to $p(\boldsymbol{y}|\boldsymbol{x})$:

$$\mathrm{I}_{\mathrm{vVUB}}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{q_\theta(\boldsymbol{y}|\boldsymbol{x})}{r(\boldsymbol{y})}\right],$$
$$\mathrm{I}_{\mathrm{vL1Out}}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)}{\frac{1}{N-1}\sum_{j\neq i} q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_j)}\right]\right].$$

With the neural approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$, $\mathrm{I}_{\mathrm{vVUB}}$ and $\mathrm{I}_{\mathrm{vL1Out}}$ no longer guarantee to be the MI upper bounds. However, both of the two estimators have good properties with a good approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$.

Theorem C.1. *If $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ satisfies $KL(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) \leq KL(p(\boldsymbol{y})\|r(\boldsymbol{y}))$, then $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \leq \mathrm{I}_{vVUB}(\boldsymbol{x};\boldsymbol{y})$.*

Proof of Theorem C.1. With the conditional $\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) \leq \mathrm{KL}(p(\boldsymbol{y})\|r(\boldsymbol{y}))$,

$$\begin{aligned}\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})}\right] = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log\left(\frac{p(\boldsymbol{y}|\boldsymbol{x})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})}\cdot\frac{q_\theta(\boldsymbol{y}|\boldsymbol{x})}{r(\boldsymbol{y})}\cdot\frac{r(\boldsymbol{y})}{p(\boldsymbol{y})}\right)\right]\\ =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{q_\theta(\boldsymbol{y}|\boldsymbol{x})}{r(\boldsymbol{y})}\right] + \mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) - \mathrm{KL}(p(\boldsymbol{y})\|r(\boldsymbol{y})) \leq \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{q_\theta(\boldsymbol{y}|\boldsymbol{x})}{r(\boldsymbol{y})}\right].\end{aligned}$$

□

Theorem C.2. *Given $N-1$ samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N-1}$ from the marginal $p(\boldsymbol{x})$, If*

$$KL(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) \leq \mathbb{E}_{\boldsymbol{x}_i\sim p(\boldsymbol{x})}\left[KL\left(p(\boldsymbol{y})\|\frac{1}{N-1}\sum_{i=1}^{N-1} q_\theta(\boldsymbol{y}|\boldsymbol{x}_i)\right)\right],$$

then $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \leq \mathrm{I}_{vL1Out}(\boldsymbol{x};\boldsymbol{y})$.

Proof. Assume we have N sample pairs $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$ drawn from $p(\boldsymbol{x},\boldsymbol{y})$, then

$$\begin{aligned}\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) =& \mathbb{E}_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\sim p(\boldsymbol{x},\boldsymbol{y})}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i)}{p(\boldsymbol{y}_i)}\right]\right]\\ =& \mathbb{E}_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\sim p(\boldsymbol{x},\boldsymbol{y})}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\log\left(\frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i)}{q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)}\cdot\frac{q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)}{\frac{1}{N-1}\sum_{j\neq i} q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_j)}\cdot\frac{\frac{1}{N-1}\sum_{j\neq i} q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_j)}{p(\boldsymbol{y}_i)}\right)\right]\right]\\ =& \mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) + \mathrm{I}_{\mathrm{vVUB}}(\boldsymbol{x};\boldsymbol{y}) - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\left(p(\boldsymbol{y})\|\frac{1}{N-1}\sum_{j\neq i} q_\theta(\boldsymbol{y}|\boldsymbol{x}_j)\right)\right].\end{aligned}$$

Apply the condition in Theorem C.2 to each $N-1$ combination of $\{\boldsymbol{x}_j\}_{j\neq i}$, we conclude $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \leq \mathrm{I}_{\mathrm{vL1Out}}(\boldsymbol{x};\boldsymbol{y})$. □

Theorem C.1 and Theorem C.2 indicate that if the approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ is good enough, the estimators $\mathrm{I}_{\mathrm{vVUB}}$ and $\mathrm{I}_{\mathrm{vL1Out}}$ can remain as MI upper bounds. Based on the analysis in Section B, when implemented with neural networks, the approximation can be far more accurate to preserve the variational estimators as MI upper bounds.

D. Implementation Details

vCLUB with Gaussian Approximation When $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ is parameterized by $\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}(\boldsymbol{x}),\boldsymbol{\sigma}^2(\boldsymbol{x})\cdot\mathbf{I})$, then given samples $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$, we denote $\boldsymbol{\mu}_i = \boldsymbol{\mu}(\boldsymbol{x}_i)$ and $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}(\boldsymbol{x}_i)$. Moreover, $\boldsymbol{\mu}_i = [\mu_i^{(1)},\mu_i^{(2)},\ldots,\mu_i^{(D)}]^{\mathrm{T}}$, $\boldsymbol{\sigma}_i =$

$[\sigma_i^{(1)}, \sigma_i^{(2)}, \ldots, \sigma_i^{(D)}]^{\mathrm{T}}$, are D-dimensional vectors as $\boldsymbol{y}_i = [y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(D)}]^{\mathrm{T}}$. Then the conditional distribution

$$q_\theta(\boldsymbol{y}_j|\boldsymbol{x}_i) = \prod_{d=1}^{D} (2\pi(\sigma_i^{(d)})^2)^{-1/2} \exp\left\{-\frac{(y_j^{(d)} - \mu_i^{(d)})^2}{2(\sigma_i^{(d)})^2}\right\}. \tag{19}$$

Therefore, the log-ratio

$$\begin{aligned}
\log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i) - \log q_\theta(\boldsymbol{y}_j|\boldsymbol{x}_i) =& \log\left(\prod_{d=1}^{D}(2\pi(\sigma_i^{(d)})^2)^{-1/2}\right) + \log\left(\prod_{d=1}^{D}\exp\left\{-\frac{(y_i^{(d)} - \mu_i^{(d)})^2}{2(\sigma_i^{(d)})^2}\right\}\right) \\
&- \log\left(\prod_{d=1}^{D}(2\pi(\sigma_i^{(d)})^2)^{-1/2}\right) - \log\left(\prod_{d=1}^{D}\exp\left\{-\frac{(y_j^{(d)} - \mu_i^{(d)})^2}{2(\sigma_i^{(d)})^2}\right\}\right) \\
=& \sum_{d=1}^{D}\left\{-\frac{(y_i^{(d)} - \mu_i^{(d)})^2}{2(\sigma_i^{(d)})^2}\right\} - \sum_{d=1}^{D}\left\{-\frac{(y_j^{(d)} - \mu_i^{(d)})^2}{2(\sigma_i^{(d)})^2}\right\} \\
=& -\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\mathrm{Diag}[\boldsymbol{\sigma}_i^{-2}](\boldsymbol{y}_i - \boldsymbol{\mu}_i) + \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\mathrm{Diag}[\boldsymbol{\sigma}_i^{-2}](\boldsymbol{y}_j - \boldsymbol{\mu}_i),
\end{aligned}$$

where $\mathrm{Diag}[\boldsymbol{\sigma}_i^{-2}]$ is a $D \times D$ diagonal matrix with $(\mathrm{Diag}[\boldsymbol{\sigma}_i^{-2}])_{d,d} = (\sigma_i^{(d)})^{-2}$, $d = 1, 2, \ldots, D$. The vCLUB estimator can be calcuated by

$$\begin{aligned}
\hat{\mathrm{I}}_{\mathrm{vCLUB}} =& \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i) - \log q_\theta(\boldsymbol{y}_j|\boldsymbol{x}_i)] \\
=& -\frac{1}{2}\left\{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\mathrm{Diag}[\boldsymbol{\sigma}_i^{-2}](\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right\} + \frac{1}{2}\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\mathrm{Diag}[\boldsymbol{\sigma}_i^{-2}](\boldsymbol{y}_j - \boldsymbol{\mu}_i)\right\}.
\end{aligned}$$

E. Detailed Experimental Setups

Information Bottleneck: For the experiment on information bottleneck, we follow the setup from Alemi et al. (2016). The parameters $\mu_\sigma(\boldsymbol{x})$ and $\Sigma_\sigma(\boldsymbol{x})$ are the output from a MLP with layers $784 \to 1024 \to 1024 \to 2K$, where K is the size of the bottleneck. We set $K = 256$. For the variational classifier to implement the Barber-Agakov MI lower bound, the structure is set to a one-layer MLP. The batch size is 100. We set our learning rate to 10^{-4}, with an exponential decay rate of 0.97 and a decay step of 1200.

Domain Adaptation: The network is constructed as follows. Both feature extractors ($i.e.$, E_c and E_d) are nine-layer convolutional neural network with leaky ReLU non-linearities. The content classifier C and the domain discriminator D are a one-layer and a two-layer MLPs, respectively. Images from each domain are normalized using Gaussian normalization.

Classifier C	Discriminator D	Extractor (both E_c and E_d)
Content feature $\boldsymbol{z}_c^s$	Domain feature $\boldsymbol{z}_d$	Input data $\boldsymbol{x}$
MLP output $C(\boldsymbol{z}_c^s)$ with shape 10	MLP, 64 ReLU MLP output $D(\boldsymbol{z}_d)$ with shape 2	3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 2×2 max pool, stride 2, dropout, $p = 0.5$, Gaussian noise, $\sigma = 1$ 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 2×2 max pool, stride 2, dropout, $p = 0.5$, Gaussian noise, $\sigma = 1$ 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 global average pool, output feature with shape 64

F. Numerical Results of MI Estimation

We report the numerical results of MI estimation quality in Table 3. The detailed setups are provided in Section 4.1. Our CLUB estimator has the lowest estimation error when the ground-truth MI value goes larger.

	Gaussian					Cubic				
MI true value	2	4	6	8	10	2	4	6	8	10
VUB	3.85	15.33	34.37	61.25	95.70	2.09	10.38	25.56	47.84	77.59
NWJ	1.67	7.20	17.46	33.26	55.34	1.10	5.54	14.68	30.25	51.07
MINE	1.61	6.66	16.01	29.60	49.87	1.53	6.58	17.4	34.20	59.46
NCE	0.59	2.85	8.56	19.66	37.79	0.45	1.89	6.70	17.48	35.86
L1Out	**0.13**	**0.11**	0.75	4.65	17.08	2.30	5.58	8.92	8.27	7.19
CLUB	0.15	0.12	**0.70**	**4.53**	**16.57**	**2.22**	**5.89**	8.25	**8.23**	**6.93**
CLUBSample	0.38	0.44	1.31	5.30	17.63	2.37	**5.89**	**8.07**	8.87	7.54

Table 3. MSE of MI estimation