Appendix of Perceptual Indistinguishability-Net (PI-Net): Facial Image Obfuscation with Manipulable Semantics

Jia-Wei Chen[1,3] Li-Ju Chen[3] Chia-Mu Yu[2] Chun-Shien Lu[1,3]
[1]Institute of Information Science, Academia Sinica [2]National Yang Ming Chiao Tung University
[3]Research Center for Information Technology Innovation, Academia Sinica
{jiawei, lijuchen}@citi.sinica.edu.tw chiamuyu@nycu.edu.tw lcs@iis.sinica.edu.tw

As a number of notations are used throughout the paper, their definitions are summarized in the table below.

Notation	Definition
m	Number of facial attributes
J	Number of clusters
N	Number of samples
$\mathcal{S} \subset \mathbb{R}^m$	The semantic space that formed by m attributes
$\mathcal{X} \subset \mathbb{R}^d$	The latent space that learned in GANs
$\mathcal{Y} \subset \mathbb{R}^k$	The space of transformed code
$\mathcal{I}$	The image space
$\mathcal{Z}$	The set of outcomes from the probability function
$\mathcal{D}$	The training dataset
$s \in \mathcal{S}$	The semantic vector with each entry indicates whether an attribute exists or not
$x \ \in \mathcal{X}$	The latent code
$x' \in \mathcal{X}$	The latent code of adjacent image
$x_i^a \in \mathcal{X}$	The anchor of the i-th data
$x_i^p \in \mathcal{X}$	The positive sample of the i-th data
$x_i^n \in \mathcal{X}$	The negative sample of the i-th data
$I \in \mathcal{I}$	The input image
$d_{\mathcal{X}}$	The distance metric for latent space $\mathcal{X}$
$\rho(x, x')$	The perceptual distance between two latent codes x and x'
K	The mechanism for assigning the probability distribution to each latent code
$K(x)$	The probability distribution over x
$K(x)(Z)$	The probability that the obfuscated latent code belongs to the set Z when the original latent code is x
H	The mechanism that satisfies $d_{\mathcal{Y}}$-privacy
M	The mechanism that satisfies $\Delta\ d_{\mathcal{X}}$-privacy
G	The generator in GANs
$F:\ \mathcal{I} \to \mathcal{X}$	The GAN inversion model that map $\mathcal{I}$ to $\mathcal{X} \in \mathbb{R}^d$
$f_S : \mathcal{I} \to \mathcal{S}$	The semantic scoring function can evaluate each image's facial attribute components
$f_A : \mathcal{X} \to \mathcal{S}$	The trained attribute classification model, where $\mathcal{X} \in\ \mathbb{R}^d$, and $\mathcal{S} \in\ \mathbb{R}^m$
$\mathcal{F}_{\mathcal{Z}}$	σ-algebra over $\mathcal{Z}$
$\mathcal{P}(\mathcal{Z})$	The set of probability function over $\mathcal{Z}$
$D_{\epsilon,k}$	The probability density function for sampling the noise with ϵ privacy budget in k-dimensional space
Δ	The sensitivity
ϵ	The privacy budget
μ	The marginal threshold in triplet loss
ω, θ	The model weight of decoding and encoding network
α, β	The learning rate of decoding and encoding network
C_j	The perceptual distance of j-th cluster

In Section 4, we claim that the noise injection mechanism H satisfies $d_{\mathcal{Y}}$-privacy and M satisfies ϵ-PI, both without a proof. We particularly note that though Fan [11] proposed the original design of H but did not provide the proof. Here, we first provide a proof that H satisfies $d_{\mathcal{Y}}$-privacy. Based on such a result, we provide a formal proof that M satisfies ϵ-PI.

Lemma 1. If $H : \mathcal{Y} \to \mathcal{P}(\mathcal{Y})$ samples $\boldsymbol{y}$ from a given $\boldsymbol{y_0}$ with the following probability density function (PDF):

$$D_{\epsilon,k}\left(\boldsymbol{y_0}\right)\left(\boldsymbol{y}\right) = C_{\epsilon,k} e^{-\epsilon \cdot d_{\mathcal{Y}}(\boldsymbol{y}_0, \boldsymbol{y})},$$

then H satisfies $d_{\mathcal{Y}}$-privacy, where $\mathcal{P}(\mathcal{Y})$ is the set of probability measures over $\mathcal{Y}$, $C_{\epsilon,k} = \frac{1}{2}\left(\frac{\epsilon}{\sqrt{\pi}}\right)^k \frac{\left(\frac{k}{2}-1\right)!}{(k-1)!}$, and $d_{\mathcal{Y}}$ is the k-dimensional Euclidean distance.

Proof. For unifying the symbol usage in Definition 3.3 and Definition 4.3, we substitute $\boldsymbol{y}$ with $\boldsymbol{z} \in \mathcal{Y}$, which is the output sampled from the PDF. After that, we have

$$D_{\epsilon,k}\left(\boldsymbol{y_0}\right)\left(\boldsymbol{z}\right) = C_{\epsilon,k} e^{-\epsilon \cdot d_{\mathcal{Y}}(\boldsymbol{y}_0, \boldsymbol{z})}.$$

The probability of sampling the output $\boldsymbol{z}$ belonging to the set Z at given $\boldsymbol{y_0}$ can be computed as:

$$H(y)(Z) = \int_Z D_{\epsilon,k}\left(\boldsymbol{y_0}\right)\left(\boldsymbol{z}\right) d\boldsymbol{z},$$

where y is identical to $\boldsymbol{y_0}$, $\forall Z \in \mathcal{F}_{\mathcal{Y}}$, and $\mathcal{F}_{\mathcal{Y}}$ is a σ-algebra over $\mathcal{Y}$. By triangular inequality, we derive

$$\begin{aligned}\int_Z D_{\epsilon,k}\left(\boldsymbol{y_0}\right)\left(\boldsymbol{z}\right) d\boldsymbol{z} &= \int_Z C_{\epsilon,k} e^{-\epsilon \cdot d_{\mathcal{Y}}(\boldsymbol{y}_0, \boldsymbol{z})}\, d\boldsymbol{z} \leq \int_Z C_{\epsilon,k} e^{-\epsilon \cdot \left(d_{\mathcal{Y}}\left(\boldsymbol{y'}, \boldsymbol{z}\right) - d_{\mathcal{Y}}\left(\boldsymbol{y}_0, \boldsymbol{y'}\right)\right)}\, d\boldsymbol{z} \\ &= e^{\epsilon \cdot d_{\mathcal{Y}}\left(\boldsymbol{y}_0, \boldsymbol{y'}\right)} \int_Z C_{\epsilon,k} e^{-\epsilon \cdot d_{\mathcal{Y}}\left(\boldsymbol{y'}, \boldsymbol{z}\right)}\, d\boldsymbol{z} \\ &= e^{\epsilon \cdot d_{\mathcal{Y}}\left(\boldsymbol{y}_0, \boldsymbol{y'}\right)} \int_Z D_{\epsilon,k}\left(\boldsymbol{y'}\right)\left(\boldsymbol{z}\right) d\boldsymbol{z}.\end{aligned}$$

Thus, we can conclude $H(y)(Z) \leq e^{\epsilon \cdot d_{\mathcal{Y}}\left(y, y'\right)} H(y')(Z)$. According to the definition of metric privacy (Definition 3.3), H satisfies $d_{\mathcal{Y}}$-privacy.

□

Theorem 1. *If H satisfies $d_{\mathcal{Y}}$-privacy, then in the case of sensitivity $0 \leq \Delta \leq 1$, $M : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$ defined as $M(x) = (H \circ f)(x) = H(f(x))$ satisfies ϵ-PI.*

Proof. According to Lemma 1, H has the following properties as it satisfies $d_{\mathcal{Y}}$-privacy:

$$H(y)(Z) \leq e^{\epsilon \cdot d_{\mathcal{Y}}(y, y')} H(y')(Z).$$

Let $y = f(x)$ and $y' = f(x')$, we have

$$\ln \left| \frac{H(f(x))(Z)}{H(f(x'))(Z)} \right| \leq \epsilon \cdot d_{\mathcal{Y}}(y, y').$$

Next, by substituting $H(f(x))$ with $M(x)$ and according to the definition of Δ-sensitivity (Definition 3.4), we have

$$\ln \left| \frac{M(x)(Z)}{M(x')(Z)} \right| \leq \epsilon \cdot \Delta d_{\mathcal{X}}(x, x').$$

Therefore, in the case of $0 \leq \Delta \leq 1$, we can derive:

$$\ln \left| \frac{M(x)(Z)}{M(x')(Z)} \right| \leq \epsilon \cdot d_{\mathcal{X}}(x, x'),$$

which implies that M satisfies ϵ-PI according to Definition 4.3. □

Intuitively, ϵ-PI refers to that the closer the perceptual distance between the latent codes, the closer the probability of producing the same obfuscated output, thus making it more difficult for an adversary to distinguish between true codes. Therefore, if the perceptual distance between the two latent codes increases after the latent codes are transformed by encoding network f, that is, $\Delta > 1$, the probability of being distinguished by an adversary increases.

In Theorem 1, we proved that M satisfies ϵ-PI, when H satisfies $d_{\mathcal{Y}}$-privacy and $0 \leq \Delta \leq 1$. However, as discussed in Section 4, because of the need to calculate the sensitivity of each cluster Δ_j, it is difficult to apply constraints in the training phase to the encoding network f to achieve $0 \leq \Delta_j \leq 1$. (For brevity, the subscript j for the cluster index is omitted below.)

We use the clipping function $\tilde{f} : \mathcal{X} \rightarrow \mathcal{Y}$, which is defined as $\tilde{f}(x) = f(x') + \boldsymbol{g}/(\|\boldsymbol{g}\|/C)$ to bound the sensitivity of the encoding network f over x , where $\boldsymbol{g} = f(x) - f(x')$ and x and x' are adjacent. That is, if $\|f(x) - f(x')\| \geq C$, then the $f(x)$ is clipped to get $\tilde{f}(x)$ to ensure that $\|\tilde{f}(x) - f(x')\| \leq C$, where $C = \Delta\|x - x'\|$. We prove that H combined with the $\tilde{f}$ satisfies ϵ-PI in case of configurable sensitivity being set to $0 \leq \Delta \leq 0.5$.

Theorem 2. *Given the range adjacent to latent code x, that is, $\|x - x'\| \leq \beta$, if H satisfies $d_{\mathcal{Y}}$-privacy and uses the $\tilde{f}$ as a clipping function, $M(x) = (H \circ \tilde{f})(x)$ satisfies ϵ-PI when the configurable sensitivity is set to $0 \leq \Delta \leq 0.5$.*

Proof. We first define the clipped $f(x)$ as $\tilde{f}(x)$, which implies that $\tilde{f}(x) - f(x') = \boldsymbol{g}/(\|\boldsymbol{g}\|/C) = \frac{f(x)-f(x')\cdot C}{\|f(x)-f(x')\|}$.

Here x' is randomly sampled from those latent codes adjacent to x, to be used as an anchor point to estimate the local sensitivity around x. We substitute x' with x^a for the anchor point to avoid subsequent semantic conflicts. Thus we have

$$\tilde{f}(x) = f(x^a) + \frac{f(x) - f(x^a) \cdot C_1}{\|f(x) - f(x^a)\|}, \quad C_1 = \Delta\|x - x^a\|.$$

For all other latent codes x' adjacent to x, the clipped value of x' is given by

$$\tilde{f}(x') = f(x^a) + \frac{f(x') - f(x^a) \cdot C_2}{\|f(x') - f(x^a)\|}, \quad C_2 = \Delta\|x' - x^a\|.$$

Let $\hat{\boldsymbol{g_1}} = \frac{f(x)-f(x^a)}{\|f(x)-f(x^a)\|}$ and $\hat{\boldsymbol{g_2}} = \frac{f(x')-f(x^a)}{\|f(x')-f(x^a)\|}$ be normalized unit vectors. Based on triangle inequality, we derive

$$\|\tilde{f}(x) - \tilde{f}(x')\| = \|C_1 \cdot \hat{\boldsymbol{g_1}} - C_2 \cdot \hat{\boldsymbol{g_2}}\| \leq C_1\|\hat{\boldsymbol{g_1}}\| + C_2\|\hat{\boldsymbol{g_2}}\| = C_1 + C_2 \leq 2\Delta\beta.$$

Let $c = \tilde{f}(x)$ and $c' = \tilde{f}(x')$, we can quickly follow the proofs of Lemma 1 and Theorem 1 to derive

$$\begin{aligned} \int_Z C_{\epsilon,k} e^{-\epsilon \cdot d_{\mathcal{Y}}(\boldsymbol{c_0},\boldsymbol{z})}\, d\boldsymbol{z} &\leq \int_Z C_{\epsilon,k} e^{-\epsilon \cdot \left(d_{\mathcal{Y}}\left(\boldsymbol{c'},\boldsymbol{z}\right) - d_{\mathcal{Y}}\left(\boldsymbol{c_0},\boldsymbol{c'}\right)\right)}\, d\boldsymbol{z} \\ &= e^{\epsilon \cdot d_{\mathcal{Y}}\left(\boldsymbol{c_0},\boldsymbol{c'}\right)} \int_Z C_{\epsilon,k} e^{-\epsilon \cdot d_{\mathcal{Y}}\left(\boldsymbol{c'},\boldsymbol{z}\right)}\, d\boldsymbol{z}, \end{aligned}$$

and then we have

$$H(c)(Z) \leq e^{\epsilon \cdot d_{\mathcal{Y}}\left(c,c'\right)} H(c')(Z).$$

Since $c = \tilde{f}(x)$ and $c' = \tilde{f}(x')$, we have

$$H(\tilde{f}(x))(Z) \leq e^{\epsilon \cdot d_{\mathcal{Y}}\left(c,c'\right)} H(\tilde{f}(x'))(Z).$$

After substitute $H(\tilde{f}(x))(Z)$ with $M(x)(Z)$, and based on $d_{\mathcal{Y}}(c, c') = \|\tilde{f}(x) - \tilde{f}(x')\| \leq 2\Delta\beta$, we derive

$$\ln\left|\frac{M(x)(Z)}{M(x')(Z)}\right| \leq \epsilon \cdot 2\Delta\beta.$$

Therefore, setting the configurable sensitivity to $0 \leq \Delta \leq 0.5$, M satisfies ϵ-PI. □