MaSS: Multi-attribute Selective Suppression for Utility-preserving Data Transformation from an Information-theoretic Perspective

Yizhuo Chen [*] [1] [2] **Chun-Fu (Richard) Chen** [*] [2] **Hsiang Hsu** [2] **Shaohan Hu** [2] **Marco Pistoia** [2] **Tarek Abdelzaher** [1]

Abstract

The growing richness of large-scale datasets has been crucial in driving the rapid advancement and wide adoption of machine learning technologies. The massive collection and usage of data, however, pose an increasing risk for people's private and sensitive information due to either inadvertent mishandling or malicious exploitation. Besides legislative solutions, many technical approaches have been proposed towards data privacy protection. However, they bear various limitations such as leading to degraded data availability and utility, or relying on heuristics and lacking solid theoretical bases. To overcome these limitations, we propose a formal information-theoretic definition for this utility-preserving privacy protection problem, and design a data-driven learnable data transformation framework that is capable of selectively suppressing sensitive attributes from target datasets while preserving the other useful attributes, regardless of whether or not they are known in advance or explicitly annotated for preservation. We provide rigorous theoretical analyses on the operational bounds for our framework, and carry out comprehensive experimental evaluations using datasets of a variety of modalities, including facial images, voice audio clips, and human activity motion sensor signals. Results demonstrate the effectiveness and generalizability of our method under various configurations on a multitude of tasks. Our source code is available at this URL.

1. Introduction

The recent rapid advances and wide adoption of machine learning technologies are largely attributed to not only the explosive growth in raw computing power, but also the unprecedented availability of large-scale datasets, for example, the monumental computer vision dataset ImageNet (Deng et al., 2009), the large multi-lingual web corpus Common Crawl (2023), and the widely used UCI HAR dataset (Anguita et al., 2013). While the vast amount of data serves as the rich basis for machine learning algorithms to learn from, the ubiquitous data collection and usage have drawn serious privacy concerns since people's private and sensitive information could be leaked through inadvertent mishandling as well as deliberate malicious exploitation. Therefore, various regulatory policies, such as GDPR and CCPA, have been drafted and put in place to guardrail the handling and usage of data. While such legislative solutions do generally help mitigate the privacy concerns, they also tend to pose blanket restrictions that result in degraded data availability. Therefore, there has been a growing interest in developing more sophisticated, flexible technical solutions.

Towards this goal, many techniques have been proposed. One of the most well-known studies is the protection against membership inference attacks, also known as Differential Privacy (Dwork et al., 2014; Mironov, 2017; Abadi et al., 2016). It focuses on preventing attackers from differentiating between two neighboring sets of samples by observing the change in the distribution of output statistics. Another widely discussed notion of privacy is the protection against attribute inference attacks, often referred to as Information-theoretic (IT) Privacy (Hukkelås et al., 2019; Bertran et al., 2019; Huang et al., 2018; Hsu et al., 2020). This line of work aims at transforming a dataset to remove or suppress its sensitive attributes while preserving its utility for downstream tasks. In this paper, we focus our discussion on providing IT Privacy protection.

Developing a data transformation framework for IT Privacy presents multiple challenges. Specifically, we identified 5 desired properties for an IT Privacy data transformation framework, which can be summarized as *SUIFT*: 1) *S*ensitivity suppression: the capability to suppress annotated sensitive attributes from the dataset; 2) *U*tility preservation:

[*] Equal contribution [1] Department of Computer Science, University of Illinois Urbana-Champaign, USA [2] Global Technology Applied Research, JPMorgan Chase, USA. Correspondence to: Yizhuo Chen <yizhuoc@illinois.edu>, Chun-Fu (Richard) Chen <richard.cf.chen@jpmchase.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Table 1. A summary of related works on IT Privacy. ●, ○ and ◐ indicate *satisfied*, *not satisfied* and *partially satisfied*, respectively. S, U, I, F, and T are abbreviations of *Sensitivity suppression*, *Utility preservation*, *Invariance of sample space*, *Feature management without annotation*, and *Theoretical basis* respectively. Apart from our method, MaSS (to be introduced shortly, and described in detail in Section 4), none of the listed methods fully satisfy all components of SUIFT (discussed in detail in Section 2 and Appendix B).

Method	S	U	I	F	T
DeepPrivacy (Hukkelås et al., 2019)	●	◐	●	◐	○
CiaGAN (Maximov et al., 2020)	●	◐	●	◐	○
Hsu et al. (2020)	●	○	●	◐	●
ALR (Bertran et al., 2019)	●	●	●	○	●
PPDAR (Wu et al., 2020)	●	●	●	○	○
BDQ (Kumawat & Nagahara, 2022)	●	●	●	○	○
ALFR (Edwards & Storkey, 2016)	◐	●	○	●	◐
LAFTR (Madras et al., 2018)	◐	●	○	●	◐
GAP (Huang et al., 2018)	●	○	●	●	◐
MSDA (Malekzadeh et al., 2019)	●	●	●	●	◐
SPAct (Dave et al., 2022)	○	●	●	●	○
MaSS (our method)	●	●	●	●	●

the capability to preserve specifically annotated useful attributes in the dataset to facilitate downstream usage; 3) *I*nvariance of sample space: keeping the transformed data in the original space as the input data, to enable plug-in usability for pretrained off-the-shelf models and to deliver better re-usability for the community; 4) *F*eature management without annotation: the capability to manage unannotated generic features in the dataset, by either suppressing them or preserving them when they are considered either useful or sensitive. 5) *T*heoretical basis: all the proposed components of the data transformation frameworks being entirely driven by a unified information-theoretic basis to ensure safety.

Various techniques have been proposed towards the goal of SUIFT in IT Privacy, as summarized in Table 1. However, each of them is limited in missing some of the desired properties of SUIFT. For example, Bertran et al. (2019), Wu et al. (2020), and Kumawat & Nagahara (2022) can only ensure the predictability in the transformed data for attributes that have already been explicitly annotated for preservation; no considerations are given to managing data's unannotated attributes. On the other hand, Huang et al. (2018), Malekzadeh et al. (2019) and Madras et al. (2018) do account for unannotated attributes, but their designs for unannotated attributes preservation are mostly heuristic-driven and lack rigorous theoretical bases, which could limit their applicability, especially for scenarios involving highly sensitive information.

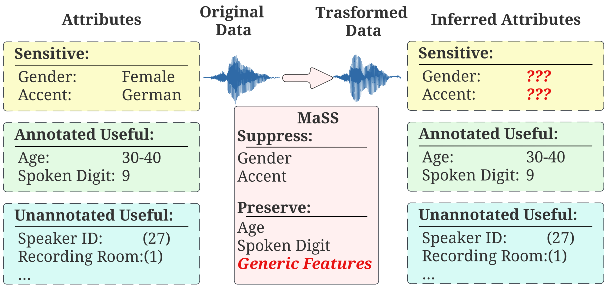

Figure 1. An illustrative use case of MaSS: The original data sample is a voice clip of a person speaking a digit, where its attributes "gender" and "accent" are considered as sensitive, while its "age" and "spoken digit" are annotated as useful. We are also interested in preserving generic features of the data. For example, the voice clip may contain attributes such as "speaker ID" or "recording room" that could prove to be useful down the road, but are not necessarily explicitly annotated yet at the time of processing. After the transformation of MaSS, sensitive attributes can no longer be accurately inferred, but the other useful attributes are preserved in the transformed data.

To address these limitations, in this paper we present MaSS, a **M**ulti-**a**ttribute **S**elective **S**uppression framework that aims at satisfying all 5 components of SUIFT. Specifically, we formulate the IT Privacy as an optimization problem from the perspective of information theory, and then convert the optimization problem into a fully differentiable trainable framework parameterized by neural networks, with sound analyses on the design derivation and operational bounds. MaSS is capable of suppressing multiple selected sensitive attributes, and preserving multiple useful attributes regardless of whether they are annotated or not. An illustrative use case of MaSS is shown in Figure 1. We also compare MaSS with various baselines extensively on three datasets of different modalities, namely voice recordings, human activity motion sensor signals, and facial images, and show its practical effectiveness under various configurations.

The contributions of this paper are summarized as follows: 1) We propose MaSS, an information theory driven data transformation framework satisfying all 5 identified desirable properties in IT Privacy, namely SUIFT; 2) We provide rigorous theoretical analyses on the design derivation and operational bounds of our proposed multi-attribute data transformation framework; and 3) We experimentally evaluate MaSS extensively on voice audio, human activity motion sensor signal, and facial image datasets, and demonstrate its effectiveness and generalizability.

Omitted proofs, details on experiment setups and training, and additional results are included in the Appendix.

2. Related Works

Privacy-preserving mechanisms. A privacy-preserving mechanism ensures privacy by randomizing a function of data in order to thwart unwanted inferences. There are two selections of the functions that lead to different privacy notions. If the function is the output of a query over a database, the privacy notion is termed differential privacy (DP) (Dwork et al., 2006), which requires the results of a query be approximately the same for small perturbations of data, and can usually be achieved by additive noise mechanisms (e.g., Gaussian, Laplacian or exponential noise (Dwork et al., 2014; Sun et al., 2020; Zhang et al., 2018; Abadi et al., 2016)). Different from DP, if the function is a conditional distribution that anonymizes sensitive information in the data while preserving non-sensitive information, it leads to the other privacy notion called information-theoretic (IT) privacy. The motivation behind IT privacy is to improve the data quality after anonymization with the additional information of the utility attributes. See Hsu et al. (2021a) for a more detailed discussion on the two privacy notions. Since our goal is to not only suppress the sensitive attributes but also preserve the data utility concurrently, the MaSS framework falls within the field of IT privacy.

IT Privacy protection for annotated attributes. By reviewing related studies for IT Privacy and examining the requirements of downstream applications, we identified 5 desirable properties of IT Privacy mechanisms, namely *SUIFT* as summarized in Section 1 and Table 1. Nevertheless, previous studies proposed for IT Privacy are limited in certain properties. For instance, DeepPrivacy (Hukkelås et al., 2019) employs a CGAN, conditioned on image background and pose features, to synthesize anonymized facial images. To further ensure de-identification, CiaGAN (Maximov et al., 2020) proposes to condition the CGAN on an identity control vector, creating images with fabricated identities. Nevertheless, these methods prioritize visual quality of the generated images over the preservation of utilities, for both annotated and unannotated useful attributes, undermining the data's usefulness for downstream ML tasks. (Hsu et al., 2020) proposes to suppress sensitivity by only locating and obfuscating information-leaking features, but is also limited in providing a mechanism to quantify and preserve the utilities. To explicitly preserve useful annotated attributes, ALR (Bertran et al., 2019) ensures that annotated useful attributes remain predictable in anonymized data, while thwarting inference of sensitive attributes from an information-theoretic perspective. PPDAR (Wu et al., 2020) extends this approach by introducing a cross-entropy-based suppression and preservation loss. This idea is further blended with a prior-based suppression loss by BDQ (Kumawat & Nagahara, 2022). Despite their advancements in preserving annotated useful attributes, these studies do not consider managing the unannotated attributes in the data.

Unannotated attributes management based on heuristics. In the neighboring field of fair representation learning, ALFR (Edwards & Storkey, 2016) proposes to preserve the unannotated attributes by minimizing the ℓ_2 reconstruction loss, while selectively suppress and preserve annotated attributes. Building upon this, LAFTR (Madras et al., 2018) introduces a fairness metrics driven optimization objective for suppression. However, these studies focused on suppressing only one binary sensitive attribute to achieve fairness. In contrast, in IT Privacy literature, GAP (Huang et al., 2018) advocates for suppressing multiple sensitive attribute, and simultaneously contraining the ℓ_2 reconstruction loss. (Malekzadeh et al., 2019) further combines ℓ_2 reconstruction loss with information theoretic losses for annotated attributes. On the other hand, Dave et al. targets their work at suppressing the unannotated attributes of the data, utilizing contrastive learning technique, while ensuring the predictability of annotated attributes. Despite the practical relevance of their handling of unannotated attributes, these works fall short in providing a robust theoretical foundation regarding the derivation and operational bounds of their design, raising concerns in scenarios demanding high safety assurances. We discuss related works in more detail in Appendix B.

3. Problem Formulation

In this paper, we focus on a multi-attribute dataset comprised of original data X, a set of M sensitive attributes $S = (S_1, S_2, \ldots, S_M)$, a set of N annotated useful attributes $U = (U_1, U_2, \ldots, U_N)$, and a set of unannotated useful attributes or generic features F. However, our access is limited to the observable joint distribution $P(X, U, S)$, as opposed to the intrinsic joint distribution $P(X, U, S, F)$. We base our work on pragmatic assumptions that S, U are random variables following finite categorical distributions, allowing the mutual information between S, U, and X to be bounded. Additionally, we presuppose that with the given X, the corresponding annotated attributes S, U are entirely determined (i.e., $P(S_i|X)$ and $P(U_j|X)$ are degenerate distributions). For broad applicability, we do not make assumptions regarding the dimension or distribution family for F and X. Furthermore, we do not assume independence between F and other variables, which means that F may correlate with the joint distribution of X, S, U.

Our goal of IT Privacy is then formulated as finding the optimal data transformation $P_\theta(X'|X)$, where the random variable X' is the transformed data, and the strongest unannotated useful attribute extractor $P_\eta(F|X)$ by solving the following constrained optimization problem:

$$\begin{aligned} \max_{\theta,\eta} \quad & I(X'; F) \\ \text{s.t.} \quad & I(X'; S_i) \leq m_i \text{ and } I(X'; U_j) \geq n_j, \end{aligned} \tag{1}$$

where $I(\cdot, \cdot)$ is Shannon mutual information, $i \in 1 \ldots M$, $j \in 1 \ldots N$, $P_\theta(X'|X)$ and $P_\eta(F|X)$ are parameterized by θ, η respectively. By solving this optimization problem, we try to ensure that, at least n_j nats (the counterpart of bits

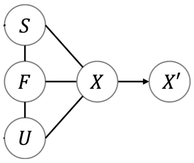

Figure 2. The Markov chain of all variables. F is correlated with U, S, X. X' is only dependent on X.

with Napierian base) information is preserved for U_j in the transformed data X', at most m_i nats information is leaked for S_i in X', and the information preserved for F in X' is maximized when the most informative F is extracted from X. For clarity, the Markov Chain of variables U, S, F, X, and X' corresponding to our problem formulation is summarized in Figure 2.

3.1. Operational Bounds

In preparation for solving our optimization problem formulated in Equation 1, a thorough comprehension of its operational bounds is imperative. Specifically, we will elucidate formally that the parameters m_i and n_j must be chosen under certain constraints to ensure the solvability of Equation 1. Moreover, it will be established that the optimization objective $I(X'; F)$ has an upperbound which can not be exceeded.

Theorem 3.1. *For the Markov Chain shown in Figure 2, there exists a solution to the optimization problem defined in Equation 1, only if for any pair of (m_i, n_j), $i \in 1 \dots M$, $j \in 1 \dots N$, it satisfies:*

$$n_j \leq m_i + I(X; U_j | S_i), \quad n_j \leq I(X; U_j) \text{ and } m_i \geq 0. \tag{2}$$

Under the assumptions that $P(S_i|X)$ and $P(U_j|X)$ are degenerate distributions, Equation 2 can be simplified to

$$n_j \leq m_i + H(U_j | S_i), \quad n_j \leq H(U_j) \text{ and } m_i \geq 0. \tag{3}$$

where $H(\cdot)$ is Shannon entropy.

Besides, for any m_i, $i \in 1 \dots M$, $I(X'; F)$ is upper bounded by

$$I(X'; F) \leq H(X|S_i) + m_i. \tag{4}$$

Please refer to Appendix A.1 for the proof. It is important to note that the values in Equation 3, specifically $H(U_j|S_i)$ and $H(U_j)$, are independent of our model's parameters and can be computed prior to training to assess solvability.

To understand the requirement of $n_j \leq m_i + H(U_j|S_i)$ in Equation 3 intuitively, consider a facial image dataset with two attributes "hair color" and "age". The high correlation between these attributes is evident, as older individuals are more likely to have white or gray hair. Should "age" be suppressed with a small m_{age}, the "hair color" information in the facial image must be correspondingly sacrificed to prevent inadvertently disclosing "age" information. The extent of this sacrifice is intuitively determined by the certainty with which "age" predicts "hair color".

To intuitively understand Equation 4, revert to the example we discussed above. When suppressing "age", certain features that were in X no longer reside in X', such as hair color and wrinkles, etc. This results in a necessary sacrifice of the information of F contained in X'. The extent of sacrifice is determined by the certainty with which "age" determines the image.

4. Data-driven Learnable Data Transformation Framework

Building upon our problem formulation, we design a learnable data-driven data transformation framework as an approximation to Equation 1, which we call Multi-attribute Selective Suppression (abbreviated as MaSS). Notably, we adopt neural networks as conditional probability approximators in our framework, and design our training objectives to be fully differentiable, allowing gradient descent based optimization. MaSS can be flexibly implemented with various neural network structures to adapt to different application requirements. The overarching architecture of MaSS is depicted in Figure 3. In the subsequent sections, we elaborate on the modules of MaSS in detail.

4.1. Data Transformation

The data transformation module takes in the original data X and outputs the transformed data X'. In line with Bertran et al., we parameterize $P_\theta(X'|X)$ as a neural network $X' = g_\theta(X, a)$, wherein a is a noise variable sampled from a multi-variate unit Gaussian distribution, serving as the source of randomness for X'.

4.2. Sensitive Attributes Suppression

Given the transformed data, we calculate a suppression loss $L_{S,i}$ for each of the sensitive attributes, which is differentiable and can be minimized to achieve the constraint $I(X'; S_i) \leq m_i$ mentioned in Equation 1. Next, we discuss in depth on the derivation of $L_{S,i}$.

The direct computation of $I(X'; S_i)$ is infeasible because of the intractability of $P(S_i|X')$. Consequently, we incorporate an *adversarial* neural network $P_{\phi_i}(S_i|X')$ as an estimation of $P(S_i|X')$, where ϕ_i is trained with the cross-entropy loss used in traditional supervised learning method:

$$\begin{aligned} L_{CE}(S_i) &= \mathbb{E}_{P(X)P_\theta(X'|X)} \left[H(P(S_i|X), P_{\phi_i}(S_i|X')) \right], \\ \phi_i &= \arg\min_{\phi_i} L_{CE}(S_i), \end{aligned} \tag{5}$$

where $H(\cdot, \cdot)$ denotes cross-entropy and the expectation is

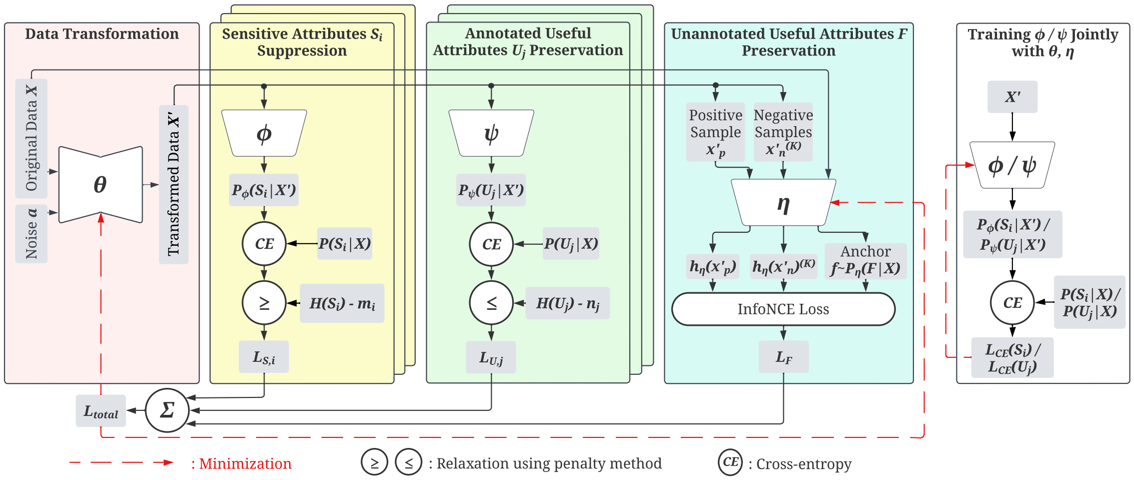

Figure 3. The overall architecture of MaSS. The data transformation module converts the original data into a transformed version. Then the transformed data is sent to both the sensitive attributes suppression module and the annotated useful attributes preservation module, to calculate a relaxed suppression or preservation loss for each attribute respectively. Additionally, the original and transformed data are sent to the unannotated useful attributes preservation module to calculate a contrastive loss. Finally, these losses are aggregated to minimize θ and η jointly. ϕ, ψ are optimized with traditional supervised learning.

estimated using mini-batch during training. Under the assumption that S_i can be fully determined given X, $P(S_i|X)$ effectively refers to the deterministic ground-truth label of each sample.

With the help of ϕ_i, the mutual information $I(X'; S_i)$ can be estimated as

$$\begin{aligned} I(X'; S_i) &\approx \mathbb{E}_{P(X)P_\theta(X'|X)P(S_i|X)}\left[\log \frac{P_{\phi_i}(S_i|X')}{P(S_i)}\right] \\ &= H(S_i) - L_{CE}(S_i), \end{aligned} \tag{6}$$

where $H(S_i)$ is a constant for each dataset and can be calculated before training. Consequently, we can also convert the constraint $I(X'; S_i) \leq m_i$ in Equation 1 to

$$H(S_i) - m_i \leq L_{CE}(S_i). \tag{7}$$

Following Bertran et al., we relax the Constraint 7 to a differentiable loss $L_{S,i}$ function eligible for gradient descent using the penalty method, which can be written as

$$\begin{aligned} d_{S,i} &= \min(L_{CE}(S_i) + m_i - H(S_i), 0), \\ L_{S,i} &= d_{S,i}^2 + |d_{S,i}|. \end{aligned} \tag{8}$$

4.3. Annotated Useful Attributes Preservation

The annotated useful attributes preservation module follows a symmetric design and derivation as the annotated sensitive attributes suppression module. Analogously, a differentiable preservation loss $L_{U,j}$ is calculated for each useful attribute to achieve the constraint $I(X'; U_j) \geq n_j$ in Equation 1. A *collaborative* neural network $P_{\psi_j}(U_j|X')$ is also introduced to estimate $P(U_j|X')$, which is trained with cross-entropy loss:

$$\begin{aligned} L_{CE}(U_j) &= \mathbb{E}_{P(X)P_\theta(X'|X)}[H(P(U_j|X), P_{\psi_j}(U_j|X'))], \\ \psi_j &= \arg\min_{\psi_j} L_{CE}(U_j), \end{aligned} \tag{9}$$

Following the same derivation as Section 4.2, we can convert the constraint $I(X'; U_j) \geq n_j$ to

$$H(U_j) - n_j \geq L_{CE}(U_j). \tag{10}$$

which can also be relaxed into a differentiable loss $L_{U,j}$ using penalty method:

$$\begin{aligned} d_{U,j} &= \max(L_{CE}(U_j) + n_j - H(U_j), 0) \\ L_{U,j} &= d_{U,j}^2 + |d_{U,j}|. \end{aligned} \tag{11}$$

In order to accelerate the training process, we further propose to pre-train an attribute inference network on original data X for each S_i, U_i, denoted as $\phi_{i,0}$ and $\psi_{j,0}$ respectively, using the cross-entropy loss. And then we initialize the transformed data attribute inference models ϕ_i and ψ_j with $\phi_{i,0}$ and $\psi_{j,0}$ respectively, so that they can converge faster during training.

Note that, different from our method, previous studies such as Bertran et al. propose to freeze the useful attribute inference model ψ_j during training after it is initialized with $\psi_{j,0}$. However, we abandoned this strategy, because a frozen useful attribute inference model will introduce a noticeable error in estimating $I(X'; U_j)$. Specifically, the estimation error will be $KL(P(U_j|X')||P_{\psi_{j,0}}(U_j|X'))$, where

$P_{\psi_{j,0}}(U_j|X')$ denotes $P(U_j|X')$ estimated with the **frozen** useful attribute inference network $\psi_{j,0}$, $KL(\cdot||\cdot)$ is Kullback–Leibler divergence. This error can be large and even unbounded because $\psi_{j,0}$ is trained to approximate $P(U_j|X)$ rather than $P(U_j|X')$. Please refer to Appendix A.2 for proof and analysis.

4.4. Unannotated Useful Attributes Preservation

The unannotated useful attributes preservation module aims at calculating and maximizing a differentiable loss approximating the negative $I(X';F)$, without any assumption on the distribution family of F and X'. Consequently, we can not approximate $I(X';F)$ using the aforementioned method in Section 4.2 because it requires the assumption that $P(F|X')$ follows a finite categorical distribution.

Moreover, approximating $I(X';F)$ using ℓ_2 reconstruction loss $\|X'-F\|_2$ (or $\|X'-X\|_2$), as did in Huang et al. (2018), Malekzadeh et al. (2019), Edwards & Storkey (2016) and Madras et al. (2018) is also infeasible since it also requires the assumptions that $P(F|X')$ (or $P(X|X')$) follows a fully factorized Gaussian distribution where each element of F (or X) is only dependent on the corresponding element of X' at the same position. In addition, we will also show empirically in Section 5.1 that ℓ_2 reconstruction loss hinders the overall performance seriously due to its unrealistic assumptions.

In order to approximate $I(X';F)$ without assumption on the distribution family of X' or F, we choose the InfoNCE loss L_F (Oord et al., 2018) as the approximation of the shifted negative $I(X';F)$. InfoNCE is known as an effective approximation method for mutual information, regardless of the distribution family of the random variables, and is stable in mini-batch based training. We also tried other mutual information estimator, e.g., MINE (Belghazi et al., 2018), and empirically compared them in Section 5.1.

To calculate L_F, we first sample one anchor f, one positive sample x'_p, and K negative samples $x_n'^{(K)}$ given a specific realization of X, from the conditional distribution $P_\eta(F|X)$, $P_\theta(X'|X)$ and the marginal distribution $P_\theta(X')$ respectively. Then we take expectation over X and all possible sampling of $f, x'_p, x_n'^{(K)}$ to calculate L_F as

$$L_F = \mathbb{E}\left[\log \frac{\mathcal{F}(f,x'_p)}{\mathcal{F}(f,x'_p)+\sum_{x'_n\in x_n'^{(K)}}\mathcal{F}(f,x'_n)}\right], \quad (12)$$

where $\mathcal{F}$ is a score function defined in the same way as SimCLR (Chen et al., 2020), which can be written as

$$\mathcal{F}(f,x') = e^{\cos(f,h(x'))/\tau}, \quad (13)$$

where τ is the temperature hyper-parameter. $h(x')$ is a feature extractor trained jointly with data transformation module, θ. Note that unlike SimCLR, our loss do not sample negative samples from $P(F)$. As proved in Oord et al. (2018), we can approximate $I(X';F)$ as

$$I(X';F) \approx -L_F + \log(K+1). \quad (14)$$

In order to further encourage the transformed data X' to remain in the original sample space of X, we propose to use a single neural network η to parameterize both $h_\eta(x')$ and $P_\eta(F|X)$. This symmetric design can also reduce the number of parameters and hence stabilize the training. Importantly, an alternative interpretation of this design is to apply the InfoNCE loss on X and X' to estimate and maximize $I(X',X)$.

Aligned with pretraining the attribute inference networks, our unannotated useful attributes extractor η is also initialized with η_0 pretrained using InfoNCE loss on the original dataset X. In the pretraining stage we use one sample in the mini-batch as both the anchor and the positive sample and use the other samples in the mini-batch as negative samples.

Analogous to L_F, which is anchored in F space, we can define another InfoNCE loss L'_F anchored in the X' space and use both losses for training. A more detailed elaboration on the calculation and the advantage of InfoNCE loss is presented in Appendix C.1.

4.5. Module Aggregation

Aggregating the losses calculated from all modules above, we convert our original constrained optimization problem defined in Equation 1 into the following differentiable optimization problem:

$$\min_{\theta,\eta} L_{\text{total}} = \frac{L_F+L'_F}{2} + \lambda\left(\sum_i L_{S,i} + \sum_j L_{U,j}\right) \quad (15)$$

where λ is a hyper parameter controlling the degree of relaxation. Equation 15 effectively recovers the constrained optimization problem defined in Equation 1 when $\lambda \to \infty$,.

5. Evaluation

In this section, we present our experimental evaluation of MaSS against several baselines methods using multiple datasets of varying modalities.

5.1. Experimental Setup

Datasets. The evaluation of MaSS is exhaustively conducted on three multi-attribute benchmark datasets of different modalities, namely the AudioMNIST (Becker et al., 2018) dataset for recorded human voices, the Motion Sense (Malekzadeh et al., 2019) dataset for human activity sensor signals, and the Adience (Eidinger et al., 2014) dataset for facial images. We use the raw data points for training on Motion Sense and Adience, whereas we convert the raw data points to feature embeddings for AudioMNIST using state-of-the-art feature extractor HuBERT-B (Hsu et al., 2021b) for training efficiency.

Baselines. We compare our method with 5 baselines, namely ALR (Bertran et al., 2019), GAP (Huang et al., 2018), MSDA (Malekzadeh et al., 2019), BDQ (Kumawat & Nagahara, 2022), and PPDAR (Wu et al., 2020). All 6 methods rely on adversarial training a sensitive attribute inference model. However, ALR, BDQ, and PPDAR do not consider the preservation of unannotated useful attributes, whereas GAP and MSDA do, using a ℓ_2 heuristic loss. Notwithstanding, GAP does not consider the preservation of annotated useful attribute.

Evaluation Metrics. This paper is focused on suppressing sensitive attributes while preserving useful attributes, rather than generating high quality synthetic data. Therefore, we adopt classification accuracy for each attribute on evaluation set as our metric to measure the effectiveness of the suppression or preservation. Specifically, for sensitive attributes, we report the classification accuracy of the adversarially trained classifier ϕ_i. For useful attributes, to ensure a fair comparison with baselines, we report the classification accuracy of a classifier tuned on the transformed data X' and its attributes U_j. The performance is considered better when the sensitive attributes' accuracies are lower and the useful attributes' accuracies are higher.

Furthermore, since the datasets we use are unbalanced, we adopt the classification accuracy of the majority classifier as a lower reference value, which can also be interpreted as the accuracy of *guessing* the attribute without accessing X' (Asoodeh et al., 2018; Liao et al., 2019). On the other hand, we also adopt the accuracy of the $\phi_{i,0}$ and $\psi_{j,0}$ on original data X in the evaluation set as a upper reference value of classification accuracy, which reflects the classification accuracy when no attributes are suppressed.

Based on the lower and upper reference values of classification accuracy, we introduce a noval normalized metric for our task, namely Normalized Accuracy Gain (NAG), which is defined as $\text{NAG} = \max\left(0, \frac{Acc - Acc_{\text{guessing}}}{Acc_{\text{no\_suppression}} - Acc_{\text{guessing}}}\right)$, where Acc denotes classification accuracy. NAG is inherently non-negative, with NAG $= 0$ suggesting that $Acc \leq Acc_{\text{guessing}}$. We consider all $Acc \leq Acc_{\text{guessing}}$ as equally effective, which indicates that this attribute is completely suppressed from X'. NAG can be seen as a more informative indicator of how the classification accuracy of each attribute is increased or decreased. Therefore, we only report NAG throughout the main paper for clarity, while the corresponding results and reference values measured in classification accuracy are also shown in Appendix E.

Table 2. Comparison of the NAG between MaSS, ablations and baselines on Motion Sense. We suppress gender, ID, while preserve activity as if unannotated useful attribute.

Method	Normalized Accuracy Gain		
	gender (↓)	ID (↓)	activity (↑)
ALR	0.0828	0.0432	0.7704
GAP	0.0053	0.0314	0.8379
MSDA	0.0063	0.0708	0.8418
BDQ	0.1178	0.0613	0.7426
PPDAR	0.0000	0.0000	0.6912
MaSS-NF	0.0000	0.0000	0.7275
MaSS-ℓ_2	0.0085	0.0260	0.8156
MaSS-MINE	0.3294	0.1271	0.6847
MaSS	0.0000	0.0026	0.8977

In order to evaluate the performance of MaSS on preserving unannotated useful attributes, we conceal the labels (annotations) of certain annotated attributes during training and only use these labels for evaluation.

Hyperparameters. Throughout our experiments, λ is simply set to 1. When an attribute is suppressed we simply set its mutual information constraint m as 0. Unless otherwise noted, we set the n of all preserved annotated attributes as the maximal value permitted by Equation 3.

Additional detailed descriptions of the datasets, model structures, and optimization process are elaborated in Appendix D. Next we present and discuss our experimental results.

5.2. Evaluation on Human Activity Sensor Signals

We first experiment on the human activity sensor signal dataset, Motion Sense. Initial experiment focuses on suppressing gender and ID attributes, while concealing the labels of the activity attribute, treating activity as an unannotated attribute for preservation. This setup mirrors scenarios aspiring to eliminate sensitive identity-related information from a dataset lacking explicit annotation on non-sensitive attributes. Apart from the 5 baselines described above, we also compare MaSS with 3 ablations to examine the unannotated useful attributes preservation module of MaSS: 1) removing the InfoNCE loss (denoted as MaSS-NF); 2) replacing InfoNCE loss to ℓ_2 reconstruction loss (denoted as MaSS-ℓ_2); and 3) replacing InfoNCE loss to a negative mutual information estimated using MINE (Belghazi et al., 2018) (denoted as MaSS-MINE). Results as shown in Table 2 demonstrate that MaSS attains the highest NAG on the activity attribute compared with all baselines and ablations. Additionally, in comparison to GAP, MSDA and MaSS-ℓ_2, our method showcases a higher NAG on activity and a reduced NAG on both suppressed attributes. We be-

Table 3. Comparison of the NAG between MaSS, ablations and baselines on AudioMNIST. We suppress gender, accent, age, ID, while preserve digit as if an unannotated attribute.

Method	Normalized Accuracy Gain				
	gender (↓)	accent (↓)	age (↓)	ID (↓)	digit (↑)
ALR	0.0000	0.0000	0.0000	0.0004	0.1036
GAP	0.0000	0.0000	0.0000	0.0281	0.9485
MSDA	0.0000	0.0000	0.0000	0.0074	0.9451
BDQ	0.0000	0.0000	0.0000	0.0112	0.5565
PPDAR	0.0000	0.0000	0.0000	0.0016	0.2839
MaSS-NF	0.0000	0.0001	0.0000	0.0000	0.1846
MaSS-ℓ_2	0.0008	0.0001	0.0020	0.0306	0.9517
MaSS-MINE	0.0076	0.0000	0.0112	0.0434	0.5031
MaSS	0.0000	0.0000	0.0000	0.0029	0.9675

lieve it is because the unrealistic assumptions made by ℓ_2 reconstruction loss overly restrict the flexibility of the data transformation. Moreover, MaSS outperforms MaSS-MINE in all attributes, which can be partly attributed to the instability of MINE mutual information estimator in the training process of our task. These results underscore MaSS's proficiency in maintaining a superior balance between preserving meaningful features and suppressing sensitive attributes.

We further experiment with suppressing gender, while preserving ID as annotated, and preserving activity as unannotated. Please refer to Appendix E.1 for results and corresponding analysis.

5.3. Evaluation on Voice Audio Dataset

Next, the application of MaSS is extended to the AudioMNIST dataset. The initial experiment involves the suppression of gender, accent, age, and ID attributes while treating digit as an unannotated attribute for preservation. Results of this experiment are shown in Table 3. We can observe that MaSS achieves the highest NAG on digit compared with all baselines and ablations, as well as a lower or equal NAG on suppressed attributes compared with GAP, MSDA and MaSS-ℓ_2, further substantiating the limitation of the ℓ_2 heuristic reconstruction loss and the effectiveness of MaSS.

In the subsequent experiment, we aim to suppress gender, accent, and age, while preserve digit as annotated and ID as unannotated. This scenario emulates conditions wherein the dataset encompasses both sensitive and useful annotated attributes, alongside with to-be-preserved unannotated attributes. It is observable from the results shown in Table 4 that MaSS secures the highest NAG on ID, along with a NAG on digit that is comparably high to other methods. Notably, although MSDA's NAG on ID is close to MaSS, it adversely bears higher NAG across all suppressed attributes.

Table 4. Comparison of the NAG between MaSS and baselines on AudioMNIST. We suppress gender, accent, age, while preserve digit as annotated useful attribute, and preserve ID as if an unannotated attribute.

Method	Normalized Accuracy Gain				
	gender (↓)	accent (↓)	age (↓)	ID (↑)	digit (↑)
ALR	0.0000	0.0000	0.0056	0.7032	0.9994
GAP	0.0000	0.0000	0.0000	0.7036	0.9579
MSDA	0.0015	0.0013	0.0323	0.8428	0.9981
BDQ	0.0000	0.0007	0.0013	0.4038	0.9980
PPDAR	0.0000	0.0000	0.0000	0.7027	0.9983
MaSS	0.0000	0.0000	0.0000	0.8514	0.9983

Table 5. Comparison of the NAG for different configurations of MaSS on AudioMNIST. ✓ denotes that this attribute is suppressed, while all other attributes are preserved as annotated useful attributes.

Method	MaSS Suppressed Attributes					Normalized Accuracy Gain				
	gender	accent	age	ID	digit	gender	accent	age	ID	digit
MaSS	✓					0.0000	0.9342	0.9574	0.9632	0.9972
	✓	✓				0.0000	0.0000	0.9199	0.9372	0.9987
	✓	✓	✓			0.0000	0.0000	0.0000	0.8680	0.9964
	✓	✓	✓	✓		0.0000	0.0000	0.0000	0.0017	0.9953

We next conduct an ablation experiment of different configurations of suppressed and preserved attributes using MaSS on AudioMNIST. The configurations and their corresponding results are shown in Table 5. We can see that MaSS consistently achieves NAG $= 0$ for most of the suppressed attributes, alongside with high NAG for preserved attributes.

We also conducted ablation experiments of varying mutual information constraints m and n, as well as an experiment comparing our method and SPAct (Dave et al., 2022). The results and analyses can be found in Appendix E.2.

5.4. Evaluation on Facial Images

Finally, we apply MaSS to Adience, suppressing gender while treating age and activity as unannotated attributes that should be preserved. The results shown in Table 6 reveal that, among all methods with NAG $= 0$ for gender, MaSS accomplishes the highest NAG for the preserved attributes.

Additionally, we also empirically show that the transformed facial images can be accurately exploited by off-the-shelf pre-trained landmark detection model PIPNet (Jin et al., 2021). The NME (Normalized Mean Error) (Jin et al., 2021) of PIPNet between transformed Adience and original Adience is 3.30%, in comparison with the 3.94% NME of PIPNet between original WLFW dataset (Wu et al., 2018) and ground truth label. The comparable performance showed that transformed Adience dataset can be accurately exploited by pre-trained PIPNet.

Table 6. Comparison of the NAG between MaSS and baselines on Adience. We suppress gender, while preserve age, ID as if unannotated useful attributes.

Method	Normalized Accuracy Gain		
	gender (↓)	age (↑)	ID (↑)
ALR	0.0128	0.0023	0.0128
GAP	0.0000	0.4907	0.5616
MSDA	0.3114	0.7928	0.8461
BDQ	0.0026	0.0000	0.0075
PPDAR	0.0000	0.0000	0.0000
MaSS	0.0000	0.7418	0.7662

Visualized transformed images, together with additional results on suppressing age, and an ablation study on retraining the sensitive attribute inference model are shown in Appendix E.3.

Apart from the above mentioned 3 datasets, we additionally experiment MaSS on a tabular dataset (Marketing Campaign, 2023) . The experimental setup, results and analysis are shown in Appendix E.4, which similarly validate the generalizability and effectiveness of MaSS.

6. Conclusion

In this paper, we present MaSS, a generalizable and highly configurable data-driven learnable data transformation framework that is capable of suppressing sensitive/private information from data while preserving its utility. Compared to existing privacy protection techniques that have similar objectives, MaSS is superior by satisfying all 5 desired properties of SUIFT. We thoroughly evaluated MaSS on three datasets of different modalities, namely voice recordings, human activity motion sensor signals, and facial images, and obtained promising results that demonstrate MaSS' practical effectiveness under various tasks and configurations.

Impact Statement

We believe that there is no ethical concern or negative societal consequence related to this work. Our work benefits the protection of people's privacy in that it is proposed to suppress sensitive attributes in the datasets while preserving their potential utility for downstream tasks.

Disclaimer

This paper was prepared for informational purposes by the Global Technology Applied Research center of JPMorgan Chase & Co. This paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J. L., et al. A public domain dataset for human activity recognition using smartphones. In *European Symposium on Artificial Neural Networks, Computational Intelligence And Machine Learning*, volume 3, pp. 3, 2013.

Asoodeh, S., Diaz, M., Alajaji, F., and Linder, T. Estimation efficiency under privacy constraints. *IEEE Transactions on Information Theory*, 65(3):1512–1534, 2018.

Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., and Samek, W. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*, 2018.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 531–540, 10–15 Jul 2018.

Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M., Reeves, G., and Sapiro, G. Adversarially learned representations for information obfuscation and inference. In *International Conference on Machine Learning*, pp. 614–623. PMLR, 2019.

CCPA. https://oag.ca.gov/privacy/ccpa. Accessed: 2023-09-27.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Common Crawl. Common crawl - open repository of web crawl data. https://commoncrawl.org/, 2023.

Dave, I. R., Chen, C., and Shah, M. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20164–20173, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Edwards, H. and Storkey, A. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

Eidinger, E., Enbar, R., and Hassner, T. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014.

GDPR. https://gdpr-info.eu/. Accessed: 2023-09-27.

Hsu, H., Asoodeh, S., and Calmon, F. Obfuscation via information density estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 906–917. PMLR, 2020.

Hsu, H., Martinez, N., Bertran, M., Sapiro, G., and Calmon, F. P. A survey on statistical, information, and estimation—theoretic views on privacy. *IEEE BITS the Information Theory Magazine*, 1(1):45–56, 2021a.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021b.

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. Generative adversarial privacy. *arXiv preprint arXiv:1807.05306*, 2018.

Hukkelås, H., Mester, R., and Lindseth, F. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pp. 565–578. Springer, 2019.

Jin, H., Liao, S., and Shao, L. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129:3174–3194, 2021.

Kumawat, S. and Nagahara, H. Privacy-preserving action recognition via motion difference quantization. In *European Conference on Computer Vision*, pp. 518–534. Springer, 2022.

Liao, J., Kosut, O., Sankar, L., and du Pin Calmon, F. Tunable measures for information leakage and applications to privacy-utility tradeoffs. *IEEE Transactions on Information Theory*, 65(12):8043–8066, 2019.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Malekzadeh, M., Clegg, R. G., Cavallaro, A., and Haddadi, H. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*, pp. 49–58, 2019.

Marketing Campaign. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis, 2023. Accessed: 2023-12-20.

Maximov, M., Elezi, I., and Leal-Taixé, L. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5447–5456, 2020.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Sun, M., Wang, Q., and Liu, Z. Human action image generation with differential privacy. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020.

Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., and Zhou, Q. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2129–2138, 2018.

Wu, Z., Wang, H., Wang, Z., Jin, H., and Wang, Z. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2126–2139, 2020.

Zhang, X., Ji, S., and Wang, T. Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*, 2018.

Appendix

A. Proofs

A.1. Proof of Theorem 3.1

Proof. **Proof for Equation 2.** For the Markov Chain shown in Figure 2, for any $i \in 1 \dots M$, and $j \in 1 \dots N$, if both $I(X'; S_i) \leq m_i$ and $I(X'; U_j) \geq n_j$ hold, then we have

$$\begin{aligned} m_i + I(X; U_j|S_i) &\geq I(X'; S_i) + I(X; U_j|S_i) \\ &= I(X'; S_i) + I(X', X; U_j|S_i) \\ &= I(X'; U_j, S_i) - I(X'; U_j|S_i) + I(X', X; U_j|S_i) \\ &= I(X'; U_j, S_i) + I(X; U_j|X', S_i) \\ &= I(X'; U_j) + I(X'; S_i|U_j) + I(X; U_j|X', S_i) \\ &\geq I(X'; U_j) \\ &\geq n_j \end{aligned} \tag{16}$$

which proves the first inequation. Following Data Processing Inequality, we can also have

$$n_j \leq I(X'; U_j) \leq I(X; U_j) \tag{17}$$

which proves the second inequation. Finally, we also have

$$m_i \geq I(X'; S_i) \geq 0 \tag{18}$$

which proves the third inequation.

Proof for Equation 3. Under the assumption that U, S are fully determined given X ($P(S_i|X), P(U_j|X)$ are degenerate distributions), we can have

$$H(S_i|X) = 0, \qquad H(U_j|X) = 0 \tag{19}$$

for any $i \in 1 \dots M$, and $j \in 1 \dots N$. Since adding a condition can not increase the entropy, we can also have

$$0 \leq H(U_j|X, S_i) \leq H(U_j|X) = 0 \tag{20}$$

Therefore we have

$$H(U_j|X, S_i) = 0 \tag{21}$$

Inserting $H(U_j|X, S_i) = 0$ and $H(U_j|X) = 0$ into the inequations 2, we can further convert them to the inequations 3 as

$$\begin{aligned} n_j &\leq m_i + I(X; U_j|S_i) \\ &= m_i + H(U_j|S_i) - H(U_j|X, S_i) \\ &= m_i + H(U_j|S_i), \end{aligned} \tag{22}$$

and

$$\begin{aligned} n_j &\leq I(X; U_j) \\ &= H(U_j) - H(U_j|X) \\ &= H(U_j). \end{aligned} \tag{23}$$

Proof for Equation 4. For the Markov Chain shown in Figure 2, according to Data Processing Inequality, we have

$$I(X'; X) - I(X'; F) = I(X'; X|F) \geq 0 \tag{24}$$

Therefore, we have

$$I(X'; F) \leq I(X'; X) \tag{25}$$

We can also have

$$\begin{aligned} I(X';X) &= H(X') - H(X'|X) \\ &= H(X') - H(X'|X,S_i) \\ &\leq H(X') - H(X'|X,S_i) + H(X|X',S_i) \\ &= H(X') + H(X|S_i) - H(X'|S_i) \\ &= I(X';S_i) + H(X|S_i) \\ &\leq H(X|S_i) + m_i \end{aligned} \tag{26}$$

□

A.2. Proof and Analysis for the Estimation Error of $I(X';U_j)$ with Frozen Useful Attribute Inference Network

Let $P_{\psi_{j,0}}(U_j|X')$ and $I_{\psi_{j,0}}(X';U_j)$ denote the conditional distribution of U_j given X' and the mutual information between U_j and X' estimated with the **frozen** useful attribute inference network $\psi_{j,0}$. For the Markov Chain shown in Figure 2, $I_{\psi_{j,0}}(X';U_j)$ is calculated as

$$I_{\psi_{j,0}}(X';U_j) = \mathbb{E}_{P(X)P_\theta(X'|X)P(U_j|X)}[\log \frac{P_{\psi_{j,0}}(U_j|X')}{P(U_j)}] \tag{27}$$

Therefore, we can prove

$$\begin{aligned} I(X';U_j) - I_{\psi_{j,0}}(X';U_j) &= \mathbb{E}_{P(X)P_\theta(X'|X)P(U_j|X)}[\log \frac{P(U_j|X')}{P_{\psi_{j,0}}(U_j|X')}] \\ &= \mathbb{E}_{P_\theta(X')P(U_j|X')}[\log \frac{P(U_j|X')}{P_{\psi_{j,0}}(U_j|X')}] \\ &= KL(P(U_j|X')||P_{\psi_{j,0}}(U_j|X')) \end{aligned} \tag{28}$$

where $KL(P(U_j|X')||P_{\psi_{j,0}}(U_j|X'))$ can be large and even unbounded, because $\psi_{j,0}$ is trained to approximate $P(U_j|X)$ rather than $P(U_j|X')$. Therefore, this strategy is not adopted in our design.

B. Additional Descriptions of Related Works

In this section we present additional discussions on related works, especially on why they are categorized as not satisfying or partially satisfying certain properties of SUIFT. DeepPrivacy (Hukkelås et al., 2019) and CiaGAN (Maximov et al., 2020) are considered partially satisfying U and F because the CGAN based framework would prioritize the visual quality of generated samples (whether they are differentiable by discriminator) over the preservation of useful information (whether generated samples contain the same useful information as their original prototypes). Besides, they do not provide theoretical justification for the CGAN based framework. (Hsu et al., 2020) is considered partially satisfying F because it does not enforce the preservation of generic features explicitly. On the contrary it proposes to only obfuscate information-leaking features, while keep other features unaltered, which implicitly preserves the generic features. In addition, (Hsu et al., 2020) also does not consider the existence of useful annotated attributes. ALR (Bertran et al., 2019) proposes a rigorous information-theoretic framework for annotated attributes, but does not incorporate unannotated attributes in the discussion. PPDAR (Wu et al., 2020) and BDQ (Kumawat & Nagahara, 2022) also propose effective solutions for annotated attributes preservation or suppression. But they do not take unannotated attributes and a detailed theoretical foundation into consideration as well.

For the works that consider the management of unannotated attributes, ALFR (Edwards & Storkey, 2016) and LAFTR (Madras et al., 2018) are proposed in fairness literature, which are designed to release a compact representation for downstream tasks, and only suppress a binary sensitive attribute. GAP (Huang et al., 2018) does not consider the preservation of annotated useful attributes. And importantly, all of ALFR, LAFTR, GAP and MSDA (Malekzadeh et al., 2019) do not theoretically justify their design for unannotated attributes in the same framework proposed for annotated attributes. SPAct (Dave et al., 2022) does not consider the existence of annotated senstive attributes and does not theoretically justify its design.

C. Additional Description of Proposed Method

C.1. InfoNCE Contrastive Learning

InfoNCE contrastive learning loss (Oord et al., 2018) is a classical contrastive learning loss, which learns useful representations of data by making the representations of positive samples (similar or related samples) closer while pushing the representations of negative samples further apart from the anchor. The sampling strategy in our framework is as follows. Suppose we have $K+1$ samples $\{x_i\}_{i=1}^{K+1}$ in a mini-batch. We first pass them through the feature extractor $P_\eta(F|X)$ and data transformation module to sample a batch of $\{f_i\}_{i=1}^{K+1}$ and $\{x'_i\}_{i=1}^{K+1}$ respectively. Then suppose we choose the j-th feature f_j as the anchor. Then the corresponding x'_j would be designated as positive sample, and all other $x'_{i\neq j}$ are designated as negative samples. After sampling, we calculate the contrastive learning loss as Equation 14 in our paper. For training stability, in our implementation each of $K+1$ features in a batch is used as anchor once and then averaged.

An analogous InfoNCE contrastive learning loss L'_F is anchored in X' space, which is defined as

$$L'_F := \mathbb{E}_{x\sim P(X)}\mathbb{E}_{x'\sim P_\theta(X'|X)}\mathbb{E}_{f_p\sim P_\eta(F|X)}\mathbb{E}_{f_n^{(K)}\sim P_\eta(F)}[\log \frac{\mathcal{F}(f_p, x')}{\mathcal{F}(f_p, x') + \sum_{f_n\in f_n^{(K)}} \mathcal{F}(f_n, x')}] \tag{29}$$

where x', f_p, $f_n^{(K)}$ are the anchor, the positive sample, and the negative samples respectively.

Compared with ℓ_2 reconstruction loss, our contrastive learning loss is advantageous in that it does not presuppose the distributions of F, X', making it broadly applicable across various domains like images, language, and sensor signals. Moreover, its superior empirical effectiveness is demonstrated in our experiments.

D. Additional Experimental Setup

D.1. Datasets

We next introduce the datasets. The Adience dataset, consisting of 26580 facial images, was originally published to help study the recognition of age and gender. Each face image has 3 attributes: ID, gender and age. We filter out the IDs with only one image. For the rest of data points, we split them into training and evaluation set as 7:3, and ensure that for each ID there is at least one image in training set and one image in evaluation set. Data points used in our experiment contains 1042 different DataIDs, 8 age groups, and 2 gender classes. The images are resized to 80*80, converted to grayscale images, and normalzied to 0-1 in our experiments.

The AudioMNIST dataset contains audio recordings of spoken digits (0-9) in English from 60 speakers. The dataset contains 8 attributes, from which we used 5 most representative attributes for our experiments, namely gender, accent, age, ID, spoken digits, with 2, 16, 18, 60, 10 classes, respectively. There are 30,000 audio clips in total. We split the data into 24000, 6,000 for training and evaluation. The audio data are transformed to feature embeddings by HuBERT-B feature extractor and normalized to unit L2-norm.

The Motion Sense dataset contains the accelerometer and gyroscope data for human doing 6 daily activities. It contains 5 attributes, form which we used 3 most representative attributes for our experiments, namely gender, ID, and activity, with 2, 24, 6 classes respectively. Following (Malekzadeh et al., 2019), we did not use ”sit” and ”stand up” activity in experiments. We used the same split and data pre-processing method as (Malekzadeh et al., 2019), which resulted in 74324 segmented data points. Specifically, we used ”trail” split strategy as described in (Malekzadeh et al., 2019), and we only used the magnitude of gyroscope and accelerometer as input. Signals are normalized to 0-mean and 1-std, and then cut into 128-length clips.

D.2. Model Structures and Optimization

We elaborated the model structures and optimization methods used for our experiments in Table 7. For faster convergence and training stability, we design the ϕ, ψ, η models used in facial image experiments as a fixed FaceNet (Schroff et al., 2015) backbone followed by learnable 3-layer MLPs, and design the θ model of facial image experiment as U-Net (Ronneberger et al., 2015). For the same reason, we add residual structures from input of the first layer to the output of the second layer for 3-layer MLP θ models used in audio and human activity experiments.

Table 7. Model structures and optimization methods used for our experiments.

Experiment	Audio	Human activity	Facial image
Dataset	AudioMNIST	Motion Sense	Adience
# total data points	30000	74324	26580
Training-evaluation split	4:1	7:4	7:3
Optimizer	AdamW (Loshchilov & Hutter, 2019)		
Learning rate	0.0001		
Weight decay	0.001		
Learning rate scheduler	Cosine scheduler		
Epoch	2000	200	4000
θ model structure	3-layer MLP	3-layer MLP	U-Net
ϕ, ψ, η model structure	3-layer MLP	6-layer Convolutional NN	Fixed FaceNet backbone followed by learnable 3-layer MLP

Table 8. Comparison of the accuracy and NAG between MaSS, ablations and baselines on Motion Sense. We suppress gender, ID, while preserve activity as if unannotated useful attribute.

Method	Accuracy (Normalized Accuracy Gain)		
	gender (↓)	ID (↓)	activity (↑)
No suppression	0.9817 (1.0000)	0.9026 (1.0000)	0.9764 (1.0000)
Guessing	0.5699 (0.0000)	0.0533 (0.0000)	0.4663 (0.0000)
ALR	0.6040 (0.0828)	0.0900 (0.0432)	0.8593 (0.7704)
GAP	0.5721 (0.0053)	0.0800 (0.0314)	0.8937 (0.8379)
MSDA	0.5725 (0.0063)	0.1134 (0.0708)	0.8957 (0.8418)
BDQ	0.6184 (0.1178)	0.1054 (0.0613)	0.8451 (0.7426)
PPDAR	0.5698 (0.0000)	0.0498 (0.0000)	0.8189 (0.6912)
MaSS-NF	0.5699 (0.0000)	0.0508 (0.0000)	0.8374 (0.7275)
MaSS-ℓ_2	0.5734 (0.0085)	0.0754 (0.0260)	0.8823 (0.8156)
MaSS-MINE	0.7056 (0.3294)	0.1613 (0.1271)	0.8156 (0.6847)
MaSS	0.5686 (0.0000)	0.0555 (0.0026)	0.9242 (0.8977)

E. Additional Experiments Results

E.1. Evaluation on Human Activity Sensor Signals

In addition to the experimental results measured in NAG shown in Table 2, we also show the experimental results measured in accuracy below in Table 8.

We also conducted an experiment where we suppress gender, while preserve ID as annotated attribute, and preserve activity as unannotated attribute. We set the n for ID as 1.6, which meets the requirements of Equation 3. The results are shown in Table 9. We can observe that MaSS achieved lowest NAG on gender as well as comparable NAG on the other preserved attributes. This outcome stems from the fact the sensitive attribute gender is determined by ID, therefore when we suppress gender, the information retained for ID is inherently limited as Equation 3. MaSS is explicitly aware of this limit and is adjusted to preserve only limited amount of information for ID. In contrast other baselines can only heuristically trade-off between suppressing and preserving.

E.2. Evaluation on Voice Audio Dataset

In addition to the experimental results measured in NAG shown in Table 3 and Table 4, we also show the experimental results measured in accuracy below in Table 10 and Table 11 respectively.

Table 9. Comparison of the classification accuracy and NAG between MaSS and baselines on Motion Sense. We suppress gender, while preserve ID as annotated useful attribute, and preserve activity as if an unannotated attribute.

Method	Accuracy (Normalized Accuracy Gain)		
	gender (↓)	ID (↑)	activity (↑)
No suppression	0.9817 (1.0000)	0.9026 (1.0000)	0.9764 (1.0000)
Guessing	0.5699 (0.0000)	0.0533 (0.0000)	0.4663 (0.0000)
ALR	0.8258 (0.6214)	0.6147 (0.6610)	0.8966 (0.8436)
GAP	0.6599 (0.2186)	0.6628 (0.7176)	0.9378 (0.9243)
MSDA	0.6418 (0.1746)	0.6360 (0.6861)	0.9030 (0.8561)
BDQ	0.7092 (0.3383)	0.6583 (0.7124)	0.9269 (0.9030)
PPDAR	0.7830 (0.5175)	0.5680 (0.6060)	0.8867 (0.8242)
MaSS	0.5870 (0.0415)	0.5931 (0.6356)	0.9168 (0.8832)

Table 10. Comparison of the classification accuracy and NAG between MaSS, ablations and baselines on AudioMNIST. We suppress gender, accent, age, ID, while preserve digit as if an unannotated attribute.

Method	Accuracy (Normalized Accuracy Gain)				
	gender (↓)	accent (↓)	age (↓)	ID (↓)	digit (↑)
No suppression	0.9962 (1.0000)	0.9843 (1.0000)	0.9657 (1.0000)	0.9808 (1.0000)	0.9975 (1.0000)
Guessing	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0167 (0.0000)	0.1000 (0.0000)
ALR	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0171 (0.0004)	0.1930 (0.1036)
GAP	0.8000 (0.0000)	0.6828 (0.0000)	0.1663 (0.0000)	0.0438 (0.0281)	0.9513 (0.9485)
MSDA	0.8000 (0.0000)	0.6833 (0.0000)	0.1665 (0.0000)	0.0238 (0.0074)	0.9482 (0.9451)
BDQ	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0275 (0.0112)	0.5995 (0.5565)
PPDAR	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0182 (0.0016)	0.3548 (0.2839)
MaSS-NF	0.8000 (0.0000)	0.6833 (0.0001)	0.1658 (0.0000)	0.0152 (0.0000)	0.2657 (0.1846)
MaSS-ℓ_2	0.8002 (0.0008)	0.6833 (0.0001)	0.1683 (0.0020)	0.0462 (0.0306)	0.9542 (0.9517)
MaSS-MINE	0.8015 (0.0076)	0.6833 (0.0000)	0.1757 (0.0112)	0.0585 (0.0434)	0.5515 (0.5031)
MaSS	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0195 (0.0029)	0.9683 (0.9675)

We also compare our method with SPAct (Dave et al., 2022). Since SPAct does not consider preserving unannotated useful attributes. Therefore we compare it in a scenerio where we only have annotated attributes. We can observe that MaSS achieved slightly lower NAG on digit compared with SPAct, but significantly lower NAG on all sensitive attributes (up to 5%), which shows that our method may achieve a better trade-off between suppression and preservation.

We also conducted experiments to show the effect of varying the constraint on sensitive attributes suppression (m). We take gender, accent, age and ID as sensitive attributes and take digit as annotated useful attribute on the AudioMNIST dataset. We fix $m = 0$ for gender, accent and age and $n = 2.3$ for digit (its maximal value). Then we vary m for ID from 0 to 1.46 (its maximal value). The results are shown in Table 13. We can observe that as m increases, the constraint is gradually loosened, which results in the gradually increasing accuracy and NAG for ID.

Another experiment is to vary constraint n for digit on AudioMNIST, while suppress gender, accent, age, ID with fixed $m = 0$. The results are shown in Table 14. We can observe when n_{digit} is large enough, as n_{digit} increases, the constraint posed by annotated attribute preservation module is gradually taken into effect, which gradually turns digit from an unannotated useful attribute (protected by unannotated useful attribute preservation module) to an annotated useful attribute (protected mostly by the annotated useful attribute module), and consequently gradually increases the accuracy and NAG of digit.

Table 11. Comparison of the classification accuracy and NAG between MaSS and baselines on AudioMNIST. We suppress gender, accent, age, while preserve digit as annotated useful attribute, and preserve ID as if an unannotated attribute.

Method	Accuracy (Normalized Accuracy Gain)				
	gender (↓)	accent (↓)	age (↓)	ID (↑)	digit (↑)
No suppression	0.9962 (1.0000)	0.9843 (1.0000)	0.9657 (1.0000)	0.9808 (1.0000)	0.9975 (1.0000)
Guessing	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0167 (0.0000)	0.1000 (0.0000)
ALR	0.7995 (0.0000)	0.6832 (0.0000)	0.1712 (0.0056)	0.6947 (0.7032)	0.9970 (0.9994)
GAP	0.8000 (0.0000)	0.6828 (0.0000)	0.1663 (0.0000)	0.6950 (0.7036)	0.9597 (0.9579)
MSDA	0.8003 (0.0015)	0.6837 (0.0013)	0.1925 (0.0323)	0.8292 (0.8428)	0.9958 (0.9981)
BDQ	0.8000 (0.0000)	0.6835 (0.0007)	0.1677 (0.0013)	0.4060 (0.4038)	0.9957 (0.9980)
PPDAR	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.6942 (0.7027)	0.9960 (0.9983)
MaSS	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.8375 (0.8514)	0.9960 (0.9983)

Table 12. Comparison of the classification accuracy and NAG between MaSS and SPAct on AudioMNIST. We suppress gender, accent, age, id, while preserve digit as annotated useful attribute.

Method	Accuracy (Normalized Accuracy Gain)				
	gender (↓)	accent (↓)	age (↓)	ID (↓)	digit (↑)
No suppression	0.9962 (1.0000)	0.9843 (1.0000)	0.9657 (1.0000)	0.9808 (1.0000)	0.9975 (1.0000)
Guessing	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0167 (0.0000)	0.1000 (0.0000)
SPAct	0.8087 (0.0442)	0.6833 (0.0001)	0.1753 (0.0108)	0.0707 (0.0560)	0.9948 (0.9970)
MaSS	0.8000 (0.0000)	0.6833 (0.0000)	0.1662 (0.0000)	0.0183 (0.0017)	0.9933 (0.9953)

Table 13. Varying the suppression constraint m for ID on AudioMNIST. We suppress gender, accent, age, ID, while preserve digit as if an annotated useful attribute.

Method	m_{ID}	Accuracy (Normalized Accuracy Gain)				
		gender (↓)	accent (↓)	age (↓)	ID (↓)	digit (↑)
No suppression	-	0.9962 (1.0000)	0.9843 (1.0000)	0.9657 (1.0000)	0.9808 (1.0000)	0.9975 (1.0000)
Guessing	-	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0167 (0.0000)	0.1000 (0.0000)
MaSS	0.0	0.8000 (0.0000)	0.6833 (0.0000)	0.1662 (0.0000)	0.0183 (0.0017)	0.9933 (0.9953)
	0.3	0.8000 (0.0000)	0.6833 (0.0000)	0.1665 (0.0000)	0.0598 (0.0447)	0.9938 (0.9959)
	0.6	0.8000 (0.0000)	0.6833 (0.0000)	0.1668 (0.0002)	0.1120 (0.0988)	0.9940 (0.9961)
	0.9	0.8000 (0.0000)	0.6833 (0.0000)	0.1670 (0.0004)	0.1493 (0.1376)	0.9937 (0.9957)
	1.2	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.1963 (0.1863)	0.9928 (0.9948)
	1.46	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.2597 (0.2520)	0.9937 (0.9957)

E.3. Evaluation on Facial Images

In addition to the experimental results measured in NAG shown in Table 6, we also show the experimental results measured in accuracy below in Table 15.

In the next experiment we demonstrate the performance of MaSS on Adience with different attribute to suppress. We can observe from Table 16 that MaSS achieved 0 NAG for suppressed attributes as well as acceptable NAG for preserved unannotated attributes.

The visualization results for both original and transformed data in the Adience dataset are depicted in Figure 4. Observing the second row, we can see that the gender information has been effectively removed from the images. Similarly, the third row demonstrates the removal of age information from the images, highlighting the efficacy of our approach in suppressing

Table 14. Varying the preservation constraint n for digit on AudioMNIST. We suppress gender, accent, age, ID, while preserve digit as if an annotated useful attribute.

Method	n_{digit}	Accuracy (Normalized Accuracy Gain)				
		gender (↓)	accent (↓)	age (↓)	ID (↓)	digit (↑)
No suppression	-	0.9962 (1.0000)	0.9843 (1.0000)	0.9657 (1.0000)	0.9808 (1.0000)	0.9975 (1.0000)
Guessing	-	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0167 (0.0000)	0.1000 (0.0000)
MaSS	0.0	0.8000 (0.0000)	0.6833 (0.0000)	0.1667 (0.0000)	0.0192 (0.0026)	0.9685 (0.9677)
	1.8	0.8000 (0.0000)	0.6833 (0.0000)	0.1657 (0.0000)	0.0207 (0.0041)	0.9683 (0.9675)
	1.9	0.8000 (0.0000)	0.6833 (0.0000)	0.1658 (0.0000)	0.0178 (0.0012)	0.9725 (0.9721)
	2.0	0.8000 (0.0000)	0.6833 (0.0000)	0.1658 (0.0000)	0.0163 (0.0000)	0.9733 (0.9731)
	2.1	0.8000 (0.0000)	0.6833 (0.0000)	0.1642 (0.0000)	0.0182 (0.0015)	0.9823 (0.9831)
	2.2	0.8000 (0.0000)	0.6833 (0.0000)	0.1665 (0.0000)	0.0202 (0.0036)	0.9885 (0.9900)
	2.3	0.8000 (0.0000)	0.6833 (0.0000)	0.1662 (0.0000)	0.0183 (0.0017)	0.9933 (0.9953)

Table 15. Comparison of the classification accuracy and NAG between MaSS and baselines on Adience. We suppress gender, while preserve age, ID as if unannotated useful attributes.

Method	Accuracy (Normalized Accuracy Gain)		
	gender (↓)	age (↑)	ID (↑)
No suppression	0.9774 (1.0000)	0.9321 (1.0000)	0.9382 (1.0000)
Guessing	0.5240 (0.0000)	0.2892 (0.0000)	0.0284 (0.0000)
ALR	0.5298 (0.0128)	0.2907 (0.0023)	0.0400 (0.0128)
GAP	0.5240 (0.0000)	0.6047 (0.4907)	0.5393 (0.5616)
MSDA	0.6652 (0.3114)	0.7989 (0.7928)	0.7982 (0.8461)
BDQ	0.5252 (0.0026)	0.2892 (0.0000)	0.0352 (0.0075)
PPDAR	0.5231 (0.0000)	0.2892 (0.0000)	0.0284 (0.0000)
MaSS	0.5240 (0.0000)	0.7661 (0.7418)	0.7255 (0.7662)

Table 16. Comparison of the Accuracy and NAG for different configurations of MaSS on Adience. ✓ denotes that this attribute is suppressed, while all other attributes are preserved as unannotated useful attributes.

Method	MaSS Suppressed Attributes			Accuracy (Normalized Accuracy Gain)		
	gender	age	ID	gender	age	ID
No suppression				0.9774 (1.0000)	0.9321 (1.0000)	0.9382 (1.0000)
Guessing	✓	✓	✓	0.5240 (0.0000)	0.2892 (0.0000)	0.0284 (0.0000)
MaSS	✓			0.5240 (0.0000)	0.7661 (0.7418)	0.7255 (0.7662)
		✓		0.7985 (0.6054)	0.2892 (0.0000)	0.5005 (0.5189)

specific attributes.

Although we would not release the labels of sensitive attributes to the public, here we conducted an ablation experiment with the assumption that the attacker can access the ground truth labels of sensitive attributes as an oracle and retrain the discriminator on transformed data. The results are shown in Table 2. We can observe that, using MaSS, the accuracy of the retrained discriminator is higher than adversarial discriminator but is still significantly lower than the discriminator trained using original data.

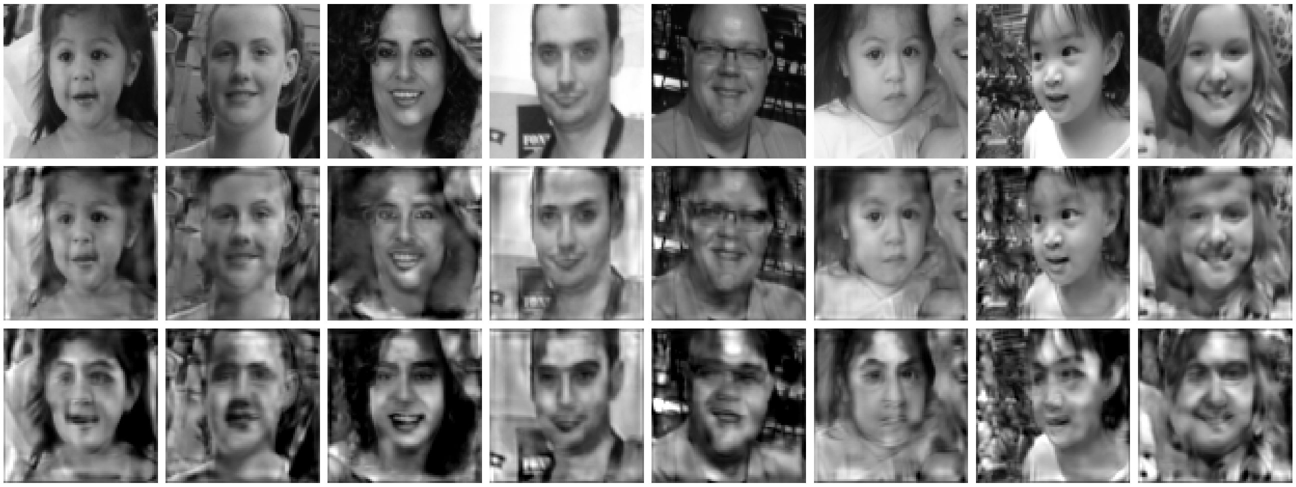

Figure 4. The visualization of the original data and transformed data in Adience dataset. The first row presents the original facial images, while the second and third rows show the transformed images with gender and age suppressed respectively. Other attributes are preserved as unannotated.

Table 17. Comparison of the accuracy and NAG between a trained-from-scratch discriminator and adversarial discriminator on the Adience dataset. We suppress gender, while preserve age, ID as if unannotated useful attributes.

Method	Accuracy (Normalized Accuracy Gain)
	gender (↓)
No suppression	0.9774 (1.0000)
Guessing	0.5240 (0.0000)
MaSS (discriminator retrained with oracle)	0.6029 (0.1740)
MaSS (adversarial discriminator)	0.5240 (0.0000)

E.4. Evaluation on Tabular Marketing Campaign Dataset

We further evaluate MaSS on the tabular Marketing Campaign (2023) dataset and compare the effectiveness of MaSS to MaSS-ℓ_2. We first convert the categorical attributes into one-hot vectors and normalize the continuous attributes by their ranges. Note that, the ℓ_2 reconstruction loss applied to one-hot vectors can be interpreted as a 0-1 loss. During training MaSS, we adopt Gumbel-Softmax for the categorical attributes to keep their differentiability and the flexibility to convert them back to the original value.

Tabular data is slightly different from other data types we experimented in the main paper, the utility and sensitive attributes (columns) are also a part of data X rather than separated attributes. Thus, we first left out the utility columns as separated attributes (U) and train the MaSS over remaining columns to generate a transformed data (X'). Then, we evaluate the classification accuracy of utility attributes (U) using the transformed X'; meanwhile, we separate the sensitive column out from the transformed data X' and then using the remaining columns to predict original sensitive columns. The results are shown in Table 18. We can observe that MaSS achieved both higher NAG for response and lower NAG for education compared with MaSS-ℓ_2, which further validate the generaizability and effectiveness of our framework.

Table 18. Comparison of the classification accuracy and NAG between MaSS and ablation on the Marketing Campaign dataset. We suppress education, while preserve response as annotated useful attribute.

Method	Accuracy (Normalized Accuracy Gain)	
	education (↓)	response (↑)
No suppression	0.5223 (1.0000)	0.8973 (1.0000)
Guessing	0.4732 (0.0000)	0.8504 (0.0000)
MaSS-ℓ_2	0.4933 (0.4094)	0.8728 (0.4776)
MaSS	0.4621 (0.0000)	0.9084 (1.2367)