Decouple-and-Sample: Protecting sensitive information in task agnostic data release

Abhishek Singh, Ethan Garza, Ayush Chopra,
Praneeth Vepakomma, Vivek Sharma, and Ramesh Raskar

Massachusetts Institute of Technology
abhi24@mit.edu

Abstract. We propose *sanitizer*, a framework for secure and task-agnostic data release. While releasing datasets continues to make a big impact in various applications of computer vision, its impact is mostly realized when data sharing is not inhibited by privacy concerns. We alleviate these concerns by sanitizing datasets in a two-stage process. First, we introduce a *global decoupling* stage for decomposing raw data into sensitive and non-sensitive latent representations. Secondly, we design a *local sampling* stage to synthetically generate sensitive information with differential privacy and merge it with non-sensitive latent features to create a useful representation while preserving the privacy. This newly formed latent information is a task-agnostic representation of the original dataset with anonymized sensitive information. While most algorithms sanitize data in a task-dependent manner, a few task-agnostic sanitization techniques sanitize data by censoring sensitive information. In this work, we show that a better privacy-utility trade-off is achieved if sensitive information can be synthesized privately. We validate sanitizer's effectiveness by outperforming state-of-the-art baselines on the existing tasks and demonstrating tasks that are not possible using existing techniques. Our code and benchmark is available at https://github.com/splitlearning/sanitizer

1 Introduction

Releasing datasets has resulted in methodological advancements in computer vision [14, 62] and machine learning (ML). However, the advancement is still limited by datasets that comply with modern privacy standards. The goal of this work is to alleviate privacy concerns in dataset release when it contains sensitive information. Since datasets are released independently of downstream tasks, we focus on the problem of protecting sensitive information in a task-agnostic manner. We refer to this problem as **sanitization**. Releasing sanitized datasets could galvanize the research community to make progress in the areas where raw data access is not feasible.

As a motivating example, consider a hospital with a dataset of face images where "ethnicity" and "age" of every face is a sensitive detail. The hospital is enabled to share the dataset with untrusted parties for several applications if we can *sanitize* all images in the dataset. To understand the benefit of sharing

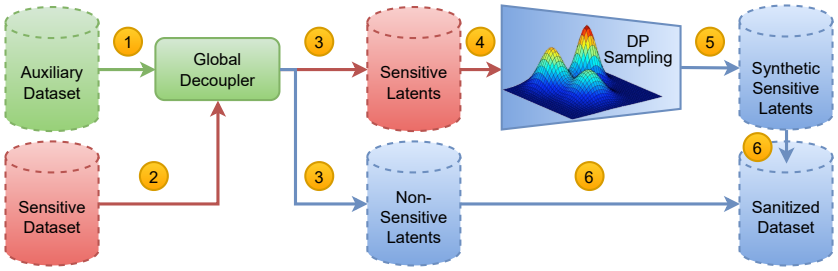


Fig. 1: Sanitizer pipeline First, we learn a latent model (global decoupler) of the data distribution using non-sensitive auxiliary dataset (in green). Next, we use the latent model to decouple sensitive (in red) and non-sensitive (in blue) information from the sensitive dataset to learn the distribution of sensitive latents. We synthetically generate sensitive latents by sampling from the distribution. Finally, we get the sanitized dataset by combining non-sensitive and synthetically generated sensitive latents.

the dataset, we list the following use-cases that also motivate our experiments in Section 4.

UC1: A crowd-sourcing company can build a facial recognition model for medical diagnostics [36, 61, 10] from the sanitized dataset. This model will be deployed on cloud, therefore the prediction will be performed over sanitized images.

UC2: A group of researchers can develop a model of capturing keypoints from face images. Unlike UC1, they want the model to predict over unsanitized images. Hence sanitized images should be photo-realistic to prevent a domain mismatch.

UC3: The hospital wants to share a sanitized dataset with a company to build an ML model to predict "age". Similar to UC2, the hospital would perform prediction on unsanitized images hence the sanitized dataset should be photo-realistic. However, unlike UC2, prediction attribute "age" is also a sensitive attribute requiring privacy.

Since there can be many such use-cases, it is impractical to assume that the hospital knows all use-cases in advance before releasing the dataset. Therefore, the goal of *Sanitizer* is to transform the dataset by anonymizing sensitive information without the knowledge of the downstream use-cases. In addition to learning ML models, being task agnostic allows *sanitizer* to do inference queries on sensitive datasets such as counting the number of faces with "smiling" attribute, or counting X-ray images with "lung cancer". Trivially cropping the sensitive parts from the image is not feasible because the pixels that reveal the sensitive information are present everywhere in a face image. Furthermore, unlike face images, identifying sensitive information visually may not be possible. For instance, several recent works [29, 30, 5, 4, 73] show ways in which sensitive demographics can be leaked leak from biomedical images using ML models.

Many works in sanitizing data have a different scope from the one considered here. Typically identity [17] of individuals is treated as sensitive information. While this notion protects privacy, we *only* focus on a specific set of sensitive attributes. For example - in UC1, it is acceptable to share face images as

long as "ethnicity" and "age" can be protected. In works that do consider specific sensitive attributes, their notion of utility is typically task-dependent as in [46, 44, 33]. Although *sanitizer* can be used for such problems due to its task-agnostic approach, not exploiting the knowledge about downstream tasks comes at the expense of a relatively lower utility. Existing works specifically in sanitization [23, 31] protect sensitive information by *censoring* it. Unlike censoring based approaches, our main idea is to share synthetically generated sensitive information. While the data receiver can not infer the original sensitive information, our approach allows them to learn from anonymized sensitive information.

To design *sanitizer*, we posit that sensitive data can be anonymized by replacing sensitive information with a synthetic one. However, for images, this synthetic replacement is not trivial to perform since the sensitive information is not localized in a region and sensitive attributes and non-sensitive attributes can share the same parts of data (ex. - race and gender). Therefore, we introduce a *latent model* that exclusively isolates sensitive information into a smaller subspace. We learn the latent model using publicly available datasets and then use the model to isolate the sensitive information from the sensitive dataset. Next, we learn a generative model of the isolated sensitive information and synthesize sensitive latents by sampling from the model. We merge these samples with the non-sensitive latent representation to obtain sanitized data. We visualize the whole pipeline in Figure 1.

Contributions: *First,* we introduce a joint optimization framework for isolating sensitive information from data. *Second*, we design a mechanism for anonymizing sensitive image datasets. *Third*, we empirically demonstrate various applications of sanitization to show the benefit of *sanitizer* over existing approaches, *Fourth*, we release a benchmark and dataset of sanitized representations obtained from baselines for rigorous attacks and defense evaluation in the future.

2 Problem Formulation

Terminology: Consider a data holder A with access to a dataset $\mathbf{D}^A = \{\mathbf{X}, \mathbf{Y}\}$ with N data points. Let $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ represents a pair of sample and *set* of labels $(\mathbf{x}, \mathbf{y})$ describing distinct attributes of $\mathbf{x}$. For instance, if $\mathbf{x}$ is a face image of an individual, the *set* $\mathbf{y}$ may include the age, gender, and ethnicity of the individual. For A, certain attributes in the label set $\mathbf{y}$ represent sensitive information (called $\mathbf{y}_S$) while others are non-sensitive ($\mathbf{y}_{NS}$) such that $\mathbf{y} = \{\mathbf{y}_S \cup \mathbf{y}_{NS}\}$. These sensitive attributes are A's secrets that prevent A from sharing $\mathbf{D}^A$. While A can release $(\mathbf{x}, \mathbf{y}_{NS})$, an attacker can guess $\mathbf{y}_S$ using $\mathbf{x}$ by exploiting correlation between $\mathbf{x}$ and $\mathbf{y}_S$. Hence, to release $\mathbf{D}^A$ for arbitrary downstream tasks, sanitization techniques transform every sample in $\mathbf{D}^A$ from $(\mathbf{x}, \mathbf{y})$ to $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ resulting in sanitized dataset $\tilde{\mathbf{D}}^A$ that can be shared with untrusted parties. The key challenge is anonymizing sensitive information while maximally retaining the utility. We assume that an auxiliary dataset $\mathbf{D}^{aux}$ from the same distribution as $\mathbf{D}^A$ is publicly accessible to all parties.

Sanitizer Overview: We perform sanitization in a two stage process: i) *Global decoupling* and ii) *Local sampling*. *Global decoupling* stage learns a latent model (parameterized by θ, ϕ) of data using $\mathbf{D}^{aux}$ for decoupling raw data ($\mathbf{x}$) into sensitive ($\mathbf{z}_S$) and non-sensitive ($\mathbf{z}_{NS}$) latents. We assume that the auxiliary dataset and the sensitive dataset come from the same distribution $p(\mathbf{x}, \mathbf{y})$. This stage does not require access to A's dataset ($\mathbf{D}^A$) and hence can be performed independently making this stage *global* since the same model can be utilized by different sensitive-data owners. We discuss the design of the *global decoupler* in Section 3.1. *Local sampling* stage learns a generative model ($f(\psi, \cdot)$) of sensitive latents in $\mathbf{D}^A$. We obtain the sensitive latents in $\mathbf{D}^A$ using the *global decoupler* from the first stage. Finally, we obtain the sanitized dataset by merging every non-sensitive latent with independently sampled sensitive-latent. We discuss the *local sampling* stage in Section 3.2.

Threat Model: We assume that the *untrusted* data-receiver can act as an attacker by utilizing auxiliary dataset $\mathbf{D}^{aux}$, parameters of *global decoupler* (θ, ϕ) and *local sampler* $(f(\psi, \cdot))$ as a side information. The side information allows the attacker to generate a mapping between $\mathbf{D}^{aux}$ and its sanitized version $\tilde{\mathbf{D}}^{aux}$. The attacker's goal is to recover A's sensitive attribute $\mathbf{y}_S$ from the sanitized dataset $\tilde{\mathbf{D}}^A$. Since $\mathbf{D}^{aux}$ and $\mathbf{D}^A$ come from the same distribution, the attacker can model the problem of inferring the sensitive attributes as an ML problem. By learning a mapping between sanitized samples ($\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}^{aux}$, $\tilde{\mathbf{y}} \in \tilde{\mathbf{Y}}^{aux}$) and sensitive attributes ($\mathbf{y}_S \in \mathbf{Y}_S^{aux}$) using $\mathbf{D}^{aux}$, the attacker can attempt inferring sensitive attributes from $\tilde{\mathbf{D}}^A$. This threat model is different from differential privacy [17] which seeks to protect identifiability.

Defining information leakage: Information leakage for sanitization has been typically defined statistically [15, 37, 57]; however, estimating these statistics requires estimation of probability distributions making it intractable for non-linear queries over higher-dimensional datasets (images). Alternatively, leakage can be quantified by simulating an attacker's performance by making some assumptions. The goal of sanitization is to minimize the distinguishability of the original sensitive attributes $\mathbf{y}_S$ from other possible values (domain($\mathbf{y}_S$)) sensitive attributes can take. For example, if "ethnicity" is a sensitive attribute then, informally, leakage is the likelihood of the attacker's correct *estimate* about the race of the sanitized face image. Formally, this can be modeled by a change in belief over the sensitive attribute before (prior $p(\mathbf{y}_S)$) and after (posterior $p(\mathbf{y}_S|\tilde{\mathbf{x}})$) observing the sanitized sample $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. A similar notion is formalized in the pufferfish framework [26] where the quantity $\frac{p(\mathbf{y}_{s_j}|\tilde{\mathbf{x}},\theta)}{p(y_{s_i}|\tilde{\mathbf{x}},\theta)} / \frac{p(y_{s_j})}{p(y_{s_i})}$ for all possible sets of secrets (y_{s_i}, y_{s_j}) and for all possible priors on data θ is bounded by e^ϵ where ϵ is a privacy parameter. Note that satisfying this definition requires modeling various possible data evolution and attacker scenarios. We focus only on a single type of attacker described in the threat model and therefore use a data-driven approach to quantify leakage. This data-driven adversary learns the joint distribution $p(\tilde{\mathbf{X}}, \mathbf{Y}_S)$ using the side information. Finally, the leakage of the sanitized datasets is evaluated as the difference between the accuracy of the adversary to correctly estimate the sensitive information $p(\mathbf{y}_S|\tilde{\mathbf{x}}, \theta)$ and the

estimation of an uninformed adversary $p(\mathbf{y}_S|\theta)$. We note that existing works in sanitization [31, 23] use the same criterion to evaluate information leakage.

Desiderata: In both stages of *sanitizer*, we have two desirable properties corresponding to privacy and utility; in total, we get four desirable properties that we elaborate on now. The first stage is *global decoupling* where we learn to separate a sample $\mathbf{x}$ into sensitive and non-sensitive latent $\mathbf{z}_S$ and $\mathbf{z}_{NS}$ respectively. Therefore, the desirable property **P1** requires $(\mathbf{z}_S, \mathbf{z}_{NS})$ to be independent. In other words, **P1** requires non-sensitive latent $\mathbf{z}_{NS}$ does not leak information about $\mathbf{z}_S$. This property can be achieved trivially by sharing all zeroes therefore to enforce utility, we desire property **P2** that requires $p(\mathbf{x}|\mathbf{z}_S, \mathbf{z}_{NS})$ to be maximum. Property **P2** requires $(\mathbf{z}_S, \mathbf{z}_{NS})$ to be useful enough for describing the original sample $\mathbf{x}$. This completes the privacy-utility desiderata for the *global decoupling* stage. The *local sampling* stage focuses only on transforming the sensitive latent, therefore, both privacy and utility desideratum only focus on $\mathbf{z}_{NS}$. For the privacy desiderata **P3** in this stage, we require that sanitized sensitive latents $\tilde{\mathbf{z}}_S$ and $\tilde{\mathbf{z}}'_S$ obtained from $\mathbf{x}, \mathbf{x}' \in \mathbf{X}$ respectively are indistinguishable from each other. **P3** enforces that identifying original data sample based only on the sensitive information should not be possible, i.e. $p(\tilde{\mathbf{z}}_S \sim f_\psi(\mathbf{z}_S)) = p(\tilde{\mathbf{z}}_S \sim f_\psi(\mathbf{z}'_S))$. For the example of a face image with sensitive "ethnicity", **P3** requires synthetically generated $\tilde{\mathbf{z}}_S$ should be independent of the original "ethnicity" of the sample. We can trivially solve **P3** by sharing only zeroes, therefore to ensure utility, we introduce property **P4** that requires the *distribution* of original sensitive and synthetic sensitive latents to be the same. Specifically, the property implies $p(\mathbf{z}_S) = p(\tilde{\mathbf{z}}_S)$. Next, we model these desiderata to design our technique.

3 Method

We sanitize a sensitive dataset in a two-stage process. The first stage is *global decoupling* where we learn a latent model of the data distribution using auxiliary dataset $\mathbf{D}^{aux}$. Our goal is to learn a latent model of data that maximally *decorrelates* $\mathbf{z}_S$ and $\mathbf{z}_{NS}$ for every sample $\mathbf{x}$ (P1) and *preserves* all details of the sample in $\mathbf{z}$ (P2). We achieve this goal by designing *global-decoupler* in Section 3.1. The second stage is *local sampling* where we learn the distribution of the sensitive portion $\mathbf{Z}_S$ of our sensitive dataset $\mathbf{D}^A$. Our goal is to sample from the distribution $\tilde{\mathbf{Z}}_S$ such that estimating original sensitive attribute $\mathbf{Y}_S$ is not

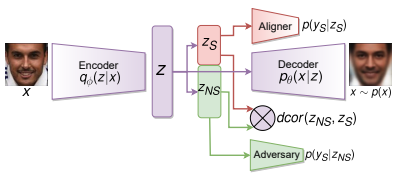


Fig. 2: Architecture for the proposed global-decoupler. The encoder samples $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}))$ partitioned into $(\mathbf{z}_S, \mathbf{z}_{NS})$. Aligner encourages $\mathbf{z}_S$ to carry information relevant to $\mathbf{y}_S$. We use adversary to reduce information between $\mathbf{z}_{NS}$ and $\mathbf{y}_S$. Finally, we minimize distance correlation between $\mathbf{z}_S$ and $\mathbf{z}_{NS}$.

feasible (P3) and the distribution of $p(\mathbf{Z}_S)$ and $p(\tilde{\mathbf{Z}}_S)$ is similar (P4). We achieve this goal by designing *DP-sampling* mechanism in Section 3.2. We summarize the overall pipeline in Figure 1.

3.1 Global decoupling for isolating sensitive information

Our goal is to design a latent model of the data distribution $p(\mathbf{x}, \mathbf{y})$ such that $\mathbf{x}$ can be decoupled into latents $\mathbf{z}_S$ and $\mathbf{z}_{NS}$. We call this latent model as *global decoupler*. We design it by integrating four components - i) generative model, ii) aligner, iii) decorrelator and iv) adversarial training. We describe the architecture of *global-decoupler* in Figure 2. Generative models ([18, 28, 52]) are being increasingly used to perform such latent modeling . Specifically, we build upon VAE [28] since they provide the flexibility of modeling data with constraints on the probability density of the latent space $p(\mathbf{z})$. Given a dataset $\mathbf{X}$, VAEs [28, 53] model the distribution of samples $p(\mathbf{x}), \forall \mathbf{x} \in \mathbf{X}$ by learning parameters ϕ of approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and θ for the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$. β-VAE [21] improves the disentanglement between the components of $\mathbf{z}$ sampled from $q_\phi(\mathbf{z}|\mathbf{x})$ by regularizing the KL divergence between the prior $p(\mathbf{z})$ and approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$. To improve disentanglement between every $\mathbf{z}_i$, existing works such as Factor-VAE[27] and TCVAE[9] regularize the total correlation of $q(\mathbf{z})$ measured by $\mathsf{KL}(q(\mathbf{z})||\prod_{i=1}^{m} q(\mathbf{z}_i))$ where KL refers to the KL divergence and m is the total number of components of $\mathbf{z}$. However, high degree of disentanglement between *every component* can hinder the reconstruction quality [21]. Therefore, instead of disentangling every pair of $(\mathbf{z}_i, \mathbf{z}_j)$, we propose a new regularized *global-decoupler* to focus on the disentanglement of $(\mathbf{z}_S, \mathbf{z}_{NS})$ instead.

A key characteristic of VAE is that the decoupled latent representations are unordered. Hence, there is no explicit control on which dimensions encode what semantic attributes. This is a challenge for our work that ideally requires that representations encoding the sensitive attributes be contiguous for decoupling. Intuitively, we decouple the vector $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ into $\mathbf{z}_S$ and $\mathbf{z}_{NS}$ with additional regularization constraints that encourage independence between $\mathbf{z}_S$ and $\mathbf{z}_{NS}$. Formally, we reformulate the original VAE objective with an *aligner* $g_u(\cdot)$ parameterized by u to estimate $\mathbf{y}_S$ from $\mathbf{z}_S$, the intuition is that the *aligner*'s gradient flow will encourage $q_\phi(\cdot)$ to maximize relevant information between $\mathbf{y}_S$ and $\mathbf{z}_S$. Since all latents are known to be correlated with each other to a certain extent, we need to prevent leakage of $\mathbf{y}_S$ in $\mathbf{z}_{NS} \sim q(\mathbf{z}_{NS}|\mathbf{x})$. Unlike FactorVAE [27] or TCVAE [9] that regularize the total correlation disentangling each dimension, we propose to regularize correlation between sensitive($q(\mathbf{z}_S)$) and non-sesntive($q(\mathbf{z}_{NS})$) latents. We re-formulate the objective for $q_\phi(\cdot)$ to minimize distance correlation [63] between $q(\mathbf{z}_S)$ and $q(\mathbf{z}_{NS})$. To motivate the use of distance correlation, we note that directly estimating probability density is intractable for high dimensional representations, various measures such as HSIC [19], MMD [6] and distance correlation [67] are used. Distance correlation between n samples of two vectors $\mathbf{x}$ and $\mathbf{y}$ can be obtained as following:

$$dcorr(\mathbf{x}, \mathbf{y}) = \frac{dcov(\mathbf{x}, \mathbf{y})}{\sqrt{dcov(\mathbf{x}, \mathbf{x}) * dcov(\mathbf{y}, \mathbf{y})}}$$

where $dcov()$ is the sample distance covariance analogue of covariance defined as $dcov(\mathbf{x}, \mathbf{y}) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \hat{\mathbf{x}}_{j,k} \hat{\mathbf{y}}_{j,k}$. Here $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are obtained by computing double centered euclidean distance matrices of $\mathbf{x}$ and $\mathbf{y}$. In particular, we use distance correlation ($dcorr$) because it can measure nonlinear correlations between samples from random variables of arbitrary dimensions ($\mathbf{z}_S$ and $\mathbf{z}_{NS}$ can have different dimensionality), allows for efficient gradient computation and does not require any kernel selection or parameter tuning, unlike HSIC and MMD. We do note that $dcorr$ is measured as a sample statistic and hence larger sample size is desirable for the unbiased sample statistic to represent the population notion of the distance correlation. To prevent information leakage of $\mathbf{y}_S$ from $\mathbf{z}_S$, we use a proxy attacker network $h_v(\cdot)$ that is trained adversarially to learn parameters v which constrains $\mathbf{z}_{NS}$ to not carry information relevant to $\mathbf{y}_S$. The final objective can be summarized as:

$$L_1(\theta, \phi, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[logp_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{1}$$

$$L_2(\phi, u) = \ell_1(g_u(\mathbf{z}_i \sim q_\phi(\mathbf{x})|_{i \leq k}), \mathbf{y}_S) \tag{2}$$

$$L_3(\phi) = dcorr(\mathbf{z}_i \sim q_\phi(\mathbf{x})|_{i \leq k}, \mathbf{z}_i \sim q_\phi(\mathbf{x})|_{k < i \leq m}) \tag{3}$$

$$L_4(\phi, v) = \ell_2(h_v(\mathbf{z}_i \sim q_\phi(\mathbf{x})|_{k < i \leq m}), \mathbf{y}_S) \tag{4}$$

Here k and m are the dimensionalities of vectors $\mathbf{z}_S$ and $\mathbf{z}$, respectively. L_1 is the β-VAE [21] formulation of VAE's evidence lower bound where the parameter β encourages disentanglement between every component of $\mathbf{z}$. Increasing β favors the property P1 (by encouraging independence) but hurts the property P2 (by reducing reconstruction). L_2 is the objective for training the parameters of the aligner model. However, L_1 does not prevent $\mathbf{z}_{NS}$ from leaking information about $\mathbf{y}_S$. Hence, we optimize L_4 adversarially to prevent information leakage. Finally, we minimize distance correlation between $\mathbf{z}_S$ and $\mathbf{z}_{NS}$ to prevent $\mathbf{y}_{NS}$ from encoding information about $\mathbf{z}_S$ and encourage decoupling $\mathbf{z}_S$ and $\mathbf{z}_{NS}$. Jointly optimizing L_1, L_2, L_3 and L_4 helps achieve properties P1 and P2. We validate each component's benefit via ablation studies in Section 5.

ℓ_1, ℓ_2 can be cross-entropy or ℓ_p-norm (often $p = 2$) depending upon $\mathbf{y}_S$. The parameters ϕ, θ, u, v are trained jointly with following objective:

$$\min_{\theta,\phi,u} \alpha_1 L_1(\theta, \phi, \beta) + \alpha_2 L_2(\phi, u) + \alpha_3 L_3(\phi) - \alpha_4 \min_{v} L_4(\phi, v) \tag{5}$$

where $\beta, \alpha_1, \alpha_2, \alpha_3, \alpha_4$ are scalar hyper-parameters that yield a trade-off between the privacy (property P1) and utility (property P2). We reiterate that this stage only accesses auxiliary dataset $\mathbf{D}^{aux}$ for the training and evaluation. Hence the parameters of the global-decoupler do not leak any sensitive information.

3.2 Local sampling for synthesizing sensitive latents

In this stage, we design the *DP-sampling* mechanism to sanitize every sample in the sensitive dataset. A sanitized sample $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ is obtained from $(x, y) \in \mathbf{D}^A$

by extracting the sensitive and non-sensitive latents $(\mathbf{z}_S, \mathbf{z}_{NS})$ from the global-decoupler and replacing the sensitive latent with a synthetic one $(\tilde{\mathbf{z}}_S)$. To satisfy the privacy desiderata of our second stage P3, $\tilde{\mathbf{z}}_S$ is sampled independently of $\mathbf{z}_S$. In contrast to prior works [22, 32, 34] that focus on censoring sensitive attributes, a key benefit of our mechanism is in anonymizing the sensitive information for individual data points while enabling downstream tasks that may benefit from joint distribution $p(\mathbf{z}_S, \mathbf{y}_S)$. Motivated by the use-case UC3, we demonstrate in Section 4 under experiment E5 on how to train an ML model on a sanitized dataset that predicts the sensitive attributes of unsanitized images. To motivate our DP-sampling mechanism, we first discuss a trivial suppression-based mechanism and a naive DP mechanism.

a) Suppression: The key is to explicitly remove sensitive information by replacing $\mathbf{z}_S$ with a zero vector (i.e. $\tilde{\mathbf{z}}_S$ is a zero vector). While this approach censors the sensitive latent, it is not possible to learn the distribution $p(\mathbf{x}, \mathbf{y}_S)$ under this mechanism resulting in a violation of our desired property P4. Therefore, we sanitize the sensitive information $\mathbf{z}_S$ using DP mechanisms.

b) DP-Obfuscation: The key idea is to add privacy calibrated noise to $\mathbf{z}_S$. This calibration can be formalized using DP where a mechanism $\mathcal{M}$ is ϵ-differentially private [17] if for every neighboring datasets X, X' and every output set $S \subset Range(\mathcal{M})$, the following inequality holds: $\mathbb{P}(\mathcal{M}(X) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(X') \in S)$. Here, we use the laplace mechanism [15] that adds noise sampled from a laplace distribution with variance as the ℓ_1-sensitivity of the query q. In the context of our work, the ℓ_1 sensitivity is defined as an identity function. Hence, to bound the sensitivity, we fix a pre-defined range $[a, b] \in \mathbb{R}$ in which $\mathbf{z}_S$ can lie, giving us sensitivity as $||a - b||_1$. Any data sample $\mathbf{x}$ that results in $\mathbf{z}_S \notin [a, b]$ is truncated to the closest vector in the range. To summarize, the mechanism can be described as $f(\mathbf{z}_S) = \mathbf{z}_S + \Delta$ where Δ is sampled from a laplace distribution, i.e. $\Delta \sim Lap(0, \frac{||a-b||_1}{\epsilon})$. While this approach allows us to have a trade-off between the privacy property P3 and utility property P4 by controlling ϵ, we get a sub-optimal trade-off due to two reasons: i) truncation step discards the values outside the range $[a, b]$ ii) the noise is added independently to every component of $\mathbf{z}_S$ therefore throwing away the structure present in the distribution of $\mathbf{z}_S$ as shown in Figure 3. This motivates developing a more utility conducive mechanism that can utilize structure in the latent space of $\mathbf{z}_S$.

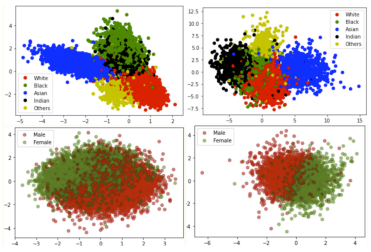


Fig. 3: Latent space visualization of $\mathbf{Z}_S$ by plotting its two components with the color of their sensitive attribute. We keep sensitive attributes as "race" and "gender" for the plots in the first and second row respectively using UTKFace [76].

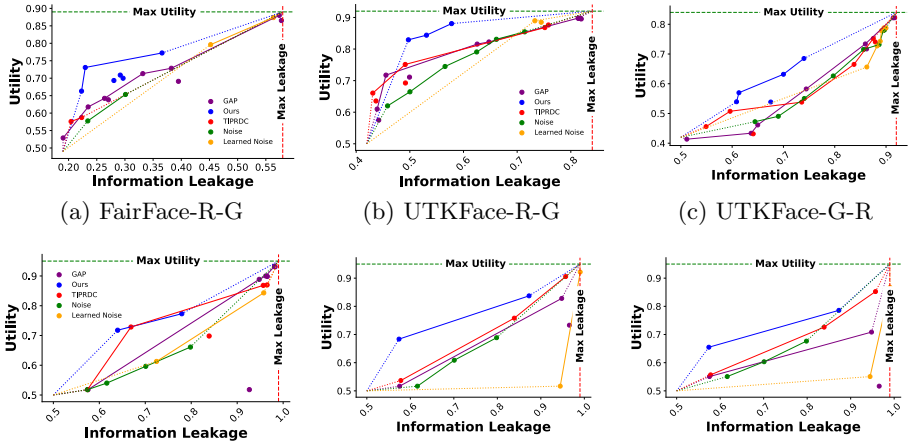
c) DP-Sampling: Our goal is to utilize the structure present in the distribution of sensitive latents $p(\mathbf{z}_S)$ as shown in Figure 3. We can observe the points

to be clustered around their respective sensitive attribute. Therefore, *instead of adding uniform and independent noise, we propose to learn the distribution of* $\mathbf{z}_S$ *and sample from it.* Since the data is low dimensional, learning a Gaussian mixture model [39] suffices to model the distribution. However, sampling from the mixture model could leak sensitive information since our threat model considers the parameters of the sampling model are accessible to the attacker. Therefore, we learn the covariance matrix of $\mathbf{Z}_S$ and perturb it in a differentially private manner before sampling from it. If the learned model of the data satisfies DP then due to the post-processing invariance property [17] of DP, the samples obtained from this model would also satisfy DP. We learn the GMM model (parameterized by ψ) using the sensitive dataset $(\mathbf{Z}_S, \mathbf{Y}_S)$. Such a sampling scheme also provides the benefit of sampling labeled pairs $(\tilde{\mathbf{Z}}_S, \tilde{\mathbf{Y}}_S)$ which is required for performing supervised learning. We utilize RONGauss mechanism [8] to learn a differentially private covariance matrix of the sensitive latent dataset $\mathbf{Z}_S$. For each $\mathbf{z}_S \in \mathbf{Z}_S$, the mechanism performs random orthonormal projection to a lower-dimensional. We learn the mean and covariance for each category in a differentially private manner. Low dimensional projection improves utility by reducing the perturbation required for the same amount of privacy. Finally, we obtain synthetic sample $\tilde{\mathbf{z}}_S$ by sampling from the Gaussian model and reprojecting it back to the original dimensionality of $\mathbf{z}_S$. Formally, this can be written as $\tilde{\mathbf{z}}_S, \tilde{\mathbf{y}}_S \sim f_\psi(\tilde{\mathbf{z}}_S, \tilde{\mathbf{y}}_S)$, here ψ is learned using original sensitive dataset $\mathbf{Z}_S$. We note that, unlike standard GMM, here we use only a single mode from the GMM for every unique class. This is a more accurate description of the data since every sample $\mathbf{z}_S$ is uniquely associated with a single sensitive attribute $\mathbf{y}_S$. While the RONGauss mechanism learns and samples the whole data space for providing a uniform privacy guarantee, we only sample sensitive latents $\mathbf{z}_S$ instead of $\mathbf{z}$ since the goal is to protect sensitive attributes and not uniform privacy. We note that this DP-sampling does not give a uniform privacy guarantee on $\tilde{\mathbf{Z}}_S$ since information about sensitive attributes can leak from $\mathbf{Z}_{NS}$ too. We developed global-decoupler to address this specific issue. Our proposed sampling scheme can be extended to synthetic data release [64] by treating every component in $\mathbf{z}$ as sensitive (i.e. $k = m$) and presents interesting future work.

Remark: In this section, we presented our two-stage sanitization process. The main advantage of separating *sanitizer* as a two-stage process is that developing *global decoupling* is a one-time procedure and can be performed by a third party that distributes the trained model to different data owners A's which can apply sanitizing mechanisms individually. This modular process is an efficient way to release sensitive datasets if there are multiple A's involved. Furthermore, we believe that future works can improve either of the two stages independently.

4 Experiments

In this Section, we compare *sanitizer* with different baselines under multiple experimental setups. Each experimental setup is focused on simulating a unique use case. For all experiments, we use CelebA [35], UTKFace [76] and Fairface [25].



(a) FairFace-R-G (b) UTKFace-R-G (c) UTKFace-G-R

(d) CelebA-G-M (e) CelebA-G-S (f) CelebA-G-C

Fig. 4: E1: Privacy-utility trade-off evaluation on different datasets: We plot sensitive information leakage as a proxy for privacy and one of the task attributes as a measure of utility for the sanitized dataset. Each point in this plot corresponds to training a *sanitizer* model and then evaluating its performance by training the adversary model and utility model on the sanitized dataset. *Sanitizer* performs better than all existing methods on all three datasets. Solid line represents the pareto-optimal curve for different methods. The dotted lines are extrapolation towards lowest and highest utility and leakage that is achievable trivially.

We split the dataset into $\mathbf{D}^{aux}$ for the first stage and $\mathbf{D}^{A}$ for the second stage. First, we train all techniques using $\mathbf{D}^{aux}$ and obtain sanitized dataset. Then we perform leakage assessment by training an adversary to learn a mapping between sanitized dataset and sensitive information. We discuss the rationale for such an adversary in Section 2. Finally, we evaluate the utility of the sanitized data based upon the experimental setup. Since our goal is task-agnostic data release, the *utility attribute is only used* after the sanitized dataset is released.

4.1 Baselines and Evaluation

Baselines: We compare against state-of-the-art visual sanitization techniques GAP [22] and TIPRDC [32], and introduce new baselines for exhaustive comparison. *GAP [22]:* is trained adversarially to maximize loss for a proxy adversary trying to infer sensitive attributes on the sanitized images. We replace the architecture proposed in the original paper with CNN architecture used in *sanitizer* to improve their results for higher dimensional image datasets. ii) *Learned Noise:* is built upon the TCNND architecture described in GAP [22] where a small dimensional noise is fed to a decoder that sanitizes the image by adding the noise vector. iii) *TIPRDC [32]:* is used as a baseline without any modification. iv) *Noise:* baseline sanitizes data by adding Gaussian noise in the pixel space; which is equivalent to the DP baseline used in TIPRDC [31].

Evaluation Metrics: We evaluate different techniques by comparing the *privacy-utility trade-off*. Here the utility is measured by the data receiver's test

Method	Fairface-R ↑	CelebA ↑	UTKFace-R ↑	UTKFace-G ↑
TIPRDC [32]	0.441	0.465	0.453	0.443
GAP [22]	0.447	0.442	0.450	0.434
Noise	0.438	0.422	0.435	0.431
Adversarial Noise [22]*	0.432	0.422	0.420	0.439
Ours	**0.476**	**0.483**	**0.476**	**0.487**

Table 1: E1: Privacy-Utility comparison: We report area under the curve (higher is better) to compare the privacy-utility trade-offs between *sanitizer* and baselines. UTKFace-R and UTKFace-G refer to the setup where *race* and *gender* is the sensitive attribute. Our method outperforms all baselines in all experiments.

accuracy on the downstream task using sanitized dataset. For measuring privacy loss, we use the technique described in Section 2. Specifically, we quantify information leakage from the dataset by comparing the performance of an adversary inferring sensitive information from the sanitized dataset. We simulate a strong adversary that dynamically adapts to a sanitization scheme. This adaptation is modeled by a pretrained adversary model that is finetuned on the *sanitized* dataset and then evaluated on the sanitized test set. This privacy loss acts as a lower bound on the worst-case privacy loss. Since inferring sensitive attribute is similar to learn an optimal classifier, the difficulty in giving upper bound on the privacy loss is similar to upper bounding generalization error in ML models. Hence, our evaluation uses a similar approach of using test set accuracy. Inspired by [54] we quantify privacy-utility trade-offs curves by different techniques using area under the pareto-optimal curve (AuC). Higher AuC value denotes a better privacy-utility trade-off. Fore more experimental details and results, we refer the reader to supplementary material.

4.2 Experimental Setup and Results

Experiment E1 *Multi-category sensitive and Binary utility*: We test the use-case **UC1** by evaluating the privacy-utility trade-off on a task where sensitive information is multi-category "race" (fine-grained) and downstream utility task is "gender" (coarse). Intuitively, we should get a good trade-off from all techniques that can share coarse-grained data while obfuscating fine-grained sensitive detail.

Experiment E2 *Binary sensitive and Multi-category utility*: We use the setup as E1 but use "race" (coarse) as sensitive attribute and "gender" (fine-grained) as utility attribute. Intuitively, we expect degradation in an overall trade-off in comparison to E1. We perform the experiments on UTKFace [76] dataset and call this configuration *UTKFace-G*.

Experiment E3 *Single sensitive and Multiple utility*: We use the same setup as E1 but evaluate multiple utility tasks. We use CelebA [35] with sensitive attribute as "gender" and utility as "mouth open", "smiling" and "high cheekbone".

Experiment E4 *Learning transferable models*: We evaluate the use-case **UC2** by training a ML model on sanitized images and evaluate it on real images (non-sanitized). This setup is similar to Classification Accuracy Score(CAS) in

	UTKFace		CelebA	
	Utility	Leakage	Utility	Leakage
Uniform Noise	0.667	0.501	0.576	0.712
GAP [22]	0.615	0.499	0.723	0.686
Adversarial Noise [22]	0.801	0.695	0.746	**0.663**
Ours	**0.86**	**0.474**	**0.9022**	0.6955

Table 2: E4, Classification Accuracy Score (CAS) evaluation: We train a classifier on privatized data samples and evaluate them on non-privatized samples.

	UTKFace		CelebA	
	Utility	Leakage	Utility	Leakage
Suppression	0.208	0.498	0.7042	0.7177
Obfuscation	0.208	0.491	0.62	0.7129
DP-Sampling	**0.521**	**0.474**	**0.817**	**0.6955**

Table 3: E5, CAS for learning a model of sensitive attribute. We experiment with different mechanisms in the *local sampling* stage.

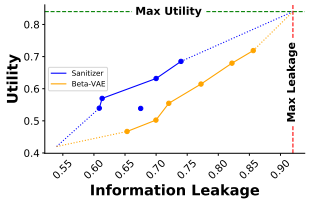


Fig. 5: Comparing β-VAE and global-decoupler by plotting the privacy-utility trade-off.

Aligner	Dcorr	Adv	Leakage ↓	Utility ↑
✗	✗	✓	0.6259	0.5474
✗	✓	✓	0.6238	0.5394
✓	✗	✗	0.6816	0.5137
✓	✗	✓	0.6318	0.5335
✓	✓	✗	0.6752	0.5386
✓	✓	✓	**0.6132**	**0.5698**

Table 4: Ablation on global-decoupler by cutting its different components. ✓ and ✗ denotes the presence and absence of the respective components.

the generative modeling community [49]. Note that it is not possible to include TIPRDC baseline since their output is constrained to embedding space.

Experiment E5 *Learn sensitive attribute distribution*: We test the use-case **UC3** of learning a ML model over the distribution $p(\mathbf{X}, \mathbf{Y}_S)$ while protecting individual sensitive information. We train the data-receiver's ML model on $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_S)$ and the attacker on $(\tilde{\mathbf{X}}, \mathbf{Y}_S)$. We evaluate data-receiver on $(\mathbf{X}, \mathbf{Y}_S)$ and the attacker on $(\tilde{\mathbf{X}}, \mathbf{Y}_S)$. This setup is not possible for our baselines since they censor sensitive information.

Results: For E1 and E2, we plot the privacy-utility trade-off for all techniques in Figure 4 and Table 1. *Sanitizer* obtains a better privacy-utility trade-off consistently. For E3, we compare trade-off by evaluating on multiple downstream tasks and observe sanitizer's consistent better performance. We posit that the consistent improvement is due to explicit modeling of different privacy-utility constraints in *global-decoupler*. We compare the results for E4 in Table 2. Unlike previous experiments, here *sanitizer* achieves substantial gap in comparison to other techniques. We believe that synthetically replacing sensitive information allows *sanitizer* to produce realistic sanitized samples. While the leakage is slightly larger on CelebA dataset, the relative improvement in utility is much larger. For E5, we compare three mechanisms proposed in Section 3.2 in Table 3. Under the same privacy budget, the proposed DP-Sampling technique achieves much better performance in both utility and leakage. Finally, we emphasize that E5 is not possible for baselines and is achieved only by the design of *sanitizer* and results for E4 and E5 validate that using *sanitizer* significantly improves performance for use-cases UC2 and UC3 (Section 1).

5 Discussion

Here, we analyze the design of global-decoupler by performing ablation study and discuss architectural limitations associated with it.

i) Ablation study for global-decoupler: We perform ablation on each of the components described in the architecture in Figure 2. We measure the sensitive information leakage and utility by comparing performance with and without each component in the objective given in Section 3. We can interpret this ablation study as keeping $\alpha_i = 0$ for the i'th component during the *global decoupling* stage. We enumerate the results in Table 4. We note that the presence of all components provides the best trade-off between the leakage and utility.
The global-decoupler is built upon β-VAE; therefore, we compare the trade-off between the two. We utilize latent space interpolation to randomize the sensitive attribute. First, we train a β-VAE model and obtain the mean representation of the sensitive latent by $z_{S_i} = \frac{1}{n_i} \sum_{j \in S_i} z_j \ \sim q_\phi(z|x_j)$ where S_i refers to a unique sensitive attribute category with n_i number of samples in the dataset. Finally, to randomize sensitive information in a given sample x, we transform the original latent $z = q_\phi(x)$ to obtain a sanitized latent $\tilde{z} = z - z_{S_i} + z_{S_j}$. For performing sensitive attribute randomization, i is the sensitive category of z, and j is chosen uniformly from the set of all categories including i. Finally, we obtain $\tilde{\mathbf{x}} = p_\theta(\tilde{z})$ as a sanitized transformation of x. We evaluate this technique on UTKFace dataset [76] with same experimental setup as E2. We model the attacker same way as explained in Section 4 and show trade-off curves in Figure 5.

ii) Architectural Limitations: The key goal of this work is to introduce a systematic framework and mechanisms for sanitization that could be useful for as many downstream tasks as possible under the privacy-utility trade-off. Here, we note two key limitations of the presented results, emerging from the generative modeling framework: i) *input sample size* - This limitation stems from the need for sufficient data points to learn a latent model of the data that generalizes between $\mathbf{D}^{aux}$ and $\mathbf{D}^{A}$. Designing latent models that can capture the distribution with a minimum number of samples is an active area of research in few-shot learning which will improve the impact of our results but is orthogonal to the scope of our work. ii) *output sample quality* - This limitation can be improved using hierarchical latent variable models [66, 51] and we consider this as part of the future work. We believe that improvement in representation capacity can improve trade-offs even further for *sanitizer*.

6 Related Work

First, we discuss prior work in privacy-preserving data release for task-dependent and task-independent setups. *Next*, we draw parallels to techniques in fairness and conditional generation of images.

Task-dependent data release techniques transform data that is conducive to a particular task. Several techniques use central DP [17] as a formal privacy definition to answer aggregate queries. The queries can be summary statistics such as mean, median [16, 17] or learning a ML model [1], sharing gradients

in federated learning [70]. Recent work in adversarial learning has resulted in techniques for task-specific latent representation [34, 71, 20, 55, 33, 46, 45, 56, 40, 60]. While we share the same goal of protecting sensitive information, our work differs in its task-independent formulation.

Task-independent techniques share data in a non-interactive manner. Similar to central DP, several works consider identification as sensitive information, however without a trusted curator. This modified central DP setup is referred to as local-DP [48]. While variants of local-DP for attribute privacy exist, their focus is primarily on protecting dataset statistics [75], different rows of a dataset [2] or task-dependent [12, 43]. Local-DP based generative models [24, 72, 65, 8, 77] learn data distribution privately to release samples. While we focus on specific sensitive information, we build upon the sampling strategy used in RonGauss [8] to sample sensitive data. TIPRDC [31] and GAP [23] are task-agnostic techniques that protect sensitive information by censoring it. While we solve a similar problem, our sampling-based approach allows performing certain tasks (eg. **E5** in Section 4) that are not possible with the censoring-based approach.

Fairness techniques aim to make predictive models unbiased with respect to *protected groups*. Among different approaches [7] used for fairness, works in censoring information [74, 69, 3, 75] related to *protected groups* is the closest approach to our work. However, we differ significantly from the censoring approach because we release anonymized sensitive information instead of censoring it. Furthermore, the goal of the sanitization problem is to maximally retain original data insofar that all biases would exist after sanitization. While the objective and evaluation for the fairness community are different, we note that Sarhan et al. [58] use a similar objective as *sanitizer* by utilizing variational inference with orthogonality constraint for preventing leakage. However, they do not provide anonymization since two correlated vectors can be orthogonal.

Conditional generation which has a similar problem setup to sanitization [59, 11]. While this has led to some relevant work in privacy, the techniques typically handcraft the objective to be task-specific for identity [13, 42, 41, 47, 50]. In contrast, *sanitizer* is agnostic of target utility and only depends upon sensitive attributes. Some recent works utilize uncertainty-based metrics [68, 38] to fuse all sensitive attributes in latent space using adversarial training but hence, unlike *sanitizer*, generate highly unrealistic images (hence not being task agnostic) due to high uncertainty.

7 Conclusion

In this work, we presented sanitizer: a framework for minimizing sensitive information leakage to facilitate task-agnostic data release. We achieve this goal through a two-stage process - i) *global decoupling* for learning a latent model of data and ii) *local sampling* for securely synthesizing sensitive information. While our approach improves the privacy-utility trade-off, future work includes technique that allow privacy guarantees for the non-sensitive latent.

Acknowledgements: This work was supported by NSF award number 1729931.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (Oct 2016), http://dx.doi.org/10.1145/2976749.2978318
2. Acharya, J., Bonawitz, K., Kairouz, P., Ramage, D., Sun, Z.: Context aware local differential privacy. In: International Conference on Machine Learning. pp. 52–62. PMLR (2020)
3. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., Pohl, K.M.: Bias-resilient neural network (2019)
4. Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., Kuo, P.C., et al.: Reading race: Ai recognises patient's racial identity in medical images. arXiv preprint arXiv:2107.10356 (2021)
5. Betzler, B.K., Yang, H.H.S., Thakur, S., Yu, M., Da Soh, Z., Lee, G., Tham, Y.C., Wong, T.Y., Rim, T.H., Cheng, C.Y., et al.: Gender prediction for a multiethnic population via deep learning across different retinal fundus photograph fields: Retrospective cross-sectional study. JMIR medical informatics **9**(8), e25165 (2021)
6. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics **22**(14), e49–e57 (2006)
7. Caton, S., Haas, C.: Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053 (2020)
8. Chanyaswad, T., Liu, C., Mittal, P.: Ron-gauss: Enhancing utility in non-interactive private data release. Proceedings on Privacy Enhancing Technologies **2019**(1), 26–46 (2019)
9. Chen, R.T., Li, X., Grosse, R., Duvenaud, D.: Isolating sources of disentanglement in variational autoencoders. arXiv:1802.04942 (2018)
10. Chen, S., Pan, Z.x., Zhu, H.j., Wang, Q., Yang, J.J., Lei, Y., Li, J.q., Pan, H.: Development of a computer-aided tool for the pattern recognition of facial features in diagnosing turner syndrome: comparison of diagnostic accuracy with clinical workers. Scientific reports **8**(1), 1–9 (2018)
11. Chen, Y.C., Shen, X., Lin, Z., Lu, X., Pao, I., Jia, J., et al.: Semantic component decomposition for face attribute manipulation. In: CVPR (2019)
12. Cheng, J., Tang, A., Chinchali, S.: Task-aware privacy preservation for multi-dimensional data. arXiv preprint arXiv:2110.02329 (2021)
13. Chhabra, S., Singh, R., Vatsa, M., Gupta, G.: Anonymizing k-facial attributes via adversarial perturbations. arXiv preprint arXiv:1805.09380 (2018)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., 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)
15. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography. pp. 265–284. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2006). https://doi.org/10.1007/11681878_14
16. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference (2006)
17. 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)

18. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv:1406.2661 (2014)
19. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: International conference on algorithmic learning theory. pp. 63–77. Springer (2005)
20. Hamm, J.: Minimax filter: Learning to preserve privacy from inference attacks. Journal of Machine Learning Research **18**(129), 1–31 (2017), http://jmlr.org/papers/v18/16-501.html
21. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework (2016)
22. Huang, C., Kairouz, P., Chen, X., Sankar, L., Rajagopal, R.: Context-Aware Generative Adversarial Privacy. Entropy **19**(12), 656 (Dec 2017). https://doi.org/10.3390/e19120656, http://arxiv.org/abs/1710.09549, arXiv: 1710.09549
23. Huang, C., Kairouz, P., Chen, X., Sankar, L., Rajagopal, R.: Generative adversarial privacy. CoRR (2018)
24. Jordon, J., Yoon, J., Schaar, M.v.d.: PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees (Sep 2018), https://openreview.net/forum?id=S1zk9iRqF7
25. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age. arXiv:1908.04913 (2019)
26. Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. ACM TODS (2014)
27. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
29. Korot, E., Pontikos, N., Liu, X., Wagner, S.K., Faes, L., Huemer, J., Balaskas, K., Denniston, A.K., Khawaja, A., Keane, P.A.: Predicting sex from retinal fundus photographs using automated deep learning. Scientific reports **11**(1), 1–8 (2021)
30. Kumar, D., Verma, C., Dahiya, S., Singh, P.K., Raboaca, M.S.: Cardiac diagnostic feature and demographic identification models: A futuristic approach for smart healthcare using machine learning (2021)
31. Li, A., Duan, Y., Yang, H., Chen, Y., Yang, J.: Tiprdc: Task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. In: ACM SIGKDD (2020)
32. Li, A., Duan, Y., Yang, H., Chen, Y., Yang, J.: Tiprdc: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 824–832 (2020)
33. Li, A., Guo, J., Yang, H., Chen, Y.: Deepobfuscator: Adversarial training framework for privacy-preserving image classification (2019)
34. Liu, C., Chakraborty, S., Mittal, P.: Deeprotect: Enabling inference-based access control on mobile sensing applications. CoRR (2017)
35. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset. Retrieved August (2018)
36. Loos, H.S., Wieczorek, D., Würtz, R.P., Malsburg, C.v.d., Horsthemke, B.: Computer-based recognition of dysmorphic faces. European Journal of Human Genetics **11**(8), 555–560 (2003)

37. Makhdoumi, A., Fawaz, N.: Privacy-utility tradeoff under statistical uncertainty. In: Annual Allerton Conference on Communication, Control, and Computing (Allerton) (2013)
38. Martinsson, J., Zec, E.L., Gillblad, D., Mogren, O.: Adversarial representation learning for synthetic replacement of private attributes. arXiv preprint arXiv:2006.08039 (2020)
39. McLachlan, G.J., Lee, S.X., Rathnayake, S.I.: Finite mixture models. Annual review of statistics and its application **6**, 355–378 (2019)
40. Mireshghallah, F., Taram, M., Ramrakhyani, P., Tullsen, D.M., Esmaeilzadeh, H.: Shredder: Learning noise to protect privacy with partial DNN inference on the edge. CoRR **abs/1905.11814** (2019), http://arxiv.org/abs/1905.11814
41. Mirjalili, V., Raschka, S., Ross, A.: Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. IEEE Access **7**, 99735–99745 (2019)
42. Mirjalili, V., Raschka, S., Ross, A.: Privacynet: semi-adversarial networks for multi-attribute face privacy. IEEE Transactions on Image Processing **29**, 9400–9412 (2020)
43. Murakami, T., Kawamoto, Y.: {Utility-Optimized} local differential privacy mechanisms for distribution estimation. In: 28th USENIX Security Symposium (USENIX Security 19). pp. 1877–1894 (2019)
44. Osia, S.A., Shamsabadi, A.S., Sajadmanesh, S., Taheri, A., Katevas, K., Rabiee, H.R., Lane, N.D., Haddadi, H.: A hybrid deep learning architecture for privacy-preserving mobile analytics. IEEE Internet of Things Journal (2020)
45. Osia, S.A., Shamsabadi, A.S., Sajadmanesh, S., Taheri, A., Katevas, K., Rabiee, H.R., Lane, N.D., Haddadi, H.: A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics. IEEE Internet of Things Journal **7**(5), 4505–4518 (May 2020). https://doi.org/10.1109/JIOT.2020.2967734, http://arxiv.org/abs/1703.02952, arXiv: 1703.02952
46. Osia, S.A., Taheri, A., Shamsabadi, A.S., Katevas, K., Haddadi, H., Rabiee, H.R.: Deep private-feature extraction (2018)
47. Othman, A., Ross, A.: Privacy of facial soft biometrics: Suppressing gender but retaining identity. In: European Conference on Computer Vision. pp. 682–696. Springer (2014)
48. Raskhodnikova, S., Smith, A., Lee, H.K., Nissim, K., Kasiviswanathan, S.P.: What can we learn privately. In: FOCS (2008)
49. Ravuri, S., Vinyals, O.: Classification accuracy score for conditional generative models. arXiv:1905.10887 (2019)
50. Raynal, M., Achanta, R., Humbert, M.: Image obfuscation for privacy-preserving machine learning. arXiv preprint arXiv:2010.10139 (2020)
51. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
52. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
53. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML (2014)
54. Roy, P.C., Boddeti, V.N.: Mitigating information leakage in image representations: A maximum entropy approach. In: CVPR (2019)
55. Roy, P.C., Boddeti, V.N.: Mitigating information leakage in image representations: A maximum entropy approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

56. Samragh, M., Hosseini, H., Triastcyn, A., Azarian, K., Soriaga, J., Koushanfar, F.: Unsupervised information obfuscation for split inference of neural networks. arXiv preprint arXiv:2104.11413 (2021)
57. Sankar, L., Rajagopalan, S.R., Poor, H.V.: An information-theoretic approach to privacy. In: Allerton Conference on Communication, Control, and Computing (Allerton) (2010)
58. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. In: European Conference on Computer Vision. pp. 746–761. Springer (2020)
59. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: CVPR (2017)
60. Singh, A., Chopra, A., Sharma, V., Garza, E., Zhang, E., Vepakomma, P., Raskar, R.: Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks. arXiv:2012.11025 (2020)
61. Stephen, I.D., Hiew, V., Coetzee, V., Tiddeman, B.P., Perrett, D.I.: Facial shape analysis identifies valid cues to aspects of physiological health in caucasian, asian, and african populations. Frontiers in psychology **8**, 1883 (2017)
62. Su, N.M., Crandall, D.J.: The affective growth of computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9291–9300 (2021)
63. Székely, G.J., Rizzo, M.L., Bakirov, N.K., et al.: Measuring and testing dependence by correlation of distances. The annals of statistics (2007)
64. Tao, Y., McKenna, R., Hay, M., Machanavajjhala, A., Miklau, G.: Benchmarking differentially private synthetic data generation algorithms. arXiv preprint arXiv:2112.09238 (2021)
65. Torkzadehmahani, R., Kairouz, P., Paten, B.: Dp-cgan: Differentially private synthetic data and label generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
66. Vahdat, A., Kautz, J.: Nvae: A deep hierarchical variational autoencoder. Advances in Neural Information Processing Systems **33**, 19667–19679 (2020)
67. Vepakomma, P., Singh, A., Zhang, E., Gupta, O., Raskar, R.: Nopeek-infer: Preventing face reconstruction attacks in distributed inference after on-premise training. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–8. IEEE (2021)
68. Wang, H.P., Orekondy, T., Fritz, M.: Infoscrub: Towards attribute privacy by targeted obfuscation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3281–3289 (2021)
69. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5310–5319 (2019)
70. Wei, K., Li, J., Ding, M., Ma, C., Yang, H.H., Farokhi, F., Jin, S., Quek, T.Q., Poor, H.V.: Federated learning with differential privacy: Algorithms and performance analysis. IEEE Transactions on Information Forensics and Security **15**, 3454–3469 (2020)
71. Wu, Z., Wang, Z., Wang, Z., Jin, H.: Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 606–624 (2018)
72. Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J.: Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739 (2018)

73. Yi, P.H., Wei, J., Kim, T.K., Shin, J., Sair, H.I., Hui, F.K., Hager, G.D., Lin, C.T.: Radiology "forensics": determination of age and sex from chest radiographs using deep learning. Emergency Radiology **28**(5), 949–954 (2021)
74. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning. pp. 325–333. PMLR (2013)
75. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
76. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5810–5818 (2017)
77. Zhang, Z., Wang, T., Honorio, J., Li, N., Backes, M., He, S., Chen, J., Zhang, Y.: Privsyn: Differentially private data synthesis (2021)