MASS: MULTI-ATTRIBUTE SELECTIVE SUPPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

The recent rapid advances in the development and deployment of machine learning technologies largely depend on the vast richness of data available today, in terms of both the quantity and the rich content contained within. For example, biometric data such as images and voices could reveal people's attributes like age, gender, sentiment, and origin, whereas location/motion data could be used to infer people's activity levels, transportation modes, and life habits. Along with the new services and applications enabled by such technological advances, various governmental policies are put in place to regulate such data usage and protect people's privacy and rights. As a result, data owners often opt for simple data obfuscation (e.g., blur people's faces in images) or withholding data altogether, which leads to severe data quality degradation and greatly limits the data's potential utility.

Aiming for a sophisticated mechanism which gives data owners fine-grained control while retaining the maximal degree of data utility, we propose *Multi-attribute Selective Suppression*, or *MaSS*, a general framework for performing precisely targeted data surgery to simultaneously *suppress* any selected set of attributes while *preserving* the rest for downstream machine learning tasks. MaSS learns a data modifier through adversarial games between two sets of networks, where one is aimed at suppressing selected attributes, and the other ensures the retention of the rest of the attributes via general contrastive loss as well as explicit classification metrics. We carried out an extensive evaluation of our proposed method using multiple datasets from different domains including facial images, voice audio, and video clips, and obtained highly promising results in MaSS' generalizability and capability of drastically suppressing targeted attributes (e.g., reducing inference on such attributes to random guess) while imposing virtually no impact on the data's usability in other downstream ML tasks.

1 INTRODUCTION

The recent rapid advances in machine learning (ML) can be largely attributed to powerful computing infrastructures as well as the availability of large-scale datasets, such as ImageNet1K (Deng et al., 2009) for computer vision, WMT (Foundation, 2019) for neural machine translation, and LibriLight (Kahn et al., 2020) for speech recognition. Studies have shown that ML models trained on large-scale datasets can usually prove effective in many additional downstream tasks (Brown et al., 2020). On the other hand, ethical concerns have been raised surrounding proper data usage in issues like data privacy (Liu et al., 2021), data minimization (Goldsteen et al., 2021), etc. Therefore, if there are more data available and can be used without worrying whether or not the data is handled properly, the ML models can be further improved by more data and help the ML community to advance on many domains.

Attempting to balance between model performance and proper data usages, a common approach usually taken is to simply modify the data to remove its "sensitive" attributes, and experimentally demonstrate that the targeted sensitive attributes are indeed removed. What's crucially important but usually omitted here, however, is the preservation of the "total utility" of the data, because the suppression operation oftentimes also negatively impact, or even completely destroy, the other "non-sensitive" attributes, hence greatly damaging the dataset's potential future utility. For example, DeepPrivacy (Hukkelås et al., 2019) *is* able to demonstrate its privacy protection capability, but the modified data it produces can no longer be utilized for additional downstream tasks like sentiment analysis, age detection, or gender classification. Since data is one of the main driving forces for the

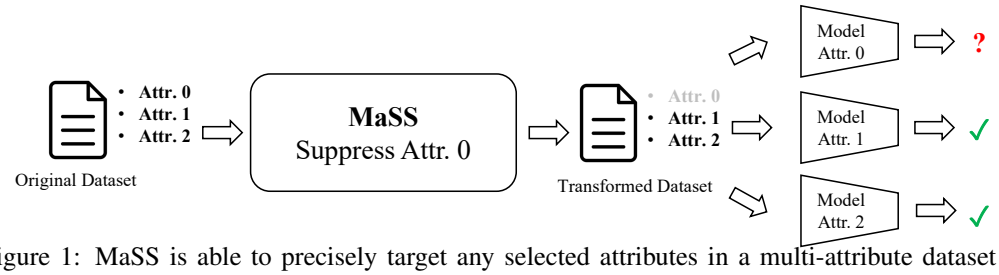

Figure 1: MaSS is able to precisely target any selected attributes in a multi-attribute dataset for suppression while leaving the rest of the attributes intact for any potential downstream ML-based analytic tasks. For example as illustrated by the diagram, when operating on the original multi-attribute dataset and configured to suppress Attr. 0, MaSS is able to transform the dataset such that the model for detecting Attr. 0 is unable to reliably detect Attr. 0 from the transformed data, while the models for Attr. 1 and 2 are not affected.

rapid advancement of machine learning research, we argue that the ideal scenario would be to have the flexibility of selecting an arbitrary set of attributes and only suppressing *them* while leaving all the other attributes completely intact. In this way, the community could unleash the potential utility of the modified data to develop more advanced algorithms.

Towards this exact goal, we present Multi-attribute Selective Suppression (or MaSS) in this paper, to enable such capability of precise attributes suppression for multi-attribute datasets. The high-level objective of MaSS is also illustrated in Figure 1, where MaSS is configured to suppress Attr. 0 without knowing in advance that Attr. 1 and 2 will be used for downstream tasks. After the data transformation performed by MaSS, Attr. 0 becomes suppressed, nondetectable by its corresponding machine learning model, but at the same time, Attr. 1 and 2 are left intact, and still can be extracted by their corresponding ML models. As a concrete example, suppose we are working with a facial-image dataset which contains attributes like age, gender, and sentiment, where, let us assume, age and gender are considered sensitive. Then, MaSS would transform this facial-image dataset such that age and gender information could no longer be not be inferred by the corresponding ML models, whereas sentiment information could still be extracted from the transformed data.

The contributions of our work are threefolds,

1. We propose the novel MaSS framework to enable the powerful flexibility of precise suppression of arbitrary, selective data attributes.
2. We employ multiple learning mechanisms in MaSS to enable its attribute-specific as well as generic feature preservation capabilities, which help it achieve satisfactory data utility protection both with and without the prior knowledge about downstream tasks.
3. We thoroughly validate MaSS using a wide range of multi-attribute datasets, including image, audio, and video. All our results demonstrate MaSS' strong performance in its intended selective attribute suppression and preservation.

2 Related Works

Data Privacy. A large body of work has studied methods applying generative adversarial networks (GANs) to generate and modify facial features in images, so these identity-related sensitive features can be de-identified. DeepPrivacy (Hukkelås et al., 2019) proposed to use a conditional generative adversarial network to generate realistic anonymized faces, while considering the existing background and a sparse pose annotation. To further ensure the face anonymization using GAN-based methods, CIAGAN (Maximov et al., 2020) proposed an identity control discriminator to control which fake identity is used in the anonymizaiton process by introducing an identity control vector. Instead of generating the entire faces for anonymization, Li et al. proposed to apply conditional GAN to only identify and modify the five identity-sensitive attributes. To also enable face anonymization with the selected semantic attributes manipulation, PI-Net (Chen et al., 2021) proposed to generate realistic looking faces with the selected attributes preserved. The above works usually focuses on suppression only while the future data utilities are not considered. Our approach not only suppresses the attributes but also preserves the data utility concurrently. On the other hand, SPAct tried to sup-

press the multiple attributes in a video through contrastive learning while preserving the utility for action recognition; however, their approach lacks the flexibility to handle individual attributes but can only process all attributes at once and limits to the action recognition dataset; while our method is fully configurable and validated in different data domains. Moriarty et al. proposed the method to suppress the biometric information while preserving its utility; however, their approach requires the information of downstream task while our method does not.

Dataset Distillation/Condensation. Dataset distillation and condensation aim to create a smaller version of a large dataset which can be used to train a model whose performance can be as good as training on the original large dataset. Therefore, training can be relatively quicker, e.g., Neural Architecture Search (NAS) methods require a lot of iterations of a whole dataset to find out the best model. Otherwise, NAS usually needs to use a proxy dataset/model for the approximation results from large dataset/model (Zhao et al., 2021; Cazenavette et al., 2022; Wang et al., 2022). On the other hand, a recent work shows the condensed dataset also conceals some attributes from the original data (Dong et al., 2022) but still remains effective for the original task. In contrast to these approaches, our proposed method tries to keep the truthfulness of data as much as possible, so we do not reduce the amount of data. Moreover, our method is designed to preserve the generic features rather than the specific task, which could be covered by the generic feature we preserved.

Self-supervised/Contrastive Learning. Self-supervised learning has been explored by the community to allow a machine learning model to learn a good data representation by designing pretext tasks instead of human annotations (Feng et al., 2019), or contrastive learning which tries to maximize the agreement between positive pairs (Chen et al., 2020; He et al., 2020; Grill et al., 2020), or clustering-based methods to generate pseudo labels for data (Caron et al., 2020; 2018), or mask autoencoder to predict the masked patches by the remaining patches (He et al., 2022). Then, they show that the feature representations are usually good for many different downstream tasks. Our goal is to keep as much information when suppressing the selected attributes without having prior knowledge; thus, self-supervised and contrastive learning methods facilitate our requirement to extract generic features without label information from downstream tasks.

3 Proposed Method

3.1 Problem Definition

Consider a multi-attribute dataset $\mathbf{X}$ with size N, and the set A of different attributes, where each data point $\mathbf{x}$'s value for attribute $a \in A$ is $a_{\mathbf{x}}$. Each of the different attributes can be learned by inferencing on $\mathbf{X}$, hence our objective is to transform $\mathbf{X}$ to suppress any selective subset of attributes $S \subseteq A$ such that no attributes in S can be reliably inferred from the transformed dataset $\mathbf{X}'$. At the same time, we need to make sure the rest of the attributes $R = A \setminus S$ are preserved and can still be inferred from $\mathbf{X}'$.

In practice, when a data owner would like to transform their data, we assume the subset S of attributes targeted for suppression is always predetermined. However, it is not always known in advance what the entire set A of attributes are, and consequently which set R of attributes need to be preserved. Therefore, for generalizability, we consider the case of an unknown R at the time of data transformation. Of course, if R happens to be given, our solution needs to be able to take advantage of this extra information as well.

3.2 Proposed Framework

We propose a Generative Adversarial Network (GAN)-based solution in tackling the data transformation problem. Our framework consists of three major components: a *Data Modifier*, a *Suppression Branch*, and a *Preservation Branch*, as depicted in Fig. 2. The data modifier G is the generator while both branches are the discriminators in the GAN framework. The data modifier tries to generate new data such that the similarity between original data and modified data are maximized and minimized via the suppression branch and the preservation branch, respectively. In a nutshell, the data modifier learns a transformation that is to be applied to the original data vectors, where the learned transformation is jointly regularized by both the suppression and preservation branches to

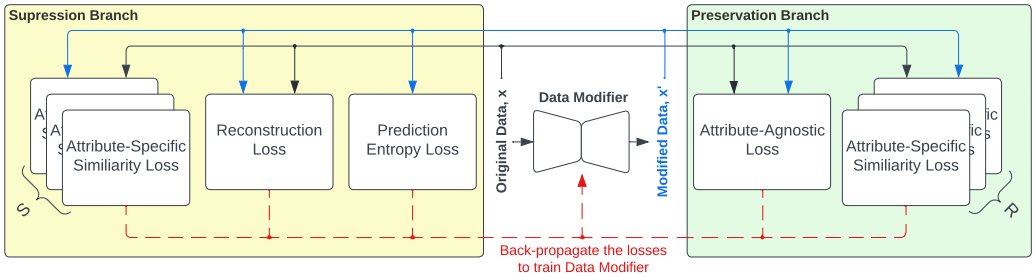

Figure 2: **Overview of MaSS.** MaSS contains three components: the data modifier, the suppression branch, and the preservation branch. The data modifier G is trained by optimizing the losses in both the suppression and the preservation branches.

ensure all targeted attributes in S are indeed suppressed in the transformed data, while all other attributes, explicitly specified or not, are left intact as much as possible. We next discuss all three components in more detail.

3.2.1 Data Modifier

We design the data modifier to be conditional on its original input in learning instance-specific transformations via a multi-layer perceptron (MLP) with a residual shortcut. In other words, based on its input $\mathbf{x}$, the data modifier G learns an additive modification to be applied to $\mathbf{x}$. Therefore, the transformed data stays in the same embedding space as the original, which leads to faster optimization convergence. Specifically, for a normalized original data vector $\mathbf{x}$, its normalized transformed version is computed as

$$\mathbf{x}' = G(\mathbf{x}) := n(\mathbf{x} + n(\text{MLP}(\mathbf{x}))), \tag{1}$$

where $n(\cdot)$ is the normalization function $n(\mathbf{v}) = \begin{cases} \frac{\mathbf{v}}{||\mathbf{v}||}, & ||\mathbf{v}|| \neq 0 \\ 0, & \text{o.w.} \end{cases}$.

During the optimization, the parameters in the data modifier are trained by back-propagating the losses designed in both branches.

3.2.2 Suppression Branch

The suppression branch is designed to make the targeted attributes in S as unrecognizable as possible. It utilizes the corresponding set of inference models pretrained on the original dataset $\mathbf{X}$. Each pretrained model corresponds to a specific attribute $s \in S$ and is composed of a feature extractor F^s, which converts raw input data to a feature vector $\mathbf{z}^s_{\mathbf{x}} = F^s(\mathbf{x})$, and a projector P^s, which maps the features to the attribute's label $\mathbf{p}^s_{\mathbf{x}} = P^s\left(\mathbf{z}^s_{\mathbf{x}}\right)$, where $\mathbf{z}^s_{\mathbf{x}}$ is the feature representation and $\mathbf{p}^s_{\mathbf{x}}$ is the prediction logit.

During the training of the data modifier, the suppression branch guides the data modifier to degrade these pretrained models' recognition accuracies on the transformed data $\mathbf{X}'$. It does so by either measuring the similarity of features from the pretrained models or by comparing the prediction results against the ground truth labels, and then adding a corresponding penalty. Specifically, for each targeted attribute $s \in S$, the feature-similarity loss function is defined as

$$L^s_{\text{sim}} = w^s \cdot \frac{1}{N} \sum_{\mathbf{x} \in \mathbf{X}} \text{sim}^s\left(\mathbf{x}', \mathbf{x}\right), \tag{2}$$

where w^s is the weight for attribute s and $\text{sim}^s(\cdot,\cdot)$ defines a similarity measure. For example, we can use the cosine similarity between feature vectors $\text{sim}^s_{\cos}\left(\mathbf{x}', \mathbf{x}\right) = \cos\left(\mathbf{z}^s_{\mathbf{x}'}, \mathbf{z}^s_{\mathbf{x}}\right)$, or the negative KL-divergence between the original and the transformed logits $\text{sim}^s_{\text{KL}}\left(\mathbf{x}', \mathbf{x}\right) = -D_{\text{KL}}\left(\mathbf{p}^s_{\mathbf{x}'} \middle\| \mathbf{p}^s_{\mathbf{x}}\right)$, or the cross-entropy loss between the transformed logits and the attribute's ground truth value $\text{sim}^s_{\text{CE}}\left(\mathbf{x}', \mathbf{x}\right) = -D_{\text{KL}}\left(\mathbf{p}^s_{\mathbf{x}'} \middle\| \mathbf{1}_{s_{\mathbf{x}}}\right)$, where $s_{\mathbf{x}}$ is $\mathbf{x}$'s ground truth value for attribute s and $\mathbf{1}_{s_{\mathbf{x}}}$ is the corresponding one-hot vector.

The above similarity-based metrics only help guide the data modifier to lead the pretrained models towards incorrect predictions about S on the transformed data. However, in terms of attribute suppression, an even stronger condition would be to reduce such predictions to random guesses. Therefore, we introduce an additional loss term to maximize the entropy H of the predictions about S on the transformed data such that the predictions look like random guesses. Combined with Eq. 2, the total loss on the suppression branch is

$$L^s = w^s \cdot \frac{1}{N} \sum_{\mathbf{x} \in \mathbf{X}} \text{sim}^s\left(\mathbf{x}', \mathbf{x}\right) - h^s \cdot \frac{1}{N} \sum_{\mathbf{x} \in \mathbf{X}} H\left(\mathbf{p}^s_{\mathbf{x}'}\right), \tag{3}$$

where h^s is the weight for the entropy loss.

Moreover, to regularize the learned data transformation to be as small as possible, we also add an $L2$ reconstruction loss term in the suppression branch

$$L_{\text{rec}} = w_{\text{rec}} \cdot \frac{1}{N} \sum_{\mathbf{x} \in \mathbf{X}} ||\mathbf{x}' - \mathbf{x}||_2. \tag{4}$$

3.2.3 Preservation Branch

While the suppression branch guides the data modifier to decrease the confidence of machine learning models on certain attributes S from the data, the preservation branch is responsible for guarding the data against said suppression and erasure such that maximum utility can be preserved through the transformation, in the sense that all the attributes $R = A \setminus S$ not targeted by the suppression branch should remain recognizable from the transformed data $\mathbf{X}'$, just as they were from the original data $\mathbf{X}$. In the proposed method, we design two types of losses: one is *attribute-agnostic* in the preservation branch, which is applicable to any attribute as the loss is defined in an agnostic way; another one is *attribute-specific*, which if the downstream tasks have been defined and we know the set of R, then we embedded them into the loss.

Attribute-Agnostic. As previously discussed, it is oftentimes difficult to identify in advance the set A of *all* attributes associated with a dataset, which means the set R might not necessarily be defined even when S is specified. The absence of R indicates that we have no knowledge about what attributes downstream tasks might want to detect from the transformed dataset.

To tackle this *attribute-agnostic* scenario, we employ self-supervised techniques, where a generic feature representation of the data is learned without any specific attribute or task information. For our preservation branch in particular, we adopt the SimCLR (Chen et al., 2020) contrastive-learning-based approach to pretrain an extractor F^{r_*} and a projector P^{r_*} that maximize the similarity between embedding pairs originated from the same data point (i.e., positive pairs) and minimize that of different data points (i.e., negative pairs). Instead of considering two different transformations like SimCLR does, we treat an original data point $\mathbf{x}$ and its transformed version $\mathbf{x}'$ as a positive pair, and all the rest as negative. We enforce this relationship by applying normalized temperature-scaled cross entropy (NT-Xent) loss function introduced in SimCLR; therefore, semantically, the transformed data could preserve more generic features. For the original data point $\mathbf{x}$-based positive pair $(\mathbf{x}, \mathbf{x}')$, the NT-Xent loss $l^{r_*}_{\mathbf{x}}$ is

$$l^{r_*}_{\mathbf{x}} = -\log \frac{e^{\cos\left(\mathbf{p}^{r_*}_{\mathbf{x}}, \mathbf{p}^{r_*}_{\mathbf{x}'}\right)/\tau}}{\sum_{\mathbf{y} \in \mathbf{X}} \left[e^{\cos\left(\mathbf{p}^{r_*}_{\mathbf{x}}, \mathbf{p}^{r_*}_{\mathbf{y}'}\right)/\tau} + \mathbb{1}_{\mathbf{y} \neq \mathbf{x}} \cdot e^{\cos\left(\mathbf{p}^{r_*}_{\mathbf{x}}, \mathbf{p}^{r_*}_{\mathbf{y}}\right)/\tau} \right]}, \tag{5}$$

and similarly for the transformed data point $\mathbf{x}'$-based positive pair $(\mathbf{x}', \mathbf{x})$, the NT-Xent loss $l^{r_*}_{\mathbf{x}'}$ is

$$l^{r_*}_{\mathbf{x}'} = -\log \frac{e^{\cos\left(\mathbf{p}^{r_*}_{\mathbf{x}'}, \mathbf{p}^{r_*}_{\mathbf{x}}\right)/\tau}}{\sum_{\mathbf{y} \in \mathbf{X}} \left[e^{\cos\left(\mathbf{p}^{r_*}_{\mathbf{x}'}, \mathbf{p}^{r_*}_{\mathbf{y}}\right)/\tau} + \mathbb{1}_{\mathbf{y} \neq \mathbf{x}} \cdot e^{\cos\left(\mathbf{p}^{r_*}_{\mathbf{x}'}, \mathbf{p}^{r_*}_{\mathbf{y}'}\right)/\tau} \right]}, \tag{6}$$

where $\mathbf{p}^{r_*}_{\mathbf{x}} = P^{r_*}(F^{r_*}(\mathbf{x}))$ and $\mathbf{p}^{r_*}_{\mathbf{x}'} = P^{r_*}(F^{r_*}(\mathbf{x}'))$ are the attribute-agnostic logits computed from the original and transformed data points, respectively; τ controls the contribution proportion between the positive and the negative samples, with a higher τ value indicating a bigger contribution from

the positive samples, and vice versa; and $\mathbb{1}_{\mathbf{y}\neq\mathbf{x}}$ is the indicator function $\mathbb{1}_{\mathbf{y}\neq\mathbf{x}} = \begin{cases} 1, & \mathbf{y} \neq \mathbf{x} \\ 0, & \text{o.w.} \end{cases}$. The total loss for the preservation branch is thus aggregated across all data points

$$L^{r_*} = w^{r_*} \cdot \frac{1}{2N} \sum_{\mathbf{x}\in\mathbf{X}} (l_{\mathbf{x}}^{r_*} + l_{\mathbf{x}'}^{r_*}), \tag{7}$$

which enables the preservation branch to enforce the underlying representations of the transformed data points in $\mathbf{X}'$ to reach maximal agreement with their corresponding origins from $\mathbf{X}$ while maintaining their discriminative characteristics with each other.

Attribute-Specific. In addition to the attribute-agnostic case discussed above, there could also be *attribute-specific* situations, where the set R of attributes that needs to be preserved is explicitly defined in advance for the data transformation. Therefore, we can pretrain the inference models on the original dataset $\mathbf{X}$ like in the suppression branch. To account for this extra information, R, we formulate the loss function similar to Eq. 2, except that here we want to maximize the similarities as opposed to minimizing them, as follows,

$$L^{r} = -w^{r} \cdot \frac{1}{N} \sum_{\mathbf{x}\in\mathbf{X}} \text{sim}^{r}(\mathbf{x}', \mathbf{x}), \tag{8}$$

where w^r is the weight for each attribute $r \in R$.

Finally, collecting all the loss terms from the data modifier, the suppression branch, as well as the attribute-agnostic and attribute-specific components of the preservation branch, we have the overall optimization loss

$$L = \sum_{s\in S} L^{s} + L_{\text{rec}} + L^{r_*} + \sum_{r\in R} L^{r}, \tag{9}$$

which is similar to GAN that the data modifier tries to minimax the loss.

4 Experimental Evaluation

We next present and discuss our experimental evaluation of MaSS on its ability to perform selective attribute suppression and preservation when carrying out data transformations.

4.1 Experimental Setup

We evaluated MaSS on three multi-attribute datasets of different domains, namely Adience (Eidinger et al., 2014) for facial images, AudioMNIST (Becker et al., 2018) for voice recordings, and PA-HMDB (Wu et al., 2020) for video clips. For all datasets, we converted all their raw data points to feature embeddings via state-of-the-art neural networks as the input $\mathbf{X}$. Table 1 lists the feature extractors used for each dataset, as well as the corresponding feature dimension and the architecture of MLP used in the data modifier. All our experiments were designed to examine MaSS' performance on attribute suppression and preservation quality, and not whether it could generate high quality synthetic data. Therefore, each transformed $\mathbf{x}'$ would stay in the same feature space as its corresponding $\mathbf{x}$. We next briefly introduce the datasets, as well as the training and evaluation protocols. More details can be found in Appx. A.

Adience. The Adience image dataset was originally published to help study the recognition of age and gender. Each image is also associated with a DataID. In total, the dataset used in our experiment contains 1,089 different DataIDs, 8 age groups, and 2 gender classes. We split the dataset into 2,815 for training and 2,525 for validation.

AudioMNIST. The AudioMNIST dataset contains audio recordings of spoken digits (0-9) in English from 60 speakers. In addition to speakerID and spoken digits, the dataset also contain accent and gender attributes. Therefore, we use AudioMNIST as a 4-attribute dataset, namely speakerID, spoken digits, accent, and gender, with 60, 10, 18, and 2 classes, respectively. There are 30,000 audio clips in total. We split the data into 18,000, 6,000, and 6,000 for training, validation, and testing, respectively.

Table 1: Model configuration for each dataset.

Dataset	Feature Extractor	Feature Dimension	MLP in Data Modifier
Adience	FaceNet (Schroff et al., 2015)	512	512-256-128-256-512
AudioMNIST	HuBERT-L (Hsu et al., 2021)	1024	1024-256-128-256-1024
PA-HMDB	R3D-18 (Tran et al., 2018)	512	512-256-128-256-512

Table 2: Results on Adience under different configurations. The checkmark (✓) denotes that the particular loss is used in the optimization. DataID is selected as the suppression target.

	Loss Configuration					Top-1 Accuracy (%)		
	L_{rec}	L^{s_1}	L^{r_*}	L^{r_1}	L^{r_2}	DataID (s_1)	Age (r_1)	Gender (r_2)
Original	-	-	-	-	-	90.8	89.1	97.4
MaSS	✓	✓	-	-	-	0.0	33.8	73.5
	✓	✓	✓	-	-	0.6	78.5	95.7
	✓	✓	✓	✓	-	0.7	86.1	95.9
	✓	✓	✓	✓	✓	0.6	86.9	96.7

PA-HMDB. The PA-HMDB51 dataset is subset of HMDB51 and it contains 6 attributes: action, skin color, face, gender, nudity, and relationship (Kuehne et al., 2011; Wu et al., 2020). The action label is annotated at video-level while the other non-action attributes are at frame-level. Nonetheless, PA-HMDB51 only contains about 500 videos, it is only used for evaluation. There are 51 different action classes. The other 5 non-action attributes are all binary. As there is no training data for the non-action attributes in PA-HDMB, we used the VISPR dataset (Orekondy et al., 2017) for training them, and used HMDB51 for the action attribute. MaSS was evaluated on both VISPR and PA-HMDB51.

Training and Evaluation Protocols. Before we can start training MaSS' data modifier G, we first need to train the models for the attributes we intend to suppress in the suppresion branch, as well as the attribute-agnostic model for the preservation branch, and potentially also the attribute-specific models if the corresponding labels for those attributes are available. Subsequently, each of these pre-trained models is used in either the suppression or the preservation branch during the training of the data modifier G via optimizing Eq. 9 defined in Sec. 3. Note that during training G, the pre-trained models in both branches are fixed without any update. For evaluation, we use the trained data modifier to generate new data and feed them through the pre-trained models and examine their performance. For both Adience and AudioMNIST, we use top-1 accuracy as the metric and report the results on each attribute. For PA-HMDB51, we follow the same practice as PA-HMDB (Wu et al., 2020) to aggregate the performance of the 5 non-action attributes via macro-average of classwise mean average precision (cMAP) and use top-1 accuracy for action. More details on the model training can be found in Appx. B.

4.2 Comparison to Baselines

First, we examine the effects of each of the loss terms for each of the datasets. Table 2 shows the performance on Adience where DataID is targeted for suppresion, and the others are expected to be preserved. When only imposing attribute suppression (L_{rec} and L^{s_1}), MaSS did successfully degrade the performance of DataID. But at the same time, the performance on age and gender deteriorated significantly. When we added the attribute-agnostic loss (L^{r_*}), MaSS still achieved good suppression on DataID. But at the same time it greatly improved the recognition accuracy on age and gender. Note that, we do not use any information from age and gender attributes during the training. Moreover, when the age information (L^{r_1}) was incorporated in MaSS, the performance of age detection on the transoformed dataset almost returned to the same level as that of the original dataset. The same trend can be observed for the gender attribute (L^{r_2}) as well.

A similar outcome can also be observed for the AudioMNIST dataset (Table 3), where SpeakerID was treated as the suppression target, and all the rest of the attributes were expected to be preserved. When MaSS only used suppression (L_{rec} and L^{s_1}), SpeakerID score did reduce, but so did the

Table 3: Results on AudioMNIST under different configurations. The checkmark (✓) denotes that the particular loss term is used in the optimization. SpeakerID is selected as the suppression target.

	Loss Configuration						Top-1 Accuracy (%)			
	L_{rec}	L^{s_1}	L^{r_*}	L^{r_1}	L^{r_2}	L^{r_3}	SpeakerID (s_1)	Digit (r_1)	Accent (r_2)	Gender (r_3)
Original	-	-	-	-	-	-	95.6	99.8	99.3	96.5
MaSS	✓	✓	-	-	-	-	0.0	26.2	45.3	53.5
	✓	✓	✓	-	-	-	1.7	67.4	68.7	88.2
	✓	✓	✓	✓	-	-	1.7	99.7	68.4	80.0
	✓	✓	✓	✓	✓	-	1.7	99.7	95.1	86.6
	✓	✓	✓	✓	✓	✓	1.7	99.6	95.7	98.4

Table 4: Results on VISPR and PA-HMDB under different configurations. The checkmark (✓) denotes that the particular loss term is used in the optimization. Metrics for the action attribute is Top-1 Accuracy (%) while cMAP (%) is used for the other 5 non-action attributes. MaSS is configured to suppress the non-action attributes.

	Loss Configuration				VISPR	PA-HMDB	
	L_{rec}	L^{s_1}	L^{r_*}	L^{r_1}	Non-action Attrs. (s_1)	Action (r_1)	Non-action Attrs. (s_1)
Original	-	-	-	-	81.8	58.7	79.7
MaSS	✓	✓	-	-	41.8	12.6	70.3
	✓	✓	✓	-	36.3	52.1	63.2
	✓	✓	✓	✓	38.6	58.0	63.4

other attributes. Adding the attribute-agnostic loss (L^{r_*}) brought significant improvement for the performance of all the other three attributes. Finally, when the attribute information were available (L^{r_1}, L^{r_2} and L^{r_3}), MaSS was able to boost the preservation performance to near perfection without reducing the suppression quality on SpeakerID.

Table 4 shows the performance of MaSS on the video dataset. With only suppression (L_{rec} and L^{s_1}), MaSS indeed lowered the cMAP on both VISPR and PA-HMDB. However, the performance on the action attribute also dropped significantly. By adding the attribute-agnostic loss (L^{r_*}), MaSS was able to significantly improve the accuracy of the action attribute without utilizing any label information. We do notice a slight decrease in the performance of other attributes, which also aligns with the observation in SPAct (Dave et al., 2022), When the action label (L^{r_1}) was incorporated, MaSS was able to preserve the same action accuracy while suppressing the other attributes.

All experimental results above show that MaSS is highly configurable and effective even without knowing the downstream tasks in advance. Moreover, when the information of downstream task is available, MaSS can boost the performance on the corresponding specific attributes.

4.3 Comparison to Other Methods

As MaSS includes both suppression and preservation, there is limited prior work for direct comparison since most of them focus on suppression only. Therefore, we first compared MaSS with the heuristic methods, e.g., perturbation on the original data, additive noise in the feature space for both the Adience and AudioMNIST datasets. For Adience, we further compared MaSS to CIA-GAN (Maximov et al., 2020) even though it only performed suppression. Lastly, we compared MaSS to SPAct on PA-HMDB. More details on the comparisons can be found in Appx. D.

Table 5 shows the comparison on Adience. As most methods were designed to suppress DataID, we configured MaSS to suppress DataID and preserve the others with attribute-agnostic setting for comparison. First, adding noise in the feature domain lowered the accuracy on the DataID attribute but it also degraded the performance on age and gender. On the other hand, for all the methods which manipulate data in the original domain, including Gaussian blurring with different kernel sizes and standard deviations, downsampling and upsampling back to the original size, and obfuscating the various face area where the face is detected by MTCNN (Schroff et al., 2015), they led to similar results since they did not learn what features should be preserved for downstream

Table 5: Comparison to other methods on Adience. Top-1 accuracy (%) is reported. Features and Raw Data denote that the modifications are made on the feature level and the image level, respectively.

Method		DataID (↓)	Age (↑)	Gender (↑)
Gaussian Noise	Features, $\sigma = 0.5$	50.8	56.8	88.1
	Features, $\sigma = 1.0$	0.5	26.8	54.9
Guassian Blur	Raw Data, $k = 11$, $\sigma = 10.0$	39.4	49.0	86.5
	Raw Data, $k = 21$, $\sigma = 15.0$	1.0	19.3	66.2
Downsample	Raw Data, 8×	11.0	30.3	78.7
	Raw Data, 4×	79.3	74.7	94.3
Obfuscation	Raw Data, Face area, ratio = 1.0	0.8	17.1	59.5
	Raw Data, Face area, ratio = 0.36	4.0	29.8	81.6
	Raw Data, Face area, ratio = 0.09	57.7	64.1	93.9
CIAGAN		1.1	17.8	66.9
MaSS	$L_{rec} + L^{s_1} + L^{r_*}$	0.6	78.5	95.7

Table 6: Comparison with other methods on AudioMNIST. Top-1 accuracy (%) is reported.

Method	SpeakerID (↓)	Digit (↑)	Accent (↑)	Gender (↑)
White Noise (Raw Data)	4.7	23.3	23.4	32.2
Masking (Raw Data)	1.8	10.6	3.9	80.0
MaSS ($L_{rec} + L^{s_1} + L^{r_*}$)	1.7	67.4	68.7	88.2

tasks. By modifying faces in the image space, CIAGAN (Maximov et al., 2020) was able to lower the performance on DataID again. However, it still wasn't capable of preserving any features for the other attributes. In contrast, MaSS not only suppressed DataID but also preserved much more data utility such that the data could still perform well on age and gender detection tasks, even though no additional information on age or gender was made available to MaSS.

The comparative study results on AudioMNIST are listed in Table 6. The heuristic approaches were capable of suppressing SpeakerID, as well as the other attributes indiscriminatively. Nonetheless, MaSS can retain the other attributes without knowing the downstream tasks while still achieving SpeakerID suppression. Lastly, comparing to SPAct (Dave et al., 2022) on PA-HMDB[1], MaSS achieved competitive suppression ratio of cMAP on both datasets (VISPR: 55.6% vs. 57% and PA-HMDB: 20.7% vs. 16%) for the other 5 non-action attributes. In addition, MaSS also provides the flexibility to configure which attributes to suppress while SPAct can only target all five attributes at the same time.

Additional ablation studies of MaSS, including suppressing different attributes, effects of loss weights and different similarity measurements, can be found in Appx. C.

5 Conclusion

In this paper, we proposed MaSS to selectively suppress attributes while preserving other attributes; moreover, with the proposed attribute-agnostic approach, the preservation can be achieved without foreseeing the downstream tasks, which expands the usability of the proposed algorithm. We validated our method in three datasets in different domains, including facial images, voice audio and video clips, and all results showed that our method is promising in attribute suppression and preservation. Because we validated MaSS in the feature space rather than original data space, an interesting future direction is to integrate with generative models, e.g., GAN or diffusion models, to convert the modified feature vectors into their original representation (image, audio, etc.). This is beyond the scope of our paper and we leave this as a future work.

[1]We compared SPAct based on their Table 8 in the paper, which is the closest to our setting. Because the baseline performance is different, we show here the relative suppression results.

Code of Ethics and Ethics statement. Our work is to selectively suppress the attributes in multi-attribute data while preserving their potential utility such that the sensitive information in the data could be minimized and hence the data could be used to benefit the community. We believe that there are no ethical concerns related to this work.

Reproducibility Statement. We provided the training and evaluation details in the main paper and appendix. Our source codes and models will be publicly available to help better understand the settings of training and evaluation.

REFERENCES

Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10718–10727, June 2022.

Jia-Wei Chen, Li-Ju Chen, Chia-Mu Yu, and Chun-Shien Lu. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6478–6487, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5378–5396. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/dong22c.html.

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

Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10356–10366, 2019. doi: 10.1109/CVPR.2019.01061.

Wikimedia Foundation. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news, 2019. URL http://www.statmt.org/wmt19/translation-task.html.

Abigail Goldsteen, Gilad Ezov, Ron Shmelkin, Micha Moffie, and Ariel Farkash. Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, pp. 1–15, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu (eds.), *Advances in Visual Computing*, pp. 565–578, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33720-9.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673, 2020. https://github.com/facebookresearch/libri-light.

Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

Jingzhi Li, Lutong Han, Ruoyu Chen, Hua Zhang, Bing Han, Lili Wang, and Xiaochun Cao. Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3891–3899, 2021.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

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

Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Bill Moriarty, Chun-Fu Richard Chen, Shaohan Hu, Sean Moran, Marco Pistoia, Vincenzo Piuri, and Pierangela Samarati. Utility-preserving biometric information anonymization. In *European Symposium on Research in Computer Security*, pp. 24–41. Springer, 2022.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*, pp. 3686–3695, 2017.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6964–6974, June 2021.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12196–12205, June 2022.

Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=mSAKhLYLSsl.