Deep Fair Clustering for Visual Learning

Peizhao Li
Brandeis University
peizhaoli@brandeis.edu

Han Zhao
Carnegie Mellon University
han.zhao@cs.cmu.edu

Hongfu Liu
Brandeis University
hongfuliu@brandeis.edu

Abstract

Fair clustering aims to hide sensitive attributes during data partition by balancing the distribution of protected subgroups in each cluster. Existing work attempts to address this problem by reducing it to a classical balanced clustering with a constraint on the proportion of protected subgroups of the input space. However, the input space may limit the clustering performance, and so far only low-dimensional datasets have been considered. In light of these limitations, in this paper, we propose Deep Fair Clustering (DFC) to learn fair and clustering-favorable representations for clustering simultaneously. Our approach could effectively filter out sensitive attributes from representations, and also lead to representations that are amenable for the following cluster analysis. Theoretically, we show that our fairness constraint in DFC will not incur much loss in terms of several clustering metrics. Empirically, we provide extensive experimental demonstrations on four visual datasets to corroborate the superior performance of the proposed approach over existing fair clustering and deep clustering methods on both cluster validity and fairness criterion.

1. Introduction

With the prevalence of machine learning based decision support systems in high-stakes domains, including college admission, loan approval, bail/parole judgments, and hiring processes, it has already become an imperative object of study to guarantee that individuals are treated equally in such automated decision processes [9, 8, 15]. The area of algorithmic fairness is precisely devoted to the study of algorithms and learning systems to ensure fairness in the execution of these processes while at the same time preserving utility as much as possible. Consider race as a sensitive attribute in visual applications such as facial recognition. In this case, a learning system may have a high recognition accuracy for one group of race, but a considerably lower accuracy for another group, due to the potential imbalance distributions between these two protected subgroups [4]. In order to mitigate such biases in recognition performance, fair and accurate recognition systems are nowadays particularly desirable to serve for many real-world scenarios.

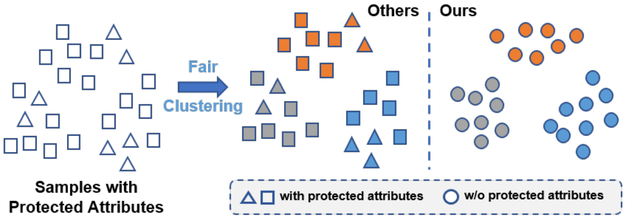


Figure 1. Comparison of existing fair clustering methods and ours. Existing methods seek to ensure that the proportion of samples with different sensitive attributes are approximately equal in each cluster. As a comparison, deep fair clustering learns fair representations that are amenable for cluster analysis under the constraint that they are statistically independent of sensitive attributes.

Cluster analysis is a classic and widely used technique in unsupervised learning to discover the underlying structure of data. In the context of algorithmic fairness, fair clustering aims to hide sensitive attributes during data partition by balancing the distribution of protected subgroups in each cluster. Chierichettiss *et al.* [7] propose a pioneering fairlet by employing a pre-processing technique to partition original data into chunks, followed by a k-center based algorithm, which encourages clusters with balanced demographic groups. In light of the expensive computation involved in fairlet, Backurs *et al.* [2] provide a scalable fair clustering algorithm with approximate fairlet decomposition that runs in nearly linear time. Wang and Davidason [31] extend the single sensitive attribute setting to the one with multi-state protected variable. Additionally, Suman *et al.* [3] propose to transform any given clustering solution to a fair one with multiple protected subgroups. Many other work also exist that are based on the classical clustering algorithm with fair constraints [39, 27, 18, 22].

Despite the abundant existing literature on fair clustering, most of them cast it as a classical balanced clustering problem [25] with equal proportions of protected subgroups in each cluster as a constraint. Furthermore, existing work mostly focuses on empirical validations with low-dimensional input data, and their performance on large-scale and high-dimensional visual data has not been fully understood, especially in the settings where sensitive attributes are deeply embodied in images that act as a background or entity style. Often the case for visual data, the original input space does not contain high-level and abstract representations that can be used to discover the intrinsic and underlying clustering structure for cluster analysis.

Motivated by the aforementioned limitations of existing fair clustering methods, in this paper, we propose Deep Fair Clustering (DFC) to learn both fair and effective representations that are also amenable for cluster analysis. We summarize main contributions in this paper as follows.

- We consider fair clustering in the context of deep representation learning for large-scale and high-dimensional visual learning. To the best of our knowledge, this important setting has not been studied so far. Towards the cluster validity and fairness in clustering results, our approach seeks to find feature mappings that are amenable for structure discovery and can simultaneously filter out the sensitive attributes.
- We propose a minimax optimization formulation to encourage cluster analysis to be independent of sensitive attributes. Furthermore, we provide a theoretical analysis showing that the fairness constraint in DFC does not cause much utility loss according to cluster validity, which is a desirable property to have.
- We evaluate DFC on four real-world visual datasets. Extensive experimental results demonstrate the superior performance of the proposed method over several competitors in terms of both cluster validity and fairness measurement. Additionally, we present indepth explorations, including representation visualization, fairness evaluation, and parameter analysis, to help better understand our fair clustering method.

2. Related Work

We discuss related work on fair machine learning and recent clustering methods, and specifically on fair clustering and deep clustering, which are most relevant to our work.

As argued by a series of related work [9, 8, 15], many existing learning systems suffer from unfair issues in terms of sensitive attributes (*e.g.*, sex, race, gender), due to the bias in data and models, and it is hence urgent to mitigate the inadvertently encoded bias and unfairness in these systems. Recently, fairness in machine learning has been widely discussed under different problem settings [10]. Seeking fair representations and eliminating the effect caused by sensitive attributes are requisite tasks to achieve fairness in learning systems. Of course, fairness is often at odds with utility maximization [38], so the goal here is to find a good balance between these two. One line of work to achieve this goal is through learning fair representations using adversarial training techniques [24, 33, 37], where the representations are learned to filter out information related to sensitive attributes. This line of work mostly focus on the supervised learning setting where the utility is often characterized by the accuracy of target tasks. In unsupervised learning, in particular in cluster analysis, Chierichetti *et al.* [7] pioneer this novel concept by embedding fairness constraints and encouraging clusters to have balanced sensitive demographic subgroups. They propose fairlet by employing a pre-processing technique to first partition the original data into chunks, followed by a k-center algorithm to achieve fairness. Schmidt *et al.* [29] extend fairlet for the object of k-means clustering. However, the fairlet procedure used in these algorithms requires at least quadratic running time and cannot scale to large datasets. To this end, Backurs *et al.* [2] propose a tree metric based near-linear time algorithm for fairlet decomposition. Beyond fairlet, Bera *et al.* [3] solve the vanilla clustering problem first and then make a fair assignment, which transforms a wide range of clustering objectives to their fair solutions. Ziko *et al.* [39] derive a KL-based fairness penalty and impose it on any desired demographic proportions within each cluster. Subsequently, fair k-center algorithm for multiple groups is proposed [27], and Kleindessner *et al.* [18] also introduce fair constraints into spectral clustering. More recently, Wang and Davidason [31] extend to multiple protected subgroups and incorporate neural network to learn a discriminative but fair assignment function. Although existing work are pioneering, they sacrifice clustering quality in pursuit of fairness. Moreover, clustering on large-scale high-dimensional visual data under the fairness constraint remains unexplored, and our work makes a step forward in this direction.

Deep clustering aims to learn clustering-favorable representations for complex intrinsic structure discovery. Towards clustering objective functions, deep clustering brings significant performance improvements via learnable feature transformation. Xie *et al.* [32] train an encoder with the KL-divergence term to make assignments confidently and prevent large clusters and feature collapse. Similarly, Yang *et al.* [34] learn a subspace through deep learning and clustering simultaneously. One step further, Caron *et al.* [5] extend the pipeline to an end-to-end training on large scale visual datasets and group features with a standard clustering algorithm. Other interesting work continually bring novel insights to this line of research, including pair-wise similarity deep clustering [6], deep learning with spectral clustering [20], minimizing relative entropy [11], deep transfer clustering [13], and deep clustering with generative model [26], adversarial learning [30], and Gaussian mixture model [35]. However, fairness in deep clustering has not been well addressed so far. In order to get a desirable and fair partition, it is essential to design applicable fair constraints for deep clustering training so as to hide sensitive attributes for clustering assignments.

3. Problem Definition and Motivations

In this paper, we focus on partitional cluster analysis that aims to separate a set of data points X with categorical sensitive attribute G into K disjoint subgroups, where all the data points are sampled i.i.d. from an underlying distribution $\mathcal{D}$, and $G = G(X)$ takes value in $[M]$[1]. Formally, let $C = C(X)$ be the clustering assignment given by a (randomized) clustering algorithm C applied to data X. As a motivating example, in social network analysis where each data point corresponds to a vectorial description of a user, the sensitive attribute G could be the race or gender of an user. In this context, the goal of fair clustering is to partition data points into K disjoint clusters while simultaneously ensures that sensitive attribute has no influence on the partitioning result, *i.e.*, $C(X)$ is statistically independent of G. In other words, the sensitive attribute G has no influence on the clustering result given by the algorithm C.

Note that our definition of fairness in the context of cluster analysis is stronger than that of the existing work on this topic [3, 39, 31], where the goal is usually to ensure that the proportion of the protected subgroup membership G is approximately equal in all the clusters. To see this, note that when G is a categorical variable, under the condition that $C(X)$ is independent of G, we have:

$$\mathbb{E}_{X\sim\mathcal{D}}[G \mid C(X) = c] = \mathbb{E}_{X\sim\mathcal{D}}[G], \ \forall c \in [K] \ , \qquad (1)$$

which shows our fairness criterion implies the one used in several of the existing fair clustering work.

On the other hand, the quality of clustering is often sensitive to the metric space in which the analysis is conducted. Specifically, for high-dimensional visual data, it is often not desirable to directly conduct cluster analysis based on the pixel-level distance between input images. Recent advances in representation learning [21, 19, 14, 16] have demonstrated that deep neural networks are capable of extracting hierarchical features that can often facilitate large-scale classification, segmentation and detection tasks. Motivated by these observations, in this work we propose a clustering algorithm that can leverage the rich representations given by deep neural nets for clustering and simultaneously being fair w.r.t. a given sensitive attribute in a strong sense.

4. Deep Fair Clustering

In this section we propose deep fair clustering, where the fair and clustering-favorable representations can be obtained by a unified framework. The goal is to learn feature representations that are not only free of sensitive attributes, but also are favorable for the following cluster analysis.

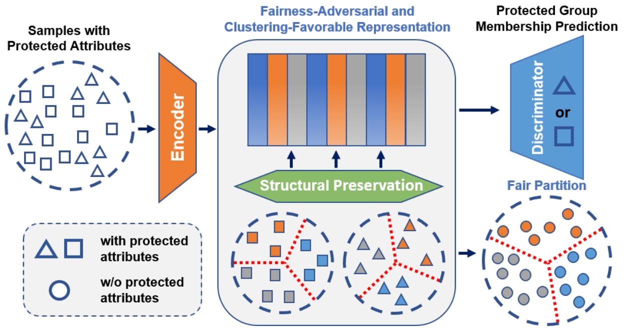

Samples with Protected Attributes
Encoder
Fairness-Adversarial and Clustering-Favorable Representation
Structural Preservation
Protected Group Membership Prediction
Discriminator
or
Fair Partition
with protected attributes
w/o protected attributes

Figure 2. Overview of Deep Fair Clustering. The proposed method incorporates adversarial fairness to achieve group invariant cluster assignment and structural preservation to deliver partition with guaranteed cluster validity.

4.1. Framework Overview

Figure 2 shows the overall framework of the proposed Deep Fair Clustering (DFC) for fair and clustering-favorable representation learning. To generate fair representations, DFC uses the min-max game strategy with the encoder producing the expected representation and the discriminator predicting protected subgroup membership. Fair representations are achieved when the discriminator cannot distinguish between representations from different protected subgroups. However, this is not enough to guarantee the representation is effective for data partition. Due to the unsupervised nature of cluster analysis, we propose a self-supervised structural preservation objective to guide representation learning. We assume that if a pair of data points in one protected subgroup is similar to each other, then they should be assigned the same cluster label in fair clustering as well. To achieve this, we separately conduct the clustering algorithm on each protected subgroup and obtain the pseudo soft assignments to guide deep fair clustering.

4.2. Objective Function

Let $\mathcal{D}$, $\mathcal{F}$, and $\mathcal{A}$ represent the functions of discriminator, encoder, and cluster assignment, respectively. We have the hidden representation $Z = \mathcal{F}(X)$. The assignment $\mathcal{A} : \mathcal{Z} \to \mathbb{R}^{N\times K}$ is a kernel function, and the soft assignment is obtained by $P = \mathcal{A}(Z)$. $\mathcal{D} : P \to \mathbb{R}^{N\times M}$ is the protected subgroup membership discriminator. The objective function of the proposed Deep Fair Clustering consists of three parts, fairness-adversarial loss, structural preservation loss, and clustering regularizer. In what follows, we provide details on these three separate terms and the clustering label assignment strategy for the proposed model.

Fairness-Adversarial Loss. When the sensitive attribute G is a categorical variable, we aim to achieve our fairness definition by encouraging the partition to be statistically independent of each sensitive attribute. To achieve this goal, we expect the distribution of soft assignments for samples belonging to different protected subgroups to

[1] We use $[M]$ to denote the set $\{1, \ldots, M\}$.

be close, *i.e.*, to minimize the distribution divergence on cluster assignments between different subgroups. By reducing the divergence, the partitioning across all subgroups approach to the same, so that DFC enables each cluster to hold equivalent fractions of samples from different protected subgroups and executes cluster analysis unrelated to sensitive attributes. We implement a method based on the adversarial training technique as it delivers favorable performance in both learning and approaching a given distribution [12]. The fairness-adversarial loss can be written as:

$$\mathcal{L}_f := \mathcal{L}(\mathcal{D} \circ \mathcal{A} \circ \mathcal{F}(X),\ G) \ , \tag{2}$$

where $\mathcal{L}$ is the cross-entropy loss function, and $\circ$ denotes the function composition. Here we maximize $\mathcal{D}$ with the probability of assigning the correct membership label for each sample, and expect assignments obtained by $\mathcal{F}$ and $\mathcal{A}$ to confuse $\mathcal{D}$ simultaneously. $\mathcal{F}$, $\mathcal{A}$ and $\mathcal{D}$ conduct the minimax game for fairness-adversarial loss that ensures partition distributions over all the subgroups similar to each other.

Structural Preservation Loss. The above fairness-adversarial term encourages invariant soft assignments *w.r.t.* different protected subgroups and leads to a model that is invariant to sensitive attributes. However, there is a degenerate solution where the representation function $\mathcal{F}$ reduces to a constant function if we solely optimize $\mathcal{L}_f$. Such constant mapping will destroy the existing structure for cluster analysis in the original data. To this end, in order to capture clustering-favorable representation under the constraint of fairness and unsupervised nature, we move a step further towards the goal of preserving the clustering structure inside each subgroup. Specifically, we expect the subsets of deep fair clustering for each protected subgroup hold similar cluster structures with the partition individually on each protected subgroup. From this point of view, we propose a structural preservation loss $\mathcal{L}_s$ as follows:

$$\mathcal{L}_s := \sum_{g \in [M]} \left\| \hat{P}_g \hat{P}_g^{\top} - P_g P_g^{\top} \right\|^2 \ , \tag{3}$$

where $\hat{P}_g$ and P_g are the soft assignment on g-th protected subgroup in the individual partition and deep fair clustering.

Clustering Regularizer. Inspired by other deep clustering [5, 6], we employ a clustering regularizer to strengthen the predicted confidence and prevent large clusters. Different from [32], we carry out the clustering regularizer on each protected subgroup to avoid that one cluster is dominated by instances with the same sensitive attribute. The auxiliary target distribution Q is firstly calculated by:

$$q_k = \frac{(p_k)^2 / \sum_{x \in X_g} p_k}{\sum_{k' \in [K]} ((p_{k'})^2 / \sum_{x \in X_g} p_{k'})} \ , \tag{4}$$

where p_k is the probability of sample x assign to k-th cluster, and X_g is the subgroup that the instance belong to. The clustering constrained loss is KL divergence between soft assignment and auxiliary target distribution:

$$\mathcal{L}_c := KL(P \,||\, Q) = \sum_{g \in [M]} \sum_{x \in X_g} \sum_{k \in [K]} p_k \log \frac{p_k}{q_k} \ . \tag{5}$$

Clustering Assignment. We follow [32] to use Student t-distribution for assignment, which can be calculated as:

$$p_k = \frac{(1 + \|z - c_k\|^2 / \alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k' \in [K]} (1 + \|z - c_{k'}\|^2 / \alpha)^{-\frac{\alpha+1}{2}}} \ , \tag{6}$$

where α is the degree of freedom of Student's t-distribution. p_k represents the assigned confidence between representation z and cluster c_k. To sum up, the overall objective function for the proposed deep fair clustering can be written as the following minimax saddle point problem:

$$\max_{\mathcal{F}, \mathcal{A}} \quad \alpha_f \mathcal{L}_f - \alpha_s \mathcal{L}_s - \mathcal{L}_c \ , \tag{7}$$

$$\min_{\mathcal{D}} \quad \alpha_f \mathcal{L}_f \ , \tag{8}$$

where α_f and α_s are trade-off hyper-parameters.

4.3. Theoretical Analysis

In the above, we use the adversarial fairness term as a regularizer to find clusters that are invariant w.r.t. the sensitive attribute. In this subsection, we provide formal analysis on the impact of fairness constraint on the clustering quality. In the limit of perfectly fair clustering, the quality of clustering results on the whole population is *almost as good as* that on each protected subgroup. In a nutshell, this result implies that in a sense our algorithm does not incur extra quality loss under our fairness constraint.

Assume that each data point x has a corresponding external label $y \in \mathcal{Y}$. For example, in the case of digit clustering, the external label y corresponds to the digit label of the input image. To this end, we use variable Y to denote the external label of X. Given a clustering assignment $C(X)$, this defines a joint distribution over Y and $C(X)$ from which we use the mutual information $I(Y; C(X))$ as a measure of the quality of clustering algorithm. It is easy to verify that $I(Y; C(X))$ is invariant under label shuffling, which makes it a particularly suited measure in cluster analysis. Intuitively, the higher the mutual information $I(Y; C(X))$, the more aligned the clustering result is with the structure given by the external label. Based on these notations, we have the main theorem of deep fair clustering written as follows [2]:

Theorem 4.1. Let $X \sim \mathcal{D}$ be the input random variable of the clustering algorithm C. If the clustering assignment $C(X)$ is independent of the sensitive attribute G, then

$$I(Y; C(X)) \leq \sum_{g \in [M]} \Pr(G = g) \cdot I(Y_g; C(X_g)), \tag{9}$$

[2] The detailed proof is provided in the supplementary material.

Table 1. Characteristics of four datasets used in experiments. Due to the discrepancy in the ratio of the number of instances in different protected subgroups and the number of classes, the maximum value of *Balance* and *Entropy* are different across four datasets.

Dataset	Type	# Instances	# Classes	# Dim.	Sensitive Attribute	Max of Bal.	Max of Ent.
MNIST-USPS	digital	67,291	10	1,024	source of digital	0.120	2.303
Color Reserve MNIST	digital	120,000	10	1,024	original or reversed	1.000	2.303
MTFL	face	2,000	2	50,176	with or w/o glasses	1.000	0.693
Office-31(A&W)	object	3,612	31	50,176	domain source	0.273	3.434

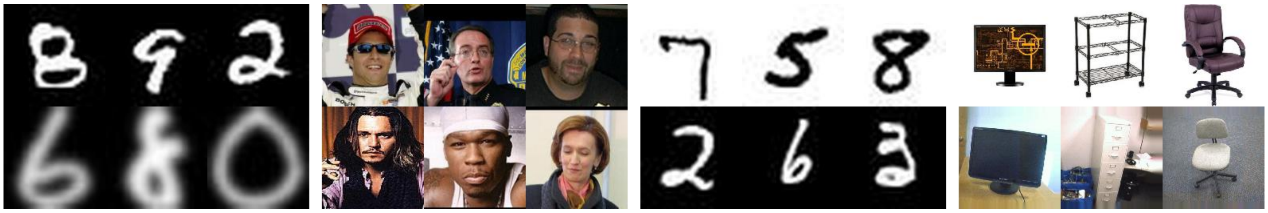


Figure 3. Visual examples in four datasets. From left to right column, three columns represent one dataset: MNIST-USPS, MTFL, MNIST, and Office-31. The upper and lower row present images with different protected attributes: from MNIST or from USPS; with glasses or without glasses; reversed one or the original one; select from *Amazon* or select from *Webcam*.

where X_g is the input random variable in the g-th protected subgroup and Y_g is the corresponding external label of X_g.

Realize that without any assumption, the mutual information is neither convex nor concave in terms of the joint distribution over $C(X)$ and Y. As a comparison, under our fairness constraint, i.e., clustering assignment $C(X)$ is independent of the sensitive attribute G, then Theorem 4.1 claims that the mutual information $I(C(X);Y)$ becomes convex, which implies that the overall clustering quality can be bounded by a convex combination of the individual clustering quality when applying the same algorithm $C(\cdot)$ to each protected subgroup. Similarly, if the entropy of cluster assignment is chosen to measure the balance of clustering (as one of the measurements we conduct in experiments), we have a nicer proposition as follows:

Proposition 4.1. Let $X \sim \mathcal{D}$ be the input random variable of the clustering algorithm C. If the clustering assignment $C(X)$ is independent of the sensitive attribute G, then

$$H(C(X)) = \sum_{g \in [M]} \Pr(G = g) \cdot H(C(X_g)). \tag{10}$$

In words, Proposition 4.1 shows that under the fairness constraint, if the clusters are well-balanced in size when we apply the same algorithm $C(\cdot)$ individually on each demographic subgroup, the clusters will also be balanced if instead we use all the data for clustering.

5. Experimental Analysis

In this section, we provide extensive experimental results on four datasets. Recent methods of deep clustering and fair clustering are used for comparison to demonstrate the superior of deep fair clustering in terms of cluster validity and fairness measurement, respectively. Finally, we provide in-depth explorations for our method in four aspects.

5.1. Setup for Experiments

Dataset. Four public datasets are simulated to evaluate deep fair clustering. Some characteristics and visual examples of these four datasets are shown in Table 1 and Figure 3. (1) *MNIST-USPS*. We construct *MNIST-USPS* dataset using all the training digital samples from MNIST[3] and USPS[4] dataset, and set the sample source as the protected attribute (MNIST or USPS). (2) *Color Reverse MNIST*. We reverse images from MNIST via *pixel*=255−*pixel* and concatenate original images to build this dataset. Here the sample source from reversed or original one is set as the protected attribute. (3) *MTFL*. Multi-task Facial Landmark (MTFL) dataset [36] consists of 12,995 images used for facial recognition and landmarks detection. It also labels every image with gender, smiling, with or w/o glasses, and the degree of the pose. We set the glasses-wearing as the protected attribute and measure the clustering validity toward gender. We randomly sample 1,000 images with and without glasses to build the balanced dataset for fair clustering. (4) *Office-31*. This dataset [28] is originally used for domain adaptation. It consists of images with 31 different categories collected from three distinct domains: *Amazon*, *Webcam*, and *DSLR*. Each domain contains all the categories but with different shooting angles, lighting conditions, or the form of presentation, *etc*. We select *Amazon* and *Webcam* for our experiments because these two has the largest domain divergence, and set the domain source as the protected attribute.

Implementation. We employ convolutional variational autoencoder for *MNIST-USPS* and *Color Reverse MNIST* datasets. We pretrain the autoencoder with sample self-construction and use the encoder for fair feature learning. The encoder consists of four convolutional layers followed with batch normalization and non-linear activation. The

[3] http://yann.lecun.com/exdb/mnist/

[4] https://www.kaggle.com/bistaumanga/usps-dataset

Table 2. Quantitative results on *MNIST-USPS* and *Color Reverse MNIST* datasets in terms of cluster validity and fairness.

MNIST-USPS

Method	Acc	NMI	Balance	Entropy
AE [17]	0.624	0.504	0.017	2.294 / 2.053
DEC [32]	0.586	0.686	0.000	2.082 / 1.735
DAC [6]	**0.757**	**0.703**	0.000	2.188 / 0.000
ClGAN [26]	0.343	0.212	0.000	2.096 / 1.912
ScFC [2]	0.176	0.053	**0.120**	2.219 / 2.219
SpFC [18]	0.162	0.048	0.000	2.196 / 1.902
FCC [39]	0.560	0.461	0.032	2.288 / 2.193
FAlg [3]	0.621	0.496	**0.093**	**2.295** / **2.227**
Ours	**0.825**	**0.789**	0.067	**2.301** / **2.265**

Color Reverse MNIST

Method	Acc	NMI	Balance	Entropy
AE [17]	0.357	0.352	0.005	1.850 / 1.645
DEC [32]	**0.401**	**0.480**	0.000	1.774 / 1.384
DAC [6]	0.309	0.193	0.623	**2.291** / **2.289**
ClGAN [26]	0.166	0.034	0.000	0.669 / 1.995
ScFC [2]	0.268	0.105	**1.000**	2.277 / 2.277
SpFC [18]	0.137	0.020	0.000	1.767 / 2.035
FCC [39]	0.349	0.295	0.021	1.904 / 1.638
FAlg [3]	0.295	0.206	0.667	2.253 / 2.287
Ours	**0.577**	**0.679**	**0.763**	**2.294** / **2.301**

decoder is the same but reversed architecture. For *MTFL* and *Office-31*, considering their complex and high resolution images, we use ImageNet-pretrained ResNet50 [14] as the feature extractor. We resize images to 32 × 32 and set batch size as $b = 512$ for two digital datasets, and 224×224 with batch size $b = 128$ for the latter two. We set trade-off hyper-parameters in overall objective function default to 1. For *Office-31*, to ensure the model convergence, we calculate $\alpha_s{=}(512/128)^2{\cdot}(31/10)$ with respect to the number of clusters and batch size in *MNIST-USPS*. We set initial learning rate $lr_{init}{=}1e-4$ for Adam optimizer, then gradually adjust learning rate by $lr = lr_{init}(1+10t)^{-0.75}$, where t is the training process changed from 0 to 1 linearly, and train till its convergence. The learning rate for the discriminative classifier is 10 times the learning rate for encoder and centroids. α in Student's t-distribution is set to 1.

Competitive Methods. Except feature extraction conjunct with K-means [1], we include three deep clustering methods, Deep Embedding Clustering [32], Deep Adaptive Clustering [6], and ClusterGAN [26], and four fair clustering algorithms, Scalable Fair Clustering [2], Spectral Fair Clustering [18], Fairness Constraints Clustering [39], and Fair Algorithm for Clustering [3] for comparison, thanks to their open-source codes. Notably, we use the same encoder architecture for all comparable methods for feature extraction to achieve fair comparison. For Spectral Fair Clustering, we construct adjacency weight graph by Gaussian kernel, and randomly select $5,000$ samples from *MNIST-USPS* and *Color Reverse MNIST* datasets for the experiment, respectively, because the original size of adjacency matrix is exceeded 32GB the maximum of experimental memory.

Table 3. Quantitative results on *MTFL* and *Office-31* dataset in terms of cluster validity and fairness.

MTFL

Method	Acc	NMI	Balance	Entropy
ResNet50 [14]	0.648	0.176	0.406	**0.693** / 0.509
DEC [32]	0.520	0.030	0.711	0.660 / 0.576
DAC [6]	0.563	0.002	0.950	0.665 / 0.673
ClGAN [26]	**0.727**	0.161	0.490	0.600 / **0.675**
ScFC [2]	0.627	0.030	**1.000**	0.666 / 0.576
SpFC [18]	0.565	0.040	0.836	0.683 / 0.652
FCC [39]	0.658	0.174	0.531	**0.693** / 0.578
FAlg [3]	0.660	**0.181**	0.666	0.689 / 0.613
Ours	**0.719**	**0.190**	**0.986**	**0.693** / **0.693**

Office-31

Method	Acc	NMI	Balance	Entropy
ResNet50 [14]	0.641	0.691	0.000	3.299 / 2.775
DEC [32]	0.546	0.604	0.000	3.063 / 2.937
DAC [6]	0.063	0.041	0.054	1.199 / 0.918
ClGAN [26]	0.516	0.536	0.000	3.341 / 3.079
ScFC [2]	0.090	0.056	**0.273**	3.248 / 3.248
SpFC [18]	0.096	0.109	0.000	3.260 / 3.055
FCC [39]	0.652	0.693	0.173	**3.401** / **3.400**
FAlg [3]	**0.689**	**0.713**	**0.196**	3.343 / 3.337
Ours	**0.692**	**0.718**	0.117	**3.422** / **3.403**

Metrics. We conduct four measurements in experiments, where *Accuracy* and *Normalized Mutual Information* (*NMI*) evaluate the cluster validity by comparing the obtained partition and external ground true labels. Balance and Entropy measure the balancing degree of discovered clusters towards the protected attribute in each dataset. These four measurements are calculated as follows:

$$\begin{aligned}
\textit{Accuracy} &= \frac{\sum_{i=1}^{n} \mathbb{1}_{y_i=\mathrm{map}(\hat{y}_i)}}{n}, \\
\textit{NMI} &= \frac{\sum_{i,j} n_{ij} \log \frac{n \cdot n_{ij}}{n_{i+} \cdot n_{+j}}}{\sqrt{(\sum_i n_{i+} \log \frac{n_{i+}}{n})(\sum_j n_{j+} \log \frac{n_{+j}}{n})}}, \\
\textit{Balance} &= \min_i \frac{\min_g |\mathcal{C}_i \cap X_g|}{n_{i+}}, \\
\textit{Entropy} &= -\sum_i \frac{|\mathcal{C}_i \cap X_g|}{n_{i+}} \log \frac{|\mathcal{C}_i \cap X_g|}{n_{i+}},
\end{aligned}$$

where $\mathbb{1}$ is the indicator function, and $\mathrm{map}(\hat{y}_i)$ is permutation mapping function that maps each cluster label $\hat{y}_i$ to the ground truth label y_i, $\mathcal{C}_i$ and X_g denote the i-th cluster and the g-th protected subgroup, n_{ij}, n_{i+} and n_{+j} represent the co-occurrence number and cluster size of i-th and j-th clusters in the obtained partition and ground truth, respectively, and n is the total data instance number.

Accuracy and *NMI* are two positive metrics to evaluate the cluster quality in terms of classification and information theory, where a larger value indicates better performance. *Balance* measures the minimum ratio of numbers of samples from different protected subgroups across all the clusters, and the upper bound for *Balance* is determined by the

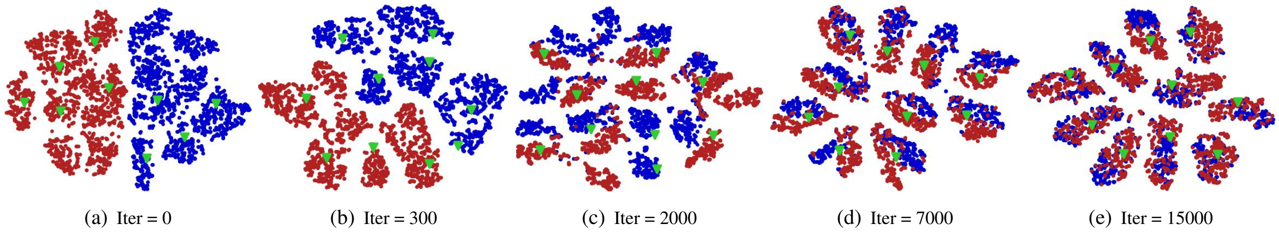


Figure 4. Feature space visualization on *Color Reverse MNIST*. Samples from different protected subgroups are colored in either red or blue. Green inverted triangles represent centroids. With the training goes on, features perform clustering structures, and each cluster contains features with different colors. The final feature space is fair to protected attributes and becomes clustering-favorable as well.

given data distribution. During clustering, one protected subgroup might get empty in one cluster and *Balance* goes to zero, which makes this metric too harsh for fair clustering evaluation. Here we add *Entropy* as a compliment that reflects the distribution of predictions as well as how unfairness the clusters are suffering from. Low entropy reveals that one subgroup of samples only assigns to a few clusters, while other clusters are with great unfairness.

5.2. Quantitative Evaluation

We provide quantitative evaluation in Table 2 and Table 3, where red indicates the highest result, and blue indicates the second highest one. As shown in tables, deep clustering methods achieve better clustering results compare to fair clustering on *Accuracy* and *NMI*. However, these methods are not primarily designed for fair clustering, they perform poorly in terms of *Balance* and *Entropy*. When there exist feature distribution divergences across different protected subgroups, deep clustering methods that only rely on unsupervised trained unfair representations easily partition given data on the basis of protected attributes and lead to unfairness. For fair clustering methods, these methods add fair constraints on original features before or after some clustering algorithm. Although these methods work well on low-dimensional datasets, they fail on high-dimensional large-scale visual data where protected attributes are inserted deeply in images which act as a background or entity style instead of a clear attribute indicator. As pointed out, current fair clustering methods sacrifice the clustering quality to achieve fairness. Note that although the fairlet based method Scalable Fair Clustering always achieves the best and maximum balance value, it performs weakly on clustering validity, and this is meaningless if we only consider fairness and no matter how the quality of partition is. The ultimate aim for fair clustering should be the object that finds an informative as well as a fair partition. It is much difficult to valid on clustering quality if only focus on clustering constraints on representation space without any learnable and clustering information preserved feature transformation. Moreover, these methods only re-assign data points with fair constraint and assume initial centroids are well applicable for fair partition. However, in our experiments, the maintain of centroids requires extensive reassignment to achieve clustering validity. Differently, in deep fair clustering, the moving of centroids along the training process can better preserve clustering structures under fair constraints. In general, deep fair clustering consistently achieves promising results across all four real-world datasets in terms of both clustering quality and fairness.

Table 4. Classification accuracy on protected subgroup prediction with unfair and fair representations. The lower, the better.

	Unfair Representations	Fair Representations
MNIST-USPS	0.995	0.712
Color Reverse MNIST	1.000	0.521
MTFL	0.982	0.919
Office-31	0.997	0.841

5.3. In-depth Exploration

In this subsection, we present exploration inside the deep fair clustering model with regard to four different aspects. We firstly visualize feature transformations during the training process, and show that learned features are clustering favorable and balanced in both. Next, by using an additional classifier, we demonstrate that the trained encoder successfully filter information owned to sensitive attributes within outputted features. Subsequently, guiding by the clustering structure in each protected subgroups, we reveal the worst and best class alignment across protected subgroup and compare them to deep fair clustering. Finally, we conduct experiments with various settings for hyper parameters and hereby verify the effect of each proposed component.

Feature Visualization. We use *Color Reverse MNIST* dataset for visualization due to the balanced sensitive attribute in each cluster via t-SNE [23] along the training process. As shown in Figure 4, the encoder at the very beginning, separates features according to their protected attribute as denotes by the color red and blue. That is because the protected attributes play a dominate role in feature representations initially. Also, the site for centroids leads the partition at that time only reflect protected attribute distribution. As the training goes on with the aim of the proposed objective function, the clustering structure starts to emerge and features with different protected attributes begin to merge. In the end, every cluster contains features that

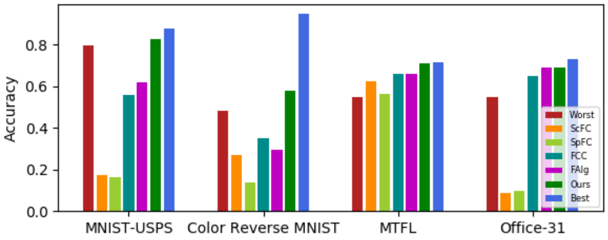
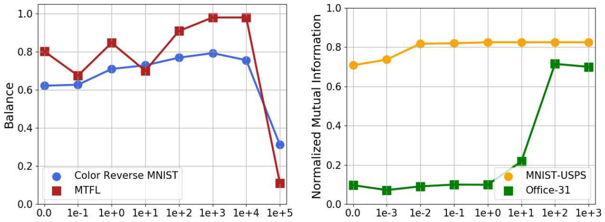


Figure 5. Clustering accuracy comparison on fair clustering with the best & the worst class matching. Deep fair clustering continually outperforms worst matching by a large margin and approach closely to best matching on some datasets.



(a) α_f. Y-axis denotes Balance.

(b) α_s. Y-axis denotes NMI.

Figure 6. Parameter analysis on α_f and α_s. (a) α_s is set as 0 with varied α_f to show how adversarial fairness contributes to balance. (b) α_f is set as 1 with varied α_s to show the enhancement on clustering quality brought by structural preservation.

come from different protected attributes and achieve fairness within each cluster. At the same time, the clustering structures for features are well performed, and centroids are learned to be at the center of each cluster of features, both of which lead features to be clustering-favorable.

Fair Representations. Here we showcase our fair representations is effective not only evaluated by external sensitive attributes, but also validated by the well-trained classifier. Taking original images as input, we firstly train a classifier conjunct an encoder (same architecture with previous experiments for each dataset) jointly from scratch using the membership of different protected subgroups as labels. We report training accuracy in Table 4 on two different representations with the same type of classifier. Results show that membership for each protected subgroup can be well recognized by the classifier with unfair representations as the input. The classifier achieves accuracy close to 100% on four datasets, which means protected attributes can be easily detected by label-oriented representations, and these representations can be seen as an unfair one. Next, to show the fairness in our proposed method, we use representations outputted by the fair encoder obtained by deep fair clustering to serve as input for the classifier, then train the classifier using memberships as labels the same to the previous experiment. With the same procedure the classifier performance on our learned representation drops. The gap demonstrates that representations in deep fair clustering correlate slighter with protected attributes, where even the well-trained classifier is hard to detect different memberships.

Label Matching. Fair clustering problem can also be viewed as a cluster label matching problem, where the complete dataset is split into several protected subgroups by the sensitive attributes. Followed by the partition on each subdatasets, we obtain the final clustering achieved by matching the cluster labels among several partitions. However, we cannot guarantee the cluster across different partition match correctly since there is no guidance in the matching process. When K is large, the matching suffers from a considerable high risk of cluster misalignment. Here we approximate the clustering accuracy of the best and worst matching by the above procedure, where the calculation is illustrated in the supplementary material. Figure 5 shows the clustering accuracy comparison on fair clustering the best and the worst matching performance. For ScPC and SpFC, they, unfortunately, perform even worse than the worst matching with unfair deep representation. This verifies one of our motivations for fair and clustering-favorable representation learning. On the contrary, deep fair clustering consistently surpasses the worst matching and is approximate to the accuracy of best matching on *MNIST-USPS*, *MTFL* and *Office-31*. It demonstrates that deep fair clustering is capable of capturing the similarity between samples across different protected subgroups, and the delivered partition is effective according to both clustering quality and fairness.

Parameter Analysis. It is straightforward to see that adversarial fairness and structural preservation contribute to clustering balance and quality, which are controlled by the hyper-parameter α_f and α_s, respectively. We conduct experiments on several choices of parameters in our objective function and report the corresponding performance in Figure 6. When $\alpha_s = 0$, our model removes the structural preservation term and α_f controls the fair representation learning. A large value of α_f leads to a more balanced partition. On the contrary, α_s plays a crucial role in making the learned representations favorable to partition, where a large value of α_s indicates a high-quality data partition.

6. Conclusion

In this paper, we considered fairness in deep clustering for large-scale high-dimensional visual learning, and proposed a deep fair clustering method to learn a fair and clustering-favorable representation for clustering and simultaneously to hide sensitive attributes. Moreover, we theoretically analyzed the performance guarantee of clustering quality with fairness constraint. Extensive experimental results on four visual datasets demonstrated the superior performance of our proposed methods over the recent fair clustering and deep clustering methods in terms of both the performance of cluster validity and fairness measurement.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007. 6

[2] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of The 36th International Conference on Machine Learning*, 2019. 1, 2, 6

[3] Suman K Bera, Deeparnab Chakrabarty, and Maryam Negahbani. Fair algorithms for clustering. *arXiv preprint arXiv:1901.02393*, 2019. 1, 2, 3, 6

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 1

[5] 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*, 2018. 2, 4

[6] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of The IEEE International Conference on Computer Vision*, 2017. 2, 4, 6

[7] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, 2017. 1, 2

[8] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018. 1, 2

[9] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018. 1, 2

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012. 2

[11] Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of The IEEE International Conference on Computer Vision*, 2017. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2014. 4

[13] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of The IEEE International Conference on Computer Vision*, 2019. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 6

[15] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 2019. 1, 2

[16] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 3

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6

[18] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. In *Proceedings of The 36th International Conference on Machine Learning*, 2019. 1, 2, 6

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 3

[20] Marc T. Law, Raquel Urtasun, and Richard S. Zemel. Deep spectral clustering learning. In *Proceedings of The 34th International Conference on Machine Learning*, 2017. 2

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of The IEEE*, 1998. 3

[22] Tongliang Liu, Dacheng Tao, and Dong Xu. Dimensionality-dependent generalization bounds for k-dimensional coding schemes. *Neural Computation*, 28(10):2213–2249, 2016. PMID: 27391679. 1

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 7

[24] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. 2

[25] Mikko I Malinen and Pasi Fränti. Balanced k-means for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition*, 2014. 1

[26] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks. In *Proceedings of The AAAI Conference on Artificial Intelligence*, 2019. 2, 6

[27] Clemens Rösner and Melanie Schmidt. Privacy Preserving Clustering with Constraints. In *Proceedings of The 45th International Colloquium on Automata, Languages, and Programming*, 2018. 1, 2

[28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of The European Conference on Computer Vision*, 2010. 5

[29] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means clustering. *arXiv preprint arXiv:1812.10854*, 2018. 2

[30] Zhiqiang Tao, Hongfu Liu, Jun Li, Zhaowen Wang, and Yun Fu. Adversarial graph embedding for ensemble clustering. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019. 2

[31] Bokun Wang and Ian Davidson. Towards fair deep clustering with multi-state protected variables. *arXiv preprint arXiv:1901.10053*, 2019. 1, 2, 3

[32] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of International Conference on Machine Learning*, 2016. 2, 4, 6

[33] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *Proceedings of The IEEE International Conference on Big Data*, 2018. 2

[34] Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of The 34th International Conference on Machine Learning*, 2017. 2

[35] Linxiao Yang, Ngai-Man Cheung, Jiaying Li, and Jun Fang. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *Proceedings of The IEEE International Conference on Computer Vision*, 2019. 2

[36] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Proceedings of The European Conference on Computer Vision*, 2014. 5

[37] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019. 2

[38] Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *arXiv preprint arXiv:1906.08386*, 2019. 2

[39] Imtiaz Masud Ziko, Eric Granger, Jing Yuan, and Ismail Ben Ayed. Clustering with fairness constraints: A flexible and scalable approach. *arXiv preprint arXiv:1906.08207*, 2019. 1, 2, 3, 6