Deep Fair Clustering via Maximizing and Minimizing Mutual Information: Theory, Algorithm and Metric

Pengxin Zeng*, Yunfan Li*, Peng Hu, Dezhong Peng, Jiancheng Lv, Xi Peng†

College of Computer Science, Sichuan University, China

{zengpengxin.gm, yunfanli.gm, penghu.ml, pengx.gm}@gmail.com;
{pengdz, lvjiancheng}@scu.edu.cn

Abstract

Fair clustering aims to divide data into distinct clusters while preventing sensitive attributes (e.g., gender, race, RNA sequencing technique) from dominating the clustering. Although a number of works have been conducted and achieved huge success recently, most of them are heuristical, and there lacks a unified theory for algorithm design. In this work, we fill this blank by developing a mutual information theory for deep fair clustering and accordingly designing a novel algorithm, dubbed FCMI. In brief, through maximizing and minimizing mutual information, FCMI is designed to achieve four characteristics highly expected by deep fair clustering, i.e., compact, balanced, and fair clusters, as well as informative features. Besides the contributions to theory and algorithm, another contribution of this work is proposing a novel fair clustering metric built upon information theory as well. Unlike existing evaluation metrics, our metric measures the clustering quality and fairness as a whole instead of separate manner. To verify the effectiveness of the proposed FCMI, we conduct experiments on six benchmarks including a single-cell RNA-seq atlas compared with 11 state-of-the-art methods in terms of five metrics. The code could be accessed from https://pengxi.me.

1. Introduction

Clustering plays an important role in machine learning [19, 27–29, 34, 42, 43], which could partition data into different clusters without any label information. It has been widely used in many real-world applications such as multi-view learning [35, 39], image segmentation [24], and bioinformatics [20]. In practice, however, the data might be confounded with sensitive attributes (*e.g.*, gender, race, etc., also termed as **group information**) that probably over-
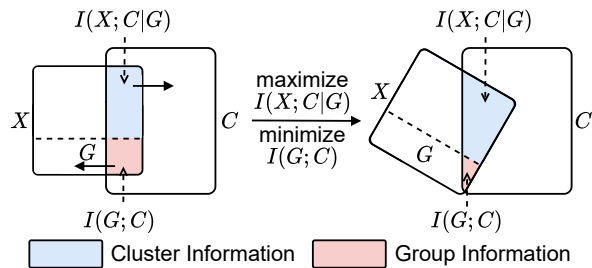
Figure 1. Illustration of our basic idea using information diagrams, where X, G, and C denote the inputs, sensitive attributes, and clustering assignments, respectively. To cluster data, we expect the overlap between non-group information and cluster information (*i.e.*, the conditional mutual information $I(X; C|G)$) to be maximized. Meanwhile, to prevent sensitive attributes from dominating the clustering results, we expect the overlap between group information and cluster information (*i.e.*, the mutual information $I(G, C)$) to be minimized.

whelm the intrinsic semantic of samples (also termed as **cluster information**). Taking single-cell RNA clustering as a showcase, standard methods would partition data based on sequencing techniques (group information) instead of intrinsic cell types (cluster information), since cells sequenced by different techniques would result in different expression levels [36] and most clustering methods cannot distinguish these two kinds of information. The case is similar in many automatic learning systems where the clustering results are biased toward sensitive attributes, which would interfere with the decision-making [9, 12, 18]. Notably, even though these sensitive attributes are known in prior, it is daunting to alleviate or even eliminate their influence, *e.g.*, removing the "gender" information from the photos of users.

As a feasible solution, fair clustering aims to hide sensitive attributes from the clustering results. Commonly, a clustering result is considered fair when samples of different sensitive attributes are uniformly distributed in clusters so

that the group information is protected. However, it would lead to a trivial solution if the fairness is over-emphasized, *i.e.*, all samples are assigned to the same cluster. Hence, in addition to fairness, balance and compactness are also highly expected in fair clustering. Specifically, a balanced clustering could avoid the aforementioned trivial solution brought by over-emphasized fairness, and the compactness refers to a clear cluster boundary.

To achieve fair clustering, many studies have been conducted to explore how to incorporate fairness into lustering [3, 4, 8, 22, 23, 38, 44, 46]. Their main differences lie in *i)* the stage of fairness learning, and *ii)* the depth of the model. In brief, [3, 8] incorporate the fairness in a pre-processing fashion by packing data points into so-called *fairlets* with balanced demographic groups and then partitioning them with classic clustering algorithms. [22, 46] are in-processing methods that formulate fairness as a constraint for clustering. As a representative of post-processing methods, [4] first performs classic clustering and then transforms the clustering result into a fair one by linear programming. Different from the above shallow models, [23, 38, 44] propose performing fair clustering in the latent space learned by different deep neural networks to boost performance. Although promising results have been achieved by these methods, almost all of them are heuristically and empirically designed, with few theoretical explanations and supports. In other words, it still lacks a unified theory to guide the algorithm design.

In this work, we unify the deep fair clustering task under the mutual information theory and propose a novel theoretical-grounded deep fair clustering method accordingly. As illustrated in Fig. 1, we theoretically show that clustering could be achieved by maximizing the conditional mutual information (CMI) $I(X; C|G)$ between inputs X and cluster assignments C given sensitive attributes G. Meanwhile, we prove that the fairness learning could be formulated as the minimization of the mutual information (MI) $I(G; C)$. In this case, sensitive attributes will be hidden in the cluster assignments and thus fair clustering could be achieved. To generalize our theory to deep neural networks, we additionally show a deep variant could be developed by maximizing the mutual information $I(X; X')$ between the input X and its approximate posterior X'. Notably, some deep clustering methods [17, 26] have been proposed based on the information theory. However, they are remarkably different from this work. To be exact, they ignored the group information. As a result, the group information will leak into cluster assignments, leading to unfair partitions. In addition, we prove that our mutual information objectives intrinsically correspond to four characteristics highly expected in deep fair clustering, namely, compact, balanced, and fair clusters, as well as informative features.

Besides the above contributions to theory and algorithm, this work also contributes to the performance evaluation. To be specific, we notice that almost all existing methods evaluate clustering quality and fairness separately. However, as fair clustering methods usually make a trade-off between these two aspects, such an evaluation protocol might be partial and inaccurate. As an improvement, we design a new evaluation metric based on the information theory, which simultaneously measures the clustering quality and fairness. The contribution of this work could be summarized as follows:

- We formulate deep fair clustering as a unified mutual information optimization problem. Specifically, we theoretically show that fair clustering could be achieved by maximizing CMI between inputs and cluster assignments given sensitive attributes while minimizing MI between sensitive attributes and cluster assignments. Moreover, the informative feature extraction could be achieved by maximizing MI between the input and its approximate posterior.

- Driven by our unified mutual information theory, we propose a deep fair clustering method and carry out extensive experiments to show its superiority on six fair clustering benchmarks, including a single-cell RNA atlas.

- To evaluate the performance of fair clustering more comprehensively, we design a novel metric that measures the clustering quality and fairness as a whole from the perspective of information theory.

2. Related Work

To alleviate or even eliminate the influence of sensitive attributes, many efforts have been devoted on fair clustering [1, 5, 7, 11, 21, 30]. Based on how fairness is incorporated, the existing works could be roughly divided into three categories, namely, pre-processing, in-processing, and post-processing methods. In brief, the pre-processing methods endow classic clustering methods with hand-craft fairness constraints. For example, Chierichetti et al. [8] first divide data points into several subsets (*i.e.*, *fairlets*) with the fairness constraint, and then employ a classic clustering algorithm on these *fairlets* to obtain the data partition. However, the fairlets construction requires at least quadratic running time, which is daunting in practice. To improve the scalability, Backurs et al. [3] employ a tree metric to approximate the fairlets construction in nearly linear time. Different from the pre-processing methods, the in-processing methods recast fairness as a constraint for joint optimization with the clustering objective. For example, Kleindessner et al. [22] recast the fairness as a linear constraint and embed it into the spectral clustering. Ziko et

al. [46] propose a variational framework by integrating fairness as a Kullback-Leibler (KL) term into the classic clustering methods. Opposite from the pre-processing methods, the post-processing methods [4] aim to transform the given clustering result into a fair one by solving a linear programming problem.

Motivated by the success of deep clustering [13, 14, 25, 37, 41], some studies have been carried out on deep fair clustering. For example, Wang et al. [38] propose learning a fair embedding by forcing the cluster centers to be equidistant from group centers, which could handle an arbitrary number of sensitive attributes. Li et al. [23] make a step forward to explore fair clustering on large-scale and high-dimensional visual data by incorporating fairness through adversarial training. Very recently, Zhang et al. [44] generate fair pseudo cluster assignments to guide the model optimization.

Although promising results have been achieved by these methods, their success partially relies on some tricks like pre-clustering and data augmentation which are clumsy in practical use. Besides, most existing works are designed heuristically and empirically, with few theoretical explanations and supports. Different from these studies, the proposed FCMI is built upon information theory, of which the working mechanism is interpretable. We reveal that deep fair clustering could be achieved by maximizing *i)* the mutual information $I(X; X')$ between the input X and its approximate posterior X', and *ii)* the conditional mutual information $I(X; C|G)$ between X and cluster assignments C given sensitive attributes G, while minimizing the mutual information $I(G; C)$. Both theoretical analysis and experimental results demonstrate the effectiveness of our method.

3. Method

In this section, we first give the mathematical definition of fair clustering. After that, we elaborate on how to learn compact, balanced, and fair clusters, as well as informative features through a unified information theory. Finally, we summarize the implementation of the proposed algorithm.

3.1. Problem Definition

For a given dataset $X = \{x_1, x_2, \ldots, x_N\}$ with sensitive attributes $G = \{g_1, g_2, \ldots, g_N\}$, fair clustering aims to partition X into K disjoint clusters with cluster assignments $C = \{c_1, c_2, \ldots, c_N\}$ by alleviating or even eliminating the influence of G, where $g_i \in [1, 2, \ldots, T]$, $c_i \in [1, 2, \ldots, K]$, $g_i = j$ means that sample x_i belongs to the *j-th* group, N is the data size, and T is the group number. As pointed out by [23], a clustering result is considered absolutely fair if the cluster assignments only depend on the semantics and are independent of the sensitive attributes. Based on such a heuristic definition, we mathematically formulate fair clustering as follows.

Definition 1 *Fair Clustering. Let $\tilde{g}_t = \{x_i | g_i = t\}$ and $\tilde{c}_k = \{x_i | c_i = k\}$ be the set of samples belonging to group t and cluster k respectively, and $p_{\tilde{g}_t, \tilde{c}_k} = \frac{1}{N} |\tilde{g}_t \cap \tilde{c}_k|$ be the joint probability density of groups and clusters, the clustering result is absolutely fair if*

$$p_{\tilde{g}_t, \tilde{c}_k} = p_{\tilde{g}_t} p_{\tilde{c}_k}, \forall t, k \tag{1}$$

where $p_{\tilde{g}_t} = \frac{1}{N} |\tilde{g}_t|$ and $p_{\tilde{c}_k} = \frac{1}{N} |\tilde{c}_k|$ denote the marginal probability densities of groups and clusters respectively.

Notably, fairness could be trivially achieved by assigning all samples to the same cluster, but obviously, it is not a reasonable solution. Hence, in addition to fairness, a good fair clustering method is also expected to embrace compactness and balance. Specifically, a clustering result is compact if the within-cluster distance is much smaller than the between-cluster distance, and a balanced clustering could avoid the aforementioned trivial solution due to over-emphasized fairness. For a deep fair clustering method, apart from the above three clustering characteristics, it also aims to learn informative features with redundancy removal for better clustering performance.

3.2. Deep Fair Clustering via Maximizing and Minimizing Mutual Information

As mentioned above, deep fair clustering has four objectives, namely, compact, balanced, fair clusters, and informative features, which are daunting to formulate and optimize jointly. In this paper, we theoretically show that these four diverse objectives could be derived from a unified theory, *i.e.*, maximizing and minimizing mutual information. In brief, fairness could be achieved by minimizing the mutual information between sensitive attributes G and cluster assignments C, *i.e.*, $I(G; C)$; compact and balanced clusters could be obtained by maximizing the conditional mutual information $I(X; C|G)$; and informative features could be learned through maximizing the mutual information $I(X; X')$, where X' denotes the approximate posterior from the prior. In the following, we will present the mathematical details for fair clustering and informative feature learning in turn.

3.2.1 Fair Clustering via $\max I(X; C|G)$ and $\min I(G; C)$

We begin with the discussion about conditional mutual information $I(X; C|G)$ and mutual information $I(G; C)$. In the following, we will prove why simultaneously maximizing $I(X; C|G)$ and minimizing $I(G; C)$ can make clusters compact, balanced, and fair. First, we formulate the objective function as follows:
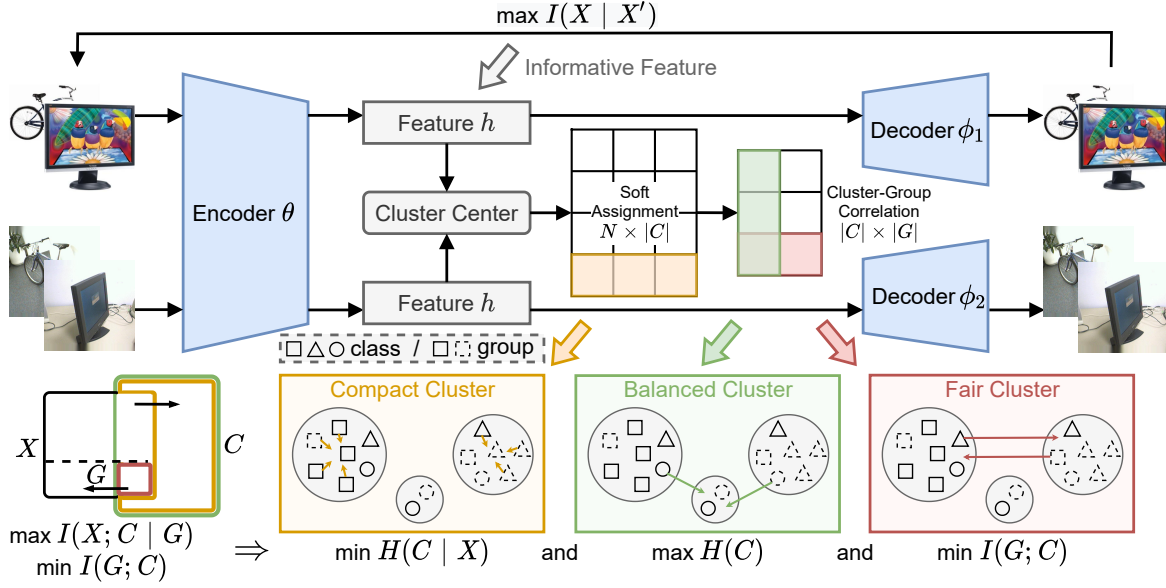
$$\max I(X; C|G) - \gamma I(G; C), \tag{2}$$

Figure 2. Overview of the proposed FCMI. In brief, FCMI extracts informative features by maximizing the mutual information $I(X; X')$ between samples X and the corresponding posterior X', implementing as an auto-encoder. To learn compact and balanced clusters in the hidden space, we maximize the conditional mutual information $I(X; C|G)$ between X and the cluster assignments C given sensitive attributes G. Meanwhile, we minimize the mutual information $I(G; C)$ to endow the clusters with robustness against sensitive attributes.

where $\gamma > 0$ is a trade-off parameter. By the definition of mutual information, we have

$$
\begin{aligned}
& I(X; C|G) - \gamma I(G; C) \\
&= H(C|G) - H(C|X, G) - \gamma I(G; C) \\
&= (H(C) - I(G; C)) - H(C|X, G) - \gamma I(G; C) \\
&= -H(C|X, G) + H(C) - (1 + \gamma)I(G; C).
\end{aligned}
\tag{3}
$$

Due to the over-high computational complexity, the first term $H(C|X, G)$ is intractable. To overcome this challenge, we theoretically show that $H(C|X, G)$ is exactly $H(C|X)$ which is more computationally efficient. Specifically, as sensitive attributes G are known in prior, we have $H(G|X) = H(G|C, X) = 0$. Hence, $H(C|X, G)$ could be rewritten as:

$$
\begin{aligned}
H(C|X, G) &= H(C|X) - I(G; C|X) \\
&= H(C|X) - (H(G|X) - H(G|C, X)) \\
&= H(C|X) = -\frac{1}{N} \sum_{i,k} c_{ik} \log c_{ik},
\end{aligned}
\tag{4}
$$

where c_{ik} is the probability of sample x_i being assigned to the k-th cluster, which is computed according to its distance to the cluster centers (more details provided later in Supplementary Sec. 2). As can be seen, the minimization of $H(C|X, G)$ pushes each sample to its corresponding cluster center and away from the others, *i.e.*, the model is encouraged to produce **compact** clusters.

The second term $H(C)$ is the entropy of cluster assignments over all samples, namely,

$$
H(C) = -\sum_k p_{\tilde{c}_k} \log p_{\tilde{c}_k},
\tag{5}
$$

where $p_{\tilde{c}_k} = \frac{1}{N} \sum_i c_{ik}$ refers to the marginal probability density function of clusters. The maximization of $H(C)$ punishes over-large or small clusters to avoid trivial solutions due to over-emphasized fairness, which leads to **balanced** clusters.

The third term $I(G; C)$ refers to the mutual information between the cluster assignments and the sensitive attributes, which is computed as

$$
I(G; C) = \sum_{t,k} p_{\tilde{g}_t, \tilde{c}_k} \log \frac{p_{\tilde{g}_t, \tilde{c}_k}}{p_{\tilde{g}_t} p_{\tilde{c}_k}},
\tag{6}
$$

where $p_{\tilde{g}_t} = \frac{1}{N} \sum_i \mathbb{1}_{x_i \in \tilde{g}_t}$ is the group marginal probability density function, and $p_{\tilde{c}_k, \tilde{g}_t} = \frac{1}{N} \sum_i \mathbb{1}_{x_i \in \tilde{g}_t} c_{ik}$ denotes the joint probability density function of clusters and groups. Notably, $I(G; C)$ is a convex function and it reaches the minimum point $I(G; C) = 0$ *i.f.f.* $p_{\tilde{g}_t, \tilde{c}_k} = p_{\tilde{g}_t} p_{\tilde{c}_k}$, which exactly corresponds to the absolute fairness defined in Definition 1. Hence, the minimization of $I(G; C)$ encourages the data partition to be fair against sensitive attributes, thus leading to **fair** clusters. Note that solely optimizing $I(X; C|G)$ would not lead to fair clustering, see Fig. 1.

Based on the above analyses, Eq. 2 could be decomposed

into the following two objectives, namely,

$$\mathcal{L}_{clu} = -H(C) + H(C|X, G)$$
$$= \sum_k p_{\tilde{c}_k} \log p_{\tilde{c}_k} - \frac{1}{N} \sum_{i,k} c_{ik} \log c_{ik}, \quad (7)$$

and

$$\mathcal{L}_{fair} = I(G; C) = \sum_{t,k} p_{\tilde{g}_t, \tilde{c}_k} \log \frac{p_{\tilde{g}_t, \tilde{c}_k}}{p_{\tilde{g}_t} p_{\tilde{c}_k}}. \quad (8)$$

3.2.2 Informative Feature Learning via $\max I(X; X')$

The above theoretical analysis has shown that compactness, balance, and fairness could be derived from the unified perspective of mutual information. In this section, we reveal that the informative feature could also be learned via maximizing the mutual information $I(X; X')$ between the prior X and the corresponding approximate posterior X'. Without loss of generality, we take X' as the auto-encoder reconstruction for the given X in the following analysis and our implementation.

To be specific, given $I(X; X') = H(X) - H(X|X')$, the maximization of $I(X; X')$ is equivalent to the minimization of the conditional entropy $H(X|X') = -\mathbb{E}_{p(X,X')}[\log p(X|X')]$, since $H(X)$ is a constant. However, as the probability density function $p(X|X')$ cannot be accessed directly, we alternate to minimize its upper bound, *i.e.*,

$$-\mathbb{E}_{p(X,X')}[\log q(X|X')]$$
$$\geq -\mathbb{E}_{p(X,X')}[\log q(X|X')] - D_{KL}(p(X|X')||q(X|X'))$$
$$= -\mathbb{E}_{p(X,X')}[\log p(X|X')], \quad (9)$$

where $q(X|X')$ could be any distribution with a known probability density function and D_{KL} is KL divergence. Without loss of generality, we assume it obeys Gaussian distribution $N(X; X', \sigma^2 I)$, then

$$-\mathbb{E}_{p(X,X')}[\log q(X|X')] \propto \mathbb{E}_{p(X,X')}[||X - X'||^2] + a, \quad (10)$$

where a is a constant. As a result, the upper bound could be optimized by minimizing the following reconstruction loss, *i.e.*,

$$\mathcal{L}_{rec} = ||X - X'||^2 = ||X - \Phi(\theta(X))||^2, \quad (11)$$

where θ denotes a shared encoder, and Φ denotes a multi-branch decoder that reconstructs samples with different attributes separately. Formally, $\Phi(h_i) = \phi_{g_i}(h_i), g_i \in [1, 2, \ldots, T]$, where ϕ_{g_i} is the group-specific decoder which reconstructs samples from the g_i-*th* group, and $h_i = \theta(x_i)$ denotes the feature extracted by the encoder. In other words, we use a multi-branch decoder to recover the group information that has been removed in the hidden space by $\mathcal{L}_{\text{fair}}$ for better reconstruction.

As $I(X; X') \leq I(X; h)$, the maximization of $I(X; X')$ intrinsically increase the lower bound of $I(X; h)$, and thus help the auto-encoder to extract **informative** features h from the raw inputs.

3.3. The Objective Function and Algorithm Details

To summarize, we unify the deep fair clustering task from the perspective of information theory. With the above theoretical analyses, we arrive at our loss by combining Eq. 7, 8 and 11, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha(\mathcal{L}_{clu} + (1 + \gamma)\mathcal{L}_{fair})$$
$$= \mathcal{L}_{rec} + \alpha\mathcal{L}_{clu} + \beta\mathcal{L}_{fair}, \quad (12)$$

3.4. The Proposed Evaluation Metrics for Fair Clustering

To evaluate the fairness of clustering results, most existing studies adopt the *Balance (Bal.)* metric [8, 38] which is defined as the ratio between the largest and smallest sensitive groups in a cluster. However, the distributions of other groups are ignored when there are more than two groups (e.g., if there were 3, 4, 18, 20 samples from each group, the *Bal.* will be 3/20 which is the same when it becomes 3, 11, 11, 20.) To address this issue, we propose a novel fairness measurement, dubbed *Minimal Normalized Conditional Entropy (MNCE)* as below.

Definition 2 *Minimal Normalized Conditional Entropy (MNCE). Given N data points X with sensitive groups G from clusters C, MNCE is defined as the minimal group entropy in each cluster divided by the global group entropy. Formally,*

$$\text{MNCE} = \frac{\min_k (H(G|\tilde{c}_k))}{H(G)}$$
$$= \frac{\min_k \left(-\sum_t \frac{|\tilde{g}_t \cap \tilde{c}_k|}{|\tilde{c}_k|} \log \frac{|\tilde{g}_t \cap \tilde{c}_k|}{|\tilde{c}_k|}\right)}{-\sum_t \frac{|\tilde{g}_t|}{N} \log \frac{|\tilde{g}_t|}{N}} \in [0, 1], \quad (13)$$

where $\tilde{g}_t = \{x_i|g_i = t\}$ and $\tilde{c}_k = \{x_i|c_i = k\}$ denote the set of samples belonging to the t-th group and k-th cluster respectively, $H(G|\tilde{c}_k)$ is the conditional entropy of sensitive attributes given the cluster assignments, and $H(G)$ denotes the entropy of sensitive attributes.

As $H(G)$ is a constant for the given dataset, we derive the following proposition from Eq. 13:

Proposition 1 *The cluster assignments are independent of the sensitive attributes if and only if MNCE = 1.*

Proof 1 *Suppose the cluster assignments are independent of the sensitive attributes, we have $p_{\tilde{g}_t, \tilde{c}_k}/p_{\tilde{c}_k} = p_{\tilde{g}_t}$, where*

$p_{\tilde{c}_k, \tilde{g}_t}$ denotes the joint probability density function of clusters and groups, $p_{\tilde{g}_t}$ and $p_{\tilde{c}_k}$ are the marginal probability density function of group and clusters, respectively. From the Bayes' theorem, one have $p_{g_t|c_k} = p_{g_t,c_k}/p_{c_k}$, and thus $p_{g_t|c_k} = p_{g_t}$. Further,

$$
\begin{aligned}
H(G|\tilde{c}_k) &= -\sum_t p_{\tilde{g}_t|\tilde{c}_k} \log p_{\tilde{g}_t|\tilde{c}_k} = -\sum_t p_{\tilde{g}_t} \log p_{\tilde{g}_t} \\
&= H(G), \forall k \in \{1, 2, \dots, K\}.
\end{aligned}
\tag{14}
$$

As a result, $\min_k(H(G|\tilde{c}_k)) = H(G)$ and MNCE $= 1$. Notably, the above deductions hold reversely as $H(G|\tilde{c}_k) \leq H(G), \forall k$.

From Proposition 1, *MNCE* reaches the maximum point $MNCE = 1$ *i.f.f.* the cluster assignments are absolutely *fair* as defined in Definition 1. Hence, a larger *MNCE* indicates a fairer clustering result.

However, a separate evaluation is less attractive as fair clustering methods usually make a trade-off between clustering quality and fairness. Alternatively, we propose the following metric to simultaneously measure the clustering quality and fairness,

Definition 3 F_β: *An overall measure for clustering quality and fairness.* Let u and v denote the clustering metric NMI and the fairness metric MNCE, F_β is defined as the harmonic mean of u and v, i.e.,

$$
\begin{aligned}
F_\beta &= 1 / \left(\frac{\beta^2}{1+\beta^2} \frac{1}{v} + \frac{1}{1+\beta^2} \frac{1}{u} \right) \\
&= \frac{(1+\beta^2)uv}{\beta^2 u + v} \in [0, 1] \ (u, v \in [0, 1]),
\end{aligned}
\tag{15}
$$

where $\beta \in [0, +\infty)$ is a hyper-parameter to adjust the weight of clustering quality and fairness. A larger β corresponds to more focus on fairness. In general, we recommend $\beta = 1$ to treat these two terms equally.

4. Experiments

In this section, we evaluate our FCMI on six fair clustering benchmarks, compared with both the classical (non-fair) and state-of-the-art fair clustering methods. In addition, we carry out a series of qualitative analyses and ablation studies to investigate the effectiveness and robustness of FCMI.

4.1. Experimental Setups

Dataset: We adopt six datasets confounded with various types of sensitive attributes for evaluations (summarized in Supplementary Table 1). Among them, the first five datasets including MNIST-USPS[1], Reverse MNIST,

HAR [2], Office-31 [33], and MTFL [45] are commonly used in fair clustering studies [23,38]. To explore the potential in practical applications, we additionally evaluate FCMI on the single-cell mouse atlas dataset that is widely used in biological analysis, where the sensitive attributes correspond to different sequence techniques [10, 15].

Implementation Details: Consistent with the previous works [23], we use a convolutional auto-encoder for MNIST-USPS and Reverse MNIST, and a fully-connected auto-encoder for handing other datasets. For MTFL and Office-31, the features extracted by ResNet50 [16] are used as the inputs. In all experiments by default, we fix the hyper-parameter $\alpha = 0.04, \beta = 0.20$ in Eq. 12 across all the datasets. The only exception is that we remove the balance constraint (*i.e.*, set $\beta = 0$) on the single-cell mouse atlas since the cells of different types are highly unbalanced. The model is trained for 300 epochs using the Adam optimizer with an initial learning rate of $1e-4$ for all datasets, with a warm-up in the first 20 epochs using the reconstruction loss defined in Eq. 11. All experiments are conducted on a Nvidia A10 GPU on the Ubuntu 18.04 platform.

Baselines: Both classic clustering methods and state-of-the-art fair clustering methods are used for comparisons. Specifically, for the classic methods, we select auto-encoder + k-means [37], DEC [40], DAC [6] and ClGAN [32] as baselines. For the shallow fair clustering methods, ScFC [3], SpFC [22], VFC [46], FAlg [4] are used for comparisons. For the deep fair clustering methods, we could only investigate the performance of DFC [23] since the code of DFDC [44] and Towards [38] are unavailable. As an alternative, we present the original results reported in their paper on the common datasets for reference. Notably, ScFC [3] and DFC [23] only support two groups and they are impractical on the HAR dataset, while other methods including our FCMI could be generalized to arbitrary group numbers.

Evaluation Metrics: In our experiments, the widely-used *ACC* and *NMI* metrics are used to investigate the clustering quality. And the previous *Balance (Bal.)* metric and the proposed *MNCE* metric are used to evaluate the fairness. In addition, we adopt the proposed measurement F_β for a comprehensive evaluation.

4.2. Quantitative Comparisons

In this section, we carry out quantitative experiments by comparing FCMI with 11 baselines. As shown in Tab. 1, although the classical clustering methods achieve competitive clustering performance in terms of *ACC* and *NMI*, they show poor results in terms of fairness. On the contrary, fair clustering methods inject fairness into clustering, leading to a debiased data partition. However, some shallow fair clustering methods such as ScFc [3] guarantee the fairness explicitly at the cost of clustering quality. Compared with

Table 1. Comparisons of FCMI with both standard and fair clustering methods on six benchmarks. The best and the second best results are marked in **bold** and <u>underline</u>, respectively. Some results are unavailable since the codes for DFDC and Toward are unpublished, and DFC and ScFC are unable to handle the HAR dataset which is consisted of multiple (≥ 3) sensitive groups.

|	MNIST-USPS					Color Reverse MNIST					HAR				
Method	ACC	NMI	Bal	MNCE	F_β	ACC	NMI	Bal	MNCE	F_β	ACC	NMI	Bal	MNCE	F_β
AE	76.3	71.8	0.0	0.0	0.0	41.0	52.8	0.0	0.0	0.0	66.3	60.7	0.0	86.9	71.5
DEC	60.0	59.4	0.0	0.0	0.0	40.7	38.2	0.0	0.0	0.0	57.1	65.5	0.0	93.7	77.1
DAC	76.3	69.9	0.0	0.0	0.0	31.4	27.1	0.0	0.0	0.0	38.2	31.5	0.0	32.4	31.9
ClGAN	38.3	35.7	0.1	1.9	3.6	20.1	9.1	2.2	14.9	11.3	52.7	44.3	0.4	0.0	0.0
ScFC	14.2	1.3	<u>11.2</u>	**95.0**	2.6	51.3	49.1	**100.0**	**100.0**	65.8	-	-	-	-	-
SpFC	20.1	15.5	0.0	0.0	0.0	11.0	2.1	0.0	0.0	0.0	19.0	0.4	0.0	0.0	0.0
VFC	58.1	55.2	0.0	0.0	0.0	38.1	42.7	0.0	0.0	0.0	62.6	66.2	25.6	98.7	<u>79.3</u>
FAlg	58.4	53.8	9.5	85.8	66.1	26.9	14.3	66.6	97.1	24.9	56.6	58.6	<u>43.2</u>	<u>99.2</u>	73.7
Towards	72.5	71.6	3.9	-	-	42.5	50.6	43.0	-	-	60.7	66.1	16.6	-	-
DFC	85.7	83.4	6.7	68.2	<u>75.0</u>	49.9	68.9	80.0	99.1	<u>81.3</u>	-	-	-	-	-
DFDC	<u>93.6</u>	<u>87.6</u>	**11.9**	-	-	<u>58.9</u>	69.0	94.6	-	-	<u>86.2</u>	**84.5**	**46.8**	-	-
FCMI(Ours)	**96.7**	**91.8**	10.7	<u>94.5</u>	**92.0**	**88.4**	**86.4**	<u>99.5</u>	<u>99.9</u>	**92.7**	**88.2**	<u>80.7</u>	40.7	**99.3**	**89.0**
	Office-31					MTFL					Mouse Atlas				
Method	ACC	NMI	Bal	MNCE	F_β	ACC	NMI	Bal	MNCE	F_β	ACC	NMI	Bal	MNCE	F_β
AE	63.8	66.8	0.0	0.0	0.0	67.2	16.0	67.8	97.3	27.5	56.1	54.5	0.6	5.7	10.3
DEC	63.3	68.6	0.0	0.0	0.0	56.7	0.6	78.0	98.9	1.1	<u>61.5</u>	<u>63.2</u>	0.7	6.3	11.4
DAC	14.0	25.2	0.0	0.0	0.0	58.9	1.4	81.5	87.9	2.7	48.3	40.3	36.0	86.4	54.9
ClGAN	52.2	54.9	0.0	0.0	0.0	**72.9**	12.6	79.1	99.0	22.4	48.3	50.8	0.8	7.2	12.6
ScFC	38.0	60.7	**26.7**	**97.7**	74.9	52.1	15.1	**100.0**	**100.0**	26.3	31.2	16.0	**63.6**	**99.9**	27.5
SpFC	9.3	11.4	0.0	0.0	0.0	65.5	0.1	75.0	98.5	0.2	21.3	3.8	0.0	0.0	0.0
VFC	65.2	69.7	20.3	86.0	77.0	68.8	8.4	88.9	99.8	15.6	45.4	49.8	0.0	0.0	0.0
FAlg	67.1	70.7	20.4	86.4	<u>77.8</u>	63.2	16.7	60.1	96.3	28.5	52.2	58.5	<u>45.2</u>	<u>92.7</u>	<u>71.7</u>
DFC	<u>69.0</u>	<u>70.9</u>	11.9	64.2	67.4	<u>72.8</u>	<u>17.6</u>	<u>97.4</u>	<u>99.9</u>	<u>30.0</u>	60.0	59.1	21.4	69.7	63.9
FCMI(Ours)	**70.0**	**71.2**	<u>22.6</u>	<u>90.6</u>	**79.7**	70.2	**19.1**	90.4	99.8	**32.0**	**65.8**	**65.4**	38.1	88.1	**75.0**
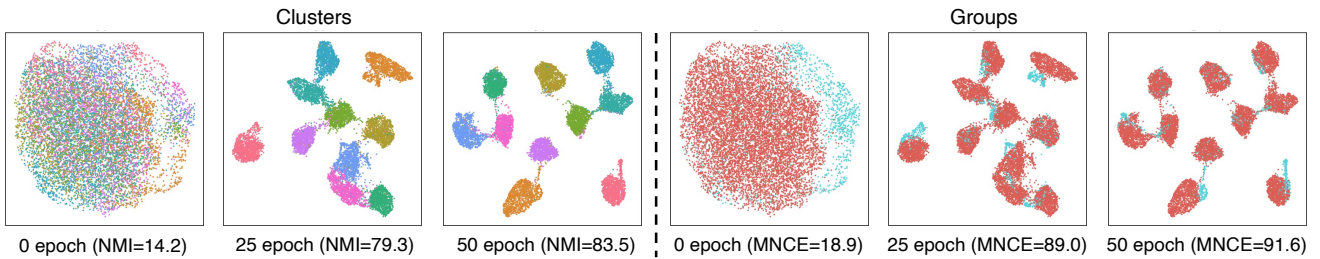
Figure 3. Visualization of the hidden representation on MNIST-USPS with the increasing training epoch. The left and right three figures are colored by classes and groups, respectively. For more visualization and comparison, please see Supplementary Fig 1.

shallow ones, deep fair clustering methods achieve a more elegant trade-off between clustering quality and group fairness. Besides the superior performance of our FCMI on clustering and fairness metrics, we would like to highlight that FCMI achieves dominance in terms of F_β. Specifically, FCMI outperforms the best competitor by 17.0%, 11.4%, 9.7%, 1.9%, 2.0%, and 3.3% in terms of F_β on six datasets respectively.

4.3. Visualization

To help understand the working mechanism of FCMI, we first visualize the hidden representation of MNIST-USPS by
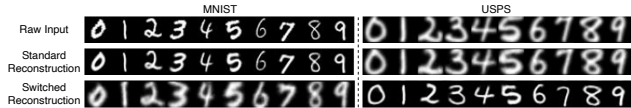
Figure 4. Visualization of the reconstructions outputted by group-specific decoders on MNIST-USPS. The rows from top to bottom refer to the raw inputs, the reconstructions from the corresponding decoder, and the reconstructions from the switched decoder.
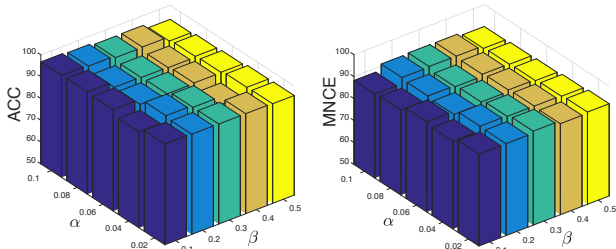

Figure 5. The influence of α and β on *ACC* and *MNCE* on MNIST-USPS. The stable performance of FCMI demonstrates its robustness against hyper-parameters.

Table 2. Ablation study of $\mathcal{L}_{clu}$ and $\mathcal{L}_{fair}$ on MNIST-USPS.

$\mathcal{L}_{clu}$	$\mathcal{L}_{fair}$	ACC	NMI	Bal	MNCE	F_β
✓	✓	**96.7**	**91.8**	**10.7**	**92.3**	**92.0**
✓		95.2	89.9	8.2	77.9	83.5
	✓	89.6	79.2	9.0	86.1	81.1

performing UMAP [31] on the learned features across the training process. As shown in Fig. 3, the data shows the group information evenly distributes in the initial feature space. As the training goes on, more compact, balanced, and fairer clusters are learned by FCMI.

Recalling in our implementation, a multi-branch decoder is used to recover the group information for better reconstruction. To verify the effectiveness of such a multi-branch decoder, we conduct experiments on MNIST-USPS and switch the decoder for different groups to reconstruct images. As shown in Fig. 4, the group information is successfully transferred from USPS to MNIST, and vice versa. In other words, this result proves that the encoder could extract the semantic information and each branch of the decoder could capture the group information.

4.4. Parameter Analysis and Ablation Study

In this section, we investigate the influence of the hyper-parameters α and β on the MNIST-USPS dataset. As shown in Fig. 5, the performance of FCMI is stable under different choices of α and β, which demonstrates its robustness against the hyper-parameters. However, when one of the mutual information $I(X;C|G)$ and $I(G;C)$ is removed
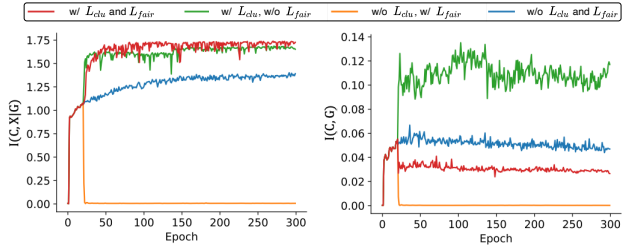

Figure 6. The evolution of mutual information $I(C, X|G)$ and $I(C, G)$ across the training process with and without $\mathcal{L}_{clu}$ and $\mathcal{L}_{fair}$.

(*i.e.*, without $\mathcal{L}_{clu}$ or $\mathcal{L}_{fair}$), our model encounters a significant drop in clustering or fairness performance as shown in Table 2. Such an ablation study verifies the effectiveness of our information theory-driven losses.

To further investigate the effectiveness of $\mathcal{L}_{clu}$ and $\mathcal{L}_{fair}$, we visualize the evolution of mutual information $I(X;C|G)$ and $I(G;C)$ across the training process on the HAR dataset. For comparisons, we compute the mutual information when we remove one or both losses. As demonstrated in Fig. 6, both $I(X;C|G)$ and $I(G;C)$ increase in the first 20 epochs since the informative features contain the cluster and group information at the preliminary learning stage. After that, with $\mathcal{L}_{clu}$ and $\mathcal{L}_{fair}$ (red line), the model becomes more powerful to differentiate different clusters while alleviating the influence of the group information, compared with the baseline when only $\mathcal{L}_{rec}$ is used (blue line). Without $\mathcal{L}_{fair}$ (green line), group information will leak into the cluster assignments resulting in unfair data partitions. Without $\mathcal{L}_{clu}$ (orange line), the model would collapse by solely minimizing mutual information $I(G;C)$.

5. Conclusion

In this paper, we build a novel deep fair clustering method (FCMI) and theoretically show that it could achieve compact, balanced, and fair clusters, as well as informative features. In addition, we design a novel evaluation metric that measures the clustering quality and fairness as a whole. Extensive experimental results demonstrate the superiority of our method over 11 baselines on six benchmarks including a single-cell RNA-seq atlas.

Acknowledgement

References

[1] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 267–275, 2019. 2

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra Perez, and Jorge Luis Reyes Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*, pages 437–442, 2013. 6

[3] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR, 2019. 2, 6

[4] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 6

[5] Brian Brubach, Darshan Chakrabarti, John Dickerson, Samir Khuller, Aravind Srinivasan, and Leonidas Tsepenekas. A pairwise fair and community-preserving approach to k-center clustering. In *International Conference on Machine Learning*, pages 1178–1189. PMLR, 2020. 2

[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*, pages 5879–5887, 2017. 6

[7] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *International Conference on Machine Learning*, pages 1032–1041. PMLR, 2019. 2

[8] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 5

[9] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020. 1

[10] Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018. 6

[11] Ian Davidson and SS Ravi. Making existing clusterings fairer: Algorithms, complexity results and insights. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3733–3740, 2020. 2

[12] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[13] 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*, pages 5736–5745, 2017. 3

[14] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, pages 373–382. Springer, 2017. 3

[15] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018. 6

[16] 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*, pages 770–778, 2016. 6

[17] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International conference on machine learning*, pages 1558–1567. PMLR, 2017. 2

[18] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019. 1

[19] Zhao Kang, Chong Peng, Qiang Cheng, and Zenglin Xu. Unified spectral clustering with optimal graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1

[20] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019. 1

[21] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*, pages 3448–3457. PMLR, 2019. 2

[22] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International Conference on Machine Learning*, pages 3458–3467. PMLR, 2019. 2, 6

[23] Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9070–9079, 2020. 2, 3, 6

[24] Sheng Li, Kang Li, and Yun Fu. Temporal subspace clustering for human motion segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 4453–4461, 2015. 1

[25] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *2021 AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 3

[26] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2023. 2

[27] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012. 1

[28] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceed-*

ings of the 27th International Conference on International Conference on Machine Learning, pages 663–670, 2010. 1

[29] Yijing Luo, Bo Han, and Chen Gong. A bi-level formulation for label noise learning with spectral cluster discovery. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2605–2611, 2021. 1

[30] Sepideh Mahabadi and Ali Vakilian. Individual fairness for k-clustering. In *International Conference on Machine Learning*, pages 6586–6596. PMLR, 2020. 2

[31] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 8

[32] 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*, volume 33, pages 4610–4617, 2019. 6

[33] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 6

[34] Xiaobo Shen, Weiwei Liu, Ivor Tsang, Fumin Shen, and Quan-Sen Sun. Compressed k-means for large-scale clustering. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 1

[35] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. From ensemble clustering to multi-view clustering. In *IJCAI*, 2017. 1

[36] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21(1):1–32, 2020. 1

[37] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 3, 6

[38] Bokun Wang and Ian Davidson. Towards fair deep clustering with multi-state protected variables. *arXiv preprint arXiv:1901.10053*, 2019. 2, 3, 5, 6

[39] Qi Wang, Mulin Chen, Feiping Nie, and Xuelong Li. Detecting coherent groups in crowd scenes by multiview clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):46–58, 2018. 1

[40] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016. 6

[41] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4066–4075, 2019. 3

[42] Yingzhen Yang, Jiashi Feng, Nebojsa Jojic, Jianchao Yang, and Thomas S Huang. ℓ^0 -sparse subspace clustering. In *European conference on computer vision*, pages 731–747. Springer, 2016. 1

[43] Jinfeng Yi, Lijun Zhang, Rong Jin, Qi Qian, and Anil Jain. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *International conference on machine learning*, pages 1400–1408. PMLR, 2013. 1

[44] Hongjing Zhang and Ian Davidson. Deep fair discriminative clustering. *arXiv preprint arXiv:2105.14146*, 2021. 2, 3, 6

[45] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014. 6

[46] Imtiaz Masud Ziko, Jing Yuan, Eric Granger, and Ismail Ben Ayed. Variational fair clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11202–11209, 2021. 2, 3, 6