1. Datasets Used for Evaluation

To demonstrate the versatility of our method, we use six datasets confounded with various types of sensitive attributes for evaluations. For clarity, we summarize the statical characteristics of the datasets in Tab. 1. Among them, the first four datasets, i.e., MNIST-USPS, Reverse MNIST, Office-31 [8], MTFL [11] are with visual modality and the last two (HAR [1] and Mouse Atlas [4, 5]) are signal-vector datasets. Notably, some existing works [2, 3] could only handle the dataset with bi-sensitive-attribute (group number is two), whereas our method is generalizable to the case of arbitrary group numbers, e.g., HAR with 30 groups.

Table 1. Datasets used for evaluations.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Samples</th>
<th>#Clusters</th>
<th>Semantic</th>
<th>#Groups</th>
<th>Sensitive Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST-USPS</td>
<td>67,291</td>
<td>10</td>
<td>Digit</td>
<td>2</td>
<td>Domain Source</td>
</tr>
<tr>
<td>Reverse-MNIST</td>
<td>120,000</td>
<td>10</td>
<td>Digit</td>
<td>2</td>
<td>Background Color</td>
</tr>
<tr>
<td>HAR</td>
<td>10,299</td>
<td>6</td>
<td>Activity</td>
<td>30</td>
<td>Subject</td>
</tr>
<tr>
<td>Office-31</td>
<td>3,612</td>
<td>31</td>
<td>Category</td>
<td>2</td>
<td>Domain Source</td>
</tr>
<tr>
<td>MTFL</td>
<td>2,000</td>
<td>2</td>
<td>Gender</td>
<td>2</td>
<td>w/ or w/o Glass</td>
</tr>
<tr>
<td>Mouse Atlas</td>
<td>6,954</td>
<td>11</td>
<td>Cell Type</td>
<td>2</td>
<td>Sequence Technique</td>
</tr>
</tbody>
</table>

2. Algorithm Implementation Details for FCMI

To elaborate on the working flow of our FCMI, we provide its pseudo code below. While the previous works [7, 9, 10] resort to some techniques such as layer-wise pre-training, pre-clustering and data augmentation, our method achieves state-of-the-art performance with end-to-end training initialized by a simple warm-up step.

Algorithm 1 Deep Fair Clustering via Minimizing and Maximizing Mutual Information.

**Input:** Samples $X = \{x_i\}_{i=1}^N$ with sensitive attributes $G = \{g_i\}_{i=1}^N$, warmup_epoch = 20, max_epoch = 300.
1: for epoch < max_epochs do
2: Extract low-dimensional features via $h = \theta(X)$
3: Compute cluster centers $U$ by applying k-means on $h$
4: for sampled mini-batch $x = \{x_j\}_{j=1}^n$ do
5: if epoch < warmup_epochs then
6: Compute the overall loss $L = L_{rec}$ by Eq. 11 in our main paper.
7: else
8: Compute soft cluster assignments $c_{ik}$ by Eq. 1
9: Compute $L_{rec}$, $L_{clu}$ and $L_{fair}$ by Eq. 11, Eq. 7 and Eq. 8 in our main paper.
10: Compute the overall loss $L$ by Eq. 12 in our main paper.
11: end if
12: Update encoder $\theta$ and decoder $\Phi$ to minimize $L$ via Stochastic Gradient Descent (SGD).
13: end for
14: end for

The remaining problem is how to compute the soft cluster assignments $c_{ik}$. To prove the effectiveness of the proposed unified information theory itself, without bells and whistles, we adopt the vanilla k-means to compute $c_{ik}$, namely,

$$c_{ik} = \frac{\exp(\cos(h_i, u_k)/\tau)}{\sum_j \exp(\cos(h_i, u_k)/\tau)},$$

where $U = \{u_1, u_2, ..., u_K\}$ are obtained cluster centers and $\tau = 0.1$ is the temperature to control the softness. In our implementation, we apply K-means at the beginning of each epoch to update the clustering centers $U$. This would...
only additionally introduce approximately 15% computational burden if we use the GPU implementation provided by faiss [6].

3. Visual Comparisons

To better show the superiority of the proposed FCMI, we conduct visualization on MNIST-USPS and Color Reverse MNIST in Fig. 1, comparing with two most competitive baselines, i.e., AE (a standard clustering method) and DFC (a fair clustering method).

From the visualizations, one could see that the standard clustering method AE fails to eliminate the influence of sensitive attributes, leading to unfair data partition. DFC is able to hide sensitive attributes from the clustering assignment. However, it fails to capture the intrinsic semantics accurately. Namely, it fails to distinguish digits 4 and 9 in MNIST and USPS, and it mixes all digits except digit 8 on Color Reverse MNIST. In contrast, our FCMI successfully clusters data based on the digits except for digits 4 and 9 on Color Reverse MNIST. Moreover, our FCMI hides the sensitive attributes more thoroughly, as shown in the better mixing of digits from MNIST and Color Reverse MNIST (the last row).

4. Broader Impact Statement

The fairness problem hinders the utilization of clustering in a wide range of real-world applications such as image segmentation, biological analysis, and information retrieval since the real-world data might be confounded with sensitive attributes, e.g., color, gender, race, and RNA-sequencing technique. Thus, it is crucial to achieve fair clustering and could bring many benefits, e.g., preventing discrimination against certain individuals; reducing the cost of human labor to discover samples with similar semantics without prejudice. However, we should not ignore the potential negative impact of this work. More specifically, our method is based on a deep neural network, which inherits the black-box nature of artificial neural networks, thus would hinder its wide adoption in mission-critical applications.

References