



HAL
open science

Graph Cuts with Arbitrary Size Constraints Through Optimal Transport

Chakib Fettal, Lazhar Labiod, Mohamed Nadif

► **To cite this version:**

Chakib Fettal, Lazhar Labiod, Mohamed Nadif. Graph Cuts with Arbitrary Size Constraints Through Optimal Transport. 2023. hal-03917041v2

HAL Id: hal-03917041

<https://hal.science/hal-03917041v2>

Preprint submitted on 11 May 2023 (v2), last revised 7 Feb 2024 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Graph Cuts with Arbitrary Size Constraints Through Optimal Transport

Chakib Fettal
Centre Borelli UMR 9010
Université Paris Cité
CDC Informatique
chakib.fettal@u-paris.fr

Lazhar Labiod
Centre Borelli UMR 9010
Université Paris Cité
lazhar.labiod@u-paris.fr

Mohamed Nadif
Centre Borelli UMR 9010
Université Paris Cité
mohamed.nadif@u-paris.fr

Abstract

Image Clustering is an important task in computer vision and machine learning in general, and new applications are constantly appearing. A common way of obtaining an image dataset partition is through graph cuts, which are also used as a component in more complex clustering paradigms such as subspace clustering. One drawback of classical min-cut algorithms is that they tend to produce small groups, which is why more balanced variants have risen, including normalized and ratio cuts. We believe, however, that with these variants the balance constraints are too restrictive for some applications (long-tailed clustering), while not being restrictive enough for others (when searching for perfectly balanced partitions), since the constraint is not hard. Here, we propose a new graph cut algorithm for partitioning with arbitrary size constraints. We formulate the graph cut problem as a constrained Gromov-Wasserstein problem, and our algorithm is slower than the classical spectral clustering algorithm by only a ratio of $\mathcal{O}(\log(n))$ while being more efficient in practice. We demonstrate the performance of our approach on several balanced and imbalanced (long-tail) datasets.

1 Introduction

Clustering is an important task in machine learning and computer vision. Intuitively, the task of image clustering boils down to grouping images into clusters such that the images within the same clusters are similar to each other, while those in different clusters are dissimilar. Applications are diverse and wide ranging, including, for example, content-based image retrieval [Bhunia et al., 2020, Lee et al., 2022, Bhunia et al., 2021], image annotation [Cheng et al., 2018, Cai et al., 2013], and image indexing [Cao et al., 2013]. Consequently, much research has been dedicated to image clustering [Chang et al., 2017, Ji et al., 2017, Elhamifar and Vidal, 2013, Ji et al., 2019].

A popular way of formulating the image clustering problem is through the minimum graph cut (min-cut) problem where the graph is created based on the input images. However, in practice, the min-cut problem suffers from the formation of some small groups which leads to bad performance. As a result, other versions of min-cut were proposed that take into account the size of the resulting groups, in order to make the partition more balanced. This notion of size is variable, for example, in the Normalized Cut (ncut) problem [Shi and Malik, 2000], size refers to the total volume of a group, while in the Ratio Cut (rcut) problem [Hagen and Kahng, 1992], it refers to the cardinality of a group. A common method for solving the ncut and rcut problems is that of the spectral clustering algorithm [Von Luxburg, 2007, Ng et al., 2001] which is popular due to it often showing good empirical performance and being somewhat efficient. The spectral clustering algorithm variants are present in many image clustering frameworks, such as for subspace clustering [Agrawal et al., 1998]

where a spectral clustering algorithm is applied to a learned subspace affinity matrix to obtain a partition of the points according to the subspaces in which they lie.

However, some restrictions that apply to the spectral clustering algorithms and to most approaches tackling the $ncut$ and $rcut$ problems in general do exist. A first one is that the balance constraint is not strict enough, meaning that even if we include the size regularization into the min-cut problem, the groups are still not necessarily of similar size, which is why several truly balanced clustering algorithms have been proposed in the literature [Chen et al., 2017, 2019, Li et al., 2018]. Another problem is that the balance constraint is too restrictive for many real world datasets, for example, a recent trend in computer vision is to propose approaches dealing with long-tailed datasets which are datasets that contain head classes that represent most of the overall dataset and then have tail classes that represent a small fraction of the overall dataset [Xu et al., 2022, Zhu et al., 2014]. Some approaches do propose integrating generic size constraints to the objective like in [Genevay et al., 2019, Höppner and Klawonn, 2008, Zhu et al., 2010], however these approaches directly deal with the input images (or data in general) instead of graphs.

In this paper, we propose a novel framework that introduces generic and at the same time stricter size constraints to the min-cut problem using Optimal Transport. To sum up, the main contributions of this work are :

- We formulate a problem for obtaining graph cuts that are balanced for an arbitrarily defined notion of size instead of specifically the volume or cardinality as is traditionally done in spectral clustering. We also propose a more general formulation of graph cuts with cluster size constraints which can help when dealing with perfectly balanced datasets and heavily imbalanced datasets such as long-tailed datasets.
- We then propose a solution for said problem through optimal transport using an approach reminiscent of the simplex algorithm and analyze its computational complexity. Links with existing works are also studied.
- Comprehensive experiments on balanced and long-tailed data sets using two variants we named OT- $ncut$ and OT- $rcut$ showcase the effectiveness of the proposed method compared to the most common min-cut algorithms (that use spectral clustering) both in terms of obtaining the desired cluster sizes as well as clustering performance. We release the code of our algorithm ¹ for reproducibility.

The rest of this paper is organized as follows : Preliminaries are presented in Section 2. Some related work is discussed in section 3. The OT-cut problem and algorithm along with their analysis and links to prior research are given in section 4. We present experimental results and analysis in section 5. Conclusions are then given in section 6.

2 Related Work

Our work is related with balanced clustering, as the latter is a special case of it, as well as with the more generic problem of constrained clustering.

Balanced Clustering A common class of constrained clustering problems is balanced clustering where we wish to obtain a partition with clusters of the same size. For example, in [Li et al., 2018], authors proposed to leverage the exclusive lasso on the k -means and min-cut problems to regulate the balance degree of the clustering results. [Lin et al., 2019] proposed a simplex algorithm to solve a minimum cost flow problem similar to k -means. In [Chen et al., 2017], authors proposed a self-balanced min-cut algorithm for image clustering implicitly using exclusive lasso as a balance regularizer in order to produce balanced partitions.

Constrained Clustering Some clustering approaches with generic size constraints, which can be seen as an extension of balanced clustering, also exist. In [Zhu et al., 2010], a heuristic algorithm to transform size constrained clustering problems into integer linear programming problems was proposed. Authors in [Ganganath et al., 2014] introduced a modified k -means algorithm which can be used to obtain clusters of preferred sizes. Clustering paradigms based on OT generally offer the

¹<https://github.com/chakib401/ot-cut>

possibility to set a target distribution for resulting partitions. [Genevay et al., 2019] proposed a deep clustering algorithm through optimal transport with entropic regularization. In [Fettal et al., 2022], authors proposed a way to perform biclustering which is an extension of clustering to bipartite graphs through Optimal Transport while choosing the size of the resulting biclusters.

3 Preliminaries

In what follows, $\Delta^n = \{\mathbf{p} \in \mathbb{R}_+^n \mid \sum_{i=1}^n p_i = 1\}$ denotes the n -dimensional standard simplex. $\Pi(\mathbf{w}, \mathbf{v}) = \{\mathbf{Z} \in \mathbb{R}_+^{n \times k} \mid \mathbf{Z}\mathbf{1} = \mathbf{w}, \mathbf{Z}^\top \mathbf{1} = \mathbf{v}\}$ denotes the transportation polytope, where $\mathbf{w} \in \Delta^n$ and $\mathbf{v} \in \Delta^k$ are the marginals of the joint distribution \mathbf{Z} and $\mathbf{1}$ is a vector of ones, its size can be inferred from the context. Matrices are denoted with uppercase boldface letters, and vectors with lowercase boldface letters. For a matrix \mathbf{M} , its i -th row is \mathbf{m}_i . Tr refers to the trace of a square matrix. Finally, \otimes denotes the Kronecker product.

3.1 Graph Cuts

Graph Cut Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with an weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ with $n = |\mathcal{V}|$, a cut is a partition of its vertices \mathcal{V} into two disjoint subsets \mathcal{A} and $\bar{\mathcal{A}}$. The value of a cut is given by

$$\text{cut}(\mathcal{A}) = \sum_{v_i \in \mathcal{A}, v_j \in \bar{\mathcal{A}}} w_{ij}. \quad (1)$$

Minimum k -cut Problem The goal of the minimum k -cut problem is to find a partition $(\mathcal{A}_1, \dots, \mathcal{A}_k)$ of the set of vertices \mathcal{V} into k different groups that is minimal in some metric. Intuitively, we wish for the edges between different subsets to have small weights, and for the edges within a subset have large weights. Formally, it is defined as

$$\text{min-cut}(\mathbf{W}, k) = \min_{\mathcal{A}_1, \dots, \mathcal{A}_k} \sum_{i=1}^k \text{cut}(\mathcal{A}_i). \quad (2)$$

This problem can also be stated as a trace minimization problem by representing the resulting partition $\mathcal{A}_1, \dots, \mathcal{A}_k$ using an assignment matrix \mathbf{X} such that for each row i , we have that

$$x_{ij} = \begin{cases} 1 & \text{if vertex } i \text{ is in } \mathcal{A}_j, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

This condition is equivalent to introducing two constraints which are $\mathbf{X} \in \{0, 1\}^{n \times k}$ and $\mathbf{X}\mathbf{1} = \mathbf{1}$. The minimum k -cut problem can then be formulated as

$$\text{min-cut}(\mathbf{W}, k) = \min_{\substack{\mathbf{X} \in \{0, 1\}^{n \times k} \\ \mathbf{X}\mathbf{1} = \mathbf{1}}} Tr(\mathbf{X}^\top \mathbf{L}\mathbf{X}), \quad (4)$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ refers to the graph Laplacian of the graph \mathcal{G} and \mathbf{D} is the diagonal matrix of degree of \mathbf{W} , i.e., $d_{ii} = \sum_j w_{ij}$.

Normalized Cut Problem In practice, solutions to the minimum k -cut problem do not yield satisfactory partitions due to the formation of small groups of vertices. Consequently, versions of the problem that take into account some notion of "size" for these groups have been proposed. The most commonly used one is normalized cut [Shi and Malik, 2000] :

$$\text{ncut}(\mathbf{W}, k) = \min_{\mathcal{A}_1, \dots, \mathcal{A}_k} \sum_{i=1}^k \frac{\text{cut}(\mathcal{A}_i)}{\text{vol}(\mathcal{A}_i)}, \quad (5)$$

since $\text{vol}(\mathcal{A}_i) = \mathbf{x}_i^\top \mathbf{D}\mathbf{x}_i$, then this problem can also be stated as a trace minimization problem :

$$\text{ncut}(\mathbf{W}, k) = \min_{\substack{\mathbf{X}\mathbf{1} = \mathbf{1} \\ \mathbf{X} \in \{0, 1\}^{n \times k}}} Tr \left(\frac{\mathbf{X}^\top \mathbf{L}\mathbf{X}}{\mathbf{X}^\top \mathbf{D}\mathbf{X}} \right). \quad (6)$$

A special case of the normalized graph cut is recovered by setting $\mathbf{D} = \mathbf{I}$ in problem 6. This problem is referred to as the ratio cut problem due to the different groups being normalized by their cardinality instead of their volumes :

$$\text{rcut}(\mathbf{W}, k) = \min_{\mathcal{A}_1, \dots, \mathcal{A}_k} \sum_{i=1}^k \frac{\text{cut}(\mathcal{A}_i)}{|\mathcal{A}_i|}, \quad (7)$$

and similarly to the normalized cut, since $|\mathcal{A}_i| = \mathbf{x}_i^\top \mathbf{x}_i$, we can formulate the ratio cut problem as a trace minimization problem :

$$\text{rcut}(\mathbf{W}, k) = \min_{\substack{\mathbf{X} \in \{0, 1\}^{n \times k} \\ \mathbf{X}\mathbf{1} = \mathbf{1}}} Tr \left(\frac{\mathbf{X}^\top \mathbf{L}\mathbf{X}}{\mathbf{X}^\top \mathbf{X}} \right). \quad (8)$$

Spectral Clustering for the Normalized & Ratio Cuts A common approach to solving the normalized graph cut problems, spectral clustering, replaces the partition constraints on \mathbf{X} with a form of semi-orthogonality constraints. In the case of rcut, we have

$$\text{ncut}(\mathbf{W}, k) = \min_{\substack{\mathbf{X} \in \mathbb{R}^{n \times k} \\ \mathbf{X}^\top \mathbf{X} = \mathbf{I}}} \text{Tr}(\mathbf{X}^\top \mathbf{L} \mathbf{X}). \quad (9)$$

On the other hand for ncut, the partition matrix \mathbf{X} is substituted with $\mathbf{H} = \mathbf{D}^{1/2} \mathbf{X}$ and a semi-orthogonality constraint is placed on this \mathbf{H} , i.e.,

$$\text{ncut}(\mathbf{W}, k) = \min_{\substack{\mathbf{H} \in \mathbb{R}^{n \times k} \\ \mathbf{H}^\top \mathbf{H} = \mathbf{I}}} \text{Tr}(\mathbf{H}^\top \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} \mathbf{H}). \quad (10)$$

A solution \mathbf{H} for the ncut problem is formed by stacking the first k -eigenvectors of the symmetrically normalized Laplacian $\mathbf{L}_s = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$ as its columns, and then applying a clustering algorithm such as k -means on its rows and assign the original data points accordingly [Ng et al., 2001]. The principle is the same for the spectral rcut algorithm.

3.2 Optimal Transport

Discrete optimal transport The goal of the optimal transport problem is to find a minimal cost transport plan \mathbf{X} between a source probability distribution of \mathbf{w} and a target probability distribution \mathbf{v} . Here we are interested in the discrete Kantorovich formulation of OT. When dealing with discrete probability distributions, said formulation is

$$\text{OT}(\mathbf{C}, \mathbf{w}, \mathbf{v}) \triangleq \min_{\mathbf{X} \in \Pi(\mathbf{w}, \mathbf{v})} \langle \mathbf{C}, \mathbf{X} \rangle, \quad (11)$$

where $\mathbf{C} \in \mathbb{R}^{n \times k}$ is the cost matrix, and c_{ij} quantifies the effort needed to transport a probability mass from \mathbf{w}_i to \mathbf{v}_j .

Discrete Gromov-Wasserstein Distance The discrete Gromov-Wasserstein (GW) distance [Mémoli, 2011] is an extension of optimal transport to the case where the source and target distributions are defined on different metric spaces :

$$\text{GW}(\mathbf{M}, \mathbf{M}', \mathbf{w}, \mathbf{v}) \triangleq \min_{\mathbf{X} \in \Pi(\mathbf{w}, \mathbf{v})} \langle L(\mathbf{M}, \mathbf{M}') \otimes \mathbf{X}, \mathbf{X} \rangle \quad (12)$$

where $\mathbf{M} \in \mathbb{R}^{n \times n}$ and $\mathbf{M}' \in \mathbb{R}^{k \times k}$ are similarity matrices defined on the source space and target space respectively, and $L : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a divergence measure between scalars, $L(\mathbf{M}, \mathbf{M}')$ is the $n \times n \times k \times k$ tensor of all pairwise divergences between the elements of \mathbf{M} and \mathbf{M}' .

4 Graph Cuts with Arbitrary Size Constraints via OT

Graph Cuts via Optimal Transport As already mentioned, the good performance of the normalized cut algorithm comes from the normalization by the volume of each group in the cut. However, the size constraint is not a hard one, meaning that obtained groups are not of exactly the same volume. This leads us to propose to replace the volume normalization by a strict balancing constraint as follows :

$$\min_{\mathcal{A}_1, \dots, \mathcal{A}_k} \sum_{i=1}^k \text{cut}(\mathcal{A}_i) \quad \text{such that} \quad \text{vol}(\mathcal{A}_1) = \dots = \text{vol}(\mathcal{A}_k). \quad (13)$$

Similarly to the ncut problem, this problem can be formulated as a trace minimization problem :

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times k}} \text{Tr}(\mathbf{X}^\top \mathbf{L} \mathbf{X}) \quad \text{such that} \quad \mathbf{X} \mathbf{1} = \mathbf{D} \mathbf{1}, \quad \mathbf{X}^\top \mathbf{1} = \frac{1}{k} \mathbf{1}. \quad (14)$$

This problem is combinatorial and thus may appear difficult to optimize. However, this is problem can be slightly modified to become an instance of the Gromov-Wasserstein problem, to which exists relatively efficient heuristics exist. Specifically, the volume constraint can be dropped by rather defining \mathbf{X} to be an element of the transportation polytope instead of being a partition matrix. The degrees are also normalized by dividing them by their total sum and then representing them as proportions instead of absolute quantities, yielding the following problem :

$$\min \text{Tr}(\mathbf{X}^\top \mathbf{L} \mathbf{X}) \quad \text{such that} \quad \mathbf{X} \in \Pi \left(\frac{1}{\sum_i d_{ii}} \mathbf{D} \mathbf{1}, \frac{1}{k} \mathbf{1} \right) \quad (15)$$

This formulation is a special case of the Gromov-Wasserstein problem for a source space whose similarity matrix is $\mathbf{C} = \mathbf{L}$ and whose destination similarity matrix is $\mathbf{C}' = \mathbf{I}$.

Note that a ratio cut version can easily be obtained by replacing the volume constraint with

$$|A_1| = \dots = |A_k| \quad (16)$$

in problem 14, and similarly in problem 15, by setting $\mathbf{D} = \mathbf{I}$, giving rise to :

$$\min Tr(\mathbf{X}^\top \mathbf{L} \mathbf{X}) \quad \text{such that} \quad \mathbf{X} \in \Pi\left(\frac{1}{n} \mathbf{1}, \frac{1}{k} \mathbf{1}\right) \quad (17)$$

Graph Cuts with Size Constraints From the previous problem, it is easy to see that target distribution does not need to be uniform, and thus any distribution can be considered, leading to further applications like long-tailed dataset clustering. Another observation is that any notion of size can be considered and not only the volume or cardinality of the formed node groups. We formulate the generic optimal transport graph cut problem as :

$$\min_{\mathbf{X} \in \Pi(\pi^s, \pi^t)} Tr(\mathbf{X}^\top \mathbf{L} \mathbf{X}) \equiv \min_{\mathbf{X} \in \Pi(\pi^s, \pi^t)} \langle \mathbf{L} \mathbf{X}, \mathbf{X} \rangle, \quad (18)$$

where π_i^s is the relative 'size' of the element i and π_j^t is the desired relative 'size' of the group j . Through the form that uses the Frobenius product, it is easy to see how our problem is related to the Gromov-Wasserstein problem.

While using the Laplacian matrix makes the problem resemble the graph cut problem more closely, in practice, we will be using a heuristic approach that consists in solving intermediate optimal transport problems (as shown in algorithm 1). Since the Laplacian matrix \mathbf{L} contains positive and negative values, it harder to interpret as a cost matrix and thus making the resolution of these intermediate problems using OT solvers unstable. Thus, during the resolution, we will be considering a cost matrix $\mathbf{M} = -\mathbf{W}$ instead of the graph Laplacian.

Transport Plans as Partition Matrices The proposed approach relies on the fact that the transport plan \mathbf{X} can be interpreted as a partition matrix. Fortunately, this interpretation can be made through the concept of h -almost hard clustering [Fettal et al., 2022] :

Definition 1 (h -almost hard clustering). *An h -almost hard clustering is a clustering whose partition matrix is $\mathbf{C} \in \mathbb{R}^{n \times k}$ such that $\|\mathbf{C}\|_0 = n + h$ and for each row c of \mathbf{C} we have that $\|c\|_0 > 0$, where $\|\cdot\|_0$ is the zero norm. When $h = 0$, we obtain a standard hard clustering with one non-zero element per row.*

The extreme points of the transportation polytope are always h -almost hard clustering (see [Peyré et al., 2019, Fettal et al., 2022] for a proof), so we add a boundary condition to our problem in order to always obtain a transport plan \mathbf{X} that can be interpreted easily as a hard partition matrix :

$$\text{OT-cut}(\mathbf{W}, k) \triangleq \min Tr(\mathbf{X}^\top \mathbf{M} \mathbf{X}) \quad \text{such that} \quad \mathbf{X} \in \text{ext}(\Pi(\pi^s, \pi^t)), \quad (19)$$

where ext is the set of extreme points of its argument. Consequently we have that :

Proposition 1. *A solution \mathbf{X} to the OT-cut problem is an h -almost hard clustering with $h \in \{0, \dots, k-1\}$.*

The OT-ncut version is obtained by setting $\pi_s = \frac{1}{\sum_i d_{ii}} \mathbf{D} \mathbf{1}$:

$$\text{OT-ncut}(\mathbf{W}, k) \triangleq \min Tr(\mathbf{X}^\top \mathbf{M} \mathbf{X}) \quad \text{such that} \quad \mathbf{X} \in \text{ext}\left(\Pi\left(\frac{1}{\sum_i d_{ii}} \mathbf{D} \mathbf{1}, \pi^t\right)\right) \quad (20)$$

while for the OT-rcut criterion, we set $\pi_s = \frac{1}{n}$, yielding :

$$\text{OT-rcut}(\mathbf{W}, k) \triangleq \min Tr(\mathbf{X}^\top \mathbf{M} \mathbf{X}) \quad \text{such that} \quad \mathbf{X} \in \text{ext}\left(\Pi\left(\frac{1}{n} \mathbf{1}, \pi^t\right)\right), \quad (21)$$

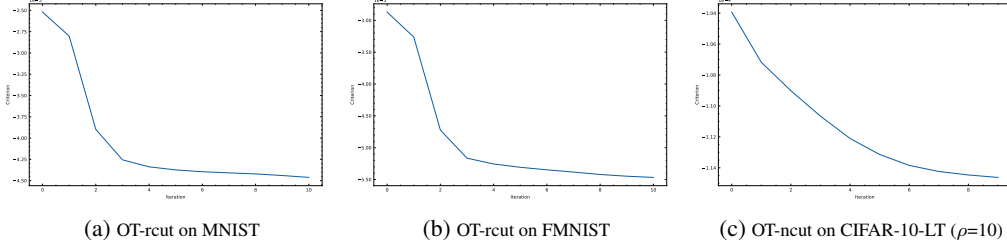


Figure 1: Evolution of Loss as function of the number of iterations.

Optimization and Complexity To solve the OT-cut problem, we propose to follow the principle of the simplex algorithm and to visit an extreme point the transportation polytope with a better objective value than the previous one at each iteration. We do this by introducing what basically comes down to a variable substitution as follows :

$$\text{OT-cut}(\mathbf{W}, k) \triangleq \min_{\mathbf{X}^{(k+1)} \in \text{ext}(\Pi(\pi^s, \pi^t))} \langle \mathbf{M}\mathbf{X}^{(k)}, \mathbf{X}^{(k+1)} \rangle. \quad (22)$$

Given a current solution $\mathbf{X}^{(k)}$, we find the next solution $\mathbf{X}^{(k+1)}$ by solving this optimal transport problem :

$$\mathbf{X}^{(k+1)} \leftarrow \arg \text{OT}(\mathbf{L}\mathbf{X}^{(k)}, \pi^s, \pi^t). \quad (23)$$

A resolution is possible by stating it as the earth-mover’s distance problem [Hitchcock, 1941] which can be solved via the network simplex algorithm. The loss function is guaranteed to decrease between consecutive iterations and it converges to a local minima in few iteration based on the initial guess $\mathbf{X}^{(0)}$, for example, in figure 1, we can see the evolution of loss function OT-rcut on MNIST, OT-rcut on FMNIST and OT-ncut on CIFAR-10-LT ($\rho = 10$). The pseudocode for the optimization procedure is presented in algorithm 1.

Possible heuristics to improve the quality of the final solution would be doing multiple runs with different initializations, or initializing the algorithm with a partition matrix obtained from a spectral-cut algorithm projected onto the transportation polytope.

Proposition 2. For a graph with $|\mathcal{E}|$ edges and n nodes, the complexity of an iteration of the proposed algorithm is

$$\mathcal{O}(kn^2 \log n)$$

Proof. To prove this, we note that in practice $n \gg k$ and that the complexity of the network simplex algorithm for some graph $\mathcal{G}_{emd} = (\mathcal{V}_{emd}, \mathcal{E}_{emd})$ is in $\mathcal{O}(|\mathcal{V}_{emd}| |\mathcal{E}_{emd}| \log |\mathcal{E}_{emd}|)$ [Orlin, 1997]. In our case, this graph has $|\mathcal{V}_{emd}| = n + k$ (since $n \gg k$, we can drop the k) and $|\mathcal{E}_{emd}| = nk$. The other operation that is performed during each iteration is the matrix multiplication $\mathbf{M}\mathbf{X}^{(k)}$ whose complexity is in $\mathcal{O}(k|\mathcal{E}|)$, in the worst case when matrix \mathbf{M} is fully dense, we have that $|\mathcal{E}| = n^2$. Note that the complexity of the spectral clustering algorithm is generally in $\mathcal{O}(kn^2)$. \square

Algorithm 1: OT-cut

Input : \mathbf{A} adjacency matrix, π^s node size distribution, π^t group size distribution, \mathbf{G}_{init} initial partition matrix.

Output : \mathbf{G} partition of the graph

$\mathbf{X}^{(0)} \leftarrow \arg \text{OT}(\mathbf{G}_{init}, \pi^s, \pi^t);$

while not converged do

 | $\mathbf{X}^{(k+1)} \leftarrow \arg \text{OT}(\mathbf{M}\mathbf{X}^{(k)}, \pi^s, \pi^t);$

end

Generate partition matrix \mathbf{G} such that each node v_i is assigned it to partition $\arg \max_i x_i;$

5 Connections with Prior Works

Gromov-Wasserstein Learning for Graphs The Gromov-Wasserstein partitioning paradigm [Xu et al., 2019] supposes that the Gromov-Wasserstein discrepancy can uncover the clustering structure

of the observed source graph \mathcal{G} when the target graph \mathcal{G}_{dc} only contains weighted self-connected isolated nodes, this means that its adjacency matrix is diagonal. The weights of this diagonal matrix as well as the source and target distribution are a special function of the node degrees. Their approach uses a regularized proximal gradient method as well as a recursive partitioning scheme and can be used in a multi-view clustering setting.

Optimal-Transport Based Biclustering Biclustering is the extension of clustering to bipartite graphs. Here, we recover the BCOT [Fettal et al., 2022] problem as a special case of OT-cut. Given a bipartite adjacency matrix \mathbf{A} , we recover their formulation through ours by considering this anti-adjacency matrix $\bar{\mathbf{A}}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}_{n \times n} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{0}_{d \times d} \end{bmatrix}, \quad \bar{\mathbf{A}} = \begin{bmatrix} \infty_{n \times n} & L(\mathbf{B}) \\ L(\mathbf{B})^\top & \infty_{d \times d} \end{bmatrix}$$

Then setting $\pi^s = [\mathbf{v}, \mathbf{w}]^\top$ and $\pi^t = [\mathbf{v}, \mathbf{w}]^\top$ and omitting the boundary condition. All in all, we have that

$$\text{BCOT}(L(\mathbf{B}), \mathbf{w}, \mathbf{v}) \equiv \text{OT-cut}(\bar{\mathbf{A}}, [\mathbf{v}, \mathbf{w}]^\top, [\mathbf{w}, \mathbf{v}]^\top)$$

OT Kernel k-Means In [Genevay et al., 2019], authors proposed an algorithm for k-means with cluster size constraints and entropic regularization. By dropping the regularization and adding a boundary constraint, one can think of the case where the adjacency matrix in our formulation is a kernel matrix and use the same principles that were used with kernel k-means [Dhillon et al., 2004] to optimize the OT graph cut criterion.

6 Experiments

We ran experiments on balanced and heavily imbalanced (long-tailed) datasets. We evaluated the clustering performance of three variants of each of OT-ncut and OT-rcut algorithms against the spectral rcut and ncut algorithms, as well as the ability of our approach to recover the desired partition distribution.

6.1 Datasets

We perform experiments on balanced datasets and long-tailed datasets. The statistical summaries of these datasets are available in table 1. The CIFAR-10-LT ($\rho = \frac{\max_i n_i}{\min_i n_i}$) variants, are generated using a long-tailed imbalance sampling method that yields a dataset whose majority class is ρ times more frequent than the minority class. Long-tailed imbalance follows an exponential decay in sample sizes across different classes.

6.2 Metrics

The evaluation is straightforward, we adopt four popular clustering metrics when dealing with the balanced datasets : clustering accuracy (CA), normalized mutual information (NMI), adjusted rand index (ARI) [Hubert and Arabie, 1985], and clustering F1 score (CF1). When dealing with the long-tailed dataset, we only use metrics that are sensitive to imbalance NMI, ARI, and CF1. When comparing the concordance of the input cluster distribution π and the cluster distribution obtained via one of our algorithms $\hat{\pi}$, we use the Kullback-Leibler divergence. The concordance of two perfectly matching distributions will be equal to zero, otherwise it will be larger .

6.3 Experimental settings

We compare two variants of our algorithm, namely, the OT-ncut and OT-rcut implemented via the Python optimal transport package (POT) [Flamary et al., 2021] to the spectral clustering variants SC-ncut and SC-rcut which were based on the implementation of the spectral clustering in the Scikit-Learn package [Pedregosa et al., 2011]. For each image dataset represented in matrix form as \mathbf{Y} , we use subspace Least Squares Regression Subspace Clustering (LSR) [Lu et al., 2012] to create the graph, we get $\mathbf{A} = \mathbf{Y}\mathbf{Y}^\top (\mathbf{Y}\mathbf{Y}^\top + \mathbf{I})^{-1}$. Note that all experiments are run five times. In the results tables, base OT-cuts variants use random initialization. The variants that end with * use the

Table 1: Characteristics of the datasets. The balance score is the ratio of the number of occurrences of the most frequent class over that of the least frequent class.

| Dataset | #Images | Dimension | #Classes | Balance |
|---|---------|-----------|----------|---------|
| MNIST [Deng, 2012] | 60,000 | 784 | 10 | 1.0 |
| Fashion-MNIST [Xiao et al., 2017] | 60,000 | 334 | 10 | 1.0 |
| KMNIST [Clanuwat et al., 2018] | 60,000 | 784 | 10 | 1.0 |
| CIFAR-10 [Krizhevsky et al., 2009] | 50,000 | 1024 | 10 | 1.0 |
| CIFAR-10-LT ($\rho = 5$) | 25,423 | 1024 | 10 | 5.0 |
| CIFAR-10-LT ($\rho = 10$) | 20,431 | 1024 | 10 | 10.0 |
| CIFAR-10-LT ($\rho = 20$) | 17,023 | 1024 | 10 | 20.0 |
| CIFAR-10-LT ($\rho = 100$) [Cao et al., 2019] | 12,406 | 1024 | 10 | 100.0 |

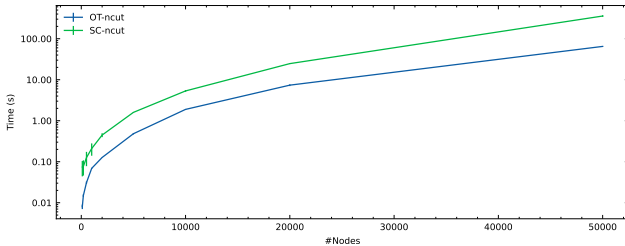


Figure 2: Training times of OT-ncut in seconds (log scale) over subsets of different sizes of MNIST.

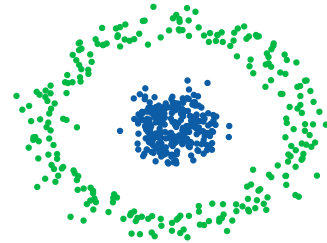


Figure 3: OT-rcut clusters on a toy dataset (RBF kernel with $\gamma = 3$).

ground truth target distribution. Finally, for variants ending in $_{SC}$, we choose the initial transport plan $\mathbf{X}^{(0)}$ by first obtaining a partition matrix through the corresponding spectral clustering algorithm, i.e., spectral ncut (SC-ncut) for OT-ncut and spectral rcut (SC-rcut) to OT-rcut. We perform 10 iterations of our algorithm to fine-tune the initial guesses of spectral cuts and perform 20 iterations when using random initialization.

6.4 Results

The results on the two dataset types are presented in what follows. We emphasize that the most important metric to see is the relative difference instead of the absolute value of the metrics since **the goal is not to learn a better graph** over the dataset but rather to get a **better cut over said graph**.

Performance on balanced datasets For balanced datasets, the results are reported in table 3. One of our two approaches yields the best results in all 16 cases. Furthermore, each one of them improves over the results of their spectral counterpart, exceeding them in 31 out of 32 cases. Notably, our OT-rcut variant achieves a significant improvement over the spectral ratio cut algorithm. In figure 1, we see the results of OT-rcut on a toy dataset (two circles) using an RBF kernel.

Table 2: Image clustering performance on the imbalanced (long-tail) datasets. Values are the averages over five runs. **Standard deviations were not reported due to being negligible (≤ 0.1).**

| | CIFAR-10-LT ($\rho = 5$) | | | CIFAR-10-LT ($\rho = 10$) | | | CIFAR-10-LT ($\rho = 20$) | | | CIFAR-10-LT ($\rho = 100$) | | |
|-------------------|----------------------------|------------|-------------|-----------------------------|------------|-------------|-----------------------------|------------|-------------|------------------------------|------------|-------------|
| | NMI | ARI | CF1 | NMI | ARI | CF1 | NMI | ARI | CF1 | NMI | ARI | CF1 |
| rcut | 0.1 | -0.0 | 3.3 | 0.1 | -0.0 | 4.0 | 0.1 | -0.0 | 4.6 | 0.1 | -0.0 | 5.8 |
| OT-rcut | 11.6 | 7.3 | 20.7 | 12.1 | 7.8 | 19.8 | 11.4 | 7.5 | 17.9 | 9.8 | 5.7 | 13.7 |
| OT-rcut $_{SC}$ | 11.1 | 6.4 | 20.8 | 10.6 | 6.5 | 18.7 | 11.3 | 7.4 | 17.1 | 9.8 | 5.8 | 13.7 |
| OT-rcut $_{SC}^*$ | 11.2 | 6.1 | 19.5 | 10.5 | 5.4 | 16.6 | 10.8 | 5.4 | 14.6 | 11.6 | 5.6 | 14.3 |
| ncut | 10.2 | 5.6 | 19.1 | 10.5 | 6.2 | 18.0 | 10.6 | 5.8 | 16.4 | 12.7 | 6.8 | 14.6 |
| OT-ncut | 12.0 | 8.3 | 21.3 | 10.1 | 7.3 | 18.9 | 10.6 | 7.9 | 17.3 | 8.4 | 6.9 | 13.8 |
| OT-ncut $_{SC}$ | 10.8 | 7.5 | 20.7 | 10.8 | 7.5 | 18.6 | 10.5 | 7.8 | 16.2 | 10.4 | 8.3 | 14.8 |
| OT-ncut $_{SC}^*$ | 10.4 | 5.9 | 20.4 | 10.4 | 5.6 | 18.0 | 10.6 | 5.7 | 16.4 | 10.9 | 5.6 | 13.1 |

Table 3: Clustering performance on balanced image datasets. Values are the averages over five runs. **Standard deviations were not reported due to being negligible (≤ 0.1).** OT-rcut* has the same results since the ground truth sizes are uniform, similarly, OT-rcut_{SC} has the also same results due to SC-rcut returning a bad guess that is equivalent to a random initialization.

| | MNIST | | | | FMNIST | | | | KMNIST | | | | CIFAR 10 | | | |
|------------------------|-------------------------|------|------|------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|-------------|
| | ACC | NMI | ARI | CF1 | ACC | NMI | ARI | CF1 | ACC | NMI | ARI | CF1 | ACC | NMI | ARI | CF1 |
| SC-ncut | 40.2 | 34.7 | 17.6 | 37.6 | 53.4 | 53.2 | 36.5 | 51.6 | 37.3 | 30.4 | 19.8 | 35.2 | 22.2 | 10.1 | 5.6 | 21.5 |
| OT-ncut _{SC} | 48.2 | 36.4 | 27.1 | 48.2 | 47.8 | 51.0 | 35.5 | 47.0 | 43.6 | 33.2 | 24.2 | 42.8 | 21.5 | 11.5 | 6.4 | 21.3 |
| OT-ncut* _{SC} | 41.5 | 35.5 | 25.0 | 41.1 | 56.3 | 53.0 | 40.7 | 56.0 | 44.7 | 33.6 | 24.2 | 44.5 | 23.0 | 10.8 | 5.8 | 23.0 |
| SC-rcut | 11.2 | 0.0 | -0.0 | 2.0 | 10.0 | 0.0 | 0.0 | 1.8 | 10.0 | 0.0 | 0.0 | 1.8 | 10.0 | 0.0 | 0.0 | 1.8 |
| OT-rcut | 38.3 | 32.3 | 20.7 | 38.2 | 54.3 | 53.9 | 39.0 | 54.3 | 41.6 | 33.1 | 22.3 | 41.6 | 23.8 | 11.7 | 6.4 | 23.8 |
| OT-rcut _{SC} | Same results as OT-rcut | | | | | | | | | | | | | | | |
| OT-rcut* _{SC} | Same results as OT-rcut | | | | | | | | | | | | | | | |

Table 4: The Kullback-Leibler divergence between the desired input target distribution and the one obtained using OT-cut variants.

| | OT-ncut* | OT-ncut | OT-rcut* | OT-rcut |
|------------------------------|----------|----------|----------|---------|
| MNIST | 3.00e-09 | 2.30e-09 | 0.0 | 0.0 |
| FMNIST | 1.70e-09 | 2.60e-09 | 0.0 | 0.0 |
| KMNIST | 2.50e-09 | 4.10e-09 | 0.0 | 0.0 |
| CIFAR-10 | 3.70e-09 | 3.10e-09 | 0.0 | 0.0 |
| CIFAR-10-LT ($\rho = 5$) | 0.0 | 2.53e-07 | 1.62e-08 | 0.0 |
| CIFAR-10-LT ($\rho = 10$) | 1.48e-08 | 6.70e-09 | 1.08e-08 | 0.0 |
| CIFAR-10-LT ($\rho = 20$) | 7.00e-09 | 1.26e-08 | 7.07e-08 | 0.0 |
| CIFAR-10-LT ($\rho = 100$) | 8.62e-08 | 5.14e-08 | 7.80e-08 | 0.0 |

Performance on long-tailed datasets Table 2 presents the results obtained on the long-tailed datasets. In all cases, OT-ncut and OT-rcut outperform their spectral clustering counterparts, yielding the best performance in 11 out of 12 cases. Notably, the improvement of OT-rcut over SC-rcut is particularly significant, consistent with the findings in the balanced case.

Concordance of the Desired & Resulting Cluster Sizes To evaluate our algorithm’s ability to produce a partition with the desired group size distribution, we use the Kullback-Leibler (KL) divergence metric. Specifically, we compare the distribution obtained by our OT-rcut and OT-ncut variants against the target distribution specified as a hyperparameter (π). Table 4 presents the KL divergence values for both variants on various datasets. Our approaches achieve near-perfect performance on most datasets. Notably, OT-rcut is able to fully recover the desired group sizes.

Running Time As Figure 2 shows, OT-cut with random initialization is more scalable than the spectral ncut algorithm, significantly outspeeding it on all subsets of MNIST despite being theoretically more complex. This is due to the fact that our algorithm needs few iterations to converge.

7 Conclusion

In this paper we proposed a new graph cut algorithm for partitioning with arbitrary size constraints through optimal transport. This approach generalizes the concept of the normalized and ratio cut to arbitrary size distributions and this for any notion of size. The proposed algorithm works well when used in conjunction with a classical spectral graph cut algorithm as a post-processing step to obtain some desired distribution. Experiments on balanced and imbalanced datasets showed the effectiveness of our approach both in terms of clustering performance, computational speed, as well as its ability to recover partitions that almost perfectly match the desired ones.

References

- Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, 1998.
- Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9779–9788, 2020.
- Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4247–4256, 2021.
- Xiao Cai, Feiping Nie, Weidong Cai, and Heng Huang. New graph structured sparsity model for multi-label image annotations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 801–808, 2013.
- Jie Cao, Zhiang Wu, Junjie Wu, and Wenjie Liu. Towards information-theoretic k-means clustering for image indexing. *Signal Processing*, 93(7):2026–2037, 2013.
- Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.
- 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.
- Xiaojun Chen, Joshua Zhexue Haung, Feiping Nie, Renjie Chen, and Qingyao Wu. A self-balanced min-cut algorithm for image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2061–2069, 2017.
- Xiaojun Chen, Renjie Chen, Qingyao Wu, Yixiang Fang, Feiping Nie, and Joshua Zhexue Huang. Labin: balanced min cut for large-scale data. *IEEE transactions on neural networks and learning systems*, 31(3): 725–736, 2019.
- Qimin Cheng, Qian Zhang, Peng Fu, Conghuan Tu, and Sen Li. A survey and analysis on automatic image annotation. *Pattern Recognition*, 79:242–259, 2018.
- Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.
- Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, 2004.
- Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.
- Chakib Fettaf, Lazhar Labiod, and Mohamed Nadif. Efficient and effective optimal transport-based biclustering. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021.
- Nuwan Ganganath, Chi-Tsun Cheng, and K Tse Chi. Data clustering with cluster size constraints using a modified k-means algorithm. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 158–161. IEEE, 2014.
- Aude Genevay, Gabriel Dulac-Arnold, and Jean-Philippe Vert. Differentiable deep clustering with cluster size constraints. *arXiv preprint arXiv:1910.09036*, 2019.
- Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

- Frank L Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of mathematics and physics*, 20(1-4):224–230, 1941.
- Frank Höppner and Frank Klawonn. Clustering with size constraints. In *Computational Intelligence Paradigms*, pages 167–180. Springer, 2008.
- Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.
- Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. *Advances in neural information processing systems*, 30, 2017.
- Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5374–5384, 2022.
- Zhihui Li, Feiping Nie, Xiaojun Chang, Zhigang Ma, and Yi Yang. Balanced clustering via exclusive lasso: A pragmatic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Weibo Lin, Zhu He, and Mingyu Xiao. Balanced clustering: A uniform model and fast algorithm. In *IJCAI*, pages 2987–2993, 2019.
- Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, pages 347–360. Springer, 2012.
- Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.
- Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.
- James B Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.
- Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.
- Yue Xu, Yong-Lu Li, Jiefeng Li, and Cewu Lu. Constructing balance from imbalance for long-tailed image recognition. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022.
- Shunzhi Zhu, Dingding Wang, and Tao Li. Data clustering with size constraints. *Knowledge-Based Systems*, 23(8):883–889, 2010.
- Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2014.