Learning Common Semantics via Optimal Transport for Contrastive Multi-view Clustering

Qian Zhang†, Lin Zhang†, Ran Song*, Runmin Cong, Yonghuai Liu, and Wei Zhang

Abstract—Multi-view clustering aims to learn discriminative representations from multi-view data. Although existing methods show impressive performance by leveraging contrastive learning to tackle the representation gap between every two views, they share the common limitation of not performing semantic alignment from a global perspective, resulting in the undermining of semantic patterns in multi-view data. This paper presents CSOT, namely Common Semantics via Optimal Transport, to boost contrastive multi-view clustering via semantic learning in a common space that integrates all views. Through optimal transport, the samples in multiple views are mapped to the joint clusters which represent the multi-view semantic patterns in the common space. With the semantic assignment derived from the optimal transport plan, we design a semantic learning module where the soft assignment vector works as a global feature weighting strategy to treat samples differently according to their semantic significance, which improves the effectiveness of cross-view contrastive representation learning. Extensive experimental results demonstrate that CSOT achieves the state-of-the-art clustering performance.

Index Terms—Multi-view clustering, semantic alignment, optimal transport, contrastive learning.

I. INTRODUCTION

In many scenarios, data are typically collected via various sensors or from different perspectives, resulting in the widespread availability of multi-view data that reflect multiple properties of objects. Humans are capable of integrating properties from multiple views to distinguish among different objects, which however remains a challenge for deep models. As an unsupervised technique for exploring the underlying semantic patterns in multi-view data, multi-view clustering (MVC) has attracted increasing attention in the community of machine learning.

This work was supported in part by the National Natural Science Foundation of China under Grants 61919141, 62076148, and U22A2057, in part by the National Science and Technology Major Project under Grant 2022ZD0012002, in part by the Shandong Excellent Young Scientists Fund Program (Overseas) under Grant 2022HWYQ-042, in part by the Young Taishan Scholars Program of Shandong Province No.tsqn201909029, and in part by Project for Self-Developed Innovation Team of Jinan City under Grant 2021GXRC038.

Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, and Wei Zhang are with the School of Control Science and Engineering, Shandong University, Jinan 250061, China.

Yonghuai Liu is with the Department of Computer Science, Edge Hill University, Ormskirk L39 4QF, UK.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org provided by the author. The material includes detailed demonstration of the proposed method and additional experimental results. Contact ransong@sdu.edu.cn for further questions about this work.

†These authors contributed equally to this work.

*Corresponding author: Ran Song (Email: ransong@sdu.edu.cn).

Traditional MVC methods [1]–[6] usually leverage the original multi-view features or shallow features to obtain the consensus information among multi-view data. Although they are computationally efficient owing to a small number of algorithmic parameters, relatively simple objective functions, simplified formalization of the problems at hand, and limited data for training, the clustering performance of such methods often falls short due to the presence of noise, redundancy, appearance and disappearance in the data and thus inexpressive features.

In recent years, a number of deep MVC methods have been proposed, which utilize deep learning techniques to extract multi-view features and have achieved competitive performance on several public datasets [7]–[13]. The state-of-the-art deep MVC methods [14]–[20] usually learn both view-specific and view-common features for clustering in a self-supervised manner, where the features of each view are extracted via a view-specific encoder before being passed into the intra-view reconstruction and cross-view contrastive learning modules. For instance, Lin et al. [14] utilized a within-view reconstruction loss to learn view-specific representations and a mutual information-based contrastive loss to handle the representation bias across multiple views. To remedy the conflict that the intra-view reconstruction loss aims to learn view-specific information while the cross-view contrastive loss attempts to preserve representation consistency across all views, Xu et al. [17] proposed a multi-level feature learning framework where the two losses were computed with low-level and high-level features, respectively. More recently, Trosten et al. [18] built a unified framework for deep MVC containing both single-view and cross-view self-supervised learning branches. Yan et al. [19] proposed to learn a consensus representation for multiple views based on the structural relationship among samples, and then performed the structure-guided contrastive learning to exploit the complementary information of similar samples.

However, there exist two main issues in previous deep MVC methods:

1) The essence of MVC is to identify the common semantic patterns across all the views prone to the semantic bias caused by the gaps among feature distributions of multiple views. Mapping all the views into a unified feature space enables the exploration of semantic associations among multiple views from a global perspective, which helps to handle the cross-view semantic bias. However, previous methods focus on directly improving the pair-wise consistency between every two views, making it difficult to correctly find all the common semantic patterns from multiple views.

2) Existing deep MVC methods are lacking a global perspective to align the representations across multiple views, which often falls short due to the presence of noise, redundancy, appearance and disappearance in the data and thus inexpressive features.
This work in three main research areas including multi-view gaps in clustering multi-view data, as illustrated in Fig. 1. The method can effectively mitigate the side effects of the view and the joint clusters in the common space, the proposed allows CSOT to benefit from the hard sample mining strategy. With such an estimation, we design a novel semantic-aware vector. Moreover, we estimate the semantic significance of each sample based on the soft semantic assignment matrix. Then for each sample, we introduce a semantic-consistent constraint to preserve the consistency between its soft semantic assignment in the common space and the predicted label constraint to preserve the consistency between its soft semantic assignment in the common space and the predicted label.

To address the above two issues, we propose to learn Common Semantics via Optimal Transport (CSOT) for contrastive learning-based MVC. CSOT first creates a unified feature space (i.e. common space in this paper) by integrating multi-view features, which provides a global platform for discriminating samples from multiple views. Although Yan et al. [19] implicitly built a common space via feature concatenation for multi-view representation learning, they only mined the instance-level consistency between the concatenated representation and the view-specific one. Instead, we employ the joint clusters, i.e. the clusters in the common space, as a comprehensive representation of common semantics for all views. The joint clusters involve rich view-common information to represent semantic patterns in a more uniform manner, and thus the cross-view semantic bias caused by the view gaps can be significantly reduced. To pass the semantic knowledge in the common space back to each view, we adopt optimal transport (OT) as an efficient solution for transporting each sample to the joint clusters. Consequently, the semantic assignment for each sample in the common space can be easily obtained. Then for each sample, we introduce a semantic-consistent constraint to preserve the consistency between its soft semantic assignment in the common space and the predicted label vector. Moreover, we estimate the semantic significance of each sample based on the soft semantic assignment matrix. With such an estimation, we design a novel semantic-aware re-weighting mechanism to adaptively adjust the significance of each view during contrastive representation learning, which allows CSOT to benefit from the hard sample mining strategy.

By exploring the semantic correlations among each view and the joint clusters in the common space, the proposed method can effectively mitigate the side effects of the view gaps in clustering multi-view data, as illustrated in Fig. 1. The main contributions of this paper are summarised as follows:

1) We propose CSOT, a novel deep contrastive framework for MVC that involves OT-based semantic information from the common space, to bridge the view gaps and improve the learning of cross-view common semantics.
2) We propose a semantic-aware re-weighting mechanism for effective cross-view contrastive representation learning based on the semantic significance of each sample estimated by CSOT.
3) Extensive experiments on several MVC datasets demonstrate that CSOT outperforms the state-of-the-art methods by a large margin.

II. RELATED WORKS

This section briefly reviews existing methods related to this work in three main research areas including multi-view clustering, optimal transport, and contrastive learning.

A. Multi-view Clustering

There are several kinds of traditional MVC methods, including subspace-based methods [2], [4], [5], kernel-based methods [1], [3], graph-based methods [6], [21], [22], spectral-based methods [23], [24], etc. For example, Jia et al. [23] designed a novel tailored tensor low-rank representation for handling multi-view spectral clustering. These methods differ in their assumptions about the underlying structure of the data and the way in which they combine information from multiple views. However, most of these methods suffer from poor representation ability of the redundant, inconsistent, or unique multi-view data, which limits the practical application in real-world scenarios.

In recent years, deep learning has demonstrated excellent performance in clustering various types of data, including but not limited to images, texts, and graphs [25], [26]. This has largely driven deep learning-based methods to gradually become the mainstream approach in the MVC domain. Learning joint representation across all the views has been considered to be a promising solution for MVC. For example, Yin et al. [27] proposed a multi-view variational autoencoder network to deal with MVC, where the shared latent embeddings and the multi-view combination weights can be learned simultaneously to obtain the final clustering assignment. Huang et al. [28] addressed MVC via jointly learning latent representations and graph structures, which work in a mutually beneficial manner and effectively promote clustering performance. As contrastive learning (CL) has been delivering impressive results in various unsupervised tasks [29], [30], several CL-based MVC methods have emerged [14], [15], [17], [19]. These methods leverage CL to explicitly align cross-view representations, thus mitigating the side effects of inconsistency caused by view gaps. Trosten et al. [16] proposed CoMVC with a selective contrastive alignment to achieve common representation learning while preserving the view-prioritization in the final feature.
space. Xu et al. [17] designed a multi-level feature learning framework for MVC, where CL is applied on both the feature level and label level. More recently, Yan et al. [19] not only applied global and cross-view feature aggregation to explore the relationships among samples but also developed a structure-guided CL strategy to enhance the consensus representation of multi-view data.

Although the CL-based deep MVC methods have significantly boosted the clustering performance, they neither explore the cross-view semantic correlations in a global perspective nor make full use of the semantically significant samples, which may affect the process of finding common semantic patterns for MVC. In this paper, we aim to overcome these two drawbacks via performing explicit semantic learning in the common space and semantic-aware sample re-weighting based on optimal transport.

B. Optimal Transport

Optimal transport (OT), also known as Monge-Kantorovich transport, is a mathematical theory that studies the transportation of mass or probability between two or more given distributions [34]. It is based on the idea of finding the most efficient and cost-effective way to transport a certain amount of material or probability from one location to another [35]. The concept of OT has been widely used in computer vision for various applications, such as person re-identification [36], cross-domain alignment [37], object detection [38], and noisy learning [39]. For example, its earliest application was for image registration [40], where it aligned images with different sizes, shapes, and intensities. In the field of domain adaptation, it can be used to facilitate the transfer of knowledge from a source domain to a target domain [41]-[43]. For example, Redko et al. [43] employed OT to estimate class proportions for unlabeled target samples in multi-source adaptation scenario. Moreover, within the domain of contrastive learning, Shi et al. [44] formulated contrastive learning as an inverse optimal transport (IOT) method, employing a bilevel optimization procedure to learn representations and the coupling (probability of matching matrix) between point sets.

Recently, OT-based approaches have gained significant popularity for clustering in the context of deep representation learning. Asano et al. [45] proposed to obtain pseudo labels in a self-supervised way by resolving an OT problem with simultaneous clustering and representation learning. Caron et al. [46] developed an online clustering-based self-supervised method called SwAV, which uses OT to maximize the similarity between the features and the prototypes. Moreover, Stegmüller et al. [47] introduced a Cross-view consistency objective with an Online Clustering mechanism (CrOC) to discover and segment the semantics of the views to learn dense visual representations without labels, where the online optimization objective for computing the clusters and corresponding assignments is based on an OT formulation.

Inspired by the successful applications of OT in aligning different distributions in multiple scenarios, we attempt to utilize OT as an effective tool to map samples from each view to the joint clusters in the common space. By doing so, it is possible to effectively overcome the view gaps and achieve the goal of learning common semantics from the multi-view data.

C. Contrastive Learning

Contrastive learning, which aims to learn useful representations by contrasting positive and negative pairs of samples, has been one of the most popular approaches for self-supervised learning in recent years. With the idea of encouraging positive samples to be closer in the learned feature space while pushing negative samples further apart, CL has shown remarkable success in various domains [29], [30], [32], [48], [49]. For example, Xie et al. [29] promoted the performance of object detection by designing a CL-based framework that not only pursues image-level representation consistency but also ensures that the patch-level features are sufficiently discriminative. Giorgi et al. [48] utilized sentence-level CL to pre-train transformer-based language models, which can effectively learn universal sentence embeddings without any labeled data.

CL has also demonstrated impressive abilities in tackling multi-modal tasks by learning representations that effectively capture the underlying associations between different modalities. For example, in order to learn discriminative multi-modal video representations, Zolfaghari et al. [50] presented a cross-modal contrastive loss that takes both inter-modality and intra-modality similarities into consideration. Yuan et al. [51] proposed a multi-modal CL-based framework to exploit the intra-modality data properties and the cross-modal semantic correlations at the same time. All of these methods achieve the cross-modal feature alignment via the CL mechanism.

Given the successful application of CL in many fields, researchers have also applied it to cluster various types of data. For example, Li et al. [52] combined both instance-level and cluster-level CL for image clustering. Pan et al. [53] proposed a novel graph-level contrastive loss for multi-view graph clustering. In most of the recent state-of-the-art deep MVC methods [14], [16]-[19], cross-view CL has been an integral component, where cross-view contrastive losses are utilized to preserve the representation consistency among multiple views. Although cross-view CL has played a positive role in improving the clustering performance, previous methods tend to focus on the consistent representation learning between each two views, and usually overlook the fact that the semantic significance of samples is varied. In this paper, we make improvements in both of these two aspects to further boost the performance of MVC.

III. METHOD

In this paper, we propose to explicitly learn Common Semantics via Optimal Transport (CSOT) for multi-view clustering. As illustrated in Fig. 2, the proposed CSOT consists of three main components, including intra-view reconstruction, OT-based semantic learning, and semantic-aware contrastive feature learning. Details of CSOT are given below.
A. Preliminaries

1) Problem Formulation: In the regular setting of MVC, the samples from different views are perfectly aligned, which implies the existence of prior knowledge that multiple samples or features from all views belong to the same multi-view sample. In other words, each multi-view sample in an M-view dataset is composed of M features from different views. As such, given an M-view dataset \( \mathcal{X} = \{ \mathbf{X}^1, \cdots, \mathbf{X}^M \} \), where \( \mathbf{X}^m = \{ \mathbf{x}^m_i \in \mathbb{R}^{D_m} \}_{i=1}^N \) represents the feature set of \( N \) samples from the \( m \)-th view, the goal of MVC is to learn comprehensive representations containing both view-specific and view-common information from the multi-view data to cluster semantically consistent samples together.

Specifically, for the \( i \)-th sample in the \( m \)-th view, the encoder \( f_{enc}^m \) encodes the original input \( \mathbf{x}^m_i \) to an \( E \)-dimensional latent embedding denoted as \( \mathbf{z}^m_i \in \mathbb{R}^E \), which is then remapped to the \( D_m \)-dimensional feature space via the decoder \( f_{dec}^m \). The reconstructed sample of \( \mathbf{x}^m_i \) is represented by \( \hat{\mathbf{x}}^m_i \). The parameters of \( f_{enc}^m \) and \( f_{dec}^m \) are optimized by minimizing the distance between \( \mathbf{z}^m_i \) and \( \hat{\mathbf{x}}^m_i \). In this way, \( f_{enc}^m \) gradually learns view-specific representations for the \( m \)-th view. Simultaneously, to learn cross-view common representations, the latent embedding of the \( i \)-th sample in the \( m \)-th view (i.e. \( \mathbf{z}^m_i \)) is separately passed into the feature projection module \( f_{proj} \) and the label prediction module \( f_{pred} \), thus yielding the projected feature embedding \( \mathbf{h}^m_i \) and soft label vector \( \mathbf{y}^m_i \) respectively. After that, several consistency constraints are imposed to pull semantically similar samples in different views together at both feature and label levels. When taking a set of \( \mathbf{x}^m_i \) as input, we further define \( \mathbf{Z}^m, \mathbf{H}^m, \) and \( \mathbf{Y}^m \) as the corresponding matrices for latent embeddings, projected features, and soft labels, respectively.

2) Optimal Transport: In what follows, we briefly describe the well-known optimal transport formulation. Considering two simplex vectors \( \mathbf{\alpha} \in \Sigma_r \) and \( \mathbf{\beta} \in \Sigma_c \), \( (\Sigma_r := \{ \mathbf{x} \in \mathbb{R}^+_r \mid \mathbf{1}^\top \mathbf{x} = 1 \} \) ), the transport between \( \mathbf{\alpha} \) and \( \mathbf{\beta} \) can be mathematically formulated as:

\[
U(\mathbf{\alpha}, \mathbf{\beta}) = \{ \mathbf{Q} \in \mathbb{R}^r_{+} \times \mathbb{R}^c_{+} \mid \mathbf{Q} \mathbf{1}_r = \mathbf{\alpha}, \mathbf{Q}^\top \mathbf{1}_c = \mathbf{\beta} \},
\]

where \( \mathbf{\alpha} \) and \( \mathbf{\beta} \) can be seen as the marginal projections of matrix \( \mathbf{Q} \) onto its rows and columns, respectively. \( \mathbf{1} \) is an all-ones vector of corresponding dimension.

Generally, the OT problem can be solved by minimizing the Wasserstein distance between \( \mathbf{\alpha} \) and \( \mathbf{\beta} \) as defined in [54]:

\[
\text{OT}(\mathbf{S}, \mathbf{\alpha}, \mathbf{\beta}) = \argmax_{\mathbf{Q} \in U(\mathbf{\alpha}, \mathbf{\beta})} \text{Tr}(\mathbf{Q}^\top \mathbf{S}),
\]

where \( \mathbf{S} \) is a similarity matrix.

The discrete optimal transport formulation, in essence, is a convex optimization problem. More precisely, it is a linear programming problem. Unfortunately, this linear programming problem has a cubic computing complexity. To avoid solving this linear programming problem, the entropic regularization term \( H(\mathbf{Q}) \) is introduced into Eq. (2). Then, the coupling matrix (or joint probability) \( \mathbf{Q}^* \) mapping \( \mathbf{\alpha} \) to \( \mathbf{\beta} \) can be quantified by optimizing the following objective function:

\[
\text{OT}^\varepsilon(\mathbf{S}, \mathbf{\alpha}, \mathbf{\beta}) = \argmax_{\mathbf{Q} \in U(\mathbf{\alpha}, \mathbf{\beta})} \text{Tr}(\mathbf{Q}^\top \mathbf{S}) + \varepsilon H(\mathbf{Q}),
\]

where \( \varepsilon > 0 \) is the regularization coefficient and \( H(\mathbf{Q}) = -\sum_{ij} Q_{ij} \log Q_{ij} \) is the entropic regularization term. Such an entropic regularization term facilitates the application of scalable approximation algorithms, such as Sinkhorn’s algorithm [54], which can significantly reduce the computational complexity and thus enables efficient computation of transportation plans. The optimal \( \mathbf{Q}^* \) for Eq.(3) has been proven to be unique, taking the form \( \mathbf{Q}^* = \text{Diag}(\mathbf{u}) \exp(\mathbf{S}/\varepsilon) \text{Diag}(\mathbf{v}) \), where the values of \( \mathbf{u} \) and \( \mathbf{v} \) can be obtained by Sinkhorn’s algorithm [54].
B. OT-based Semantic Learning

In this section, we propose to utilize OT-based semantic learning to ensure that representations of semantically related instances are close to each other in the learned common space. Specifically, we start with building the common space via aggregating the features of different views:

$$h_i^c = \frac{1}{M} \sum_{m=1}^{M} h_i^n,$$

(4)

where $h_i^n$ is the projected feature of the $i$-th sample in the $m$-th view and $h_i^c$ denotes the feature of the $i$-th sample across all the $M$ views. In other words, we use the average of feature embeddings from all the views as a multi-view representation and thus create sample-centered representation in the common space. Then, $k$-means clustering is performed over the fused embeddings $\{h_i^c\}_{i=1}^{N}$, yielding $K$ joint clusters $\{c_j\}_{j=1}^{K}$ among the $N$ samples (the matrix version is denoted as C) that indicate different semantic patterns in the common space. $K$ is the number of categories in the dataset. Compared to the intra-view clusters generated by performing clustering algorithm on $\{h_i^n\}_{i=1}^{N}$, the joint clusters in the common space involve more view-common information and thus can represent the common semantics discovered from multiple views. To establish the semantic associations between a batch of samples and the joint semantics discovered from multiple views. To establish the semantic associations between a batch of samples and the joint clusters, we use OT to seek an efficient solution for transporting each sample to the joint clusters, as a result of which the optimal transport plan can be easily obtained as in Eq. (5):

$$Q^{m*} = OTc(S^m, \frac{1}{B}1_B, \frac{1}{K}1_K),$$

(5)

where $S^m = H^mC^T$ indicates the cosine similarity between the samples in the $m$-th view and the joint clusters. $B$ denotes the batch size for each training iteration and $K$ is the number of joint clusters. By estimating the Wasserstein distance between the two aforementioned distributions, we are able to find the best match between the given sample $x_i^n$ and a joint cluster $c_j$ in the common space. Note that $Q^{m*} \in \mathbb{R}^{B \times K}$ satisfies the constraint strictly, which means that the sum of each row equals $1/B$ and the sum of each column equals $1/K$.

Furthermore, by performing soft max operation on $Q^{m*}$ along the column dimension, we obtain the semantic assignment matrix $\tilde{Q}^{m*}$, which can be rewritten as:

$$\tilde{Q}^{m*} = \left( \begin{array}{cccc} \tilde{Q}^{m*}_{11K} & \tilde{Q}^{m*}_{12K} & \cdots & \tilde{Q}^{m*}_{1K} \\ \tilde{Q}^{m*}_{21K} & \tilde{Q}^{m*}_{22K} & \cdots & \tilde{Q}^{m*}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{Q}^{m*}_{B1} & \tilde{Q}^{m*}_{B2} & \cdots & \tilde{Q}^{m*}_{BK} \end{array} \right),$$

(6)

where the sum of each row equals 1. As such, each element $\tilde{Q}^{m*}_{ij}$ in $\tilde{Q}^{m*}$ can be interpreted as a straightforward probability of assignment, indicating the likelihood that the $i$-th sample is associated with the $j$-th cluster in the common space. A higher $\tilde{Q}^{m*}_{ij}$ indicates that the sample $x_i^n$ is closer to the joint cluster $c_j$ compared to other clusters. Each row vector $q_i^m = [\tilde{Q}^{m*}_{i1}, \tilde{Q}^{m*}_{i2}, \ldots, \tilde{Q}^{m*}_{iK}]$ in matrix $Q^m$ represents the soft semantic assignment for the $i$-th sample.

Note that the joint clusters in our method encourage a more global-aware semantic alignment, which discovers intrinsic associations among multi-view data based on the assignment matrix derived from the OT. As such, the semantic gap among multiple views can be gradually minimized as each joint cluster exhibits the underlying characteristics of a specific semantic class in the common space. With the soft semantic assignment in the common space, we establish a common-specific semantic constraint to encourage the label predictor to capture the semantic distribution of multi-view data. Specifically, for the $i$-th sample, we aim to maximize the consistency between the assignment vector $q_i^m$, i.e. the $i$-th row of the semantic assignment matrix $Q^{m*}$ described in Eq. (6), and the predicted label vector $y_i^m$ for the $m$-th view by minimizing the KL distance:

$$L_{kl} = \sum_{m=1}^{M} \sum_{i=1}^{B} KL(q_i^m, y_i^m),$$

(7)

where $KL(\cdot, \cdot)$ denotes the Kullback–Leibler (KL) divergence that measures the statistical similarity between two distributions.

To explicitly encode the learning of discriminative features with common semantics for multiple views, we also employ the Wasserstein distance to align the samples across different views with joint clusters. In detail, to find a match between an instance $x_i^n$ and the joint cluster $c_j$ in the common space, we employ Wasserstein minimization with an entropic constraint on the similarity distribution between them, resulting in the following OT loss:

$$L_{OT} = \sum_{i=1}^{B} \sum_{j=1}^{K} (1 - S_{ij})Q_{ij}^{m*} + \sum_{i=1}^{B} \sum_{j=1}^{K} Q_{ij}^{m*} \ln Q_{ij}^{m*},$$

(8)

where $S_{ij}$ denotes the similarity between the $i$-th sample and the $j$-th cluster in the common space. It employs self-supervision to implicitly encode the semantic structure of data into the common space, effectively enforcing the alignment of samples across different views during model training.

Moreover, to further ensure the semantic consistency among multiple views, we perform cross-view semantic learning with the idea that the predicted labels for two different views, i.e. $Y^m$ and $Y^n$, should be consistent. Therefore, we contrast the two distributions of samples from two views by minimizing the contrastive loss shown in Eq. (9):

$$l_{na}^{mn} = -\frac{1}{K} \sum_{j=1}^{K} \frac{\log \sum_{k=1}^{K} e^{d(Y_{ij}^m, Y_{jk}^n)/\tau_a}}{\sum_{k=1}^{K} \sum_{m,n} e^{d(Y_{ij}^m, Y_{jk}^n)/\tau_a} - e^{1/\tau_a}},$$

(9)

where $d(\cdot, \cdot)$ denotes the cosine similarity between two vectors, and $\tau_a$ is a hyper-parameter for semantic contrastive learning. Note that the above contrastive loss takes into account both cross-view and intra-view negative samples of the anchor sample in the $m$-th view. Then the loss for cross-view semantic learning can be computed as:

$$L_{cst} = \frac{1}{2} \sum_{m=1}^{M} \sum_{n \neq m} l_{mn}^{mn} + \sum_{m=1}^{M} \sum_{j=1}^{K} \phi_{jm} \log \phi_{jm},$$

(10)
where \( \phi^m_i = \frac{1}{B} \sum_{b=1}^{B} \mathbf{Y}^m_{i,b} \). One can see that in addition to the contrastive loss described in Eq. (9), we include an extra entropy term to spread the predictions uniformly across the \( K \) clusters \([55]\).

In total, the objective function of semantic learning in CSOT is defined as:

\[
\mathcal{L}_{sem} = \mathcal{L}_{kl} + \mathcal{L}_{OT} + \mathcal{L}_{csl}. \tag{11}
\]

### C. Semantic-aware Cross-view Contrastive Learning

With the knowledge that up-weighting hard samples contributes significantly to contrastive learning, we design a semantic-aware re-weighting mechanism to improve the model’s ability to learn cross-view common semantics for MVC, as shown in Fig. 3. Based on the OT described in Sec. III-B, the soft semantic assignments are available for estimating the significance of samples from a semantic perspective. Given a sample \( x^m_i \) from the \( m \)-th view, if it belongs to the \( k \)-th joint cluster in the common space, then the similarity between \( h^m_i \) and \( c_k \) is supposed to reflect the proportion of view-common and view-specific information in the sample. Once a sample is dominated by view-common information, it is likely to be redundant with the samples from the other views, and thus easier to be explored for learning common semantics. In contrast, the sample that contains more view-specific information is relatively hard in the case of cross-view contrastive learning, making it a semantically significant sample for MVC.

Therefore, we propose a semantic-aware re-weighting mechanism to up-weight the semantically significant (or hard) samples while down-weighting the others. The weight for sample \( x^m_i \) is defined as:

\[
w^m_i = 1 - \max(\mathbf{q}^m_i), \tag{12}
\]

where \( \mathbf{q}^m_i = [\tilde{Q}^m_{i1}, \tilde{Q}^m_{i2}, \ldots, \tilde{Q}^m_{iK}] \) represents the semantic assignment vector for the sample \( x^m_i \) and \( \max(\cdot) \) means the maximum element of a vector, such that the weight \( w^m_i \) and the maximum value of \( \mathbf{q}^m_i \) follow the inverse correlation. In other words, the higher the similarity between a sample and the corresponding joint cluster, the higher the proportion of view-common information this sample has, the easier to learn common semantics from it, and thus the lower the weight it has. In this paper, the semantic-aware cross-view contrastive loss for the \( m \)-th view and the \( n \)-th view is computed by:

\[
l^{mn}_{b} = -\frac{1}{B} \sum_{i=1}^{B} w^m_i \log \frac{1}{\sum_{j=1}^{B} \sum_{e=m,n} e^{d(h^m_i, h^n_j)/\tau_b}} e^{d(h^m_i, h^n_j)/\tau_b} - e^{1/\tau_b}, \tag{13}
\]

where \( d(\cdot, \cdot) \) represents the cosine similarity between two features, and \( \tau_b \) is the temperature factor for contrastive feature learning. Here the contrastive loss for cross-view representation learning also involves both cross-view and intra-view negative samples as in Eq. (9). Then the semantic-aware contrastive loss across all the views can be obtained via:

\[
\mathcal{L}_{csl} = \frac{1}{2} \sum_{m=1}^{M} \sum_{n \neq m} l^{mn}_{b}. \tag{14}
\]

Recently, Xu et al. [59] introduced a self-weighting mechanism that re-weights multi-view contrastive losses according to semantic clues. Specifically, this mechanism induces the model to focus on reliable pairs by down-weighting the unreliable pairs that have larger discrepancy. In contrast, the semantic-aware re-weighting mechanism proposed in this section is designed to up-weight the semantically significant samples that have lower similarity with the corresponding joint cluster, which facilitates the mining of challenging samples during multi-view contrastive learning.

As a result, by applying the semantic-aware re-weighting mechanism to the cross-view contrastive loss, our CSOT pays more attention to the semantically unique samples while reducing the influence of semantically redundant ones. This enables the model to focus on the most challenging samples during training, and thus can help the model learn cross-view common semantics in a more effective manner. Note that we do not apply the proposed semantic-aware re-weighting mechanism in the cross-view semantic learning, mainly because the contrastive loss described in Eq. (9) is calculated with the predicted label vectors, which are more semantically meaningful than the feature vectors used in Eq. (13). Therefore, there is no need to involve additional semantic constraints for computation efficiency.

### D. Joint Training of CSOT

In the previous sections, we have described in detail how we conduct semantic learning via OT and how the semantic-aware contrastive learning works in learning common semantics. The overall training objective of our CSOT framework includes three loss items:

\[
\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{sem} + \lambda_2 \mathcal{L}_{csl}, \tag{15}
\]

where \( \lambda_1 \) and \( \lambda_2 \) are the weighting factors for semantic learning and weighted contrastive feature learning, respectively. In this paper, we empirically set \( \lambda_1 \) and \( \lambda_2 \) to 0.3 and 1.0, respectively. See Sec. IV-D for more details.

In Eq. (15), \( \mathcal{L}_{rec} \) denotes the view-specific reconstruction loss, which allows effective representation learning for each view via pulling the reconstructed sample close to the original input. The implementation details of \( \mathcal{L}_{rec} \) can be found in recent deep MVC works [17], [19], [20].

The second term of Eq. (15), i.e. \( \mathcal{L}_{sem} \), is the loss of semantic learning described in Sec. III-B. As discussed above,
Algorithm 1 The training process of CSOT

Input: an M-view $\mathcal{X} = \{X^m\}_{m=1}^M$ with $N$ samples in each view, the view-specific encoders $\{f_{enc}^m\}_{m=1}^M$, the view-specific decoders $\{f_{dec}^m\}_{m=1}^M$, the feature projector $f_{proj}$, the label predictor $f_{pred}$, the number of epochs for pre-training $T_{pre}$, the number of epochs for joint training $T_{joint}$, and the number of iterations per epoch $I$. 

1: for $t \in \{1, \ldots, T_{pre}\}$ do
2:   for $i$ in $I$ do
3:     Extract the view-specific features $Z^m = f_{enc}^m(X^m)$;
4:     Reconstruct the original inputs $X^m = f_{dec}^m(Z^m)$;
5:     Compute the reconstruction loss $\mathcal{L}_{rec}$;
6:     Update $\{f_{enc}^m\}_{m=1}^M$ and $\{f_{dec}^m\}_{m=1}^M$ to minimize $\mathcal{L}_{rec}$.
7:   end for
8: end for
9: for $t \in \{1, \ldots, T_{joint}\}$ do
10:   for $i$ in $I$ do
11:     Extract the view-specific features $Z^m = f_{enc}^m(X^m)$;
12:     Reconstruct the original inputs $X^m = f_{dec}^m(Z^m)$;
13:     Compute the reconstruction loss $\mathcal{L}_{rec}$;
14:     Obtain the projected features $H^m = f_{proj}^m(Z^m)$ and the predicted label vectors $Y^m = f_{pred}^m(Z^m)$;
15:     Build the common space via Eq. (4) and obtain the joint clusters $\{c_j\}_{j=1}^K$ via k-means clustering;
16:     Obtain the optimal transport plan $Q^{ms}$ mapping the projected features $H^m$ to the joint clusters $\{c_j\}_{j=1}^K$ via Eq. (5);
17:     Compute the loss $\mathcal{L}_{sem}$ for semantic learning via Eqs. (7)-(11);
18:     Compute the loss $\mathcal{L}_{wcl}$ for semantic-aware contrastive representation learning via Eqs. (12)-(14);
19:     Compute the total loss $\mathcal{L}$ via Eq. (15);
20:     Update $\{f_{enc}^m\}_{m=1}^M$, $\{f_{dec}^m\}_{m=1}^M$, $f_{proj}$, and $f_{pred}$ simultaneously to minimize $\mathcal{L}$.
21: end for
22: end for

Output: The view-specific encoders $\{f_{enc}^m\}_{m=1}^M$, the feature projector $f_{proj}$, and the label predictor $f_{pred}$.

### Table I: Statistics of different multi-view datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#samples</th>
<th>#classes</th>
<th>#views</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST-USPS</td>
<td>5000</td>
<td>10</td>
<td>2</td>
<td>784, 784</td>
</tr>
<tr>
<td>BDGP</td>
<td>2500</td>
<td>5</td>
<td>2</td>
<td>1750, 79</td>
</tr>
<tr>
<td>CCV</td>
<td>6773</td>
<td>20</td>
<td>3</td>
<td>5000, 5000, 4000</td>
</tr>
<tr>
<td>Synthetic3D</td>
<td>600</td>
<td>3</td>
<td>3</td>
<td>3, 3, 3</td>
</tr>
<tr>
<td>Caltech</td>
<td>1400</td>
<td>7</td>
<td>5</td>
<td>40, 254, 928, 512, 1984</td>
</tr>
<tr>
<td>Fashion</td>
<td>10000</td>
<td>10</td>
<td>3</td>
<td>784, 784, 784</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>50000</td>
<td>10</td>
<td>3</td>
<td>512, 2048, 1024</td>
</tr>
<tr>
<td>NUS-WIDE</td>
<td>30000</td>
<td>31</td>
<td>5</td>
<td>64, 225, 144, 73, 128</td>
</tr>
<tr>
<td>ALOI</td>
<td>10800</td>
<td>100</td>
<td>4</td>
<td>77, 13, 64, 125</td>
</tr>
<tr>
<td>DHA</td>
<td>483</td>
<td>22</td>
<td>2</td>
<td>110, 6144</td>
</tr>
<tr>
<td>UWA</td>
<td>507</td>
<td>30</td>
<td>2</td>
<td>110, 6144</td>
</tr>
</tbody>
</table>

it promotes the training in two ways. On the one hand, the OT-based semantic assignment $Q^{ms}$ provides pseudo labels for the samples from a global perspective while the label predictor outputs cluster assignment probabilities for each single sample. The consistency between such two kinds of labels can be maintained through minimizing $\mathcal{L}_{sem}$. On the other hand, $\mathcal{L}_{sem}$ aims to preserve the label consistency between each two views by including a cross-view semantic contrastive loss term in it.

$\mathcal{L}_{wcl}$ in Eq. (15) represents the semantic-aware contrastive loss for representation learning introduced in Sec. III-C, where a re-weighting mechanism is applied to the traditional loss function. This term involves the significance estimation of each view from a global semantic perspective, and fully exploits the role of hard sample mining in learning cross-view common semantics.

In conclusion, the OT-based semantic learning can effectively explore the semantic associations among multi-view data, pushing each sample closer to the corresponding joint cluster in the common space. Based on the learned semantics, the semantic significance of each sample can be estimated for re-weighting the contrastive losses to achieve more effective cross-view feature learning. The collaboration of these two modules ultimately boosts the MVC performance.

The pseudo code for the training process of CSOT can be found in Algorithm 1.

E. Complexity Analysis

To analyze the computational complexity of CSOT, we denote the number of views and the number of clusters for each dataset as $M$ and $K$, respectively. During each training iteration, the time complexities for computing $\mathcal{L}_{rec}$ and $\mathcal{L}_{wcl}$ are $O(MB)$ and $O(MB^2)$, respectively, where $B$ stands for the mini-batch size. Since $\mathcal{L}_{sem}$ involves solving the optimal transport problem defined in Eq. (5), its computational complexity is $O(M(BK(I+1)+B+K^2))$, where $I$ denotes the number of iterations for Sinkhorn’s algorithm. Considering that $B$ and $I$ are usually much larger than 1, the final computational complexity of CSOT is $O(M(B^2+ BK+K^2))$. In the testing stage, the computational complexity is $O(MBK)$, which is linear in the number of views, the number of samples, as well as the number of clusters simultaneously.

IV. EXPERIMENTS

In this section, we report extensive experimental results on various multi-view datasets to verify the effectiveness of the proposed CSOT framework for MVC. More experimental results are additionally presented in the supplementary material.

A. Experimental Settings

1) Datasets: Eleven widely-used multi-view datasets that vary in both data volume and number of views are utilized in the experiments as listed in Table I. The MNIST-USPS dataset, created based on two popular handwritten digit datasets, i.e. MNIST and USPS, consists of 5,000 digital images.

Note that when we refer to the number of samples in a dataset, it usually means the number of multi-view samples, which equals the number of samples in each view.
image samples each in the two views. The BDGP \cite{8} dataset includes 2,500 samples of drosophila embryos, each of which is composed of a visual view and a textual view. The ACC \cite{9} dataset is a collection of 6,773 video samples belonging to 20 different categories. Each video sample is represented by three kinds of hand-crafted Bag-of-Words representations, including Space-Time Interest Points (STIP), Scale-Invariant Feature Transform (SIFT), and Mel-Frequency Cepstral Coefficients (MFCC). Synthetic3D \cite{10} is a small dataset with only 600 samples from three classes and each sample is represented by features in three different views. Caltech-101 \cite{11} is an image dataset of multi-view objects from 101 categories. Following \cite{17,19}, we use four sub-sets of Caltech-101 to evaluate the performance of deep MVC methods on datasets with various numbers of views. The Fashion \cite{12} dataset contains 10,000 samples of products from three views, where each view corresponds to a fashion style of the product. CIFAR10 \cite{13} is a three-view dataset that has 50,000 samples belonging to 10 categories. The three views are represented by 2014/512/1024-dimensional vectors, respectively. NUS-WIDE \cite{61} is composed of 30,000 images belonging to 31 classes of objects, and each image is represented by multiple
low-level features. ALOI \cite{62} consists of 10,800 images of small objects, along with 4 types of feature vectors for each image. DHA \cite{63} and UWA \cite{64} are two RGB-D datasets collected in the real world for human action recognition, in which each action is represented by both RGB and depth features.

2) **Implementation Details:** For all the datasets, the original representations from multiple views are first mapped to a fixed dimensional feature space via view-specific encoders. Then, the objective function defined in Eq. (15) is computed for dimensional feature space. The representations from multiple views are first mapped to a fixed dimensional feature space via view-specific encoders. Then, the objective function defined in Eq. (15) is computed for the reconstruction loss. We pre-train the view-specific encoders with the reconstruction loss for 200 epochs before optimizing the CSOT framework with the overall loss. The Adam optimizer with a learning rate of 0.0003 is adopted for optimization. The hyper-parameters $\lambda_1$ and $\lambda_2$ are empirically set to 0.3 and 1.0 respectively. In this paper, three commonly used metrics are utilized to estimate the clustering performance of CSOT, including Accuracy (ACC), Normalized Mutual Information (NMI), and Purity (PUR). In the testing stage, only the view-specific encoders and the label predictor are preserved to extract $K$-dimensional label vectors for clustering. We average these vectors across multiple views for a fused vector, with which the argmax operation is performed to produce the final results. Other settings related to the calculation of the metric scores follow the release codes of the advanced MVC methods \cite{17,19}. Higher scores indicate better performance. All experiments are implemented through PyTorch on a Ubuntu 18.04 system with an NVIDIA 1080Ti GPU.

**B. Comparison**

We compare the proposed CSOT with the following 12 state-of-the-art methods including 3 traditional MVC methods and 9 deep learning-based ones.

- **LMVSC** \cite{2} solves large-scale MVC in linear time by integrating smaller view-specific graphs built with only a small number of instances.
- **CDIMC-net** \cite{57} boosts MVC by using graph embedding and a self-paced learning strategy, leading to a focus on the local structure and confidence of samples.
- **EAMC** \cite{58} is an end-to-end adversarial-attention network proposed for multi-modal clustering.
- **CONAN** \cite{59} handles deep MVC with a contrastive fusion framework that maximizes the similarity between the fused common representation and each view-specific one.
- **SiMVC** \cite{16} simply combines cross-view representations by a weighted sum, in which the weights of multiple views are learnable parameters.
- **CoMVC** \cite{16} includes a selective contrastive alignment module to learn view-common representation and preserve the view-prioritization simultaneously.
- **PLCMF** \cite{5} explicitly applies a pseudo-label constraint to collective matrix factorization for preserving both intra-view and inter-view similarities.
- **MFLVC** \cite{17} learns multi-level features to conduct view-specific reconstruction, feature-level contrastive learning, and label-level discrimination separately.

- **FastMICE** \cite{6} defines random view group to capture view-wise relationships, which not only benefits the clustering robustness but also improves the efficiency.
- **SDMVC** \cite{60} conducts discriminative feature learning on global scale in a self-supervised manner, where pseudo labels help learn consistent cluster assignment.
- **AECoDDC** \cite{18} is a deep MVC method composed of view-specific autoencoders, cross-view contrastive losses, and deep divergence-based clustering.
- **GCFagg** \cite{19} explores the global relationship among multi-view representations and leverages feature aggregation to enhance the learned common representation.

The comparative results on multiple datasets are shown in Table II IV. Overall, CSOT achieves the state-of-the-art performance. Some specific observations are listed below:

1) CSOT outperforms the traditional methods (LMVSC \cite{2}, PLCMF \cite{5}, and FastMICE \cite{6}) by a large margin across all datasets. For example, compared to FastMICE \cite{6}, CSOT improves the NMI scores on the small-scale BDGP dataset and the large-scale Fashion dataset by 36.7% and 14.9%, respectively. The performance of traditional methods is generally worse than that of deep learning-based ones, because they utilize raw multi-view features in the datasets, which are usually noisy and not generalisable.

2) Among the deep MVC methods, CSOT shows the best robustness to multiple datasets. In particular, it ranks first for all three metrics on 7 of the 9 datasets including MNIST-USPS, BDGP, Synthetic3D, Fashion, CIFAR10, NUS-WIDE, and ALOI, while achieving at least the second best on the other datasets. The slightly lower scores of CSOT on CCV may be due to the fact that the original features are less discriminative, which results in joint clusters failing to accurately reflect the semantic patterns in the multi-view data.

3) CSOT effectively utilizes the information from the increased views to boost the clustering performance. Table III shows the comparative results on the Caltech dataset with increasing numbers of views. We can see that the clustering performance of CSOT is gradually improved by 24.3% on the ACC score when the number of views increases from 2 to 5. On the contrary, some competing methods (EMAC \cite{58}, CONAN \cite{59}, PLCMF \cite{5}, AECoDDC \cite{18}, and GCFagg \cite{19}) fail to benefit from the increased views. In other words, their clustering results may deteriorate when the number of views increases.

To further demonstrate the superiority of CSOT, we show in Fig. 4 the comparison between the confusion matrices generated by CSOT and those generated by the AECoDDC \cite{18}. It can be observed that the elements along the diagonal of the confusion matrices in the top row usually have higher values than those in the bottom row, indicating that the predicted labels of CSOT are more consistent with the ground-truth labels. Moreover, we compare CSOT with the state-of-the-art contrastive MVC methods on two multi-view datasets, DHA and UWA, collected in the real world for human action recognition. The experimental results listed in Table V show that CSOT achieves the highest scores in all three metrics, which demonstrate that CSOT provides good discrimination among real-world human actions.
We have also performed significance test to demonstrate the superiority of CSOT. Almost Stochastic Dominance (ASO) test \[^{[65]}\], \[^{[66]}\] is a statistical test specifically designed for deep neural models, which has been used to compare different algorithms across a variety of tasks \[^{[67]}\]–\[^{[69]}\]. In ASO test, \(\epsilon_{\text{min}} = 0\) represents that the estimated method achieves a stochastic dominance over the comparative methods and \(\epsilon_{\text{min}} < 0.5\) means an almost stochastic dominance. In this paper, we conduct the ASO test between CSOT and recent contrastive MVC methods according to the clustering results produced by 10 repeated experiments with random seeds. Table VI lists the results of significance test with a confidence level of \(\alpha = 0.05\), indicating that CSOT is significantly superior to the competing methods.

CSOT works well for two reasons. On the one hand, the OT-based semantic learning helps the model achieve better semantic alignment among multiple views in the common space. Thus the trained model can effectively extract the complementary information from multi-view data while discarding the redundant one. On the other hand, the semantic-aware re-weighting strategy fully utilizes the learned common semantics for the evaluation of the hardness of the samples. Thus the cross-view contrastive feature learning can be performed in a more effective way. As a result, the CSOT method well balances the complementarity and specificity of the multiple views, which leads to more accurate feature presentation and clustering results.

C. Ablation Studies

1) Ablation Studies on Main Components of CSOT: To evaluate the effectiveness of the OT-based semantic learning and the semantic-aware contrastive feature learning in CSOT, we perform ablation studies on 5 datasets. Based on the total loss defined in Eq. [15], we evaluate the MVC performance with different settings of the loss items. We start with removing the loss for OT-based semantic learning \(L_{\text{sem}}\). One can clearly see from Table VII that the clustering performance shows a significant degradation on all the three datasets, which indicates that utilizing OT to conduct multi-view semantic learning is highly effective. Then, we discard the semantic-aware contrastive feature learning module \(L_{\text{cl}}\) while keeping the OT-based semantic learning term \(L_{\text{sem}}\) along with the reconstruction loss \(L_{\text{rec}}\). As shown in the second and the fifth rows of each sub-table in Table VII, it also results in performance drops on all the 5 datasets. Especially for the Caltech-5V dataset, the performance drop is highly significant, with the ACC score dropping by 16.9%. This suggests that pulling cross-view representations via the semantic-aware contrastive loss is a suitable choice in MVC. To further demonstrate the priority of the proposed semantic-based re-weighting strategy, we replace the semantic-aware contrastive loss \(L_{\text{cl}}\) with the widely used contrastive loss \(L_{\text{cl}}\) without any re-weighting strategy. It can be seen that the contrastive loss used in this work allows the MVC model to achieve performance gains on all the 5 datasets. Moreover, the advantage of our semantic-aware contrastive loss becomes more and more pronounced when the number of views increases, which demonstrates that the re-weighting strategy helps the model learn discriminative features for multi-view data more effectively.

### Table V

Comparison of state-of-the-art contrastive MVC methods on DHA and UWA.

<table>
<thead>
<tr>
<th>Evaluation metrics</th>
<th>DHA</th>
<th>UWA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>0.744</td>
<td>0.793</td>
</tr>
<tr>
<td>NMI</td>
<td>0.208</td>
<td>0.088</td>
</tr>
<tr>
<td>PUR</td>
<td>0.723</td>
<td>0.723</td>
</tr>
</tbody>
</table>

### Table VI

Results of significance test on MNIST-USPS and BDGP. \(\epsilon_{\text{min}}^{\text{ACC}}\) and \(\epsilon_{\text{min}}^{\text{NMI}}\) denote the ASO score for ACC and NMI respectively.

<table>
<thead>
<tr>
<th>Evaluation metrics</th>
<th>MNIST-USPS</th>
<th>BDGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>(\epsilon_{\text{min}}^{\text{ACC}})</td>
<td>(\epsilon_{\text{min}}^{\text{ACC}})</td>
</tr>
<tr>
<td>NMI</td>
<td>(\epsilon_{\text{min}}^{\text{NMI}})</td>
<td>(\epsilon_{\text{min}}^{\text{NMI}})</td>
</tr>
<tr>
<td>CSOT (\rightarrow) MFLVC</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>CSOT (\rightarrow) AECoDDC</td>
<td>5.55e-5</td>
<td>7.37e-4</td>
</tr>
<tr>
<td>CSOT (\rightarrow) GCFAgg</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table VII

Ablation studies of the main components of CSOT on various datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Method</th>
<th>ACC</th>
<th>NMI</th>
<th>PUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST-USPS</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td>CCV</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.295</td>
<td>0.295</td>
<td>0.321</td>
</tr>
<tr>
<td>Synthtic3D</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.300</td>
<td>0.300</td>
<td>0.324</td>
</tr>
<tr>
<td>Caltech-5V</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.975</td>
<td>0.975</td>
<td>0.975</td>
</tr>
<tr>
<td>DHA</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.697</td>
<td>0.697</td>
<td>0.718</td>
</tr>
</tbody>
</table>

### Table VIII

Ablation studies of optimal transport on various datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Method</th>
<th>ACC</th>
<th>NMI</th>
<th>PUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST-USPS</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td>CCV</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.295</td>
<td>0.295</td>
<td>0.321</td>
</tr>
<tr>
<td>Synthtic3D</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.300</td>
<td>0.300</td>
<td>0.324</td>
</tr>
<tr>
<td>Caltech-5V</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.975</td>
<td>0.975</td>
<td>0.975</td>
</tr>
<tr>
<td>DHA</td>
<td>w/o OT-related losses w/ cosine similarity w/ OT</td>
<td>0.697</td>
<td>0.697</td>
<td>0.718</td>
</tr>
</tbody>
</table>
2) Ablation Studies on Optimal Transport in CSOT: In CSOT, OT facilitates mapping samples from each view to the joint clusters in the common space, which is crucial for bridging the semantic gaps among multiple views. To comprehensively analyze the performance of CSOT, we carried out ablation studies of OT and listed the results in Table VIII. We use the cosine similarity matrix as a substitute for the OT plan to accomplish the matching between samples and the joint clusters. This means that only $L_{OT}$ defined in Eq. 7 is removed, while other losses can still be computed using the soft semantic assignment based on the cosine similarity. The experimental results exhibit a decrease in clustering performance on all the 5 datasets, especially on the DHA dataset, where the ACC score decreases by 5.4%. Such results demonstrate the superiority of OT in matching samples from each view to the joint clusters for MVC. Furthermore, we move out all OT-related losses in CSOT for comparison, which unsurprisingly leads to further degradation of the clustering performance. This also verifies the effectiveness of learning common semantics via OT in MVC.

D. Parameter Sensitivity Analysis

CSOT has two weighting parameters to balance the losses and two temperature parameters to control the strength of samples in contrastive learning. To figure out the effect of such hyper-parameters on clustering performance, we implement CSOT under different settings, and present the results in Fig. 5. We employ a grid search strategy in the range of [0.1, 1.0] to explore the best setting of the weighting parameters $\lambda_1$ and $\lambda_2$. From Figs. 5a and 5b, we can see that the NMI score of CSOT remains at a high level over both datasets when $\lambda_1$ is less than 0.5 and $\lambda_2$ is greater than 0.7. In this work, we empirically set $\lambda_1$ and $\lambda_2$ to 0.3 and 1.0, respectively. Then we apply the same strategy to analyze the effect of temperature parameters $\tau_1$ and $\tau_2$. We can see from Figs. 5c and 5d that CSOT is insensitive to either $\tau_1$ or $\tau_2$ in the range of [0.4, 0.6]. We thus set both of them to 0.5 in the experiments.
E. Convergence Analysis

To investigate the convergence of CSOT, we conduct experiments on 4 benchmark datasets and record the variations of the losses and the metric scores during the CSOT training phase. Fig. 6 presents the corresponding variation curves on MNIST-USPS, BDGP, Caltech-5V, and Fashion for the CSOT training phase. We can observe that as the training proceeds, the value of the total loss computed by Eq. 15 gradually decreases and converges to a low level, while the metric scores of MVC progressively increase to a steady state. This trend suggests that the loss functions of CSOT constrain the model to continuously optimize its parameters for the improvement of the clustering performance.

In addition, we show in Fig. 7 the evolution of feature distribution for each view during the whole training process including both the pre-training and the CSOT training phases. The features are visualized via t-distributed stochastic neighbor embedding (t-SNE) [70], which projects the learned high-dimensional deep features to a lower 2D space. Taking the MNIST-USPS and BDGP datasets as examples, we can observe that as the training proceeds, the samples within each cluster are gradually gathered together and the boundaries between different clusters become more pronounced. This indicates that CSOT allows the model to learn distinguishable features from multi-view data, regardless of the cross-view semantic gaps that may separate positive samples in different clusters.

F. Effect of Different Clustering Schemes

We also explore the effect of different clustering schemes. Once the training process shown in Fig. 2 ends, we can obtain the trained feature projector and label predictor for testing. As a result, both the projected features and the predicted labels are available in the testing stage, providing two different schemes to acquire the final clustering results. One is to perform the k-means clustering on the projected features, and the other is to apply argmax operation on the predicted label vectors. We can see from Fig. 8 that utilizing the predicted label vectors to obtain the final clustering results is a better choice. This is mainly because CSOT provides explicit semantic supervision from the common space for the label prediction, which forces...
the predicted labels to reflect the common semantic patterns in the multi-view data.

Moreover, we show in Fig. 8b the clustering performance produced by the predicted labels for each view and the fusion of them. It is evident that the accuracy of multi-view clustering surpasses that of single-view clustering, indicating that CSOT can effectively handle multi-view data for higher clustering accuracy.

G. Cross-view Retrieval

To further highlight the advantages of CSOT in learning common semantics from multi-view data, we evaluate the performance of the trained model in the cross-view retrieval task. In cross-view retrieval, the feature of a sample in a query view is used to retrieve the paired samples in other views, allowing us to calculate the Mean Average Precision (MAP).

Due to the presence of view gaps, the common semantics play a vital role in the process of cross-view feature matching. Table IX reports the MAP scores achieved by CSOT and two recent contrastive MVC methods, MFLVC [17] and GCFAgg [19]. The experimental results demonstrate that CSOT outperforms both competing methods by a large margin in the cross-view retrieval task, no matter which view is selected as a query. This verifies that CSOT effectively overcomes the interference of view gaps and learns multi-view common semantics.

V. Conclusion

This paper proposes a novel framework, namely CSOT, which learns common semantics via OT to boost the performance of contrastive multi-view clustering. Compared with previous work, the most distinctive feature of our approach is that it provides a more global perspective for exploring cross-view common semantics by mapping multi-view data into a unified feature space. Specifically, we employ OT to conduct semantic learning in the common space, which enables the model to effectively capture the common semantics regardless of view gaps. Based on the learned common semantics, we design a semantic-aware re-weighting mechanism to increase the influence of semantically unique samples during cross-view contrastive feature learning. Experimental results demonstrate that our method effectively learns the underlying semantic patterns of multi-view data, achieving state-of-the-art clustering performance. However, CSOT has a drawback in that it assumes that the multi-view samples in the dataset are complete and fully aligned, ignoring the fact that real-world multi-view datasets are likely to be incomplete or partially aligned. Therefore, how to improve CSOT to make it suitable for incomplete or partially aligned multi-view data will be investigated in our future work.

REFERENCES


