Toward Generalized Multistage Clustering: Multiview Self-Distillation

Jiatai Wang, *Student Member, IEEE*, Zhiwei Xu, *Member, IEEE*,
Xin Wang, *Senior Member, IEEE*, and Tao Li, *Member, IEEE*

***Abstract*—Existing multistage clustering methods independently learn the salient features from multiple views and then perform the clustering task. Particularly, multiview clustering (MVC) has attracted a lot of attention in multiview or multimodal scenarios. MVC aims at exploring common semantics and pseudo-labels from multiple views and clustering in a self-supervised manner. However, limited by noisy data and inadequate feature learning, such a clustering paradigm generates overconfident pseudo-labels that misguide the model to produce inaccurate predictions. Therefore, it is desirable to have a method that can correct this pseudo-label mistraction in multistage clustering to avoid bias accumulation. To alleviate the effect of overconfident pseudo-labels and improve the generalization ability of the model, this article proposes a novel multistage deep MVC framework where multiview self-distillation (DistilMVC) is introduced to distill dark knowledge of label distribution. Specifically, in the feature subspace at different hierarchies, we explore the common semantics of multiple views through contrastive learning and obtain pseudo-labels by maximizing the mutual information between views. Additionally, a teacher network is responsible for distilling pseudo-labels into dark knowledge, supervising the student network and improving its predictive capabilities to enhance its robustness. Extensive experiments on real-world multiview datasets show that our method has better clustering performance than the state-of-the-art (SOTA) methods.**

***Index Terms*—Hierarchical contrastive learning, multistage clustering, multiview self-distillation, mutual information between views.**

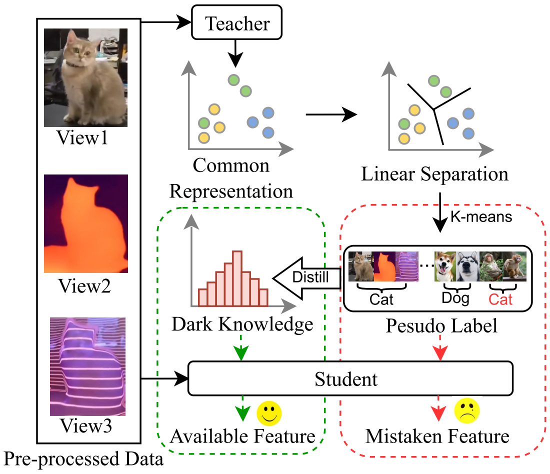


Fig. 1. Overconfident pseudo-labels used in MVC and their distillation. The multiview data instances are learned to achieve a common representation of views. However, pseudo-labels obtained from common representation learning are often overconfident for this multiview scenario. Distillation after labeling, obtains dark knowledge, a new self-supervised signal that contains richer semantic information compared to pseudo-labels, can better guide the multistage clustering and significantly improve the quality of clustering.

I. Introduction

TRADITIONAL clustering methods [24], [29], [33], [40], [41], [49], [51], [60], [76] have been used with specific machine learning techniques in various tasks. Among them, clustering algorithms [9], [42], [43], [67] based on deep learning have emerged due to their powerful generalization capability and scalability. These algorithms jointly learn the parameters of some specific neural networks and assign the features extracted to clusters. Among them, one-stage deep clustering methods [47], [65], [96] work end-to-end for feature learning and are easy to lock in low-level features. On the other hand, the multistage deep clustering method [71], [84] performs multiple rounds of feature extraction under the supervision of the pseudo-labels obtained through self-learning, where the labels are used to guide the training of a prediction model for clustering. The overall process of multistage deep clustering fits exactly into the self-supervised paradigm of model training guided by the intrinsic structure of data, which helps to achieve enhanced feature learning and clustering performance. According to Cover's [14] theorem, complex data are more likely to be linearly separable when they are projected to a high-dimensional representation space, and this theory provides a base for the feasibility of such pseudo-label-based training. The pseudo-labels learned are used as a priori or self-supervised signal to guide the training of clustering model [67], [75], [79], [83], [83], [84]. Recently, multistage clustering methods have become a focus of research [67].

Received 13 November 2022; revised 8 June 2023, 8 February 2024, and 19 June 2024; accepted 1 October 2024. Date of publication 11 November 2024; date of current version 4 June 2025. This work was supported in part by the National Science Foundation of China under Grant 61962045 and Grant 62272248, in part by the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region under Grant NJYT23104, and in part by the Basic Scientific Research Expenses Program of Universities directly under Inner Mongolia Autonomous Region under Grant JY20220273 and Grant JY20240002. *(Corresponding authors: Zhiwei Xu; Tao Li.)*

Jiatai Wang and Tao Li are with the College of Computer Science, Nankai University, Tianjin 300350, China, and also with the Haihe Laboratory of Information Technology Application Innovation (ITAI), Tianjin 300459, China (e-mail: 1120240357@mail.nankai.edu.cn; litao@nankai.edu.cn).

Zhiwei Xu is with the Haihe Laboratory of Information Technology Application Innovation (ITAI), Tianjin 300459, China, and also with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xuzhiwei2001@ict.ac.cn).

Xin Wang is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: x.wang@stonybrook.edu).

Digital Object Identifier 10.1109/TNNLS.2024.3479280

Data in the real world are mostly collected from different (types of) sensors or feature extractors. Multiview clustering (MVC), one of the multistage clustering problems, has been proposed to explore the common semantics among different views and investigate the effectiveness of pseudo-labeling for self-supervision [15], [30], [57], [80]. However, MVC suffers from some drawbacks and constraints when applied to multimodal or multiviews. As the number of views increases, each view introduces its unique form of noise, exacerbating the overall noise in the dataset. If pseudo-labels are directly extracted from noisy data features and used by the predictor, it will lead to a low-entropy state in the data representation and overconfidence [5]. Ultimately, the model is not taken as a well-calibrated predictor [23]. Furthermore, the pseudo-labels for cross-entropy penalize all negative predictions regardless of their logical foundations. Significant features related to negative prediction are neglected. Thus, it is a challenge to avoid the damaging impact of false pseudo-labels during feature learning and correct the inaccurate bootstrapping [65], [84].

To address this challenge, we first comprehensively study multistage deep learning methods in computer vision and find that the use of knowledge distillation (KD) can deplete the negative impact of pseudo-labels and considerably enhance model performance in both supervised and unsupervised settings [2], [8], [34], [37], [64]. In this case, a teacher network iteratively optimizes the student network by replacing pseudo-labels with the KD results, i.e., the implicit feature distribution (dark knowledge [27]), so that the student can comprehensively distinguish similarities and differences among samples and extract significant features (see Fig. 1). Based on these observations, we propose a novel multistage MVC framework based on multiview self-distillation (DistilMVC), which leverages contrastive learning techniques to optimize the backbone network of our design (the teacher network and the student network) and build a bond between the teacher and the student by a self-distillation process. More specifically, we leverage contrastive learning to achieve better representations of unlabeled view data, as well as establish a feature space for the self-distillation process. This ensures semantic consistency across views through multiscale mutual information maximization, even with significant differences between views. In addition, the dark knowledge learned in the self-distillation process can align the feature space of the student network with that updated by the teacher network in the iterative process and alleviates the overconfidence of the conventional MVC schemes associated with pseudo-labels.

To summarize:

1) We explore the use of KD in MVC and propose a multiview self-distillation technology that transforms overconfident pseudo-labels into dark knowledge, reducing the impact of false pseudo-labels on multiview feature learning. As dark knowledge contains essential hierarchical information that is not included in pseudo-labels, using it as a supervision indicator can generalize the multiview representation learning.
2) We propose a contrastive method to learn multiview semantics in feature spaces from different hierarchies. In a low-dimensional latent space, we directly maximize the mutual information with invariant information clustering (IIC), and in a high-dimensional subspace, we raise the lower bound of mutual information according to the fixed point related to the scale of negative samples. This can accordingly improve the self-supervised learning multiview representation performance for MVC.
3) Based on the proposed multiview self-distillation technology, we introduce a new multistage framework, which uses dark knowledge instead of pseudo-labels as a supervision indicator and thus generalize MVC capability.
4) Experiments on eight real-world image datasets demonstrate that DistilMVC outperforms the state-of-the-art (SOTA) clustering performance and can achieve strong robustness.

To the authors' best knowledge, DistilMVC[1] is the first method to incorporate KD into self-supervised feature learning of MVC, providing a novel solution for high-quality MVC methods. This allows MVC models to be embedded into the physical world to learn more consistent representation in broad scenarios in a self-supervised way.

II. Related Work

In this section, we briefly review three lines of related work, deep MVC, contrastive learning, and KD.

A. Deep MVC

As the mainstream type of enhanced multistage clustering approaches, MVC has attracted increasingly wide attention from researchers. Traditional MVC methods [24], [29], [33], [40], [41], [49], [51], [60], [72], [73], [76], [93] have a number of limitations, including high complexity, slow speed, and difficult deployment in real-world scenarios. SimpleMKKM [50] improves clustering by learning optimal coefficients for neighborhood mask matrices, simplifying parameter settings, and achieving global optimization, but ignores the cross-view semantics. SL-CAUBG [90] replaced single-view anchors with cross-view consensus anchors and unified bipartite graphs to improve clustering performance, whereas it increased the complexity cost. Furthermore, FAMKKM [70] integrates the basic partition guided by original kernel information to reduce computational complexity. In recent years, deep learning-based MVC methods [1], [3], [44], [47], [48], [75], [81], [82], [84], [87], [88], [89] have received more and more attention. They exploit the excellent representation ability from multiview data latent clustering patterns. Such methods can be roughly divided into two categories, namely, one-stage and multistage methods. Most of the one-stage methods [47], [65], [96] are designed to work end-to-end. Synchronizing feature learning and clustering taken by this kind of method can effectively reduce the multistage error accumulation and better support streaming data processing. S^{3}OCNet [42] belongs to one-stage clustering and thus does not generate pseudo-labels that can be used in the iterative optimization process,

[1] The code is available at https://github.com/TitusWjt/DistilMVC.git.

but rather optimizes the clustering model directly through backpropagation. Some one-stage methods [63], [69] combine multiview learning and K-means in one step to minimize information loss in clustering and improve performance. However, such methods heavily rely on the quality of feature initialization and may prematurely lock in low-level features of different views leading to fall into local optima [67]. The multistage methods [71], [75], [84] follow the self-supervised learning paradigm, first pretraining for feature learning and then fine-tuning according to different proxy tasks or algorithms. One-stage methods are likely to latch onto low-level features because of their dependence on initialization, so the multistage method with pretraining usually has a better performance in providing higher accuracy (ACC).

The proposed DistilMVC is a multistage MVC framework that requires pretraining to obtain rich prior knowledge, which avoids relying on low-level features in the clustering learning process. Almost all MVC methods do not take into account the inaccurate guidance from the use of pseudo-labels and thus suffer from model degradation. To address this issue, we replace pseudo-labels with dark knowledge from the perspective of KD. Recently, Li et al. [38] used dual attention layers and dual contrastive learning losses to learn view-specific prototypes and model view relationships, whereas it introduces higher computational complexity during iterative optimization of prototypes and data imputation. OPMC [77] also requires additional computational steps or methods for precise matching of structural information between views. Chen et al. [11] learned features through contrastive cluster assignments but ignored semantic consistency, facing challenges related to significant view heterogeneity. On the contrary, DistilMVC excels in capturing and distinguishing subtle differences in multiview data, showing better performance with large-scale and complex data. It ensures semantic consistency across views through multiscale mutual information maximization, even with significant differences between views.

B. Contrastive Learning

Contrastive learning [10], [12], [13], [21], [26] is an essential method for unsupervised learning [6]. Its major goal is to maximize feature space similarity between positive samples while reducing the distance between negative samples. In the field of computer vision, contrastive learning methods have produced excellent results [67]. For example, SimClR [12] or MoCo [26] minimize the InfoNCE loss function [55] to maximize the lower bound of mutual information. Since the processing of negative samples is very cumbersome, the follow-up works, such as BYOL [21], SimSiam [13], and DINO [10], have successfully transformed the contrastive task into a prediction task without defining negative samples and achieved amazing results.

Previous work simply constructs positive and negative samples based on data augmentation. For example, Yan et al. [85] reduce the impact of false negatives in samples by using transition probabilities to accurately identify and minimize the similarity between truly dissimilar sample pairs. Although these studies have shown that consistency could be learned by maximizing the mutual information of different views, they ignore the mutual information at different hierarchies. In contrast, our method aims to learn shared semantics from multiple views. DistilMVC first constructs two independent subspaces and defines positive and negative samples according to the feature matrix in each subspace, respectively, and then uses the InfoNCE loss to maximize the lower bound of mutual information of different views.

C. Knowledge Distillation

KD is a model compression method in which a smaller student model relies on a pretrained teacher model to obtain performance close to or even surpassing the teacher model. In order to help students learn more semantic information, minimizing the loss of the output class probability (soft label) of the teacher model [27] can make the soft label contain rich dark knowledge. Recently, Li et al. [39] proposed a distillation strategy, adaptive sharpening (ADS), which adaptively filters high-confidence predictions through a soft threshold to address the issue of overconfidence. However, ADS must rely on a small amount of manually labeled data for semi-supervised learning, limiting its application in MVC.

The differences between this work and existing KD studies are as follows. DistilMVC adopts a self-distillation [28], [56], [86], [95] method that does not require a pretrained model of the teacher network, nor does it need to detach the gradient of the teacher network. In DistilMVC, the student network and the teacher network do collaborative training, and the teacher network relies on the momentum update [26] of the student network parameters, which is conducive to maintaining consistent semantic information for high-dimensional features. The proposed method extracts the dark knowledge from high-dimensional features, supervises the learning of the student network, and improves the generalization ability of the model [53]. To the best of authors' knowledge, this is the first work that applies KD to MVC, which optimizes pseudo-labels quality and improves the clustering performance.

III. Revisiting KD Used in Multistage Learning Tasks

A multistage deep learning task [46], [67], including multistage MVC [54], [71], [84], leverages K-means and other basic clustering methods [2] to convert high-dimensional features into pseudo-labels to guide learning tasks. However, the distance measures in high-dimensional spaces are not reliable due to dimensional catastrophes, imbalanced data distribution, and noise pollution [22], [62], [68], leading to the overconfidence in K-means or other basic clustering methods and thus the biased pseudo-labeling. As the noise accumulates, the obtained pseudo-labels [56], [83] lose intracluster and intercluster associations, degrading the model prediction performance (low-entropy prediction).

Inspired by the fact that KD is feasible to tackle low-entropy prediction problems [56], [92], we explore the use of KD in multistage learning tasks. More specifically, we perform five experiments, three of which are supervised tasks and two are

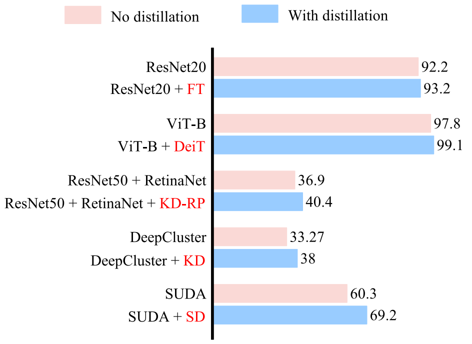


Fig. 2. Comparison of learning performance of visual tasks with or without distillation. In this figure, we display the performance improvements of different feature extractors with an additional distillation process. The performance improves in the cases of using the convolution-based ResNet [25], the self-attention-based ViT [16], the object detection network RetinaNet [46], the convolutional neural network (CNN)-based deep clustering [9], and the unsupervised domain adaptation [20].

unsupervised tasks, and incorporate a KD method into each task. The specific experimental settings are shown in Table I. The corresponding distillation methods are as follows.

1) FT [34] uses convolutional operations to transfer dark knowledge.
2) DeiT [64] proposes the distillation token and uses its representation with the teacher model's dark knowledge to compute the distillation loss.
3) KD-RP [37] exploits the differences in student and teacher networks to guide dark KD.
4) KD [2] provides additional information about semantic similarity to model learning through the use of dark knowledge generated by self-distillation.
5) SD [8] exploits self-distillation to learn effective representations to group point clouds in the target domain.

The experimental results are shown in Fig. 2, with the corresponding distillation methods highlighted in red. The five tasks can all improve the performance of their backbone networks after exploiting the KD. Compared with pseudo-labels, dark knowledge from the teacher contains the similarity information between classes [91], providing richer semantics.

In Section IV, we consider this observation and leverage self-distillation in multistage MVC. For an MVC task, the view data of the same instance are usually weakly correlated, and the view data of different instances are sometimes correlated. These correlations cannot be represented by a pseudo-label. Thus, there is a need to construct a teacher network to capture the distribution of features, which can serve as a self-supervised signal to guide the iteration of the student network. Through training the student, the student's predictions come to match the feature distribution of the teacher and achieve a stable convergence. In other words, when the inputs are going to be noisy, we hope the stable dark knowledge can provide some improvement on the predictions of the student.

IV. Proposed MVC With Self-Distillation Method

Multiview data introduce more features, and thus, overconfident pseudo-labels are poor to represent these features accompanied by more noise, which results in existing multistage clustering methods being difficult to adapt to this MVC scenario.

To solve the abovementioned issues and alleviate the overconfidence of pseudo-labels while learning the common semantics of different views, we propose a novel technique, the multiview distillation technique. Its contrastive method to learn multiview semantics from different hierarchies is present in the first place. Then, we incorporate this technique into a novel multistage MVC framework (DistilMVC).

A. Framework Overview

Given a multiview dataset $\mathcal{X} = \{X^v \in \mathbb{R}^{N \times D_v}\}_{v=1}^{V}$, where each view takes N samples. V denotes the number of views, $v \in \{1, \dots, V\}$. D_v denotes the dimension of the vth view sample X^v, and $k \in \{1, \dots, K\}$ is the number of categories to cluster (see Fig. 3). Overall, DistilMVC projects a given dataset into a feature space wherein information consistency and stability with self-distillation are guaranteed by involving three joint learning objectives.

1) To reconstruct the views and build the feature space, DistilMVC is equipped with an autoencoder for each view, and the encoder and decoder for the view v are denoted by f_v and g_v, respectively. A within-view reconstruction loss is used to learn a view-specific representation so that the trivial feature is abandoned.
2) To thoroughly understand the data and provide a feature space for distillation, a student network (w_s) contains a predictor (w_p) as a cluster head, and a teacher network (w_t) is included, shared by all views and applied to extract multiview features and project the original features to the feature spaces of different hierarchies. DistilMVC learns common semantics by maximizing the mutual information of the feature spaces with different hierarchies. Specifically, the student network and the teacher network will construct two independent high-dimensional subspaces and indirectly improve the lower boundary of mutual information through contrastive learning in their respective subspaces. At the same time, we introduce IIC [31] to directly maximize the mutual information of low-dimensional features.
3) To combat the overconfidence of pseudo-labels, we use the dark knowledge output by the teacher as a new self-supervised signal. The predictor w_p converts the features of w_s into probability distributions and uses them as soft labels for distillation. Specifically, the teacher network outputs k-dimensional features and converts 1-D pseudo-labels into k-dimensional dark knowledge by adjusting the temperature and adding a Softmax activation function. The dark knowledge obtained by the final distillation is used as the ground truth, and the Kullback-Leibler divergence (KL) is used to measure its similarity to the output of the student network.

B. Reconstruction Loss

Deep autoencoders can capture the salient features of data and have applications in many unsupervised

TABLE I

Backbone Settings for Different Vision Tasks and Their Corresponding Improved KD Methods

	Dataset	Backbone	Distillation	Metrics
Image classification	CIFAR10[36]	ResNet20[25]	FT	Accuracy
Image classification	CIFAR10[36]	ViT-B[16]	DeiT-B	Accuracy
Object detection	CoCo[46]	ResNet20+RetinaNet[25, 45]	KD-RP	Average Precision
Image classification	CIFAR10[36]	Deep Cluster[9]	KD	Accuracy
3D vision classification	PointDA-10 [59]	PointNet+DGCNN[58, 74]	SD	Average Precision

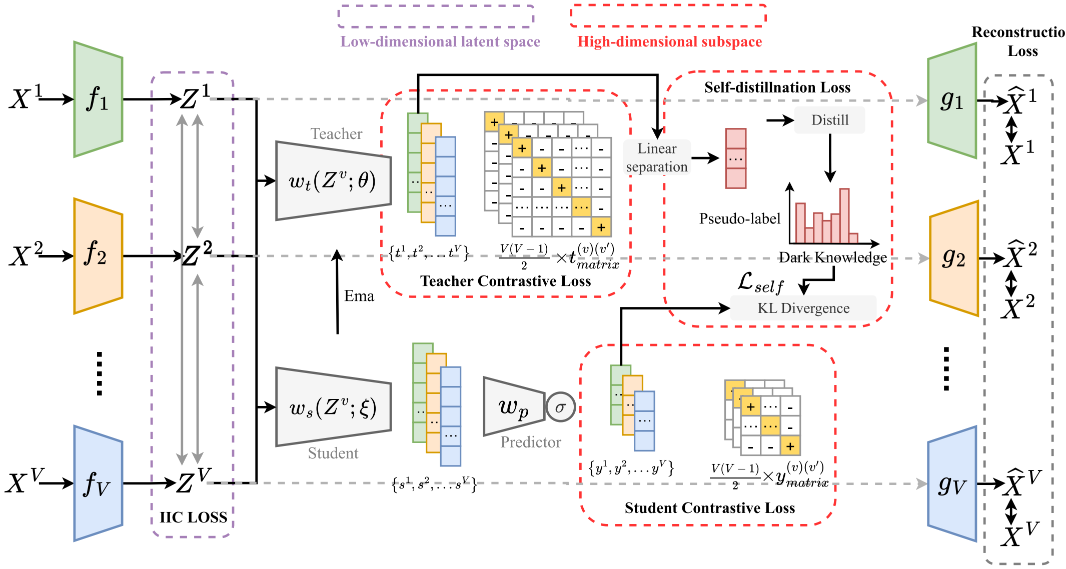


Fig. 3. Framework of the proposed DistilMVC. The encoder f_v and decoder g_v learn latent representation Z^v for the vth view by reconstructing X^v (Section IV-B). The student network w_s and the teacher network w_t capture hierarchical representations through contrastive learning in their subspaces, and the latent representations $\{Z^1, Z^2, \ldots, Z^v\}$ maximize the mutual information pairwise (Section IV-C). The probability distribution of the obtained features of the student network will be compared with the dark knowledge of the teacher network to calculate KL divergence (Section IV-D), where "Ema" denotes the exponential moving average, and the teacher network is updated with momentum by the parameters of the student network.

domains [61], [94]. Therefore, we minimize

$$\mathcal{L}_{\text{rec}} = \sum_{v=1}^{V}\sum_{n=1}^{N}\left\| X_n^v - g^v\left(f^v\left(X_n^v\right)\right)\right\|_2^2 = \sum_{v=1}^{V}\sum_{n=1}^{N}\left\| X_n^v - g^v\left(Z_n^v\right)\right\|_2^2 \quad (1)$$

to enable the autoencoder to convert heterogeneous multiview data into a cluster-friendly latent representation Z^v. For the vth view, X_n^v represents the nth feature vector. The learned latent representation is defined as Z^v, and Z_n^v denotes the nth latent representation. $\hat{X}^v$ is the reconstructed view of Z^v. This design can make the autoencoder maintain the respective diversity of views, avoid the trivial solution, and prevent model collapse, which is the basis for improving the performance of MVC.

C. Contrastive Loss

For the model to perform feature learning effectively, the teacher network and the student network project the low-dimensional representation $\{Z^1, Z^2, \ldots, Z^v\}$ into the higher dimensional spaces $\{t^1, t^2, \ldots, t^v\}$ and $\{y^1, y^2, \ldots, y^v\}$ at different hierarchies, respectively. To enable effective feature learning at different hierarchies, we take the following procedures: 1) optimizing $\mathcal{L}_{\text{stu}}$ and $\mathcal{L}_{\text{tea}}$ to indirectly raise the lower bound of mutual information between views and 2) optimizing $\mathcal{L}_{\text{IIC}}$ to directly maximize the mutual information between views. We propose an objective function for learning common semantics

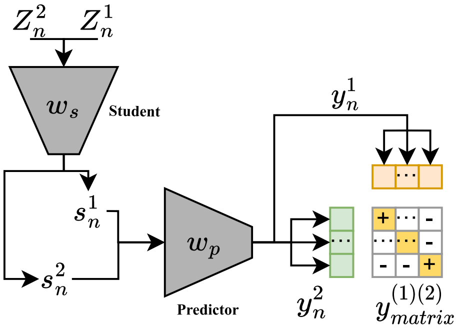


Fig. 4. Calculation of student contrastive loss. A group of shared deep neural networks w_s and w_p are used to extract features from different views. The predictor w_p is used to project the features into high-dimensional subspaces, where y_n^1 and y_n^2 denote the pseudo-labels generated by Softmax operations in this contrastive learning. The feature matrix $y_{\text{matrix}}^{(1)(2)}$ is obtained by multiplying y_n^1 and y_n^2, to learn common semantics.

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{stu}} + \mathcal{L}_{\text{tea}} + \mathcal{L}_{\text{IIC}}. \quad (2)$$

Each component of this objective function will be described in detail in the following.

1) Student Contrastive Loss: Fig. 4 shows how contrastive learning is used in the student network in the example case of

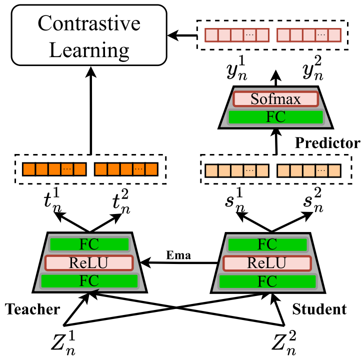


Fig. 5. Illustration of the model structure of the student network and teacher network.

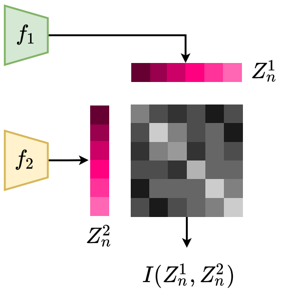


Fig. 6. Calculation of mutual information between two views. The mutual information of Z_n^1 and Z_n^2 can be directly obtained on a joint probability distribution matrix $P_{Z_n^1, Z_n^2}$. The matrix can be calculated by approximating Z_n^1 and Z_n^2 as two independent discrete probability distributions.

$V = 2$. Given a batch of n (Z_n^1, Z_n^2) pairs, a student network is trained to predict which of the $n \times n$ possible (Z_n^1, Z_n^2) pairings across a batch actually occurred. To do this, w_p learns the multiview embedding space feature matrix $y_{\text{matrix}}^{(v)(v')}$ by maximizing the cosine similarity of y_n^1 and y_n^2 of n positive sample pairs on the diagonal while simultaneously minimizing the cosine similarity of the embeddings of $(n^2 - n)$ negative sample pairs. The pairwise similarity in the feature matrix is measured by cosine similarity as

$$\cos\left(y_n^v, y_m^{v'}\right) = \frac{\left(y_n^v\right)\left(y_m^{v'}\right)^\top}{\left\|y_n^v\right\|\left\|y_m^{v'}\right\|} \tag{3}$$

where $n, m \in [1, N]$, $v, v' \in [1, V]$, $N > K$, and $v \neq v'$. In order to optimize the pairwise similarity, without loss of generality, given the sample pairs y_n^v and $y_n^{v'}$, we optimize the symmetric cross-entropy loss

$$\begin{aligned}\ell_y^{(v)(v')} &= -\frac{1}{2K}\sum_{k=1}^{K}\log \\ &\times \frac{\exp\left(\cos\left(y_k^v, y_k^{v'}\right)/\tau_s\right)}{\sum_{m=1}^{K}\left[\exp\left(\cos\left(y_k^v, y_m^v\right)/\tau_s\right) + \exp\left(\cos\left(y_k^v, y_m^{v'}\right)/\tau_s\right)\right]}\end{aligned} \tag{4}$$

where τ_s is the student network temperature parameter that controls the softness of the distribution. Since we wish to identify all positive pairs of the entire dataset, the contrastive loss of sample pairs s_n^v and $s_n^{v'}$ needs to be computed on all views, which we extend to $V \geq 2$ as follows:

$$\mathcal{L}_{\text{stu}} = \sum_{v=1}^{V}\sum_{v\neq v'}\ell_y^{(v)(v')} - H(Y). \tag{5}$$

In (5), we add an additional entropy balance term

$$H(Y) = -\sum_{v=1}^{V}\left[P(y^v)\log P(y^v) + P\left(y^{v'}\right)\log P\left(y^{v'}\right)\right]. \tag{6}$$

This regularization term avoids the trivial solution and prevents all sample points from clustering into the same class.

2) Teacher Contrastive Loss: As shown in Fig. 5, both the teacher network and the student network use the same feature learning methods. The only distinction is that the teacher network does not require an additional regularization term to prevent model collapse. The goal of the teacher network is to provide a supervised signal for the optimization of the student network while providing high-dimensional features $\{t^1, t^2, \ldots, t^v\}$ for linear separation. Specifically, we still learn the mutual information of high-dimensional subspace through contrastive learning and provide high-dimensional features to cover the correlations and probability distributions among samples. This also establishes feature space for subsequent self-distillation.

We give the sample pair t_n^v and $t_n^{v'}$ to optimize the symmetric cross-entropy loss as

$$\begin{aligned}\ell_t^{(v)(v')} &= -\frac{1}{2N}\sum_{n=1}^{N}\log \\ &\times \frac{\exp\left(\cos\left(t_n^v, t_n^{v'}\right)/\tau_t\right)}{\sum_{m=1}^{N}\left[\exp\left(\cos\left(t_n^v, t_m^v\right)/\tau_t\right) + \exp\left(\cos\left(t_n^v, t_m^{v'}\right)/\tau_t\right)\right]}\end{aligned} \tag{7}$$

where τ_t is the temperature parameter. Considering all views on the dataset, we give the optimization objective of the teacher network as

$$\mathcal{L}_{\text{tea}} = \sum_{v=1}^{V}\sum_{v\neq v'}\ell_t^{(v)(v')}. \tag{8}$$

3) IIC Loss: Minimizing InfoNCE [55] at high-dimensional hierarchies can be seen as maximizing the lower bound of mutual information indirectly. That is, $I(y^v, y^{v'}) \geq \log(n^2 - n) - \mathcal{L}_{\text{stu}}$, where $I(y^v, y^{v'})$ denotes the mutual information between s^v and $s^{v'}$, $(n^2 - n)$ is the number of negative samples, and similarly $I(t^v, t^{v'}) \geq \log(n^2 - n) - \mathcal{L}_{\text{tea}}$. Different from the above methods, we directly maximize the mutual information between different views in low-dimensional hierarchies

$$\mathcal{L}_{\text{IIC}} = -\sum_{v=1}^{V}\sum_{v\neq v'}\sum_{n=1}^{N} I\left(Z_n^v,\ Z_n^{v'}\right) \tag{9}$$

where I represents the mutual information. As shown in Fig. 6, according to IIC [31], we approximate Z_n^v and $Z_n^{v'}$ into two independent discrete distributions and further obtain the joint probability distribution of Z_n^v and $Z_n^{v'}$. Therefore, I is directly calculated by

$$\mathbb{E}_{P_{Z_n^1,Z_n^2}}\left(P_{Z_n^1,Z_n^2}\log\frac{P_{Z_n^1,Z_n^2}}{P_{Z_n^1}P_{Z_n^2}}\right). \tag{10}$$

D. Self-Distillation Loss

To make better use of the learned common semantics for clustering, we need to add some interactions for the two independent student and teacher subspaces for fine-tuning. The teacher network and the student network use the same network structure, but the network parameters are different. The teacher network is updated in the form of a moving average [26], introducing a momentum encoder to provide a regression target for the student network. θ and ξ denote the learnable parameters of teacher network w_t and student network w_s, respectively. The parameter θ is ξ exponential moving average. With the target momentum being $\mu \in [0, 1]$, the parameter θ is updated with

$$\theta \leftarrow \mu\theta + (1-\mu)\xi. \tag{11}$$

We do not use the soft labels output by the teacher network directly as the distribution required for distillation because such probability distributions do not contain obvious clustering information. We will first use the cluster information contained in the high-level features to improve the clustering effect of semantic labels, and a new cluster center C can be obtained by optimizing the following objectives:

$$\min_{\{\mathbf{C}^v\}_{v=1}^V}\sum_{n\in\mathcal{X}}\sum_{m=1}^{K}\sum_{v=1}^{V}\left\|\theta z_n^v - c_m^v\right\|_2^2 = \min_{\mathbf{C}}\sum_{n\in\mathcal{X}}\sum_{m=1}^{K}\|t_n - c_m\|_2^2 \tag{12}$$

where θ is the parameter of the teacher network, $\mathbf{C} \in \mathbb{R}^{K\times\sum_{v=1}^V d_v}$, $c_m = (c_m^1, c_m^2, \ldots, c_m^V) \in \mathbb{R}^{K\times\sum_{v=1}^V d_v}$, and d_v is the dimension of t_n. This step is more efficient with the K-means algorithm, so we can linearly separate the t_n according to the cluster center c to get the V group of pseudo-labels $\{\mathbf{P}^v = \text{argmin}_m \|t_n^v - c_n^v\|_2^2\}_{v=1}^V$. The Softmax activation function will be stacked to the predictor's final layer, and s_{nm}^v is defined as the probability that the nth sample is clustered into the mth cluster for the vth view, so there are also V groups of probability distributions $\{\mathbf{l}^v = \text{argmax}_m\, y_{nm}^{(v)}\}_{v=1}^V$. However, $\mathbf{P}^v$ and $\mathbf{I}^v$ are not aligned, so we need to define a loss matrix $\mathbf{M} \in \mathbb{R}^{K\times K}$ to help us correct $\mathbf{P}^v$ [43], $\tilde{\mathbf{m}}_{nm} = \sum_{n\in\mathcal{X}} \nVdash[l_i^v = n]\nVdash[l_i^v = m]$, element $\mathbf{m}_{nm} = \max_{n,m}\tilde{\mathbf{m}}_{nm} - \tilde{\mathbf{m}}_{nm}$. The alignment problem will be treated as a maximum matching problem

$$\min_{\mathbf{A}}\sum_{i=1}^{K}\sum_{j=1}^{K} m_{ij}a_{ij}$$
$$\text{s.t. } \mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{I}_K \tag{13}$$

where $\mathbf{A} \in \mathbb{R}^{K\times K}$ is a Boolean matrix, and (13) is optimized using the Hungarian algorithm [32] to get $\{\mathbf{P}^{*v}\}_{v=1}^V$. Here, the dark knowledge is defined as $[(1-\tau_d)\mathbf{P}^{*v} + \tau_d u]$, and we use the KL divergence distillation model

$$\begin{aligned}\mathcal{L}_{\text{self}} &= \sum_{v=1}^{V} D_{\text{KL}}\big([(1-\tau_d)\mathbf{P}^{*v} + \tau_d u], y^v\big)\\ &= -\sum_{v=1}^{V}\big[(1-\tau_d)\mathbf{P}^{*v} + \tau_d u\big]\log\frac{[(1-\tau_d)\mathbf{P}^{*v} + \tau_d u]}{y^v}\end{aligned} \tag{14}$$

where τ_d is a distillation factor, and u is a Gaussian distribution. y^v is a sharp distribution whereas the dark knowledge is a smooth distribution, and thus, the above KL divergence can make them form a confrontation, effectively preventing the model from collapsing. Empirically, we set $\tau_d = 0.1$.

E. Training and Inference

$\mathcal{L}_{\text{rec}}$ is the reconstruction loss of the autoencoder, and $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{self}}$ implement feature learning and label distillation, respectively. A dynamic balance factor is usually used to measure the loss throughout the training process [21]. But in practice, we have found that simply adding together all these losses works well, so there is no need to set the balance factor.

During the pretraining stage, we fed the dataset $\mathcal{X}$ to DistilMVC and use $(\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{con}})$ as the objective function for training. Learning different hierarchies of mutual information can provide rich semantic knowledge, which lays the foundation for subsequent distillation. The pretrained model is loaded and fine-tuned by optimizing $(\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{self}})$ to alleviate the wrong traction of pseudo-labels and improve the clustering performance. In our design, even though the initial teacher may not necessarily be accurate, such a weak teacher can still work because our KD is essentially a regularization process [17], [91]. Hence, our self-distillation loss is equal to

$$\begin{aligned}\mathcal{L}_{\text{self}} &= \sum_{v=1}^{V} H\big([(1-\tau_d)\mathbf{P}^{*v} + \tau_d u], y^v\big)\\ &= \sum_{v=1}^{V}[(1-\tau_d)\mathbf{P}^{*v} + \tau_d u]\log y^v\\ &= \sum_{v=1}^{V}(1-\tau_d)H(\mathbf{P}^{*v}, y^v) + \tau_d H(u, y^v)\\ &= \sum_{v=1}^{V}(1-\tau_d)H(\mathbf{P}^{*v}, y^v) + \tau_d(D_{\text{KL}}(u, y^v) + H(u)).\end{aligned} \tag{15}$$

Since the entropy $H(u)$ is constant, (14) is equal to

$$\mathcal{L}_{\text{self}} = \sum_{v=1}^{V}((1-\tau_d)H(\mathbf{P}^{*v}, y^v) + \tau_d D_{\text{KL}}(u, y^v)). \tag{16}$$

$\mathcal{L}_{\text{self}}$ not only minimizes the prediction error $(H(\mathbf{P}^{*v}, y^v))$ between the teacher and the student, but also includes $D_{\text{KL}}(u, y^v)$ as a regularization term for a label smoothing regularization. This term penalizes predictions that deviate from the distribution u, thereby reducing the student's overconfidence in any specific class even though the input instance does not belong to that class. As a result, the student's depiction will not be exclusively fixed on one class, but will be distributed among all possible classes. This approach alleviates the dependence on multiview data with noise, boosting the robustness and of the model. τ_d adjusts the weight of the regularization term, with a higher τ_d value means a stronger smoothing effect. For a weak teacher, we adapt τ_d value as well. Therefore, our self-distillation process can improve the generalization of the student in case the teacher is unreliable or even quite weak.

In the inference stage, we fed the multiview data to DistilMVC, and the predictor w_p in the student network will obtain the probability distribution of all view clusters $\{y_{nm}^{(v)}\}_{v=1}^{V}$, which is weighted and summed on each view to get the final clustering result, $\text{argmax}_m((1/V)\ \text{sum}_{v=1}^{V} y_{nm}^{v})$. The detailed steps are summarized in Algorithm 1.

Algorithm 1 Model Training With DistilMVC

```
Input: dataset $\mathcal{X}$; pretraining epochs $T_p$; fine-tuning
       epochs $T_t$; view number $V$; cluster number $K$;
       encoder $f^v$; decoder $g^v$; predictor $w_p$; student
       $w_s$; teacher $w_t$.
Output: clustering assignments $\mathcal{Y}$
Initialize:
Sample multi-view data $\{X^v\}_{v=1}^{V}$ from $\mathcal{X}$
Get the latent representation by $Z^v = f^v(X^v)$
Get the student representations by $y^v = w_p(w_s(Z^v))$
Get the teacher representations by $t^v = w_t(Z^v)$
(Procedure 1) Pretraining Stage:
for $epoch = 1$ to $T_p$ do
    if the $f^v$ and $g^v$ are not convergence then
        Get the reconstruction loss $\mathcal{L}_{rec}$ by Eq. (1)
    else
        Get the student contrastive loss $\mathcal{L}_{stu}$ by Eq. (5)
        Get the teacher contrastive loss $\mathcal{L}_{tea}$ by Eq. (8)
        Get the invariant information clustering loss $\mathcal{L}_{IIC}$
        by Eq. (9)
        Update $f^v, g^v, w_p, w_s, w_t$ through gradient
        descent to minimize Eqs. (1) and (2).
    end
end
(Procedure 2) Fine-tuning Stage:
Load model weights and $\mathcal{X}$
for $epoch = 1$ to $T_t$ do
    Distill dark knowledge by Eq. (14)
    Get the overall loss $\mathcal{L}$ by Eqs. (1), (2) and (14).
    Update $f^v, w_p, w_s, w_t$ through minimizing $\mathcal{L}$
    Momentum update $w_s \leftarrow w_t$ by Eq. (11)
end
(Procedure 3) Inference Stage:
    Compute predictions by $y^v = w_p(w_s(f^v(X^v)))$
    Get cluster assignment by
$\mathcal{Y} = \text{argmax}\left(\frac{1}{V}\ sum_{v=1}^{V} y^v\right)$
```

V. Experiments

In this section, we evaluate the proposed DistilMVC method on eight widely used multiview datasets and compare it with eight SOTA clustering methods.

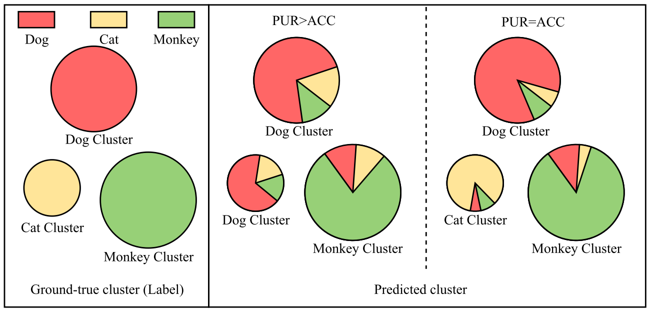


Fig. 7. Relation between PUR and ACC values. PUR = ACC indicates that there is a one-to-one correspondence between the predicted labels of clusters and their ground-true labels. When PUR > ACC, there exist duplicated clusters. Since the proportion of "dog" in the predicted clusters is larger, there are two clusters marked with the label "dog."

A. Datasets and Experimental Settings

1) Comparisons With State of the Arts: The comparison methods include three traditional methods (i.e., MVC-LFA [71], IMVTST-MVI [76], and SL-CAUBG [90]) and seven deep methods (i.e., CDIMC-net [75], EAMC [96], SiMVC [65], CoMVC [65], COMPLETER [47], SURE [88], and MFLVC [84]). For all methods, we use the recommended model structure and parameters for fair comparisons.

2) Datasets: In our experiments, we used eight datasets: Scene [18], MNIST-USPS [57], BDGP [7], Fashion [78], Caltech-2V, Caltech-3V, Caltech-4V, and Caltech-5V. To evaluate the robustness of DistilMVC over the number of views, Caltech [19] as a multiview RGB image dataset is disassembled into Caltech-2V, Caltech-3V, Caltech-4V, and Caltech-5V. Table II describes the datasets used in more detail. As the most popular dataset used in MVC, their feature dimensions are adequate to capture essential characteristics and conform to domain standards, thereby supporting the reliability of our experimental results.

3) Experimental Implementation: We conduct all the experiments on the platform of Ubuntu 16.04 with Tesla P100 graphics processing units (GPUs) and 32G memory size. Our model and baseline are built on the PyTorch 1.11.0 framework. Based on extensive ablation studies, the batch size is set to 128, and the epochs for the two phases of pretraining and fine-tuning were set to 150 and 50, respectively. The temperature parameters τ_s, τ_t, and τ_d are fixed to 0.5, 1.0, and 0.1, respectively. We use Adam optimizer [35] with the default parameters to train our model and set the initial learning rate as 0.0001. The structure of the autoencoder for the vth view is defined as $X^v - \text{Fc}_{512} - \text{Fc}_{1024} - \text{Fc}_{2048} - \text{Fc}_{512} - Z^v - \text{Fc}_{512} - \text{Fc}_{2048} - \text{Fc}_{1024} - \text{Fc}_{512} - \hat{X}^v$, where Fc_{512} denotes a fully connected neural network with 512 neurons, and each layer is followed by a rectified linear unit (ReLU) layer. As shown in Fig. 5, the teacher network structure and the student network structure have two linear layers each, and the ReLU activation function is added in the middle.

4) Evaluate Metrics: The clustering performance is evaluated with three metrics: ACC, normalized mutual information (NMI), and purity (PUR). More details on these indicators can be found in [4]. A higher value of these evaluation indicators can reflect a better clustering performance.

TABLE II
DATASET SUMMARY

Datasets	Sample	Type	Views	# of categories	Dimension
Caltech-2V	1,400	WM and CENTRIST	2	7	40/254
Scene	4,485	PHOG and GIST	2	15	20/59
MNIST-USPS	5,000	Two styles of digital images	2	10	784/784/784
BDGP	2,500	Visual and textual views	2	5	1750/79
Caltech-3V	1,400	WM, CENTRIST, and LBP	3	7	40/254/928
Caltech-4V	1,400	WM, CENTRIST, LBP, and GIST	4	7	40/254/928/512
Caltech-5V	1,400	WM, CENTRIST, LBP, GIST, and HOG	5	7	40/254/928/512/1984
Fashion	10,000	Three styles of images [84]	3	10	784/784/784

TABLE III
PERFORMANCE COMPARISONS ON FOUR DUAL-VIEW DATASETS. THE FIRST-BEST RESULTS ARE INDICATED IN RED, AND THE SECOND-BEST RESULTS ARE INDICATED IN BLUE

Datasets	Caltech-2V			Scene			MNIST-USPS			BDGP		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
MVC-LFA [71](2019)	0.462	0.348	0.496	0.357	0.391	0.384	0.768	0.675	0.768	0.564	0.395	0.612
CDIMC-net [75](2020)	0.515	0.480	0.564	0.346	0.374	0.351	0.620	0.676	0.647	0.884	0.799	0.885
EAMC [96] (2020)	0.419	0.256	0.427	0.250	0.319	0.263	0.735	0.837	0.778	0.681	0.480	0.697
IMVTST-MVI[76](2021)	0.409	0.398	0.540	0.340	0.312	0.181	0.669	0.592	0.717	0.981	0.950	0.982
SiMVC [65](2021)	0.508	0.471	0.557	0.289	0.281	0.293	0.981	0.962	0.981	0.704	0.545	0.723
CoMVC [65](2021)	0.466	0.426	0.527	0.306	0.303	0.314	0.987	0.976	0.989	0.802	0.670	0.803
COMPLETER [47](2021)	0.599	0.572	0.612	0.391	0.415	0.401	0.989	0.971	0.989	0.960	0.950	0.963
SURE [88](2022)	0.548	0.471	0.580	0.417	0.426	0.441	0.992	0.977	0.992	0.907	0.794	0.907
MFLVC [83](2022)	0.606	0.528	0.616	0.401	0.428	0.443	0.995	0.985	0.995	0.989	0.966	0.976
SL-CAUBG [90](2023)	0.594	0.484	0.609	0.405	0.381	0.411	0.991	0.979	0.991	0.984	0.958	0.989
DistilMVC(ours)	0.619	0.533	0.619	0.428	0.432	0.448	0.996	0.987	0.996	0.991	0.971	0.991

TABLE IV
PERFORMANCE COMPARISON OVER FOUR MULTIVIEW DATASETS. THE SYMBOL "–" DENOTES UNKNOWN RESULTS, AS COMPLETER AND SURE MAINLY FOCUS ON TWO-VIEW CLUSTERING

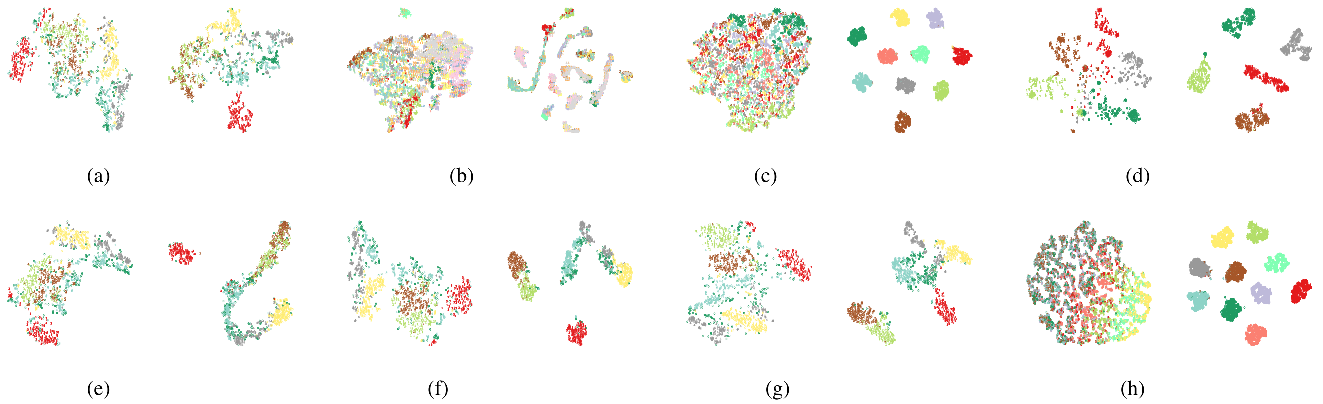
Datasets	Caltech-3V			Caltech-4V			Caltech-5V			Fashion		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
MVC-LFA [71](2019)	0.551	0.423	0.578	0.609	0.522	0.636	0.741	0.601	0.747	0.791	0.759	0.794
CDIMC-net [75](2020)	0.528	0.483	0.565	0.560	0.564	0.617	0.727	0.692	0.742	0.776	0.809	0.789
EAMC [96](2020)	0.389	0.214	0.398	0.356	0.205	0.370	0.318	0.173	0.342	0.614	0.608	0.638
IMVTST-MVI[76](2021)	0.558	0.445	0.576	0.687	0.610	0.719	0.760	0.691	0.785	0.632	0.648	0.635
SiMVC [65](2021)	0.569	0.495	0.591	0.619	0.536	0.630	0.719	0.677	0.729	0.825	0.839	0.825
CoMVC [65](2021)	0.541	0.504	0.584	0.568	0.569	0.646	0.700	0.687	0.746	0.857	0.864	0.863
COMPLETER [47](2021)	–	–	–	–	–	–	–	–	–	–	–	–
SURE[88](2022)	–	–	–	–	–	–	–	–	–	–	–	–
MFLVC [83](2022)	0.631	0.566	0.639	0.733	0.652	0.734	0.804	0.703	0.804	0.992	0.980	0.992
SL-CAUBG [90](2023)	0.603	0.437	0.634	0.658	0.489	0.658	0.810	0.674	0.812	0.975	0.971	0.981
DistilMVC(ours)	0.650	0.575	0.663	0.809	0.695	0.809	0.824	0.709	0.824	0.993	0.982	0.993

B. Experimental Results and Analysis

Tables III and IV list the clustering performances of all methods on eight datasets, from which we obtain the following observations.

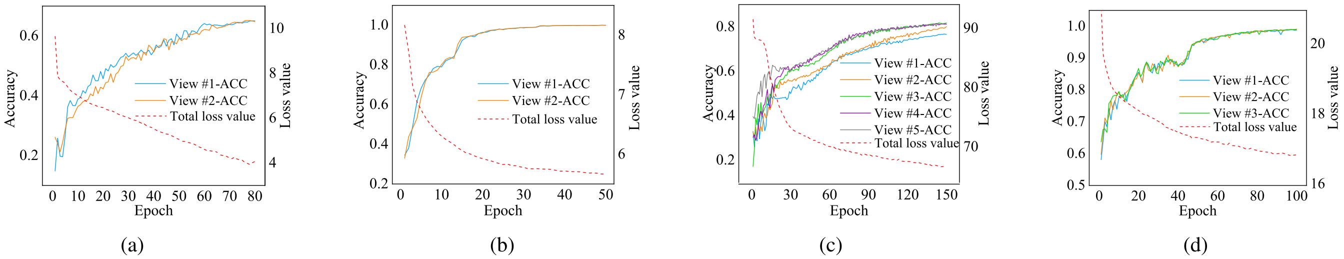
1) Our DistilMVC achieves the best performance on all datasets. Compared with the second-best method, DistilMVC has a significant improvement, especially surpassing 7.6% on the Caltech-4V dataset.
2) COMPLETER and SURE suffer from missing and unaligned data problems, respectively, so we evaluated the above two methods using complete and aligned data and found that they still significantly underperformed DistilMVC.
3) On the Caltech dataset, DistilMVC shows considerable improvement as the number of views increases.
4) PUR calculates the proportion of the samples in a cluster with the ground-true label [4]. ACC only concerns about the best-matched cluster with the ground-true label [32].

Therefore, the case that some clusters share the same label will lead to PUR $>$ ACC [52]. Our DistilMVC obtains the same value for both ACC and PUR on all six datasets, which indicates that there is a strict one-to-one relation between the predicted clusters by DistilMVC and the ground-true clusters, i.e., no cluster's labels are duplicated, ensuring that the semantics of each predicted cluster are independent of each other (see Fig. 7). In contrast, the PUR values of all other


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 8. Visualization on eight datasets via t-SNE [66]. For each dataset, we visualize the fused representation of different views and the fused representation obtained by the student network after DistilMVC training. (a) Caltech-2V. (b) Scene. (c) MNIST-USPS. (d) BDGP. (e) Caltech-3V. (f) Caltech-4V. (g) Caltech-5V. (h) Fashion.



(a) (b) (c) (d)

Fig. 9. Clustering ACC of DistilMVC. The x-axis denotes the training epochs on four datasets, and the left and right y-axes denote the clustering ACC and corresponding loss value, respectively. (a) Convergence on Caltech-2V. (b) Convergence on BDGP. (c) Convergence on Caltech-5V. (d) Convergence on fashion.

methods are higher than their ACC values. This also confirms the robustness of our method.

The reasons for the above observations can be explained as follows.

1) None of the baselines take into account the over-confident traction of inaccurate pseudo-labels, resulting in limited clustering quality.
2) COMPLETER and SURE suffer from a lack of deep mining of mutual information at different hierarchies.
3) With the increase in data views, not only it is the inherent noise of the data introduced, but also it leads to the mistakes of positive sample pairs as negative ones. DistilMVC can filter out some of the inconsistent noise and provide more stable clustering according to smooth dark knowledge rather than pseudo-labels. The Fashion dataset has only the least number of views, i.e., three views, and has fewer noises compared to the other datasets. All models thus easily capture the underlying patterns and achieve advanced clustering results, leading to the improvement of DistilMVC on the baselines is limited.
4) PUR values of all other methods are higher than their ACC values, which means different predicted clusters share the same label.

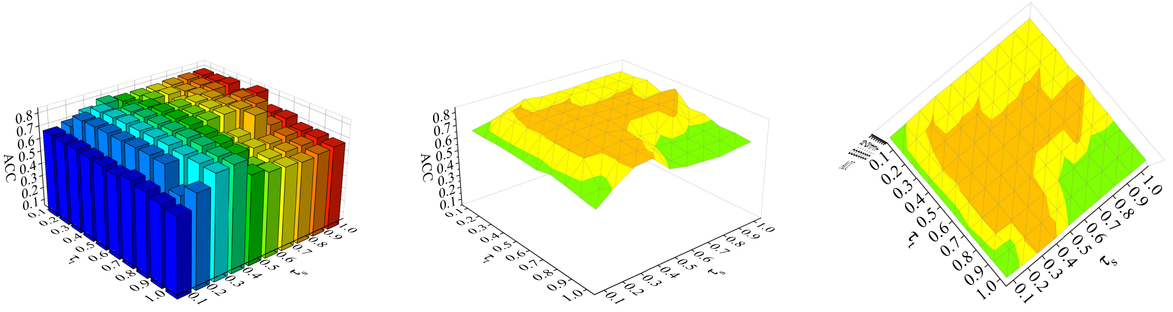
Over-confident pseudo-labels generated by baselines provide incorrect clustering directions. On the other hand, DistilMVC uses dark knowledge instead of pseudo-labels to provide a more precise guide for self-supervised clustering and thus corrects the false clustering directions, while using the Hungarian algorithm to ensure that the label of each cluster is distinct. Hence, the ground-true cluster labels and predicted cluster labels have one-to-one correspondence. This is the core idea of multiview self-distillation.

Unlike traditional and existing deep MVC approaches, our DistilMVC targets to further optimize the pseudo-label learning. The overconfidence of pseudo-labels is alleviated by self-distillation, and robust clustering results are obtained by learning different hierarchies of mutual information to enforce the consistency of different views. In addition to the clustering performance, the visualization of the learned available features is shown in Fig. 8. All datasets except Caltech-2V eventually converge well, and Caltech-2V has poor clustering due to its large number of views and small number of samples. We also find that the data distribution becomes more compact and independent through training, and the clustering density is higher, indicating that our multiview self-distillation method achieves an effective improvement in clustering performance.

C. Model Analysis

1) Convergence Analysis: We investigate the convergence of DistilMVC by reporting the loss value and the corresponding clustering performance with increasing epochs. As shown in Fig. 9, one could observe that the loss remarkably decreases in the first 20 epochs, and meanwhile, the ACC of different views continuously increases and tends to be smooth and consistent.

2) Parametric Analysis: The temperature hyperparameters τ_s [see (4)] and τ_t [see (7)] are used to control the shape of the distribution. As shown in Fig. 10(a), we change their values in the range of [0.1, 1.0] and the interval is 0.1.



(a) (b) (c)

Fig. 10. Clustering performance of DistilMVC on the Caltech-5V dataset with different parameters τ_s and τ_t, including (a) 3-D bar graph and (b) and (c) 3-D surface graph. In the (b) and (c) 3-D surface graphs, the green region, yellow region, and orange region indicate that the ACC is in the ranges (0.6, 0.7], (0.7, 0.8], and (0.8, 0.9], respectively.

TABLE V

ABLATION STUDIES ON LOSS COMPONENTS ON CALTECH-2V, CALTECH-3V, CALTECH-4V, AND CALTECH-5V. "✓" DENOTES DISTILMVC WITH THE COMPONENT, AND "*" INDICATES THE METHOD OF ADDING SELF-DISTILLATION ON THE ORIGINAL MODEL

	$\mathcal{L}_{rec}$		$\mathcal{L}_{con}$		$\mathcal{L}_{self}$	Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
		$\mathcal{L}_{tea}$	$\mathcal{L}_{stu}$	$\mathcal{L}_{IIC}$		ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
(1)	✓					0.3043	0.1965	0.3043	0.1614	0.0162	0.1614	0.1429	0.0043	0.1429	0.1429	0.0101	0.2450
(1*)	✓				✓	0.4886	0.3099	0.5036	0.4736	0.3285	0.4993	0.4621	0.3634	0.4786	0.4914	0.3798	0.5243
(2)		✓	✓	✓		0.5286	0.4365	0.5464	0.5071	0.4357	0.5114	0.5393	0.4395	0.5279	0.7350	0.5902	0.7350
(2*)		✓	✓	✓	✓	0.5136	0.4513	0.5414	0.4793	0.4461	0.5129	0.5086	0.4954	0.5400	0.7321	0.5910	0.7921
(3)	✓	✓		✓		0.3864	0.3090	0.3864	0.1429	0.0009	0.1429	0.1436	0.0018	0.1436	0.3543	0.2414	0.3657
(3*)	✓	✓		✓	✓	0.5507	0.4472	0.5514	0.5871	0.5175	0.5921	0.6271	0.5768	0.6271	0.7600	0.6929	0.7600
(4)	✓		✓	✓		0.5650	0.5033	0.5871	0.6200	0.5270	0.6286	0.7250	0.6528	0.7350	0.7643	0.6904	0.7643
(4*)	✓		✓	✓	✓	0.5621	0.5214	0.5686	0.5836	0.5039	0.6029	0.6671	0.6158	0.6821	0.7443	0.6522	0.7443
(5)	✓	✓	✓			0.5814	0.5055	0.5921	0.6364	0.5654	0.6536	0.7971	0.6838	0.7971	0.7971	0.6838	0.7971
(5*)	✓	✓	✓		✓	0.5843	0.5327	0.5864	0.6371	0.5649	0.6543	0.8057	0.6954	0.8057	0.8057	0.6954	0.8057
(6)	✓	✓	✓	✓		0.5779	0.4958	0.5921	0.6343	0.5659	0.6536	0.7993	0.6863	0.7993	0.8171	0.6930	0.8171
(6*)	✓	✓	✓	✓	✓	0.6192	0.5329	0.6192	0.6500	0.5751	0.6629	0.8086	0.6951	0.8086	0.8236	0.7090	0.8239

In Fig. 10(c), the orange region belongs to the temperature comfort zone, accounting for 37.04% of the total region and is in the center. The dark knowledge in this region contains rich semantic information, i.e., the KL divergence between the dark knowledge and the output distribution of the student network is lower, which also proves that DistilMVC can bring high-quality supervision to the student network. The yellow and green regions account for 45.73% and 17.23% of the total region, respectively, and are distributed at the edges. The yellow region is between the orange region and the green region, which is a buffer zone, and the clustering performance decreases slightly in this region. The green region proves that the temperatures τ_s and τ_t are too large or too small, which will obviously reduce the clustering performance, so our choice needs to avoid the green region. The reasons are as follows: 1) when τ_s and τ_t are close to 1 at the same time, they will enter the green region. The reason is that the temperature τ_s and τ_t are too large and the distribution is too smooth, so the model fails to learn the focus and collapses and 2) when τ_t is 0.1, it will enter the green region. The reason is that the temperature τ_t is too small and the distribution is too peak, so the model will pay special attention to difficult negative samples, making it difficult for the model to converge or the learned features to generalize.

3) Ablation Experiment: We perform the ablation study to demonstrate the importance of each component of our method. As shown in Table V, we designed six sets of schemes on four datasets with different numbers of views and observed the following results.

1) All losses play an integral role in DistilMVC.
2) A significant improvement is obtained after introducing the self-distillation method on (1), (3), (5), and (6), which further proves that our method can effectively mitigate the problem of overconfidence in pseudo-labels and thus improve the clustering performance.
3) The addition of self-distillation in (2) and (4) leads to model degradation.
4) Comparing (1) and (6), we can see that optimizing the loss $\mathcal{L}_{\text{con}}$ can lead to a huge improvement, proving the effectiveness of our proposed method for maximizing mutual information at different hierarchies.
5) The above four observations hold for all datasets, which also demonstrates the robustness of our method.

The reasons for the above observations can be explained as follows.

1) $\mathcal{L}_{\text{rec}}$ establishes the feature space for feature learning, $\mathcal{L}_{\text{con}}$ learns features by maximizing mutual information at different hierarchies, and $\mathcal{L}_{\text{self}}$ improves error prediction by reducing the confidence of the model, and each of the three components is responsible for and reinforces each other.
2) The pseudo-labels are derived from the high-dimensional features learned by the teacher network, and the self-distillation method can transform the

pseudo-labels into dark knowledge, improving the quality of the supervised signal.

3) View reconstruction is conducive to maintaining the complementarity between views, which is the basis of feature learning. If $\mathcal{L}_{\text{rec}}$ is skipped and $\mathcal{L}_{\text{con}}$ is directly optimized, complementary information will be lost. Therefore, for (2), the features learned by the teacher network are not linearly separable due to the lack of complementary information, so they are not suitable for distillation. For (4), teacher networks are not involved in learning, and inaccurate distillation can provide more false labels to student networks.
4) Optimized $\mathcal{L}_{\text{con}}$ can maximize mutual information at different hierarchies from teacher, student, and encoder, which greatly facilitates consistent learning.
5) DistilMVC has strong generalization ability and robustness. Thus, multiview self distillation is well-suited for feature learning and clustering in stages for highly qualified clustering.

VI. Conclusion

In this article, we propose a novel and flexible DistilMVC, which can handle all kinds of multiview data to enable effective MVC. Based on a self-distilled architecture, DistilMVC can effectively alleviate false predictions caused by overconfidence in pseudo-labels, and when combined with a feature learning method of different hierarchies of mutual information, it achieves SOTAs on eight datasets. Thus, it solves a persistent nuisance of MVC: the pseudo-labels obtained by feature learning are not adequate for self-supervised signals. Such a unified framework will provide novel insight for the community to understand MVC. In the future, we plan to further explore the potential of our theory and framework for other multiview learning tasks, such as incomplete MVC, cross-modal retrieval, and 3-D reconstruction.

References

[1] M. Abavisani and V. M. Patel, “Deep multimodal subspace clustering networks,” *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1601–1614, Dec. 2018.

[2] M. Adnan, Y. A. Ioannou, C.-Y. Tsai, and G. W. Taylor, “Domain-agnostic clustering with self-distillation,” 2021, *arXiv:2111.12170*.

[3] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, “Self-supervised learning by cross-modal audio-video clustering,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9758–9770.

[4] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, “A comparison of extrinsic clustering evaluation metrics based on formal constraints,” *Inf. Retr.*, vol. 12, no. 4, pp. 461–486, Aug. 2009.

[5] E. Arazo, D. Ortego, P. Albert, N. E. O’Connor, and K. McGuinness, “Pseudo-labeling and confirmation bias in deep semi-supervised learning,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[6] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[7] X. Cai, H. Wang, H. Huang, and C. Ding, “Joint stage recognition and anatomical annotation of Drosophila gene expression patterns,” *Bioinformatics*, vol. 28, no. 12, pp. i16–i24, Jun. 2012.

[8] A. Cardace, R. Spezialetti, P. Z. Ramirez, S. Salti, and L. D. Stefano, “Self-distillation for unsupervised 3D domain adaptation,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4155–4166.

[9] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.

[10] M. Caron et al., “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[11] J. Chen, H. Mao, W. L. Woo, and X. Peng, “Deep multiview clustering by contrasting cluster assignments,” 2023, *arXiv:2304.10769*.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[13] X. Chen and K. He, “Exploring simple Siamese representation learning,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Jun. 2021, pp. 15750–15758.

[14] T. M. Cover, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,” *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965.

[15] C. Deng, Z. Lv, W. Liu, J. Huang, D. Tao, and X. Gao, “Multi-view matrix decomposition: A new scheme for exploring discriminative information,” in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3438–3444.

[16] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.

[17] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, and Z. Liu, “SEED: Self-supervised distillation for visual representation,” 2021, *arXiv:2101.04731*.

[18] F. Li and P. Pietro, “A Bayesian hierarchical model for learning natural scene categories,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2005, pp. 524–531.

[19] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories,” in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2004, p. 178.

[20] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 2, Jul. 2015, pp. 1180–1189.

[21] J.-B. Grill et al., “Bootstrap your own latent-a new approach to self-supervised learning,” in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[22] L. J. Gunn, F. Chapeau-Blondeau, M. D. McDonnell, B. R. Davis, A. Allison, and D. Abbott, “Too good to be true: When overwhelming evidence fails to convince,” *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 472, no. 2187, Mar. 2016, Art. no. 20150748.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. 34th Intl. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 1321–1330.

[24] J. Guo and J. Ye, “Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 118–125.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[27] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, *arXiv:1503.02531*.

[28] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, “Learning lightweight lane detection CNNs by self attention distillation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1013–1021.

[29] M. Hu and S. Chen, “One-pass incomplete multi-view clustering,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3838–3845.

[30] P. Hu, X. Peng, H. Zhu, J. Lin, L. Zhen, and D. Peng, “Joint versus independent multiview hashing for cross-view retrieval,” *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 4982–4993, Oct. 2021.

[31] X. Ji, A. Vedaldi, and J. Henriques, “Invariant information clustering for unsupervised image classification and segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9865–9874.

[32] R. Jonker and T. Volgenant, “Improving the Hungarian assignment algorithm,” *Oper. Res. Lett.*, vol. 5, no. 4, pp. 171–175, 1986.

[33] Z. Kang et al., “Partition level multiview subspace clustering,” *Neural Netw.*, vol. 122, pp. 279–288, Feb. 2020.

[34] J. Kim, S. Park, and N. Kwak, “Paraphrasing complex network: Network compression via factor transfer,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[36] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[37] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang, "Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, 2022, pp. 1306–1313.

[38] H. Li, Y. Li, M. Yang, P. Hu, D. Peng, and X. Peng, "Incomplete multi-view clustering via prototype-based imputation," 2023, *arXiv:2301.11045*.

[39] J. Li, Y. Pan, and I. W. Tsang, "Taming overconfident prediction on unlabeled data from hindsight," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 1–13, 2023.

[40] L. Li, Z. Wan, and H. He, "Incomplete multi-view clustering with joint partition and graph learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 589–602, Jan. 2023.

[41] S.-Y. Li, Y. Jiang, and Z.-H. Zhou, "Partial multi-view clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, 2014, pp. 1968–1974.

[42] S. Li, F. Liu, L. Jiao, P. Chen, and L. Li, "Self-supervised self-organizing clustering network: A novel unsupervised representation learning method," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1–15, 2022.

[43] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 10, pp. 8547–8555.

[44] Z. Li, Q. Wang, Z. Tao, Q. Gao, and Z. Yang, "Deep adversarial multi-view clustering network," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2952–2958.

[45] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[47] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "COMPLETER: Incomplete multi-view clustering via contrastive prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2021, pp. 11174–11183.

[48] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, "Dual contrastive prediction for incomplete multi-view representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4447–4461, Apr. 2023.

[49] J. Liu et al., "A novel consensus learning approach to incomplete multi-view clustering," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107890.

[50] X. Liu, "Hyperparameter-free localized simple multiple kernel K-means with global optimum," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 1–11, 2022.

[51] X. Liu et al., "Efficient and effective regularized incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2634–2646, Feb. 2020.

[52] C. D. Manning, *Introduction to Information Retrieval*. Rockland, MA, USA: Syngress Publishing, 2008.

[53] H. Mobahi, M. Farajtabar, and P. L. Bartlett, "Self-distillation amplifies regularization in Hilbert space," in *Proc. Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 3351–3361.

[54] F. Ntelemis, Y. Jin, and S. A. Thomas, "Information maximization clustering via multi-view self-labelling," *Knowl.-Based Syst.*, vol. 250, Aug. 2022, Art. no. 109042.

[55] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[56] S. Park et al., "Improving unsupervised image clustering with robust learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12278–12287.

[57] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "COMIC: Multi-view clustering without parameter selection," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5092–5101.

[58] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[59] C. Qin, H. You, L. Wang, C.-C. J. Kuo, and Y. Fu, "PointDAN: A multi-scale 3D domain adaption network for point cloud representation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[60] N. Rai, S. Negi, S. Chaudhury, and O. Deshmukh, "Partial multi-view clustering using graph regularized NMF," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2192–2197.

[61] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3210–3221, Jul. 2018.

[62] M. Steinbach, L. Ertöz, and V. Kumar, "The challenges of clustering high dimensional data," in *New Directions in Statistical Physics*. Berlin, Germany: Springer, 2004, pp. 273–309.

[63] C. Tang, Z. Li, J. Wang, X. Liu, W. Zhang, and E. Zhu, "Unified one-step multi-view spectral clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 6449–6460, Jun. 2023.

[64] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[65] D. J. Trosten, S. Løkse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1255–1265.

[66] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[67] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. V. Gool, "SCAN: Learning to classify images without labels," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 268–285.

[68] Ulrike Von Luxburg et al., "Clustering stability: An overview," *Found. Trends Mach. Learn.*, vol. 2, no. 3, pp. 235–274, 2010.

[69] X. Wan et al., "One-step multi-view clustering with diverse representation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 1, 2024, doi: 10.1109/TNNLS.2024.3378194.

[70] J. Wang et al., "Fast approximated multiple kernel K-means," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 6171–6180, Nov. 2024.

[71] S. Wang et al., "Multi-view clustering via late fusion alignment maximization," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 3778–3784.

[72] Y. Wang and L. Wu, "Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering," *Neural Netw.*, vol. 103, pp. 1–8, Jul. 2018.

[73] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, Oct. 2018.

[74] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Oct. 2019.

[75] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and G.-S. Xie, "CDIMC-Net: Cognitive deep incomplete multi-view clustering network," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3230–3236.

[76] J. Wen et al., "Unified tensor framework for incomplete multi-view clustering and missing-view inferring," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10273–10281.

[77] Y. Wen et al., "Unpaired multi-view graph clustering with cross-view structure matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 1–15, 2024.

[78] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[79] C. Xu, Z. Guan, W. Zhao, H. Wu, Y. Niu, and B. Ling, "Adversarial incomplete multi-view clustering," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3933–3939.

[80] C. Xu, D. Tao, and C. Xu, "Multi-view self-paced learning for clustering," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3974–3980.

[81] J. Xu, Y. Ren, G. Li, L. Pan, C. Zhu, and Z. Xu, "Deep embedded multi-view clustering with collaborative training," *Inf. Sci.*, vol. 573, pp. 279–290, Sep. 2021.

[82] J. Xu et al., "Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9234–9243.

[83] J. Xu et al., "Deep incomplete multi-view clustering via mining cluster complementarity," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 8761–8769.

[84] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He, "Multi-level feature learning for contrastive multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 16051–16060.

[85] W. Yan, Y. Zhang, C. Tang, W. Zhou, and W. Lin, "Anchor-sharing and clusterwise contrastive network for multiview representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 9, 202, doi: 10.1109/TNNLS.2024.3357087.

[86] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2859–2868.

[87] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2021, pp. 1134–1143.

[88] M. Yang, Y. Li, P. Hu, J. Bai, J. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1055–1069, Jan. 2023.

[89] M. Yin, W. Huang, and J. Gao, "Shared generative latent representation learning for multi-view clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 6688–6695.

[90] S. Yu et al., "Sparse low-rank multi-view subspace clustering with consensus anchors and unified bipartite graph," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 22, 2023, doi: 10.1109/TNNLS.2023.3332335.

[91] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3903–3911.

[92] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Sep. 2020, pp. 13876–13885.

[93] P. Zeng, M. Yang, Y. Lu, C. Zhang, P. Hu, and X. Peng, "Semantic invariant multi-view clustering with fully incomplete information," 2023, *arXiv:2305.12743*.

[94] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2402–2415, May 2022.

[95] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3713–3722.

[96] R. Zhou and Y.-D. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14619–14628.

Jiatai Wang (Student Member, IEEE) received the M.S. degree from Inner Mongolia University of Technology, Hohhot, China, in 2024. He is currently pursuing the Ph.D. degree with the College of Computer Science, Nankai University, Tianjin, China.

He has authored several articles in high-impact journals in the computer vision field, such as Institute of Engineering and Technology (IET) computer vision. His interests are focused on unsupervised learning in the computer vision (CV) field.

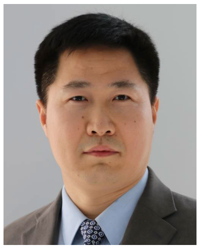
Zhiwei Xu (Member, IEEE) received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2002, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2018.

From 2020 to 2021, he held a visiting postdoctoral position at the Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY, USA. He is currently a Professor with the Haihe Laboratory of Information Technology Application Innovation, Tianjin, China, while working as an Adjunct Professor with the Institute of Computing, Chinese Academy of Sciences. His research interests include in-network data compact representation, learning, and related security and privacy problems.

Xin Wang (Senior Member, IEEE) received the B.S. degree in telecommunications engineering and the M.S. degree in wireless communications engineering from Beijing University of Posts and Telecommunications, Beijing, China, respectively, and the Ph.D. degree in electrical and computer engineering from Columbia University, New York, NY, USA, in 2001.

She was a Member of Technical Staff of mobile and wireless networking with the Bell Labs Research, Lucent Technologies, Murray Hill, NJ, USA, and an Assistant Professor with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, USA. She is currently an Associate Professor with the Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY, USA. Her research interests include algorithm and protocol design in wireless networks and communications, mobile and distributed computing, as well as big data analysis and machine learning.

Dr. Wang achieved the NSF Career Award in 2005 and ONR Challenge Award in 2010. She has served on the executive committee and technical committee of numerous conferences and funding review panels and serves as an Associate Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING.



Tao Li (Member, IEEE) received the Ph.D. degree in computer science from Nankai University, Tianjin, China, in 2007.

He currently works as a Professor with the College of Computer Science, Nankai University. His main research interests include heterogeneous computing, machine learning, and the Internet of Things.

Dr. Li is a member of ACM and a Distinguished Member of CCF.