Hierarchical Mutual Information Analysis: Towards Multi-View Clustering in the Wild

Jiatai Wang*†, Zhiwei Xu*‡, Xuewen Yang§, Xin Wang¶, and Li Tao*
*Haihe Lab of ITAI, Tianjin, China
†College of Data Science and Application, Inner Mongolia University of Technology, Huhhot, China
‡Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
§InnoPeak Technology, Inc, Palo Alto, California, USA
¶Department of Electrical and Computer Engineering, Stony Brook University, New York, U.S.A
wangjiatai@hl-it.cn, xuzhiwei2001@ict.ac.cn, xuewen.yang@protonmail.com, x.wang@stonybrook.edu, cslitao@126.com

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10651355

***Abstract*—Multi-view clustering (MVC) can explore common semantics from multiple views and has been extensively used to support management with unsupervised training data. However, the issue of spatio-temporal asynchronism often leads to multi-view data being missing or unaligned in the real world. This limit poses significant challenges in learning consistent representations. This paper proposes a deep MVC framework where data recovery and alignment are fused hierarchically from an information-theoretic perspective, maximizing the mutual information among different views and ensuring the consistency of their latent spaces. To address the issue of missing views, we use dual prediction for instance-level alignment. While leveraging contrastive reconstruction enhances the mutual information of features within the same class for class-level alignment. This could be the first attempt to view recovery and alignment can be solved simultaneously in a unified theoretical framework. Extensive experiments show that our method outperforms baseline methods even in the cases of missing and unaligned views.**

***Index Terms*—Multi-view clustering, Missing and unaligned views, Mutual information**

I. Introduction

The quality of training data is critical for multi-modal learning, particularly in practical applications of large multi-modal models [1] where massive data are frequently gathered from different sources or multiple views. In this case, semantics serves as meaningful, shared information extracted from different views, even if the data representation (such as pixels in images and words in text) may vary across views. As an important unsupervised technology for multi-modal data management, multi-view clustering (MVC) [2] aims to mine common semantics to improve learning efficiency. Despite potential differences in multi-view data presentation, significant progress has been made. The success of all existing work [3]–[5] is supported under the assumptions of the completeness of data and the consistency of different views strictly. However, these two assumptions (see Fig 1(a)) would inevitably be violated in the real world. Firstly, a Partially View-unaligned Problem (PVP) always exists due to the asynchronism of data transmission. Even worse, PVP even coexists with the Partially Sample-missing Problem (PSP) (see Fig 1(b)), while the data transmission of one view fails. It is a practical need and a challenge to learn common semantics and alleviate the impacts of PVP and PSP to ensure learning consistency.

Previous MVC methods can be roughly classified into three subgroups, i.e., i) MVC methods for complete data [3]–[5] strive to learn discriminative representations by utilizing the consistency and complementary information from different views; ii) PVP-oriented MVC methods [6]–[8], which build cross-view mappings at the instance level in an unsupervised manner; and iii) PSP-oriented MVC methods [9]–[13] which utilize existing views by way of mathematical derivation or model prediction to recover lost views. Although existing MVC methods have achieved important progress [6], [7], [9]–[11], [14] in solving PSP or PVP, few solve PSP and PVP jointly. Yang et al. found that the correspondence of negative pairs in contrastive learning might be false, i.e., false negatives, and propose SURE [15] to handle that. SURE imputes the missing sample by the weighted sum of its peers in the same view. It re-aligns samples through contrastive learning, establishing class-level correspondences to unify the handling of PVP and PSP. However, two challenges remain: i) the inability to solve PVP and PSP independently and simultaneously, resulting in the need for alignment to rely heavily on fill performance; and ii) its class-level imputing and alignment can not provide detailed information for each instance in the dataset, and degrade its clustering accuracy. From information theory, since missing and unaligned data exist in different hierarchies, the proposed model must also be designed using different hierarchies of mutual information. Hierarchies are different types of data representation and processing methods are conducted, allowing the model to capture and integrate information on multiple scales. Mutual information and the conditional entropy between different views measure multi-view data's consistent and inconsistent semantics, respectively [10]. Therefore, we aim to base both PVP and PSP respective solutions on mutual information enhancement algorithms to enhance each other in a unified mutual information framework.

This work was supported by the National Science Foundation of China (61962045, 62272248, 62062055, 61902382, and 61972381); the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region NJYT23104; the Open Project Fund of State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences (CARCHA202108) and the Natural Science Foundation of Tianjin of China (21JCZDJC00740, 21JCYBJC00760, 23JCQNJC00010). (Corresponding Author: Zhiwei Xu.)

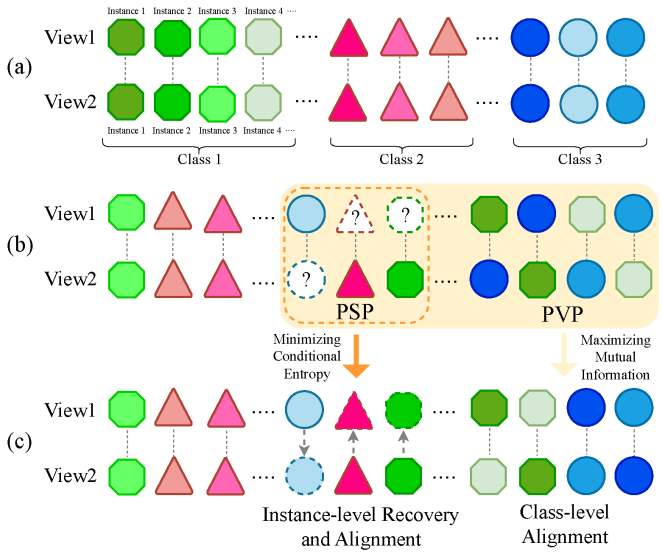


Fig. 1. Illustrative examples of the PVP and PSP. Taking bi-view data as a showcase, we use two rows of polygons to denote two views, where each column of polygons represents a pair of instances that may be incomplete or unaligned. Polygons with the same shape belong to one class (or "category"), and the same color is a pair of aligned instances. The "?" denotes that the view sample is missing. (a) Multi-view data: Ideally, the data has no missing or unaligned samples. (b) Incomplete data: Due to the complexity of data collection and transmission in practice, there are missing and unaligned view samples in the raw data, leading to PVP and PSP. (c) Instance-level recovery and alignment: recovers the missing samples of the corresponding views based on the existing view samples in the same instance. Class-level alignment: Minimize the distance between samples of the same class while maximizing the distance between samples of different classes.

We propose a novel incomplete MVC framework, multi-view clustering via maximizing hierarchical mutual information (HmiMVC), to conquer the challenges of PVP and PSP caused by the lack of correspondence between views. HmiMVC projects a raw dataset into a hierarchical latent space wherein information consistency is guaranteed. As shown in Fig. 1(c), we solve PVP by maximizing the mutual information between already aligned views through contrastive learning to achieve class-level alignment. To solve both PSP and PVP, we introduce dual prediction to predict the missing data while minimizing the conditional entropy, constituting a natural instance-level alignment. Finally, we merge two levels of alignment strategies into a unified reconstruction process in a hierarchically consistent way to avoid model collapse. The main contributions of this paper are:

- We address PVP and PSP simultaneously in a mutual information framework, enabling data recovery and alignment to be mutually reinforcing by novelly parallelizing class-level and instance-level strategies.
- From an information-theoretic insight, the proposed HmiMVC method has a novel loss function that achieves information consistency and data restorability using a contrastive loss and a dual prediction loss.
- Extensive experiments demonstrate that HmiMVC boosts mutual information and achieves state-of-the-art clustering effectiveness.

Ensuring the completeness and alignment of multi-view data is imperative for industrial processes' safe and stable operation. By re-aligning and filling samples, HmiMVC can further be applied to other applications, including but not limited to cross-modal retrieval, autonomous driving, etc.

II. RELATED WORK

A. Multi-view Clustering.

The assumption of data completeness is the foundation of almost all MVC techniques. However, view correspondence and instance completeness are lost upon violation of the assumption, resulting in PVP and PSP. Researchers have achieved substantial progress in addressing PSP using various methods [9]–[13]. PVP remains a relatively unexplored issue despite these advancements, as highlighted in recent studies [7]. However, PVP can only find the optimal alignment path using the traditional Hungarian algorithm [16]. Subsequently, PVC [7] has implemented differentiable Hungarian algorithms, and MvCLN [6] has achieved the class-level alignment using robust contrastive loss. Although SURE [15] has extended the filling of missing views based on MvCLN, class-level imputation does not provide detailed information about the data samples, and the quality of the alignment directly relies on the performance of the imputing, which increases the risk to use the features learned on the mistaken samples. In addition, Wen et al. [8] used the structural information of each view to refine the alignment relations, thus alleviating the need for MVC for the paired samples. Similarly, Zeng et al. [17] discover the existence of invariant semantic distributions across views, and design a training process without paired samples. However, this type of approach is sensitive to the features and distribution of the data, especially when there are complex nonlinear relationships between different views.

Considering the limitations of the existing work, our HmiMVC implements, for the first time, independent and simultaneous data recovery and alignment and supports this process according to hierarchical mutual information. On the other hand, HmiMVC blends class-level and instance-level strategies to capture details and features from the data, thus improving the model's ability to learn more complex patterns and discriminative information. Since the priori information of sample pairs is provided, our model is easier to converge, especially in the case of noisy data or large cross-view discrepancy.

B. Contrastive Learning.

Contrastive learning [18]–[24] is an essential method for unsupervised learning [25], [26] and requires the pre-definition of positive and negative samples. Its primary objective is to maximize the similarity in feature space among positive samples while increasing the distance between negative ones. This approach enhances the model's ability to recognize similar samples and facilitates more accurate classification. For example, SimClR [18] or MoCo [19] minimize the InfoNCE loss

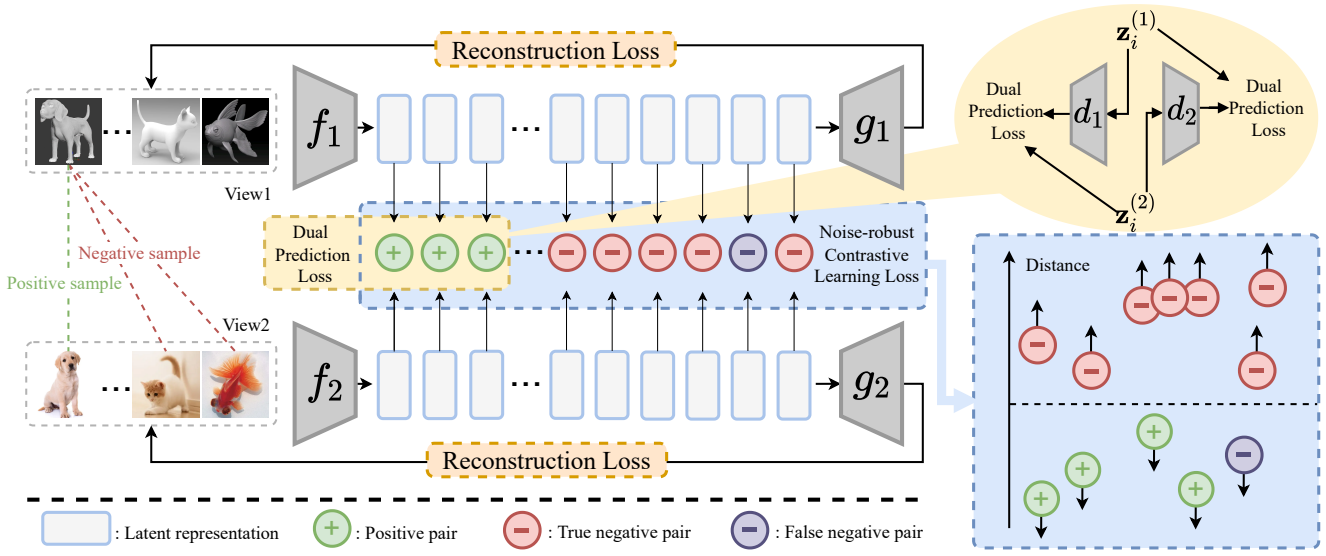


Fig. 2. Overview of HmiMVC. Bi-view data is used as a showcase in this figure. Our method contains three joint learning objectives, i.e., noise-robust contrastive learning, dual prediction, and reconstruction. Specifically, the noise-robust contrastive learning objective learns class-level alignment relations from aligned positive and unaligned negative samples. Dual prediction allows for constructing instance-level alignment and recovering missing views from one of its existing views. The goal of reconstruction loss is to maintain the diversity of views and project all views into view-specific spaces.

function [27] to maximize the lower bound of mutual information. Since the processing of negative samples is cumbersome, BYOL [20], SimSiam [21], and DINO [22] have successfully transformed the contrastive task into a prediction task without defining negative samples and achieved amazing results. Therefore, no matter what kind of contrastive learning method, its essence is to maximize the mutual information between positive samples. Although existing studies [3], [19], [28] have shown that consistency could be learned by maximizing the mutual information of different views, they ignore it at different hierarchies. Lin et al. demonstrated that inconsistency in learning can be defined in terms of conditional entropy. Consequently, our strategy of learning mutual information from other hierarchies can be seen as an effective means to alleviate inconsistent learning [9], [10].

In contrast, HmiMVC uses robust contrastive learning to reduce the impact of false-positive samples to solve PVP and minimizes the conditional entropy to cope with PSP. Additionally, our method is specifically designed for handling missing and unaligned data, whereas the existing contrastive learning works ignore this practical problem.

III. Method

In this section, we propose a new deep multi-view clustering method, HmiMVC, to learn the representation of incomplete and unaligned multi-view samples in different hierarchies. For clarity, we will first introduce the proposed loss function and then elaborate on each objective.

A. Notations and Motivation

A multi-view dataset $\overline{\mathbf{X}}_N = \left\{\mathbf{X}_{N_x}^{(v)}, \mathbf{S}_{N_s}^{(v)}, \mathbf{W}_{N_w}^{(v)}\right\}_{v=1}^{V}$ includes N samples across V views, where $v \in [1, V]$ denotes the view index. $\left\{\mathbf{X}_{N_x}^{(v)}\right\}_{v=1}^{V} = \left\{\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \ldots, \mathbf{x}_{N_x}^{(v)}\right\}_{v=1}^{V}$ denotes the complete alignment data used for training, where N_x is the number of complete and aligned instances. $\left\{\mathbf{S}_{N_s}^{(v)}\right\}_{v=1}^{V}$/$\left\{\mathbf{W}_{N_w}^{(v)}\right\}_{v=1}^{V}$ denotes the data with PVP/PSP, where N_s is the number of unaligned instances and N_w is the the number of missing instances($N = N_x + N_s + N_w$). So the missing rate of data $\alpha = \frac{N_w}{N}$, the unaligned rate $\beta = \frac{N_s}{N}$, and the proportion of complete and aligned data is $\gamma = \frac{N_x}{N}$.

With the above definitions, we assume that the data used for clustering has both missing and unaligned cases ($N_s > 0, N_s > 0$), which are fully utilized to build the model by $\mathbf{X}_{N_x}$, so that the model fills up the missing parts of $\mathbf{S}_{N_s}$ and realigns $\mathbf{W}_{N_w}$ during clustering process. To accomplish this, existing MVC methods project the original features into the feature space, fusing the features of all views $\left\{\mathbf{z}_N^{(v)}\right\}_{v=1}^{V}$ to obtain a common representation of all views. Clustering the fused features directly transforms the multi-view clustering task into a single-view clustering task. One of the persisting challenges is that some methods prioritize data filling over alignment, resulting in outcomes that are heavily influenced by the chosen filling strategy. Therefore, we hope to design an algorithm capable of handling the missing and unaligned problems simultaneously and in parallel at different hierarchies. As shown in Fig. 2, we propose that HmiMVC implements class-level alignment and instance-level filling at different hierarchies, respectively, and it consists of three learning objectives:

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{\text{pre}} + \mathcal{L}_{\text{rec}}, \tag{1}$$

where $\mathcal{L}_{cl}$, $\mathcal{L}_{pre}$, and $\mathcal{L}_{rec}$ are noise-robust contrastive loss,

dual prediction loss, and view reconstruction loss, respectively.

B. Class-level Alignment.

At this hierarchy, we maximize mutual information across views to ensure consistency learning and thus understand the alignment relation [7]. The most intuitive idea is to realize this process through contrastive learning, positive pairs are defined as samples that are aligned across different views but belong to the same class, whereas negative pairs may include samples from different classes or unaligned samples within the same view. Through this mechanism, the model can discover and reinforce the consistency of category information across different views. As shown in Fig. 2, we use $\left(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}\right)$ as positive pairs ($i < N_x$), and stochastically select cross-view samples to form negative pairs $\left(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)}\right)$. Considering the impact of false-negative samples [15], mathematically,

$$\mathcal{L}_{cl} = \frac{1}{2N_c} \sum_{i=1}^{N_c} \left(Y\mathcal{L}_i^{pos} + (1-Y)\mathcal{L}_i^{neg}\right), \tag{2}$$

where N_c represents the total number of sample pairs, and $Y = 1/0$ for positive/negative pairs. Then the feature contrastive loss between $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_i^{(2)}$ for the i-th pair of positive samples is formulated as:

$$\mathcal{L}_i^{\text{pos}} = \left\| f_1\left(\mathbf{x}_i^{(1)}\right) - f_2\left(\mathbf{x}_i^{(2)}\right) \right\|_2^2 = \left\| \mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)} \right\|_2^2, \tag{3}$$

where f_v and $\mathbf{z}_i^{(v)}$ denote the encoder and the latent representation of $\mathbf{x}_i^{(v)}$, respectively. We aim to maximize $\left(\mathbf{z}_i^{(1)}, \mathbf{z}_j^{(2)}\right)$'s distance in a latent space by minimizing

$$\mathcal{L}_i^{neg} = \frac{1}{\tau} \max\left(\tau \left\| \mathbf{z}_i^{(1)} - \mathbf{z}_j^{(2)} \right\|_2^{\frac{1}{2}} - \left\| \mathbf{z}_i^{(1)} - \mathbf{z}_j^{(2)} \right\|_2^{\frac{3}{2}}, 0\right)^2, \tag{4}$$

where τ is the temperature parameter computed only once at the initial state with $\tau = \frac{1}{N_{pos}} \sum d\left(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}\right) + \frac{1}{N_{neg}} \sum d\left(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)}\right)$, N_{pos} and N_{neg} denote the number of positive and negative pairs ($N_c = N_{pos} + N_{neg}$), respectively. In the inference phase we have $\sum_{v_1}^{V} \sum_{v_2 \neq v_1}^{V} C\left(\mathbf{s}_i^{(v1)}, \mathbf{s}_j^{(v2)}\right) = V(V-1)$, realigning $\mathbf{s}_i^{(v1)}$ and $\mathbf{s}_j^{(v2)}$ by computing the Euclidean distance $C(.)$.

C. Instance-level Recovery and Alignment.

To address the PSP in PVP, we mitigate the inconsistency between views by minimizing the conditional entropy at another hierarchy of the feature space [10]. Specifically, we achieve the above goal by training dual prediction networks so that views can predict each other. In this way, we can predict missing views from existing views. This module fills in the missing data and achieves the natural instance-level alignment. The metric of Normalized Mutual Information (NMI) can describe this process. According to its definition, $\text{NMI}(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}) = \frac{H(\mathbf{z}_i^{(1)}) - H\left(\mathbf{z}_i^{(1)} | \mathbf{z}_i^{(2)}\right)}{H(\mathbf{z}_i^{(1)}) + H(\mathbf{z}_i^{(2)})}$, minimizing the conditional entropy $H\left(\mathbf{z}_i^{(1)} \mid \mathbf{z}_i^{(2)}\right)$ maximizes NMI. According to the variational inference [9], we introduce a network $d_{(v)}$ that minimizes the conditional entropy approximately by minimizing

$$H\left(\mathbf{z}^{(1)} \mid \mathbf{z}^{(2)}\right) = -\mathbb{E}_{\mathcal{P}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}}\left[\log \mathcal{P}\left(\mathbf{z}^{(1)} \mid \mathbf{z}^{(2)}\right)\right]. \tag{5}$$

Maximizing $\mathbb{E}_{\mathcal{P}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}}\left[\log \mathcal{Q}\left(\mathbf{z}^{(1)} \mid \mathbf{z}^{(2)}\right)\right]$ by neglecting the constants derived from the Gaussian distribution is equivalent to minimize

$$\mathbb{E}_{\mathcal{P}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}} \left\| \mathbf{z}_i^{(1)} - d^{(2)}\left(\mathbf{z}_i^{(2)}\right) \right\|_2^2, \tag{6}$$

where $d^{(2)}$ could be a parameterized model which maps $\mathbf{z}^{(2)}$ to $\mathbf{z}^{(1)}$, as shown in Fig. 2. Further, we have

$$\mathcal{L}_{\text{pre}} = \left\| d^{(1)}\left(\mathbf{z}_i^{(1)}\right) - \mathbf{z}_i^{(2)} \right\|_2^2 + \left\| d^{(2)}\left(\mathbf{z}_i^{(2)}\right) - \mathbf{z}_i^{(1)} \right\|_2^2. \tag{7}$$

After the model converges, we predict the missing view and form a natural instance-level alignment, *i.e.*, $\mathbf{w}_i^{(1)} = d_2\left(f_{(2)}\left(\mathbf{w}_i^{(2)}\right)\right)$, where $\mathbf{w}_i^{(1)}$ is the missing sample predicted to be recovered by the representation of $\mathbf{w}_i^{(2)}$. So $\mathbf{w}_i^{(1)}$ and $\mathbf{w}_i^{(2)}$ are aligned on the instance-level.

D. Reconstruction for Hierarchical Consistencies

For each view, we feed it into an autoencoder for learning the latent representation $\mathbf{z}^{(v)}$ by minimizing

$$\mathcal{L}_{rec} = \frac{1}{2N_x} \sum_{i=1}^{N} \sum_{v=1}^{2} \left\| \mathbf{x}_i^{(v)} - g_v\left(\left[\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}\right]\right) \right\|_2^2, \tag{8}$$

where g_v denotes the decoder for the v-th view. As a result, the conflict between the reconstruction objective and two consistency objectives is alleviated, and trivial solutions are avoided.

IV. EXPERIMENTS

This section evaluates the proposed HmiMVC on four widely-used multi-view datasets and compares it with three state-of-the-art clustering methods.

TABLE I
DATASET SUMMARY

Datasets	Size	# of categories	Dimension
Scene-15	4485	15	20/59
Deep Animal	10158	50	4096/4096
MNIST-USPS	5000	10	784/784
Caltech101-20	2386	20	1984/512

A. Experiment Setup

Datasets. Four widely-used datasets are used in our experiments as shown in Table I. 1) Scene-15 [29] consists of 4,485 images distributed over 15 Scene categories, and we use two views of PHOG [30] and GIST [31] features, 20D and

59D feature vectors, respectively. 2) Deep Animal consists of 10,158 images from 50 classes and includes two types of 4096-dim features of [32] extracted by DECAF [33] and VGG19 [34] respectively as two views. 3) MNIST-USPS [35] is a popular handwritten digit dataset containing 5,000 samples with two different styles of digital images. 4) Caltech101 [36] consists of 9,144 images following [9] with the views of HOG [30] and GIST features.

Implementation. All our datasets are reshaped into vectors. We set the dimension of the autoencoder and prediction model to dim-1024-1024-1024-10-1024-1024-1024-dim, where dim is the dimension of the input data and a batch normalization layer and a ReLU layer follows each layer. Autoencoders for all views in HmiMVC are implemented using fully connected neural networks with similar architecture. MLPs are used to implement instance-level recovery and alignment, and each MLP has three linear layers with ReLU activation functions added in the middle of each layer. Based on extensive ablation studies, the batch size is 1024 and the training epochs are 150, respectively. We utilize Adam optimizer [37] with default parameters and a learning rate 0.0001.

To further verify the generalization of HmiMVC among different datasets, we conduct experiments in complex scenarios where both PVP and PSP coexist. We randomly select N_x, N_w, and N_s instances as training data, incomplete data, and unaligned data, respectively. The ratio of unaligned data β and missing data α is both set to 0.25 to simulate PVP and PSP.

Comparison methods. We chose three classical and state-of-the-art methods to compare with HmiMVC, including robuSt mUlti-view clusteRing with incomplEte information (SURE) [15], inCOMPlete muLti-view clustEring via conTrastivE pRediction (COMPLETER) [9] for PSP-only, and Multi-view Contrastive Learning with Noise-robust loss (MvCLN) [6] for PVP-only. Since few methods in the community work in our experimental settings, we had to add constraints to the selected methods to ensure the fairness of the comparison. Due to COMPLETER not handling unaligned data, we establish the alignment relationship directly with the Hungarian algorithm. The alignment modules of SURE and MvCLN must rely on the complete data that has been filled, so 50% of the data is filled before alignment. MvCLN cannot handle missing views, so we compute the mean of the same view to fill in the missing samples.

Evaluation metrics. The clustering effectiveness is evaluated by three metrics, $i.e.$, Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI).

B. Convergence Analysis

We investigate the convergence of HmiMVC by reporting the loss value and the corresponding clustering performance with increasing epochs. As shown in Fig. 3, one could observe that the loss remarkably decreases in the first 20 epochs, and various evaluation metrics continuously increase and tend to be smooth and consistent. Furthermore, we show t-sne [38] visualizations of the obtained representations on four datasets as shown in Fig. 4. The missing rate α and unaligned rate β=0.25 are also fixed to 0.25 in the experiments. As the epoch number increases, the common representations learned by HmiMVC become more compact and independent, and the clustering density is higher. Even in the case of extremely chaotic data, the boundary relationship between categories can be found, and clustering can be completed.

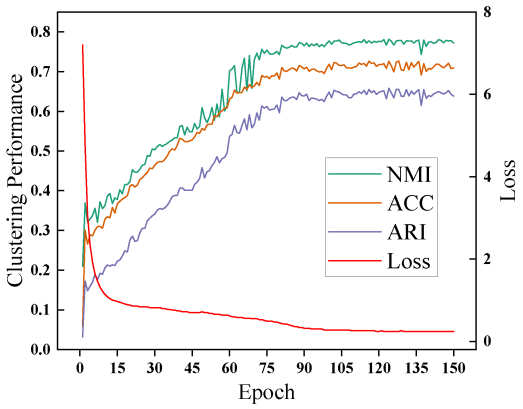


Fig. 3. Convergence analysis of clustering performance and loss values.

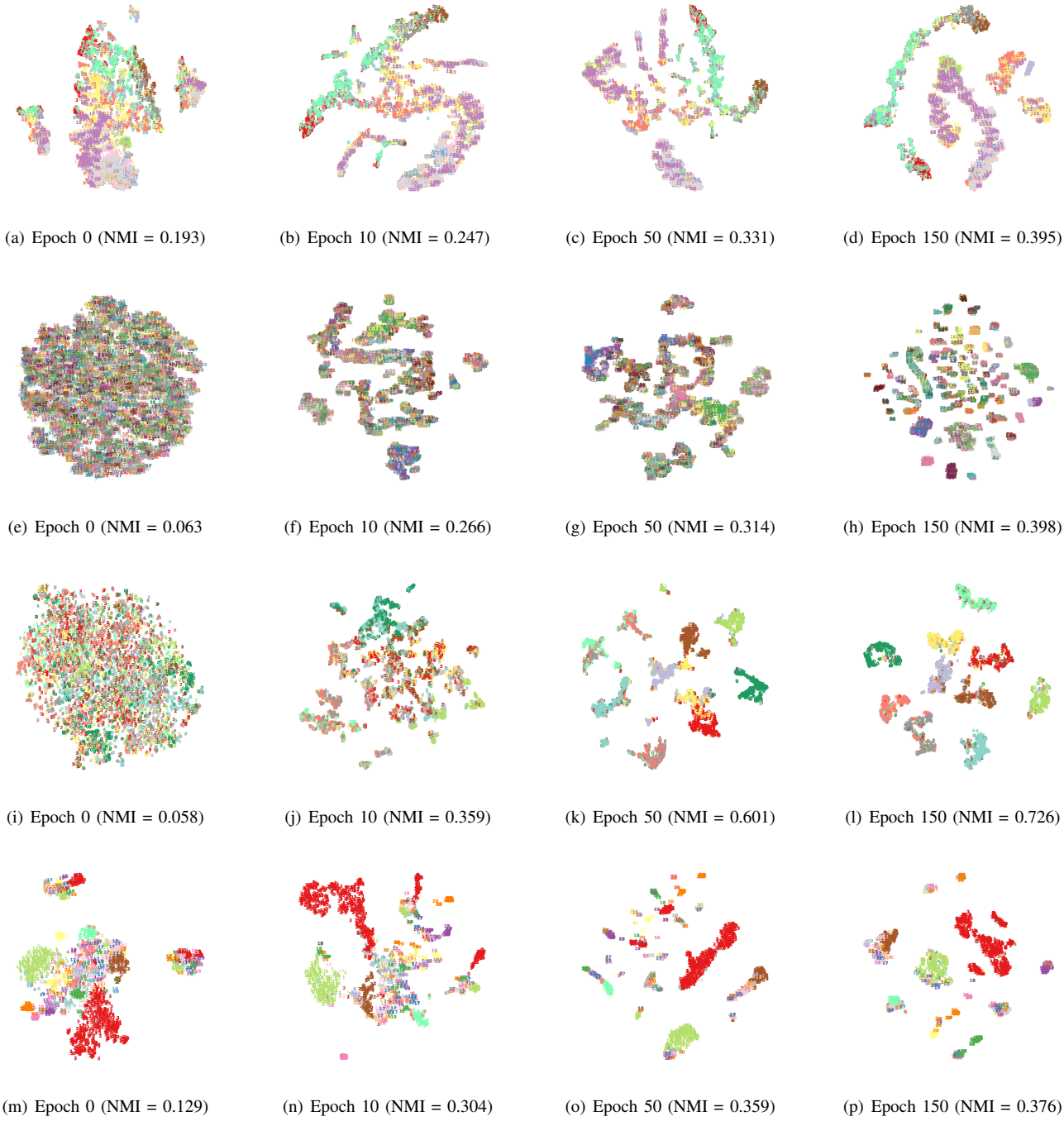
C. Comparisons with State of the Arts

We evaluate all methods in Sec IV-A, where missing rate α=0.25 (denoted by Incomplete) and β=0.25 (denoted by Unaligned). The clustering performance of all methods on four datasets is depicted in Table II, from which we obtain the following observations: (1) HmiMVC is robust on different datasets. (2) Our HmiMVC obtains the best performance on all datasets. Compared with the second-best methods, HmiMVC has considerable improvements, its NMI completely outperforms all baselines, especially on dataset Scene-15, Deep Animal, and Caltech101. (3) The improvements obtained by the previous SOTA method (i.e., SURE) are limited.

The reasons for the above observations can be explained as follows: (1) The design of hierarchical consistency makes the model more adaptable to different hierarchies of data variations and noise, improving generalization over different datasets. (2) HmiMVC uses hierarchic mutual information as an optimization objective to capture the correlation between different views, which improves the clustering performance. (3) Compared to SURE, our HmiMVC reduces the dependence of view alignment on view imputation, thus mitigating the negative risk of inaccurately filled views.

D. Visualization Verification on the Validity of HmiMVC

To further verify the necessity and validity of the HmiMVC design, we visualized the clustering results on MNIST-USPS dataset to test the performance of all baselines under the same experimental setup as Sec IV-C. Fig. 5 shows the results on MNIST-USPS with different methods, from which we could have the following observations: COMPLETER does not form boundary-separated clusters due to the interference of unaligned views on the original algorithm. And MvCLN



(a) Epoch 0 (NMI = 0.193) (b) Epoch 10 (NMI = 0.247) (c) Epoch 50 (NMI = 0.331) (d) Epoch 150 (NMI = 0.395)

(e) Epoch 0 (NMI = 0.063 (f) Epoch 10 (NMI = 0.266) (g) Epoch 50 (NMI = 0.314) (h) Epoch 150 (NMI = 0.398)

(i) Epoch 0 (NMI = 0.058) (j) Epoch 10 (NMI = 0.359) (k) Epoch 50 (NMI = 0.601) (l) Epoch 150 (NMI = 0.726)

(m) Epoch 0 (NMI = 0.129) (n) Epoch 10 (NMI = 0.304) (o) Epoch 50 (NMI = 0.359) (p) Epoch 150 (NMI = 0.376)

Fig. 4. Visualization of common representations during training. In detail, the figure shows the visualization results of the representations of Scene-15 (a-d), Deep Animal (e-h), MNIST-USPS (i-l), and Caltech101 (m-p) datasets when the epochs are 0, 10, 50, and 150, respectively.

and SURE's formed an incorrect number of clusters, which contained many erroneous samples. Since NMI measures the similarity between the distribution of clustering results and the distribution of real classes, our HmiMVC fully utilizes the mutual information of different hierarchies to mine the consistent distribution of views. HmiMVC has a clearer clustering structure than other methods, and the number of clusters equals the number of real labeled classes.

Additionally, we visualized the recovered images of SURE vs. HmiMVC and found that the noise is depleted (see Fig. 6). The potential propagation of noise or errors introduced during the imputation process to the alignment process and information distortion may occur during alignment. The instance-level imputation of SURE is more helpful in accurately predicting the missing views; the noise introduced during the imputation process is reduced. In contrast, HmiMVC achieves algorith-

TABLE II
THE PERFORMANCE COMPARISON ON MULTIPLE-VIEW DATASETS. THE 1st BEST RESULTS ARE INDICATED IN RED AND *italic*.

Incomplete Type	Datasets	Sence-15			Deep Animal			MNIST-USPS			Caltech101		
	Evaluation metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
missing rate 25% unaligned rate 25% 50% of complete data	COMPLETER [9](2021)	0.382	0.147	0.218	0.387	0.274	0.161	0.540	0.485	0.356	0.370	0.147	0.219
	MvCLN [6](2021)	0.355	0.385	0.193	0.366	0.250	0.150	0.698	0.810	0.590	0.325	0.175	0.169
	SURE [15](2022)	0.319	0.392	0.195	0.339	0.251	0.159	0.653	*0.830*	0.661	0.358	0.257	0.234
	HmiMVC (Ours)	*0.395*	*0.398*	*0.221*	*0.398*	*0.286*	*0.198*	*0.726*	0.782	*0.678*	*0.376*	*0.258*	*0.249*

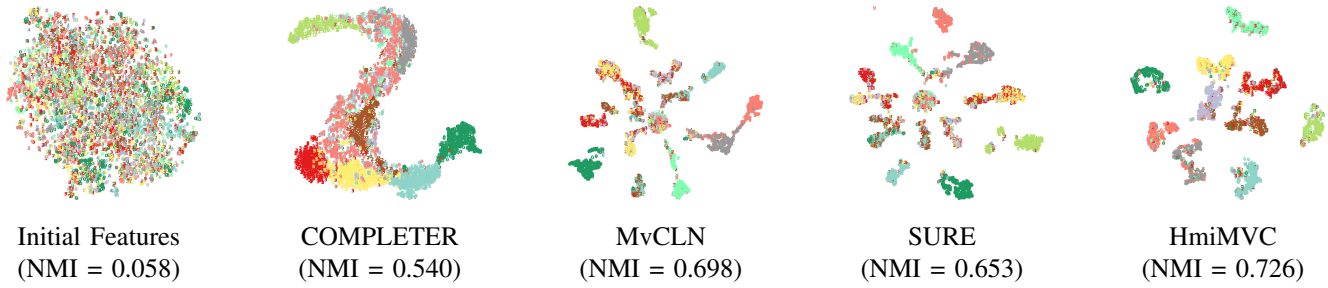


Fig. 5. T-sne [38] visualization on the MNIST-USPS dataset with all baselines.

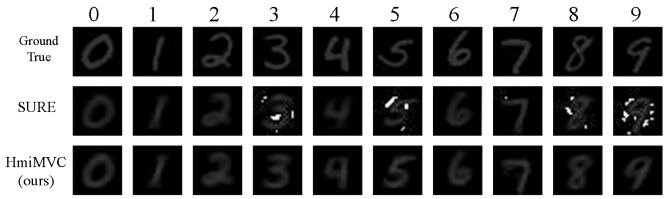


Fig. 6. Data recovery on noisy MNIST UPS datasets. Line 1 is the anchor view, while Lines 2 and 3 are the SURE and HmiMVC view recovery results, respectively.

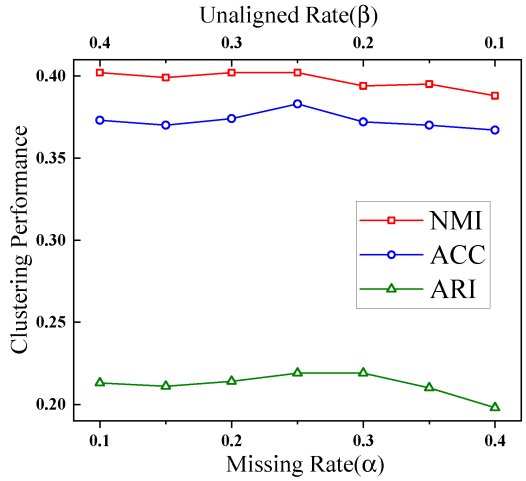


Fig. 7. Parameter analysis of performance comparisons with different missing rates (α) and unaligned rates (β).

mic decoupling and alleviates the dependence between the imputation process and the alignment process, the data from each view can better preserve their original features, thereby reducing the likelihood of noise propagation and accumulation.

E. Model Analysis

1) Parametric Analysis: As shown in Figure 7, with the completion rate γ being fixed at 0.5, our clustering performance is always relatively stable no matter how α and β vary ($\alpha + \beta = 0.5, \alpha > 0, \beta > 0$). This demonstrates that HmiMVC is robust because class-level alignment and instance-level recovery realize the public solution of PVP and PSP in a unified mutual information framework.

2) Ablation experiment: We conduct the ablation study on MNIST-USPS to demonstrate the importance of each component of our method. As shown in Table III, all losses play an integral role in HmiMVC. It should be pointed out that optimizing $\mathcal{L}_{cl}$ or $\mathcal{L}_{pre}$ separately may lead to meaningless solutions. To solve this problem, we must optimize both $\mathcal{L}_{cl}$ and $\mathcal{L}_{pre}$ during view refactoring to avoid trivial solutions.

TABLE III
ABLATION STUDY.

moudules of HmiMVC	NMI	ACC	ARI
(1)$\mathcal{L}_{pre}$	0.169	0.211	0.092
(2)$\mathcal{L}_{rec}$	0.329	0.457	0.242
(3)$\mathcal{L}_{cl}$	0.485	0.541	0.234
(4)$\mathcal{L}_{rec} + \mathcal{L}_{pre}$	0.486	0.523	0.347
(5)$\mathcal{L}_{pre} + \mathcal{L}_{cl}$	0.539	0.607	0.413
(6)$\mathcal{L}_{rec} + \mathcal{L}_{cl}$	0.693	0.760	0.675
(7)$\mathcal{L}_{rec} + \mathcal{L}_{pre} + \mathcal{L}_{cl}$	0.726	0.782	0.678

V. CONCLUSION

This paper proposes HmiMVC to provide a hierarchically consistent framework for handling PVP and PSP. HmiMVC achieves consistency in learning across views by maximizing

the hierarchical mutual information and minimizing the conditional entropy, bridging the gap between existing methods. To the best of our knowledge, this is the first work that combines class-level and instance-level alignment strategies and enables HmiMVC to achieve state-of-the-art performance in practice by handling PSP and PVP problems in parallel. We experimentally show that our loss could mitigate or eliminate the noise introduced during pairwise construction. This framework trains feature extractors and predictors, which can be used in feature compression, unsupervised labeling, and cross-modal feature retrieval. We thus suggest that people embed our model into the physical world to learn more consistent representation in broad scenarios and promote data-driven decision-making. In the future, we plan to extend this work to include more views.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] G. Chao, S. Sun, and J. Bi, "A survey on multi-view clustering," *arXiv preprint arXiv:1712.06246*, 2017.

[3] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He, "Multi-level feature learning for contrastive multi-view clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 051–16 060.

[4] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.

[5] X. Yang, J. Jiaqi, S. Wang, K. Liang, Y. Liu, Y. Wen, S. Liu, S. Zhou, X. Liu, and E. Zhu, "Dealmvc: Dual contrastive calibration for multi-view clustering," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 337–346.

[6] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1134–1143.

[7] Z. Huang, P. Hu, J. T. Zhou, J. Lv, and X. Peng, "Partially view-aligned clustering," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2892–2902, 2020.

[8] Y. Wen, S. Wang, Q. Liao, W. Liang, K. Liang, X. Wan, and X. Liu, "Unpaired multi-view graph clustering with cross-view structure matching," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[9] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "Completer: Incomplete multi-view clustering via contrastive prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 174–11 183.

[10] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, "Dual contrastive prediction for incomplete multi-view representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[11] J. Wang, Z. Xu, X. Yang, D. Guo, and L. Liu, "Self-supervised image clustering from multiple incomplete views via constrastive complementary generation," *IET Computer Vision*, vol. 17, no. 2, pp. 189–202, 2023.

[12] S. Deng, J. Wen, C. Liu, K. Yan, G. Xu, and Y. Xu, "Projective incomplete multi-view clustering," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[13] W. Lv, C. Zhang, H. Li, X. Jia, and C. Chen, "Joint projection learning and tensor decomposition-based incomplete multiview clustering," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[14] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and G.-S. Xie, "Cdimc-net: Cognitive deep incomplete multi-view clustering network," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3230–3236.

[15] M. Yang, Y. Li, P. Hu, J. Bai, J. C. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[16] R. Jonker and T. Volgenant, "Improving the hungarian assignment algorithm," *Operations Research Letters*, vol. 5, no. 4, pp. 171–175, 1986.

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

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[21] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[22] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[23] R. Jiang, P. Ishwar, and S. Aeron, "Hard negative sampling via regularized optimal transport for contrastive representation learning," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[24] R. Jiang, T. Nguyen, P. Ishwar, and S. Aeron, "Supervised contrastive learning with hard negative samples," *arXiv preprint arXiv:2209.00078*, 2022.

[25] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[26] Q. Yan, T. Hu, Y. Sun, H. Tang, Y. Zhu, W. Dong, L. Van Gool, and Y. Zhang, "Towards high-quality hdr deghosting with conditional diffusion models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[27] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[28] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9865–9874.

[29] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 524–531.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[31] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145–175, 2001.

[32] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "Comic: Multi-view clustering without parameter selection," in *International conference on machine learning*. PMLR, 2019, pp. 5092–5101.

[36] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.