Dual Contrastive Prediction for Incomplete Multi-view Representation Learning

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, Xi Peng

Abstract—In this article, we propose a unified framework to solve the following two challenging problems in incomplete multi-view representation learning: i) how to learn a consistent representation unifying different views, and ii) how to recover the missing views. To address the challenges, we provide an information theoretical framework under which the consistency learning and data recovery are treated as a whole. With the theoretical framework, we propose a novel objective function which jointly solves the aforementioned two problems and achieves a provable sufficient and minimal representation. In detail, the consistency learning is performed by maximizing the mutual information of different views through contrastive learning, and the missing views are recovered by minimizing the conditional entropy through dual prediction. To the best of our knowledge, this is one of the first works to theoretically unify the cross-view consistency learning and data recovery for representation learning. Extensive experimental results show that the proposed method remarkably outperforms 20 competitive multi-view learning methods on six datasets in terms of clustering, classification, and human action recognition. The code could be accessed from https://pengxi.me.

Index Terms—Multi-view Learning, Contrastive Prediction, View Missing, Multi-view Clustering, Multi-view Representation Learning.

1 Introduction

In real-world applications, data is usually presented in multiple views or modalities, which often exhibit a variety of heterogeneous properties. To narrow down such a heterogeneous gap, multi-view representation learning (MvRL) [1], [2] aims to learn a function f that maps the multi-view data into a lower-dimensional space wherein a common representation is learned to facilitate the downstream tasks like clustering [3]–[13], classification [14]–[17], and human action recognition [18]. To achieve this goal, the key to MvRL is learning the consistency across different views.

The success of the consistency learning relies on an implicit data assumption, *i.e.*, all views are available for every data point. In practice, however, such an assumption is probably unsatisfactory due to the complexity in data collection and transmission, and therefore leads to the so-called incomplete multi-view problem (IMP). To solve the IMP, numerous methods [19]–[24] have been proposed in recent times, which aim to answer: i) how to learn a consistent representation across different views? and ii) how to recover the missing views for the incomplete data? Although the existing works have achieved promising performance, almost all of them treat the above two problems as two irrelevant tasks and solve them separately, thus leading to a suboptimal solution.

Based on the observation shown in Fig. 1, we theoretically show that the cross-view consistency learning and data recovery could be treated as two sides of one coin. With our theoretical results, we propose a general objective function which jointly solves these two challenging problems. We prove that the rep-

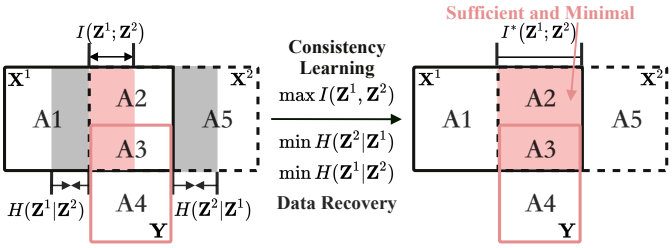


Fig. 1: Our key observation and basic idea. In the figure, $\mathbf{X}^1$ and $\mathbf{X}^2$ denote two views of a given dataset, and the corresponding representations are denoted by $\mathbf{Z}^1$ and $\mathbf{Z}^2$, respectively. The information contained in $\mathbf{X}^1$ =A1 $\cup$ A2 $\cup$ A3 and $\mathbf{X}^2$ = A2 $\cup$ A3 $\cup$ A5 are represented by the solid and dotted rectangles, respectively. The area under the red rectangular box $(\mathbf{Y} = \text{A3} \cup \text{A4})$ indicates the task-relevant information. Specifically, A1 $\big(H(\mathbf{X}^1|\mathbf{X}^2)\big)$ and A5 $\big(H(\mathbf{X}^2|\mathbf{X}^1)\big)$ indicate the view-specific information of $\mathbf{X}^1$ and $\mathbf{X}^2$. A2 $\big(I(\mathbf{X}^1;\mathbf{X}^2|\mathbf{Y})\big)$ and A3 $\big(I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y})\big)$ together denote the mutual information of $\mathbf{X}^1$ and $\mathbf{X}^2$, where A3 contains task-relevant information while A2 is task-agnostic. A4 $\big(H(\mathbf{Y}|\mathbf{X}^1,\mathbf{X}^2)\big)$ indicates the task-relevant information that is unavailable from the input data. From information theory, we propose to quantify the cross-view consistency and cross-view recoverability using the mutual information $I(\mathbf{Z}^1;\mathbf{Z}^2)$ (red area) and the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$ (grey area), respectively. To learn a consistent representation and recover the missing views, we maximize the mutual information $I(\mathbf{Z}^1;\mathbf{Z}^2)$ while minimizing the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$. Subtly, the two objectives could mutually boost and jointly optimizing both could achieve a sufficient (*i.e.*, A3 $\in \mathbf{Z}^i$) and minimal (*i.e.*, (A1 $\cup$ A5) $\notin \mathbf{Z}^i$ and A2 $\in \mathbf{Z}^i$) representation.

- *Y. Lin, Y. Gou, J. Lv, and X. Peng are with College of Computer Science, Sichuan University, Chengdu 610065, China. E-mail: { linyijie.gm, gouyuanbiao, pengx.gm }@gmail.com, lvjiancheng@scu.edu.cn*
- *X. Liu is with Department of Computer Science, Wake Forest University, Winston Salem, NC, 27109, USA. E-mail: liux16@wfu.edu.*
- *J. Bai is with TAL AI Lab, Beijing 100080, China. E-mail: baijinfeng1@tal.com.*

This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0104500; in part by NSFC under Grant U21B2040, U19A2078, 62176171, 61836006; in part by the Sichuan Science and Technology Planning Project under Grant 2022YFQ0014; in part by the Fund of Sichuan University Tomorrow Advancing Life; in part by the 111 Project under grant B21044.
(Corresponding author: Xi Peng.)

resentation learned by the objective function is sufficient and minimal. A sufficient representation refers to that the adequate information is learned for the downstream tasks, and a minimal representation denotes that all task-irrelevant information is removed with a fixed gap. To implement our idea, we propose a novel incomplete multi-view representation learning method, termed Dual Contrastive Prediction (DCP). To be specific, DCP projects the high-dimensional data into a latent space wherein the cross-view consistency and data recoverability are guaranteed by three joint losses. In short, a within-view reconstruction loss is used to learn the view-specific representations while preserving the original information; a dual contrastive loss is designed to learn the cross-view consistency by maximizing mutual information $I(\mathbf{Z}^1;\mathbf{Z}^2)$; and a dual prediction loss is proposed to recover the missing views through minimizing the conditional entropy $H(\mathbf{Z}^1|\mathbf{Z}^2)$ and $H(\mathbf{Z}^2|\mathbf{Z}^1)$. In summary, the contributions and novelties are given as follows:

- We provide a novel insight to the community that the cross-view consistency learning and data recovery are with intrinsic connections in the framework of information theory. Such a theoretical framework is remarkably different from existing MvRL studies which treat the consistency learning and data recovery as two separate problems.
- Under our information theoretical framework, we propose DCP which achieves the information consistency and data recoverability through a dual contrastive loss and a dual prediction loss, respectively.
- To utilize available label information, DCP designs and utilizes the instance- and category-level contrastive loss to enhance the separability of representations.
- We theoretically and experimentally prove that DCP could learn a sufficient and minimal representation for three tasks, *i.e.*, clustering, classification, and human action recognition.

2 Related Work

In this section, we briefly review three topics related to this work, *i.e.*, incomplete multi-view representation learning, contrastive learning, and information theory in multi-view learning.

2.1 Incomplete Multi-view Representation Learning

Incomplete multi-view representation learning (IMvRL) aims to learn a shared representation from the data with missing views. Based on the way of utilizing the cross-view information, IMvRL methods could be roughly divided into four categories, *i.e.*, matrix factorization based methods [25]–[28], kernel learning based methods [23], [29], spectral clusterinliuxinwangg based methods [22], and deep learning based methods [14], [30]. In brief, the matrix factorization based methods often project the incomplete data into a low-dimensional common subspace which satisfies the low-rankness constraint. For example, PVC [26] utilizes ℓ_F-norm and ℓ_1-norm to reduce the influence of missing data, and DAIMC [25] establishes a consensus representation matrix with the help of $\ell_{2,1}$-norm. As a typical kernel learning based method, MKKM-IK [29] proposes a multi-kernel algorithm through an iterative optimization manner. Through spectrum analysis, PIC [22] learns a common representation by constructing a consistent Laplacian graph from the incomplete views. Recently, some deep learning based methods [14], [18]–[20] utilize GAN to generate the missing views and then learn the shared subspace for all views.

This work is remarkably different from the existing approaches in the given aspects. First, almost all these methods [19], [20], [22], [23], [25]–[29] treat the consistency learning and data recovery as two separate problems. In contrast, our DCP unifies the consistency learning and data recovery into the framework of information theory [31]. Second, different from the works like [22], [23], [29], our method attempts to directly recover the missing data rather than the similarity relation, thus embracing higher explainability. In addition, it should also be pointed out that this paper is also different from [30] by the following three aspects. First, most of the methods including [30] lack a theoretical understanding of the success of the learned representation, whereas we prove that DCP could learn a sufficient and minimal representation. Second, [30] is a clustering method, whereas our method employs a novel loss function to learn the representation for unsupervised and supervised tasks including but not limited to clustering, classification, and human action recognition, as verified in our experiments. Third, [30] is designed for bi-view data, whereas DCP could solve the missing view problem with more than two views.

2.2 Contrastive Learning

As one of the most effective unsupervised learning paradigms, contrastive learning [32]–[39] has made huge developments in the field of representation learning. The basic idea of contrastive learning is to seek a latent space wherein the similarity between positive pairs is maximized while the similarity between negative pairs is minimized. Recently, some studies have shown that the success of contrastive learning could be attributed to the maximization of mutual information [36]. To be more precise, the widely-used InfoNCE loss is a lower bound of mutual information, *i.e.*, $I\left(\mathbf{Z}^1;\mathbf{Z}^2\right) \geq \log(N) - \mathcal{L}_{\mathrm{NCE}}$, where N is the number of negative pairs. Hence, MoCo [40] and CPC [37] could be interpreted as examples that maximizes the mutual information with the InfoNCE loss.

The differences between the existing contrastive learning studies and our work are as below. First, most of the existing contrastive learning methods employ data augmentations to generate different views, whereas our method directly handles the data from multiple views. As a result, they will not face and cannot handle the incomplete data problem. Second, unlike the existing studies [30], [33], [40] only perform contrastive learning at the instance level, our method simultaneously conducts instance- and category-level contrastive learning to obtain more discriminative representations. Although [36] also utilizes predictive learning to enhance contrastive learning, it is remarkably different from our work in the following aspects. On the one hand, the method and the goal is different. In short, [36] aims to enhance the performance of contrastive learning, whereas this work aims at recovering the missing data through dual predictive learning. On the other hand, our theoretical result is also different from [36]. In brief, we prove that the data recovery and consistency learning could mutually boost through contrastive learning and dual prediction.

2.3 Information Theory in Multi-view Learning

In recent years, some efforts have been devoted to the problem of multi-view representation learning based on information

theory [35], [36], [41]–[45]. Information bottleneck theory [46] provides a unified theoretic explanation for these works. More specifically, given the multi-view input $\mathbf{X}^1$ and $\mathbf{X}^2$, the quality of representation $\mathbf{Z}$ w.r.t. tasks $\mathbf{Y}$ could be characterized by the shared information between the representation and input $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2)$, and the shared task information between the representation and tasks $I(\mathbf{Z},\mathbf{Y})$. Ideally, a good representation is encouraged to maximize the task information $I(\mathbf{Z},\mathbf{Y})$ (sufficient) while minimizing the information from the raw data $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2)$ (minimal). Since most methods have successfully obtained sufficient task information, the major difference among them is the optimization strategy of $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2)$. For example, the supervised information bottleneck [43], [44] utilizes label information to minimize the superfluous information that $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2) = I(\mathbf{X}^1\mathbf{X}^2;\mathbf{Y})$. InfoMax [35], [41], [45], [47] aims to maximize $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2)$ in an unsupervised manner, thus preserving some superfluous information contained in raw data $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2) \geq I(\mathbf{X}^1;\mathbf{X}^2) > I(\mathbf{X}^1\mathbf{X}^2;\mathbf{Y})$ accordingly.

Different from the aforementioned methods [35], [41], [43]–[45], [47], DCP could be one of the first works that explicitly discards the irrelevant information and theoretically proves that the learned representation is sufficient $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2) = I(\mathbf{X}^1\mathbf{X}^2;\mathbf{Y})$ and minimal $I(\mathbf{Z};\mathbf{X}^1\mathbf{X}^2) = I(\mathbf{X}^1;\mathbf{X}^2)$. Moreover, we extend the multi-view information theory into the view-missing scenarios and reveal that the cross-view consistency learning and data recovery are with intrinsic connections under the framework of information theory.

3 The Proposed Method

In this section, we first introduce our information theoretical framework where the consistency learning and data recovery are unified. In addition, we prove that the theoretical framework could learn a sufficient and minimal representation. After that, we elaborate on the proposed Dual Contrastive Prediction method (DCP) which could learn multi-view representation from incomplete data.

3.1 Theoretical Results

In this section, we show that the consistency learning and data recovery could be regarded as a whole, and jointly optimizing them could learn a sufficient and minimal representation. In the following, we will first define the consistency learning and data recovery from the view of information theory. Then, we will prove that the consistency learning and data recovery could mutually boost. After that, we will propose a general objective function which is provable to learn a sufficient and minimal representation.

In the following, we denote $\mathbf{X}^1$ and $\mathbf{X}^2$ as two different views of the same instance whose label is $\mathbf{Y}$. The representation $\mathbf{Z}^i$ of view $\mathbf{X}^i$ is obtained through a deterministic mapping $f^{(i)}$: $\mathbf{Z}^i = f^{(i)}(\mathbf{X}^i)$, where $i = 1, 2$. We use $H(A)$ to denote the entropy, $H(A|B)$ to denote the conditional entropy, $I(A;B)$ to denote mutual information, and $I(A;B|C)$ to denote conditional mutual information for the variable A, B, and C. Considering the consistency learning and data recovery challenges, we have the following definitions:

Definition 1 (Cross-view Consistency)**.** *Two view-specific representations $\mathbf{Z}^i$ and $\mathbf{Z}^j$ are consistent if $I(\mathbf{Z}^i;\mathbf{Z}^j) \geq I(\mathbf{Z}^i;\mathbf{Z}')$ and $I(\mathbf{Z}^i;\mathbf{Z}^j) \geq I(\mathbf{Z}'';\mathbf{Z}^j)$ for any $\mathbf{Z}' \in \mathcal{T}(\mathbf{X}^j)$ and $\mathbf{Z}'' \in \mathcal{T}(\mathbf{X}^i)$, where $\mathcal{T}(\mathbf{X}^v)$ is the set of possible latent representations of the v-th view $\mathbf{X}^v$.*

Notice the mutual information $I(\mathbf{Z}^i;\mathbf{Z}^j)$ is represented by the red area in Fig. 1. By definition, it is expected to learn cross-view consistent representations by maximizing $I(\mathbf{Z}^i;\mathbf{Z}^j)$.

Definition 2 (Cross-view Recoverability)**.** *A representation $\mathbf{Z}^i$ is recoverable w.r.t. $\mathbf{Z}^j$ if $H(\mathbf{Z}^i|\mathbf{Z}^j) \leq H(\mathbf{Z}^i|\mathbf{Z}')$ for any $\mathbf{Z}' \in \mathcal{T}(\mathbf{X}^j)$. $\mathbf{Z}^i$ is fully recovered from $\mathbf{Z}^j$ i.i.f. $H(\mathbf{Z}^i|\mathbf{Z}^j) = 0$.*

Similarly, the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$ is represented by the grey area in Fig. 1. The representation $\mathbf{Z}^i$ is expected to be recovered from $\mathbf{Z}^j$ when $H(\mathbf{Z}^i|\mathbf{Z}^j)$ is minimized.

It should be pointed out that the above multi-view recoverability is task-oriented instead of general purposed, *i.e.*, only the view-shared instead of all information would be recovered to facilitate the downstream tasks, which include but not limited to clustering and classification. As illustrated in Fig. 1, one could easily see that our task-oriented recovery model will only restore the view-consistent information. Based on previous definitions, we then show that the cross-view consistency learning and data recoverability are equivalent from the information-theoretic point of view.

Theorem 1 (Equivalence of Consistency and Recoverability)**.** *Representations $\mathbf{Z}^i$ and $\mathbf{Z}^j$ are cross-view consistent if and only if $\mathbf{Z}^i$ is recoverable w.r.t. $\mathbf{Z}^j$ and $\mathbf{Z}^j$ is recoverable w.r.t. $\mathbf{Z}^i$.*

Proof. Let $\mathbf{Z}^i$ and $\mathbf{Z}^j$ be two consistent representations. Then we have, $I(\mathbf{Z}^i;\mathbf{Z}^j) \geq I(\mathbf{Z}^i;\mathbf{Z_0})$, for $\mathbf{Z_0} \in \mathcal{T}(\mathbf{X}^j)$. By the property of mutual information, $I(\mathbf{Z}^i;\mathbf{Z}^j) = H(\mathbf{Z}^i) - H(\mathbf{Z}^i|\mathbf{Z}^j)$. Hence we can rewrite the inequality as

$$H(\mathbf{Z}^i) - H(\mathbf{Z}^i|\mathbf{Z}^j) \geq H(\mathbf{Z}^i) - H(\mathbf{Z}^i|\mathbf{Z_0}), \tag{1}$$

and consequently,

$$H(\mathbf{Z}^i|\mathbf{Z}^j) \leq H(\mathbf{Z}^i|\mathbf{Z_0}), \tag{2}$$

for $\mathbf{Z_0} \in \mathcal{T}(\mathbf{X}^j)$. By definition, $\mathbf{Z}^i$ is recoverable w.r.t. $\mathbf{Z}^j$. Similarly, we can show $\mathbf{Z}^j$ is recoverable w.r.t. $\mathbf{Z}^i$. Following the argument backwards would trivially prove the other direction of the statement.

Theorem 1 indicates that the cross-view consistency and data recovery could be treated as two sides of one coin. On the one hand, the data recoverability could be further improved because maximizing $I(\mathbf{Z}^i;\mathbf{Z}^j)$ could increase the view-shared information. On the other hand, the view-inconsistent information will be discarded through minimizing $H(\mathbf{Z}^i|\mathbf{Z}^j)$, thus improving the consistency.

Based on the theoretical observation, we propose a general objective function which jointly optimizes the consistency learning and data recovery. Formally,

$$\begin{aligned} &\max I\left(\mathbf{Z}^1;\mathbf{Z}^2\right) \\ &\text{s.t. } \min H\left(\mathbf{Z}^1|\mathbf{Z}^2\right),\ \min H\left(\mathbf{Z}^2|\mathbf{Z}^1\right) \end{aligned} \tag{3}$$

The above objective could achieve a sufficient and minimal multi-view representation based on the commonly used assumption of multi-view data [42] as below:

Assumption 1 (Equal Sufficiency of Multi-view Data)**.** *The sufficiency of each view is approximately equivalent for downstream tasks, i.e., $I(\mathbf{X}^1;\mathbf{Y}) = I(\mathbf{X}^2;\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y})$ is held.*

According to the above sufficiency assumption, we could derive the following multi-view property.

Proposition 1. $I(\mathbf{X}^1;\mathbf{Y}|\mathbf{X}^2) = I(\mathbf{X}^2;\mathbf{Y}|\mathbf{X}^1) = 0$.

Proof. Based on the chain rule of mutual information, we have

$$I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{Y}|\mathbf{X}^2), \tag{4}$$
$$I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y}) = I(\mathbf{X}^2;\mathbf{Y}) - I(\mathbf{X}^2;\mathbf{Y}|\mathbf{X}^1). \tag{5}$$

With Assumption 1, we have $I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{Y}) = I(\mathbf{X}^2;\mathbf{Y})$. Then, $I(\mathbf{X}^1;\mathbf{Y}|\mathbf{X}^2) = I(\mathbf{X}^2;\mathbf{Y}|\mathbf{X}^1) = 0$ is obtained.

Based on the above assumption and proposition, we are ready to prove the sufficient and minimal property of the obtained representation.

Definition 3 (Sufficient Representation)**.** *The representations $\mathbf{Z}^1$ and $\mathbf{Z}^2$ are sufficient if $I(\mathbf{Z}^1;\mathbf{Y}) = I(\mathbf{Z}^2;\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y})$. Sufficient representations guarantee adequate information of the input for the downstream tasks, i.e., A3 $\in \mathbf{Z}^i$ in Fig. 1.*

Definition 4 (Minimal Representation)**.** *The representations $\mathbf{Z}^1$ and $\mathbf{Z}^2$ are minimal if $I(\mathbf{Z}^1;\mathbf{X}^1|\mathbf{Y}) = I(\mathbf{Z}^2;\mathbf{X}^2|\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{X}^2|\mathbf{Y})$. Minimal representations are expected to remove task-irrelevant information (A1 $\cup$ A5 $\notin \mathbf{Z}^i$ in Fig. 1) and retain a fixed gap $I(\mathbf{X}^1;\mathbf{X}^2|\mathbf{Y})$, i.e., A2 $\in \mathbf{Z}^i$ in Fig. 1.*

Theorem 2 (Sufficient and Minimal Multi-view Representations)**.** *The optimizers $\mathbf{Z}^1_{sm}$ and $\mathbf{Z}^2_{sm}$ of Eq. (3) are sufficient and minimal multi-view representations.*

Proof. We first prove the representations $\mathbf{Z}^i_{sm}$ are sufficient (A3 $\in \mathbf{Z}^i_{sm}$), where $i = \{1, 2\}$. As $\mathbf{Z}^i = f^{(i)}(\mathbf{X}^i)$, then the information flow $\mathbf{X}^i \to \mathbf{Z}^i$ [36] could be described by the Markov chains.

Without loss of generality, we consider the case of $i = 1$, and $i = 2$ is with the similar proof. In detail, based on the Markov chain and the Data Processing Inequality [48], $I(\mathbf{Z}^1;\mathbf{Z}^2)$ is maximized at $I(\mathbf{X}^1;\mathbf{X}^2)$. As $\mathbf{X}^1 \to \mathbf{Z}^1$ and $\mathbf{Z}^1_{sm}$ attempts to maximize $I(\mathbf{Z}^1;\mathbf{Z}^2)$, then $I(\mathbf{Z}^1_{sm};\mathbf{X}^2) = I(\mathbf{X}^1;\mathbf{X}^2)$. Hence we have

$$I(\mathbf{Z}^1_{sm};\mathbf{X}^2;\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y}), \tag{6}$$
$$I(\mathbf{Z}^1_{sm};\mathbf{X}^2|\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{X}^2|\mathbf{Y}). \tag{7}$$

Performing the chain rule of mutual information on Eq. (6) gives:

$$I\left(\mathbf{Z}^1_{sm};\mathbf{Y}\right) = I(\mathbf{X}^1;\mathbf{Y}) - I(\mathbf{X}^1;\mathbf{Y}|\mathbf{X}^2) + I\left(\mathbf{Z}^1_{sm};\mathbf{Y}|\mathbf{X}^2\right). \tag{8}$$

Proposition 1 has shown that $I(\mathbf{X}^1;\mathbf{Y}|\mathbf{X}^2) = 0$, and we only need to prove $I\left(\mathbf{Z}^1_{sm};\mathbf{Y}|\mathbf{X}^2\right) = 0$. To be specific, with the Markov chain $\mathbf{X}^2 \to \mathbf{Z}^2$ and the Data Processing Inequality, we have

$$I\left(\mathbf{Z}^1_{sm};\mathbf{Y}|\mathbf{X}^2\right) \le I\left(\mathbf{Z}^1_{sm};\mathbf{Y}|\mathbf{Z}^2\right). \tag{9}$$

As $H(\mathbf{Z}^1|\mathbf{Z}^2)$ is minimized at $\mathbf{Z}^1_{sm}$, then $I(\mathbf{Z}^1_{sm};\mathbf{Y}|\mathbf{Z}^2) = 0$. Hence

$$I(\mathbf{Z}^1_{sm};\mathbf{Y}|\mathbf{X}^2) \le I\left(\mathbf{Z}^1_{sm};\mathbf{Y}|\mathbf{Z}^2\right) = 0. \tag{10}$$

Combining Assumption 1, Proposition 1, Eq. (8), and Eq. (10), we have

$$I(\mathbf{Z}^1_{sm};\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{Y}) = I(\mathbf{X}^1;\mathbf{X}^2;\mathbf{Y}), \tag{11}$$

implying $\mathbf{Z}^1_{sm}$ is a sufficient representation.

Following, we prove that $\{\mathbf{Z}^i_{sm}\}^2_{i=1}$ are minimal (A1 $\cup$ A5 $\notin \mathbf{Z}^i_{sm}$ and A2 $\in \mathbf{Z}^i_{sm}$). By applying the chain rule of mutual information, we have

$$I\left(\mathbf{Z}^1;\mathbf{X}^1 \mid \mathbf{Y}\right) = I\left(\mathbf{Z}^1;\mathbf{X}^1;\mathbf{X}^2 \mid \mathbf{Y}\right) + I\left(\mathbf{Z}^1;\mathbf{X}^1 \mid \mathbf{X}^2,\mathbf{Y}\right). \tag{12}$$

Based on the Markov chain $\mathbf{X}^1 \to \mathbf{Z}^1$, we have

$$I(\mathbf{Z}^1;\mathbf{X}^1;\mathbf{X}^2|\mathbf{Y}) = I(\mathbf{Z}^1;\mathbf{X}^2|\mathbf{Y}). \tag{13}$$

Then, Eq. (12) could be rewritten as

$$I\left(\mathbf{Z}^1;\mathbf{X}^1 \mid \mathbf{Y}\right) = I\left(\mathbf{Z}^1;\mathbf{X}^2 \mid \mathbf{Y}\right) + I\left(\mathbf{Z}^1;\mathbf{X}^1 \mid \mathbf{X}^2,\mathbf{Y}\right). \tag{14}$$

Based on the Markov chains $\mathbf{X}^2 \to \mathbf{Z}^2$, we have

$$I(\mathbf{Z}^1_{sm};\mathbf{X}^1|\mathbf{Z}^2,\mathbf{Y}) \ge I(\mathbf{Z}^1_{sm};\mathbf{X}^1|\mathbf{X}^2,\mathbf{Y}). \tag{15}$$

Since $H(\mathbf{Z}^1|\mathbf{Z}^2)$ is minimized at $\mathbf{Z}^1_{sm}$, then

$$I(\mathbf{Z}^1_{sm};\mathbf{X}^1|\mathbf{Z}^2,\mathbf{Y}) = 0. \tag{16}$$

Combining Eq. (15) and Eq. (16) gives

$$I(\mathbf{Z}^1_{sm};\mathbf{X}^1|\mathbf{X}^2,\mathbf{Y}) = 0. \tag{17}$$

Substituting Eq. (14) with Eq. (7) and Eq. (17), and we end up with

$$I\left(\mathbf{Z}^1_{sm};\mathbf{X}^1 \mid \mathbf{Y}\right) = I\left(\mathbf{Z}^1_{sm};\mathbf{X}^2 \mid \mathbf{Y}\right) = I\left(\mathbf{X}^1;\mathbf{X}^2 \mid \mathbf{Y}\right), \tag{18}$$

showing $\mathbf{Z}^1_{sm}$ is a minimal representation.

3.2 The Loss Function

Eq. (3) is a general form of sufficient and minimal MvRL, and there are a variety of implementations. In this paper, we propose a novel loss function which employs dual contrastive learning and dual prediction to achieve information consistency and data recoverability, respectively. In brief, we first project the raw data into a latent feature space through a within-view reconstruction. Meanwhile, the consistent representation is obtained by maximizing the mutual information across different views through contrastive learning, and the data recoverability is guaranteed by minimizing the conditional entropy of different views through dual prediction.

As illustrated in Fig. 2, our method consists of three joint learning objectives, namely, a within-view reconstruction loss $\mathcal{L}_{rec}$, a cross-view dual contrastive loss $\mathcal{L}_{cl}$, and a cross-view dual prediction loss $\mathcal{L}_{pre}$. To sum up, our loss function is as below:

$$\mathcal{L} = \mathcal{L}_{cl} + \lambda_1 \mathcal{L}_{pre} + \lambda_2 \mathcal{L}_{rec}, \tag{19}$$

where the parameters λ_1 and λ_2 balance the importance of $\mathcal{L}_{pre}$ and $\mathcal{L}_{rec}$. In our experiments, these two balance factors are fixed to 0.1.

3.2.1 Within-view Reconstruction Loss

Without loss of generality, we take bi-view data as an example. For a given dataset $\bar{\mathbf{X}} = \{\bar{\mathbf{X}}^{(1,2)}, \bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}\}$ with n samples, $\bar{\mathbf{X}}^{(1,2)}$, $\bar{\mathbf{X}}^{(1)}$, and $\bar{\mathbf{X}}^{(2)}$ denote the examples presented in both views, the first view only, and the second view only, respectively. In other words, $\bar{\mathbf{X}}^{(1,2)}$ represents the set of complete data with m samples and $\mathbf{X}^v$ is the v-th view of complete samples, *i.e.*, $\bar{\mathbf{X}}^{(1,2)} = \{\mathbf{X}^1, \mathbf{X}^2\}$. During the training stage, we optimize our method in an end-to-end fashion on the complete samples

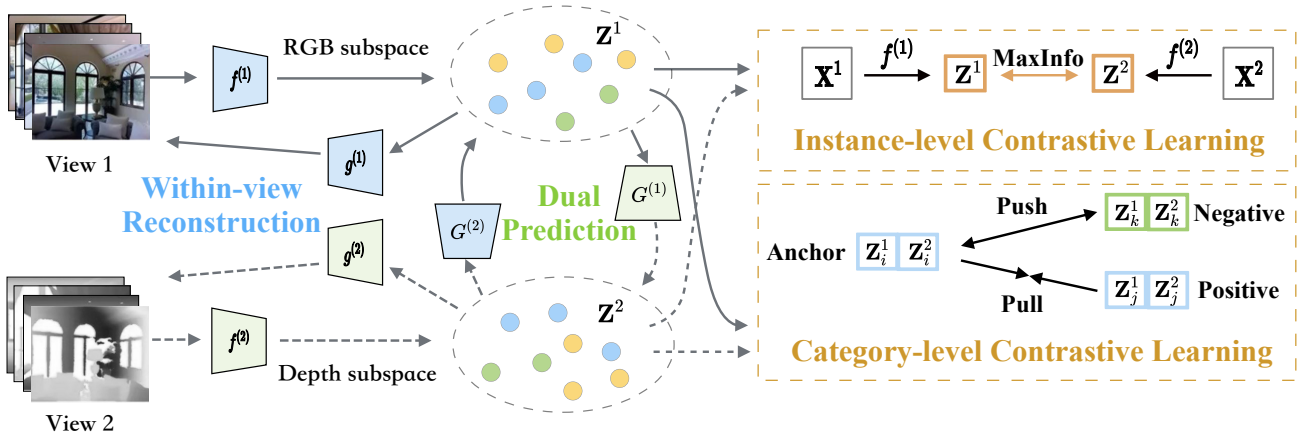


Fig. 2: Overview of DCP. In the figure, we use RGB and depth data as a showcase. As shown, DCP consists of three joint losses, *i.e.*, within-view reconstruction, dual cross-view contrastive learning, and cross-view dual prediction. Specifically, the within-view reconstruction loss projects each view into a view-specific subspace through an autoencoder. Dual contrastive learning objectives are constituted by the instance-level and category-level contrastive learning. In short, the instance-level contrastive learning loss aims to maximize the mutual information $I(\mathbf{Z}^1;\mathbf{Z}^2)$ for enhancing the cross-view consistency. The category-level contrastive learning loss aims to minimize the distance between an anchor (obtained by concatenating the view-specific representations) and a real within-class positive, while maximizing the distance between the anchor and a negative from the misclassified class. The dual prediction loss aims to recover one view from another view through the dual prediction $G^{(1)}$ and $G^{(2)}$.

$\bar{\mathbf{X}}^{(1,2)}$. In the testing stage, we feed the whole dataset including the incomplete ones $\{\bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}\}$ into the network and obtain the representations for all views.

To model the correlation across different views, we first learn a latent subspace for each view individually. To be specific, the v-th view is passed through a view-specific autoencoder to obtain the representation $\mathbf{Z}^v$ by minimizing

$$\mathcal{L}_{rec} = \sum_{v=1}^{2}\sum_{t=1}^{m}\left\|\mathbf{X}_t^v - g^{(v)}\left(\mathbf{Z}_t^v\right)\right\|_2^2, \tag{20}$$

where $\mathbf{X}_t^v$ denotes the t-th sample of $\mathbf{X}^v$, and $g^{(v)}$ is the decoder for the v-th view. The representation $\mathbf{Z}_t^v$ is defined by

$$\mathbf{Z}_t^v = f^{(v)}\left(\mathbf{X}_t^v\right), \tag{21}$$

where $f^{(v)}$ is the encoder for the view v. Thanks to Eq. (20), our method could preserve as much as possible view information, while avoiding trivial solutions.

3.2.2 Dual Contrastive Learning Loss

To overcome the consistency learning challenge in incomplete multi-view representation learning, we utilize contrastive learning to maximize the consistency of multi-view. The proposed dual cross-view contrastive learning consists of instance- and category-level contrast loss. Formally,

$$\mathcal{L}_{cl} = \mathcal{L}_{icl} + \mathcal{L}_{ccl}, \tag{22}$$

where the instance-level contrastive loss $\mathcal{L}_{icl}$ tries to learn an informative and consistent representation for different views without the help of labels and the category-level contrastive loss $\mathcal{L}_{ccl}$ aims to enhance the separability using the label information. Next, we will elaborate on these two objectives.

Instance-level Contrastive Learning: In the latent feature space learned by the autoencoder, we conduct contrastive learning to maximize the consistency across different views. Since most existing contrastive learning studies [37], [40] have attributed their success to maximize the lower bound of mutual information, we propose to maximize the mutual information between the representations of different views directly. Mathematically,

$$\mathcal{L}_{icl} = -\sum_{t=1}^{m}\left(I\left(\mathbf{Z}_t^1;\mathbf{Z}_t^2\right) + \alpha\left(H\left(\mathbf{Z}_t^1\right) + H\left(\mathbf{Z}_t^2\right)\right)\right), \tag{23}$$

where I and H denote the mutual information and entropy, respectively. The balance parameter α is fixed to 9 for regularizing the entropy in our experiments.

We design this loss with the following goals. On the one hand, a larger entropy $H(\mathbf{Z}^i)$ denotes a more informative representation. The reason could be explained by information theory [48], *i.e.*, the information entropy is the average amount of information of an event. Since the representation $\mathbf{Z}^i$ is conditioned on $\mathbf{X}^i$, *i.e.*, $\mathbf{Z}^i = f^{(i)}(\mathbf{X}^i)$, a larger entropy could retain more desirable information from $\mathbf{X}$. On the other hand, the maximization of $H(\mathbf{Z}^i)$ could avoid the trivial solution which would assign all instances to the same cluster.

To calculate the basic measures of information in $\mathcal{L}_{icl}$, we first define the probability distribution. More specifically, each element of $\{\mathbf{Z}^i\}_{i=1}^2$ is treated as a cluster assignment probability [47], [49] by stacking a softmax activation function on the last layer of the encoder. In other words, $\{\mathbf{Z}^i\}_{i=1}^2$ could be regarded as the distribution of two discrete cluster assignment variables z and z' over D classes, where D is the dimension of representations. In practice, D could be larger than the real cluster number, *i.e.*, so-called over-clustering. Therefore, we could define the joint probability distribution $\mathcal{P}(z, z')$ via $\mathbf{P} \in \mathcal{R}^{D\times D}$, *i.e.*,

$$\mathbf{P} = \frac{1}{m}\sum_{t=1}^{m}\mathbf{Z}_t^1\left(\mathbf{Z}_t^2\right)^\top. \tag{24}$$

Further, the marginal probability distributions $\mathcal{P}(z = d)$ and $\mathcal{P}(z' = d')$ could be defined by $\mathbf{P}_d$ and $\mathbf{P}'_d$, respectively. They

can be obtained by summing over the d-th row and the d'-th column of the joint probability distribution matrix $\mathbf{P}$. To sum up, Eq. (23) could be calculated through

$$\mathcal{L}_{icl} = -\sum_{d=1}^{D}\sum_{d'=1}^{D}\mathbf{P}_{dd'}\ln\frac{\mathbf{P}_{dd'}}{\mathbf{P}_{d}^{\alpha+1}\cdot\mathbf{P}_{d'}^{\alpha+1}}, \tag{25}$$

where $\mathbf{P}_{dd'}$ (*i.e.*, $\mathcal{P}(z=d,z'=d')$) is the d-th row and d'-th column of $\mathbf{P}$, and α is the same balance factor as defined in Eq. (23).

Category-level Contrastive Learning: To further enhance the separability of the representation, we propose a category-level contrastive loss function $\mathcal{L}_{ccl}$ that utilizes available label information to guide the representation learning in the supervised scenario. For the unsupervised settings, $\mathcal{L}_{ccl}$ will be removed from our loss function due to the unavailability of labels. To be specific, we enforce $\mathcal{L}_{ccl}$ on the common representations $\mathbf{Z}_t = [\mathbf{Z}_t^1;\mathbf{Z}_t^2]$, where $[;]$ is the concatenation operation. Formally,

$$\mathcal{L}_{ccl} = \sum_{t}^{m}\left[\mathbb{E}_{\mathbf{Z}\sim\mathcal{T}(y)}S(\mathbf{Z},\mathbf{Z}_t) - \mathbb{E}_{\mathbf{Z}\sim\mathcal{T}(gt)}S(\mathbf{Z},\mathbf{Z}_t)+\gamma\right]_+, \tag{26}$$

where gt denotes the ground truth of $\mathbf{Z}_t$, $S(\mathbf{Z},\mathbf{Z}_t)=\mathbf{Z}^T\mathbf{Z}_t$ is the dot product similarity function, $\mathcal{T}(gt)$ is the set of common representations from the ground truth label gt, and $\mathcal{T}(y)$ is the set of representations from the prediction y. The predicted label y is decided by

$$y = \arg\max_{y\in\mathcal{Y}}\mathbb{E}_{\mathbf{Z}\sim\mathcal{T}(y)}S(\mathbf{Z},\mathbf{Z}_t). \tag{27}$$

Our basic idea is to penalize the misclassification. To be specific, a misclassified instance should be close to the real instead of predicted within-class instances. The non-negative constant γ in Eq. (26) is used as a margin to control these two cases. Specifically, for a correctly predicted instance, $\gamma = 0$, and otherwise $\gamma = 1$. In other words, the category-level contrastive loss would not contribute to the training for the correctly predicted instances.

Comparing to the traditional triple loss [50], $\mathcal{L}_{ccl}$ is with the following differences. First, for each anchor, we directly choose the centroid of the true class and misclassified class as positive pair and negative pair, respectively. In contrast, [50] has to exhaustively choose the hard negative pairs for guaranteeing performance. Second, we only construct negative pairs in the misclassified class instead of all classes, thus leading to higher efficiency. Moreover, to speed up convergence, we compute $\mathcal{L}_{ccl}$ in a mini-batch fashion.

3.2.3 Dual Prediction Loss

To overcome the data recovery challenge in incomplete multi-view representation learning, we propose a dual prediction mechanism as shown in Fig. 2. To be specific, in the latent subspace parametrized by the autoencoders, the view-specific representation are mutually predicted by minimizing the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j) = -\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}[\log\mathcal{P}(\mathbf{Z}^i|\mathbf{Z}^j)]$. As it is intractable to directly calculate such expectations, we use a common approximative approach by importing a variational distribution $\mathcal{Q}(\mathbf{Z}^i|\mathbf{Z}^j)$ and then maximizing the lower bound of $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}[\log\mathcal{P}(\mathbf{Z}^i|\mathbf{Z}^j)]$, *i.e.*, $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log\mathcal{Q}\left(\mathbf{Z}^i\mid\mathbf{Z}^j\right)\right]$.

Such variational distributions $\mathcal{Q}$ can be Gaussian [51] or Categorical distribution [52]. For simplicity, we assume the distribution $\mathcal{Q}$ as a Gaussian distribution $\mathcal{N}\left(\mathbf{Z}^i\mid G^{(j)}\left(\mathbf{Z}^j\right),\sigma\mathbf{I}\right)$, where $\sigma\mathbf{I}$ is the variance matrix, and $G^{(j)}(\cdot)$ is a parametrized model which recovers $\mathbf{Z}^i$ from $\mathbf{Z}^j$ as shown in Fig. 2. By ignoring the constants derived from the Gaussian distribution, the maximization of the expectation $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log\mathcal{Q}\left(\mathbf{Z}^i\mid\mathbf{Z}^j\right)\right]$ is equivalent to

$$\min\ \mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left\|G^{(j)}\left(\mathbf{Z}^j\right)-\mathbf{Z}^i\right\|_2^2. \tag{28}$$

Considering the bi-view data, the dual prediction loss is with the following form:

$$\mathcal{L}_{pre} = \left\|G^{(1)}(\mathbf{Z}^1)-\mathbf{Z}^2\right\|_2^2 + \left\|G^{(2)}(\mathbf{Z}^2)-\mathbf{Z}^1\right\|_2^2. \tag{29}$$

It should be pointed out that the dual prediction loss alone may lead to trivial solutions that $\mathbf{Z}^1$ and $\mathbf{Z}^2$ converge to the same constant. To avoid such case, the within-view reconstruction loss is helpful as discussed above.

After the training stage, we feed the whole dataset including the incomplete ones $\{\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)}\}$ into the network and obtain the representations for all views. More precisely, for the samples with missing views $(\bar{\mathbf{X}}^{(1)},\bar{\mathbf{X}}^{(2)})$, we recover the missing representations $\hat{\mathbf{Z}}^{(i)}$ from the existing representations $\bar{\mathbf{Z}}^{(j)}$ through the dual prediction, *i.e.*,

$$\hat{\mathbf{Z}}^{(i)} = G^{(j)}\left(\bar{\mathbf{Z}}^{(j)}\right) = G^{(j)}\left(f^{(j)}(\bar{\mathbf{X}}^{(j)})\right), \tag{30}$$

where $\bar{\mathbf{Z}}^{(j)}$ is the representations of $\bar{\mathbf{X}}^{(j)}$. Afterwards, we derive the common representation $\mathbf{Z}$ by simply concatenating all view-specific representations. More specifically, for the complete samples, $\mathbf{Z}=[\mathbf{Z}^1;\mathbf{Z}^2]$; and for the incomplete samples, $\mathbf{Z}=\left[\hat{\mathbf{Z}}^{(1)};\bar{\mathbf{Z}}^{(2)}\right]$ or $\mathbf{Z}=\left[\bar{\mathbf{Z}}^{(1)};\hat{\mathbf{Z}}^{(2)}\right]$.

3.3 Contrastive Prediction for More than Two Views

In this section, we utilize dual contrastive prediction to solve the incomplete multi-view problem with more than two views. As shown in Fig. 3, there are two general formulations to solve this problem through dual contrastive prediction, namely, the core view based approach (DCP-CV) and the complete graph based approach (DCP-CG).

Given a dataset $\{\mathbf{X}^i\}_{i=1}^V$ with V views, the DCP-CV selects one important view (*e.g.*, $\mathbf{X}^1$ as shown in Fig. 3(a)) as the centre, and conduct dual contrastive prediction between $\mathbf{X}^1$ and other views $\mathbf{X}^j$ as follows:

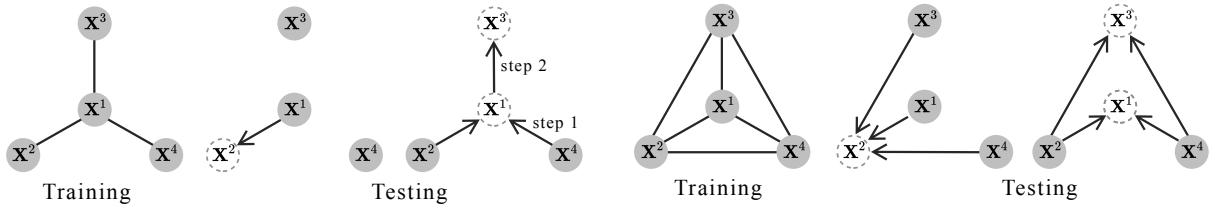
$$\mathcal{L}_{\text{DCP-CV}} = \sum_{i=2}^{V}\mathcal{L}_{ip}\left(\mathbf{Z}^1,\mathbf{Z}^i\right) + \lambda_2\sum_{i=1}^{V}\mathcal{L}_{rec}(\mathbf{Z}^i) + \mathcal{L}_{ccl}, \tag{31}$$

where $\mathcal{L}_{ip} = \mathcal{L}_{icl} + \lambda_1\mathcal{L}_{pre}$, λ_1 and λ_2 are scalars defined in Eq. (19). Category-level contrastive loss $\mathcal{L}_{ccl}$ is implemented on the common representation $[\mathbf{Z}^1;\cdots;\mathbf{Z}^V]$ where $[;]$ is the concatenation operation.

Alternatively, the DCP-CG conducts instance-level contrastive learning and dual prediction on all possible view pairs as shown in Fig. 3(b). Formally,

$$\mathcal{L}_{\text{DCP-CG}} = \sum_{1\le i<j\le V}\mathcal{L}_{ip}\left(\mathbf{Z}^i,\mathbf{Z}^j\right) + \lambda_2\sum_{i=1}^{V}\mathcal{L}_{rec}(\mathbf{Z}^i) + \mathcal{L}_{ccl}. \tag{32}$$

Although DCP-CV and DCP-CG are both capable of learning sufficient and minimal representations, we recommend the latter due to the following reasons. On the one hand, for consistency



(a) Core View (DCP-CV) (b) Complete Graph (DCP-CG)

Fig. 3: Contrastive prediction for more than two views. The solid and dotted circles denote the available views and missing views, respectively. The lines indicate training with $\mathcal{L}_{icl}$ and $\mathcal{L}_{pre}$, and the arrows indicate the recovery of missing view .

learning, DCP-CG will capture more information because all view pairs are included and exploited in the learning phase. In detail, the DCP-CG attempts to maximize the mutual information between views for $v(v-1)/2$ times, while the DCP-CV only maximizes $v-1$ times as shown in Fig. 3. On the other hand, the DCP-CG shows more robustness in performing data recovery than the DCP-CV, which is later verified in Section 4.5. Below are the theoretical explanations for the robustness of the DCP-CG. First, as shown in the middle column of Fig. 3(b), the DCP-CG could easily recover the missing views by averaging the predictions of other available views. Second, as shown in the right column of Fig. 3(a), the DCP-CV may encounter an error accumulation problem. To be specific, when both $\mathbf{X}^1$ and $\mathbf{X}^3$ are missing, one needs to first recover the core view $\mathbf{X}^1$ and recover $\mathbf{X}^3$ afterward. Besides, DCP-CV needs to choose an important view as the core view which is difficult when the prior knowledge is unavailable. In our implementation, we randomly select one view as the core view for simplicity.

4 Experiments

In this section, we compare the proposed method with 20 methods on six widely-used multi-view datasets in terms of three different tasks include clustering, classification, and human action recognition.

4.1 Experimental Settings

We conduct experiments on the following widely-used dataset:

- **LandUse-21** [53]: The dataset contains 2,100 satellite images from 21 categories, and we use the PHOG and LBP features as two views.
- **Caltech101-20** [54]: The dataset consists of 2,386 images of 20 subjects, and we use the HOG and GIST features as two views.
- **Scene-15** [55]: The dataset contains 4,485 images distributed over 15 scene categories. Similarly, we use the PHOG and GIST features as two views.
- **Noisy MNIST** [56] The dataset is a multi-view version of MNIST. It uses the original MNIST images as view 1 and randomly selects within-class images with Gaussian noise as view 2. We use a 20k subset of Noisy MNIST consisting of 10k validation images and 10k testing images because most of the baselines cannot handle such a large dataset.
- **UWA** (UWA3D Multi-view Activity) [57]: The dataset is collected by Kinect sensors with RGB and depth features. It contains 660 action sequences, *i.e.*, 11 actions performed by 12 subjects with five repetitions per action.
- **DHA** (Depth-included Human Action dataset) [58]: The dataset contains 483 video clips of 23 categories with RGB and depth features.

In the clustering and classification evaluations, we compare the proposed method with the following 12 approaches: (1) **DCCA** (Deep Canonically Correlated Analysis) [59] maps multiple features into a common space by neural networks and concatenates the low-dimensional features as the common representation. (2) **DCCAE** (Deep Canonically Correlated Autoencoders) [56] utilizes autoencoders to learn the common space for different views and concatenates each low-dimensional feature together as the final representation. (3) **BMVC** (Binary Multi-view Clustering) [60] jointly learns the collaborative discrete representations and binary cluster structures. (4) **AE^2-Nets** (Autoencoder in Autoencoder Networks) [61] integrates information from heterogeneous views by nested autoencoders. (5) **ITML** (Information-Theoretic Metric Learning) [62] uses a Mahalanobis distance function and Bregman optimization to learn the metric. (6) **PVC** (Partial Multi-View Clustering) [26] uses non-negative matrix factorization to project each view into a low-dimensional space. (7) **EERIMVC** (Efficient and Effective Regularized Incomplete Multi-view Clustering) [23] learns a consensus clustering matrix by completing the missing elements in the kernel matrix. (8) **DAIMC** (Doubly Aligned Incomplete Multi-view Clustering) [25] learns a common space using non-negative matrix factorization and $L_{2,1}$-Norm regularized regression. (9) **IMG** (Incomplete Multi-Modal Visual Data Grouping) [28] learns a compact global structure in the low-dimensional space by using a Laplacian graph of the complete instances. (10) **UEAF** (Unified Embedding Alignment Framework) [63] projects features of different views into a common space by using a novel reverse graph regularization term. (11) **PIC** (Perturbation-oriented Incomplete Multi-view Clustering) [22] learns the common representation based on spectral perturbation theory. (12) **CPMNets** (Cross Partial Multi-View Networks) [14] classifies partial multi-view data by focusing on the completeness and versatility of learned representations.

For the above baselines, we use the recommended parameters and network structures for a fair comparison. In brief, for DCCA and DCCAE, we fix the dimension of latent representation to 10. For EERIMVC, we choose the Gauss kernel to construct kernel matrices and search the optimal parameter λ from 2^{-15} to 2^{15} with an interval of 2^3. For BMVC, we set the dimension of binary code to 128. For ITML, we concatenate the original multi-view features as the input. It should be pointed out that, the first five baseline methods could only handle complete data and we use the mean value to fill the missing views.

To evaluate the performance of handling incomplete data, we manually create data with different missing rates and use the same incomplete data for all the tested methods. To be specific, given a dataset with v views, we randomly select m samples as incomplete data and randomly remove $1 \sim v-1$ views from each of them. The missing rate η is defined as $\eta = m/n$, where n is the number of all examples.

4.2 Network Architectures and Implementation Details

Our method contains two training modules, *i.e.*, view-specific autoencoders, and dual prediction networks. For these two modules, we employ a fully-connected network structure with batch normalization layer and ReLU activation on all datasets. To generate the over-clustering representations, we use a Softmax activation at the last layer of the encoders and prediction modules.

To be specific, we set the dimensionality of the encoders to $K-1024-1024-1024-D$ for Caltech101-20, Scene-15, Landuse-21, and Noisy MNIST dataset, where K is the dimension of raw data and D is the dimension of latent space. For the human action recognition datasets DHA and UWA, we set the dimensionality of the encoders to $K-2048-512-D$ for the RGB view and $K-1024-512-D$ for the Depth view. Note that the decoders mirror the encoders. The dimensionality of the dual predictors is fixed to $D-128-256-128-256-128-D$ for all datasets. In practice, D is set to 128, 128, 128, 64, 40, and 32 for Caltech101-20, Scene-15, UWA, DHA, Noisy MNIST, and Landuse-21, respectively.

For evaluation, we simply concatenate all view-specific representations as the common representation. For the clustering task, we feed the common representation into k-means like the traditional fashion [22], [23], [25], [26], [28], [56], [63]–[67]. For the supervised tasks (*i.e.*, classification and action recognition), we use Eq. (27) to predict the label.

We carry out experiments on an Ubuntu 18.04 OS with an NVIDIA 2080Ti GPU in PyTorch 1.2 [68]. We use Adam optimizer [69] with the initial learning rate of 0.0001 for all datasets. The batch-size is set to 256 for LandUse-21, Caltech101-20, Scene-15, and Noisy MNIST; 128 for DHA and 200 for UWA. The maximal training epoch is fixed to 500 for LandUse-21, Caltech101-20, Scene-15, and Noisy MNIST; and 2000 for DHA and UWA. The trade-off parameters α, λ_1 and λ_2 are fix to 9, 0.1, and 0.1 for all datasets. For the time cost evaluation, our method takes about 50 seconds to train a model on LandUse-15, 60 seconds on Caltech101-20, 80 seconds on Scene-15, 500 seconds on Noisy MNIST, 50 seconds on UWA, and 40 seconds on DHA.

4.3 Experiments on Clustering

In this section, we evaluate our method with 10 state-of-the-art multi-view clustering methods. In this experiment, our DCP does not include the category-level contrastive loss $\mathcal{L}_{ccl}$ for optimization. Instead, we first pre-train the autoencoders by $\mathcal{L}_{icl}$ and $\mathcal{L}_{rec}$ with 100 epochs which stabilizes the training of the dual prediction and after the warm-up, we train the whole networks with Eq. (19).

For a comprehensive analysis, we utilize three widely-used clustering metrics including ACC (Accuracy), NMI (Normalized Mutual Information), and ARI (Adjusted Rand Index) for performance evaluation. In general, a higher value indicates a better clustering performance. We investigate the effectiveness of all methods by increasing the missing rate η from 0 to 0.8 with a gap of 0.1. The clustering results are obtained by repeating each method with five random dataset partitions and initializations.

From the results in Fig. 4 and Appendix A, one could observe that: i) when data is complete, DCP achieves competitive performance on all datasets, which verifies the effectiveness of our method on the complete multi-view data; ii) our method significantly outperforms all the tested baselines in almost all settings; iii) with the increase of the missing rate, the performance degradation of the compared methods are much larger than that of ours. For example, our method and the most competitive baseline PIC achieve the NMI of 68.06% and 67.93% when $\eta = 0$ on Caltech101-20. While, with the increase of η, our method is remarkably superior to PIC, *e.g.*, 67.39% vs. 64.32% when $\eta = 0.5$.

4.3.1 Parameter Analysis

In this section, we conduct the parameter sensitivity analysis and ablation study on the Caltech101-20 dataset. In the experiment, we fix the missing rate η to 0.5.

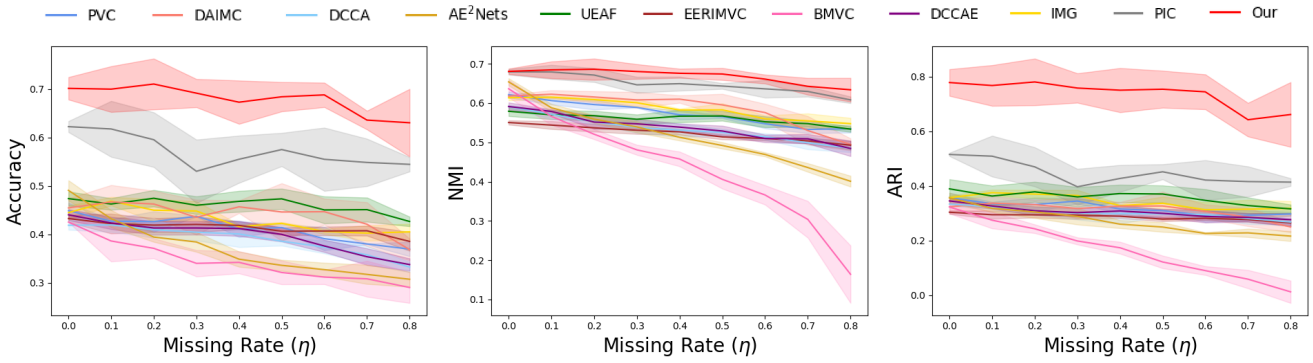
As discussed above, our method contains three balance parameters, namely, the dual prediction trade-off parameter λ_1, the reconstruction trade-off parameter λ_2, and the entropy trade-off parameter α. Although DCP with the fixed value of these parameters has shown promising performance, it is still important to explore the influence of these parameters and the full potential of our method. To the end, we first evaluate the influence of λ_1 and λ_2. As shown in Fig. 6, we change the value of λ_1 and λ_2 in the range of $\{0.01, 0.1, 1, 10, 100\}$. From the results, one could observe that our method is not sensitive to λ_1. In addition, a good choice of λ_2 (0.1 or 1) will remarkably improve the performance of our method and the best result is achieved when $\lambda_2 = 1$.

Next, we evaluate the influence of α by investigating the relations among α, the information entropy of representations $H(Z^i)$, and the clustering performance (*i.e.*, ACC, NMI, and ARI). In the evaluations, we fix λ_1 and λ_2 to 0.1 based on the above parameter analysis and increase the value of α from 0 to 30 with an interval of 0.5. As shown in Fig. 7, with increasing α, $H(Z^i)$ synchronously grows while the clustering performance generally first improves and then degrades. Such a result could be explained by the InfoMin principle [43], *i.e.*, the irrelevant information might be thrown away when the mutual information between views is reduced. To be specific, the increased entropy will enlarge the mutual information and thus the clustering performance is improved. When α further increases, the mutual information will contain more task-irrelevant information, thus leading to more redundancy that will suppress the performance of the downstream tasks.

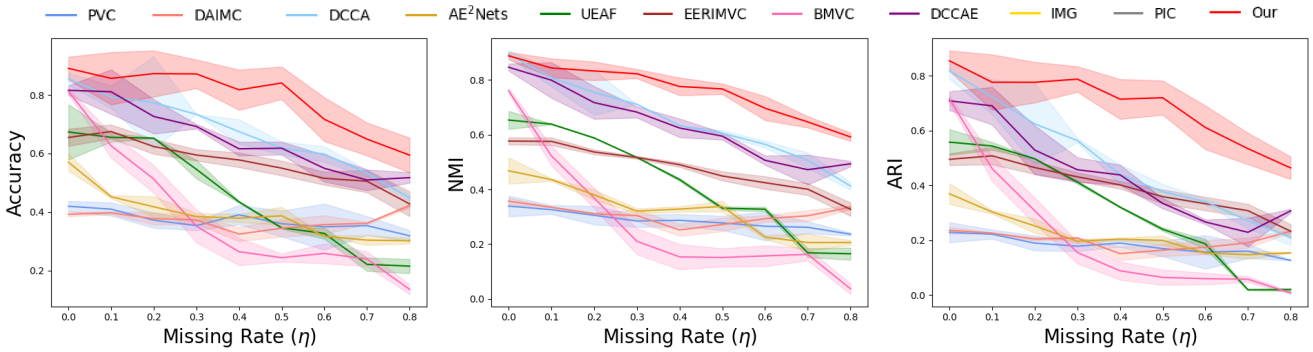
4.3.2 Visualization on Data Recovery

In this section, we carry out visual experiments on the Noisy MNIST dataset by passing the recovered representation through the decoder to recover the missing views. The experiment is designed to verify our theoretical results, *i.e.*, only the information shared by views (*i.e.*, A2 and A3 in Fig. 1) will be recovered or equivalently DCP could learn sufficient and minimal representations. In addition, the evaluation will also show that our method could explicitly infer the representation rather than the similarity of the missing views.

In the experiments, the missing rate η is set as 0.5 and some recovered examples are shown in Fig. 8. In the figure, we recover view 1 from view 2 (the top three rows) and view 2 from view 1

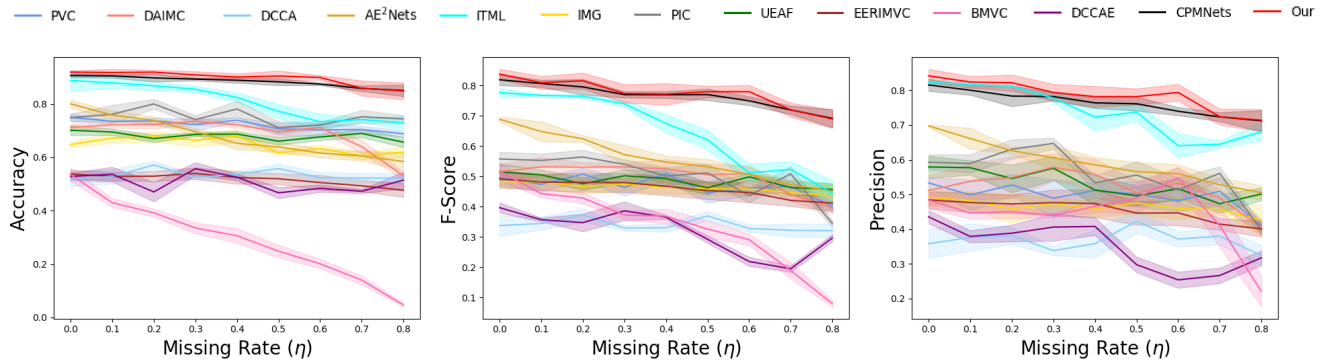


(a) Caltech101-20

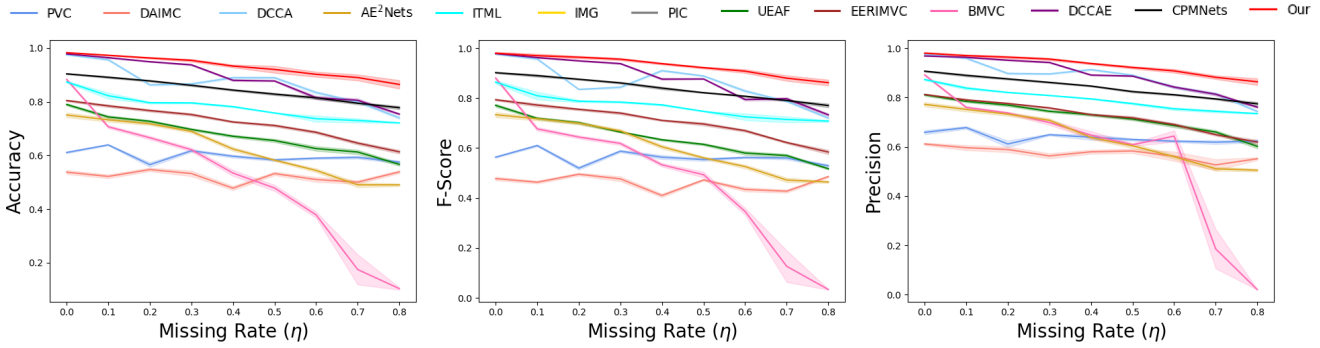


(b) Noisy MNIST

Fig. 4: Clustering performance comparisons on (a) Caltech101-20 and (b) Noisy MNIST with different missing rates (η).



(a) Caltech101-20



(b) Noisy MNIST

Fig. 5: Classification performance comparisons on (a) Caltech101-20 and (b) Noisy MNIST with different missing rates (η).

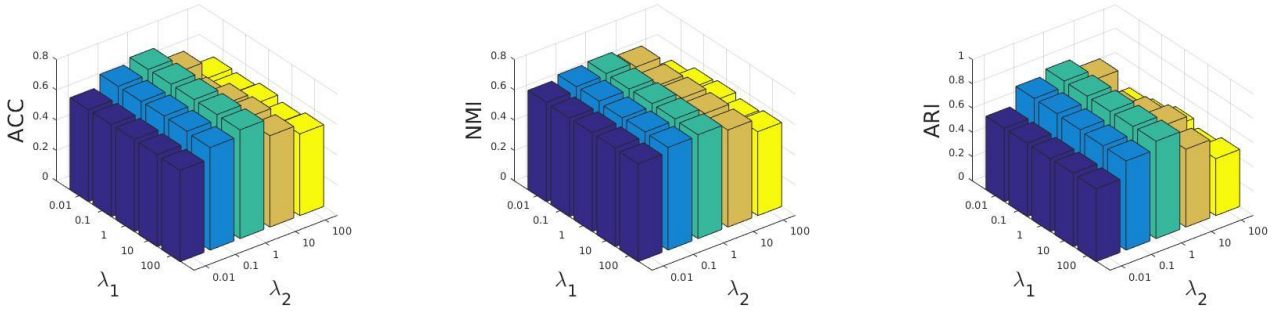


Fig. 6: Parameter analysis on Caltech101-20.

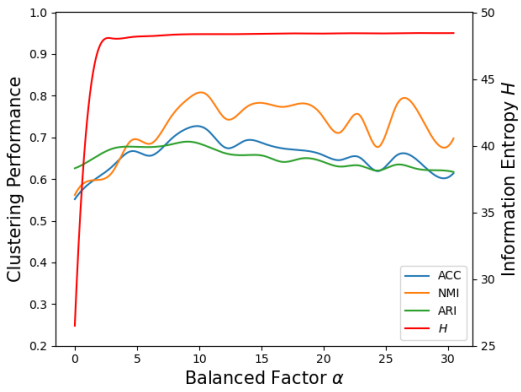


Fig. 7: Clustering results of our method with the increasing α on Caltech101-20. The x-axis denotes the parameter α, the left and right y-axis denote clustering performance and information entropy, respectively.

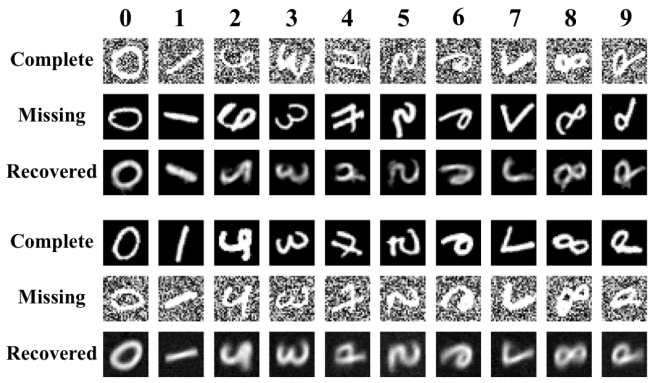


Fig. 8: Data recovery on the Noisy MNIST dataset. Row 1 and 4 are complete views, Row 2 and 5 are missing views, and Row 3 and 6 are the recovered results from the complete view.

(the bottom three rows), respectively. From the results, one could have the following observations. In the first case (*i.e.*, the first three rows), although the recovered images are inferred from the complete ones, they are more similar to the missing images instead of the complete ones. In the second case (*i.e.*, the last three rows), the contrary results are obtained. These results show that DCP recovers the important characteristics (digits) while discarding the superfluous information like noises, which is consistent with our theoretical result. More specifically, on the one hand, the recovered images will sufficiently remain the shared information (*i.e.*, digits) of two views through maximizing the mutual information. On the other hand, the recovered images will discard the inconsistent information (noisy background) by minimizing the conditional entropy.

4.4 Experiments on Classification

In this section, we evaluate the effectiveness of our method for classification task comparing with 12 state-of-the-art multi-view representation learning algorithms. We vary the missing rate η from 0 to 0.8 with a gap of 0.1. The classification results are obtained by repeating each method with five random initializations and dataset partitions. For a comprehensive analysis, three widely-used classification metrics including Accuracy, F-score, and Precision are used. A higher value of these metrics indicates a better classification performance.

From Fig. 5 and Appendix A, one could observe that: i) DCP achieves the highest performance among all baselines in most of the settings; ii) Although some baselines achieve competitive results when the missing rate is low, their results degenerate significantly with increasing η. Taking the results on Noisy MNIST as an example (see Fig. 5(b)), DCP and DCCAE obtain an Accuracy of 98.30% and 97.84% when $\eta = 0$, respectively. However, when $\eta = 0.8$, DCP performs remarkably superior to DCCAE, *i.e.*, 86.47% vs. 75.74%.

4.4.1 Ablation Studies

We carry out the following ablation study to investigate the role of our four losses, *i.e.*, the instance-level contrastive loss $\mathcal{L}_{icl}$, the category-level contrastive loss $\mathcal{L}_{ccl}$, the dual prediction loss $\mathcal{L}_{pre}$, and the reconstruction loss $\mathcal{L}_{rec}$. From Table 3, one could observe that all loss terms play indispensable roles. It should be pointed out that category-level contrastive loss $\mathcal{L}_{ccl}$ utilizes the label information thus cannot be applied for unsupervised clustering tasks. As shown in the bottom three lines of Table 3, the separability of the representation could be further enhanced by incorporating category-level contrastive loss $\mathcal{L}_{ccl}$ with the instance-level contrastive loss $\mathcal{L}_{icl}$.

TABLE 1: Multi-view clustering performance with more than two views. † denotes different variations of our method DCP.

Missing Type	Methods	Caltech101-20			Scene-15			LandUse-21		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Incomplete	EERIMVC [23] (TPAMI'20)	42.72	49.77	27.96	33.45	30.13	16.36	22.05	22.60	7.76
	BMVC [60] (TPAMI'19)	36.46	45.46	20.38	28.73	26.78	10.16	20.14	21.36	5.26
	DCP (bi-view)	68.44	**67.39**	75.44	39.50	**42.35**	**23.51**	23.07	27.00	**11.13**
	DCP-CV†	**69.76**	67.10	74.65	38.92	40.59	22.87	24.03	30.32	10.04
	DCP-CG†	68.58	66.85	**76.41**	**40.18**	41.81	23.46	**25.09**	**31.20**	10.95
Complete	EERIMVC [23] (TPAMI'20)	42.63	54.03	30.16	40.41	38.39	23.41	25.25	30.87	12.00
	BMVC [60] (TPAMI'19)	48.11	59.23	35.95	40.35	43.76	24.29	25.61	29.80	12.80
	DCP (bi-view)	70.18	68.06	**76.88**	41.07	45.11	24.78	26.23	30.65	**13.70**
	DCP-CV†	70.34	69.29	76.10	40.86	44.42	24.79	26.37	32.38	13.06
	DCP-CG†	**70.58**	**69.59**	76.83	**41.81**	**45.23**	**25.84**	**26.66**	**32.74**	13.50

TABLE 2: Multi-view classification performance with more than two views. † denotes different variations of our method DCP.

Missing Type	Methods	Caltech101-20			Scene-15			LandUse-21		
		ACC	Precision	F-score	ACC	Precision	F-score	ACC	Precision	F-score
Incomplete	EERIMVC [23] (TPAMI'20)	51.92	46.11	45.33	45.81	44.31	39.71	35.33	31.50	30.31
	BMVC [60] (TPAMI'19)	52.51	49.41	38.62	34.62	42.12	31.38	28.52	39.73	28.56
	CPMNets [14] (TPAMI'20)	87.11	78.27	75.81	48.42	48.65	45.41	31.76	31.59	29.50
	DCP (bi-view)	**90.48**	78.20	**78.00**	65.70	64.80	63.40	57.71	57.40	55.60
	DCP-CV†	87.09	74.00	71.60	67.53	66.20	65.60	55.14	55.80	53.40
	DCP-CG†	87.49	**81.00**	76.00	**68.14**	**67.20**	**66.20**	**58.81**	**59.20**	**57.20**
Complete	EERIMVC [23] (TPAMI'20)	53.59	48.09	49.44	53.71	54.69	47.45	37.47	35.49	33.90
	BMVC [60] (TPAMI'19)	76.61	73.17	68.29	42.72	40.65	34.57	43.57	45.92	41.01
	CPMNets [14] (TPAMI'20)	91.55	82.82	83.34	56.11	58.35	52.72	42.00	42.77	39.60
	DCP (bi-view)	91.93	84.20	83.80	75.10	74.00	73.40	69.67	70.80	70.40
	DCP-CV†	92.85	85.20	82.80	75.44	74.40	74.30	72.24	72.40	71.80
	DCP-CG†	**93.31**	**87.80**	**84.20**	**75.59**	**74.80**	**74.40**	**73.14**	**73.20**	**72.80**

TABLE 3: Ablation study on the Caltech101-20 dataset, where "✓" indicates the used component. The results in the middle column and right column denote the clustering results and classification results, respectively.

$\mathcal{L}_{pre}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{icl}$	$\mathcal{L}_{ccl}$	ACC	NMI	ARI	ACC	Precision	F-score
✓				38.61	37.65	26.50	12.92	11.00	5.00
	✓			33.65	31.60	16.43	61.17	42.80	42.60
		✓		46.69	58.03	41.86	74.44	41.80	41.20
✓	✓			54.70	52.63	43.49	63.64	**52.40**	41.46
	✓	✓		55.75	59.35	58.88	76.01	43.40	43.20
✓		✓		64.59	62.11	71.07	76.07	44.20	44.60
✓	✓	✓		**68.44**	**67.39**	**75.44**	**76.32**	49.40	**47.60**
✓	✓		✓	–	–	–	86.19	77.80	70.60
✓	✓	✓	✓	–	–	–	**90.48**	**78.20**	**78.00**

4.5 Experiments with More than Two Views

In this section, we evaluate the effectiveness of our method on multi-view data in both clustering and classification tasks. For the Caltech101-20 dataset, we use the HOG, GIST, and LBP features. For the Scene-15 and LandUse-21 datasets, we use the PHOG, LBP, and GIST features. For our core view based approach (DCP-CV), we choose the HOG, GIST, and PHOG as the core view for Caltech101-20, Scene-15, and LandUse-21, respectively. To verify the generalization of DCP, we use the same network and hyper-parameters as described in Section 4.2.

We test all methods in two settings, *i.e.,* missing rate $\eta = 0.5$ (marked as Incomplete) and $\eta = 0$ (marked as Complete). From the result in Table 1 and Table 2, one could observe that: i) in these two settings, the DCP-CG outperforms the DCP-CV on all three datasets, showing that the former is a better choice as discussed in Section 3.3; ii) when the views are complete, the DCP-CG and the DCP-CV both achieve better results than DCP (bi-view), indicating that the representation quality could be further improved with increasing view number; iii) although the missing scenario of triple-view data is more complex than bi-view data, the DCP-CG achieves comparable results with DCP (bi-view), demonstrating the scalability of our method to the multi-view setting.

4.6 Experiments on Human Action Recognition

In this section, we compare our method with eight state-of-the-art multi-view action recognition baselines on two datasets. More specifically, (1) **SVM** (Support Vector Machine) [70] is a traditional classifier which constructs hyperplanes in a predetermined high-dimension space. (2) **VLAD** (Action Vector of Local Aggregated Descriptor) [71] learns the representation by aggregating local features and the video spatio-temporal content. (3) **TSN** (Temporal Segment Networks) [72] combines a sparse temporal sampling with the video-level supervision. (4) **WDMM** (Weighted Depth Motion Maps) [73] uses the linear aggregation of spatio-temporal information to recognize from the depth view. (5) **AMGL** (Auto-Weight Multiple Graph Learning) [74] is a supervised multi-view classification method which learns an optimal weight for each graph. (6) **MLAN** (Multi-view Learning with Adaptive Neighbours) [75] performs semi-supervised classification using an adaptive graph-based learning method. (7) **GMVAR** (Generative Multi-View Human Action Recognition) [18] explores the latent connections in both intra- and cross-view using an adversarial generative network. (8) **GVCA** (Generative View-Correlation Adaptation) [76] employs a semi-supervised data augmentation mechanism to enhance the action recognition performance.

To evaluate SVM in the multi-view scenario, we concatenate the RGB and depth features as the input like [18], [76] do. In

TABLE 4: Human action recognition performance on the UWA dataset, where "–" denotes the method cannot handle such scenarios. In addition, RGB (R), Depth (D), and R+D denote the performance with single RGB view, single depth view, and RGB-Depth view, respectively. A→B indicates that the view A is generated by the view B. The best and the second-best result is indicated in red and blue color, respectively.

Method	RGB	R→D	Depth	D→R	R+D
SVM [70]	69.44	68.53	34.92	34.33	72.72
VLAD [71]	71.54	-	-	-	-
TSN [72]	71.01	-	-	-	-
WDMM [73]	-	-	46.58	-	-
AMGL [74]	69.17	71.54	39.92	35.96	68.53
MLAN [75]	67.19	67.19	33.28	33.61	66.64
GMVAR [18]	-	73.53	-	50.35	76.28
GVCA [76]	-	-	-	-	77.08
Ours: (Mean)	79.92	79.69	50.39	50.16	77.95
Ours: (Best)	81.10	81.88	51.18	51.18	80.31

TABLE 5: Human action recognition performance on the DHA dataset.

Method	RGB	R→D	Depth	D→R	R+D
SVM [70]	66.11	70.24	78.92	78.18	83.47
VLAD [71]	67.13	-	-	-	-
TSN [72]	67.85	-	-	-	-
WDMM [73]	-	-	81.05	-	-
AMGL [74]	64.61	59.05	72.84	67.33	74.89
MLAN [75]	67.91	67.91	72.96	72.83	76.13
GMVAR [18]	-	69.72	-	83.48	88.72
GVCA [76]	-	-	-	-	89.31
Ours: (Mean)	78.43	79.50	79.26	80.99	89.26
Ours: (Best)	82.64	81.40	82.23	83.05	90.08

experiments, we use TSN [72] and WDMM [73] to extract the RGB features and depth features, respectively. More specifically, each video is divided into five segments and a snippet is randomly chosen from each segment. In accordance with [18], we sample three snippets from each video. For the RGB view of the snippets, we use the ResNet-101 pre-trained on ImageNet to produce the class scores for each snippet and the class scores are with the dimension of 6,144. For the depth view of the snippets, we utilize WDMM and follow the same scheme used in [18], [76] to obtain a 110-dimensional feature vector. In the evaluation, 50% samples are used for training and the remainder is used for testing.

The action recognition accuracy is reported in Table 4 and 5. As our method could recover the missing view from its correspondence, we show its result in terms of R→D and D→R. For SVM, AMGL, and MLAN which could only handle the complete data, we use the mean feature as the "generated" view. From the results, one could observe that DCP achieves the best results in almost all settings. Especially, its performance superiority is quite notable when the RGB view is available, *i.e.*, R→D and RGB. In detail, the accuracy of DCP is 12% and 9% higher than the best baseline in the case of RGB and R→D, respectively. Moreover, the results in R+D illustrate that our method could further improve the accuracy compared with the baselines.

5 Conclusion

This paper theoretically unifies two seemingly separate challenges in incomplete multi-view representation learning, namely, consistency learning and missing data recovery. We show the two problems are equivalent from the perspective of information theory and could be treated as two sides of one coin rather than two independent objectives. Our proposed method DCP elegantly achieves both objectives by jointly optimizing a dual contrastive loss and a dual prediction loss. The experiments show that DCP achieves superior performance in both complete and incomplete scenarios over existing approaches and is scalable to the multi-view (more than two) settings. One open question worth discussing is the role of complementary principles in incomplete multi-view representation learning. More specifically, as shown in the visualization results of data recovery, only the shared information among different views will be recovered. Hence, a question naturally arises on whether preserving the view-specific information during recovery would lead to a better representation. We think such a question is worthy to explore and would inspire some novel insights to the community.

Acknowledgements

The authors would like to thank the associate editor and reviewers for the constructive comments and valuable suggestions that remarkably improve this work.

References

[1] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data ai: framework, application and case studies," *Front. Inf. Technol. Electron. Eng.*, vol. 22, no. 12, pp. 1551–1558, 2021.
[2] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021.
[3] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 46–58, 2018.
[4] X. Yang, C. Deng, T. Liu, and D. Tao, "Heterogeneous graph attention network for unsupervised multiple-target domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1992–2003, 2022.
[5] C. Xu, D. Tao, and C. Xu, "Multi-view self-paced learning for clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3974–3980.
[6] Z. Kang, W. Zhou, Z. Zhao, J. Shao, M. Han, and Z. Xu, "Large-scale multi-view subspace clustering in linear time," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4412–4419.
[7] C. Lu, S. Yan, and Z. Lin, "Convex sparse spectral clustering: Single-view to multi-view," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2833–2843, 2016.
[8] C. Xu, D. Tao, and C. Xu, "Multi-view learning with incomplete views," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5812–5825, 2015.
[9] Z. Tao, H. Liu, S. Li, Z. Ding, and Y. Fu, "From ensemble clustering to multi-view clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2843–2849.
[10] 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.*, vol. 97, 2019, pp. 5092–5101.
[11] 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.*, 2021, pp. 1134–1143.
[12] M. Yang, Y. Li, P. Hu, J. Bai, J. C. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.
[13] X. Peng, Y. Li, I. W. Tsang, H. Zhu, J. Lv, and J. T. Zhou, "Xai beyond classification: Interpretable neural clustering," *Journal of Machine Learning Research*, vol. 23, no. 6, pp. 1–28, 2021.
[14] 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.*, 2020.
[15] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2155–2161.
[16] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," in *Proc. Int. Conf. Learn. Representations*, 2021.

[17] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022.

[18] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6211–6220.

[19] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu, "Partial multi-view clustering via consistent gan," in *Proc. - IEEE Int. Conf. Data Min.*, 2018, pp. 1290–1295.

[20] Y. Jiang, Q. Xu, Z. Yang, X. Cao, and Q. Huang, "DM2C: deep mixed-modal clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5880–5890.

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

[22] H. Wang, L. Zong, B. Liu, Y. Yang, and W. Zhou, "Spectral perturbation meets incomplete multi-view data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3677–3683.

[23] X. Liu, M. Li, C. Tang, J. Xia, J. Xiong, L. Liu, M. Kloft, and E. Zhu, "Efficient and effective regularized incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2634–2646, 2020.

[24] X. Li and S. Chen, "A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[25] M. Hu and S. Chen, "Doubly aligned incomplete multi-view clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2262–2268.

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

[27] W. Shao, L. He, and S. Y. Philip, "Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization," in *Eur. Conf. on Mach. Learn. and Princ. and Pract. of Knowl. Discov. in DB*, 2015, pp. 318–334.

[28] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2392–2398.

[29] X. Liu, M. Li, L. Wang, Y. Dou, J. Yin, and E. Zhu, "Multiple kernel k-means with incomplete kernels," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2259–2265.

[30] 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.*, 2021, pp. 11 174–11 183.

[31] K. Sridharan and S. M. Kakade, "An information theoretic framework for multi-view learning," in *Annu. Conf. Learn. Theory*, 2008, pp. 403–414.

[32] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.

[33] 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.

[34] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6706–6716.

[35] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.

[36] Y. H. Tsai, Y. Wu, R. Salakhutdinov, and L. Morency, "Self-supervised learning from a multi-view perspective," in *Proc. Int. Conf. Learn. Representations*, 2021.

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

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

[39] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1255–1265.

[40] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[41] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Representations*, 2019.

[42] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2020.

[43] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6827–6839.

[44] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, 2014.

[45] Z. Wan, C. Zhang, P. Zhu, and Q. Hu, "Multi-view information-bottleneck representation learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 11, 2021, pp. 10 085–10 092.

[46] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *physics/0004057*, 2000.

[47] X. Ji, A. Vedaldi, and J. F. Henriques, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9864–9873.

[48] M. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.

[49] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, "Deep clustering with sample-assignment invariance prior," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4857–4868, 2020.

[50] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[51] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[52] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Annu. Meet. Assoc. Comput. Linguist.*, 2019, pp. 2978–2988.

[53] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL Int. Conf. Adv. Inf.*, 2010, pp. 270–279.

[54] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2750–2756.

[55] F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 524–531.

[56] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[57] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2430–2443, 2016.

[58] Y. Lin, M. Hu, W. Cheng, Y. Hsieh, and H. Chen, "Human action recognition and retrieval using sole depth information," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 1053–1056.

[59] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[60] Z. Zhang, L. Liu, F. Shen, H. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1774–1782, 2019.

[61] C. Zhang, Y. Liu, and H. Fu, "Ae2-nets: Autoencoder in autoencoder networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2577–2585.

[62] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 227, 2007, pp. 209–216.

[63] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and H. Liu, "Unified embedding alignment with missing views inferring for incomplete multi-view clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5393–5400.

[64] Y. Yang, J. Feng, N. Jojic, J. Yang, and T. S. Huang, "ℓ^0-sparse subspace clustering," in *Eur. Conf. Comput. Vis.*, 2016, pp. 731–747.

[65] W. Liu, X. Shen, and I. W. Tsang, "Sparse embedded k-means clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3319–3327.

[66] 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. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3438–3444.

[67] Z. Huang, P. Hu, J. T. Zhou, J. Lv, and X. Peng, "Partially view-aligned clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020.

[68] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[70] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
[71] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. C. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3165–3174.
[72] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.
[73] R. Azad, M. Asadi-Aghbolaghi, S. Kasaei, and S. Escalera, "Dynamic 3d hand gesture recognition by learning weighted depth motion maps," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1729–1740, 2018.
[74] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 1881–1887.
[75] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2408–2414.
[76] Y. Liu, L. Wang, Y. Bai, C. Qin, Z. Ding, and Y. Fu, "Generative view-correlation adaptation for semi-supervised multi-view learning," in *Eur. Conf. Comput. Vis.*, 2020, pp. 318–334.

Yijie Lin received the bachelor's degree in computer science and technology from Sichuan University, Chengdu, China, in 2020. He is currently pursuing the PhD degree with the School of Computer Science, Sichuan University. His research interest includes multi-modal learning.

Yuanbiao Gou Yuanbiao Gou received the B.E. degree from the School of Software Engineering, Sichuan University, Chengdu, China, in 2019. He is currently pursuing the PhD degree with the School of Computer Science, Sichuan University. His current research interest is image processing.

XiaotianLiu Xioatian Liu received the B.S. degree in Mathematics and Computer Science from Wake Forest University in 2020. He is currently pursuing the M.S. degree at the Department of Computer Science, Wake Forest University. His current research interest includes partial multi-view learning and tensor decompositions.

Jinfeng Bai received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2015. He is currently working in Tomorrow Advancing Life Education Group (TAL) and is in charge of the AI laboratory. His research interests cover speech recognition, speech synthesis, image text recognition, sequence learning, natural language processing, computer vision and other fields.

Jiancheng Lv (Member, IEEE) received the Ph.D. degree in computer science and engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2006. He is currently a Professor with the Data Intelligence and Computing Art Laboratory, College of Computer Science, Sichuan University, Chengdu. His research interests include neural networks, machine learning, and big data.

Xi Peng is currently a full professor at College of Computer Science, Sichuan University. His current interests mainly focus on machine learning and multi-media analysis. On these areas, he has authored more than 70 articles in JMLR, TPAMI, IJCV, ICML, NeurIPS, and so on. Dr. Peng has served as an Associate Editor for four journals such as " IEEE Trans on SMC: Systems", a Guest Editor for four journals such as "IEEE Trans. on Neural Network and Learning Systems".

Supplementary Material of Dual Contrastive Prediction for Incomplete Multi-view Representation Learning

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, Xi Peng

In this supplementary material, we present the additional experimental results and new visualization results to further support our theoretical conclusions.

Appendix A
Experiments on Clustering and Classification

In this section, we present the clustering and classification results of the LandUse-21 and Scene-15 datasets. As shown in Fig. 9 and Fig. 10, DCP has achieved the superior performance in almost all settings over existing approaches.

Appendix B
Complementarity between $\mathcal{L}_{pre}$ and $\mathcal{L}_{icl}$

During training, the dual prediction loss $\mathcal{L}_{pre}$ overcomes the data recovery problem through a instance-level formulation similar to contrastive learning. A natural question to ask is whether $\mathcal{L}_{pre}$ is redundant to our model when $\mathcal{L}_{icl}$ can help to learn consistent representations already? In short, the dual prediction loss $\mathcal{L}_{pre}$ is complementary to $\mathcal{L}_{icl}$, which is indispensable to our model. In the following, we will theoretically show and experimentally verify the necessity of $\mathcal{L}_{pre}$.

From information theory: It should be pointed out that the prediction loss $\mathcal{L}_{pre}$ is proposed to recover the missing views through minimizing the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$, and the instance-level contrastive loss $\mathcal{L}_{icl}$ is designed for the consistency learning through maximizing mutual information $I(\mathbf{Z}^1;\mathbf{Z}^2)$. In short, $\mathcal{L}_{pre}$ and $\mathcal{L}_{icl}$ are specific training losses of the general objective function Eq. (3):

$$\begin{aligned} &\max I\left(\mathbf{Z}^1;\mathbf{Z}^2\right) \\ &\text{s.t. } \min H\left(\mathbf{Z}^1|\mathbf{Z}^2\right),\ \min H\left(\mathbf{Z}^2|\mathbf{Z}^1\right) \end{aligned} \tag{3}$$

As shown in Theorem 1 of manuscript, the cross-view consistency learning and cross-view data recoverability are equivalent from the information-theoretic point of view. More specifically, the cross-view consistency and data recoverability could be treated as two sides of one coin. On the one hand, the data recoverability could be further improved because maximizing $I(\mathbf{Z}^i;\mathbf{Z}^j)$ will increase the view-shared information. On the other hand, the view-inconsistent information will be discarded throughout minimizing $H(\mathbf{Z}^i|\mathbf{Z}^j)$, thus improving the consistency. Moreover, Theorem 2 has shown that optimizing Eq. (3) could achieve sufficient and minimal multi-view representations. To sum up, $\mathcal{L}_{pre}$ and $\mathcal{L}_{icl}$ are complementary to each other and will mutually boost.

From experimental results: To verify the above theoretical results, we investigate the impact of $\mathcal{L}_{pre}$ towards $\mathcal{L}_{icl}$ by ablating $\mathcal{L}_{pre}$. In detail, we use the same network and balance parameters described in Section 4.2, and test it in the following two settings, *i.e.,* missing rate $\eta = 0.5$ (marked as Incomplete) and $\eta = 0$ (marked as Complete). As shown in Table 6, $\mathcal{L}_{pre} + \mathcal{L}_{icl}$ is consistently superior to $\mathcal{L}_{icl}$ in both settings, verifying our claim that $\mathcal{L}_{pre}$ is complementary to $\mathcal{L}_{icl}$.

Appendix C
Experiments on Unbalanced Multi-view Data

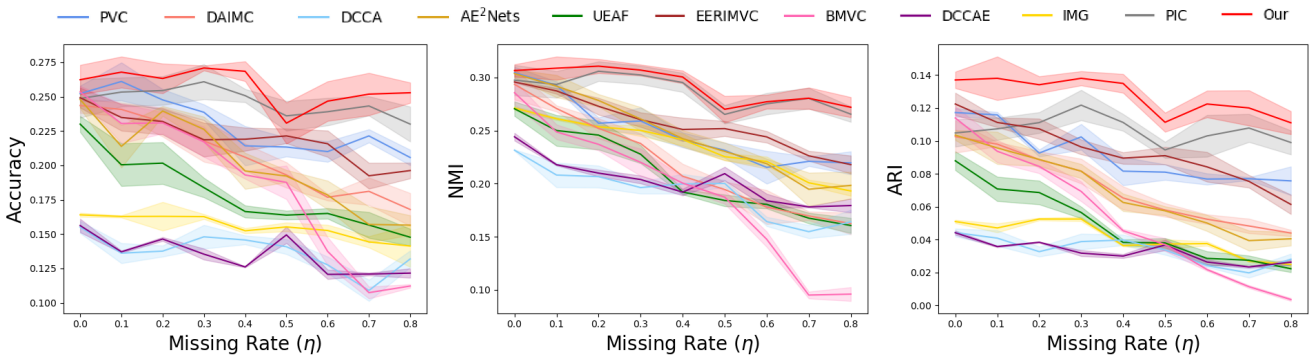
Real world data gathered from multiple sensors may exhibit different levels of view quality, causing degradation in the model's performance. In this section, we show that utilizing dual prediction loss $\mathcal{L}_{pre}$ can be helpful to address the problem of unbalanced view quality.

When one view is of low quality, *e.g.,* $\mathbf{X}^2$ is noisy, one could assume that $\mathbf{X}^2$ contains more task irrelevant information $H(\mathbf{X}^2|\mathbf{Y})$ (*i.e.,* A2 $\cup$ A5 in Fig 1). As the dual prediction loss $\mathcal{L}_{pre}$ is designed for data recovery through minimization of the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$, most task-irrelevant information $H(\mathbf{X}^2|\mathbf{X}^1)$ (*i.e.,* A5 in Fig 1) will be discarded via optimizing $\mathcal{L}_{pre}$. Meanwhile, as indicated by Theorem 1 and the above experiments, $\mathcal{L}_{pre}$ is beneficial to consistency learning and performance improvement.

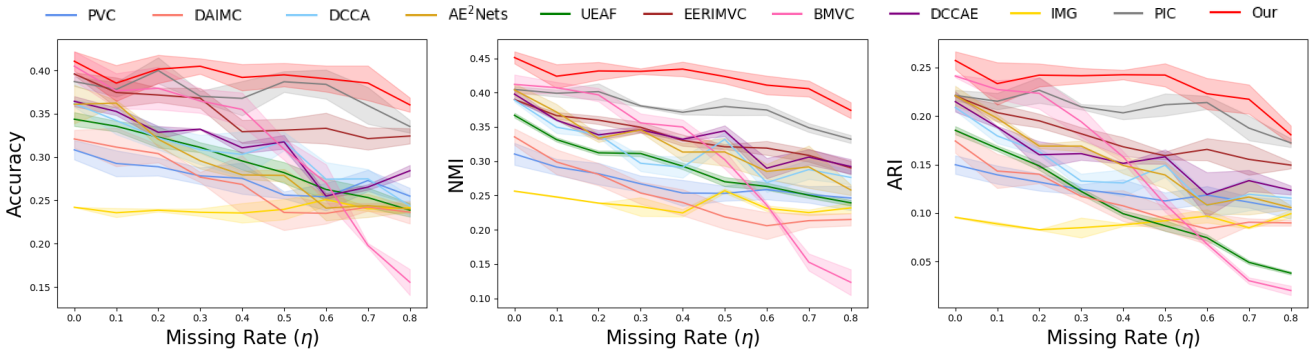
To assess our method's capacity for this problem, following CoMVC [1], we corrupt the USPS view and Edge view in MNIST-USPS and E-MNIST with additive Gaussian noise and record the models' performance in terms of clustering as the standard deviation of the noise increases.

1) E-MNIST dataset, derived from the MNIST dataset, contains the original hand-written digit and an edge-detected feature, respectively. Following [1], we use the training set containing 50,000 images. The MNIST and Edge images are with 784 (28×28) pixels. Note that E-MNIST naturally contains unbalanced view information by design.

2) MNIST-USPS dataset. Following [2], we use raw images from the MNIST and USPS datasets as two different views. For each dataset, we randomly selected 5,000 samples distributed over 10 digits to constitute the MNIST-USPS dataset. A single MNIST

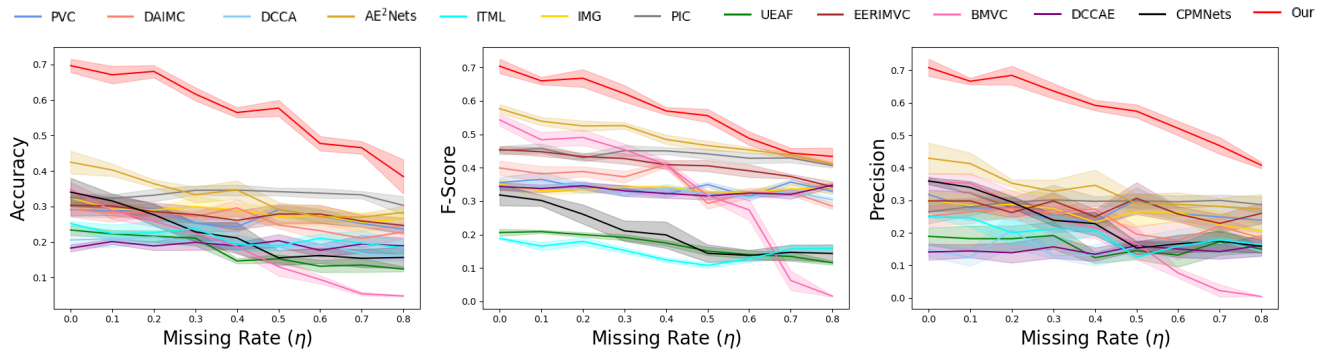


(a) LandUse-21

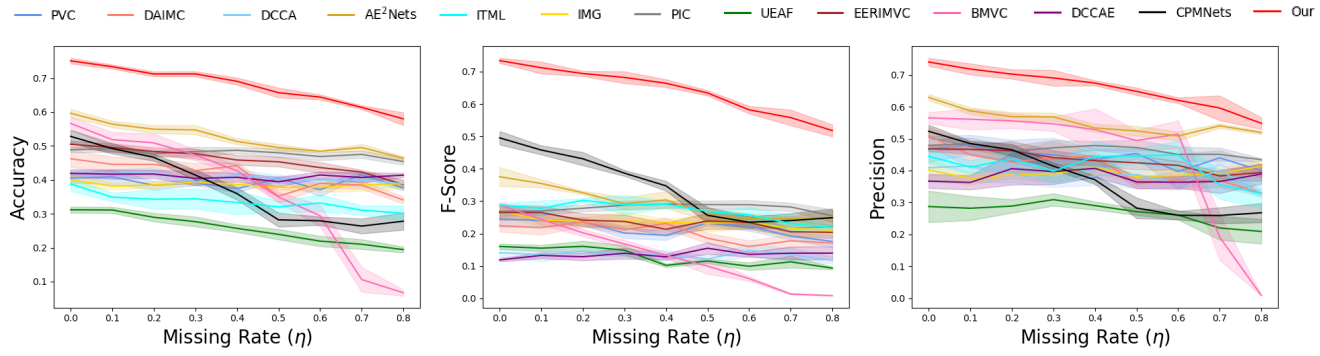


(b) Scene-15

Fig. 9: Clustering performance comparisons on (a) LandUse-21 and (b) Scene-15 with different missing rates (η).

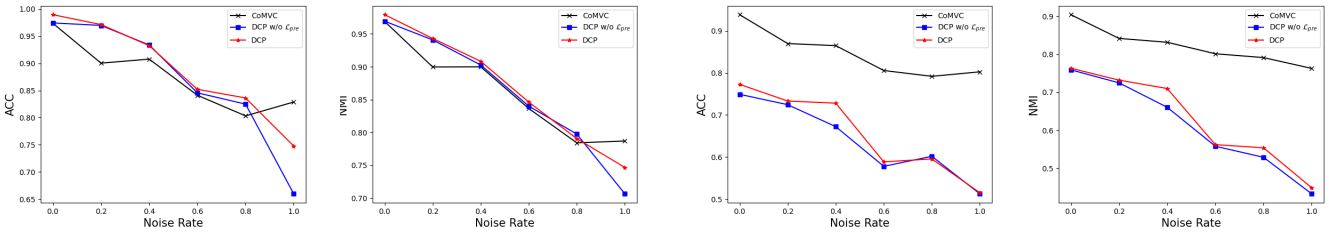


(a) LandUse-21



(b) Scene-15

Fig. 10: Classification performance comparisons on (a) LandUse-21 and (b) Scene-15 with different missing rates (η).



(a) MNIST-USPS (b) E-MNIST

Fig. 11: Clustering ACC and NMI on MNIST-USPS and E-MNIST, with increasing levels of Gaussian noise added to the second view.

TABLE 6: Clustering performance by ablating $\mathcal{L}_{pre}$. As shown, $\mathcal{L}_{pre}$ improves the performance of $\mathcal{L}_{icl}$ in two different settings on three datasets.

Missing	Methods	Caltech101-20			Scene-15			LandUse-21		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Incomplete	$\mathcal{L}_{icl}$	46.69	58.03	41.86	38.07	40.62	21.30	16.80	22.46	5.75
	$\mathcal{L}_{pre} + \mathcal{L}_{icl}$	**64.59**	**62.11**	**71.07**	**39.49**	**42.08**	**22.92**	**20.34**	**26.89**	**10.14**
Complete	$\mathcal{L}_{icl}$	70.33	68.07	**78.39**	40.87	43.79	23.29	25.31	29.91	13.86
	$\mathcal{L}_{pre} + \mathcal{L}_{icl}$	**70.79**	**68.38**	77.87	**41.28**	**44.27**	**24.16**	**26.57**	**30.22**	**14.32**

image consists of 784 (28×28) pixels, and a single USPS image consists of 256 (16×16) pixels.

For fair comparisons, we use the same networks described in Section 4.2, except the encoder for E-MNIST, where the same network of CoMVC [1] is used. The dimension D of latent representation is set to 128 and 64 for E-MNIST and MNIST-USPS, respectively. We maintain the same trade-off parameters as described in the manuscript.

As shown in Fig. 11, DCP performs better than DCP w/o $\mathcal{L}_{pre}$ in most cases, which verifies our claim that the dual prediction loss $\mathcal{L}_{pre}$ is of use to unbalanced multi-view problems. Admittedly, the performance of DCP is worse than CoMVC [1] on the E-MNIST dataset since the primary focus of DCP is different from CoMVC. To be specific, CoMVC is designed to solve this particular problem by learning a fusion weight designated for balancing the importance of each view. In contrast, our DCP is proposed to address the incomplete multi-view problem. Surprisingly, despite the mentioned "unfairness" in the comparison, DCP still achieves a slightly better performance than CoMVC on MNIST-USPS in almost all noise cases.

APPENDIX D
VISUALIZATION ON REPRESENTATIONS

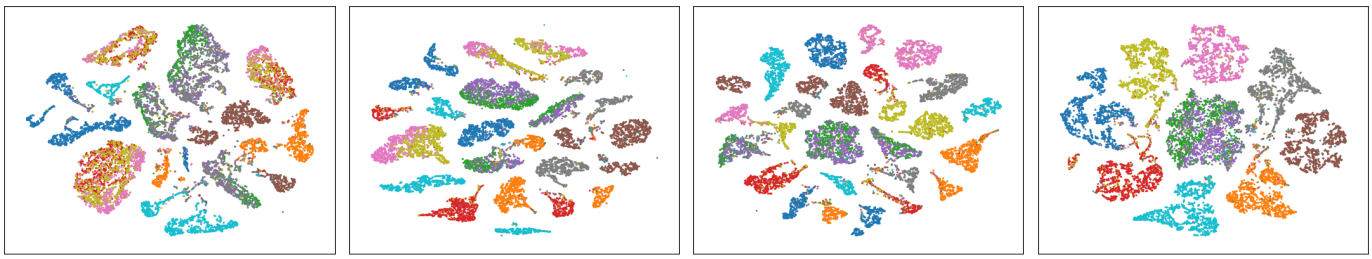
For a more comprehensive study, we employ t-SNE [3] to visually illustrate the learned representation on the Noisy MNIST dataset with the dimensionality of two. As shown in Fig. 12, the representations learned by DCP become more compact and discriminative with increasing training epochs when missing rate $\eta = 0.5$. Furthermore, we visually investigate all methods in the following two settings, *i.e.*, $\eta = 0.5$ (Fig. 13(a)) and $\eta = 0$ (Fig. 13(b)). From the results, one could observe that the representations learned by DCP are more discriminative than other methods.

REFERENCES

[1] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1255–1265.

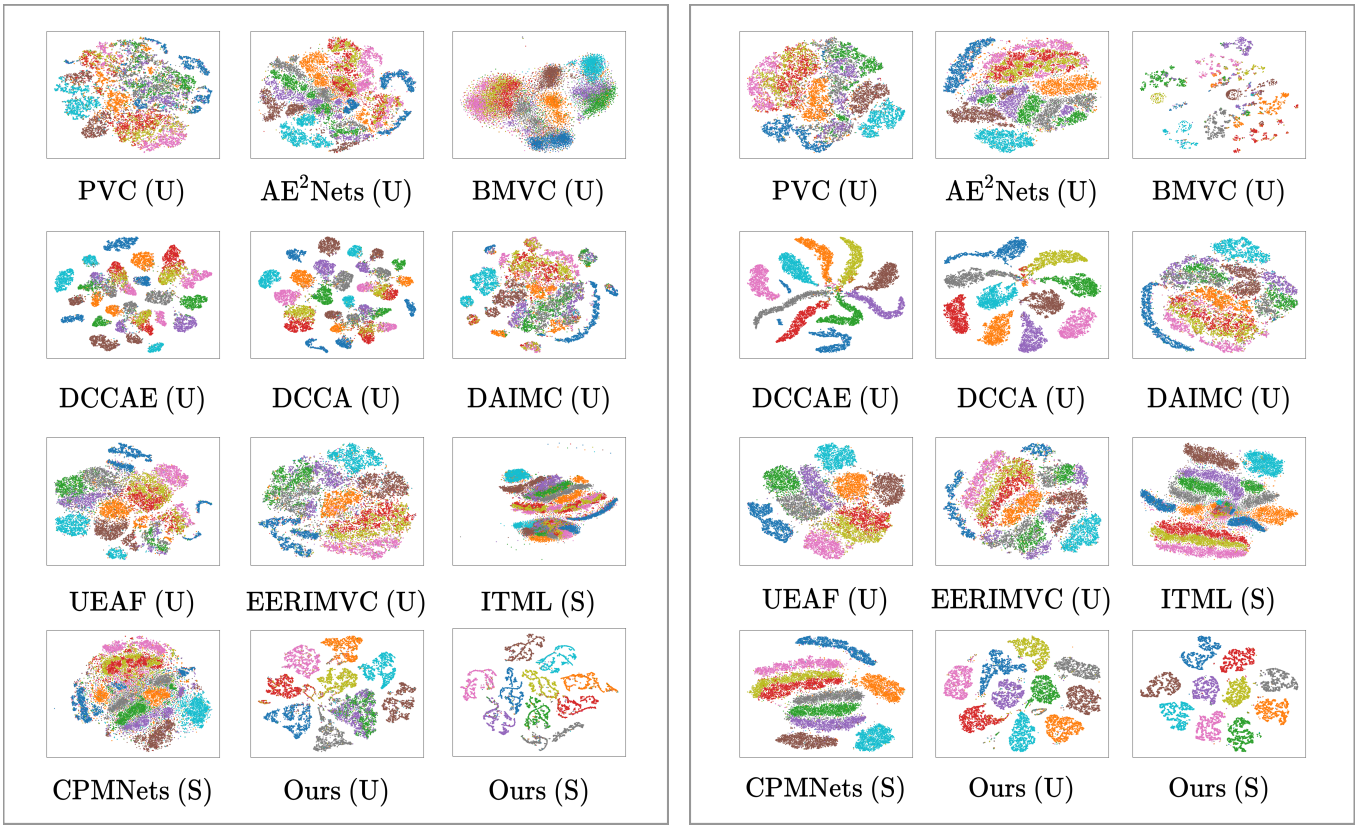
[2] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "COMIC: multi-view clustering without parameter selection," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 5092–5101.

[3] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.



(a) 20 epoch (NMI = 0.567) (b) 50 epoch (NMI = 0.651) (c) 100 epoch (NMI = 0.707) (d) 200 epoch (NMI = 0.759)

Fig. 12: t-SNE visualization on the Noisy MNIST dataset with increasing training iteration.



(a) Incomplete (b) Complete

Fig. 13: t-SNE visualization on the Noisy MNIST dataset. 'U' and 'S' denote unsupervised and supervised learning paradigm, respectively.