Incomplete Multi-view Clustering via Prototype-based Imputation

Haobin Li[†], **Yunfan Li**[†], **Mouxing Yang**, **Peng Hu**, **Dezhong Peng** and **Xi Peng**[*]
College of Computer Science, Sichuan University
{haobinli.gm, yunfanli.gm, yangmouxing, penghu.ml}@gmail.com, pengdz@scu.edu.cn, pengx.gm@gmail.com

Abstract

In this paper, we study how to achieve two characteristics highly-expected by incomplete multi-view clustering (IMvC). Namely, i) instance commonality refers to that within-cluster instances should share a common pattern, and ii) view versatility refers to that cross-view samples should own view-specific patterns. To this end, we design a novel dual-stream model which employs a dual attention layer and a dual contrastive learning loss to learn view-specific prototypes and model the sample-prototype relationship. When the view is missed, our model performs data recovery using the prototypes in the missing view and the sample-prototype relationship inherited from the observed view. Thanks to our dual-stream model, both cluster- and view-specific information could be captured, and thus the instance commonality and view versatility could be preserved to facilitate IMvC. Extensive experiments demonstrate the superiority of our method on five challenging benchmarks compared with 11 approaches. The code could be accessed from https://pengxi.me.

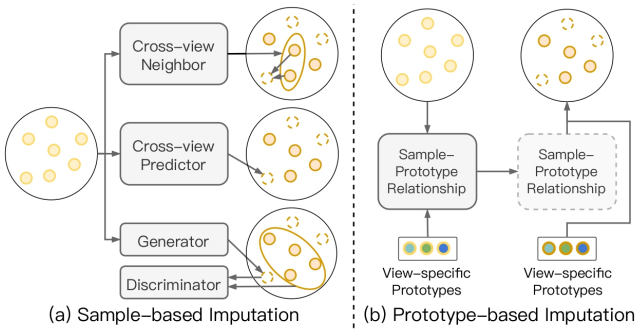


Figure 1: Our basic idea. (a) Three typical sample-based data recovery paradigms in existing IMvC studies, namely, i) neighborhood-based recovery, ii) cross-view prediction, and iii) adversarial generation. One limitation of the three paradigms is that two highly-expected characteristics in IMvC, *i.e.*, instance commonality and view versatility, are not fully explored. (b) The proposed prototype-based imputation paradigm. In brief, the data is recovered using the sample-prototype relationship inherited from the observed view and the prototypes from the missing view. Such a paradigm could recover both cluster- and view-specific information, thus preserving instance commonality and view versatility.

1 Introduction

Clustering is a fundamental tool in data analysis [Van Gansbeke *et al.*, 2020; Li *et al.*, 2021; Li *et al.*, 2022], which aims at partitioning instances into different clusters without the help of data annotations. To handle multi-view data, many efforts have been devoted to developing multi-view clustering (MvC) methods [Tao *et al.*, 2017; Hu *et al.*, 2019; Huang *et al.*, 2019; Yang *et al.*, 2021a; Yang *et al.*, 2021b]. Almost all of MvC works implicitly or explicitly take the data completeness assumption, *i.e.*, all instances exist in all views. In practice, however, the assumption is always violated due to the complexity of data collection and transmission, leading to the incomplete problem in multi-view data. For example, when building medical history for patients, the multi-view healthcare data is susceptible to be incomplete due to disease concealment in the data collection or information loss during treatment transfer.

To achieve incomplete multi-view clustering (IMvC), a feasible solution is employing the observed cross-view samples to recover the missing counterparts and then performing clustering. As shown in Fig 1(a), one of the most straightforward paradigms is using observed samples to find cross-view neighbors which are further used to recover missing samples. Such a paradigm implicitly assumes that the views could be mapped into a common space wherein the neighbors of the missing sample could be accurately identified by its cross-view counterpart. In practice, however, such an assumption is satisfied always at the cost of the *view versatility* since the view-specific information is often excluded to learn the common space. To compensate for view versatility, some studies propose capturing the view-specific information using a cross-view predictor [Lin *et al.*, 2021] or generator [Wang *et al.*, 2018]. Unfortunately, such a generative paradigm essentially learns an equivalent mapping for the whole dataset across views, which will lose the *instance commonality*, *i.e.*, within-cluster compactness and between-cluster scatterness.

Different from the aforementioned sample-based imputation methods, we propose a prototype-based imputation paradigm as shown in Fig. 1(b). Unlike existing methods that

restore the missing sample through learning a common representation for cross-view samples, we propose performing data recovery using the prototypes from the missing view and the sample-prototype relationship from the observed view. Thanks to our paradigm, the instance commonality and the view versatility can be preserved because the prototypes capture the cluster- and view-specific information. Furthermore, our invariance assumption on the sample-prototype relationship is milder than that on the cross-view representation taken in these works.

To implement the prototype-based imputation, ones have to overcome the following two technical challenges, *i.e.*, i) incorporating prototypes and samples to enhance the instance commonality, and ii) learning view-specific prototypes to preserve view versatility. To this end, we propose an incomplete multi-view clustering method based on a novel dual-stream model consisting of a dual attention layer and a dual contrastive learning loss. To be specific, the dual attention layer aims to enhance the instance commonality by representing samples and prototypes with each other. More specifically, the sample representation is learned by aggregating the sample itself and the corresponding prototype, thus enhancing the commonality of with-cluster instances. In a dual manner, the prototype representation is learned through aggregating prototype itself and the current input samples, thus integrating the historical and current information. The dual contrastive learning loss is designed to preserve view versatility, which consists of the standard contrastive learning on samples and a new bounded contrastive loss on the prototypes. Thanks to the bounded contrastive loss, the prototypes will embrace the unique view-specific information, thus preserving the view versatility. The major contributions of this paper could be summarized as follows:

1. From the standpoint of data recovery for IMvC, we proposed a novel imputation method which restores the missing samples using the prototypes and the sample-prototype relationship. Such a prototype-based imputation paradigm could preserve instance commonality and view versatility that are favorites to IMvC.
2. From the standpoint of unsupervised multi-view representation learning, we propose a novel dual-stream model which learns sample representation using prototypes and prototype representation using the input samples. Thanks to the dual-stream model, our method could learn better representation for boosting IMvC performance.
3. Extensive experiments on five benchmarks demonstrate the superiority of our method in both incomplete multi-view clustering and data recovery performance, compared with 11 baselines.

2 Related Work

In this section, we briefly review two related topics, namely, incomplete multi-view clustering and attention-based model.

2.1 Incomplete Multi-View Clustering

IMvC is a long-standing task in the multi-view learning community, which has attracted numerous studies. Based on the way to utilize the cross-view information, classic IMvC methods could be divided into three categories, including matrix factorization based [Li *et al.*, 2014; Zhao *et al.*, 2016; Shao *et al.*, 2015; Hu and Chen, 2019], kernel learning based [Bach and Jordan, 2002; Liu *et al.*, 2020], and similarity relation based [Wang *et al.*, 2019; Liu *et al.*, 2019]. To handle more complex and large-scale data, several deep IMvC methods have been developed recently. Based on the paradigm of recovering the missing data, deep IMvC methods could be divided into three categories, including i) neighborhood-based methods [Tang and Liu, 2022; Yang *et al.*, 2022b], which impute the missing data with the help of cross-view nearest neighbors, ii) predictor-based methods [Lin *et al.*, 2021; Lin *et al.*, 2022], which learn a direct mapping from observed views to missing views for data recovery, and iii) GAN-based methods [Wang *et al.*, 2018; Jiang *et al.*, 2019; Zhang *et al.*, 2020], which recover the missing data through adversarial generation.

Among the above works, deep IMvC methods are most similar to this study. However, this study is remarkably different from existing works in the following aspects. First, the existing works impute data based on the observed counterparts which might discard either instance commonality or view versatility. In contrast, the proposed prototype-based imputation paradigm performs recovery using the prototypes in the missing view and the sample-prototype relationship in the observed view, thus taking the best of both worlds. Second, to the best of our knowledge, this could be the first attention-based model in the IMvC community, showing its great potential in unsupervised data recovery and IMvC.

2.2 Attention-Based Model

The attention-based model learns better representation by focusing on regions with relevant information, which has achieved great success in various tasks such as image classification [Yu *et al.*, 2018], person re-identification [Yang *et al.*, 2022a], object detection [Woo *et al.*, 2018], neural machine translation [Vaswani *et al.*, 2017], and sentence summarization [Rush *et al.*, 2015]. Recently, some works have explored the attention mechanism in multi-view learning. For example, [Qu *et al.*, 2017] promotes using attention to collaborate different views for multi-view representation learning. [Zhou and Shao, 2018] proposes a viewpoint-aware attention model for vehicle re-identification. [Luo *et al.*, 2020] implements attention-enhanced matching confidence volume in multi-view stereo. [Yan *et al.*, 2022] introduces lateral connections by cross-view attention, and fuses multi-view information for video recognition.

The major differences between this work and previous attention-based models lie in two aspects. First, different from most existing works that focus on single-stream and instance-wise attention, the proposed dual-stream model employs a novel dual attention layer to incorporate samples and learnable prototypes with each other. Second, unlike most existing works that solely use attention for general multi-view feature fusion, the proposed dual attention layer is IMvC-oriented, which simultaneously facilitates clustering-favorable feature extraction and data recovery.

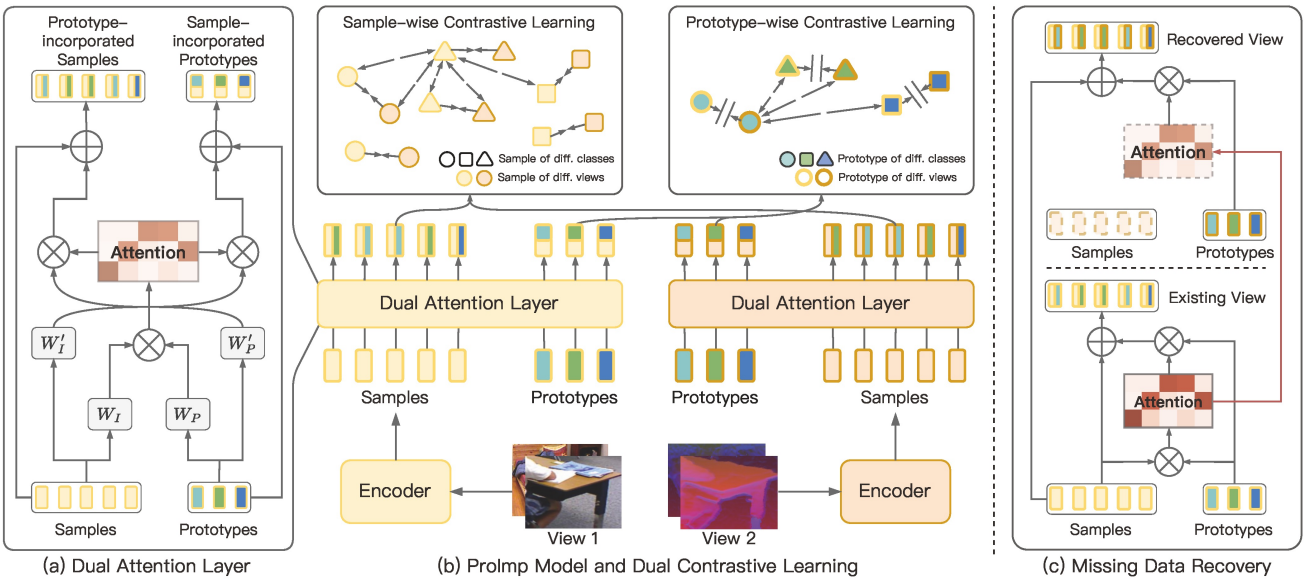


Figure 2: Overview of our ProImp method. (a) The dual attention layer. The attention is computed between samples and prototypes to incorporate each other. On the one hand, the sample representation aggregates the sample itself and the corresponding prototype, thus enhancing the commonality of with-cluster instances. On the other hand, the prototype representation aggregates the prototype itself and the current input samples, thus integrating the historical and current information. (b) The ProImp model and dual contrastive learning objective. To optimize the entire model as well as learnable view-specific prototypes, in addition to conducting standard contrastive learning on samples, we dually contrast prototypes with a new bounded contrastive loss to preserve view versatility. (c) The prototype-based missing data recovery. The missing samples are recovered with the attention inherited from the observed view and prototypes in the missing view, which enjoys both instance commonality and view versatility. Meanwhile, samples from the observed view are skip-connected to introduce instance consistency in the recovered data.

3 Method

In this section, we propose a dual-stream model dubbed ProImp to achieve incomplete multi-view clustering. As illustrated in Fig. 2, ProImp is composed of a dual attention layer to model the relationship between samples and prototypes, as well as a dual contrastive learning loss to learn attention and view-specific prototypes. For data recovery, ProImp adopts the prototype-based imputation paradigm to preserve instance commonality and view versatility. In the following, we first introduce our dual attention layer in Sec. 3.1, then elaborate on the dual contrastive learning loss in Sec. 3.2, and finally present the prototype-based imputation paradigm in Sec. 3.3.

3.1 Dual Attention Layer

Without loss of generality, we take bi-view data as an example for clarity. Let $\mathbf{X} = \{\mathbf{X}^{1,2}, \mathbf{X}^1, \mathbf{X}^2\}$ be an incomplete multi-view dataset, where $\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^{1,2}$ refer to three subsets of instances that have data observed in the first, the second, and both views. We denote the set of N complete instances as $\mathbf{X}^{1,2} = \{X^1, X^2\}$, where $X^v = \{x_1^v, x_2^v, \ldots, x_N^v\}$ denotes the samples in the v-th view.

As illustrated in Fig. 2, the dual attention is computed between samples X^v and a set of learnable prototypes $C^v = \{c_1^v, c_2^v, \ldots, c_K^v\}$, where K corresponds to the target cluster number. Mathematically, the attention A^v is computed through

$$A^v = \text{Softmax}\left((W_I^v X^v)^T W_P^v C^v / \sqrt{d}\right), \tag{1}$$

where W_I^v and W_P^v are two linear layers for samples and prototypes in $v-$th view respectively, and d is the dimension of features.

The attention A^v is then used to incorporate samples and prototypes in a dual manner. For sample representation, the corresponding prototype is aggregated to each sample, namely,

$$Z^v = X^v + W'^v_P C^v \left(A^v\right)^T, \tag{2}$$

where $Z^v = \{z_1^v, z_2^v, \ldots, z_N^v\}$ is the new representation of samples, and W'^v_P is another linear layer for prototypes. Such behavior intrinsically pulls each sample to its corresponding cluster center, thus enhancing the instance commonality favored in clustering.

Likewise, for prototype representation, features of current samples would be aggregated into prototypes, namely,

$$U^v = C^v + W'^v_I X^v A^v, \tag{3}$$

where $U^v = \{u_1^v, u_2^v, \ldots, u_K^v\}$ is the new representation of prototypes, and W'^v_I is another linear layer for samples. Such behavior enables prototypes to integrate the historical and current cluster information.

Notably, an encoder network is adopted to extract the features of samples in each view before feeding them to the dual attention layer. Here we omit it in mathematical notations for simplicity.

3.2 Dual Contrastive Learning

As discussed above, the dual attention layer outputs prototype-incorporated samples and sample-incorporated

prototypes in each view. To optimize the entire model and learnable view-specific prototypes, we conduct dual contrastive learning on samples and prototypes, respectively.

Sample-Wise Contrastive Learning

To mine instance consistency between cross-view samples, we adopt the following contrastive loss that maximizes the similarities between cross-view samples of the same instance, while minimizing those between samples of different instances, namely,

$$\mathcal{L}_S = \frac{1}{2N}\sum_{i=1}^{N}\left(\mathcal{L}_i^{1,2} + \mathcal{L}_i^{2,1}\right), \tag{4}$$

$$\mathcal{L}_i^{1,2} = -\log\frac{e^{s\left(z_i^1, z_i^2\right)/\tau_I}}{\sum_{j=1}^{N}\left[e^{s\left(z_i^1, z_j^1\right)/\tau_I} + e^{s\left(z_i^1, z_j^2\right)/\tau_I}\right]}, \tag{5}$$

where $s(\cdot,\cdot)$ denotes the cosine similarity, $\tau_I = 0.5$ is the temperature parameter, and $\mathcal{L}_i^{2,1}$ is defined similarly as $\mathcal{L}_i^{1,2}$.

Prototype-Wise Contrastive Learning

As discussed, our prototype-based imputation paradigm requires prototypes to capture view versatility. In other words, prototypes from different views should not collapse into an identical representation. To this end, instead of simply maximizing the similarities between cross-view prototypes of the same cluster, we propose to optimize their similarities to a bound. Meanwhile, to achieve a more distinct clustering, we minimize the similarities between prototypes of different clusters, which leads to the following bounded contrastive loss,

$$\begin{aligned}\mathcal{L}_P =& \frac{2}{K}\sum_{i=1}^{K}\frac{\left|s\left(u_i^1, u_i^2\right) - \alpha\right|}{\tau_P} + \frac{1}{K}\sum_{\substack{v_1=1\\v_2=1}}^{2}\sum_{i=1}^{K}\\ &\log\left[e^{\left|s\left(u_i^{v_1}, u_i^{v_2}\right)-\alpha\right|/\tau_P} + \sum_{j=1, j\neq i}^{K} e^{s\left(u_i^{v1}, u_j^{v2}\right)/\tau_P}\right],\end{aligned} \tag{6}$$

where α denotes the similarity bound and $\tau_P = 2.0$ is the temperature parameter.

Attention Regularization

Recall that the dual attention A^v defined in Eq. 1 is an $N \times K$ matrix, where A_{ij}^v intrinsically corresponds to the probability of the i-th sample belonging to the j-th cluster. To achieve more distinct clustering, we expect each sample to be confidently assigned to a certain cluster. Meanwhile, we should prevent the trivial solution where most samples are assigned to the same cluster. For these purposes, we propose the following attention regularization term, namely,

$$\mathcal{L}_R = \sum_{v=1}^{2}\sum_{j=1}^{K}\left[A_{\cdot j}^v \log A_{\cdot j}^v - \beta\sum_{i=1}^{N} A_{ij}^v \log A_{ij}^v\right], \tag{7}$$

where $A_{\cdot j}^v = \sum_{i=1}^{N} A_{ij}^v$ and β is the weight parameter to balance the sharpness and uniformity of attention.

Combining the dual contrastive learning loss and the attention regularization term, we arrive at the overall objective function of the proposed ProImp, namely,

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_P + \mathcal{L}_R. \tag{8}$$

3.3 Prototype-Based Imputation

To recover the missing samples in the incomplete data $\{\mathbf{X}^1, \mathbf{X}^2\}$, we design the following prototype-based imputation strategy as illustrated in Fig. 2(c). Specifically, given data $\mathbf{X}^1$ observed in view 1, the missing data in view 2 is recovered with attention A^1 and prototypes C^2 through

$$\hat{Z}^2 = X^1 + W'^2_P C^2 \left(A^1\right)^T, \tag{9}$$

where A^1 is the dual attention computed in view 1 according to Eq. 1, and $\hat{Z}^2$ is the recovered data in view 2. The idea behind such an attention inheritance is that the instance semantics are expected to be consistent across different views. By incorporating cluster- and view-specific information from prototypes, both instance commonality and view versatility could be preserved in the recovered data. In addition, samples from the observed view are skip-connected to the recovered data to introduce instance consistency.

Likewise, the missing data $\hat{Z}^1$ in view 1 is similarly recovered given data $\mathbf{X}^2$ observed in view 2. Let $\mathbf{Z}^1$ and $\mathbf{Z}^2$ denote the observed views in the incomplete data, the representation $\mathbf{Z} = \{\{Z^1, Z^2\}, \{\mathbf{Z}^1, \hat{Z}^2\}, \{\hat{Z}^1, \mathbf{Z}^2\}\}$ of both the observed and recovered data is concatenated and fed into the k-means algorithm to achieve clustering. Notably, though the attention itself intrinsically corresponds to the cluster assignment, it only utilizes data from a single view. Therefore, a simple concatenation operation is applied to gather multi-view information.

4 Experiment

In this section, we evaluate the proposed ProImp method on five widely-used multi-view datasets compared with 11 baselines. First, we present the experimental setting and implementation details in Sec. 4.1. Then, we compare our ProImp with state-of-the-art methods in Sec. 4.2. After that, we conduct the parameter analyses and ablation studies in Sec. 4.3. Finally, we present visualization results in Sec. 4.4.

4.1 Experimental Settings

Five multi-view datasets are used in our experiments, including Scene15 [Fei-Fei and Perona, 2005], Reuters [Amini *et al.*, 2009], NoisyMNIST [Wang *et al.*, 2015], CUB [Zhang *et al.*, 2019a], and MNIST-USPS [Peng *et al.*, 2019]. We randomly remove one view for m instances to simulate incomplete multi-view data with a missing rate of m/n, where n corresponds to the total number of instances.

The proposed ProImp is implemented in PyTorch 1.11.0 and all the experiments are conducted on an NVIDIA 3090 GPU on Ubuntu 20.04 OS. The model is trained for 150 epochs using the Adam optimizer with an initial learning rate of 1e-3, with a batch size of 1,024 on all datasets. The similarity bound α in Eq. 6 and the weight parameter in Eq. 7 are set to 0.75 and 0.02, respectively. In practice, we first warm

Setting	Method	Scene-15			Reuters			NoisyMNIST			CUB			MNIST-USPS		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Incomplete	DCCA	28.78	28.35	13.24	45.84	26.08	18.00	63.75	61.72	41.17	44.20	43.30	26.65	78.29	75.69	68.33
	DCCAE	29.01	29.13	12.86	47.04	28.00	14.48	65.42	62.87	38.32	42.33	40.87	25.46	79.53	79.19	68.40
	BMVC	32.45	30.87	11.56	32.10	6.98	2.89	30.71	19.16	10.60	29.79	20.28	6.35	43.90	39.00	21.00
	AE^2-Nets	22.44	23.43	9.56	29.08	7.55	4.84	29.88	23.78	11.81	35.87	32.00	15.90	40.90	29.30	19.70
	PMVC	25.47	25.37	11.31	29.32	7.42	4.42	33.13	25.49	14.62	57.73	54.37	38.29	60.50	47.10	39.80
	UEAF	28.95	26.92	8.37	33.32	20.06	12.19	37.45	34.42	25.71	45.80	45.25	26.88	63.32	58.86	49.23
	DAIMC	27.00	23.47	10.62	40.94	18.66	15.04	33.81	26.42	15.96	62.70	58.48	47.72	55.20	49.60	38.60
	EERIMVC	31.50	31.11	14.82	29.77	12.01	4.21	55.62	45.92	36.76	68.73	63.90	53.77	65.20	55.70	48.90
	COMPLETER	39.50	42.35	23.51	34.61	17.53	2.93	80.01	75.23	70.66	53.66	65.45	47.26	88.91	89.52	85.31
	SURE	39.60	41.58	23.49	47.18	30.89	23.32	92.34	84.99	84.31	58.33	50.37	37.44	92.34	84.99	84.31
	DSIMVC	30.56	35.47	17.24	39.87	19.61	17.13	57.47	55.12	44.08	54.57	51.35	35.04	96.71	91.82	92.98
	ProImp(Ours)	**41.58**	**42.86**	**25.31**	**51.89**	**35.54**	**28.53**	**94.86**	**87.43**	**89.08**	**73.30**	**66.38**	**54.84**	**96.81**	**91.85**	**93.06**
Complete	DCCA	36.61	39.20	21.03	47.95	26.57	12.71	89.64	88.33	83.95	55.60	56.11	43.18	87.19	91.65	86.73
	DCCAE	34.58	39.01	19.65	41.98	20.30	8.51	78.00	81.24	68.15	55.30	58.70	45.05	96.80	97.73	96.58
	BMVC	40.50	41.20	24.11	42.39	21.86	15.14	88.31	77.01	76.58	66.21	61.70	48.69	87.10	84.50	82.00
	AE^2-Nets	37.17	40.47	22.24	42.39	19.76	14.87	52.83	51.24	39.52	48.80	46.71	30.49	54.00	46.50	35.40
	PMVC	30.83	31.05	14.98	32.50	11.11	7.48	41.09	36.36	24.47	64.53	70.34	53.11	60.40	59.50	47.30
	UEAF	34.37	36.69	18.52	40.19	24.34	15.94	66.22	64.34	54.83	63.33	56.91	44.48	77.78	73.77	66.31
	DAIMC	32.09	33.55	17.42	40.78	21.15	15.98	38.40	34.66	22.98	71.57	70.69	57.89	65.10	65.50	54.20
	EERIMVC	39.60	38.99	22.06	33.21	14.28	3.90	65.66	57.60	51.34	74.00	73.05	62.41	79.00	68.10	62.40
	COMPLETER	41.07	44.68	24.78	36.20	18.87	4.75	89.08	88.86	85.47	63.57	70.18	52.92	94.46	95.63	93.20
	SURE	40.95	43.19	25.01	49.06	29.91	23.56	98.36	95.38	96.43	58.00	59.32	45.16	99.31	98.06	98.47
	DSIMVC	31.66	35.61	17.21	43.20	23.29	19.02	60.98	58.09	46.74	59.67	57.12	41.26	98.50	95.70	96.60
	ProImp(Ours)	**43.61**	**45.02**	**26.84**	**56.54**	**39.35**	**32.77**	**99.17**	**97.48**	**98.18**	**80.63**	**75.48**	**66.04**	**99.61**	**98.86**	**99.14**

Table 1: The clustering performance on four multi-view benchmarks. The best and second best results are denoted in **bold** and underline.

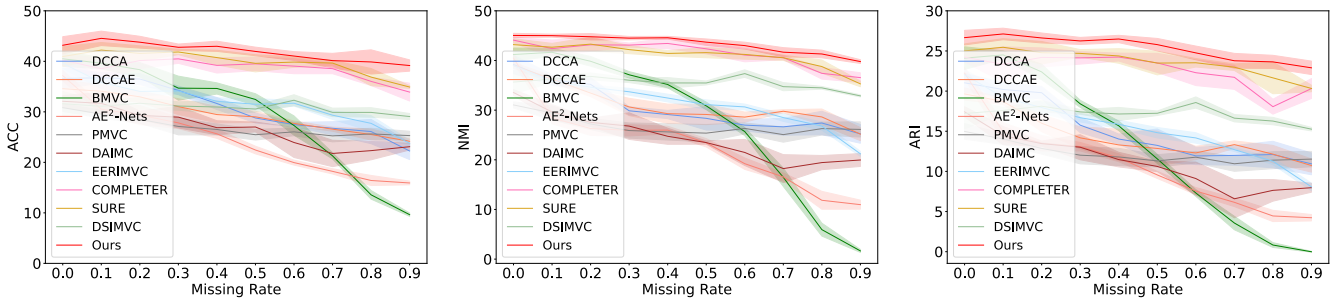


Figure 3: Clustering performance on Scene-15 under different missing rates. The colored regions denote the standard variances in five random experiments.

up the model with the sample-wise contrastive loss in Eq. 5 and the regularization term in Eq. 7 for 50 epochs. After that, we align the prototypes in different views with the Hungarian algorithm and train the model with the overall loss in Eq. 8 till the end.

Three widely-used metrics including Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) are used for evaluation. A higher value of these metrics indicates a better clustering performance.

4.2 Comparisons with State of the Arts

We compare ProImp with 11 state-of-the-art multi-view clustering baselines, including DCCA [Andrew *et al.*, 2013], DCCAE [Wang *et al.*, 2015], BMVC [Zhang *et al.*, 2018], AE^2-Nets [Zhang *et al.*, 2019b], PMVC [Li *et al.*, 2014], UEAF [Wen *et al.*, 2019], DAIMC [Hu and Chen, 2019], EERIMVC [Liu *et al.*, 2020], COMPLETER [Lin *et al.*, 2021], SURE [Yang *et al.*, 2022b], and DSIMVC [Tang and Liu, 2022].

We first evaluate ProImp and baselines under the Incomplete (with the missing rate of 50%) and Complete (with the missing rate of 0%) scenarios. Table 1 shows the average clustering performance under five random experiments. As can be seen, our ProImp significantly outperforms the state-of-the-art methods on all datasets. In particular, ProImp achieves a relatively 22% (28.53% v.s. 23.32%) and 39% (32.77% v.s. 23.56%) ARI improvement under the Incomplete and Complete scenarios on the Reuters dataset, compared with the second best method SURE. The superior performance demonstrates the effectiveness of the proposed dual-stream model, including the dual attention layer and dual contrastive learning objective.

We further explore the robustness of ProImp by increasing the missing rate from 0% to 90% with a gap of 10% on the Scene-15 dataset. Considering that the number of complete instances would be greatly reduced under large missing rates, we adjust the batch size to 128 and set the learning rate as 3e-4 in this experiment. As shown in Fig. 3, our ProImp substantially outperforms baselines under all missing rates. In addition, the performance of ProImp drops less as the miss rate increases. For example, in terms of ACC, ProImp outperforms SURE by 2.23% (43.18% v.s. 40.95%) under the com-

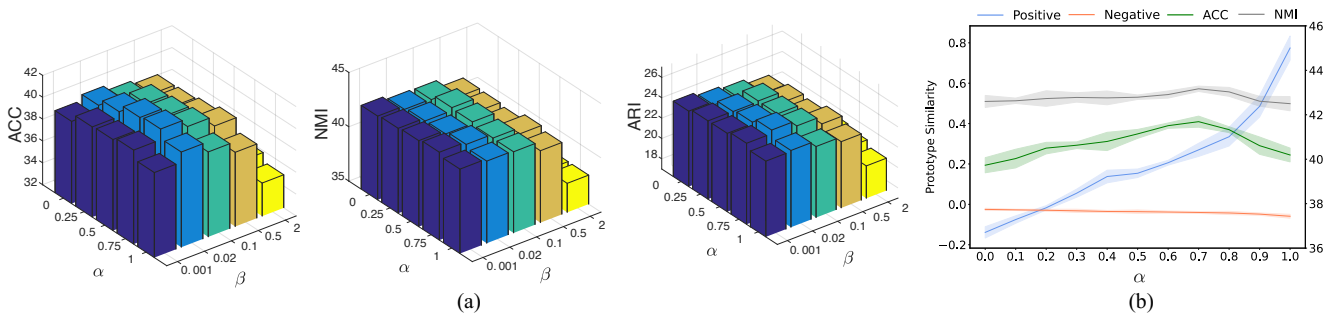


Figure 4: Parameter analyses on Scene-15. (a) The clustering performance of ProImp under different choices of the similarity bound α and the balance weight β. (b) The cross-view prototype similarity and clustering performance under different choices of α.

plete scenario, and the performance gap increases to 4.26% (39.17% v.s. 34.91%) under 90% missing rate. Such a result demonstrates the superiority of our attention imputation strategy for data recovery, as it could preserve both the view versatility and instance commonality information.

In this section, to better understand the effectiveness of each component in ProImp, we conduct a series of parameter analyses and ablation studies. In brief, we first investigate the influence of the similarity bound α in the prototype stream loss and the balance weight β in the attention regularization term. Then we perform ablation studies on each loss term. Finally, we test variant data recovery strategies.

Influence of Hyper-Parameters α and β

There are two hyper-parameters in the proposed ProImp, namely, the similarity bound α in the prototype stream loss and the balance weight β in the attention regularization term. To investigate how they influence the performance of ProImp, we change α in the range of $\{0, 0.25, 0.5, 0.75, 1\}$ and β in the range of $\{0.001, 0.02, 0.1, 0.5, 2\}$. As shown in Fig. 4a, ProImp achieves the best performance when $\alpha = 0.75$. According to the cross-view prototype similarity shown in Fig. 4b, positive prototype pairs get closer as α increases. An over-small cross-view prototype similarity would harm view consistency, and an over-large value would sacrifice the view versatility, both leading to inferior performance. As for the other parameter β, we find that ProImp achieves promising results under a reasonable range (*i.e.*, from 0.001 to 0.1). However, when the balance weight is too large, the attention between each instance and prototype tends to be equal. Such a collapsed attention would cause a significant performance drop.

Effectiveness of Each Loss Term

To explore the effectiveness of the proposed sample-wise contrastive loss, prototype-wise contrastive loss, and attention regularization, we conduct the ablation experiments on the three losses in Eq. 8. According to the results shown in Table 2, the regularization term $\mathcal{L}_R$ itself is not sufficient to learn appropriate attention. Both $\mathcal{L}_S$ and $\mathcal{L}_P$ could guide attention optimization, leading to better clustering performance. The best performance is achieved when all three losses are adopted, as both the instance commonality and view versatility are achieved.

$\mathcal{L}_S$	$\mathcal{L}_P$	$\mathcal{L}_R$	ACC	NMI	ARI
		✓	25.12	23.91	10.72
✓		✓	39.37	42.21	23.86
	✓	✓	27.73	25.29	12.20
✓	✓	✓	**41.58**	**42.86**	**25.31**

Table 2: Ablation study of three losses on Scene-15, where "✓" denotes the loss is adopted.

4.3 Parameter Analyses and Ablation Studies

Variants of Data Recovery Strategy

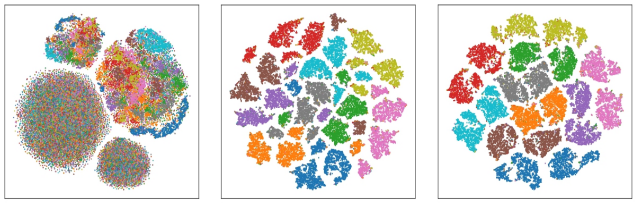
Recall that to preserve instance commonality and view versatility, we recover the missing view by the sample-prototype attention inherited from the observed view and prototypes from the missing view, namely, $\hat{Z}^2 = X^1 + W'^2_P C^2 \left(A^1\right)^T$ via Eq. 9. Here, to prove the superiority of our paradigm, we further investigate three other variants of recovery strategies on the Scene-15 dataset. Specifically,

- Prototypes from observed views: recovering by using the prototypes and sample-prototype attention from the observed view, *i.e.*, $\hat{Z}^2 = X^1 + W'^1_P C^1 \left(A^1\right)^T$;
- Prototypes from missing views only: recovering by only using prototypes from the missing view, *i.e.*, $\hat{Z}^2 = 2W'^2_P C^2 \left(A^1\right)^T$;
- Samples from observed views only: recovering by only using the observed cross-view counterparts, *i.e.*, $\hat{Z}^2 = 2X^1$;

Notably, as both sample and prototype features are L2 normalized, we scale the features of the last two variants to keep the length consistent. From the results in Table 3, one could have the following conclusions. First, replacing the missing view prototypes with those in the observed view would lose the view versatility, thus remarkably degrading the performance. Second, solely using prototypes or cross-view counterparts suffers from losing either cross-view consistency or instance commonality, resulting in sub-optimal results. In comparison, our default paradigm takes the best of both worlds, leading to the best performance.

Strategy	ACC	NMI	ARI
P. from observed views	32.23	34.44	17.73
P. from missing views only	36.32	39.06	21.14
S. from observed views only	40.17	41.43	23.66
Default	**41.58**	**42.86**	**25.31**

Table 3: Ablation study on different data recovery strategies on Scene-15. "P." denotes prototypes and "S." denotes samples.



(a) 0 epoch (NMI=0.065) (b) 50 epoch (NMI=0.805) (c) 150 epoch (NMI=0.872)

Figure 5: t-SNE visualization on the NoisyMNIST dataset across the training process.

4.4 Visualizations

In this section, we present two visualization results on the NoisyMNIST dataset with a missing rate of 0.5 to provide an intuitive understanding of the training process and data recovery performance of ProImp.

Features Learned by ProImp Across the Training Process

We conduct t-SNE visualization on features learned by ProImp at three different training epochs in Fig. 5. As can be seen, data forms four clusters at the initialization, which corresponds to the observed and recovered data from two views. After 50 epochs, data tends to form semantic clusters. However, as the prototypes are not yet matched across views, the recovered data is not semantically aligned with the observed data. At the end of the training, the gap between observed and recovered within-cluster samples is significantly narrowed, indicating a good instance commonality. Meanwhile, samples from different views are still not collapsed together, indicating that the view versatility is well preserved.

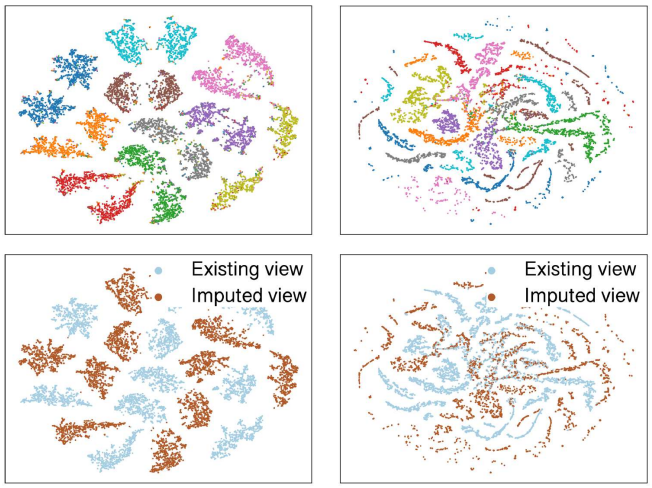
Data Recovery Performance

As discussed, a major advantage of our prototype-based imputation strategy is that it could preserve both instance commonality and view versatility in the recovered data. To prove its superiority, we visualize the observed and recovered data learned by our ProImp and the best competitor SURE in Fig. 6. From the results, one could see that i) data recovered by our ProImp forms more compact clusters, thanks to the dual-stream model which enhances the commonality between within-cluster instances, and ii) data recovered by our ProImp shows a more distinct pattern with observed data, which demonstrates that the view versatility is preserved from the view-specific prototypes.

To provide a quantitative evaluation of instance commonality, we compute the silhouette score on the data recovered by different paradigms. A larger value indicates better within-cluster compactness and between-cluster scatter. The results in Table 4 show that our prototype-based imputation strategy

Method	Silhouette Score
COMPLETER	58.52
SURE	63.31
ProImp w/o Prototype	44.08
ProImp	**65.45**

Table 4: Comparsions on instance commonality with different data recovery paradigms on NoisyMNIST.



(a) Ours (b) SURE

Figure 6: t-SNE visualization of the observed and recovered data on NoisyMNIST, compared with the best competitor SURE. Data are colored by classes and views in the first and second rows, respectively.

significantly enhances the instance commonality, surpassing existing generative and neighborhood-based paradigms.

5 Conclusion

To implement the proposed prototype-based imputation paradigm, we propose a dual-stream model by designing a dual attention layer and a dual contrastive learning loss. Thanks to the proposed model, the instance commonality and view versatility could be preserved into representation, thus boosting the IMvC performance. Extensive experiment results demonstrate the superiority of our model in both clustering and data recovery performance. In the future, we plan to extend ProImp to handle the datasets that consist of three and more views.

Acknowledgments

This work was supported by NSFC under Grant U21B2040, 62176171, U19A2078, and Sichuan Science and Technology Planning Project (Grant No. 2022YFS0128).

Contribution Statement

† Equal Contribution. ∗ Corresponding Author. Xi Peng conceived the study. Haobin Li and Yunfan Li designed the ProImp and are of equal contribution. Mouxing Yang

ran the baseline methods. Peng Hu and Dezhong Peng analyzed the results. All authors participated in the writing of the manuscript.

References

[Amini *et al.*, 2009] Massih R Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. *NeurIPS*, 2009.

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

[Bach and Jordan, 2002] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 2002.

[Fei-Fei and Perona, 2005] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[Hu and Chen, 2019] Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. *arXiv:1903.02785*, 2019.

[Hu *et al.*, 2019] Peng Hu, Dezhong Peng, Yongsheng Sang, and Yong Xiang. Multi-view linear discriminant analysis network. *IEEE Transactions on Image Processing*, 2019.

[Huang *et al.*, 2019] Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *IJCAI*, 2019.

[Jiang *et al.*, 2019] Yangbangyan Jiang, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Dm2c: Deep mixed-modal clustering. *NeurIPS*, 2019.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *AAAI*, 2014.

[Li *et al.*, 2021] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *AAAI*, 2021.

[Li *et al.*, 2022] Yunfan Li, Mouxing Yang, Dezhong Peng, Taihao Li, Jiantao Huang, and Xi Peng. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 2022.

[Lin *et al.*, 2021] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *CVPR*, 2021.

[Lin *et al.*, 2022] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Liu *et al.*, 2019] Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, En Zhu, Tongliang Liu, Marius Kloft, Dinggang Shen, Jianping Yin, and Wen Gao. Multiple kernel k k-means with incomplete kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[Liu *et al.*, 2020] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Luo *et al.*, 2020] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *CVPR*, 2020.

[Peng *et al.*, 2019] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *ICML*, 2019.

[Qu *et al.*, 2017] Meng Qu, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. An attention-based collaboration framework for multi-view network representation learning. In *CIKM*, 2017.

[Rush *et al.*, 2015] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv:1509.00685*, 2015.

[Shao *et al.*, 2015] Weixiang Shao, Lifang He, and Philip S Yu. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *ECML PKDD*, 2015.

[Tang and Liu, 2022] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *ICML*, 2022.

[Tao *et al.*, 2017] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. From ensemble clustering to multi-view clustering. In *IJCAI*, 2017.

[Van Gansbeke *et al.*, 2020] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[Wang *et al.*, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, 2015.

[Wang *et al.*, 2018] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Partial multi-view clustering via consistent gan. In *ICDM*, 2018.

[Wang *et al.*, 2019] Hao Wang, Linlin Zong, Bing Liu, Yan Yang, and Wei Zhou. Spectral perturbation meets incomplete multi-view data. *arXiv:1906.00098*, 2019.

[Wen *et al.*, 2019] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Hong Liu. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *AAAI*, 2019.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[Yan *et al.*, 2022] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid.

Multiview transformers for video recognition. In *CVPR*, 2022.

[Yang *et al.*, 2021a] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. Partially view-aligned representation learning with noise-robust contrastive loss. In *CVPR*, 2021.

[Yang *et al.*, 2021b] Xu Yang, Cheng Deng, Zhiyuan Dang, and Dacheng Tao. Deep multiview collaborative clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Yang *et al.*, 2022a] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *CVPR*, 2022.

[Yang *et al.*, 2022b] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jian Cheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Yu *et al.*, 2018] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv:1805.03508*, 2018.

[Zhang *et al.*, 2018] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Zhang *et al.*, 2019a] Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. *NIPS*, 2019.

[Zhang *et al.*, 2019b] Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In *CVPR*, 2019.

[Zhang *et al.*, 2020] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Zhao *et al.*, 2016] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, 2016.

[Zhou and Shao, 2018] Yi Zhou and Ling Shao. Aware attentive multi-view inference for vehicle re-identification. In *CVPR*, 2018.