MISSING-VIEW SELF-DISTILLATION IMPUTATION FOR INCOMPLETE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world multi-view data often contain missing views due to failures during acquisition, transmission, or storage. This makes it difficult to directly apply multi-view clustering methods that require complete samples. Existing methods typically recover missing information through cross-view prediction, neighborhood structure, or relation propagation. However, some graph-based methods rely on shared relations that may not capture specific structure of target view, potentially introducing bias into the recovered representations. To address this limitation, we propose Missing-View Self-Distillation Imputation for Incomplete Multi-View Clustering (MVSI-IMVC). Specifically, we construct a target-view self-distillation predictor. We use full-information and pseudo-missing states from complete samples and use the observed target view to provide direct supervision. Teacher predictions further guide target-view latent recovery for incomplete samples. This allows the model to learn predictive relationships specific to each target view and missing pattern. We also introduce Henze–Penrose divergence-guided fused representation learning. Together with a cross-view instance contrastive loss, it organizes completed latent representations at the instance and distribution levels and connects target-view latent recovery with fused representation learning and clustering. Experiments show that MVSI-IMVC achieves competitive clustering performance on multiple benchmark datasets.

1 INTRODUCTION

Multi-view data provide complementary descriptions of the same objects through different modalities, sensors, or feature extractors. Multi-view clustering (MVC) integrates these descriptions without class annotations, providing a way to discover structure in heterogeneous data. In recent years, deep MVC methods have typically combined view-specific representation learning with shared semantic modeling or contrastive learning to improve clustering performance (Xu et al., 2021; 2022b). However, their standard formulations generally assume that every sample is observed in all views. In practice, failures during data acquisition, transmission, or storage can leave some views missing. This prevents the direct application of methods that require complete samples (Chao et al., 2024). Incomplete multi-view clustering (IMVC) aims to learn representations from complete and incomplete samples and discover a shared clustering structure.

Existing IMVC methods can be broadly divided into traditional and deep approaches. Traditional methods mainly use matrix factorization, kernel learning, or graph learning to handle incomplete samples. Matrix factorization methods typically exploit cross-view sample correspondences to learn shared or consistent low-dimensional representations. For example, PVC encourages representations of the same sample to remain close across views in a latent subspace. DAIMC learns shared cross-view representations through weighted semi-nonnegative matrix factorization and basis alignment (Li et al., 2014; Hu & Chen, 2018). Kernel learning methods start from sample similarities and jointly optimize missing kernel matrix estimation and multi-kernel clustering. This allows the clustering objective to also constrain the completion process (Liu et al., 2017). Graph learning methods exploit observed-view information by modeling structural relationships among samples. For example, Wen et al. (2020) use low-rank representations to adaptively construct graphs for individual views. They combine spectral constraints with co-regularization to learn a common representation. These traditional methods can uncover shared cross-view structure from incomplete samples. However, their representations are usually built directly on pre-extracted input features or kernel and graph

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

structures. Their capacity for nonlinear representation is limited by the model formulation, which has motivated the development of deep IMVC methods (Xu et al., 2022a).

Deep IMVC methods use neural networks to learn nonlinear representations and recover missing information from observed views. Some approaches estimate missing views through reconstruction, semantic neighbors, or conditional generation (Wang et al., 2021; Xue et al., 2021; Tang & Liu, 2022; Wen et al., 2024). Latent graph structures and cross-view relationships are also exploited to transfer information for missing-view recovery (Pu et al., 2024; Chao et al., 2024; 2025). However, some graph-based methods rely on cross-view relation transfer or consensus information propagation to recover missing representations. Their recovery is guided by graph structure and clustering objectives. When the transferred relations differ from the target view's local structure, the recovered representation may not accurately reflect the sample's features in that view.

To address this issue, we propose Missing-View Self-Distillation Imputation for Incomplete Multi-View Clustering (MVSI-IMVC). The proposed framework is illustrated in Figure 1. We design a target-view self-distillation predictor that learns target-view latent recovery relationships from complete samples naturally present in the training set. Specifically, we construct full-information and pseudo-missing states from the same complete sample and process them with a target-view predictor. The observed target-view representation provides direct supervision for both states. The teacher prediction from the full-information state further guides target-view latent recovery from the pseudo-missing state. In this way, the target-view observations and additional multi-view context of complete samples are converted into supervision that guides target-view latent recovery for incomplete samples. This enables the model to learn predictive relationships specific to each target view and missing pattern. We further introduce Henze–Penrose divergence-guided fused representation learning to characterize statistical association between completed latent representations and their fused representation (Yasaei Sekeh & Hero, 2019). Combined with a cross-view instance contrastive loss, it regularizes the completed latent representations at the instance and distribution levels. This connects target-view latent recovery with subsequent fused representation learning and clustering.

Our main contributions are summarized as follows:

- We propose the MVSI-IMVC framework with a target-view self-distillation predictor that uses complete samples to construct full-information and pseudo-missing states. By combining target-view supervision and self-distillation, it learns target-view latent recovery relationships specific to each missing pattern.
- We introduce Henze–Penrose divergence-guided fused representation learning. Combined with a cross-view instance contrastive loss, it organizes completed latent representations at the instance and distribution levels and connects target-view latent recovery with fused representation learning and clustering.
- Experiments on multiple benchmark datasets show that the proposed model achieves competitive clustering performance.

2 Related Work

2.1 Incomplete Multi-View Clustering

Deep IMVC methods mainly follow four directions. Autoencoder-based methods such as COMPLETER and DCP combine cross-view prediction with contrastive learning to recover missing-view representations (Lin et al., 2021; 2023). DIMVC-HIA integrates completion with hierarchical imputation and alignment (Du et al., 2026). Among adversarial generation methods, GP-MVC generates missing views through view-specific adversarial networks conditioned on shared representations (Wang et al., 2021). Graph- and neighborhood-based methods include ICMVC, which transfers cross-view relations through graph convolution (Chao et al., 2024), and GHICMC, which propagates consensus representations over a global graph (Chao et al., 2025). DSIMVC imputes from semantic neighbors and selects reliable imputed samples (Tang & Liu, 2022). Among prototype-based methods, ProImp recovers missing views through view-specific prototypes and transferred sample–prototype relationships (Li et al., 2023). FreeCSL learns consensus prototypes and view-specific cluster structure without explicit imputation or alignment (Dai et al., 2025). However, cross-view imputation may

introduce semantic mismatch, and excessive cross-view consistency can neglect view-specific cluster information (Tang & Liu, 2022; Dai et al., 2025).

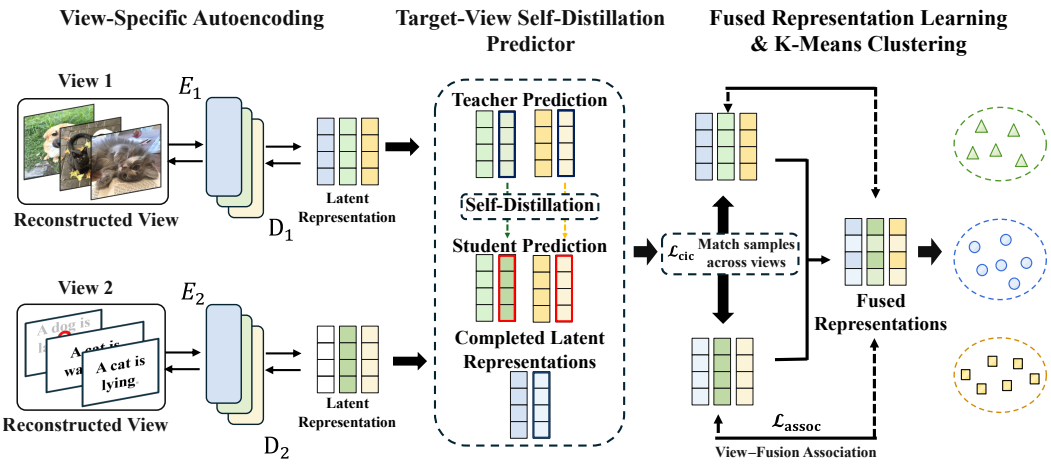


Figure 1: Overview of the proposed MVSI-IMVC framework. View-specific encoders extract latent representations, while target-view self-distillation uses complete samples to supervise missing-view recovery. Cross-view instance contrastive and Henze–Penrose divergence-based view–fusion association losses guide fused representation learning for k-means clustering.

3 METHOD

3.1 NOTATIONS AND PROBLEM DEFINITION

Let $\mathcal{D} = \{(\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(V)}, \mathbf{m}_i)\}_{i=1}^N$ contain N samples represented by V views. The binary indicator $m_i^{(v)} = 1$ means that view v of sample i is observed, whereas $m_i^{(v)} = 0$ means that it is missing. A sample is complete if all V views are observed and incomplete if at least one view is missing. We assume that every sample retains at least one observed view, i.e., $\sum_v m_i^{(v)} \geq 1$. Given the number of clusters K, the goal is to learn a fused representation for clustering from incomplete multi-view data without using class labels during representation learning.

3.2 TARGET-VIEW SELF-DISTILLATION PREDICTOR

Cross-view agreement alone does not specify the latent content to recover for each target view. We therefore supervise target-view latent recovery using both its observed representation and the teacher prediction from the full-information state. Complete samples supply both sources through paired pseudo-missing and full-information states. The student branch processes the pseudo-missing state, while the teacher branch processes the full-information state; both branches use the same predictor.

View-Specific Latent Representation Learning. We learn view-specific latents and use reconstruction to retain information from observed inputs. Each view uses an encoder E_v and a decoder D_v. For an observed slot, the encoder produces a d-dimensional simplex-valued latent representation

$$\mathbf{z}_i^{(v)} = \mathrm{softmax}(E_v(\mathbf{x}_i^{(v)})) \in \Delta^{d-1}. \tag{1}$$

For a minibatch $\mathcal{B}$, let $\mathcal{O}_v = \{i \in \mathcal{B} : m_i^{(v)} = 1\}$, and let d_v denote the input feature dimension of view v. The reconstruction loss is

$$\mathcal{L}_{\mathrm{rec}} = \sum_{v:\,|\mathcal{O}_v|>0} \frac{1}{|\mathcal{O}_v| d_v} \sum_{i \in \mathcal{O}_v} \left\| D_v(\mathbf{z}_i^{(v)}) - \mathbf{x}_i^{(v)} \right\|_2^2. \tag{2}$$

The observed-view latents provide inputs to the target-view predictors and, for complete samples, target representations for supervision.

Student prediction from the pseudo-missing state. For target view v, let P_v be a predictor with a softmax output. Given a binary availability mask $\mathbf{m}$, its input is

$$\mathbf{s}_i(\mathbf{m}) = \left[m^{(1)}\mathbf{z}_i^{(1)}; \ldots; m^{(V)}\mathbf{z}_i^{(V)}; \mathbf{m}\right]. \tag{3}$$

Latent blocks for missing views are filled with zeros, and the concatenated mask indicates which views are observed. Each predictor outputs a latent distribution in its own target view's space.

We simulate incomplete samples from complete samples so that pseudo-hidden target views remain observed in the original samples and can provide supervision. For sample i, we draw a pseudo-missing mask $\bar{\mathbf{m}}_i$ from the empirical distribution of missing patterns, retaining at least one observed and one hidden view. Let $\mathcal{H}_i = \{v : \bar{m}_i^{(v)} = 0\}$ denote the pseudo-hidden views. The student prediction for each pseudo-hidden view is obtained from the pseudo-missing state:

$$\mathbf{p}_{i,S}^{(v)} = P_v(\mathbf{s}_i(\bar{\mathbf{m}}_i)), \qquad v \in \mathcal{H}_i. \tag{4}$$

Since both the observed target latent and the prediction are simplex-valued, we match their distributions using Kullback–Leibler (KL) divergence. With $\mathrm{sg}[\cdot]$ denoting the stop-gradient operator, the recovery loss is

$$\mathcal{L}_{\text{recover}} = \mathbb{E}_{i,v\in\mathcal{H}_i} D_{\mathrm{KL}}\left(\mathrm{sg}[\mathbf{z}_i^{(v)}] \,\|\, \mathbf{p}_{i,S}^{(v)}\right). \tag{5}$$

The expectation is over sampled complete samples and their pseudo-hidden views. This supervision acts on the target view's latent distribution, encouraging target-view latent recovery to retain the information encoded in that view.

Teacher prediction from the full-information state. For the same sample, we provide all view latents to P_v and obtain the teacher prediction

$$\mathbf{p}_{i,T}^{(v)} = P_v(\mathbf{s}_i(\mathbf{1})), \tag{6}$$

where $\mathbf{1}$ denotes the all-observed mask. The teacher has access to the target view and the complete multi-view context. We ground its prediction in observed target-view latents to tie the distillation reference to the content being recovered:

$$\mathcal{L}_{\text{ground}} = \mathbb{E}_{i,v} D_{\mathrm{KL}}\left(\mathrm{sg}[\mathbf{z}_i^{(v)}] \,\|\, \mathbf{p}_{i,T}^{(v)}\right). \tag{7}$$

Here, the expectation covers all views of the sampled complete samples.

Target-view self-distillation. The recovery loss specifies the target-view latent to recover. Distillation additionally links the teacher and student predictions, allowing the teacher prediction to guide recovery of each pseudo-hidden view:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{i,v\in\mathcal{H}_i} D_{\mathrm{KL}}\left(\mathrm{sg}[\mathbf{p}_{i,T}^{(v)}] \,\|\, \mathbf{p}_{i,S}^{(v)}\right). \tag{8}$$

The stop-gradient operator treats the teacher output as a fixed target for this loss. The three terms are optimized jointly:

$$\mathcal{L}_{\text{tsp}} = \mathcal{L}_{\text{recover}} + \lambda_{\mathrm{g}}\mathcal{L}_{\text{ground}} + \lambda_{\mathrm{d}}\mathcal{L}_{\text{distill}}, \tag{9}$$

where λ_{g} and λ_{d} control the grounding and distillation terms, respectively.

We use failure replay to revisit difficult recovery cases. A small priority cache retains high-loss pseudo-missing tasks, represented as sample–mask pairs. When available, these tasks replace a small fraction of newly sampled pseudo-missing tasks.

For an actual availability mask $\mathbf{m}_i$, we retain observed latents and use the corresponding target-view predictors to recover missing latent slots:

$$\tilde{\mathbf{z}}_i^{(v)} = \begin{cases} \mathbf{z}_i^{(v)}, & m_i^{(v)} = 1, \\ P_v(\mathbf{s}_i(\mathbf{m}_i)), & m_i^{(v)} = 0. \end{cases} \tag{10}$$

These completed latent representations are used for subsequent fusion. The teacher prediction provides supervision during training; at inference, target-view prediction uses only the observed views and their mask.

3.3 HENZE–PENROSE DIVERGENCE-GUIDED FUSED REPRESENTATION LEARNING

Starting from the completed latent representations, we learn fused representations for clustering by encouraging Henze–Penrose divergence-based association between each view and the fusion. We additionally use a cross-view instance contrastive loss to encourage agreement for the same sample and separation between different instances.

Let $\mathcal{B}$ be a minibatch of $B \geq 2$ samples. We use $V \geq 2$ views and a latent dimension $d \geq 2$, chosen independently of the number of clusters K. The completed latent representations in Equation 10 satisfy $\tilde{\mathbf{z}}_i^{(v)} \in \Delta^{d-1}$.

Cross-view instance contrastive loss. Representations of the same sample in two views form a positive pair, while other samples in the opposite view provide negatives. For a view pair (u, v), define

$$s_{ij}^{(u,v)} = \exp\left(\frac{(\tilde{\mathbf{z}}_i^{(u)})^\top \tilde{\mathbf{z}}_j^{(v)}}{\|\tilde{\mathbf{z}}_i^{(u)}\|_2 \|\tilde{\mathbf{z}}_j^{(v)}\|_2}\right). \tag{11}$$

The cross-view instance contrastive loss compares each positive score with the sum of cross-view negative scores in both directions:

$$\mathcal{L}_{\text{cic}} = -\frac{1}{2B} \sum_{u<v} \sum_{i=1}^{B} \left[\log \frac{s_{ii}^{(u,v)}}{\sum_{j\neq i} s_{ij}^{(u,v)}} + \log \frac{s_{ii}^{(u,v)}}{\sum_{j\neq i} s_{ji}^{(u,v)}}\right]. \tag{12}$$

Both directions exclude the positive pair from their denominators.

Fusion of completed latent representations. To base fusion weights on observed latents, we estimate them from complete samples in the minibatch. Let $\mathcal{C} = \{i \in \mathcal{B} : m_i^{(v)} = 1 \text{ for all } v\}$ and $n_{\mathcal{C}} = |\mathcal{C}|$. For $n_{\mathcal{C}} \geq 2$, the variance over all entries of view v's complete-sample latent matrix is

$$t_v = \frac{1}{n_{\mathcal{C}} d - 1} \sum_{i\in\mathcal{C}} \sum_{a=1}^{d} \left(z_{ia}^{(v)} - \frac{1}{d}\right)^2. \tag{13}$$

The pooled mean is $1/d$ because each row sums to one. We set $w_v = t_v / \sum_u t_u$ when the variances are finite and their sum is positive; otherwise, including $n_{\mathcal{C}} < 2$, we use $w_v = 1/V$. These weights reflect coordinate concentration in the complete-sample representations and are shared across the minibatch:

$$\mathbf{f}_i = \sum_{v=1}^{V} w_v \tilde{\mathbf{z}}_i^{(v)}, \qquad w_v \geq 0, \quad \sum_v w_v = 1. \tag{14}$$

The fused representation remains in Δ^{d-1}. We compute the association loss from the fused and completed latent representations over the entire minibatch.

View–fusion association. Following the geometric mutual information construction (Yasaei Sekeh & Hero, 2019), we form a joint coordinate distribution and an independent reference. Let $\tilde{\mathbf{Z}}^{(v)}, \mathbf{F} \in \mathbb{R}^{B\times d}$ collect the completed view-specific latent representations and fused representations, respectively. Corresponding rows refer to the same sample. Choose a sample index I uniformly from the minibatch, and let $U_v, W \in \{1, \ldots, d\}$ be auxiliary coordinate variables with conditional marginals $\Pr(U_v = a \mid I = i) = \tilde{z}_{ia}^{(v)}$ and $\Pr(W = b \mid I = i) = f_{ib}$. For each sample, we use the product coupling $\Pr(U_v = a, W = b \mid I = i) = \tilde{z}_{ia}^{(v)} f_{ib}$. Averaging these per-sample outer products yields the joint coordinate table

$$\mathbf{Q}^{(v)} = \frac{1}{B} (\tilde{\mathbf{Z}}^{(v)})^\top \mathbf{F} \in \mathbb{R}^{d\times d}. \tag{15}$$

The simplex-valued representations ensure that $\mathbf{Q}^{(v)}$ has nonnegative entries summing to one. Although the coordinates are conditionally independent within each sample, mixing over the shared sample index can produce association between U_v and W.

Rows and columns of $\mathbf{Q}^{(v)}$ refer to view and fusion coordinates, respectively, with marginals $q_a^{(v)} = \sum_b Q_{ab}^{(v)}$ and $r_b = \sum_a Q_{ab}^{(v)}$. Both marginal vectors lie in $\mathbb{R}^d$, and the fusion marginal $\mathbf{r}$ is shared across views. Their product defines the independent reference $\mathbf{R}^{(v)} = \mathbf{q}^{(v)}\mathbf{r}^\top \in \mathbb{R}^{d\times d}$, whose entries are $R_{ab}^{(v)} = q_a^{(v)} r_b$. Unlike $\mathbf{Q}^{(v)}$, this reference averages over independently paired sample indices, so comparing the two measures association beyond overall coordinate usage.

View–fusion association loss. We use the equal-prior Henze–Penrose divergence between the joint table and its independent reference to encourage view–fusion association, and add marginal entropy regularization for broad batch-level coordinate usage. With $H(\mathbf{p}) = -\sum_a p_a \log p_a$, using natural logarithms and $0 \log 0 = 0$, the view–fusion association loss is

$$\mathcal{L}_{\text{assoc}} = -\sum_{v=1}^{V} \left[\frac{1}{2} \sum_{a,b} \frac{\left(Q_{ab}^{(v)} - R_{ab}^{(v)}\right)^2}{Q_{ab}^{(v)} + R_{ab}^{(v)}} + \lambda_{\text{H}} \big(H(\mathbf{q}^{(v)}) + H(\mathbf{r}) \big) \right]. \tag{16}$$

Here, $\lambda_{\text{H}} = 9$, and zero-over-zero terms are defined as zero. For each view, the divergence term is symmetric, lies in $[0, 1]$, and is zero exactly when $\mathbf{Q}^{(v)} = \mathbf{R}^{(v)}$, corresponding to independence of the constructed coordinate variables. Maximizing this term encourages their association.

The entropy term encourages broad batch-level coordinate usage without requiring each sample's representation to be uniform. Through the completed latent representations, this loss guides both encoders and target-view predictors, coupling target-view latent recovery with fused representation learning.

3.4 Overall Objective

The model is jointly optimized with the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{tsp}} + \beta \mathcal{L}_{\text{cic}} + \gamma \mathcal{L}_{\text{assoc}}. \tag{17}$$

Here, α, β, and γ are balancing coefficients. $\mathcal{L}_{\text{rec}}$ reconstructs observed views, $\mathcal{L}_{\text{tsp}}$ trains the target-view self-distillation predictor, $\mathcal{L}_{\text{cic}}$ is the cross-view instance contrastive loss, and $\mathcal{L}_{\text{assoc}}$ promotes view–fusion association with marginal entropy regularization. Finally, k-means is applied to the fused representations to obtain K clusters.

4 Experiments

4.1 Datasets

We conducted experiments on four representative datasets. **HandWritten** (Duin, 1998) contains 2,000 samples from ten digit classes, with Pixel, Fourier, and Profile features. **Scene-15** (Lazebnik et al., 2006) and **LandUse-21** (Yang & Newsam, 2010) contain 4,485 samples from 15 scene classes and 2,100 samples from 21 land-use classes, respectively. Both use GIST, PHOG, and LBP as three views. **ALOI-100** (Geusebroek et al., 2005), a subset of ALOI, contains 10,800 object images from 100 classes, with HSB, RGB, Colorsim, and Haralick features.

4.2 Comparison Methods

We compare our method with ten baselines. Fusion-kmeans applies k-means to mean-fused features. COMPLETER (Lin et al., 2021) learns cross-view predictive mappings together with contrastive representation learning to recover missing-view features. DIMVC (Xu et al., 2022a) learns from incomplete samples without explicit imputation by exploiting cluster complementarity. DSIMVC (Tang & Liu, 2022) dynamically imputes missing views from semantic neighbors and selects imputed samples through bilevel optimization. DCP (Lin et al., 2023) learns consistent representations through contrastive learning and recovers missing-view latent representations through dual prediction. ProImp (Li et al., 2023) uses view-specific prototypes and transferred sample–prototype relationships for imputation. APADC (Xu et al., 2023) performs completion-free feature learning through adaptive projection and distribution alignment. ICMVC (Chao et al., 2024) combines graph-based information

Table 1: **Clustering performance (%)** under missing rates $\{0.1, 0.3, 0.5, 0.7\}$. Higher values indicate better performance. Bold and underlined values denote the best and second-best results, respectively; OOM denotes out of memory.

	Missing rate	0.1			0.3			0.5			0.7		
	Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
LandUse-21	Fusion-kmeans	20.45	25.42	8.62	16.41	17.48	5.51	12.86	12.58	2.77	11.45	9.81	1.38
	COMPLETER(2021)	26.40	32.48	13.93	26.96	32.64	12.09	21.36	26.27	9.34	24.43	29.01	10.31
	DIMVC(2022)	24.63	30.04	10.58	23.69	29.94	10.01	22.40	27.78	9.38	21.77	26.14	7.91
	DSIMVC(2022)	18.47	19.34	5.58	17.95	18.47	5.16	18.13	18.53	5.26	17.90	17.97	5.11
	DCP(2023)	26.78	30.87	13.80	27.08	30.69	13.80	23.07	27.00	11.31	25.18	28.04	12.00
	ProImp(2023)	22.38	23.79	8.76	19.53	20.55	6.86	20.30	21.94	7.32	15.10	15.48	4.00
	APADC(2023)	22.75	31.90	9.50	18.08	24.72	7.22	15.67	21.23	5.61	15.11	20.08	4.84
	ICMVC(2024)	28.18	31.78	15.14	25.77	29.39	12.69	25.98	27.74	11.92	22.26	24.95	9.31
	GHICMC(2025)	26.86	31.14	12.81	25.72	29.26	11.32	25.15	28.57	11.26	23.53	26.55	9.72
	DDR-IMVC(2026)	28.21	34.68	14.93	28.02	33.49	14.27	27.14	31.89	13.24	25.36	28.09	10.67
	MVSI-IMVC	**28.29**	**35.27**	**15.17**	**28.29**	**34.82**	**14.81**	**27.19**	**33.43**	**13.44**	**26.86**	**29.97**	**12.59**
Scene-15	Fusion-kmeans	34.20	34.84	20.93	22.39	23.60	12.80	17.54	17.31	7.88	14.90	13.54	4.29
	COMPLETER(2021)	40.28	42.50	23.13	40.12	42.93	23.96	39.12	41.79	22.98	38.05	40.22	21.84
	DIMVC(2022)	32.95	27.41	15.61	33.51	29.42	16.75	30.65	25.21	13.64	29.58	24.11	12.85
	DSIMVC(2022)	27.65	29.74	14.11	26.73	29.36	13.94	26.40	28.03	13.04	25.31	27.04	12.43
	DCP(2023)	38.54	42.39	23.33	40.49	43.10	24.14	39.50	42.35	23.51	38.55	40.57	21.72
	ProImp(2023)	40.74	42.14	24.00	41.69	43.03	25.28	40.28	41.80	23.89	39.96	40.35	22.92
	APADC(2023)	43.70	44.20	26.00	41.80	43.10	24.30	39.90	42.40	23.80	38.50	41.10	22.80
	ICMVC(2024)	38.78	36.62	21.84	37.40	34.94	20.60	31.35	27.91	14.98	25.31	23.88	11.33
	GHICMC(2025)	41.26	43.21	24.95	40.98	43.13	25.05	40.91	42.29	24.64	38.79	40.94	22.92
	DDR-IMVC(2026)	46.16	47.62	28.69	45.53	**45.99**	28.05	44.35	43.67	26.79	42.12	41.20	24.73
	MVSI-IMVC	**46.62**	**47.68**	**29.30**	**46.47**	**45.99**	**28.44**	**44.82**	**44.33**	**27.99**	**42.21**	**41.31**	**25.25**
HandWritten	Fusion-kmeans	41.70	47.59	34.22	36.28	38.76	21.46	29.64	28.51	11.54	25.69	22.14	6.58
	COMPLETER(2021)	83.22	82.47	73.59	75.38	77.55	61.69	74.05	76.13	58.89	78.55	76.07	68.67
	DIMVC(2022)	67.13	63.17	53.16	59.43	56.49	43.19	54.80	50.50	30.76	43.82	41.54	23.80
	DSIMVC(2022)	84.35	80.32	74.38	85.64	80.71	75.94	84.73	78.82	74.13	82.71	75.35	69.85
	DCP(2023)	53.35	65.72	35.60	51.96	63.88	31.49	59.06	65.07	36.51	60.97	60.53	29.90
	ProImp(2023)	83.20	80.29	74.17	84.24	77.75	72.60	78.16	70.79	63.96	80.31	68.85	62.92
	APADC(2023)	67.43	65.34	47.18	68.95	67.28	45.98	68.85	68.61	56.43	61.77	61.97	48.26
	ICMVC(2024)	83.16	81.33	74.78	82.01	79.62	72.22	75.13	71.99	63.19	72.47	70.01	59.71
	GHICMC(2025)	96.19	92.14	92.89	96.11	91.32	90.83	**94.88**	**89.16**	**89.10**	92.73	85.85	84.71
	DDR-IMVC(2026)	96.38	92.23	91.99	96.15	91.49	91.21	94.34	88.38	87.87	90.86	82.65	81.92
	MVSI-IMVC	**96.45**	**92.34**	**92.94**	**96.25**	**91.62**	**91.85**	94.40	88.54	88.04	**93.50**	**87.29**	**86.19**
ALOI-100	Fusion-kmeans	52.37	72.31	40.79	30.63	55.28	16.10	22.48	47.33	7.46	17.39	41.82	4.94
	COMPLETER(2021)	48.19	77.96	44.25	43.03	72.43	36.73	36.16	66.89	26.52	34.55	64.06	24.97
	DIMVC(2022)	71.86	84.99	61.79	68.52	82.15	58.31	64.80	78.53	51.36	61.64	75.33	47.25
	DSIMVC(2022)	38.76	67.49	29.71	38.89	66.00	29.12	39.32	64.42	28.53	35.98	61.28	25.16
	DCP(2023)	51.85	74.88	42.73	47.38	70.54	38.38	42.37	66.30	32.36	36.02	60.75	25.40
	ProImp(2023)	68.39	83.47	62.08	45.98	73.01	38.53	32.71	65.74	24.76	29.23	62.08	19.46
	APADC(2023)	47.40	68.92	35.02	38.95	62.27	26.10	32.78	58.16	20.10	26.02	53.91	14.11
	ICMVC(2024)	68.02	80.78	56.64	68.14	80.40	55.94	67.68	78.92	53.92	49.15	70.50	38.45
	GHICMC(2025)		OOM			OOM			OOM			OOM	
	DDR-IMVC(2026)	76.02	88.35	67.82	73.18	85.82	64.36	69.87	82.34	58.09	66.94	78.60	53.20
	MVSI-IMVC	**79.82**	**90.28**	**74.67**	**75.64**	**88.38**	**69.88**	**71.08**	**86.29**	**63.77**	**68.77**	**84.65**	**62.45**

transfer with high-confidence guidance. GHICMC (Chao et al., 2025) exploits global graph propagation and hierarchical information transfer to handle missing views. DDR-IMVC (Wang et al., 2026) refines representations of incomplete samples using representations of complete samples and employs robust contrastive constraints to mitigate alignment noise.

4.3 Experimental Settings

Our model is implemented in PyTorch and trained with Adam on an NVIDIA GeForce RTX 5080 GPU. The learning rate ranges from 7.5×10^{-6} to 2×10^{-5}. We evaluate sample-level missing rates $\rho \in \{0.1, 0.3, 0.5, 0.7\}$, where ρ denotes the fraction of samples made incomplete by randomly removing views while retaining at least one observed view per incomplete sample. We apply k-

means to the fused representations using the known number of classes. Performance is evaluated by clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI), with higher values indicating better results.

4.4 Analysis of Main Results

Table 1 compares our model with ten baselines on four datasets at four missing rates using ACC, NMI, and ARI.

Overall comparison. Our model ranks first in 45 of the 48 dataset–rate–metric comparisons, including one tied result, and second in the remaining three. Its average ACC, NMI, and ARI margins over the strongest available baseline for each condition and metric are 0.85, 1.30, and 2.03 percentage points, respectively.

Performance at high missing rates. At the highest tested missing rate of 0.7, our model achieves the best result in all 12 dataset–metric comparisons, with average ACC, NMI, and ARI gains of 1.05, 2.14, and 2.96 percentage points over the strongest available baselines, respectively. The average NMI and ARI margins are the largest among the four tested missing rates.

Dataset-wise analysis. The largest average gains occur on ALOI-100. Relative to DDR-IMVC, the strongest available baseline on this dataset, our model improves ACC, NMI, and ARI by an average of 2.33, 3.62, and 6.83 percentage points across the four missing rates. On LandUse-21 and Scene-15, it achieves the best or tied-best results in all 12 comparisons per dataset.

4.5 Model Analysis

1) Ablation Study. Table 2 evaluates progressively adding the losses in Equation 17 on Scene-15 and HandWritten. Adding $\mathcal{L}_{\text{tsp}}$ to reconstruction raises ACC from 17.30% to 35.03% and from 29.60% to 69.00%, respectively. Adding $\mathcal{L}_{\text{cic}}$ further improves all three metrics on both datasets. Adding $\mathcal{L}_{\text{assoc}}$ then yields the best results, with additional ACC gains of 11.27 and 20.00 percentage points, respectively. These results support the combined use of target-view self-distillation, cross-view instance contrastive loss, and view–fusion association loss.

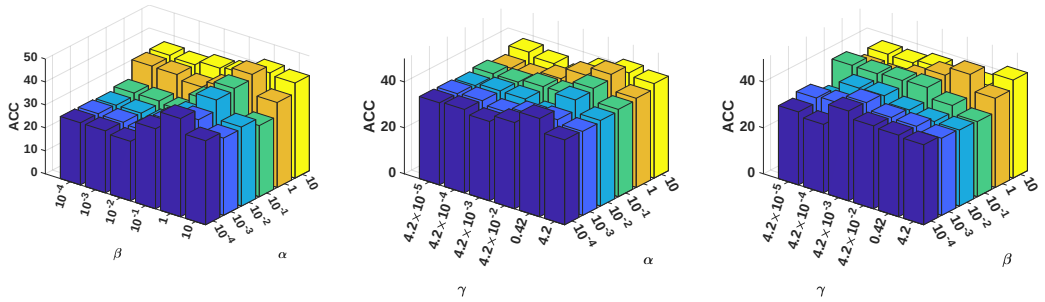
Table 2: Ablation study of the loss terms on Scene-15 and HandWritten (%). A checkmark indicates an included term. Bold and underlined values denote the best and second-best results within each dataset, respectively.

Loss Terms				Scene-15			HandWritten		
$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{tsp}}$	$\mathcal{L}_{\text{cic}}$	$\mathcal{L}_{\text{assoc}}$	ACC	NMI	ARI	ACC	NMI	ARI
✓				17.30	20.90	2.45	29.60	37.41	7.17
✓	✓			35.03	35.06	19.30	69.00	70.45	57.18
✓	✓	✓		<u>35.20</u>	<u>35.44</u>	<u>22.29</u>	<u>76.25</u>	<u>74.89</u>	<u>64.69</u>
✓	✓	✓	✓	**46.47**	**45.99**	**28.44**	**96.25**	**91.62**	**91.85**

2) Sensitivity to Loss Weights. Figure 2 evaluates loss-weight sensitivity on Scene-15 at $\rho = 0.3$, varying two coefficients while fixing the third. ACC generally improves as α and β increase from very small values. Relatively high ACC is observed across several neighboring parameter combinations, suggesting flexibility in balancing the three objectives within the tested ranges.

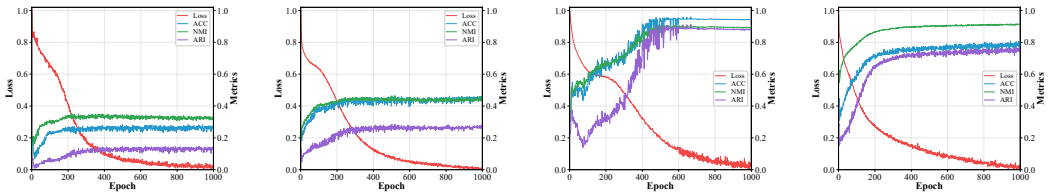
3) Convergence Analysis. Figure 3 shows the training loss and clustering metrics over 1,000 epochs on four datasets at $\rho = 0.3$. The loss generally decreases, while ACC, NMI, and ARI improve and gradually stabilize, supporting the model's empirical convergence under the tested setting.

4) Visualization Analysis. Figure 4 compares our model with nine baselines on HandWritten at $\rho = 0.5$. Our model forms compact, clearly separated clusters, whereas several baselines exhibit dispersed or overlapping groups, providing qualitative support for the discriminative structure of the learned representations.



(a) α and β (b) α and γ (c) β and γ

Figure 2: Sensitivity of ACC (%) to pairs of loss weights on Scene-15 at a missing rate of $\rho = 0.3$. The remaining weight is fixed in each panel.



(a) LandUse-21 (b) Scene-15 (c) HandWritten (d) ALOI-100

Figure 3: Training loss and clustering metrics on four datasets at a missing rate of $\rho = 0.3$. The left and right axes show the loss and clustering metrics, respectively.

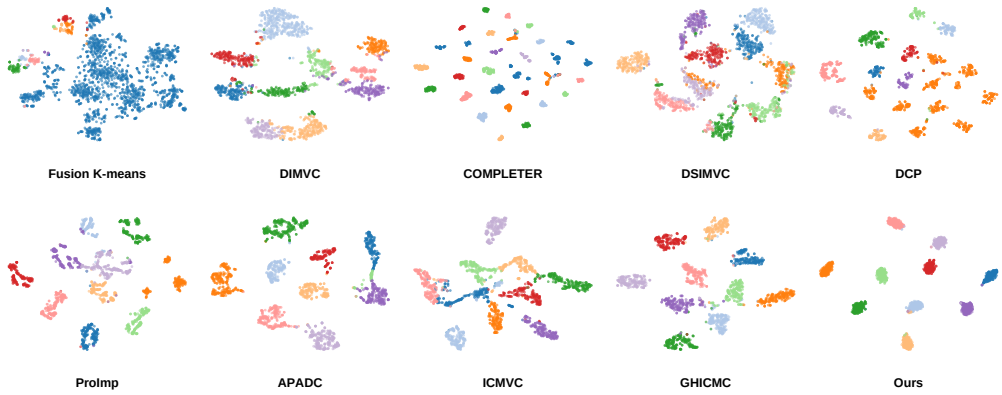


Figure 4: Visualization results of nine baseline methods and our model on HandWritten at a missing rate of $\rho = 0.5$.

5 Conclusion

We proposed MVSI-IMVC for incomplete multi-view clustering. Its target-view self-distillation predictor learns missing-view latent recovery from complete samples using target-view supervision and teacher predictions. Henze–Penrose divergence-guided fused representation learning, together with a cross-view instance contrastive loss, regularizes the completed representations at the instance and distribution levels for clustering. Experiments on four benchmark datasets show competitive performance across missing rates, with more pronounced advantages on ALOI-100 and at higher missing rates.

AI USE STATEMENT

In this work, we used large language models to assist with editing and polishing the language of this manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 11221–11229. AAAI Press, 2024. doi: 10.1609/AAAI.V38I10.29000. URL https://doi.org/10.1609/aaai.v38i10.29000.

Guoqing Chao, Kaixin Xu, Xijiong Xie, and Yongyong Chen. Global graph propagation with hierarchical information transfer for incomplete contrastive multi-view clustering. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 15713–15721. AAAI Press, 2025. doi: 10.1609/AAAI.V39I15.33725. URL https://doi.org/10.1609/aaai.v39i15.33725.

Yuzhuo Dai, Jiaqi Jin, Zhibin Dong, Siwei Wang, Xinwang Liu, En Zhu, Xihong Yang, Xinbiao Gan, and Yu Feng. Imputation-free and alignment-free: Incomplete multi-view clustering driven by consensus semantic learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5071–5081. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00478. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dai_Imputation-free_and_Alignment-free_Incomplete_Multi-view_Clustering_Driven_by_Consensus_Semantic_CVPR_2025_paper.html.

Yiming Du, Ziyu Wang, Jian Li, Rui Ning, and Lusi Li. Deep incomplete multi-view clustering via hierarchical imputation and alignment. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 20941–20949. AAAI Press, 2026. doi: 10.1609/AAAI.V40I25.39235. URL https://doi.org/10.1609/aaai.v40i25.39235.

Robert Duin. Multiple features. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/72.

Jan-Mark Geusebroek, Gertjan J. Burghouts, and Arnold W. M. Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005. doi: 10.1023/B:VISI.0000042993.50813.60. URL https://doi.org/10.1023/B:VISI.0000042993.50813.60.

Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2262–2268, 2018. doi: 10.24963/ijcai.2018/313. URL https://www.ijcai.org/proceedings/2018/313.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 2169–2178. IEEE, 2006. doi: 10.1109/CVPR.2006.68. URL https://doi.org/10.1109/CVPR.2006.68.

Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. In *Proceedings of the Thirty-Second International Joint*

Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, pp. 3911–3919. ijcai.org, 2023. doi: 10.24963/IJCAI.2023/435. URL https://doi.org/10.24963/ijcai.2023/435.

Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, pp. 1968–1974, 2014. doi: 10.1609/aaai.v28i1.8973. URL https://ojs.aaai.org/index.php/AAAI/article/view/8973.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. COMPLETER: incomplete multi-view clustering via contrastive prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 11174–11183. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01102. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lin_COMPLETER_Incomplete_Multi-View_Clustering_via_Contrastive_Prediction_CVPR_2021_paper.html.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4447–4461, 2023. doi: 10.1109/TPAMI.2022.3197238. URL https://doi.org/10.1109/TPAMI.2022.3197238.

Xinwang Liu, Miaomiao Li, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel k-means with incomplete kernels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2259–2265, 2017. doi: 10.1609/aaai.v31i1.10893. URL https://ojs.aaai.org/index.php/AAAI/article/view/10893.

Jingyu Pu, Chenhang Cui, Xinyue Chen, Yazhou Ren, Xiaorong Pu, Zhifeng Hao, Philip S. Yu, and Lifang He. Adaptive feature imputation with latent graph for deep incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14633–14641, 2024. doi: 10.1609/aaai.v38i13.29380. URL https://ojs.aaai.org/index.php/AAAI/article/view/29380.

Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21090–21110. PMLR, 2022. URL https://proceedings.mlr.press/v162/tang22c.html.

Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Trans. Image Process.*, 30:1771–1783, 2021. doi: 10.1109/TIP.2020.3048626. URL https://doi.org/10.1109/TIP.2020.3048626.

Qianqian Wang, Xurui Liao, Wei Feng, and Quanxue Gao. Debiased and denoised representation learning for incomplete multi-view clustering. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Bp3I456do5.

Jie Wen, Yong Xu, and Hong Liu. Incomplete multiview spectral clustering with adaptive graph learning. *IEEE Transactions on Cybernetics*, 50(4):1418–1429, 2020. doi: 10.1109/TCYB.2018.2884715. URL https://doi.org/10.1109/TCYB.2018.2884715.

Jie Wen, Shijie Deng, Waikeung Wong, Guoqing Chao, Chao Huang, Lunke Fei, and Yong Xu. Diffusion-based missing-view generation with the application on incomplete multi-view clustering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52762–52778. PMLR, 2024. URL https://proceedings.mlr.press/v235/wen24c.html.

Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9214–9223. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00910. URL https://doi.org/10.1109/ICCV48922.2021.00910.

Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. Deep incomplete multi-view clustering via mining cluster complementarity. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 8761–8769. AAAI Press, 2022a. doi: 10.1609/AAAI.V36I8.20856. URL https://doi.org/10.1609/aaai.v36i8.20856.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16030–16039. IEEE, 2022b. doi: 10.1109/CVPR52688.2022.01558. URL https://doi.org/10.1109/CVPR52688.2022.01558.

Jie Xu, Chao Li, Liang Peng, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *IEEE Trans. Image Process.*, 32:1354–1366, 2023. doi: 10.1109/TIP.2023.3243521. URL https://doi.org/10.1109/TIP.2023.3243521.

Zhe Xue, Junping Du, Changwei Zheng, Jie Song, Wenqi Ren, and MeiYu Liang. Clustering-induced adaptive structure enhancing network for incomplete multi-view data. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 3235–3241, 2021. doi: 10.24963/ijcai.2021/445. URL https://www.ijcai.org/proceedings/2021/445.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279. ACM, 2010. doi: 10.1145/1869790.1869829. URL https://doi.org/10.1145/1869790.1869829.

Salimeh Yasaei Sekeh and Alfred O. Hero, III. Geometric estimation of multivariate dependency. *Entropy*, 21(8):787, August 2019. ISSN 1099-4300. doi: 10.3390/e21080787. URL https://doi.org/10.3390/e21080787.