Robust Multi-View Clustering with Noisy Correspondence

Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, Peng Hu

***Abstract*—Deep multi-view clustering leverages deep neural networks to achieve promising performance, but almost all existing methods implicitly assume that all views are aligned correctly. This assumption is unrealistic in many real-world scenarios, where noise, occlusion, or sensor differences can inevitably cause misaligned data. Based on this observation, we reveal and study a practical but understudied problem in multi-view clustering (MVC), *i.e.*, noisy correspondence (NC). Considering this problem, we argue that the main challenge is to prevent the model from overfiting NC. To this end, we propose a novel Robust Multi-view Clustering with Noisy Correspondence (RMCNC) method, which alleviates the influence of the misaligned pairs from multi-view data. To be specific, we first compute a united probability with all positive pairs to learn cross-view alignment consistency, thereby alleviating the adverse impact of the individual false positives. To further mitigate the overfitting problem, we propose a noise-tolerance multi-view contrastive loss that avoids overemphasizing noisy data. Moreover, RMCNC is a unified framework, which can deal with both partially view-aligned and NC problems in multi-view clustering. To the best of our knowledge, it could be the first study on NC in multi-view clustering. The experimental results on eight benchmark datasets indicate our RMCNC achieves competitive performance and robustness. The code of RMCNC is released at https://github.com/sunyuan-cs/2024-TKDE-RMCNC.**

***Index Terms*—Noisy correspondence, multi-view clustering, contrastive learning, partially view-aligned**

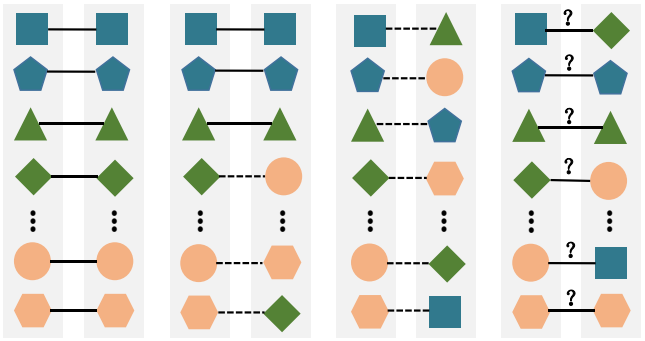
I. Introduction

With the quick advancement of information technology, data could typically be gathered from multiple views or modalities, such as images, texts, and audio, which are called multi-view data [1–3]. Different from single-view data [4–6], multi-view data could provide richer and more comprehensive descriptions of the targets from different views [7, 8], which has attracted more and more attention from academic and industrial communities. To comprehensively understand these data, multi-view clustering (MVC) [9] is an effective way to handle and analyze multi-view data, which aims to group the unlabeled multi-view instances into clusters, where each cluster has similar semantics [10–13]. Thanks to its powerful data-handling ability, it has been successfully applied in various applications, such as data mining and knowledge discovery [14, 15].

Y. Sun, Y. Qin, Y. Li, X. Peng, and P. Hu are with College of Computer Science, Sichuan University, Chengdu, China, 610044 (e-mail: sunyuan_work@163.com).

D. Peng is with College of Computer Science, Sichuan University, Chengdu, China, 610044, and also with Sichuan Newstrong UHD Video Technology Co., Ltd., Chengdu 610095, China

Corresponding author: Peng Hu (e-mail: penghu.ml@gmail.com).



(a) FVP (b) PVP (c) FUP (d) NCP

Fig. 1: Toy examples of different view alignments in multi-view data. Each shape represents an instance, each color denotes a category, and the line indicates the correspondence between different views. The question mark means that the corresponding correspondence is uncertain. (a) fully view-aligned problem (FVP), where all views are perfectly aligned; (b) partially view-aligned problem (PVP), where some known data are view-aligned and some are not; (c) fully unpaired problem (FUP), where no data are view-aligned; (d) noisy correspondence problem (NCP), where the alignment information is noisy and unknown.

Although existing MVC methods [16–18] have achieved satisfactory progress, almost all of them implicitly assume all views are correctly aligned, as shown in Figure 1(a). However, this assumption is impractical in many real scenarios, where some unaligned data may be misaligned as paired data due to unavoidable noise, occlusion, or sensor faults. Recently, to relax the assumption, some works revealed and studied a novel partially view-aligned problem (PVP) [19], where multi-view data consist of a known portion of aligned data and a portion of misaligned data, as shown in Figure 1(b). Although PVP-oriented methods could alleviate the requirement for well-aligned data, it is still time- and cost-prohibitive to obtain a considerable portion of well-aligned multi-view data, especially in healthcare and manufacturing. Moreover, some works investigated a challenging fully unaligned problem (FUP) [20], which means that all views are unaligned in multi-view data, as shown in Figure 1(c). Although these methods do not require any well-aligned data, they should align all view-unaligned data across different views [21, 22], leading to extremely high computational complexity, especially for large-scale data. In addition to high computational complexity, learning from fully view-unaligned data may have no gain or even worse performance than a single view since it is

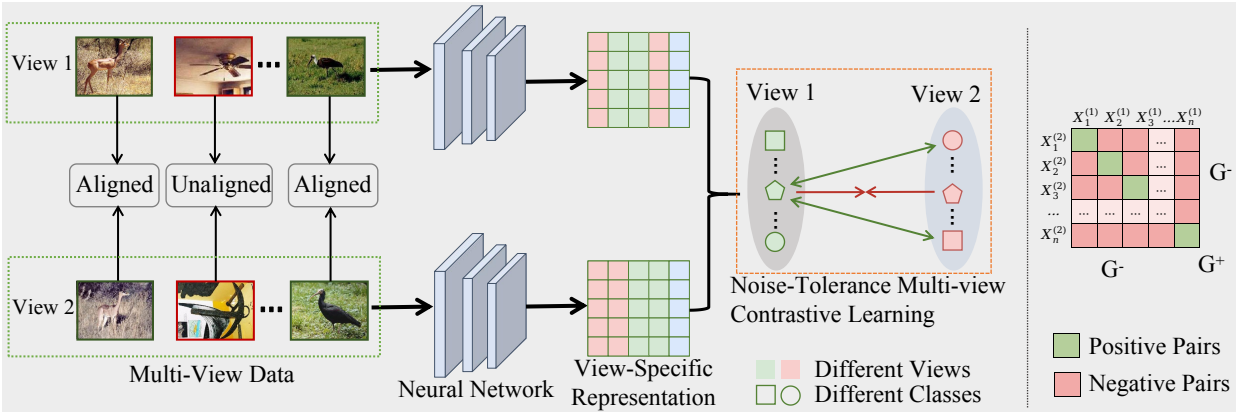


Fig. 2: The framework of the proposed RMCNC. View-specific representations are first learned from the two-view raw features through the neural network. Then, according to the alignment relationships of the cross-view positive and negative pairs, we compute the united probability with the positive pairs. Finally, we propose noise-tolerance multi-view contrastive loss to prevent overemphasizing noisy data.

hard to correctly align different views without any cross-view correspondence. Thus, FUP is too pathological to be practical in real-world scenarios. In practice, aligned and misaligned data are inextricably involved with each other in multi-view data. That is to say, some unaligned data may be mistakenly treated as aligned ones in multi-view data, thereby leading to the noisy correspondence problem (NCP) [23]. Obviously, NCP will remarkably degrade the clustering performance due to its false guidance, which however is less touched so far.

To tackle the practical and challenging problem, we propose a novel Robust Multi-View Clustering with Noisy Correspondence method (RMCNC) to robustly learn representations from noisy views, whose framework is shown in Figure 2. Specifically, inspired by contrastive learning [24–26], we first construct the positive and negative pairs across distinct views by using noisy alignments. However, we observe that NC causes traditional contrastive learning methods to overfit noisy guidance (*i.e.*, false positives and false negatives) as shown in Figure 3. To overcome this challenge, we present maximizing a united probability of all positive pairs instead of the individual probability of each positive, thereby mitigating the noise disturbance brought by NC in multi-view learning. Although the united probability could mitigate the overfitting problem, widely-used contrastive losses (*e.g.*, InfoNCE [27]) could overemphasize the hard pairs and still lead to overfitting NC, as shown in Figures 3 and 5. Finally, to alleviate the overemphasizing issue, we propose a noise-tolerance multi-view contrastive learning loss to reduce the focus on noisy pairs, thus boosting the robustness against NC. In brief, the main contributions of this paper are summarized as follows:

- We propose a new problem setting for MVC with NC, which relaxes the strong assumption of view alignment in multi-view data, embracing more practicality. To the best of our knowledge, this is the first work to achieve robust MVC under both NCP and PVP.
- We present maximizing the united probability to exploit the cross-view alignment consistency, thereby efficiently alleviating the adverse impact of the individual false positives.
- We develop a noise-tolerance multi-view contrastive loss that reduces the focus on noisy pairs and alleviates the overfitting problems, thus enhancing the robustness against NC.
- We provide a robustness analysis for the proposed loss against NC. Moreover, various experiments on the widely-used datasets validate the superiority of RMCNC.

II. Related Work

In this section, we briefly introduce some recent related works in MVC and learning with noisy correspondence.

A. Multi-View Clustering

In recent years, MVC [28, 29] has received widespread attention and has become a hot topic. In addition, to overcome the incomplete problem of multi-view data, a large number of incomplete MVC methods [30] have been proposed. Although these methods [31] have achieved pleasing performance, almost all of them heavily depend on an implicit view-alignment assumption [9, 32, 33], *i.e.*, all views are aligned correctly. In real scenarios, this assumption is easily destroyed due to unavoidable noise, occlusion, or sensor faults, thus leading to PVP [19] and FUP [20]. Partially view-aligned clustering (PVC) [19] is the first attempt to reveal this practical and challenging PVP. PVC proposes a feasible solution that redefines the Hungarian algorithm to achieve instance alignment across views. However, such instance-level alignment has a high computational complexity $\mathcal{O}(n^3)$, resulting in the inability to handle large-scale data. Multi-view contrastive learning with noise-robust loss (MvCLN) [25] considers that the essence of clustering is a one-to-many mapping, and further reveals that category-level alignment possesses higher accessibility. To this end, MvCLN proposes establishing category-level alignment under the contrastive learning framework. Although these methods could alleviate the requirement for well-aligned data, it is still expensive or even impossible to collect a considerable portion of well-aligned multi-view data. Another challenging problem (*i.e.*, FUP) assumes that all views are unaligned in multi-view data. MVC for unknown mapping

relationships (MVC-UM) [20] makes the first attempt to handle FUP, which adopts the graph-based coupling term to excavate the consistency information for all views. To explore the high-order correlations of all views, T-UMC [22] proposes a unified tensor framework that utilizes local structures and data coupling. UIMC [21] studies a more challenging real-world scenario (*i.e.*, the case of FUP and incomplete multi-view data), which proposes a joint framework to fill and realign feature data. However, FUP-oriented methods always have extremely high computational complexity due to aligning different views without any cross-view correspondence. In conclusion, PVP and FUP are respectively too ideal and pathological to be practical in real-world scenarios.

In contrast to PVC and FUC, NCP considers a more practical problem, where misaligned data is unknown and occurs randomly in multi-view data. Since the adverse influence of the misaligned data, NC could cause traditional contrastive learning methods to overfit noisy guidance, thereby leading to degraded clustering performance. To prevent the model from overfitting NC, we propose a robust-tolerance MVC method to avoid overemphasizing noisy data without all the bells and whistles.

B. Learning with Noisy Correspondence

As a new problem in the field of cross-modal retrieval, noisy correspondence is a special kind of noisy label, where mismatched pairs are incorrectly considered as matched pairs. Huang et al. [23] first reveal and study the noisy correlation problem (NCP) in cross-modal matching and propose a noisy correspondence rectifier (NCR), which can rectify matching relationships to achieve robust cross-modal retrieval. However, due to the noise effect, it may make the supervised information unreliable/uncertain, which significantly degrades the performance. To overcome the issue, Qin al et. [34] propose deep evidential cross-modal learning (DECL) to accurately estimate the uncertainty caused by noise, thereby enhancing the robustness and reliability. In addition, Hu et al. provide an unbiased estimate for cross-modal retrieval risk and derive robust cross-modal learning (RCL) [35] to endow cross-modal methods with robustness against NCP. To prevent the confirmation bias problem, Yang et al. propose bidirectional cross-modal similarity consistency (BiCro) [36] to estimate soft labels for image-text data with noisy correspondence, which can reflect true correspondence degree. In general, except cross-modal retrieval tasks [37], more and more other tasks with NC have also been widely studied, which includes audio-visual action recognition [38], video reasoning [39], graph matching [40], and person re-identification [41–43].

Although cross-modal retrieval with noisy correspondence has been studied recently, this problem has not been touched on in MVC. Such multi-view data with noisy correspondence will significantly decrease the clustering performance. To handle the practical yet challenging problem in MVC, we propose a novel RMCNC method, which makes the first attempt to perform MVC with NC.

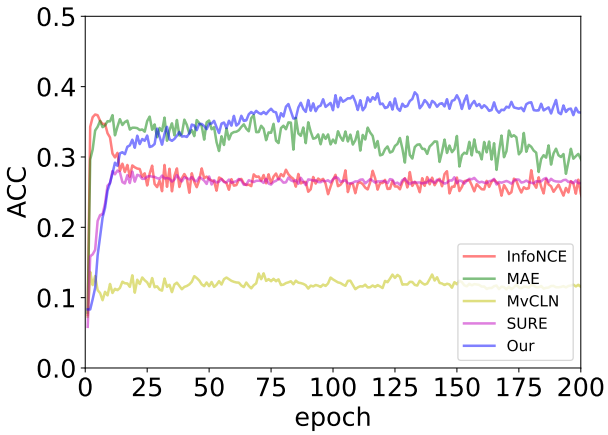


Fig. 3: Performance of different loss functions conducted on the Animal with 40% noise.

III. Methodology

In this section, we present a robust MVC with noisy correspondence (RMCNC) method to handle both PVP and NCP. We first formally introduce the formulation of PVP and NCP. Then, we give a detailed description of the proposed RMCNC. Finally, we present some theoretical analysis for our noise-tolerance clustering loss.

A. Problem Formulation

Let $\left\{\boldsymbol{X}^{(v)}\right\}_{v=1}^{V} = \left\{\boldsymbol{x}_1^{(v)}, \ldots, \boldsymbol{x}_i^{(v)}, \ldots, \boldsymbol{x}_N^{(v)}\right\}_{v=1}^{V}$ be a multi-view dataset with N instances observed from V views, where $\boldsymbol{X}^{(v)}$ denotes the data from v-th view, $\boldsymbol{x}_i^{(v)}$ represents the i-th sample from the v-th view. For clarity, we define correspondence label $\boldsymbol{L} = \{l_{ij}^{vu} | i, j = 1, 2, \cdots, N; v, u = 1, 2, \cdots, V\}$ to represent correspondences across different views, where l_{ij}^{vu} is the correspondence between $\boldsymbol{x}_i^{(v)}$ and $x_j^{(u)}$. In the multi-view dataset, it implicitly assumes that each view from the same instance is aligned, denoted as $l_{ii}^{vu} = 1$ for the v-th and u-th views from the i-th instance, unaligned otherwise, denoted as $l_{ij}^{vu} = 0$ $(i \neq j)$ for $\boldsymbol{x}_i^{(v)}$ and $\boldsymbol{x}_j^{(u)}$ from different instances. However, in real-world scenarios, the views of an instance may be misaligned due to the ubiquitous noise (*i.e.*, NCP), where $l_{ii}^{vu} = 1$ may be wrong. Mathematically, we give the general definition for NCP as follows.

Definition 1 (NCP): *The given multi-view dataset $\left\{\boldsymbol{X}^{(v)}\right\}_{v=1}^{V}$ contains the view-aligned and view-misaligned data, meaning we do not know whether $l_{ii}^{vu} = 1$ is true or false for any pair. Mathematically, NCP could be formulated as*

$$\sum_{v}^{V} \sum_{u \neq v}^{V} Q\left(\boldsymbol{x}_i^{(v)}, \boldsymbol{x}_i^{(u)}\right) < V(V-1), \forall i \in [1, N], \quad (1)$$

where Q is an alignment indicator, $Q(a, b) = 1$ if two cross-view samples (i.e., a and b) come from the same instance, and $Q(a, b) = 0$ otherwise.

B. Robust Multi-view Learning with Noisy Correspondence

For multi-view data, MVC aims to learn discriminative and consistent representations across different views, thereby facilitating intra-cluster compactness while inter-cluster scattering. Recently, inspired by the great success achieved by contrastive learning, it has been applied to MVC and achieved promising performance [24–26]. However, almost all of these methods implicitly assume that different views are perfectly aligned while ignoring the presence of ubiquitous noise, which could cause them to overfit false positives as shown in Figure 3. From the figure, one could observe that typical contrastive learning losses (*e.g.*, InfoNCE) suffer from overfitting, which is manifested as an initial increase followed by a decrease in performance, due to overemphasis on false positives. In contrast, some robust losses (*e.g.*, MAE and SURE) could achieve stable performance, indicating that robust loss could alleviate the overfitting problem. However, we also observe that the robust losses still suffer from underfitting issues, resulting in lower performance than the best of InfoNCE. Furthermore, some robust MVC methods (*e.g.*, MvCLN and SURE) have been proposed to handle PVP, however, they still cannot conquer the overfitting and underfitting problems caused by NC. To address the problems, we propose a novel noise-tolerance multi-view contrastive learning loss to improve the robustness against NC.

First, we adopt multiple view-specific neural networks $h_v(\cdot, \Theta_v)$ to project different views into the specific latent spaces (*i.e.*, $\{\boldsymbol{Z}^v\}_{v=1}^V$, where Θ_v is the parameters of the corresponding networks. To measure cross-view similarities, we calculate the cosine similarity between two features from different views as follows:

$$S(\boldsymbol{z}_i^v, \boldsymbol{z}_j^u) = \frac{\boldsymbol{z}_i^v (\boldsymbol{z}_j^u)^\top}{\|\boldsymbol{z}_i^v\| \|\boldsymbol{z}_j^u\|}, \tag{2}$$

where $\boldsymbol{z}_i^v$ and $\boldsymbol{z}_j^u$ are the view-specific representations. Thus, we could define the probability that two cross-view samples belong to the same instance as follows:

$$p_{ij}^{vu} = \frac{\exp^{\left(S(\boldsymbol{z}_i^v, \boldsymbol{z}_j^u)/\tau\right)}}{\sum_{j=1}^N \exp^{\left(S(\boldsymbol{z}_i^v, \boldsymbol{z}_j^u)/\tau\right)}}, \tag{3}$$

Inspired by contrastive learning [27, 44], we could utilize the widely-used contrastive loss InfoNCE for MVC as follows:

$$\mathcal{L}_i^{vu} = -\log \frac{\exp^{(S(\boldsymbol{z}_i^v, \boldsymbol{z}_i^u)/\tau)}}{\sum_{j=1}^N \exp^{\left(S(\boldsymbol{z}_i^v, \boldsymbol{z}_j^u)/\tau\right)}}, \tag{4}$$

where τ denotes a temperature parameter [10]. Minimizing Equation (4) is equivalent to maximizing the agreement between positive pairs while minimizing the agreement between negative pairs, thus facilitating learning discriminative representations. Like cross-entropy, InfoNCE tends to focus on hard pairs, which hardly rely on well-labeled data. However, based on the memorization effect of DNN [45, 46], the hard pairs probably are false positives, thus misleading the model to optimize in the wrong direction, *i.e.*, overfitting problem.

To alleviate the overfitting problem of InfoNCE, inspired by [47, 48], we could adopt robust MAE loss to mitigate the impact of noise. Thus, Equation (4) could be rewritten as follows:

$$\hat{\mathcal{L}}_i^{vu} = \frac{1}{N} \sum_{j=1}^N \left| l_{ij} - p_{ij}^{vu} \right|, \tag{5}$$

where $|\cdot|$ denotes the absolute value. From Equation (5), we could observe that MAE treats each sample equally, thereby leading to theoretical robustness against noisy labels [35]. This robustness is also demonstrated in Figure 3, wherein MAE achieves promising robustness against NC. Although this equal treatment brings robustness, it inevitably loses focus on information pairs, thus resulting in an underfitting issue and performance degradation [49]. In general, since InfoNCE mainly focuses on the hard samples, it easily leads to overfitting NC. Hence, InfoNCE is more suitable for clean samples. MAE treats each sample equally to enhance robustness. However, MAE lacks the ability to handle the more challenging samples, thereby leading to underfiting and poor performance. Although some PVP-oriented methods (*e.g.*, MvCLN and SURE) have achieved promising results with partially aligned views, they also will face overfitting or underfitting problems due to unknown noisy pairs.

To mitigate the noise disturbance brought by NC in multi-view data, we propose a novel robust multi-view clustering method with noisy correspondence. Specifically, we first compute a united probability with all positive pairs to learn cross-view alignment consistency, thereby alleviating the adverse impact of the individual false positives. The united probability can be formulated as

$$\begin{aligned} G_+^{vu} &= \sum_{i=1}^N \sum_{j=1}^N l_{ij} \exp\left(S(\boldsymbol{z}_i^v, \boldsymbol{z}_j^u)/\tau\right), \\ G_-^{vu} &= \sum_{i=1}^N \sum_{j=1, l_{ij}=0}^N \exp\left(S(\boldsymbol{z}_i^v, \boldsymbol{z}_j^u)/\tau\right). \end{aligned} \tag{6}$$

To further mitigate the overfitting problem, we propose a noise-tolerance multi-view contrastive loss to avoid overemphasizing noisy pairs, thereby enhancing robustness against misaligned pairs. Mathematically, we adopt an exponential loss to formulate the problem as

$$\begin{aligned} \mathcal{L}_r^{vu} &= \frac{1 - (P_+^{vu})^q}{q} = \frac{1}{q}(1 - (P^+)^q) \\ &= \frac{1}{q}\left(1 - \left(\frac{G_+^{vu}}{G_+^{vu} + G_-^{vu}}\right)^q\right), \end{aligned} \tag{7}$$

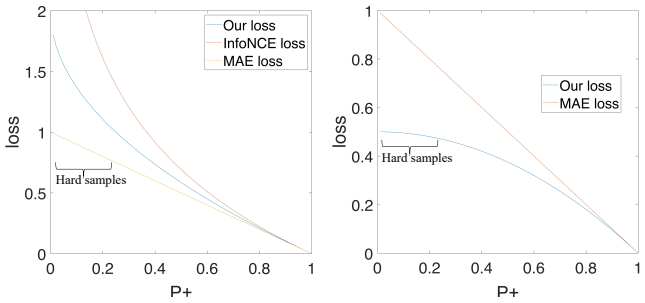
where $q > 0$ is a regulatory factor. For the q value of the loss, we can choose it based on the robustness analysis and parametric analysis. The overall loss function of RMCNC can be formulated as:

$$\mathcal{L}_r = \sum_{v=1}^V \sum_{u \neq v}^V \mathcal{L}_r^{vu} \tag{8}$$

Due to limited computing resources, we cannot employ the whole training data for training. Therefore, we conduct Monte Carlo sampling to relax the whole training set as batch data.

Intuitively, our loss mainly uses the parameter q to seek a balance between InfoNCE and MAE, which could focus

less on hard samples than InfoNCE and more on informative samples than MAE, thus mitigating the overfitting and underfitting problems. Specifically, we plot the loss curves of InfoNCE, MAE, and our RMCNC to visually illustrate the robustness of our method in Figure 4. From the figure, one could see that RMCNC reduces the relative loss for hard samples, which are often very likely to be noisy ones, when $0 < q < 1$, thus making the model focus more on clean data than noisy ones. Compared to MAE, our RMCNC could treat each sample discriminatively instead of equal treatment, thus focusing more on informative samples. More specifically, our RMCNC could pay more attention to the hard samples than MAE when $0 < q < 1$, while less attention to the easy samples when $q > 1$. Thanks to this multi-granularity attention, our RMCNC could alleviate the overemphasis on hard samples and focus more on easy samples, thus embracing robustness against NC as shown in Figure 3.



(a) 0<q<1 (b) q>1

Fig. 4: Comparison between InfoNCE, MAE, and RMCNC.

C. Robustness Analysis

To intuitively study the robustness of our $\mathcal{L}_r$, we conduct the robustness analysis on gradient qualitatively. Given the united probability P_+^{vu}, the gradient of our robust loss $\mathcal{L}_r$ is calculated as:

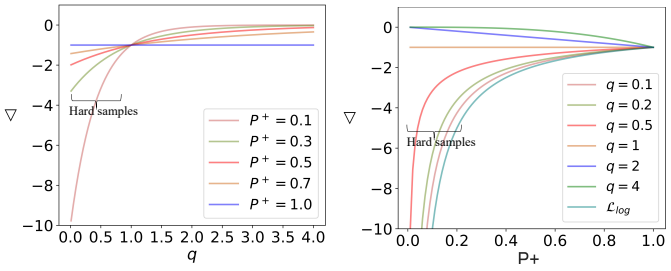
$$\nabla \mathcal{L}_r = \frac{\partial \mathcal{L}_r}{\partial P_+^{vu}} = -(P_+^{vu})^{(q-1)} \tag{9}$$

From Equation (9), it could be seen that our gradient is still regulated by the factor q to exhibit different behaviors as shown in Figure 5(a). When q ($q > 1$) is larger, $\mathcal{L}_r$ will produce a larger magnitude of the gradient (*i.e.*, $|\nabla|$) for easy samples than hard ones, which makes the model focus more on clean data than noisy ones, thus boosting the robustness against NC.

On the contrary, $\mathcal{L}_r$ with small q could focus more on the hard samples than MAE to alleviate the underfitting problem, while less than InfoNCE to handle the overfitting problem. To further explore $\mathcal{L}_r$, we exploit the negative-log-likelihood loss $\mathcal{L}_{log}$ of P_+^{vu} as a comparison. Given the united probability P_+^{vu}, the gradient of $\mathcal{L}_{log}$ is

$$\nabla \mathcal{L}_{log} = \frac{\partial \mathcal{L}_{log}}{\partial P_+^{vu}} = -\frac{1}{P_+^{vu}} \tag{10}$$

Markedly, $\mathcal{L}_r$ is equivalent to $\mathcal{L}_{log}$ when q approaches zero. As shown in Figure 5(b), the magnitude of the gradient produced by $\mathcal{L}_{log}$ is greater for hard samples than that of $\mathcal{L}_r$. That is to say, the lower the united probability, the smaller the magnitude of our RMCNC. Therefore, it indicates that our loss could not overemphasize the hard samples, thus embracing more robustness than $\mathcal{L}_{log}$. Besides, as shown in Figure 3, the performance of our RMCNC gradually increases and then tends to be stable without performance degradation as the epoch increases, which indicates that RMCNC has robustness against the noise disturbance brought by NC.



(a) ∇ versus q (b) ∇ versus P^+

Fig. 5: The behavior of ∇ with different q and P_+^{vu}.

Algorithm 1 RMCNC for MVC with noisy correspondence

Input: multi-view dataset $\boldsymbol{X}^{(v)}$; networks $h_v(\boldsymbol{X}^{(v)}, \Theta_v)$; batch size M; training epoch B; temperature parameter τ; parameter q.
Initialize: Initialize the parameters Θ;
/********************Training************************/**
while *epoch* $< B$ **do**
  for *batch =1 to M* **do**
    Sample a mini-batch $\left\{\boldsymbol{x}_i^{(v)}\right\}_{i=1}^{M}$.
    Encode $\left\{\boldsymbol{x}_i^{(v)}\right\}_{i=1}^{M}$ as the representations $\left\{\boldsymbol{z}_i^{(v)}\right\}_{i=1}^{M}$.
    Construct positive and negative pairs through l_{ij}.
    Obtain the overall loss $\mathcal{L}_r$ with Equation (8).
    Minimize $\mathcal{L}_r$ to compute Θ_v by gradient descent.
  end
end
/********************Realign************************/**
for *batch =1 to M* **do**
  Obtain the representations from each batch through $h_1(\boldsymbol{x}_i^{(1)}, \Theta_1)$ and $h_2(\boldsymbol{x}_i^{(2)}, \Theta_2)$.
  Calculate the Euclidean distance $\mathbf{D}$ of the cross-view representations.
end
for $\boldsymbol{x}_i^{(1)}$ *in* $\boldsymbol{X}^{(1)}$ **do**
  Realign multi-view data by the smallest distances.
  Obtain the common representation by Concatenating the corresponding representations.
end
Performing k-means clustering for the common representation.
Output: Clustering results.

IV. Experiments

In this section, we evaluate the performance of our RMCNC on four widely-used benchmark datasets and compare it

TABLE I: The partially view-alignend clustering results (%) on four benchmarks.

Methods	Scene-15			Caltech-101			Animal			NoisyMNIST		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CCA (NeurIPS'03)	32.73	34.24	18.80	20.06	41.56	16.62	15.34	9.72	8.24	34.46	29.83	17.89
KCCA (JMLR'02)	33.09	31.43	16.35	12.57	31.36	7.65	14.72	9.55	7.81	26.57	18.19	10.55
DCCA (ICML'13)	34.27	36.55	18.83	12.52	32.13	7.63	14.93	10.44	9.11	29.22	20.24	11.08
DCCAE (ICML'15)	33.62	36.56	18.54	11.75	30.54	6.60	21.72	25.15	9.22	27.61	19.45	10.00
LMSC (CVPR'17)	26.27	20.45	10.93	21.54	40.26	15.51	33.41	28.44	15.20	/	/	/
MvC-DMF (AAAI'17)	28.49	24.31	11.22	9.54	23.41	3.84	35.30	27.38	14.96	27.34	22.96	6.85
SwMC (IJCAI'17)	31.03	30.39	12.94	19.03	22.75	3.73	34.22	31.01	16.32	/	/	/
BMVC (TPAMI'18)	36.81	36.55	20.20	12.13	31.33	7.11	35.62	27.24	15.01	28.47	24.69	14.19
AE2-Nets (CVPR'19)	28.56	26.58	12.96	10.45	29.51	7.90	4.20	3.30	0.00	38.25	34.32	22.02
PVC (NeurIPS'20)	37.88	39.12	20.63	22.11	47.82	17.98	3.80	0.10	0.30	81.84	82.29	82.03
MvCLN (CVPR'21)	38.53	39.90	24.26	30.09	43.07	38.34	26.18	40.19	19.71	91.05	84.15	83.56
T-UMC (TCYB'22)	35.77	34.41	24.27	26.31	48.32	33.47	30.62	37.17	20.54	/	/	/
UIMC (TNNLS'22)	31.31	31.92	23.54	26.06	**49.03**	30.74	30.62	37.21	19.92	/	/	/
SURE (TPAMI'23)	40.32	40.33	23.08	30.87	44.25	39.89	27.74	40.83	19.91	95.17	88.24	89.72
Our RMCNC	**40.51**	**41.13**	**24.55**	**33.92**	48.10	**45.33**	**44.43**	**49.48**	**30.41**	**95.53**	**88.76**	**90.43**

with 14 state-of-the-art MVC methods. For a comprehensive evaluation, we employ ACC, NMI, and ARI as the evaluation metrics to measure the clustering performance. This section is organized as follows. First, we briefly introduce the ten datasets that we use in our experiments. Then we elaborate on the experimental settings, including the hyper-parameters, the evaluation scenarios, etc. Next, we present the comparative results and the ablation studies. Finally, we provide some additional analysis on different aspects of our method, such as visualization analysis, parameter sensitivity, and influence of different noise proportions.

A. Datasets

In this section, we briefly introduce six multi-view datasets and four multi-modal datasets: **Scene15** [50] is a collection of 4,485 images from 15 different categories in indoor and outdoor scenes. Following [51], we use GIST and PHOG to extract 20-dim and 59-dim features for each image, respectively. **Caltech-101** [52] contains 9,146 images from 102 different objects. We follow [9] to obtain 1,984-dim and 512-dim features of all images by HOG and GIST, respectively. **DeepAnimal** [26] comprises 10,158 images from over 50 animal classes. Similar to [26], we employ DECAF [53] and VGG19 [54] models to extract 4096-dim features from each image as its two views. **NoisyMNIST** [55] includes 70,000 images belonging to 10-digit classes. Since some competitors cannot handle large-scale data, we choose 30,000 samples at random. Following [15], we use the raw images and the corresponding images with white Gaussian noise as two views. **100Leaves** [56] has 1,600 images from 100 classes, which have three views extracted by texture histogram features, shape descriptor, and fine-scale margin. Caltech-3V [10] contains 1,400 images from 7 classes, which has the three-view data (*i.e.*, WM, CENTRIST, and LBP). **WIKI** [4] is a widely-used cross-modal dataset, including 2,866 image-text pairs from 10 semantic classes. Following [4], we adopt the 128-dimensional SIFT feature vector and the 10-dimensional topic vector to represent image-text pairs. **NUS-WIDE** [57, 58] contains 9,000 images and the corresponding tags from 10 classes. Following [59], we use the 19-layer VGGNet to extract image representation and adopt the 4,096-dimensional deep features from the fc7 layer to represent image data. And we adopt the sentence CNN to extract 300-dimensional features as text representation. **XMedia** [60] consists of 40,0000 text-image pairs with 200 classes. Similar to [59], we can obtain 4,096-dimensional deep image features and 300-dimensional text features. **XRMB** [61] contains 85,297 multi-view data (*i.e.*, the acoustic view and the articulation view) from 39 classes, whereas their features are 273 and 112 dimensions, respectively.

B. Experiment Settings

After obtaining the view-specific representations, similar to SURE [26], we conduct the semantic-level alignment scheme to realign the cross-view data. Once we obtain the realigned multi-view data, we can use the concatenated representations to perform downstream tasks, such as clustering. Specifically, our RMCNC can be conducted in three steps to achieve MVC.

Step 1: After we learn the representations $h_1(\boldsymbol{X}^{(1)}, \Theta_1)$ and $h_2(\boldsymbol{X}^{(2)}, \Theta_2)$, we can calculate the Euclidean distance $\mathbf{D} \in \mathbb{R}^{N \times N}$ between them.

Step 2: For each view, we adopt the smallest distance ($\mathbf{D}_{ij}$) corresponding representation as the correspondences of other views, thereby obtaining the realigned data.

Step 3: We generate a consensus representation by concatenating them and adopting k-means to achieve MVC.

In general, the proposed RMCNC is summarized in Algorithm 1. In our experiment settings, we adopt the neural networks ($h_v(\boldsymbol{X}^{(v)}, \Theta_v)$) to encode all views $\boldsymbol{X}^{(v)}$, where the feature dimension of each view is d^v. Specifically, for these datasets, we set the dimension of the encoders to $d^v-1024-1024-1024-512$ for the NUS-WIDE and Caltech-3V datasets, $d^v - 1024 - 1024 - 512$ for the Caltech-101, NoisyMNIST, and XRMB datasets, and $d^v - 1024 - 512$ the

TABLE II: The fully view-alignend clustering results (%) on four benchmarks.

Methods	Scene-15			Caltech-101			Animal			NoisyMNIST		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CCA (NeurIPS'03)	36.37	36.91	19.82	20.25	45.41	16.34	20.40	14.70	12.70	71.31	52.60	48.46
KCCA (JMLR'02)	37.93	37.42	21.38	21.45	45.58	17.62	25.30	15.60	11.80	96.85	92.10	93.23
DCCA (ICML'13)	36.61	39.20	21.03	27.60	47.84	30.86	23.21	17.12	13.15	89.64	88.33	83.95
DCCAE (ICML'15)	34.58	39.01	19.65	19.84	45.05	14.57	30.00	43.80	18.00	78.00	81.24	68.15
LMSC (CVPR'17)	38.46	35.50	20.54	26.87	48.80	18.06	43.91	48.20	33.40	/	/	/
MvC-DMF (AAAI'17)	30.99	31.35	15.68	24.35	44.98	14.82	44.56	55.62	30.09	74.39	63.22	49.79
SwMC (IJCAI'17)	33.89	32.98	11.78	30.74	36.07	7.75	45.98	55.45	34.96	/	/	/
BMVC (TPAMI'18)	40.74	41.67	24.19	27.59	46.43	21.28	41.01	55.22	36.36	88.31	77.01	76.58
AE2-Nets (CVPR'19)	37.17	40.47	22.24	20.79	45.01	15.89	3.80	0.10	0.30	42.11	43.38	30.42
PVC (NeurIPS'20)	38.01	39.82	21.06	21.74	49.31	18.48	3.80	0.00	0.00	87.10	92.84	93.14
MvCLN (CVPR'21)	37.90	42.31	25.58	30.41	46.90	42.99	35.30	54.20	29.40	97.30	94.16	95.31
T-UMC (TCYB'22)	37.30	35.49	26.21	26.89	49.53	33.72	31.10	37.50	20.50	/	/	/
UIMC (TNNLS'22)	32.12	32.03	23.60	26.14	**49.63**	28.22	30.17	36.72	19.45	/	/	/
SURE (TPAMI'23)	42.75	42.48	24.57	34.16	48.04	**51.45**	35.80	53.60	29.50	98.39	95.41	96.50
Our RMCNC	**43.70**	**43.31**	**26.34**	**34.63**	42.04	51.00	**46.69**	**57.70**	**37.00**	**98.65**	**96.00**	**97.04**

Scene-15, Animal, WIKI, XMedia, and 100Leaves datasets, respectively. We implement our RMCNC using PyTorch and train it on one NVIDIA GeForce GTX 3090 GPU. All comparison experiments were conducted on Nvidia GeForce RTX 3090 and Tesla V100 GPUs. For optimization, we use Adam [62] with the initial learning rate, without using any scheduler or weight decay. Furthermore, we train our RMCNC for 100 epochs and set the batch size as 1,024 for all datasets. To evaluate our RMCNC comprehensively, we exploit three different settings as follows:

Partially view-aligned: We randomly partition tested datasets into two equal parts, *i.e.*, view-unaligned set $\{\mathbf{U}^{(v)}\}_{v=1}^{V}$ and view-aligned set $\{\mathbf{A}^{(v)}\}_{v=1}^{V}$. For the view-unaligned data $\{\mathbf{U}^{(v)}\}_{v=1}^{V}$, we randomly shuffle the samples in other views except for the first view, thereby obtaining the fully view-unaligned set $\{\mathbf{U}^{(v)}\}_{v=1}^{V}$. For PVC, MvCLN, SURE, and our RMCNC, we directly apply them to the partially view-aligned data. For other baselines that cannot handle PVP, we first use PCA to project data into a latent space, and then use the Hungarian algorithm to establish the cross-view alignment relationships. Then we apply these baselines to the realigned data. **Fully view-aligned:** The PVP-oriented methods and our RMCNC still need to realign views after training because they assume the presence of unknown unaligned data in the clustering stage. For other baselines, we directly adopt them to the fully view-aligned data. **Noisy correspondence:** In this setting, since the cross-view alignment ground truths are unknown, all baselines are directly applied to the data with noisy correspondence.

C. Baselines and Evaluation Metrics

In this section, we compare our proposed RMCNC with 14 state-of-the-art MVC methods, which can be divided into three categories: fully-view-aligned methods (CCA [63], KCCA [64], DCCA [65], DCCAE [55], LMSC [66], MvC-DMF [67], SwMC [68], BMVC [9], and AE2-Nets [28]), partially-view-aligned methods (PVC [19], MvCLN [25], SURE [26]), and fully view-unaligned methods (T-UMC [22], UIMC [21]). For PVP and FVP, we compare our RMCNC with all baselines. For NCP, we only compare our RMCNC with partially-view-aligned methods and fully view-unaligned methods, as fully-view-aligned methods cannot handle these scenarios. For these methods, we mainly take the two-view datasets as an example to conduct clustering experiments. In addition, we add three methods (*i.e.*, MFLVC [10], DCP [69], and DealMVC [70]) to perform experiments on the three-view datasets.

To measure the clustering performance comprehensively, we utilize three widely-used metrics in the experiments, namely Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). For these metrics, higher values indicate better performance, and the best scores are marked in bold. Note, '/' indicates that we cannot perform the method due to out of memory. In addition, to reduce the influence of randomness on the performance, we repeat each experiment five times for all methods, including our RMCNC, and report the average results.

D. Comparison with State-of-the-Arts

We conduct extensive experiments under FVP, PVP, and NCP to evaluate the effectiveness of our RMCNC. For NCP, we artificially inject incorrect correspondence of different ratios (*i.e.*, 20%, 40%, 60%, and 80% noise rates) to investigate the robustness of tested methods against noisy correspondence. We report the average results of the five-time runs in Tables I to III. Due to the high computational complexity and running time of T-UMC and UIMC, we cannot conduct the experiments on the large-scale NoisyMNIST dataset.

Results under partially view-aligned: We demonstrate the clustering performance in the PVP scenario in Table I. From these results, we can observe that: 1) Our RMCNC achieves the best performance on all datasets with all metrics. Specifically, on the Animal dataset, our RMCNC obtains remarkable improvements of 8.81% (ACC), 8.65% (NMI), and 9.87% (ARI) compared to the second-best results. This

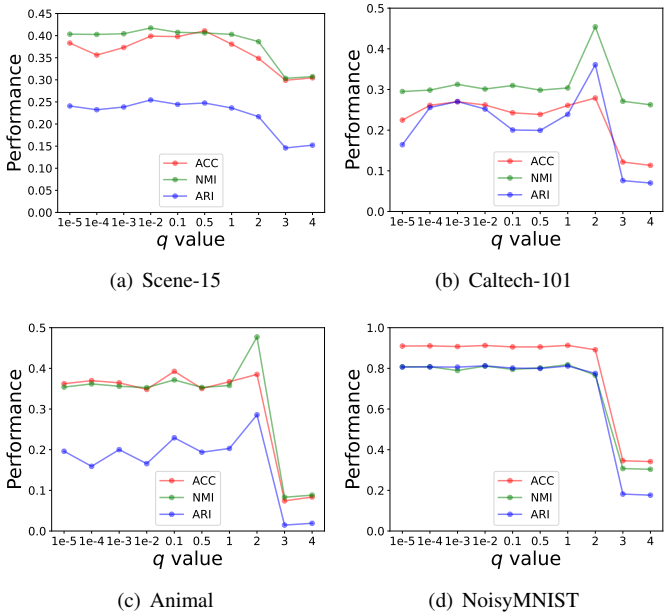
TABLE III: The clustering results (%) with different noise rates on four benchmarks.

Noise	Methods	Scene-15			Caltech-101			Animal			NoisyMNIST		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
20%	PVC (NeurIPS'20)	30.28	23.64	20.41	19.12	37.47	17.94	3.80	0.10	0.30	70.42	75.54	70.27
	MvCLN (CVPR'21)	39.31	36.68	21.16	30.43	41.78	43.09	21.96	26.87	2.43	80.26	77.34	72.25
	T-UMC (TCYB'22)	37.13	35.33	24.43	26.30	39.01	32.38	30.91	37.68	20.72	/	/	/
	UIMC (TNNLS'22)	31.29	31.84	23.42	26.79	40.72	31.22	30.48	37.00	19.35	/	/	/
	SURE (TPAMI'23)	40.02	37.27	21.50	29.68	41.35	41.71	32.26	43.13	22.40	91.00	83.08	83.06
	Our RMCNC	**41.00**	**40.09**	**25.70**	**30.60**	**45.06**	**41.77**	**39.87**	**51.00**	**30.01**	**95.29**	**88.00**	**89.89**
40%	PVC (NeurIPS'20)	22.41	13.65	20.28	13.31	27.73	15.81	3.80	0.10	0.30	43.28	51.45	39.19
	MvCLN (CVPR'21)	37.01	35.21	19.61	18.45	30.15	18.79	16.27	15.38	2.13	54.91	53.39	41.32
	T-UMC (TCYB'22)	35.51	34.78	23.52	24.02	37.67	30.29	30.23	37.18	20.79	/	/	/
	UIMC (TNNLS'22)	31.38	31.89	23.46	26.44	40.95	30.97	30.90	37.10	19.90	/	/	/
	SURE (TPAMI'23)	36.94	36.16	19.86	19.86	30.74	22.81	25.54	29.40	13.85	49.66	45.62	32.59
	Our RMCNC	**39.55**	**40.93**	**24.86**	**27.57**	**46.56**	**36.88**	**36.39**	**46.58**	**26.05**	**91.00**	**79.99**	**83.02**
60%	PVC (NeurIPS'20)	17.88	7.02	17.04	9.43	21.20	11.43	3.80	0.10	0.30	30.21	32.17	19.78
	MvCLN (CVPR'21)	30.40	33.20	17.03	10.91	23.89	5.40	12.10	14.91	1.37	40.59	34.92	23.32
	T-UMC (TCYB'22)	35.59	34.29	20.62	22.88	35.17	26.01	30.77	37.09	20.05	/	/	/
	UIMC (TNNLS'22)	31.26	31.84	21.40	21.89	39.41	26.79	30.18	36.97	19.40	/	/	/
	SURE (TPAMI'23)	31.05	30.55	15.35	10.83	23.69	4.53	16.64	18.20	5.52	35.55	29.22	17.57
	Our RMCNC	**38.36**	**38.80**	**22.97**	**24.02**	**41.59**	**28.91**	**34.91**	**43.05**	**22.73**	**84.39**	**69.47**	**62.64**
80%	PVC (NeurIPS'20)	12.91	2.52	12.19	7.33	17.13	8.88	3.80	0.10	0.30	25.62	23.58	15.26
	MvCLN (CVPR'21)	25.80	28.71	13.07	8.06	20.97	2.80	10.08	12.76	1.39	34.30	25.99	16.69
	T-UMC (TCYB'22)	32.15	30.51	18.63	14.21	28.47	12.53	25.63	29.34	14.82	/	/	/
	UIMC (TNNLS'22)	31.31	31.80	18.39	14.86	30.25	13.26	28.54	32.49	15.16	/	/	/
	SURE (TPAMI'23)	24.39	26.90	11.66	8.20	21.42	3.54	10.17	12.08	2.16	27.34	27.34	15.88
	Our RMCNC	**33.53**	**33.41**	**19.24**	**15.67**	**32.35**	**13.59**	**30.47**	**35.42**	**16.95**	**74.31**	**60.87**	**56.36**

demonstrates that our RMCNC could learn better representations to realign the view-unaligned data. 2) The clustering task on the large-scale NoisyMNIST dataset is more challenging due to the presence of unaligned data. Clearly, all fully-view-aligned methods perform poorly against PVP on the dataset. On the other hand, partially-view-aligned methods and our RMCNC could learn robust representations and achieve impressive results. In particular, our RMCNC achieves the best performance, *i.e.*, 95.53%, 88.76%, and 90.43% scores for ACC, NMI, and ARI metrics, respectively.

Results under fully view-aligned: For the fully view-aligned setting, we report the average results of all tested methods on four multi-view benchmarks in Table II. From the table, one could see that: 1) Our RMCNC outperforms all baselines in most metrics. More specifically, on the Animal dataset, our RMCNC achieves an improvement of 0.71%, 2.08%, and 0.64% for ACC, NMI, and ARI metrics, respectively. This indicates that the proposed method can enhance the representation capability of the well-aligned data. 2) Most methods generally have better results under this setting than the PVP setting. This is because FVP has more ground-truth correspondences to provide ample supervised information. 3) Most methods obtain promising performance on NoisyMNIST. The possible reason is that the large-scale well-aligned data could provide more semantic information.

Results under noisy correspondence: We test the robustness of the baselines and our RMCNC under different noise rates (*i.e.*, 20%, 40%, 60%, and 80%) on four datasets with synthetic noisy correspondences. We report the quantitative evaluation results in Table III. From the table, we could observe that: 1) RMCNC remarkably outperforms partially-view-aligned methods and fully-view-unaligned methods on all metrics. This demonstrates the superior robustness of our RMCNC against NC. 2) RMCNC achieves the best perfor-



(a) Scene-15 (b) Caltech-101

(c) Animal (d) NoisyMNIST

Fig. 6: Parameter analysis of 40% noise in terms of q.

TABLE IV: The multi-modal clustering results (%) with different noise rates on four benchmarks.

Noise	Methods	WIKI			NUS-WIDE			XMedia			XRMB		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0%	PVC (NeurIPS'20)	12.60	0.09	0.00	0.10	0.00	0.00	/	/	/	/	/	/
	MvCLN (CVPR'21)	52.72	46.03	34.14	54.13	38.31	35.73	18.24	60.32	12.58	22.37	35.6	12.4
	T-UMC (TCYB'22)	17.67	3.78	1.76	47.12	33.25	30.71	/	/	/	/	/	/
	UIMC (TNNLS'22)	16.95	4.39	1.77	55.81	42.69	33.1	/	/	/	/	/	/
	SURE (TPAMI'23)	54.21	47.00	36.34	57.44	44.82	38.32	53.59	83.89	32.87	23.49	35.54	13.06
	Our RMCNC	**59.26**	**52.99**	**45.32**	**67.59**	**53.64**	**48.99**	**79.42**	**88.57**	**71.50**	**28.66**	**41.04**	**19.23**
20%	PVC (NeurIPS'20)	12.60	0.09	0.00	0.10	0.00	0.00	/	/	/	/	/	/
	MvCLN (CVPR'21)	44.90	29.86	22.40	42.46	29.25	21.33	17.14	61.55	14.52	24.21	30.61	12.48
	T-UMC (TCYB'22)	18.08	3.92	1.84	39.11	21.05	18.93	/	/	/	/	/	/
	UIMC (TNNLS'22)	16.84	4.36	1.74	45.00	32.05	29.29	/	/	/	/	/	/
	SURE (TPAMI'23)	42.99	29.07	21.78	57.00	45.02	38.58	45.09	72.33	31.12	22.81	32.16	12.85
	Our RMCNC	**48.89**	**33.83**	**28.91**	**64.38**	**48.29**	**44.06**	**70.49**	**80.17**	**56.57**	**28.28**	**38.43**	**19.50**
40%	PVC (NeurIPS'20)	12.60	0.09	0.00	0.10	0.00	0.00	/	/	/	/	/	/
	MvCLN (CVPR'21)	35.89	17.51	12.53	36.43	18.60	13.44	13.48	47.59	9.62	21.25	23.86	10.20
	T-UMC (TCYB'22)	17.98	4.00	1.89	26.72	7.75	6.54	/	/	/	/	/	/
	UIMC (TNNLS'22)	16.67	4.25	1.76	35.01	22.04	19.29	/	/	/	/	/	/
	SURE (TPAMI'23)	35.10	17.31	12.40	39.10	21.92	16.21	43.93	64.31	30.37	22.59	29.80	11.69
	Our RMCNC	**38.50**	**20.11**	**15.80**	**59.70**	**41.61**	**37.64**	**56.17**	**69.17**	**36.62**	**27.58**	**35.24**	**18.95**
60%	PVC (NeurIPS'20)	12.60	0.09	0.00	0.10	0.00	0.00	/	/	/	/	/	/
	MvCLN (CVPR'21)	26.80	8.48	5.45	28.36	12.42	8.03	9.76	39.12	7.67	14.32	11.72	4.68
	T-UMC (TCYB'22)	18.53	4.22	2.02	16.39	1.61	1.73	/	/	/	/	/	/
	UIMC (TNNLS'22)	16.63	4.23	1.75	15.19	12.19	9.45	/	/	/	/	/	/
	SURE (TPAMI'23)	27.11	8.61	5.58	28.83	13.49	8.64	27.05	52.86	16.34	19.35	24.21	9.15
	Our RMCNC	**29.29**	**10.77**	**7.26**	**52.41**	**31.87**	**27.84**	**37.65**	**55.39**	**16.72**	**25.67**	**30.85**	**17.84**

TABLE V: Multi-view clustering performance (%) with more than two views.

Methods	100Leaves (0%)			100Leaves (50%)			Caltech-3V (0%)			Caltech-3V (50%)		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
T-UMC (TCYB'22)	63.81	82.19	53.05	38.54	58.42	25.12	39.50	28.73	18.63	39.86	29.28	19.21
UIMC (TNNLS'22)	44.15	69.06	29.20	43.40	60.75	28.30	40.79	31.91	30.85	40.95	31.52	30.78
MFLVC (CVPR'22)	4.00	30.29	2.85	2.50	13.04	0.90	61.97	**57.13**	47.57	47.86	22.63	19.83
DCP (TPAMI'23)	58.57	84.59	52.60	42.35	64.12	25.43	56.34	50.82	40.47	53.66	45.95	37.19
DealMVC (MM'23)	10.37	51.06	8.37	5.63	19.75	1.22	59.00	55.24	43.90	49.21	26.29	23.41
Our RMCNC	**77.17**	**89.88**	**69.11**	**47.07**	**67.81**	**29.36**	**65.80**	56.75	**48.50**	**58.83**	**47.16**	**37.74**

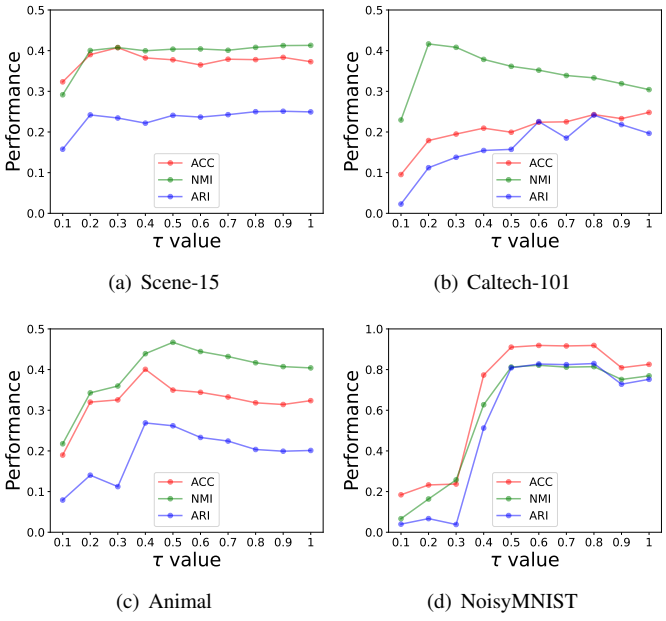
mance at different noise rates. Especially, the results under 80% noise show the stability and robustness of RMCNC. 3) The performance of most methods decreases sharply with the increase of noise rate, as misaligned pairs make the supervised information unreliable. 4) Fully-view-unaligned methods should realign the whole dataset during training, which makes them more effective for high-noise data by sacrificing efficiency. The experimental results also indicate they have better robustness for high-noise data.

Results of multi-modal clustering: We further evaluate the robustness of our RMCNC to handle with multi-modal data with the large semantic gap. Specifically, we record the multi-modal clustering results under different noise rates (*i.e.*, 0%, 20%, 40%, and 60%) on four multi-modal datasets in Table IV. From the table, one could see that: 1) Our RMCNC remarkably outperforms all comparison methods under different noise rates on all multi-modal datasets, which further shows the superior robustness of RMCNC against NC. 2) Due to the heterogeneity gap between multi-modal data, fully-view-unaligned methods have difficulty aligning the whole dataset during training, thereby leading to poor performance. 3) Due to instance-level alignment, PVP has a high computational complexity $\mathcal{O}(n^3)$, resulting in the inability to handle large-scale data, *e.g.*, XMedia and XRMB. Besides, instance-level alignment fails to deal with the heterogeneity gap, leading to disappointing results on the WIKI and NUS-WIDE datasets. 4) Although MvCLN and SURE are robust PVP-oriented methods, they still face overfitting or underfitting problems due to unknown noisy pairs, especially on the XMedia dataset.

Results with more than two views: To evaluate the clustering performance of our RMCNC on multi-view data with more than two views, we compare it with several MVC methods. To be specific, we report the clustering results under different noise rates (*i.e.*, 0% and 50%) on two three-view datasets in Table V. From the table, we could have the following observations: 1) RMCNC has a significant im-

TABLE VI: Ablation study on all datasets with different noise rates. The first and second best are marked in bold and underlined, respectively.

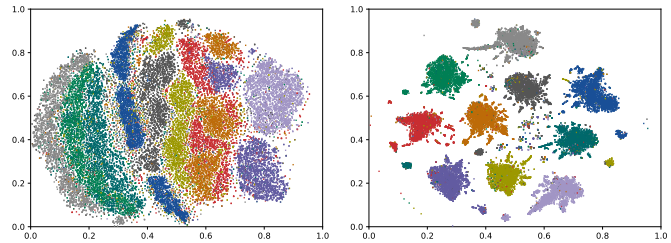
Noise	Methods	Scene-15			Caltech-101			Animal			NoisyMNIST		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
20%	InfoNCE	34.58	33.59	18.35	19.28	39.6	14.59	33.82	38.87	19.03	81.93	67.15	65.08
	MAE	36.62	37.08	22.38	21.41	43.53	17.36	33.42	40.86	18.83	38.49	30.68	14.99
	RMCNC-log	39.85	42.15	25.61	26.36	39.76	27.67	**44.58**	48.20	**30.57**	95.10	**88.28**	89.38
	RMCNC	**41.00**	**40.09**	**25.70**	**30.60**	**45.06**	**41.77**	39.87	**51.00**	30.01	**95.29**	88.00	**89.89**
40%	InfoNCE	34.14	31.29	17.13	14.25	25.80	8.60	30.45	30.15	14.01	57.78	51.93	37.92
	MAE	37.60	37.13	22.68	18.15	35.09	13.43	31.55	37.96	16.58	36.29	25.35	13.34
	RMCNC-log	38.47	40.92	24.30	23.32	31.17	21.24	**36.56**	36.28	21.44	90.88	**80.56**	80.57
	RMCNC	**39.55**	**40.93**	**24.86**	**27.57**	**46.56**	**36.88**	36.39	**46.58**	**26.05**	**91.00**	79.99	**83.02**
60%	InfoNCE	33.72	28.71	16.07	12.60	24.14	5.96	23.97	21.26	8.33	44.86	31.27	23.04
	MAE	34.53	33.78	19.38	14.25	25.80	8.60	28.11	33.32	13.58	25.72	14.78	6.78
	RMCNC-log	37.08	38.47	22.53	18.00	22.85	11.59	28.57	25.38	13.19	81.57	68.15	60.93
	RMCNC	**38.36**	**38.80**	**22.97**	**24.02**	**41.59**	**28.91**	**34.91**	**43.05**	**22.73**	**84.39**	**69.47**	**62.64**
80%	InfoNCE	30.80	28.12	**20.85**	8.52	18.41	2.74	15.23	11.9	3.06	32.93	21.68	13.70
	MAE	33.17	32.51	17.09	9.79	18.78	4.06	22.56	25.57	8.75	21.61	11.69	5.17
	RMCNC-log	**33.66**	32.86	17.91	15.08	17.61	7.98	17.68	14.64	4.69	70.70	58.11	50.25
	RMCNC	33.53	**33.41**	19.24	**15.67**	**32.35**	**13.59**	**30.47**	**35.42**	**16.95**	**74.31**	**60.87**	**56.36**



(a) Scene-15 (b) Caltech-101

(c) Animal (d) NoisyMNIST

Fig. 7: Parameter analysis of 40% noise in terms of τ.



(a) 0 Epoch (b) 40 Epoch

Fig. 8: t-SNE visualization on NoisyMNIST with 40% noise.

provement on clean three-view datasets compared with these baselines, which indicates that RMCNC could explore more comprehensive information from multiple views. 2) Under the case of 50% noise rate, RMCNC still has the best performance, which verifies the importance of leveraging robust multi-view learning with NC. 3) For 100Leaves, due to the low dimensionality (*i.e.*, 64-dimension) of the data, MFLVC and DealMVC cannot extract effective representation information, thereby resulting in extremely poor performance.

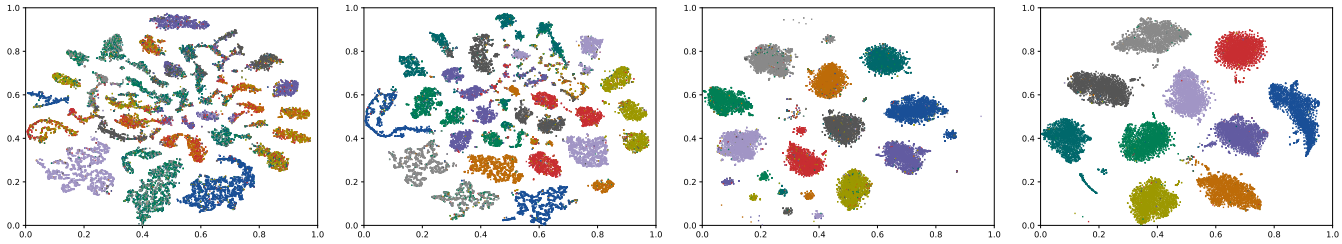
E. Ablation Study

To verify the effectiveness of our RMCNC comprehensively, we conduct some ablation studies. Specifically, we choose different loss functions (*i.e.*, InfoNCE and MAE) and perform the training experiments by using the same settings (such as network structure, optimizer, and hyper-parameters). Moreover, to further demonstrate the effectiveness of noise-tolerance contrastive learning, we replace the exponential loss with the negative-log-likelihood loss, named RMCNC-log. The loss of RMCNC-log can be defined as follows:

$$\mathcal{L}_{log} = \sum_{v=1}^{V}\sum_{u\neq v}^{V}\left(-log\left(\frac{G_{+}^{vu}}{G_{+}^{vu}+G_{-}^{vu}}\right)\right). \tag{11}$$

The ablation experiments are conducted on four datasets with different noise rates, whose evaluation results are shown in Table VI. From the results, both our RMCNC and RMCNC-log achieve better performance, which indicates maximizing total alignment probability can alleviate the adverse impact of individual false positives and improve robustness. However, for RMCNC-log, it may face the overfitting problem caused by NCs, thus degrading the clustering performance. Thanks to our exponential loss, our RMCNC achieves better performance than RMCNC-log, which further indicates that our method can overcome the overfitting caused by NC. Overall, our RMCNC outperforms three variant methods, which indicates

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3423307



(a) MvCLN (40% noise) (b) SURE (40% noise) (c) RMCNC (40% noise) (d) RMCNC (0% noise)

Fig. 9: t-SNE visualization of representations learned by MvCLN, SURE, and RMCNC on the NoisyMNIST.

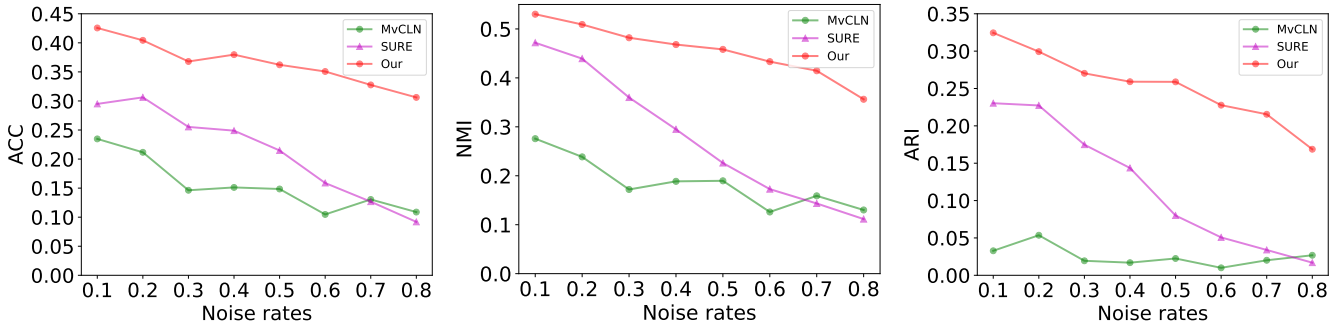


Fig. 10: Performance analysis on the Animal dataset under different noise rates.

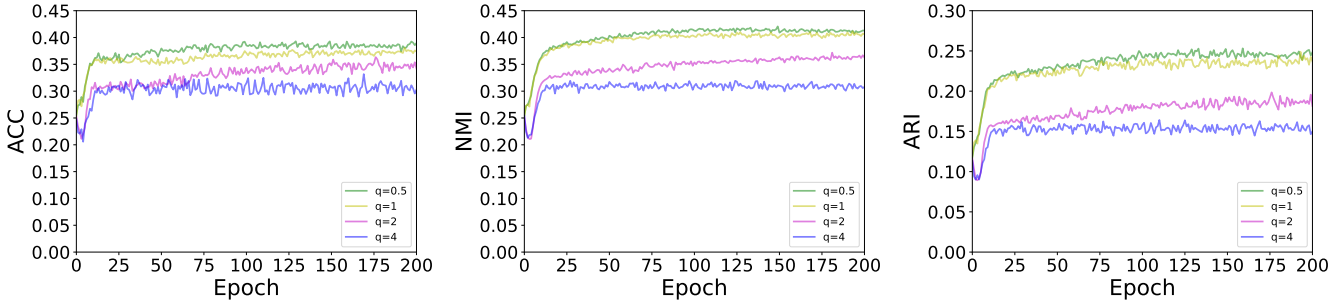


Fig. 11: The Performance against the number of epochs for training with different q values on the Scene-15 dataset.

our method can effectively mitigate the negative influence caused by NC, and endow MVC with robustness against NC. Besides, as demonstrated in Figure 3, we can observe that InfoNCE leads to the overfitting problem since it prioritizes the hard pairs during training, thus making it vulnerable to NC. As training progresses, it causes overfitting to incorrect supervision, leading to a drop in performance. MAE gives equal weight to all samples, making it more resilient to noisy data. Although MAE could alleviate the overfitting problem and achieve stable performance, it will lead to an underfitting problem due to its inertia against hard samples. Fortunately, our exponential loss finds a good balance between InfoNCE and MAE, thus enjoining robustness against NC like MAE while activeness to hard samples like InfoNCE. In general, our RMCNC can obtain better performance improvement due to our noise-tolerance multi-view contrastive learning.

F. Parameter Analysis

We analyze the influence of the parameters q and τ on the clustering performance by conducting experiments with 40% noise. First, we vary the parameter q from 10^{-5} to 4 by fixing $\tau = 0.5$. The experimental results are shown in Figure 6. From the figure, one could see that our method achieves the best performance when $q = 2$. Then, we vary the temperature parameter τ from 0.1 to 1 by fixing $q = 2$. As shown in Figure 7, we plot the performance curves of MVC with different q on four datasets with 40% noise. We can observe that our RMCNC achieves stable performance when parameter τ changes significantly. It indicates that RMCNC has a large range of τ (*i.e.*, from 0.5 to 1), which shows the ease of tuning of RMCNC.

G. Visualization Analysis

To illustrate the superiority of the representation learned by our RMCNC, we use t-SNE to visualize the representation on the NoisyMNIST dataset with 40% noise. First, we show t-SNE visualization with different epochs in Figure 8. From the figure, one could see that the data is mixed in the initialization phase, which makes it difficult to cluster different classes. After 40 epochs, the clustering structures of

learning representation become clearer, which indicates that our method can improve clustering performance. Moreover, we show t-SNE visualization with different methods under 40% noise in Figure 9. From the figure, we could find that MvCLN and SURE cannot resist noise very well, thus the data for each category is mixed together. Since our RMCNC has better robustness against NC, it has a clear clustering structure of data. In other words, RMCNC obtains large inter-cluster scatters and small intra-cluster scatters. When multi-view data is clean, we can see that our RMCNC obtains a clearer cluster structure on clean data than on noisy data. It indicates that NC destroys the discriminative information of the data, making it difficult for the model to distinguish the alignment relationship.

H. Influence of Different Noise Proportions

To make a more careful comparison, we evaluate the performance of our RMCNC on different rates of noise correspondences (*i.e.*, from 10% to 80%). The performance curves on the Animal dataset are plotted in Figure 10. Based on the figure, the conclusion can be drawn as follows: 1) As the ratio increases, the clustering performance decreases, which indicates that the correspondence noise could mislead the model and degrade its performance. 2) Our RMCNC can obtain stable results until the ratio reaches 80%, which indicates that our RMCNC can learn realigned representations by using fewer aligned samples and be immune to NC. 3) Our RMCNC obtains significant improvement in different noise proportions compared with MvCLN and SURE. Specifically, as the noise rate increases, the performance of our RMCNC drops smaller. It indicates the effectiveness and robustness of our method for resisting the negative effects of the correspondence noise.

I. Robustness Experiments with Different q

To better investigate the influence of q, we plot the performance curves with 40% noise in terms of different q (*i.e.*, from 1e-5 to 4). As shown in Figure 11, when training on Scene-15 with 40% noise, the clustering scores decrease as increasing q. Besides, compared with the comparison methods in Figure 3, our method is still robust for distinct q, which can overcome the overfitting problem.

V. Conclusion

In this paper, we address a new and challenging problem in MVC, namely noisy correspondence (NC). To tackle this problem, we propose a novel deep framework for robust MVC with NC to mitigate or even eliminate the negative effect of the misaligned pairs. To avoid the overfitting caused by NC, we maximize the total alignment probability and propose the noise-tolerance multi-view contrastive learning to endow our model with robustness against NC. To the best of our knowledge, for the first time, we propose a unified paradigm to deal with both NCP and PVP. Besides, we qualitatively show that our RMCNC can mitigate the negative influence of NCP. Extensive experiments demonstrate that RMCNC achieves state-of-the-art performance and robustness. In the future, we will study a unified framework to simultaneously solve Partially Sample-missing Problem (PSP) and NCP.

Acknowledgments

This work was supported by NSFC under Grants U21B2040, 62176171, 62102274, 62372315; by Sichuan Science and Technology Planning Project under Grants 2024NSFTD0038, 2024NSFTD0047, 24ZDZX0007, 2024NSFTD0049, 2023ZYD0143, 2024YFHZ0144, 2024YFHZ0089; by the Fundamental Research Funds for the Central Universities under Grants CJ202303 and CJ202403.

References

[1] X. Yang, C. Deng, Z. Dang, and D. Tao, "Deep multiview collaborative clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 516–526, 2023.

[2] C. Tang, K. Sun, C. Tang, X. Zheng, X. Liu, J.-J. Huang, and W. Zhang, "Multi-view subspace clustering via adaptive graph learning and late fusion alignment," *Neural Networks*, vol. 165, pp. 333–343, 2023.

[3] X. Li, Y. Sun, Q. Sun, Z. Ren, and Y. Sun, "Cross-view graph matching guided anchor alignment for incomplete multi-view clustering," *Information Fusion*, vol. 100, p. 101941, 2023.

[4] Y. Sun, Z. Ren, P. Hu, D. Peng, and X. Wang, "Hierarchical consensus hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, 2023.

[5] Y. Sun, J. Dai, Z. Ren, Q. Li, and D. Peng, "Relaxed energy preserving hashing for image retrieval," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[6] Y. Sun, D. Peng, and Z. Ren, "Discrete aggregation hashing for image set classification," *Expert Systems with Applications*, vol. 237, p. 121615, 2024.

[7] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "Comic: Multi-view clustering without parameter selection," in *ICML*, 2019, pp. 5092–5101.

[8] Z. Lin, Z. Kang, L. Zhang, and L. Tian, "Multi-view attributed graph clustering," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[9] Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1774–1782, 2018.

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

[11] E. Pan and Z. Kang, "Multi-view contrastive graph clustering," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2148–2159, 2021.

[12] Z. Kang, W. Zhou, Z. Zhao, J. Shao, M. Han, and Z. Xu, "Large-scale multi-view subspace clustering in linear time," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4412–4419.

[13] H. Tao, C. Hou, X. Liu, T. Liu, D. Yi, and J. Zhu, "Reliable multi-view clustering," in *Proceedings of the*

AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[14] C. Zhang, S. Wang, J. Liu, S. Zhou, P. Zhang, X. Liu, E. Zhu, and C. Zhang, "Multi-view clustering via deep matrix factorization and partition alignment," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4156–4164.

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

[16] Z. Kang, Z. Lin, X. Zhu, and W. Xu, "Structured graph learning for scalable subspace clustering: From single view to multiview," *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 8976–8986, 2022.

[17] X. Liu, L. Liu, Q. Liao, S. Wang, Y. Zhang, W. Tu, C. Tang, J. Liu, and E. Zhu, "One pass late fusion multi-view clustering," in *ICML*, 2021, pp. 6850–6859.

[18] R. Lin, S. Du, S. Wang, and W. Guo, "Multi-channel augmented graph embedding convolutional network for multi-view clustering," *IEEE Transactions on Network Science and Engineering*, pp. 1–12, 2023.

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

[20] H. Yu, J. Tang, G. Wang, and X. Gao, "A novel multi-view clustering method for unknown mapping relationships between cross-view samples," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2075–2083.

[21] J. Lin, X. Li, M. Chen, C. Wang, and H. Zhang, "Incomplete data meets uncoupled case: A challenging task of multiview clustering," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[22] J. Lin, M. Chen, C. Wang, and H. Zhang, "A tensor approach for uncoupled multiview clustering," *IEEE Transactions on Cybernetics*, pp. 1–14, 2022.

[23] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 406–29 419, 2021.

[24] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8547–8555.

[25] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1134–1143.

[26] M. Yang, Y. Li, P. Hu, J. Bai, J. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1055–1069, 2023.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[28] C. Zhang, Y. Liu, and H. Fu, "Ae2-nets: Autoencoder in autoencoder networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2577–2585.

[29] X. Li, Z. Ren, Q. Sun, and Z. Xu, "Auto-weighted tensor schatten p-norm for robust multi-view graph clustering," *Pattern Recognition*, vol. 134, p. 109083, 2023.

[30] G. Chao, Y. Jiang, and D. Chu, "Incomplete contrastive multi-view clustering with high-confidence guiding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 11 221–11 229.

[31] G. Chao, S. Sun, and J. Bi, "A survey on multiview clustering," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 146–168, 2021.

[32] M.-S. Chen, T. Liu, C.-D. Wang, D. Huang, and J.-H. Lai, "Adaptively-weighted integral space for fast multiview clustering," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3774–3782.

[33] Z. Yang, Q. Xu, W. Zhang, X. Cao, and Q. Huang, "Split multiplicative multi-view subspace clustering," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5147–5160, 2019.

[34] Y. Qin, D. Peng, X. Peng, X. Wang, and P. Hu, "Deep evidential learning with noisy correspondence for cross-modal retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4948–4956.

[35] P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, "Cross-modal retrieval with partially mismatched pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2023.

[36] S. Yang, Z. Xu, K. Wang, Y. You, H. Yao, T. Liu, and M. Xu, "Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency," *arXiv preprint arXiv:2303.12419*, 2023.

[37] X. Ma, M. Yang, Y. Li, P. Hu, J. Lv, and X. Peng, "Cross-modal retrieval with noisy correspondence via consistency refining and mining," *IEEE Transactions on Image Processing*, 2024.

[38] H. Han, Q. Zheng, M. Luo, K. Miao, F. Tian, and Y. Chen, "Noise-tolerant learning for audio-visual action recognition," *IEEE Transactions on Multimedia*, 2024.

[39] Y. Lin, J. Zhang, Z. Huang, J. Liu, Z. Wen, and X. Peng, "Multi-granularity correspondence learning from long-term noisy videos," in *International Conference on Learning Representations*, 2024.

[40] Y. Lin, M. Yang, J. Yu, P. Hu, C. Zhang, and X. Peng, "Graph matching with bi-level noisy correspondence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 362–23 371.

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 308–14 317.

[42] M. Yang, Z. Huang, and X. Peng, "Robust object re-

identification with coupled noisy labels," *International Journal of Computer Vision*, pp. 1–19, 2024.

[43] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," *arXiv preprint arXiv:2308.09911*, 2023.

[44] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[45] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International Conference on Machine Learning*, 2017, pp. 233–242.

[46] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, "Learning cross-modal retrieval with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5403–5413.

[47] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[48] L. Feng, T. Kaneko, B. Han, G. Niu, B. An, and M. Sugiyama, "Learning with multiple complementary labels," in *ICML*, 2020, pp. 3072–3081.

[49] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

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

[51] D. Dai and L. Van Gool, "Ensemble projection for semi-supervised image classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2072–2079.

[52] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

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

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

[55] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *ICML*, 2015, pp. 1083–1092.

[56] Y. Wang, D. Chang, Z. Fu, J. Wen, and Y. Zhao, "Partially view-aligned representation learning via cross-view graph contrastive network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[57] P. Hu, L. Zhen, D. Peng, and P. Liu, "Scalable deep multimodal learning for cross-modal retrieval," in *Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 635–644.

[58] Y. Sun, J. Dai, Z. Ren, Y. Chen, D. Peng, and P. Hu, "Dual self-paced cross-modal hashing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 184–15 192.

[59] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 394–10 403.

[60] X. He, Y. Peng, and L. Xie, "A new benchmark and approach for fine-grained cross-media retrieval," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1740–1748.

[61] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *2015 ICASSP*. IEEE, 2015, pp. 4590–4594.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, USA, 2015, pp. 1–13.

[63] A. Vinokourov, N. Cristianini, and J. Shawe-Taylor, "Inferring a semantic representation of text via cross-language correlation analysis," *Advances in Neural Information Processing Systems*, vol. 15, 2002.

[64] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, no. Jul, pp. 1–48, 2002.

[65] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[66] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Latent multi-view subspace clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4279–4287.

[67] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[68] F. Nie, J. Li, X. Li *et al.*, "Self-weighted multiview clustering with multiple graphs." in *International Joint Conferences on Artificial Intelligence*, 2017, pp. 2564–2570.

[69] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, "Dual contrastive prediction for incomplete multi-view representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4447–4461, 2023.

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