Appendix for Robust Multi-View Learning via Representation Fusion of Sample-Level Attention and Alignment of Simulated Perturbation

1. Related Work

In this section, we discuss the connections and differences between our method and related work including multi-view learning, contrastive learning, and attention mechanism.

1.1. Multi-view learning

Multi-view learning (MVL) refers to models learning comprehensive information from multiple views with matched correspondences. In this paper, we focus on deep learning based MVL methods and categorize existing methods into two types, *i.e.*, representation fusion and representation alignment. Representation fusion methods are the earliest popular in deep MVL, which aims to obtain a fused representation that is superior to representations of individual views [1, 29]. Many of these methods produce more accurate results on the fused representation than that on individual views' representations, and use it to further refine their representation learning [50, 54]. Representation alignment methods are first investigated by canonical correlation analysis based deep MVL approaches [2, 41, 53]. With the advancement of contrastive learning from self-supervised learning, an increasing number of deep MVL methods have adopted contrastive learning to capture the agreement across views [24, 26, 43, 44, 49]. To achieve the representation alignment, these contrastive MVL methods treat different views of a sample as positive pairs and maximize the similarity among their representations, thereby aiming to learn the semantic information across multiple views [7, 20, 42, 52]. Different from previous deep MVL methods, our RML performs the sample-level attention based multi-view representation fusion, and then achieves the simulated perturbation based representation alignment between the fused representations rather than between views.

1.2. Contrastive learning

Contrastive learning is a validated and effective paradigm for self-supervised representation learning [11, 37]. It usually constructs positive and negative sample pairs and encourages the model to learn discriminative representations, thereby aggregating the representations of positive sample pairs closer [8, 31]. The approaches for constructing positive sample pairs vary according to the types of data. For instance, in terms of image data, data augmentation techniques such as rotation and color filtering are typically employed to generate multiple images that are semantically consistent [10, 14]. For time-series data, adjacent samples in the sequence are used to construct positive sample pairs [32, 34]. Recently, contrastive learning has made significant progress in multi-view or multimodal domains, where different views or modalities of a sample are treated as positive sample pairs without the need for data augmentation [13, 35, 43]. Motivated by the success of data augmentation in contrastive learning [27, 45, 48], in this work, we propose a novel simulated perturbation based multi-view contrastive learning method for representation learning and downstream tasks, where the positive sample pairs are constructed by the two perturbed versions of fused representations.

1.3. Attention mechanism

Attention mechanism is an important technique initially introduced in the context of neural machine translation which enables models to selectively focus on relevant parts of the input data [3, 33]. It computes a weighted sum of input features, where the weights are dynamically determined based on the relevance of each feature to the task at hand, and this allows models to handle dependencies more effectively than traditional methods. Due to this property, attention mechanism has been integrated in many MVL applications [16, 30, 54]. Transformer [46] is one of the most popular networks in deep learning, which is built upon the attention mechanism and excels at modeling long-range dependencies between elements in sequences. Recent advances have also employed transformer-like networks to MVL [35, 39, 51], where the goal usually is to integrate and process information from multiple views such as text, image, audio, and video. However, the heterogeneous and imperfect natures of real-world multi-view data often hinder the transferability of existing successful experiences. To this end, this work proposes a robust MVL method which has a sample-level attention based multi-view fusion model using a transformer-like encoder network.

2. Implementation Details

2.1. Method details

For unsupervised multi-view clustering task, we directly utilize the model $\mathcal{F}_{\theta f}$ and minimize the loss function $\mathcal{L}_{\mathrm{RML}}$. Then, we employ the unsupervised clustering algorithm K-means [12] on the fused representations $\mathbf{Z}$ to obtain the clustering results.

For noise-label multi-view classification task, we extend our RML model $\mathcal{F}_{\theta f}$ by adding a classification head $\mathcal{H}_{\omega}$, obtaining class prediction probabilities $\mathbf{q}_i = \mathcal{H}_{\omega}(\mathcal{F}_{\theta f}(\{\mathbf{x}_i^m\}_{m=1}^V))$ through Softmax. Subsequently, we minimize the sum of $\mathcal{L}_{\mathrm{RML}}$ and cross-entropy loss on the training set. In this paper, we propose two variants for noise-label multi-view classification. The first one is formulated as follows:

$$\begin{aligned} &\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{RML}} \\ &s.t.\ \mathcal{L}_{\mathrm{CE}} = \mathcal{L}_{\mathrm{CrossEntropy}}(\mathbf{Y}, \{\mathbf{X}^m\}_{m=1}^V) \\ &\qquad\qquad = -\sum_i \mathbf{y}_i \log \mathbf{q}_i. \end{aligned} \tag{1}$$

This variant is entitled as $\mathrm{RML} + \mathcal{L}_{\mathrm{CE}}$. Furthermore, we incorporate the proposed simulated perturbations to establish multiple cross-entropy objectives, for further improving the model robustness to imperfect multi-view data. To be specific, the second variant is defined as $\mathrm{RML} + \mathcal{L}_{\mathrm{MCE}}$:

$$\begin{aligned} &\mathcal{L} = \mathcal{L}_{\mathrm{MCE}} + \lambda \mathcal{L}_{\mathrm{RML}} \\ &s.t.\ \mathcal{L}_{\mathrm{MCE}} = \mathcal{L}_{\mathrm{CrossEntropy}}(\mathbf{Y}, \{\mathbf{X}^m\}_{m=1}^V) \\ &\qquad\quad + \mathcal{L}_{\mathrm{CrossEntropy}}(\mathbf{Y}, \{\mathbf{N}^m\}_{m=1}^V) \\ &\qquad\quad + \mathcal{L}_{\mathrm{CrossEntropy}}(\mathbf{Y}, \{\mathbf{M}^m\}_{m=1}^V) \\ &\qquad = -\sum_i \left(\mathbf{y}_i \log \mathbf{q}_i + \mathbf{y}_i \log \mathbf{q}_i^N + \mathbf{y}_i \log \mathbf{q}_i^M\right), \end{aligned} \tag{2}$$

where we have $\mathbf{q}_i^N = \mathcal{H}_{\omega}(\mathcal{F}_{\theta f}(\{\mathbf{n}_i^m\}_{m=1}^V))$ and $\mathbf{q}_i^M = \mathcal{H}_{\omega}(\mathcal{F}_{\theta f}(\{\mathbf{m}_i^m\}_{m=1}^V))$, by which we make the classification model more robust to the noise perturbed data $\mathbf{n}_i^m$ as well as the unusable perturbed data $\mathbf{m}_i^m$.

For cross-modal hashing retrieval task, we apply our method RML in a plug-and-play manner to existing cross-modal hashing retrieval approaches. Specifically, we integrate our RML model $\mathcal{F}_{\theta f}$ on the top of the representation learning module of methods UCCH [17] and NRCH [47], and incorporate our optimization objective $\mathcal{L}_{\mathrm{RML}}$ as a regularization term into that of the cross-modal retrieval objective (*i.e.*, $\mathcal{L}_{\mathrm{UCCH}}$ and $\mathcal{L}_{\mathrm{NRCH}}$):

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\mathrm{UCCH}} + \lambda \mathcal{L}_{\mathrm{RML}}, \\ \mathcal{L} &= \mathcal{L}_{\mathrm{NRCH}} + \lambda \mathcal{L}_{\mathrm{RML}}. \end{aligned} \tag{3}$$

2.2. Experiment details

In this paper, we established the common model architecture of RML for the three different tasks, *i.e.*, multi-view clustering, multi-view classification, and cross-modal retrieval. This helps demonstrate the universality of our RML framework and promotes the comparable evaluation. Specifically, we leverage MLP networks and attention layer to implement the multi-view transformer fusion network $\mathcal{F}_{\theta f}$ in RML. Firstly, V parallel MLP networks are leveraged to transfer the input data $\{\mathbf{X}^m\}_{m=1}^V$ into word embeddings $\{\mathbf{E}^m\}_{m=1}^V$. For the m-th view, the MLP network can be illustrated as $\mathbf{X}^m \to Fc(D_m) - GELU - dropout(0.2) \to Fc(D_m) - dropout(0.2) \to \mathbf{E}^m$, where $Fc(D_m)$ denotes the fully-connected network with D_m neurons (D_m is the data dimensionality of the m-th view), GELU is the active function of Gaussian Error Linear Unit [15], and $\mathrm{dropout}(0.2)$ is the dropout operation [38] with the rate of 0.2. Upon $\{\mathbf{E}^m\}_{m=1}^V$, we adopt the typical transformer encoder network to obtain V encoded embeddings $\{\mathbf{F}^m\}_{m=1}^V$. Here, we use only one transformer encoder block [46] and the number of heads for multi-head attention is set to 1. Finally, we add multiple $\{\mathbf{F}^m\}_{m=1}^V$ and utilize a one-layer fully-connected MLP network to achieve the fused representation $\mathbf{Z}$. The dimensions of $\{\mathbf{E}^m, \mathbf{F}^m\}_{m=1}^V$ and $\mathbf{Z}$ are all set to 256 (*i.e.*, d_e and d are set to 256). We employ InfoNCE [31] contrastive loss to implement the optimization objective $\mathcal{L}_{\mathrm{RML}}$, where the temperature τ is set to 0.5. To train the model parameters, the optimizer we choose is Adam [21] with the learning rate of 0.0003. σ in the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ is set to 0.4.

When using K-Means clustering in our experiments, different views are concatenated to form a single one. For a fair comparison, the hyper-parameters of all comparison methods adopted the recommended settings given by the authors, and these comparison methods use the same input multi-view or multimodal data as that used in our RML.

In our cross-modal retrieval experiments, we follow the experimental settings and results in UCCH [17] to evaluate the performance of baselines and our RML. Specifically, we conduct two kinds of cross-modal retrieval task, *i.e.*, $\mathrm{Image} \rightarrow \mathrm{Text}$ and $\mathrm{Text} \rightarrow \mathrm{Image}$. Here, the ground-truth relevant samples refer to the cross-modal samples which have the same semantic category as the query sample. To evaluate the cross-modal retrieval results, the retrieval protocols adopt the same way in [17] that we measure the accuracy scores of the Hamming ranking results by Mean Average Precision (MAP), which returns the mean value of average precision scores for each query sample. In our experiments, we take MAP@ALL where all MAP scores are calculated on all retrieval results returned by tested methods. For RML+UCCH and RML+NRCH, to facilitate a fair comparison, we took the source code of UCCH and NRCH and inserted our RML module into them without introducing unnecessary changes. Since NRCH has different settings in data partitioning and pre-processing from UCCH, we treat NRCH and RML+NRCH as another set of comparison.

2.3. Dataset details

As we highly expect a MVL method which is compatible with various multi-view datasets, we conducted experiments on multiple types of multi-view or multimodal datasets to validate the effectiveness and universality of methods. We provide the detailed information of datasets as follow:

- **DHA** [23] is a repository documenting the intricacies of human motion, which captures RGB and depth image sequences as two views for each sample. Spanning across 23 unique categories, this multimodal dataset serves as a resource for the in-depth research of human motion.
- **BDGP** [6] comprises 2,500 samples of drosophila embryos which are categorized into 5 different classes. For each sample, two views of features have been extracted, including a 1,750-dimensional visual feature and a 79-dimensional textual feature.
- **Prokaryotic** [5] is a bioinformatics dataset that collects 551 prokaryotic species with three views. The dataset provides 4 species, described by textual features in the bag-of-words format, proteome compositions encoded by the frequency of amino acids, and gene repertoires using presence/absence indicators for gene families.
- **Cora** [4] consists of 2,708 scientific documents published over 7 topics, such as neural networks, reinforcement learning, and theory. Each document has a content-citation pair, that is 1,433-dimensional word content information and 2,708-dimensional citation information.
- **YoutubeVideo** [28] is a large-scale multi-view dataset with 101,499 samples from 31 classes, in which 512-dimensional cuboids histogram, 647-dimensional HOG, and 838-dimensional MISC vision features are leveraged to describe video data collected from the YouTube website.
- **WebKB** [40] is a dataset about web page information collected from the computer science departments of various universities. It comprises 1,051 samples belonging to course or non-course pages, and each sample has a fulltext view and an inlink view in web pages.
- **VOC** [9] consists of image-text pairs to form a two-modality dataset, with 5,649 instances across 20 categories. For each sample, the first modality is represented by 512-dimensional image GIST features, while the second modality is characterized by a word frequency count of 399-dimensional features.
- **NGs** [19] is a subset of the newsgroup dataset, consisting of 500 newsgroup documents and 5 categories. Each document has three views obtained through pre-processing methods, *i.e.*, supervised mutual information, partitioning around medoids, and unsupervised mutual information.
- **Cifar100** [22] is a popular image database with 50,000 samples from 100 subcategories. We follow [25] that extracts the image features through ResNet18, ResNet50, and DenseNet121 to construct three views, respectively.
- **MIRFLICKR-25K** [18] and **NUS-WIDE** [36] are two image-text datasets widely-used for cross-modal retrieval tasks (including image-to-text retrieval and text-to-image retrieval). We follow the setting in [17] to ensure a fair comparison as follows. For MIRFLICKR-25K, 18,015 image-text pairs are randomly selected as the retrieval set and the left 2,000 pairs are used as the query set, where each sample is with multiple labels from 24 semantic categories. The pretrained 19-layer VGGNet extracts the 4,096-dimensional image features and the bag-of-words (BoW) obtains 1,386-dimensional text features. For NUS-WIDE, 184,457 image-text pairs are randomly selected as the retrieval set and the remaining 2,100 pairs are the query set, belonging to 10 classes. Each pair is represented by the 4,096-dimensional VGGNet image features and 1,000-dimensional BoW text features.

3. More Experimental Results

In this appendix, we provide more experimental results to support our claims in this paper.

For noise-label multi-view classification task, Table 1 shows the results on different noise rates which further indicate the effectiveness of our RML to improve the robustness against noise labels. We provide the mean values of five independent runs

of comparison experiments as well as the corresponding standard deviation in the following Tables 2, 3, and 4. The results indicate that the improvement achieved by our method is significant.

Table 1. Performance comparison on noise-label multi-view classification

Method	DHA			BDGP			Prokaryotic			Cora			YoutubeVideo		
	ACC	Pre.	F1	ACC	Pre.	F1	ACC	Pre.	F1	ACC	Pre.	F1	ACC	Pre.	F1
noise label rate is 0%															
Trans.+$\mathcal{L}_{\mathrm{CE}}$	0.789	0.829	0.792	0.967	0.968	0.967	0.836	0.841	0.837	0.828	0.828	0.827	0.473	0.740	0.387
Trans.+$\mathcal{L}_{\mathrm{MCE}}$	0.788	0.819	0.788	0.903	0.905	0.903	0.842	0.850	0.844	0.778	0.780	0.778	0.648	0.711	0.602
RML+$\mathcal{L}_{\mathrm{CE}}$	0.712	0.815	0.670	0.959	0.959	0.959	0.854	0.860	0.855	0.772	0.775	0.767	0.759	0.761	0.758
RML+$\mathcal{L}_{\mathrm{MCE}}$	0.796	0.836	0.795	0.957	0.958	0.957	0.852	0.856	0.853	0.822	0.828	0.821	0.773	0.774	0.773
noise label rate is 10%															
Trans.+$\mathcal{L}_{\mathrm{CE}}$	0.724	0.770	0.723	0.845	0.847	0.845	0.766	0.778	0.770	0.753	0.754	0.753	0.471	0.725	0.387
Trans.+$\mathcal{L}_{\mathrm{MCE}}$	0.723	0.764	0.719	0.789	0.793	0.789	0.769	0.780	0.772	0.720	0.724	0.719	0.440	0.762	0.339
RML+$\mathcal{L}_{\mathrm{CE}}$	0.688	0.805	0.640	0.950	0.951	0.950	0.795	0.816	0.801	0.764	0.767	0.756	0.754	0.754	0.753
RML+$\mathcal{L}_{\mathrm{MCE}}$	0.727	0.798	0.710	0.867	0.868	0.867	0.776	0.796	0.782	0.792	0.797	0.788	0.766	0.767	0.765
noise label rate is 30%															
Trans.+$\mathcal{L}_{\mathrm{CE}}$	0.626	0.676	0.619	0.605	0.605	0.603	0.636	0.680	0.648	0.577	0.592	0.580	0.268	0.804	0.113
Trans.+$\mathcal{L}_{\mathrm{MCE}}$	0.618	0.656	0.609	0.600	0.605	0.599	0.617	0.687	0.636	0.548	0.564	0.551	0.475	0.706	0.406
RML+$\mathcal{L}_{\mathrm{CE}}$	0.622	0.773	0.568	0.938	0.938	0.938	0.769	0.807	0.778	0.665	0.673	0.658	0.590	0.640	0.580
RML+$\mathcal{L}_{\mathrm{MCE}}$	0.623	0.773	0.570	0.938	0.939	0.938	0.767	0.807	0.777	0.668	0.678	0.663	0.600	0.645	0.593
noise label rate is 50%															
Trans.+$\mathcal{L}_{\mathrm{CE}}$	0.457	0.487	0.448	0.437	0.441	0.435	0.473	0.594	0.505	0.400	0.432	0.407	0.266	0.804	0.112
Trans.+$\mathcal{L}_{\mathrm{MCE}}$	0.470	0.519	0.467	0.442	0.446	0.441	0.472	0.606	0.505	0.374	0.413	0.382	0.267	0.805	0.112
RML+$\mathcal{L}_{\mathrm{CE}}$	0.608	0.736	0.563	0.933	0.933	0.933	0.735	0.783	0.747	0.664	0.669	0.648	0.592	0.634	0.584
RML+$\mathcal{L}_{\mathrm{MCE}}$	0.610	0.737	0.565	0.936	0.936	0.936	0.735	0.783	0.747	0.665	0.666	0.651	0.598	0.639	0.593
noise label rate is 70%															
Trans.+$\mathcal{L}_{\mathrm{CE}}$	0.273	0.309	0.259	0.256	0.259	0.255	0.301	0.477	0.340	0.269	0.324	0.282	0.261	0.637	0.172
Trans.+$\mathcal{L}_{\mathrm{MCE}}$	0.254	0.275	0.242	0.249	0.252	0.249	0.296	0.470	0.336	0.259	0.305	0.271	0.259	0.512	0.205
RML+$\mathcal{L}_{\mathrm{CE}}$	0.421	0.649	0.330	0.886	0.890	0.885	0.402	0.547	0.437	0.600	0.630	0.591	0.586	0.623	0.580
RML+$\mathcal{L}_{\mathrm{MCE}}$	0.422	0.650	0.331	0.883	0.887	0.881	0.408	0.551	0.443	0.603	0.622	0.595	0.587	0.626	0.580

Table 2. Performance comparison of unsupervised multi-view clustering on multi-view datasets (mean $\pm$ std)

Method	DHA		BDGP		Prokaryotic		Cora		YoutubeVideo	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.656±0.029	0.798±0.001	0.443±0.029	0.573±0.041	0.562±0.022	0.325±0.006	0.363±0.041	0.172±0.043	0.199±0.002	0.194±0.001
MCN	0.758±0.021	0.800±0.017	0.957±0.026	0.901±0.041	0.528±0.025	0.287±0.014	0.386±0.017	0.184±0.032	0.183±0.002	0.187±0.001
CPSPAN	0.663±0.033	0.775±0.010	0.690±0.087	0.636±0.077	0.539±0.031	0.229±0.023	0.419±0.030	0.190±0.007	0.232±0.014	0.221±0.013
CVCL	0.662±0.063	0.754±0.033	0.907±0.078	0.785±0.009	0.526±0.049	0.281±0.032	0.483±0.007	0.310±0.003	0.273±0.005	0.258±0.002
DSIMVC	0.635±0.046	0.778±0.043	0.983±0.003	0.944±0.007	0.597±0.017	0.318±0.014	0.478±0.037	0.353±0.038	0.189±0.003	0.188±0.001
DSMVC	0.762±0.013	0.836±0.008	0.523±0.079	0.396±0.010	0.502±0.063	0.258±0.040	0.447±0.041	0.308±0.026	0.178±0.002	0.180±0.001
MFLVC	0.716±0.011	0.812±0.004	0.983±0.012	0.951±0.005	0.569±0.034	0.316±0.023	0.485±0.041	0.351±0.024	0.184±0.002	0.186±0.002
SCM	0.814±0.021	0.840±0.041	0.962±0.003	0.885±0.027	0.550±0.030	0.278±0.020	0.564±0.020	0.378±0.008	0.316±0.007	0.313±0.003
SCM$_{RE}$	0.804±0.001	0.840±0.001	0.971±0.004	0.913±0.002	0.582±0.037	0.312±0.028	0.574±0.008	0.374±0.009	0.317±0.001	0.322±0.004
RML+K-means	0.822±0.012	0.847±0.005	0.981±0.004	0.941±0.009	0.605±0.013	0.316±0.014	0.570±0.029	0.371±0.011	0.331±0.004	0.339±0.003

Table 3. Performance comparison of unsupervised multi-view clustering on multi-view datasets (mean $\pm$ std)

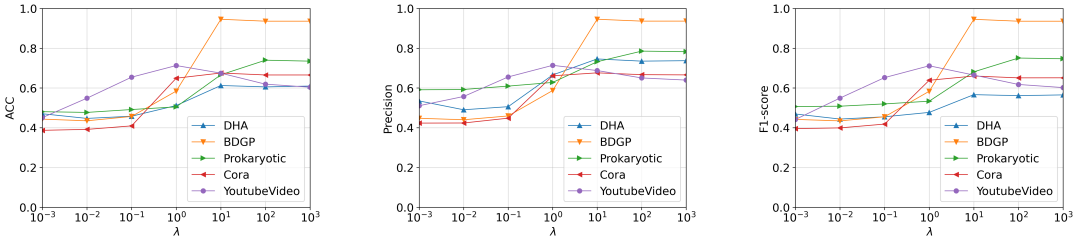
Method	WebKB		VOC		NGs		Cifar100	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.617±0.008	0.002±0.001	0.487±0.008	0.360±0.020	0.206±0.002	0.019±0.003	0.975±0.006	0.996±0.001
MCN	0.636±0.002	0.081±0.002	0.274±0.035	0.286±0.011	0.886±0.006	0.736±0.002	0.864±0.023	0.962±0.001
CPSPAN	0.771±0.021	0.166±0.042	0.452±0.022	0.488±0.017	0.352±0.002	0.215±0.015	0.918±0.014	0.982±0.002
CVCL	0.741±0.030	0.246±0.026	0.315±0.041	0.317±0.026	0.568±0.077	0.317±0.078	0.956±0.003	0.977±0.001
DSIMVC	0.702±0.014	0.250±0.013	0.212±0.017	0.204±0.011	0.630±0.062	0.502±0.059	0.895±0.011	0.969±0.005
DSMVC	0.663±0.018	0.134±0.012	0.633±0.034	0.723±0.041	0.352±0.027	0.082±0.013	0.851±0.023	0.959±0.007
MFLVC	0.672±0.021	0.245±0.014	0.292±0.004	0.280±0.001	0.908±0.000	0.802±0.000	0.877±0.018	0.964±0.009
SCM	0.689±0.017	0.094±0.021	0.607±0.046	0.622±0.043	0.968±0.004	0.900±0.012	0.999±0.001	0.999±0.000
SCM$_{RE}$	0.725±0.024	0.268±0.052	0.629±0.001	0.629±0.011	0.965±0.001	0.893±0.001	0.999±0.000	0.999±0.000
RML+K-means	0.868±0.079	0.508±0.156	0.656±0.031	0.615±0.011	0.983±0.007	0.943±0.022	0.999±0.000	0.999±0.000

Regarding hyper-parameter λ, we consider noise-label multi-view classification and cross-modal hashing retrieval tasks, where $\mathcal{L}_{\mathrm{RML}}$ is treated as a regularization term weighted by λ. The parameter analysis with the noise label rate of 50% is shown in Figure 1, where we observe stable classification performance within the range of $[10^1, 10^2, 10^3]$. For the noise-label

Table 4. Performance comparison on noise-label multi-view classification (mean ± std)

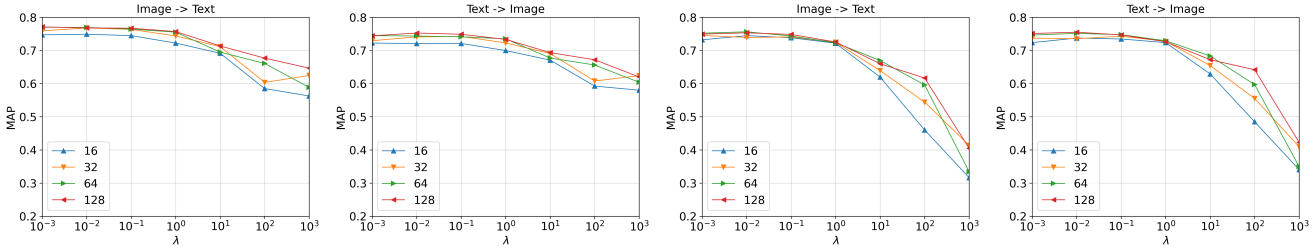
Method	DHA			BDGP			Prokaryotic			Cora			YoutubeVideo		
	ACC	Pre.	F1	ACC	Pre.	F1	ACC	Pre.	F1	ACC	Pre.	F1	ACC	Pre.	F1
noise label rate is 0%															
Trans.+$\mathcal{L}_{\text{CE}}$	0.789±0.023	0.829±0.025	0.792±0.025	0.967±0.013	0.968±0.013	0.967±0.014	0.836±0.013	0.841±0.008	0.837±0.011	0.828±0.006	0.828±0.005	0.827±0.006	0.473±0.170	0.740±0.054	0.387±0.225
Trans.+$\mathcal{L}_{\text{MCE}}$	0.788±0.034	0.819±0.037	0.788±0.035	0.903±0.010	0.905±0.010	0.903±0.010	0.842±0.019	0.850±0.014	0.844±0.017	0.778±0.009	0.780±0.009	0.778±0.009	0.648±0.020	0.711±0.007	0.602±0.028
RML+$\mathcal{L}_{\text{CE}}$	0.712±0.036	0.815±0.031	0.670±0.047	0.959±0.006	0.959±0.006	0.959±0.006	0.854±0.025	0.860±0.018	0.855±0.023	0.772±0.012	0.775±0.012	0.767±0.015	0.759±0.003	0.761±0.003	0.758±0.003
RML+$\mathcal{L}_{\text{MCE}}$	0.796±0.027	0.836±0.020	0.795±0.028	0.957±0.007	0.958±0.007	0.957±0.007	0.852±0.022	0.856±0.016	0.853±0.020	0.822±0.016	0.828±0.014	0.821±0.017	0.773±0.002	0.774±0.003	0.773±0.002
noise label rate is 10%															
Trans.+$\mathcal{L}_{\text{CE}}$	0.724±0.036	0.770±0.027	0.723±0.042	0.845±0.022	0.847±0.021	0.845±0.022	0.766±0.023	0.778±0.017	0.770±0.021	0.753±0.015	0.754±0.016	0.753±0.015	0.471±0.167	0.725±0.065	0.387±0.224
Trans.+$\mathcal{L}_{\text{MCE}}$	0.723±0.027	0.764±0.022	0.719±0.033	0.789±0.018	0.793±0.018	0.789±0.018	0.769±0.021	0.780±0.017	0.772±0.019	0.720±0.014	0.724±0.015	0.719±0.015	0.440±0.201	0.762±0.050	0.339±0.263
RML+$\mathcal{L}_{\text{CE}}$	0.688±0.031	0.805±0.036	0.640±0.045	0.950±0.013	0.951±0.013	0.950±0.013	0.795±0.021	0.816±0.010	0.801±0.017	0.764±0.006	0.767±0.008	0.756±0.013	0.754±0.006	0.754±0.006	0.753±0.006
RML+$\mathcal{L}_{\text{MCE}}$	0.727±0.027	0.798±0.037	0.710±0.043	0.867±0.023	0.868±0.024	0.867±0.024	0.776±0.013	0.796±0.004	0.782±0.010	0.792±0.021	0.797±0.023	0.788±0.027	0.766±0.003	0.767±0.003	0.765±0.003
noise label rate is 30%															
Trans.+$\mathcal{L}_{\text{CE}}$	0.626±0.073	0.676±0.075	0.619±0.074	0.605±0.021	0.605±0.016	0.603±0.018	0.636±0.045	0.680±0.034	0.648±0.041	0.577±0.017	0.592±0.013	0.580±0.016	0.268±0.001	0.804±0.001	0.113±0.001
Trans.+$\mathcal{L}_{\text{MCE}}$	0.618±0.044	0.656±0.043	0.609±0.044	0.600±0.039	0.605±0.038	0.599±0.039	0.617±0.047	0.687±0.047	0.636±0.045	0.548±0.015	0.564±0.016	0.551±0.015	0.475±0.171	0.706±0.081	0.406±0.241
RML+$\mathcal{L}_{\text{CE}}$	0.622±0.009	0.773±0.023	0.568±0.019	0.938±0.008	0.938±0.007	0.938±0.008	0.769±0.035	0.807±0.022	0.778±0.032	0.665±0.015	0.673±0.020	0.658±0.017	0.590±0.014	0.640±0.003	0.580±0.020
RML+$\mathcal{L}_{\text{MCE}}$	0.623±0.010	0.773±0.022	0.570±0.023	0.938±0.006	0.939±0.006	0.938±0.006	0.767±0.035	0.807±0.022	0.777±0.031	0.668±0.014	0.678±0.016	0.663±0.015	0.600±0.007	0.645±0.007	0.593±0.013
noise label rate is 50%															
Trans.+$\mathcal{L}_{\text{CE}}$	0.457±0.065	0.487±0.077	0.448±0.073	0.437±0.025	0.441±0.027	0.435±0.024	0.473±0.052	0.594±0.026	0.505±0.043	0.400±0.016	0.432±0.016	0.407±0.017	0.266±0.001	0.804±0.001	0.112±0.001
Trans.+$\mathcal{L}_{\text{MCE}}$	0.470±0.043	0.519±0.033	0.467±0.043	0.442±0.026	0.446±0.028	0.441±0.027	0.472±0.048	0.606±0.037	0.505±0.040	0.374±0.012	0.413±0.015	0.382±0.011	0.267±0.002	0.805±0.001	0.112±0.001
RML+$\mathcal{L}_{\text{CE}}$	0.608±0.027	0.736±0.022	0.563±0.035	0.933±0.014	0.933±0.013	0.933±0.014	0.735±0.019	0.783±0.020	0.747±0.018	0.664±0.011	0.669±0.013	0.648±0.012	0.592±0.005	0.634±0.005	0.584±0.010
RML+$\mathcal{L}_{\text{MCE}}$	0.610±0.029	0.737±0.025	0.565±0.038	0.936±0.009	0.936±0.008	0.936±0.009	0.735±0.019	0.783±0.020	0.747±0.018	0.665±0.014	0.666±0.019	0.651±0.013	0.598±0.004	0.639±0.007	0.593±0.007
noise label rate is 70%															
Trans.+$\mathcal{L}_{\text{CE}}$	0.273±0.049	0.309±0.059	0.259±0.050	0.256±0.021	0.259±0.022	0.255±0.021	0.301±0.035	0.477±0.038	0.340±0.032	0.269±0.010	0.324±0.016	0.282±0.010	0.261±0.006	0.637±0.206	0.172±0.074
Trans.+$\mathcal{L}_{\text{MCE}}$	0.254±0.060	0.275±0.072	0.242±0.061	0.249±0.016	0.252±0.019	0.249±0.018	0.296±0.018	0.470±0.016	0.336±0.013	0.259±0.008	0.305±0.014	0.271±0.008	0.259±0.007	0.512±0.239	0.205±0.076
RML+$\mathcal{L}_{\text{CE}}$	0.421±0.017	0.649±0.030	0.330±0.028	0.886±0.041	0.890±0.042	0.885±0.044	0.402±0.040	0.547±0.051	0.437±0.035	0.600±0.017	0.630±0.023	0.591±0.014	0.586±0.007	0.623±0.004	0.580±0.012
RML+$\mathcal{L}_{\text{MCE}}$	0.422±0.015	0.650±0.030	0.331±0.028	0.883±0.051	0.887±0.052	0.881±0.054	0.408±0.038	0.551±0.050	0.443±0.033	0.603±0.011	0.622±0.019	0.595±0.014	0.587±0.005	0.626±0.007	0.580±0.010

multi-view classification task, λ is set to 10^3 to emphasize $\mathcal{L}_{\text{RML}}$ in joint optimization when the noise label rates are large (*e.g.*, 30%, 50%, 70%). When the noise label rates are small (*e.g.*, 0%, 10%), λ is set to 10^0 for recommended settings. For the cross-modal hashing retrieval tasks, stable performance is observed within the range of $[10^{-3}, 10^{-2}, 10^{-1}]$ as shown in Figure 2. On cross-modal retrieval datasets MIRFLICKR-25K and NUS-WIDE, we kept λ unchanged in our comparison experiments (*i.e.*, $\lambda = 10^{-1}$).



(a) ACC (b) Pre. (c) F1

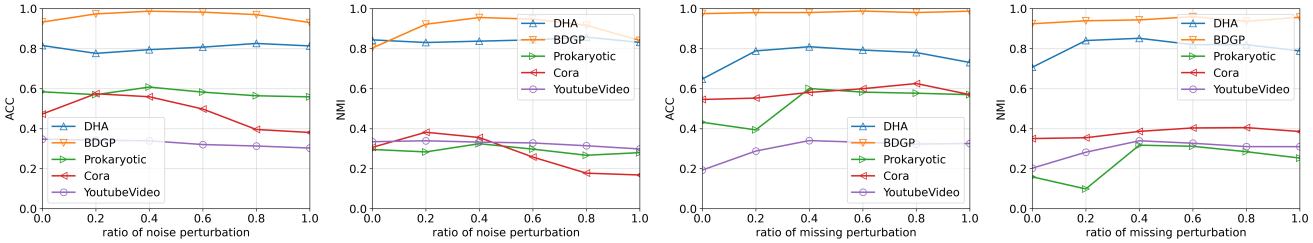
Figure 1. The hyper-parameter analysis of λ over three metrics on noise-label multi-view classification with the noise label rate of 50%.



(a) MIRFLICKR-25K (image-to-text) (b) MIRFLICKR-25K (text-to-image) (c) NUS-WIDE (image-to-text) (d) NUS-WIDE (text-to-image)

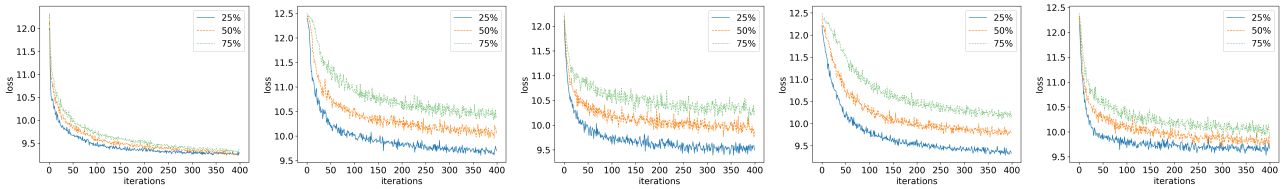
Figure 2. The hyper-parameter analysis of λ on cross-modal hashing retrieval tasks over hash code lengths of $[16, 32, 64, 128]$, including image-to-text retrieval (a,c) and text-to-image retrieval (b,d) on datasets MIRFLICKR-25K and NUS-WIDE.

Figure 3 and Figure 4 provide additional visualization results on more datasets that are unable to be shown in the main paper due to space limitations.



(a) p $vs.$ ACC (b) p $vs.$ NMI (c) r $vs.$ ACC (d) r $vs.$ NMI

Figure 3. Hyper-parameter analysis of the different ratios in our proposed simulated perturbation based multi-view contrastive learning on unsupervised multi-view clustering tasks, including noise perturbation (a-b) and unusable perturbation (c-d).



(a) DHA (b) BDGP (c) Prokaryotic (d) Cora (e) YoutubeVideo

Figure 4. The training loss values during our proposed simulated perturbation based multi-view contrastive learning, indicating that RML has well-converged optimization objective even with different perturbation ratios ($25\%, 50\%, 75\%$).

4. Potential Negative Societal Impacts

In this paper, we propose a robust multi-view learning method, which works in the field of fundamental machine learning and computer vision algorithms. It will not produce new negative societal impacts beyond what we already know.

References

[1] Mahdi Abavisani and Vishal M Patel. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1601–1614, 2018. 1

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013. 1

[3] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 1

[4] Gilles Bisson and Clément Grimal. Co-clustering of multi-view datasets: a parallelizable approach. In *ICDM*, pages 828–833, 2012. 3

[5] Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. The landscape of microbial phenotypic traits and associated genes. *Nucleic acids research*, page gkw964, 2016. 3

[6] Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, 2012. 3

[7] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *ICCV*, pages 16752–16761, 2023. 1

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 3

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 1

[11] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, pages 297–304, 2010. 1

[12] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. 2

[13] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *ICML*, pages 4116–4126, 2020. 1

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1

[15] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 2

[16] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *ICCV*, pages 4193–4202, 2017. 1

[17] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *TPAMI*, 45(3):3877–3889, 2023. 2, 3

[18] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ICMR*, pages 39–43, 2008. 3

[19] Syed Fawad Hussain, Gilles Bisson, and Clément Grimal. An improved co-similarity measure for document clustering. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 190–197, 2010. 3

[20] Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *CVPR*, pages 11600–11609, 2023. 1

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009. 3

[23] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In *ACM MM*, pages 1053–1056, 2012. 3

[24] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *TPAMI*, 45(8): 9552–9566, 2023. 1

[25] Suyuan Liu, Xinwang Liu, Siwei Wang, Xin Niu, and En Zhu. Fast incomplete multi-view clustering with view-independent anchors. *TNNLS*, 35(6):7740–7751, 2024. 3

[26] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *ICCV*, pages 754–763, 2021. 1

[27] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *IJCAI*, pages 4697–4705, 2024. 1

[28] Omid Madani, Manfred Georg, and David Ross. On using nearly-independent feature families for high precision and confidence. *Machine Learning*, 92(2-3):457–477, 2013. 3

[29] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, pages 14200–14213, 2021. 1

[30] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017. 1

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2

[32] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, pages 11205–11214, 2021. 1

[33] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *EMNLP*, pages 2249–2255, 2016. 1

[34] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, pages 6964–6974, 2021. 1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1

[36] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, pages 251–260, 2010. 3

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 1

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014. 2

[39] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 1

[40] Ting-Kai Sun, Song-Can Chen, Zhong Jin, and Jing-Yu Yang. Kernelized discriminative canonical correlation analysis. In *2007 International Conference on Wavelet Analysis and Pattern Recognition*, pages 1283–1287, 2007. 3

[41] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI*, pages 8992–8999, 2020. 1

[42] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *ICML*, pages 21090–21110, 2022. 1

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794, 2020. 1

[44] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *CVPR*, pages 1255–1265, 2021. 1

[45] David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001. 1

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 2

[47] Longan Wang, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Robust contrastive cross-modal hashing with noisy labels. In *ACM MM*, pages 5752–5760, 2024. 2

[48] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *TPAMI*, 45(5):5549–5560, 2022. 1

[49] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *CVPR*, pages 16051–16060, 2022. 1

[50] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, S Yu Philip, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *TKDE*, 35(07):7470–7482, 2023. 1

[51] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *TPAMI*, 45(10):12113–12132, 2023. 1

[52] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. GCFAgg: Global and cross-view feature aggregation for multi-view clustering. In *CVPR*, pages 19863–19872, 2023. 1

[53] Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. Category-based deep cca for fine-grained venue discovery from multimodal data. *TNNLS*, 30(4):1250–1258, 2018. 1

[54] Runwu Zhou and Yi-Dong Shen. End-to-end adversarial-attention network for multi-modal clustering. In *CVPR*, pages 14619–14628, 2020. 1