Disentangling Multi-view Representations Beyond Inductive Bias

Guanzhou Ke
guanzhouk@gmail.com
Beijing Jiaotong Univeristy
Beijing, China

Yang Yu*
yangy1@bjtu.edu.cn
Beijing Jiaotong Univeristy
Beijing, China

Guoqing Chao
guoqingchao@hit.edu.cn
Harbin Institute of Technology
Weihai, China

Xiaoli Wang
xiaoliwang@njust.edu.cn
Nanjing University of Science and Technology
Nanjing, China

Chenyang Xu
chyond.xu@gmail.com
Wuyi University
Jiangmen, China

Shengfeng He*
shengfenghe@smu.edu.sg
Singapore Management University
Singapore

ABSTRACT

Multi-view (or -modality) representation learning aims to understand the relationships between different view representations. Existing methods disentangle multi-view representations into consistent and view-specific representations by introducing strong inductive biases, which can limit their generalization ability. In this paper, we propose a novel multi-view representation disentangling method that aims to go beyond inductive biases, ensuring both interpretability and generalizability of the resulting representations. Our method is based on the observation that discovering multi-view consistency in advance can determine the disentangling information boundary, leading to a decoupled learning objective. We also found that the consistency can be easily extracted by maximizing the transformation invariance and clustering consistency between views. These observations drive us to propose a two-stage framework. In the first stage, we obtain multi-view consistency by training a consistent encoder to produce semantically-consistent representations across views as well as their corresponding pseudo-labels. In the second stage, we disentangle specificity from comprehensive representations by minimizing the upper bound of mutual information between consistent and comprehensive representations. Finally, we reconstruct the original data by concatenating pseudo-labels and view-specific representations. Our experiments on four multi-view datasets demonstrate that our proposed method outperforms 12 comparison methods in terms of clustering and classification performance. The visualization results also show that the extracted consistency and specificity are compact and interpretable. Our code can be found at https://github.com/Guanzhou-Ke/DMRIB.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**.

*Corresponding author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611794

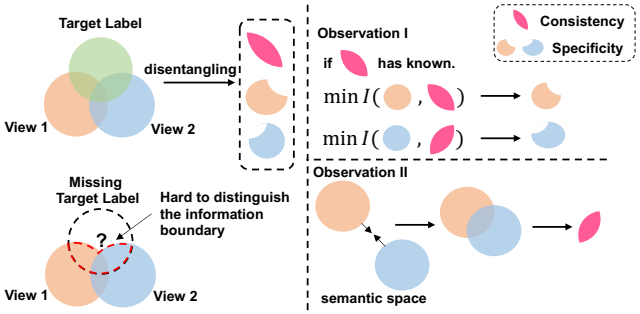

Figure 1: Illustration of the key observations and basic idea behind our method. On the left side of the graph, when target labels are available, we can disentangle consistent (pink) and specific representations (orange and blue) from comprehensive representations using the information bottleneck principle. However, when target labels (e.g., object categories) are missing, it becomes difficult to distinguish the boundaries of each part of the information. Based on this phenomenon, we found (on the right side) that if multi-view consistency is known in advance, we can obtain the remaining information by minimizing the mutual information between the comprehensive and consistent representations. Additionally, we observed that pulling the semantic distance between two views closer in a suitable semantic space helps us extract the multi-view consistency.

KEYWORDS

multi-view representation learning, disentangled representation, consistency and specificity

ACM Reference Format:
Guanzhou Ke, Yang Yu, Guoqing Chao, Xiaoli Wang, Chenyang Xu, and Shengfeng He. 2023. Disentangling Multi-view Representations Beyond Inductive Bias. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611794

1 INTRODUCTION

With the increasing availability of data from various sources, such as images, text, and sensors, it is essential to extract useful and rich information from them in multimedia applications. Multi-view

Representation Learning (MRL) [35], also known as multi-modal representation learning, is a promising approach that has gained attention in the communities. In cross-modal retrieval [26], "views" can be the pairs of images and corresponding textual descriptions referring to the same object. In autonomous driving [49], "views" can also refer to video frames of the same object captured by cameras at different positions. While MRL has proven effective in practical applications, such as clustering [7, 14], classification [36, 46], and face synthesis [40, 42, 43], understanding the underlying relationships between different view representations remains an open question. In general, multi-view comprehensive representations consist of consistent and specific representations [19], which combine in a certain pattern to form the multi-view comprehensive representation. Consistent representations refer to the shared information among different views, while view-specific representations refer to the private information of each view. Therefore, distinguishing between these two representations is an essential step in understanding multi-view representations.

One possible approach to understanding multi-view representations is to disentangle the comprehensive representation of each view and extract its consistency and specificity. Previous methods [8, 18, 33, 34] utilize the information bottleneck principle [29] to achieve this goal. The information bottleneck principle is a supervised method that aims to minimize the mutual information between the input data $\mathbf{x}$ and its embedding representation $\mathbf{z}$, while simultaneously maximizing the mutual information between $\mathbf{z}$ and target label $\mathbf{y}$. In the context of multi-view representation learning, this principle can be used to extract consistent representations by maximizing the mutual information between views, and view-specific representations can be similarly obtained by minimizing it. However, since most multi-view data is large-scale and unlabeled, these methods cannot be directly applied in unsupervised settings. In unsupervised settings, models cannot directly determine which parts of the representation correspond to consistency or specificity, which may lead to suboptimal solutions (see left part of Figure 1). For example, in the scenario of identifying clothes from different angles, we hope the model focuses on the inherent traits of the clothes (consistency), rather than who is wearing them (specificity). This task is easy to solve in the supervised setting but is challenging in the unsupervised setting where the model needs to distinguish between the boundaries of people and clothes. To address this issue, some works [9, 10, 41] introduce prior distribution assumptions for the process of disentangling consistent and view-specific representations. For example, [41] assumes that the consistency obeys the Gumbel distribution, and the specificity obeys the mixed Gaussian distribution. We argue that both the information bottleneck principle and prior distribution assumptions are strong "inductive biases". While the former requires data to have distinguishable information boundaries, the latter requires data to obey the prior distribution assumptions. However, these inductive biases may limit the scalability of downstream applications.

The main objective of this research is to investigate whether unsupervised multi-view representation disentangling can be achieved with weak inductive biases. This study is motivated by two key observations, which are illustrated in Figure 1. The first observation is that it is easier to separate specificity from comprehensive representations when consistent information is known beforehand. The second observation is that finding a suitable semantic space and narrowing the semantic distance between views in this space can help extract view consistency. Therefore, the challenge of extracting consistency is transformed into finding a suitable transformation space that satisfies two characteristics: low-level and high-level information. From the low-level perspective, the semantics associated with different view data of the same object should remain the same after data augmentation. For example, the semantics associated with the term "dog" remains the same even after color transformation is applied to views of dogs with different backgrounds. This is called transformation invariance. From the high-level perspective, all views of similar objects should have the same clustering prototype. We refer this to as clustering consistency. By narrowing the semantic distance of multi-view data in the space with transformation invariance and clustering consistency, multi-view consistency can be extracted. Based on these insights, a two-stage multi-view representation disentanglement method is proposed in this paper.

In the first stage, we employ a consistent encoder that maximizes both the transformation invariance and clustering consistency to output the consistent representation and corresponding clustering pseudo-labels. To maximize intra-view consistency, we first apply data augmentation to each view to generate new data that preserves the original semantics. Then, we map these augmented views to the same semantic space and cluster them to assign the same pseudo-label to views with the same semantics. We also introduce a maximum entropy constraint to prevent assigning all instances to the same cluster. In the second stage, we use multiple view-specific encoders to extract comprehensive representations for each view. We then minimize the upper bound of mutual information between view-consistent representations and comprehensive representations of each view to obtain specificity for each view. We adopt the VAE architecture and concatenate the pseudo-labels and view-specific representations as input to the view-specific decoders to generate data. Our approach has two benefits: First, the two-stage architecture only requires variational inference for view-specific representations; and second, utilizing pseudo-labels to control the view-generation process can improve interpretability. In other words, the pseudo-labels output from the first stage provides interpretability for the consistent representation and can control the generation of view data corresponding to the class. In addition, we extract specificity with a probabilistic approach, using a mixture of Gaussian distributions to fit the specificity distribution, which enables the sampling of data with different styles. Combining the two attributes, we can ultimately achieve data generation with specified output classes and output styles. Extensive experimental results on four multi-view datasets demonstrate our superior performances against state-of-the-art methods. The main contributions of this paper can be summarized as follows:

- We propose a two-stage unsupervised multi-view representation disentangling method that goes beyond inductive bias, requiring only the information of consistency to achieve disentanglement.
- We delve into multi-view consistency by mining view transformation invariance and clustering consistency. Ablation studies show that the quality of consistent representations

directly affects the expressive ability of the model and the quality of disentanglement.
- Our proposed method outperforms state-of-the-art methods in terms of clustering and classification performance. Visualization results also demonstrate that the extracted consistency and specificity are compact and interpretable.

2 RELATED WORK

2.1 Multi-view Representation Learning

Multi-view representation learning [35] can be broadly categorized into two groups: statistic-based and deep learning-based methods. Statistic-based methods focus on extracting view-consistent representations using techniques such as CCA-based methods [6, 26], non-negative matrix factorization methods [20, 38], and subspace methods [2, 37]. However, these methods have difficulty scaling up to high-dimensional and large-scale data scenarios [6, 26]. Therefore, a large number of deep learning-based methods have been developed in recent years [1, 23, 30, 35, 41, 43, 45, 48]. A comprehensive review [17] provides an overview of the current development status of MRL. Our method belongs to the deep learning-based approaches. In unsupervised scenarios, most previous methods use generative models such as autoencoders [1, 35, 45] or GANs [48] to extract multi-view comprehensive representations. However, there is a significant amount of redundant information in these representations that does not provide substantial help for downstream tasks. Therefore, many methods have begun to focus on the importance of disentangling multi-view representations, such as extracting view-consistent representations using contrastive learning [15, 28, 30] or separating consistent and view-specific representations based on the theory of information bottleneck [8, 41]. Unfortunately, in unsupervised environments, the model has difficulty distinguishing the boundaries between different types of information. Unlike previous methods, we propose a two-stage disentanglement strategy that first extracts view-consistent representations using self-supervised learning and then uses them as known information to extract view-specific representations by minimizing the upper bound of mutual information between consistent representations and comprehensive representations.

2.2 Disentangled Representation Learning

Disentangled Representation Learning (DRL) has emerged as a promising direction for learning independent factors in representations. This concept can be traced back to Independent Component Analysis (ICA)[44], which has since inspired various deep learning-based approaches, including InfoGAN[5] and β-VAE [13]. VAE-based methods, in particular, learn a variational distribution $q(z) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ to approximate the original data distribution $q(x)$, making them highly interpretable in statistics.

Recent efforts have focused on applying DRL to multi-view learning, where multiple views of data are available for training. For example, Xu et al.[41] proposed a multi-view DRL method that combines a generative model and a disentangled representation model to learn discriminative representations. Federici et al.[8] proposed a multi-view DRL method that applies a mutual information maximization objective to learn a joint latent space. Wang et al. [34] proposed a deep multi-view learning method that incorporates a shared disentangled representation into the learning process.

In comparison to these existing methods, our proposed method only aims to separate the specificity from the comprehensive representations. This simplification allows us to use a VAE-based architecture for the second stage of our method. Additionally, we only need to perform variational inference on the specificity, which reduces the computational complexity of our approach. Furthermore, we use pseudo-labels to assist the data reconstruction process, which helps to generate the required data under certain conditions.

3 METHODOLOGY

Give a multi-view dataset with V views $\mathcal{X} = \{X^{(1)}, X^{(2)}, \cdots, X^{(V)} | X^{(i)} \in \mathbb{R}^{N \times d_v}\}$, where d_v is the dimensionality of v-th view. The proposed method aims to extract comprehensive representations from $\mathcal{X}$ and subsequently disentangle them into view-consistent representations S and view-specific representations $P^{(v)}$. To this end, we leverage unsupervised pre-text tasks, such as contrastive learning [3, 4], to extract view-consistent representations. According to the conclusion of previous methods [8, 28], we found that maximizing the transformation invariance in contrastive learning methods is equivalent to maximizing the intra-view consistency. Furthermore, to improve the generalizability of consistent representations, we assume that a good view-consistent representation needs to satisfy two conditionals: i) transformation invariance and ii) clustering consistency. This part is depicted in section 3.1. After that, we extract comprehensive representations using the Conditional VAE (CVAE) [27] and then minimize the upper bound of mutual information between the comprehensive representation and the consistent representation for disentanglement. The benefit of this strategy is that it reduces the complexity of disentangling by reducing the number of unknown variables. This part will be discussed in section 3.2. We present the framework of the proposed method in Figure 2.

3.1 Mining Consistency

As previously mentioned, we assume that view-consistent information can be extracted through the transformation invariance and clustering consistency. Transformation invariance means that two views of an object, $X^{(1)}$ and $X^{(2)}$, should maintain the same semantic information, regardless of any transformations applied to them. For example, even after applying color jitter, the semantics of $X^{(1)}$ and $X^{(2)}$ should remain unchanged. On the other hand, clustering consistency implies that the clustering prototypes obtained from two views should remain the same after applying the same clustering mapping function. For instance, when processing different views of a dog's front and side, the clustering algorithm should assign both views to the same cluster. Next, we will examine how to extract them from multiple views data.

In contrastive learning [4, 11], the goal is to learn transformation invariance (or augmentation invariance) by reducing the semantic distance between two distinct augmentations of an image. Meanwhile, in [28], researchers have shown that the intra-view consistent information can be learned by using two different views of an object in multi-view scenarios. Thus, we design the strategy of

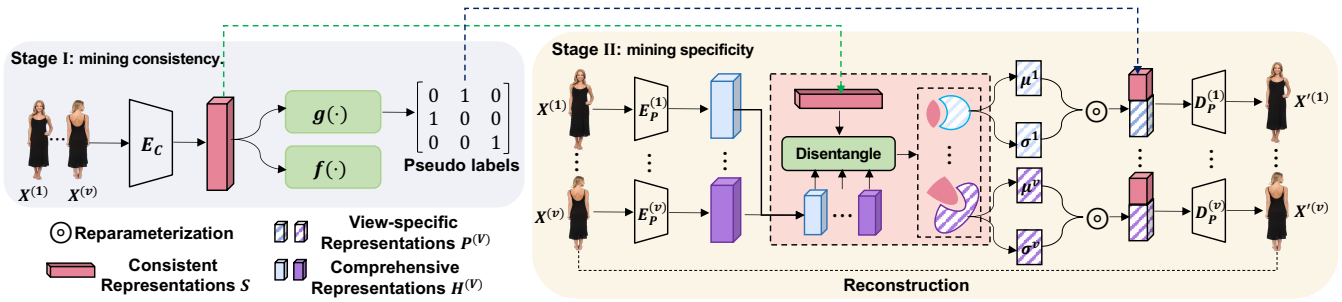

Figure 2: Illustration of the workflow of the proposed method. It includes a consistent encoder $E_c(\cdot)$, V view-specific encoders $E_p^{(i)}(\cdot)$ and decoders $D_p^{(i)}(\cdot)$, a clustering head $g(\cdot)$, a contrastive head $f(\cdot)$, and a disentanglement module. We adopt a two-stage approach to disentangle the consistency and the specificity. In the first stage, we train $E_c(\cdot)$ using contrastive and clustering heads, and then freeze its weights of the $E_c(\cdot)$. In the second stage, we use $E_p^{(i)}(\cdot)$ to learn the comprehensive representation of views, and then disentangle view-specific representations from the comprehensive representation by using the consistent representation S and clustering labels as known information. Finally, we complete the reconstruction task by concatenating the view-specific representation and clustering labels as the input to $D_p^{(i)}(\cdot)$.

maximizing multi-view transformation invariance. In essence, we apply an augmentation strategy, similar to [4], to each view. This approach yields a significant advantage in that we can obtain $2V\times$ positive and negative sample pairs. Then, we develop a multi-view contrastive learning loss, which is presented in the following:

$$\mathcal{L}_{ins} = \frac{1}{VB} \sum_{1\le i<j\le V} \sum_{k=1}^{2BV} -\log \frac{h(\boldsymbol{z}_k^{(i)}, \boldsymbol{z}_k^{(j)})}{\sum_{m=1}^{2BV} \mathbb{1}_{[m\neq k]} h(\boldsymbol{z}_k^{(i)}, \boldsymbol{z}_m^{(j)})} \tag{1}$$

where $h(\boldsymbol{a}, \boldsymbol{b}) = \exp(\text{sim}(\boldsymbol{a}, \boldsymbol{b})/\tau)$, $\text{sim}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$, B is the size of a minibatch, $\mathbb{1}_{[m\neq k]} \in \{0, 1\}$ implies an indicator function evaluating to 1 if $m \neq k$, and $\boldsymbol{z}^{(i)}$ is the contrastive vector of i-th view data x^i obtained from $\boldsymbol{z}^{(i)} = f(E_c(x^{(i)}))$, $f(\cdot)$ is the contrastive head consists of MLP and $E_c(\cdot)$ denotes the consistent encoder.

To further mine the view-consistent information, we argue that different views of an object must have consistent clustering prototypes. Inspired by [32], we encourage different views of an object and K nearest neighbors that are close in embedding space and can be classified into the same cluster. We employ a clustering head, denoted as $g(\cdot)$, to classify each sample in $\mathcal{X}$, which terminates in a softmax function to execute a soft assignment over clusters $C = \{1, \cdots, C\}$. The probability of assigning $X^{(i)}$ to cluster c is denoted as $g^c(X^{(i)})$. Therefore, we define the loss function of multi-view clustering consistency as the following:

$$\begin{aligned} \mathcal{L}_{clu} = &-\frac{1}{VB} \sum_{1\le i<j\le V} \sum_{k=1}^{B} \log\langle g(x_k^{(i)}), g(x_k^{(j)})\rangle \\ &+ \lambda_{clu} \sum_{c\in C} g'^c \log g'^c \\ &\text{with } g'^c = \frac{1}{B} \sum_{k=1}^{B} g^c(x_k^{(i)}) \end{aligned} \tag{2}$$

where λ_{clu} denotes the entropy weight, $\langle\cdot\rangle$ is the dot product operator. In Eq. (2), the first term encourages $g(\cdot)$ to make consistent cluster for a sample $x_k^{(i)}$ and its multi-view neighboring sample $x_k^{(j)}$. To avoid $g(\cdot)$ obtaining trivial solutions, the second term is used to spread the clustering results uniformly.

In practice, we have discovered that pre-training the network using Eq. (1), followed by mining the clustering consistency information using Eq. (2), produces superior results compared to training them jointly.

3.2 Mining Specificity

The goal of disentangling the multi-view representations is to separate the consistent representation S and view-specific representations $P^{(v)}$ from the multi-view comprehensive representation H. In the unsupervised setting, it is difficult for the model to distinguish which part is the consistency, and which part is the specificity. We have reconsidered the representation disentangling from the information theory perspective. Disentangling specificity from the comprehensive representation is equivalent to minimizing the upper bound of the mutual information between the view-consistent representation and comprehensive representations. Intuitively, we assume that the information of the comprehensive representation equals the sum of the consistent and view-specific information. Like solving ternary equations, solving view-specific representations will be easy when consistent and comprehensive representations are known. Therefore, we can determine the i-th view-specific representations $P^{(i)}$ by minimize the following objective:

$$\min\ I(S, H^{(i)}) + \epsilon \tag{3}$$

where ϵ is the noise contained within the view, and it is a constant. In the complete view setting, we consider ϵ negligible. Then, we have:

$$
\begin{aligned}
I(S, H^{(i)}) &= \mathbb{E}_{q(S,H^{(i)})} \log \frac{q_i(H^{(i)}|S)}{q(H^{(i)})} \\
&= \mathbb{E}_{q(S,H^{(i)})} \log \frac{q_i(H^{(i)}|S)}{r(H^{(i)})} \frac{r(H^{(i)})}{q(H^{(i)})} \\
&= \mathbb{E}_{q(S,H^{(i)})} \log \frac{q_i(H^{(i)}|S)}{r(H^{(i)})} - \mathbf{KL}[r(H^{(i)}) \| q(H^{(i)})] \\
&\leq \mathbb{E}_{q(S,H^{(i)})} \log \frac{q_i(H^{(i)}|S)}{r(H^{(i)})}
\end{aligned} \tag{4}
$$

where $q_i(\cdot)$ denotes the i-th view-specific marginal distribution, $q(a; b)$ implies the joint distribution of random variables a and b, and $r(\cdot)$ is mixed Gaussian distributions. In Eq. (4), we can clearly see that the upper bound depends on the approximation of the marginal distribution $r(H^{(i)})$ to prior $q(H^{(i)})$. Since $\mathbf{KL}[r(H^{(i)}) \| q(H^{(i)})] \geq 0$, we just need to optimize the first term. Therefore, we simplify the upper bound as the following:

$$
\min I(S, H^{(i)}) \equiv \min \mathcal{L}_{dis} = \mathbb{E}_{q(S,H^{(i)})} \log \frac{q_i(H^{(i)}|S)}{r(H^{(i)})} \tag{5}
$$

Theoretically, as long as the consistent information is accurate enough, Eq. (5) can approach the optimal solution indefinitely. This means that the quality of disentangling improves with the improvement of consistent information.

Next, we adopt the architecture of CVAE to extract comprehensive representations $H^{(i)}$. There are two advantages to this approach: i) it allows for fitting disentangled view-specific representations using a mixture of Gaussian distributions, and ii) by combining the category information output by the consistent encoder, the interpretability of the representations can be improved. We extend the CVAE to multi-view scenarios, and its loss function is as follows:

$$
\begin{aligned}
\mathcal{L}_{cave} = -\sum_{i=1}^{V} & \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|X^{(i)},S)} [\log \Phi(X^{(i)}|\mathbf{z}, S)] \\
& + \mathbf{KL}[q(\mathbf{z}|X^{(i)}, S) \| \Phi(\mathbf{z}))]
\end{aligned} \tag{6}
$$

where $\mathbf{z} = [P^{(i)}; S]$, $[\cdot]$ denotes the concatenating operation, and $P^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. Therefore, we can formulate the joint loss function of the second stage:

$$
\mathcal{L}_{spc} = \mathcal{L}_{cave} + \lambda_{dis} \mathcal{L}_{dis}. \tag{7}
$$

4 EXPERIMENTS

4.1 Dataset

We evaluate the proposed method and other competitive methods using four multi-view datasets. There are: (a) Edge-MNIST [22], which is a well-known benchmark dataset consisting of 70,000 grayscale digit images (0-9) with 28×28 pixels. The views contain the original digits and the edge-detected version, respectively; (b) Edge-FMNIST [39], which is a fashion dataset consisting of 28×28 grayscale images of clothing items. We synthesize the second view by running the same edge detector used to create Edge-MNIST; (c) COIL-20 [24], which depicts from different angles containing grayscale images of 20 items. We create a two-view dataset by randomly grouping the images for an item into two groups; (d) MVC-10 [21], which is a multi-angle clothing dataset consisting of 161,260 left-, right-, back-, and front-view with 10 categories. In our experiments, we use any three views to build the multi-view dataset. We report the dataset description in Table 1.

Table 1: Dataset Description

Dataset	#samples	#view	#class	#shape
Edge-MNIST	70,000	2	10	$(1 \times 28 \times 28)$
Edge-FMNIST	70,000	2	10	$(1 \times 28 \times 28)$
COIL-20	1,440	2	20	$(1 \times 128 \times 128)$
MVC-10	161,260	3	10	$(3 \times 224 \times 224)$

4.2 Baseline and Metrics

We compare the proposed method and the following 12 baseline clustering and classification methods, which are categorized into three types: **(a) Single-view methods:** K-means (KM) for clustering, and Support Vector Machine (SVM) for classification. Note that KM_{cat} denotes concatenating all view-specific representations. β-VAE [13] is a VAE-based method, which can obtain disentangled representation in the single-view scenario. In our settings, we select the best view as the β-VAE's input; **(b) Multi-view methods:** SCAN [32] and SimCLR [4] are self-supervised methods, and we use them to extract the consistent representation. SiMVC and CoMVC [30] are two contrastive learning-based multi-view clustering methods. EAMC [47] is an adversarial multi-view clustering method. CMC [28] is a contrastive multi-view representation learning method. MORI-RAN [15] is a contrastive fusion-based multi-view representation learning method; **(c) Multi-view disentangled methods:** Multi-VAE [41] is a VAE-based multi-view disentangled representation learning method. MIB [8] is an information bottleneck-based multi-view representation learning method. Note that MIB is limited to two views; for datasets with more than two views, we select the best two views as its input.

4.3 Implementation Details

We implement the proposed method and other non-linear comparison methods on the PyTorch 1.10 [25] platform, running on Ubuntu 18.04 LTS utilizing an NVIDIA A100 tensor core Graphics Processing Units (GPUs) with 40 GB memory size. For simplicity, we use ResNet [12] as the consistent encoder, where ResNet-18 is used for the Edge-MNIST and Edge-FMNIST datasets, and ResNet-34 is used for the COIL-20 and MVC-10 datasets. We set the dimensionality of view-specific encoders to $I - \text{Conv}^4_{32} - \text{Conv}^4_{64} - \text{Conv}^4_{128} - \text{Conv}^4_{256} - O$ for all experiments, where I and O indicate the dimension of the data's input and the encoder's output, respectively. It means that convolution kernel sizes are $4 - 4 - 4$, channels are $32 - 64 - 128 - 256$, the stride is set as 2, and the dimensionality of embedding is 256. The decoders are symmetric with the encoders. For all μ^v and σ^v, are set as 10-dimensional. We pre-train the consistent encoder using a similar way in the [4] and [32]. For the second

Table 2: Clustering results on four datasets, where "–" denotes the dataset cannot handle such scenarios. † indicates that we use Eq. (1) and (2) to enable it to handle multi-view data. The best and the second best values are highlighted in red and blue, respectively. All results are reproduced using our implemented code.

	Edge-MNIST			Edge-FMNIST			COIL-20			MVC-10		
Method	ACC_{clu}	NMI	ARI	ACC_{clu}	NMI	ARI	ACC_{clu}	NMI	ARI	ACC_{clu}	NMI	ARI
KM_{cat}	38.27±1.93	32.87±1.69	20.07±1.30	21.82±1.08	21.71±1.25	14.36±1.63	36.25±3.06	50.38±1.55	22.28±2.95	-	-	-
β-VAE (NIPS'18) [13]	57.88±0.42	52.77±0.02	48.17±0.01	40.87±0.11	39.48±0.32	39.42±0.13	18.52±0.53	58.73±0.60	34.12±0.60	34.55±1.07	30.45±0.86	11.03±1.37
SimCLR† (PRML'20) [4]	44.69±0.26	36.90±0.12	27.06±0.31	47.48±0.42	47.14±0.31	31.29±0.21	80.15±3.63	92.44±0.92	80.76±3.10	38.64±1.82	41.30±1.45	21.78±1.59
SCAN† (ECCV'20) [32]	95.38±1.26	91.44±1.28	90.19±1.73	70.34±1.95	65.82±1.86	57.19±1.44	89.11±2.30	94.57±3.42	88.26±2.37	39.08±1.53	45.14±1.32	31.65±1.18
SiMVC (CVPR'21) [30]	86.41±1.15	83.17±0.94	82.66±1.17	56.83±0.27	50.12±0.15	43.28±0.31	77.45±2.59	92.03±3.16	73.58±6.33	35.82±4.18	38.61±3.47	20.97±3.92
CoMVC (CVPR'21) [30]	95.42±1.54	90.85±2.38	89.73±2.05	58.94±2.69	53.27±2.08	49.72±3.22	90.04±1.20	95.19±1.21	91.13±1.38	39.15±2.19	43.96±1.67	25.79±1.52
EAMC (CVPR'20) [47]	66.78±0.52	63.11±0.34	59.83±1.22	55.21±1.60	62.57±1.48	48.51±1.88	67.28±3.59	75.83±4.21	68.73±6.30	41.38±1.08	42.53±0.66	29.46±1.35
CMC (ECCV'20) [28]	80.76±1.12	78.49±1.18	76.28±0.78	49.56±2.25	45.68±2.16	43.55±1.47	78.32±1.38	91.20±1.33	63.40±1.32	25.17±2.54	19.34±2.93	15.10±1.09
MORI-RAM (ICDM'22) [15]	82.13±4.11	78.25±1.68	75.14±3.52	58.33±4.48	55.19±2.70	40.86±2.23	73.58±0.71	84.72±1.36	71.95±1.19	24.48±0.29	22.39±0.49	12.29±1.05
Multi-VAE (CVPR'21) [41]	60.24±0.83	58.37±0.58	44.16±0.91	53.38±1.14	56.56±1.56	41.02±0.95	64.58±1.58	79.59±1.15	54.19±2.50	33.49±0.13	34.72±0.44	17.09±1.10
MIB (ICLR'20) [8]	53.65±5.33	48.16±7.41	36.07±3.66	54.41±4.82	53.08±6.50	44.69±6.77	51.67±5.79	83.12±3.36	56.76±4.89	-	-	-
Ours	97.71±1.24	95.82±2.07	95.08±1.28	73.06±1.47	70.39±0.28	60.44±1.44	90.64±2.24	97.36±1.07	90.68±2.33	48.91±0.56	47.67±0.73	33.65±0.66
Δ SOTA	↑2.29	↑4.37	↑4.89	↑2.72	↑4.57	↑3.25	↑1.53	↑2.17	↓0.45	↑7.53	↑2.53	↑2.00

Table 3: Classification results on four datasets, where "–" denotes the dataset cannot handle such scenarios. † indicates that we use Eq. (1) and (2) to enable it to handle multi-view data. The best and the second best values are highlighted in red and blue, respectively. All results are reproduced using our implemented code.

	Edge-MNIST		Edge-FMNIST		COIL-20		MVC-10	
Method	ACC_{cls}	F-Score	ACC_{cls}	F-Score	ACC_{cls}	F-Score	ACC_{cls}	F-Score
SVM_{cat}	42.89±0.07	41.11±0.5	53.51±0.21	53.58±0.08	10.42±0.01	8.02±0.01	-	-
β-VAE (NIPS'18) [13]	96.11±0.14	96.06±0.16	81.61±0.03	81.50±0.41	95.49±0.25	96.04±0.03	70.63±2.13	57.14±1.04
SimCLR† (PRML'20) [4]	97.97±0.02	97.95±0.03	80.09±0.03	80.06±0.02	97.19±0.11	97.00±0.11	71.61±0.20	67.02±0.16
SCAN† (ECCV'20) [32]	98.13±0.04	98.04±0.10	83.62±0.02	80.16±0.01	98.60±0.01	98.60±0.01	72.84±0.14	71.28±0.11
CMC (ECCV'20) [28]	97.53±0.03	97.50±0.02	77.11±0.14	75.78±0.28	97.31±0.01	97.31±0.01	70.19±0.89	69.79±1.42
MORI-RAM (ICDM'22) [15]	94.76±0.94	94.14±0.56	77.88±1.12	77.16±0.99	88.43±1.01	85.69±1.33	65.26±0.18	61.77±0.69
MIB (ICLR'20) [8]	90.81±1.30	90.03±0.69	75.33±0.05	73.80±0.05	59.72±2.29	53.99±2.03	-	-
Ours	99.41±0.06	99.41±0.04	85.19±0.07	84.88±0.07	99.81±0.01	99.80±0.01	79.07±0.09	79.04±0.22
Δ SOTA	↑1.28	↑1.37	↑1.57	↑3.38	↑1.21	↑1.20	↑6.20	↑7.76

stage of our method, we use Adam with default parameters and an initial learning rate of 0.0005 for training view-specific encoders and decoders for 150 epochs. For the comparing methods, we use their release codes with the settings recommended by the authors. For evaluation, we extract all latent representations, then feed them into K-means and SVM, and report their results, respectively. To eliminate the randomness, we run our method and other methods 10 times, and report their average and standard deviation values in terms of all evaluation metrics.

4.4 Evaluation Metrics

In order to evaluate clustering performance, three standard evaluation metrics are used: clustering ACCuracy (ACC_{clu}), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Readers seeking further details on these metrics are referred to [16]. It is important to note that the validation process of clustering methods is limited to cases where ground truth labels are available. For classification, ACC_{cls} and F-Score are utilized as evaluation metrics. In all cases, a higher value indicates better performance.

4.5 Comparison Results and Analysis

We evaluate our method with 12 state-of-the-art multi-view methods in terms of clustering and classification performance on four datasets, as shown in Table 2 and Table 3, where the following observations are obtained:

Our method outperforms other compared methods on all metrics of clustering and classification tasks. Especially, we achieve significant improvements over the second-best method on Edge-MNIST, Edge-FMNIST, and MVC-10 datasets. For instance, our method achieves 2.29% (97.71% − 95.42%) and 7.53% (48.91% − 41.38%) higher clustering accuracy than the second-best method on Edge-MNIST and MVC-10 datasets, respectively.

In comparison with single-view methods (KM$_{cat}$ and β-VAE), we find that simply concatenating all comprehensive representations of different views does not significantly improve the downstream performance. Furthermore, the multi-view disentangling method (Multi-VAE, MIB, and our method) significantly outperforms the single-view disentangling method. Therefore, we believe that disentangling the integrated representation is beneficial for improving the performance of downstream tasks. The essential reason is that disentangling can extract some task-independent information.

Compared with other contrastive learning-based methods, we find that using only consistent representations can achieve satisfactory results in clustering and classification tasks. At the same time, we also find that adding some specificity information appropriately can further improve the performance of downstream tasks. For example, both our method and Multi-VAE [41] concatenate the consistent representation and view-specific representations into one. We believe that this paradigm helps the model to process the information for the required part of the downstream tasks.

In the comparison results between our method and the end-to-end disentangling method (Multi-VAE [41] and MIB [8]), we find that as the complexity of data (quantity and dimension) increases, it becomes increasingly difficult to disentangle the consistency and specificity simultaneously from the comprehensive representation. In contrast, our method becomes more prominent in reducing the complexity of disentangling. These results confirm our observation that the boundary between multi-view consistency and specificity becomes unclear in the unsupervised setting. Therefore, obtaining one part of the information can improve the quality and reduce the difficulty of disentangling.

4.6 Ablation Study

We conducted a comprehensive ablation study on Edge-FMNIST, including the proposed method without pretext task $\mathcal{L}_{ins}$, the proposed method without the pseudo-label prediction $\mathcal{L}_{clu}$, and the proposed method without disentangling module $\mathcal{L}_{spc}$. We compared these components with our complete method, and the results are shown in Table 4. One intuitive result is that using the disentangling module can improve the performance of the model, with 2.72% (73.06% − 70.34%) increase in terms of ACC_{clu} metric. Additionally, we find that without $\mathcal{L}_{ins}$, the pseudo-label prediction cannot work independently, leading to poor results. At the same time, when only the disentangling module is used, its performance is relatively poor.

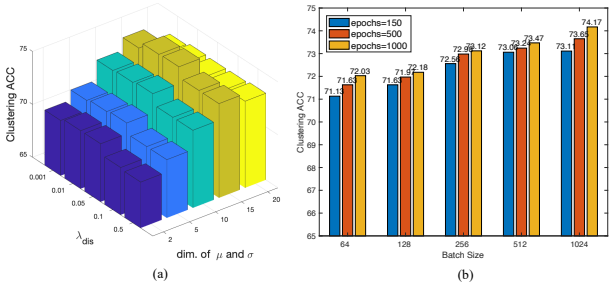

Figure 3: (a) Parameter sensitivity analysis of our method's performance to changes in the dimensionality of μ and σ and the hyperparameter λ_{dis}. (b) The effect of varying batch size and number of training epochs on clustering performance.

Table 4: Components analysis on the E-FMNIST dataset.

$\mathcal{L}_{ins}$	$\mathcal{L}_{clu}$	$\mathcal{L}_{spc}$	ACC$_{clu}$	NMI	ARI	ACC$_{cls}$	F-score
✓	✓	✓	73.06	70.39	60.44	85.19	84.88
✓		✓	49.34	50.11	42.40	80.46	78.39
	✓	✓	31.77	27.19	24.38	51.06	49.37
		✓	37.15	30.22	27.43	63.84	62.10
✓			47.48	47.14	31.29	80.09	80.06
	✓		12.81	9.57	2.05	10.05	10.05
✓	✓		70.34	65.82	57.19	83.62	80.16

This suggests that our method heavily relies on the quality of consistency, and the higher the quality of the consistent representation, the better the quality of disentangling.

4.7 Parameter Analysis

We conducted a hyperparameter analysis of the proposed method on the Edge-FMNIST dataset, including the dimensionality of μ and σ, λ_{dis}, batch size, and training epochs, as shown in Figure 3. According to the results in Figure 3(a), we find that the dimensionality of μ and σ range from 10 to 15, and λ_{dis} range from 0.01 to 0.05 can achieve better results. the dimensionality of μ and σ too low will lead to insufficient representation, while too high will produce redundant representation. λ_{dis} is the penalty coefficient of the disentanglement loss $\mathcal{L}_{dis}$, and if it is too low, disentangling may not be sufficient, while if it is too high, the model may obtain trivial solutions. On the other hand, as shown in Figure 3(b), with the increase of batch size and epochs, the performance of the proposed method will also increase. We believe that increasing training time will be more beneficial when the batch size is less than 512.

4.8 Visualization

We visualize all the representations of the proposed method on the Edge-MNIST dataset in the presented results, as shown in Figure 4 and Figure 5. In Figure 4, we use t-SNE [31] to visualize the consistent representation of Multi-VAE [41] and our method,

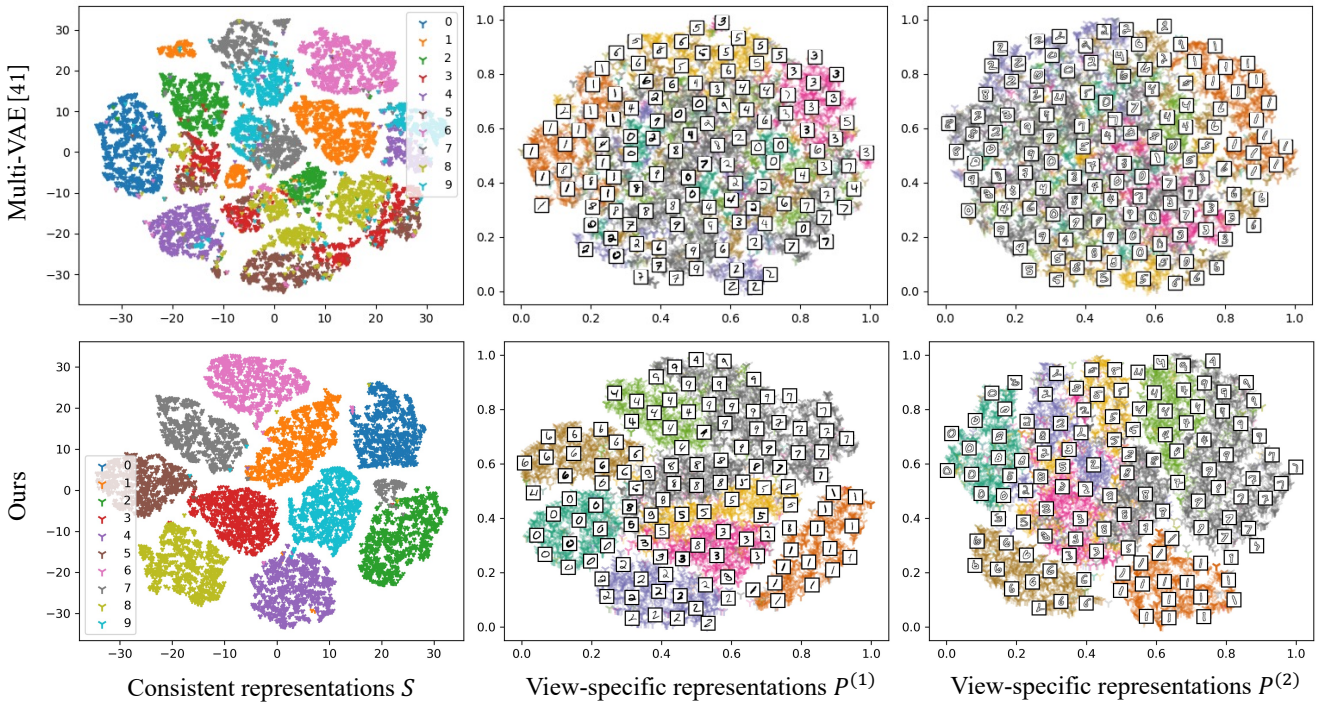

Figure 4: Visualization of representations using t-SNE [31] on the Edge-MNIST dataset. The top row displays the representations of Multi-VAE [41], while the bottom row shows the representations obtained by our proposed method. The leftmost column corresponds to the consistent representation S, while the subsequent two columns display the view-specific representations $P^{(1)}$ and $P^{(2)}$ of two different views, respectively.

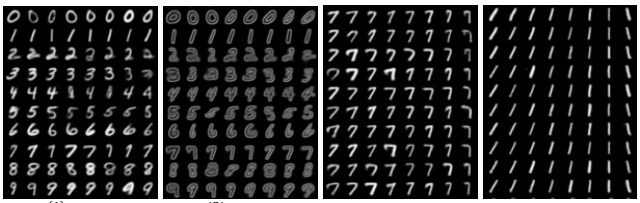

Figure 5: Visualization of disentangled representations: (a) Width variation of view-specific representation $P^{(1)}$ for different digits; (b) Angle variation of view-specific representation $P^{(2)}$ for different digits; (c) Width variation of specific digit "7" in both views; (d) Angle variation of specific digit "1" in both views.

and view-specific representations. We can see that the consistent representation extracted by our method is more compact than Multi-VAE [41]. Furthermore, in the view-specific space, we find that the specificity extracted by our method can also be divided into specific attribute regions. For example, in the view-specific representation $P^{(1)}$, we can see that there is a significant angle change for the digit "1" when the value of the x-axis changes from 0.6 to 1.0. In addition, we show the change of digit width and angle in Figure 5. Thanks to our disentanglement method, we can generate data with specific attributes and specific categories, such as digits "1" with different angle variations and digits "7" with different widths. The above results indicate that our disentanglement method can make multi-view representations have good interpretability and compactness.

5 CONCLUSION

In summary, we propose a novel two-stage disentanglement method that mines multi-view consistency by maximizing the transformation invariance and clustering consistency, and mines specificity by minimizing the mutual information between the consistent and comprehensive representations. Our method achieved superior clustering and classification performance on four datasets, and the ablation studies demonstrated the effectiveness of our disentangling module in enhancing the expressive power of concatenated representations. Moreover, the interpretability and compactness of both consistency and specificity obtained by our method were demonstrated through visualization results. In future work, we plan to extend our method to the incomplete-view scenario, as we have observed that our method can effectively generate specific views and help predict and restore missing views in such scenarios.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Beijing Jiaotong University (No. 2021JBWZB002 & No. 2023YJS113); Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097); and Singapore Ministry of Education Academic Research Fund Tier 1 (MSS23C002).

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*. PMLR, 1247–1255.

[2] Maria Brbić and Ivica Kopriva. 2018. Multi-view low-rank sparse subspace clustering. *Pattern Recognition* 73 (2018), 247–258.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* 33 (2020), 9912–9924.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems* 29 (2016).

[6] Paramveer Dhillon, Dean P Foster, and Lyle Ungar. 2011. Multi-view learning of word embeddings via cca. *Advances in neural information processing systems* 24 (2011).

[7] Si-Guo Fang, Dong Huang, Xiao-Sha Cai, Chang-Dong Wang, Chaobo He, and Yong Tang. 2023. Efficient multi-view clustering via unified and discrete bipartite graph learning. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[8] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *8th International Conference on Learning Representations, ICLR 2020*.

[9] Aviv Gabbay and Yedid Hoshen. 2021. Scaling-up disentanglement for image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6783–6792.

[10] Abel Gonzalez-Garcia, Joost Van De Weijer, and Yoshua Bengio. 2018. Image-to-image translation for cross-domain disentanglement. *Advances in neural information processing systems* 31 (2018).

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 770–778.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

[14] Dong Huang, Chang-Dong Wang, and Jian-Huang Lai. 2023. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[15] Guanzhou Ke, Yongqi Zhu, and Yang Yu. 2022. MORI-RAN: Multi-view Robust Representation Learning via Hybrid Contrastive Fusion. In *IEEE International Conference on Data Mining Workshops, ICDM 2022*, K. Selçuk Candan, Thang N. Dinh, My T. Thai, and Takashi Washio (Eds.). IEEE, 467–474.

[16] Abhishek Kumar, Piyush Rai, and Hal Daume. 2011. Co-regularized multi-view spectral clustering. *Advances in neural information processing systems* 24 (2011), 1413–1421.

[17] Yingming Li, Ming Yang, and Zhongfei Zhang. 2019. A Survey of Multi-View Representation Learning. *IEEE Trans. Knowl. Data Eng.* 31, 10 (2019), 1863–1883.

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

[19] Jing Liu, Yu Jiang, Zechao Li, Zhi-Hua Zhou, and Hanqing Lu. 2014. Partially shared latent factor learning with multiview data. *IEEE transactions on neural networks and learning systems* 26, 6 (2014), 1233–1246.

[20] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM international conference on data mining*. SIAM, 252–260.

[21] Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. 2016. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*. 313–316.

[22] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. *Advances in neural information processing systems* 29 (2016).

[23] Yuqin Lu, Jiangzhong Cao, Shengfeng He, Jiangtao Guo, Qiliang Zhou, and Qingyun Dai. 2023. Learning invariant and uniformly distributed feature space for multi-view generation. *Information Fusion* 93 (2023), 383–395.

[24] S Nene. 1996. Columbia object image library. *COIL-100. Technical Report* 6 (1996).

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019), 8026–8037.

[26] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. 251–260.

[27] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings*. Springer, 776–794.

[29] Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*. 368–377.

[30] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. 2021. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1255–1265.

[31] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[32] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. Scan: Learning to classify images without labels. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings*. Springer, 268–285.

[33] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. 2021. Multi-view information-bottleneck representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10085–10092.

[34] Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. 2019. Deep multi-view information bottleneck. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 37–45.

[35] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*. PMLR, 1083–1092.

[36] Xiaoli Wang, Liyong Fu, Yudong Zhang, Yongli Wang, and Zechao Li. 2022. MMatch: semi-supervised discriminative representation learning for multi-view classification. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 9 (2022), 6425–6436.

[37] Xiaobo Wang, Zhen Lei, Xiaojie Guo, Changqing Zhang, Hailin Shi, and Stan Z Li. 2019. Multi-view subspace clustering with intactness-aware similarity. *Pattern Recognition* 88 (2019), 50–63.

[38] Yang Wang, Lin Wu, Xuemin Lin, and Junbin Gao. 2018. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE transactions on neural networks and learning systems* 29, 10 (2018), 4833–4843.

[39] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[40] Cheng Xu, Keke Li, Xuandi Luo, Xuemiao Xu, Shengfeng He, and Kun Zhang. 2022. Fully Deformable Network for Multiview Face Image Synthesis. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[41] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. 2021. Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9234–9243.

[42] Xuemiao Xu, Keke Li, Cheng Xu, and Shengfeng He. 2020. GDFace: Gated Deformation for Multi-View Face Image Synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (Apr. 2020), 12532–12540.

[43] Yangyang Xu, Xuemiao Xu, Jianbo Jiao, Keke Li, Cheng Xu, and Shengfeng He. 2021. Multi-view face synthesis via progressive face flow. *IEEE Transactions on Image Processing* 30 (2021), 6024–6035.

[44] Howard Hua Yang and Shun-ichi Amari. 1997. Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information. *Neural computation* 9, 7 (1997), 1457–1482.

[45] Changqing Zhang, Yeqing Liu, and Huazhu Fu. 2019. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2577–2585.

[46] Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. 2022. Non-Aligned Multi-View Multi-Label Classification Via Learning View-Specific Labels. *IEEE Transactions on Multimedia* (2022), 1–13.

[47] Runwu Zhou and Yi-Dong Shen. 2020. End-to-End Adversarial-Attention Network for Multi-Modal Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 14607–14616.

[48] Runwu Zhou and Yi-Dong Shen. 2020. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14619–14628.

[49] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. 2020. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*. PMLR, 923–932.