Disentangling Multi-view Representations via Curriculum Learning with Learnable Prior

Kai Guo[1] , **Jiedong Wang**[1] , **Xi Peng**[1,2] , **Peng Hu**[1] , **Hao Wang**[1*]
[1]College of Computer Science, Sichuan University, China
[2]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, China
{kaiguo.gm, wangjd.cs, pengx.gm, penghu.ml, cshaowang}@gmail.com

Abstract

Multi-view representation learning methods typically follow a consistent-and-specific pipeline that aims at extracting latent representations for an entity from its multiple observable views to facilitate downstream tasks. However, most of them overlook the complex underlying correlation between different views. To solve this issue, we delve into a well-known property of neural networks (NNs) that NNs tend to learn simple patterns first and then hard ones. In our case, view-consistent representations are simple patterns and view-specific representations are hard. To this end, we propose to disentangle view-consistency and view-specificity and learn them gradually. Specifically, we devise a novel curriculum learning approach that adjusts the whole model to learn view-consistent representations first and then progressively view-specific representations. Besides, we saddle each view with a learnable prior that allows each view-specific representation to appropriate its distribution. Moreover, we incorporate a mixture-of-experts layer and a disentangling module to further enhance the quality of the learned representations. Extensive experiments on five real-world datasets show that the proposed model outperforms its counterparts markedly. The code is available at https://github.com/XLearning-SCU/2025-IJCAI-CL2P.

1 Introduction

Multi-view or multi-modal data such as image, text, and audio are appealing yet challenging for real-world applications. The data often contains consistent information across all views and complementary information for others compared to single-view data. Multi-view representation learning aims at learning latent representations from multi-view data and later uses them for downstream tasks, e.g., text generation [Ju *et al.*, 2021], anomaly detection [Wang *et al.*, 2023], and cross-modal retrieval [Ma *et al.*, 2024].

Over the years, various multi-view representation learning methods have been proposed. These methods generally fall into two categories: model-based multi-view methods [Xu *et al.*, 2022; Yan *et al.*, 2023; Zou *et al.*, 2024] and information-theory-based methods [Shi *et al.*, 2019; Federici *et al.*, 2020; Sutter *et al.*, 2024]. Among them, Federici et al. [2020] extended the information bottleneck principle to multi-view representation learning. Xu et al. [2022] proposed a multi-level (low-level features, high-level features, and semantic labels) representation learning framework for multi-view data. Yan et al. [2023] presented a cross-sample and cross-view aggregation method for representations of multi-view data. Sutter et al. [2024] designed a data-dependent multimodal variational prior for guiding the latent representation of each modal towards a shared aggregate posterior. However, despite the advances in these methods, there is a commonly overlooked issue to be resolved for multi-view representation learning, i.e., the intrinsic differences between view-consistency learning and view-specificity learning.

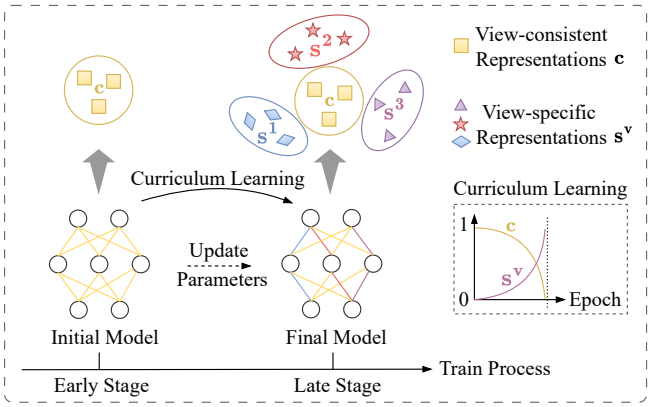


Figure 1: **Curriculum learning over view-consistency and view-specificity**. The model initially memorizes simple patterns, i.e., view-consistency c. As training progresses, the model gradually memorizes hard patterns, i.e., view-specificity s^v. Such a learning strategy is similar to humans. It facilitates the model in memorizing simple and hard patterns to enhance overall performance.

Multi-view data usually come from diverse sources or views. View-consistent representation mainly extracts common correlations across all views, and view-specific representation mines exclusive information within an independent view. Therefore, there are intrinsic differences between view-

*Corresponding author.

consistent and view-specific representation learning, i.e., the difficulty of view-consistency learning is lower than that of view-specificity. Due to the intrinsic differences, the same learning strategy may not work simultaneously for view-consistency and view-specificity learning. However, most existing methods ignore such intrinsic differences and adopt the same learning strategy for them, and finally result in a second-best performance. Moreover, neural networks have an important property that neural networks initially memorize *simple patterns* and then gradually memorize *hard patterns*. That is, neural networks initially focus on learning simple features of the data. As training progresses, the networks start to gradually model hard and complex features. Applying a simultaneously uniform learning strategy to both view-consistency (i.e., simple patterns) and view-specificity (i.e., hard patterns) would conflict with the inherent property of neural networks and then degrade the efficacy of data on downstream tasks.

To deal with the aforementioned issues, we propose to disentangle view-consistency and view-specificity, and devise an innovative **C**urriculum **L**earning model with **L**earnable **P**rior (denoted as CL2P). As shown in Figure 1, curriculum learning has scheduled parameters over epochs to adjust the model first to learn simple view-consistent representations and then to learn hard view-specific representations progressively, which we call *progressive curriculum learning*. Such a learning manner exactly fits this property of neural networks. Meanwhile, to enhance view-consistency, we saddle the model with a novel *Mixture-of-Experts* (MoE) layer. The MoE layer boosts the fusion of multi-view data. To enhance view-specificity, we further introduce a *learnable prior*. The learnable prior allows view-specific representation from each view to fit its optimal distribution. Moreover, to avoid interference between view-consistency and view-specificity, we delve into mutual information and propose to minimize the upper bound of it to disentangle view-consistent and view-specific representations.

In summary, we make the following contributions:

- We explore a well-known property of neural networks that neural networks fit simple patterns first and then hard ones. To our knowledge, this is the first work to investigate this property of neural networks in the context of multi-view representation learning.
- We propose a novel curriculum learning model called CL2P. The CL2P exhibits an innovative amalgamation of curriculum learning, mixture-of-experts learning, learnable prior, representation disentangling, and a joint loss function in wrapping up all components.
- We evaluate our CL2P using five real-world multi-view datasets. Extensive experimental results demonstrate the effectiveness of the proposed method and its superior performance in comparison to baselines.

2 Related Work

Multi-view Autoencoders. Multi-view autoencoders (MVAE) are employed to model datasets that originate from multiple sources or views. MVAEs are particularly useful for representation learning, understanding the relationships between different views, and generating missing data. Existing MVAE-based methods can be classified into two categories: model-based MVAE [Xu *et al.*, 2022; Li *et al.*, 2023; Hu *et al.*, 2023; Lu *et al.*, 2024], and information-theory-based MVAE [Shi *et al.*, 2019; Federici *et al.*, 2020; He *et al.*, 2024; Sutter *et al.*, 2024].

Our method falls into the category of information-theory-based MVAE. We leverage the principles of information theory to address the issue of optimal distributions for view-specific representations from different views, enhancing representation performance and decreasing the learning complexity of view-specificity. In addition, we propose a progressive curriculum learning strategy to tackle the heterogeneity between view-consistent and view-specific representations.

Curriculum Learning. Curriculum learning (CL) is a training strategy that trains a machine learning model from easier data to harder data, which imitates the meaningful learning order in human curricula [Bengio *et al.*, 2009; Soviany *et al.*, 2022]. To date, most CL methods are designed by following the pipeline of a difficulty measurer and a training scheduler. The difficulty measurer is used to measure the hard level of each data to decide learning priority. The training scheduler determines the timing for feeding hard data into the training process. Regarding whether these two components are designed in a data-driven automated manner, CL can be broadly categorized into two types: predefined CL and automatic CL [Wang *et al.*, 2021].

Our method proposes a progressive curriculum learning strategy that extends the concept and applications of predefined CL. In our method, the intrinsic differences (i.e., learning priority) between view-consistency and view-specificity are determined based on human prior knowledge.The training scheduler is an adaptive trade-off parameter over epochs.

Prior Learning. Prior learning (PL) is a critical part of variational inference. PL has emerged as an approach to refining variational inference models [Xu *et al.*, 2019]. It enhances the quality of learned representations by improving the evidence lower bound (ELBO). For example, Bauer et al. [2019] designed a learned acceptance function to reweight the proposal. Similarly, Takahashi et al. [2019] presented an implicit optimal prior for variational autoencoders. This implicit prior uses the density ratio trick. Building upon prior learning ideas, we novelly use a learnable prior based on pseudo-inputs. Our method is the first to extend the concept of prior learning to multi-view settings. The prior learning enables each view-specificity to discover its optimal distribution.

3 Methodology

Definition 1 (Problem Statement)**.** *Given a set of multi-view data with n samples and m views $\mathcal{D} = \{\mathbf{x}_i|\mathbf{x}_i^1, ..., \mathbf{x}_i^m\}_{i=1}^n$, the dataset is used to train a model (e.g., our CL2P). The trained model is then employed to derive high-quality view-consistent representation and view-specific representation for downstream tasks (e.g., clustering and classification).*

Overall Architecture. Figure 2 illustrates the architecture of the proposed CL2P, comprising four novel components: *Curriculum Learning*, *Mixture-of-Experts* (MoE), *Prior Learning*, and *Disentangling* module. Specifically, curriculum learning guides the whole model to first learn the

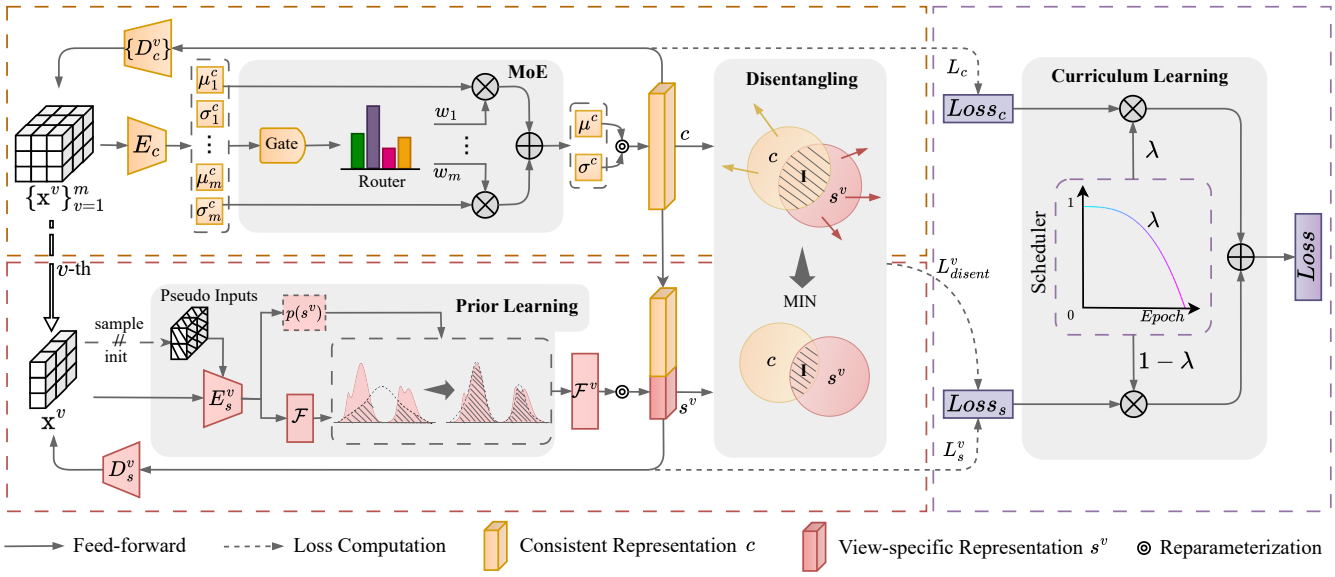


Figure 2: An overview of the proposed CL2P. It mainly consists of four parts: 1) *Curriculum learning*, which adjusts the whole model to first learn the simple view-consistent representations c and then hard view-specific representations $\{s^v\}_{v=1}^m$ progressively, 2) *Mixture-of-Experts (MoE)*, which integrates all views into a consistent representation c for multi-view data, 3) *Pseudo-inputs*, which together with view-specific representation learning drive optimal prior for different views, and 4) *Disentangling module*, which reduces the redundancy between view-consistent representation c and view-specific representation $\{s^v\}_{v=1}^m$. [Best viewed in color]

simple consistent representations and then pay attention to the hard view-specific ones progressively. In practice, the losses of view-consistency and view-specificity are adjusted by curriculum learning using an adaptive trade-off parameter λ, where λ is automatically generated by the "Scheduler" according to the training progress. In view-consistent representation learning, we use a shared encoder E_c to extract view-consistent representations from all views. Then, we use a MoE layer to fuse these representations by adaptively weighting each expert's output. Finally, the fused consistent representation c serves as input for decoders $\{D_c^v\}_{v=1}^m$, generating reconstructed views. In view-specific representation learning, we leverage a set of view-specific encoders $\{E_s^v\}_{v=1}^m$ to extract view-specific representations $\{s^v\}_{v=1}^m$. Meanwhile, we allocate the pseudo-inputs for view data $\mathbf{x}^v$. The pseudo-inputs serve as auxiliary elements for learning an optimal prior for the view-specific representation s^v. For the v-th view-specificity, s^v and c are concatenated as the input to the v-th view-specific decoder D_s^v. Moreover, to prevent the mutual dilution between view-consistency and view-specificity, we use a disentangling module to minimize the redundancy between these two aspects. For efficiency, we adopt variational autoencoders [Kingma, 2013] to construct the model. Next, we elaborate on our solutions for each component.

Property 1. *DNNs tend to prioritize memorization of simple instances first and then gradually memorize hard instances [Zhang* et al.*, 2017; Arpit* et al.*, 2017; Kumar* et al.*, 2024].*

3.1 Progressive Curriculum Learning

According to *Property* 1, DNNs tend to initially focus on learning simple patterns from the data, and then to fit more complex features within the data. We revisit this property in the context of multi-view representation learning. As aforementioned, there are intrinsic differences in the learning of view-consistency and view-specificity. In this work, we propose a progressive curriculum learning method to shift the learning attention of the networks from view-consistency to view-specificity. That is, the progressive curriculum learning method aims at first learning view-consistent representations and then gradually learning view-specific representations. To this end, we design an adaptive trade-off parameter λ to schedule the network parameters and losses produced by these two parts. During the training stage, the $Loss_c$ on learning view-consistent representations is weighted by λ, while the $Loss_s$ on learning view-specific representations is weighted by $1-\lambda$. Thus, the overall loss of model is $\lambda Loss_c + (1-\lambda) Loss_s$. The scheduler of λ is formulated as follows:

$$\lambda = 1 - \left(\frac{\mathcal{T}}{\mathcal{T}_{max}}\right)^2, \tag{1}$$

where $\mathcal{T}_{max}$ is the number of total training epochs and $\mathcal{T}$ is the current epoch. λ gradually decreases as the training epochs increase. The changes of λ ultimately affect the attention on network parameter updates.

The proposed curriculum learning shifts the learning "attention" between simple view-consistent representations and hard view-specific representations. This approach, which is closely aligned with the *Property* 1 of DNNs, is beneficial for fully extracting patterns from multi-view data. Furthermore, λ controls the parameter updating for each part, thereby preventing damage to the consistent representations while emphasizing view-specific representations at the later stages of training. Next, we delve into the details of view-consistent and view-specific representation learning.

3.2 Mixture of Consistent Posteriors

For view-consistency learning, we assign each view an "expert" to model its posterior distribution, and then combine these individual posteriors as follows

$$q_{\phi_c}(c|\{\mathbf{x}^v\}_{v=1}^m) = \sum_{v=1}^{m} w_v q_{\phi_c}(c|\mathbf{x}^v), \tag{2}$$

where ϕ_c are the trainable parameters of consistent encoder $E_c(\cdot)$. Using a Gaussian distribution as the probability distribution for view-consistency c, we refine Eq. (2) as

$$\mu^c = \sum_{v=1}^{m} w_v \mu_v^c,\ \sigma^v = \sum_{v=1}^{m} w_v \sigma_v^c, \tag{3}$$

where $[\mu^c, \sigma^v]$ is the mixture-of-posteriors ("experts"). $[\mu_v^c, \sigma_v^v]$ is the v-th posterior ("expert") and $[\mu_v^c, \sigma_v^v] = E(\mathbf{x}^v)$. $W = \left[w^1, \cdots, w^m\right]$ is the parameters of the router in the mixture-of-posteriors. The weights W are obtained using a softmax function over the outputs of a gating network:

$$W = Softmax(GatingNet(\{[\mu_v^c, \sigma_v^c]\}_{v=1}^m)). \tag{4}$$

The mixture-of-posteriors assigns posterior distributions of different views to multiple experts. Each expert focuses on a corresponding view, and the gating network integrates these learned representations. In such a manner, the approach extracts consistent information across all views.

To infer view-consistent representation c from mixture-of-posteriors, we utilize reparameterization trick

$$q_{\phi_c}(c|\{\mathbf{x}^v\}_{v=1}^m) = \mathcal{N}(\mu^c, (\sigma^c)^2) = \mu^c + \sigma^c \epsilon^c, \tag{5}$$

where $\epsilon^c \sim \mathcal{N}(0, 1)$. After obtaining the view-consistent representations c, we adopt a series of decoders $\{D_c^v(\cdot)\}_{v=1}^m$ to reconstruct view-specific content. This design allows the model to preserve the view-specific detail in the reconstructions when aligning the view-consistency across views.

Then, we can derive the ELBO of view-consistency as shown below

$$\begin{aligned}\mathcal{L}_c(\{\mathbf{x}^v\}) = &\sum_{v=1}^{m} \mathbb{E}_{q_{\phi_c}(c|\{\mathbf{x}^i\}_{i=1}^m)}\left[\log p_{\theta_c}(\mathbf{x}^v|c)\right] \\ &- \mathbf{KL}(q_{\phi_c}(c|\{\mathbf{x}^v\}_{v=1}^m)\|p(c)),\end{aligned} \tag{6}$$

where θ_c are the trainable parameters of view-consistent decoders $\{D_c^v(\cdot)\}_{v=1}^m$. $\mathbf{KL}(\cdot)$ denotes the KL divergence. During the inference stage, we discard the view-consistent decoder and solely utilize the consistent encoder $E_c(\cdot)$ to obtain the consistent representations, i.e., $c = E_c(\{\mathbf{x}^v\}_{v=1}^m)$.

3.3 View-specific Prior Learning

For view-specificity learning, we employ a set of view-specific encoders $\{E_s^v(\cdot)\}_{v=1}^m$ to obtain coarse view-specific representations s^v for each view. Then we concatenate the above-mentioned view-consistent representations and these view-specific representations, formulated as $z^v = [c, s^v]$, which is the input to the view-specific decoders $\{D_s^v(\cdot)\}_{v=1}^m$. The loss function for this process is shown as follows

$$\begin{aligned}\mathcal{L}_s(\mathbf{x}^v) = &\mathbb{E}_{q_{\phi_s}(z^v|\mathbf{x}^v)}\left[\log p_{\theta_s}(\mathbf{x}^v|z^v)\right] \\ &- \mathbf{KL}(q_{\phi_s}(s^v|\mathbf{x}^v)\|p(s^v)).\end{aligned} \tag{7}$$

We rewrite the KL divergence term in Eq. (7) to provide a more interpretable form, introducing two regularization terms instead of a single KL divergence term:

$$\begin{aligned}\mathbf{KL}(q_{\phi_s}(s^v|\mathbf{x}^v)\|p(s^v)) = &\mathbb{E}_{q(s^v)}\left[-\log p(s^v)\right] \\ &- \mathbb{H}\left[q_{\phi_s}(s^v|\mathbf{x}^v)\right],\end{aligned} \tag{8}$$

where the first is the cross-entropy between the aggregated posterior and the prior $p(s^v)$, and the second is the entropy of the variational posterior. Here, $q(s^v) = \frac{1}{n}\sum_{i=1}^n q_{\phi_s}(s^v|\mathbf{x}_i^v)$ [Makhzani *et al.*, 2015]. Typically, the prior is predefined, often chosen as a standard normal distribution. Notably, we find an optimal prior that optimizes the ELBO by minimizing the following Lagrange function:

$$\min_{p(s^v)}\ \mathbb{E}_{q(s^v)}\left[-\log p(s^v)\right] + \beta\left(\int p(s^v)\mathrm{d}s^v - 1\right). \tag{9}$$

where β is the Lagrange multiplier ensuring the proper normalization $p(s^v)$. The solution to this problem is the aggregated posterior $p^*(s^v) = \frac{1}{n}\sum_{i=1}^n q_{\phi_s}(s^v|\mathbf{x}_i^v)$, indicating that the optimal prior aligns with the aggregated posterior.

Based on the $p^*(s^v)$, we can get a new prior for view-specific representations s^v. However, directly using the aggregated posterior as a prior can lead to challenges such as overfitting, high computational complexity, and unlearnability. To mitigate these challenges, the optimal solution can be further approximated by a series of *pseudo-inputs*:

$$p(s^v) = \frac{1}{K}\sum_{k=1}^{K} q_{\phi_s}(s^v|\mathbf{u}_k^v), \tag{10}$$

where K is the number of pseudo-inputs, and $\mathbf{u}_k^v$ is a pseudo-input with the same shape as the real input $\mathbf{x}_i^v$. These pseudo-inputs are learnable parameters, optimized via backpropagation. The learned pseudo-inputs represent typical patterns in the data distribution. So they effectively act as hyperparameters for the prior distribution. Overall, by substituting Eq. (10) into Eq. (7), we derive the loss function of the v-th view-specificity. This loss harmonizes prior learning and representation learning. Learnable priors facilitate the quality of view-specific representations. During the inference stage, we solely utilize the view-specific encoder $E_s^v(\cdot)$ to obtain the v-th view-specific representations, i.e, $s^v = E_s^v(\mathbf{x}^v)$.

3.4 Representation Disentangling

For disentangling view-consistent and view-specific representations, we propose to minimize the upper bound of mutual information (MI) between s^v and c. The MI between two variables is typically formulated as

$$\mathbf{I}(s^v, c) = \mathbb{E}_{p(s^v,c)}\left[\log\frac{p(s^v, c)}{p(s^v)p(c)}\right] = \mathbb{E}_{p(s^v,c)}\left[\log\frac{q(s^v|c)}{p(s^v)}\right]. \tag{11}$$

However, the conditional probability $q(s^v|c)$ is unknown. To address this issue, we adopt contrastive log-ratio upper bound (CLUB) [Cheng *et al.*, 2020], which uses a variational distribution $q_\theta(s^v|c)$ to approximate $q(s^v|c)$, as shown below

$$\begin{aligned}\mathbf{I}_{CLUB}(s^v, c) = &\mathbb{E}_{p(s^v,c)}\left[\log q_\theta(s^v|c)\right] \\ &- \mathbb{E}_{p(s^v)}\mathbb{E}_{p(c)}\left[\log q_\theta(s^v|c)\right] \\ \geq &\mathbf{I}(s^v, c).\end{aligned} \tag{12}$$

Given Eq. (12), we then define disentangling loss $\mathcal{L}_d$ as

$$\mathcal{L}_d(\mathbf{x}^v) = \mathbf{I}_{CLUB}(s^v, c). \tag{13}$$

Objective Function. Finally, merging the Eq. (6), Eq. (7) and Eq. (13) via the adaptive parameter λ, we have the joint loss function of our CL2P as follows

$$\mathcal{L}(\{\mathbf{x}^v\}) = \lambda\mathcal{L}_c(\{\mathbf{x}^v\}) + (1-\lambda)\sum_{v=1}^{m}(\mathcal{L}_s(\mathbf{x}^v) + \mathcal{L}_d(\mathbf{x}^v)). \tag{14}$$

We denote that the joint loss function leverages an adaptive parameter λ to make the learning of c and $\{s^v\}_{v=1}^m$ conform to the *Property* 1 of neural networks. In addition, the training pseudo-code of the proposed CL2P is outlined in Algorithm 1, which describes the steps for updating the model parameters by optimizing the total loss function.

Algorithm 1 Training of the proposed CL2P.

Input: Multi-view dataset $\mathcal{D} = \{\mathbf{x}^1, \cdots, \mathbf{x}^m\}$.
Parameter: Total training epochs $\mathcal{T}_{max}$, current training epoch $\mathcal{T}$, K pseudo-inputs of each view, parameters $\theta_c, \phi_c, \theta_s, \phi_s$ of encoders and decoders.
Output: View-consistent representations c, and view-specific representations $\{s^v\}_{v=1}^m$.

Initial the K pseudo-inputs for each view
while $\mathcal{T} \leq \mathcal{T}_{max}$ **do**
  $c \leftarrow E_c(\{\mathbf{x}^v\}_{v=1}^m)$, and $\{s^v\}_{v=1}^m \leftarrow \{E_s^v(\mathbf{x}^v)\}_{v=1}^m$
  Compute the consistency-loss $\mathcal{L}_c$ using Eq. (6)
  Compute the specificity-loss $\mathcal{L}_s^v$ using Eq. (7)
  Compute the disentangling-loss $\mathcal{L}_d^v$ using Eq. (13)
  Update $\mathcal{T} \leftarrow \mathcal{T} + 1$, $\lambda \leftarrow 1 - (\frac{\mathcal{T}}{\mathcal{T}_{max}})^2$
  Compute the total loss $\mathcal{L}_{model}$ using Eq. (14) via λ
  Update $\theta_c, \phi_c, \theta_s, \phi_s \leftarrow \Delta\mathcal{L}_{model}(\theta_c, \phi_c, \theta_s, \phi_s)$
end while
return c and $\{s^v\}_{v=1}^m$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our CL2P and other competitive methods using five real-world datasets, including: **(1) Edge-MNIST** [LeCun *et al.*, 1998], which consists of 0-9 grayscale digit images. The views contain the original digit images and the edge-detected version. **(2) Edge-Fashion** [Xiao *et al.*, 2017], containing grayscale images of various fashion clothing. We generate the second view by applying the same edge-detection technique used in Edge-MNIST; **(3) Multi-COIL-20** [Nene *et al.*, 1996b], which contains grayscale images of 20 distinct objects, captured from multiple angles. We construct a three-view dataset by randomly grouping the images of each object into three groups. **(4) Multi-COIL-100** [Nene *et al.*, 1996a] which depicts from different angles containing RGB images of 100 items. Similar to Multi-COIL-20, we create a three-view dataset by randomly grouping the images of an object into three different groups; **(5) Multi-Office-31** [Saenko *et al.*, 2010], which consists of objects commonly encountered in office settings, from three distinct domains (Amazon, DSLR, and Webcam). We construct a three-view dataset with each domain serving as a view. A summary of the statistics for these datasets is shown in Table 1.

Dataset	Samples	Views	Classes	Size
Edge-MNIST	70,000	2	10	32×32
Edge-Fashion	70,000	2	10	32×32
Multi-COIL-20	480	3	20	64×64
Multi-COIL-100	2,400	3	100	64×64
Multi-Office-31	2,817	3	31	64×64

Table 1: A summary of dataset statistics.

Baselines. We compare CL2P against three categories of baseline methods, namely, i) single-view learning methods: Beta-VAE [Higgins *et al.*, 2017] and Joint-VAE [Dupont, 2018]; ii) model-based multi-view methods: MFLVC [Xu *et al.*, 2022], GCFAgg [Yan *et al.*, 2023], SCM [Luo *et al.*, 2024], and CSOT [Zhang *et al.*, 2024]; and iii) information-theory-based methods: MVAE [Wu and Goodman, 2018], DVIB [Bao, 2021], Multi-VAE [Xu *et al.*, 2021], and MRDD [Ke *et al.*, 2024]. Our method falls into the third category. Additionally, we evaluate two variants of our method: CL2P-C, which uses only view-consistent representations c, and CL2P-CS, which incorporates both view-consistent representations c and view-specific representations $\{s^v\}_{v=1}^m$.

Implementation Details. We implement the proposed method and other comparison methods on PyTorch 2.1.0, utilizing one NVIDIA A10 GPU (24 GB). We employ convolutional networks [He *et al.*, 2016] to construct our model, with a uniform encoder-decoder architecture across all datasets. Both view-consistency and view-specificity dimensions are set to 20. The number of pseudo-inputs is fixed at 250, initialized with randomly selected training data. We train our model for 200 epochs using the AdamW optimizer with the learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. We set a batch size of 128 for Edge-MNIST and Edge-Fashion, and 32 for Multi-COIL-20, Multi-COIL-100, and Multi-Office-31. For all baseline models, we use the optimal settings as recommended in the original paper for fair comparison.

4.2 Experimental Results

Task Settings. We evaluate all baseline models across five datasets for both clustering and classification tasks. For clustering, we use the k-means algorithm [Hartigan and Wong, 1979]. For classification, we apply support vector classification (SVC) [Hsu, 2003] with an 80:20 train-test split ratio. We use Accuracy (ACC) as the metric to evaluate the performance of each model for both clustering and classification tasks on five datasets. Each model is evaluated over 10 runs, and we report the average values along with the variances. Notably, for single-view methods, we select the results from the best view as the evaluation outcome. For multi-view methods limited to two views, we choose the optimal pair of views as inputs for the models.

Overall Evaluation. Tables 2 and 3 show the performance of clustering and classification tasks, respectively. From the

Methods	Edge-MNIST		Edge-Fashion		Multi-COIL-20		Multi-COIL-100		Multi-Office-31	
	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI
Beta-VAE [Higgins *et al.*, 2017]	61.36±0.58	53.92±0.93	56.36±0.50	55.21±0.95	67.19±0.41	77.80±0.90	59.48±0.60	80.62±0.35	23.98±0.48	28.12±0.26
Joint-VAE [Dupont, 2018]	24.50±0.69	14.58±0.30	33.59±0.62	24.10±0.85	51.43±0.93	64.46±0.89	44.38±0.55	70.71±0.38	18.95±0.50	27.24±0.30
MFLVC† [Xu *et al.*, 2022]	33.62±0.83	29.00±0.93	26.95±0.94	27.03±0.24	54.69±0.77	75.07±0.49	34.34±0.39	81.97±0.06	24.39±0.51	34.79±0.30
GCFAgg† [Yan *et al.*, 2023]	32.46±0.80	24.93±0.78	37.75±0.60	34.83±0.76	67.58±0.83	78.77±0.57	59.76±0.97	82.16±0.32	34.61±0.90	45.55±0.56
SCM† [Luo *et al.*, 2024]	34.57±0.92	29.58±0.95	35.23±0.55	32.16±0.20	58.59±0.83	68.08±0.53	58.29±0.73	81.72±0.39	15.53±0.57	21.01±0.40
CSOT† [Zhang *et al.*, 2024]	32.55±0.78	32.35±0.49	37.36±0.76	37.80±0.91	49.16±0.87	63.15±0.70	53.01±0.91	75.17±0.33	12.32±0.43	15.48±0.17
MVAE [Wu and Goodman, 2018]	51.12±0.61	45.15±0.62	52.67±0.78	55.25±0.71	69.86±0.98	77.32±0.98	65.98±0.70	84.74±0.61	35.25±0.97	46.24±0.94
DVIB [Bao, 2021]	29.67±0.96	20.00±0.94	31.99±0.88	23.51±0.96	52.86±0.94	67.43±0.73	48.45±0.89	73.13±0.42	15.92±0.45	22.10±0.46
Multi-VAE [Xu *et al.*, 2021]	59.60±0.73	61.71±0.56	44.19±0.27	40.69±0.29	68.35±0.82	79.91±0.71	46.74±0.94	70.01±0.66	27.69±0.72	33.30±0.58
MRDD [Ke *et al.*, 2024]	69.88±0.28	64.18±0.52	58.83±1.00	60.51±0.94	68.96±0.98	79.42±0.86	64.06±0.96	84.64±0.56	31.90±0.95	40.22±0.62
CL2P-C (Ours)	63.93±0.95	58.91±0.98	57.17±0.52	62.45±0.88	**73.40±0.61**	**83.24±0.69**	**67.68±0.81**	**85.71±0.56**	29.54±0.90	37.31±0.69
CL2P-CS (Ours)	**72.22±0.41**	**67.50±0.91**	**60.74±0.85**	**65.53±0.99**	69.79±0.21	80.39±0.91	61.42±0.90	82.74±0.42	**37.19±0.94**	**50.57±0.74**
SOTA	+2.34	+3.32	+1.91	+5.02	+3.54	+3.33	+1.70	+0.97	+1.94	+4.33

Table 2: **Clustering results** (%) on five datasets. **Bold** indicates the best results, while underline marks the second-best results. Dagger footnote † denotes that the dimensionality of the latent representations is set to 20.

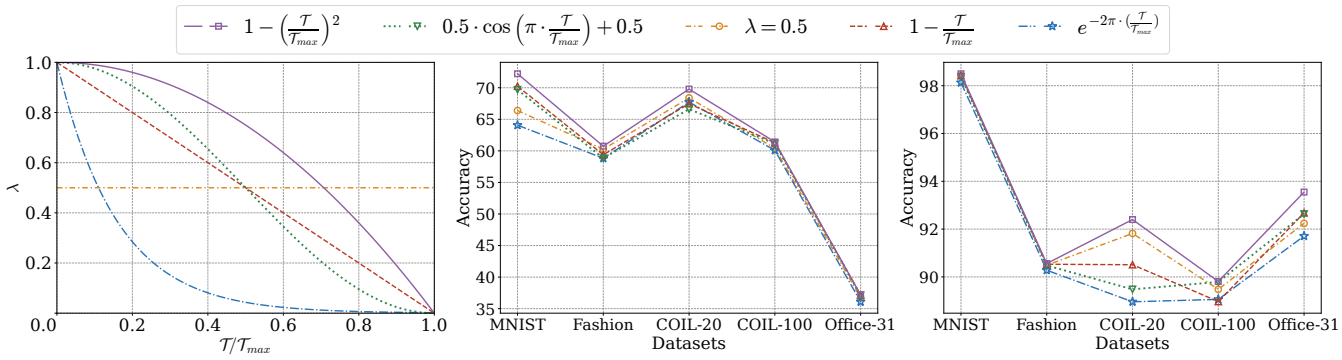
Methods	Edge-MNIST		Edge-Fashion		Multi-COIL-20		Multi-COIL-100		Multi-Office-31	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Beta-VAE [Higgins *et al.*, 2017]	94.00±0.16	93.39±0.16	80.84±0.34	80.58±0.31	88.74±0.61	86.58±0.86	79.27±0.94	75.84±0.79	40.00±0.76	36.62±0.61
Joint-VAE [Dupont, 2018]	95.30±0.22	95.26±0.22	79.82±0.32	79.57±0.24	53.68±0.61	47.28±0.58	36.20±0.85	26.48±0.88	25.67±0.98	22.08±0.87
MFLVC† [Xu *et al.*, 2022]	55.30±0.40	54.09±0.44	41.95±0.33	37.44±0.16	55.84±0.66	49.88±0.61	29.84±0.89	20.13±0.83	32.02±0.97	28.61±0.70
GCFAgg† [Yan *et al.*, 2023]	75.86±0.36	75.40±0.32	78.85±0.29	78.46±0.33	61.69±0.61	56.85±0.56	48.37±0.91	39.82±0.97	54.38±0.67	51.43±0.51
SCM† [Luo *et al.*, 2024]	89.39±0.25	89.22±0.25	81.56±0.31	81.28±0.31	78.79±0.97	78.99±0.94	63.54±0.21	55.69±0.84	21.89±0.77	19.61±0.93
CSOT† [Zhang *et al.*, 2024]	54.83±0.33	52.24±0.34	58.86±0.35	54.02±0.35	35.06±0.61	26.35±0.71	33.16±0.99	25.99±0.99	25.10±0.87	22.26±0.75
MVAE [Wu and Goodman, 2018]	97.76±0.15	97.74±0.15	88.93±0.18	88.89±0.17	93.37±0.89	91.17±0.90	92.23±0.92	91.60±0.91	88.41±0.63	88.96± 0.86
DVIB [Bao, 2021]	76.19±0.43	75.79±0.39	71.68±0.36	72.62±0.39	71.86±0.97	69.24±0.87	72.69±0.93	70.57±0.70	37.03±0.74	36.70±0.97
Multi-VAE [Xu *et al.*, 2021]	95.76±0.15	95.72±0.15	84.69±0.23	84.43±0.19	88.31±0.58	85.76±0.93	74.93±0.62	72.58±0.97	61.46±0.65	61.73±0.87
MRDD [Ke *et al.*, 2024]	98.53±0.10	98.52±0.10	88.81±0.29	88.73±0.30	94.38±0.97	95.67±0.96	91.11±0.86	90.50±0.76	82.46±0.91	81.31±0.86
CL2P-C (Ours)	**98.67±0.10**	**98.65±0.10**	89.35±0.24	89.28±0.26	**96.36±0.78**	**96.40±0.99**	**93.12±0.84**	**92.56±0.73**	73.99±0.95	73.18±0.98
CL2P-CS (Ours)	98.50±0.09	98.49±0.10	**90.56±0.28**	**90.46±0.29**	92.40±0.83	92.78±0.95	89.81±0.98	89.69±0.85	**93.42±0.95**	**93.05±0.75**
SOTA	+0.14	+0.13	+1.63	+1.57	+1.98	+0.73	+0.89	+0.96	+5.01	+4.09

Table 3: **Classification results** (%) on five datasets. **Bold** indicates the best results, while underline marks the second-best results. Dagger footnote † denotes that the dimensionality of the latent representations is set to 20.

results, we have the following observations: (1) The proposed CL2P is superior to all baseline methods. The results show the effectiveness of the proposed method. (2) CL2P-C outperforms CL2P-CS on the Multi-COIL-20 and Multi-COIL-100 datasets, which indicates that naive concatenation may degrade the overall quality of representation. (3) Single-view methods are inferior to multi-view methods in most cases. On the contrary, single-view methods sometimes outperform certain multi-view methods. The reason might be due to that some specific views are dominant and those multi-view methods struggle to handle data with significant view-specific differences. (4) Information-theory-based methods generally outperform model-based methods, indicating that information-theoretic constraints guide the model to learn higher-quality representations. Nevertheless, those methods overlook the inherent properties of neural networks and thus fail to surpass our method.

Ablation Study. We conduct an ablation study to measure the contribution of the four key components, i.e., curriculum learning (CL), prior learning (PL), disentangling learning (DL), and Mixture-of-Experts (MoE), in our CL2P-CS method. We remove each component from CL2P-CS and then evaluate the performance of each resulting variant on five datasets using the clustering performance. The ablation results are shown in Table 4. From the results, we can see that every component in our model plays an important role. This shows that all components in our model are essential. For the datasets Multi-COIL-100 and Multi-Office-31, the ablation results show slight differences. This indicates that more advanced encoders and decoders may be required to enhance the model's capability for information extraction.

Scheduler Ablation. We explore several different strategies to generate the adaptive trade-off parameter λ, including line decay ($\lambda = 1 - \frac{\mathcal{T}}{\mathcal{T}_{max}}$), cosine decay ($\lambda = 0.5 \cdot cos(\pi \cdot$

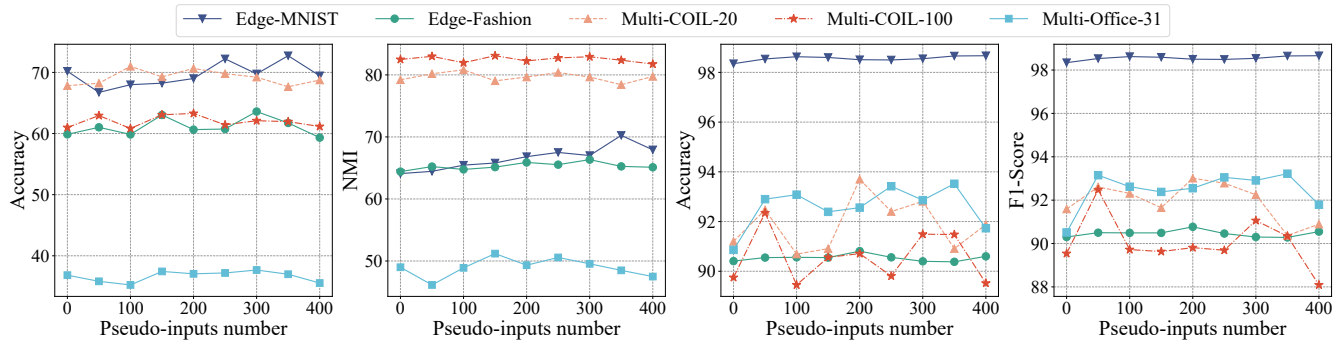


(a) λ curves against $\mathcal{T}/\mathcal{T}_{max}$ on different schedulers.

(b) Clustering accuracy (%) of different schedulers on five datasets.

(c) Classification accuracy (%) of different schedulers on five datasets.

Figure 3: Study of different schedulers on curriculum learning.



(a) Clustering accuracy (%).

(b) Clustering NMI (%).

(c) Classification accuracy (%).

(d) Classification F1-Score (%).

Figure 4: Different metrics against the number K of pseudo-inputs.

Methods	MNIST	Fashion	COIL-20	COIL-100	Office-31
CL2P-CS	**72.22**	**60.74**	**69.79**	**61.42**	**37.19**
w/o CL	66.39	60.26	68.39	60.43	36.72
w/o PL	70.19	59.87	67.83	60.98	36.83
w/o DL	63.35	58.16	67.36	60.06	36.52
w/o MoE	70.48	60.44	66.54	60.85	36.11

Table 4: Component ablation on the proposed model. Clustering accuracy scores (%) on five datasets. Best results are in **bold**.

$\frac{\mathcal{T}}{\mathcal{T}_{max}}$) + 0.5), exponential decay ($\lambda = e^{-2\pi\cdot(\frac{\mathcal{T}}{\mathcal{T}_{max}})}$), and identity ($\lambda = 0.5$), in our CL2P-CS. The results are shown in Figure 3. Figure 3(a) illustrates the variation of λ with respect to different schedulers. Figures 3(b) and 3(c) present the performance for each scheduler. From the results, we observe that a slower decay rate for parameter λ yields better performance. When λ decays too rapidly, it prematurely shifts the attention of networks away from view-consistency. The premature shift reduces the quality and effectiveness of consistency, leading to a decline in overall representations.

Study on the number of pseudo-inputs K. We investigate the effect of parameter K by conducting a grid search of the settings {0, 50, 100, 150, 200, 250, 300, 350, 400}, in the CL2P-CS method. The results are shown in Figure 4. We find that the model remains robust with a wide range of parameter values, e.g., [150, 350]. However, both excessive and insufficient numbers of pseudo-inputs lead to suboptimal performance. Additionally, certain settings may introduce side effects on the performance. Therefore, the selection of the K value is crucial to the overall performance of the model.

5 Conclusion

In this paper, we proposed a novel multi-view representation learning framework called CL2P. Specifically, we extend curriculum learning to tackle the intrinsic differences between view-consistency learning and view-specificity learning. Moreover, we devise a learnable prior for view-specific representations from different views. Additionally, we saddle the model with a mixture-of-experts layer and a disentangling module to enhance the representation quality. We evaluated CL2P and baseline methods on clustering and classification tasks. The results show that CL2P outperforms baselines markedly on five real-world datasets. In the future, we plan to integrate more advanced encoders-decoders into CL2P to enhance representational capability, particularly for multi-modal data. We also intend to explore CL2P in incomplete multi-view and semi-supervised learning settings.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (grant no. 2024YFB4710604), NSFC (grant nos. 62406209, 62472295, U24B20174, and U21B2040), Sichuan Science and Technology Planning Project (grant nos. 2024NSFTD0130, 2024NSFTD0038, and 2025ZNSFSC1486), and the Fundamental Research Funds for the Central Universities (grant nos. CJ202303, CJ202403, and YT202421).

References

[Arpit *et al.*, 2017] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR, 2017.

[Bao, 2021] Feng Bao. Disentangled variational information bottleneck for multiview representation learning. In *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part II 1*, pages 91–102. Springer, 2021.

[Bauer and Mnih, 2019] Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75. PMLR, 2019.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[Cheng *et al.*, 2020] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: a contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pages 1779–1788. PMLR, 2020.

[Dupont, 2018] Emilien Dupont. Learning disentangled joint continuous and discrete representations. *Advances in Neural Information Processing Systems*, 31, 2018.

[Federici *et al.*, 2020] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

[Hartigan and Wong, 1979] 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.

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

[He *et al.*, 2024] Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *ACM Multimedia 2024*, 2024.

[Higgins *et al.*, 2017] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: learning basic visual concepts with a constrained variational framework. In *The 5th International Conference on Learning Representations*. OpenReview.net, 2017.

[Hsu, 2003] ChihWei Hsu. A practical guide to support vector classification. *Department of Computer Science, National Taiwan University*, 2003.

[Hu *et al.*, 2023] Shizhe Hu, Guoliang Zou, Chaoyang Zhang, Zhengzheng Lou, Ruilin Geng, and Yangdong Ye. Joint contrastive triple-learning for deep multi-view clustering. *Information Processing and Management*, 60(3):103284, 2023.

[Ju *et al.*, 2021] Jiahuei Ju, Jhenghong Yang, and Chuanju Wang. Text-to-text multi-view learning for passage reranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1803–1807, New York, NY, USA, 2021. Association for Computing Machinery.

[Ke *et al.*, 2024] Guanzhou Ke, Bo Wang, Xiaoli Wang, and Shengfeng He. Rethinking multi-view representation learning via distilled disentangling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26774–26783, 2024.

[Kingma, 2013] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kumar *et al.*, 2024] Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *The Twelfth International Conference on Learning Representations*, 2024.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li *et al.*, 2023] Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. *arXiv preprint arXiv:2301.11045*, 2023.

[Lu *et al.*, 2024] Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14193–14201, 2024.

[Luo *et al.*, 2024] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view

clustering with data-level fusion. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4697–4705, 8 2024.

[Ma *et al.*, 2024] Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 2024.

[Makhzani *et al.*, 2015] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015.

[Nene *et al.*, 1996a] Sameer A Nene, Shree K Nayar, and Hiroshi Murase. Columbia object image library (coil-100). In *Technical Report, Department of Computer Science, Columbia University CUCS-006-96*, 1996.

[Nene *et al.*, 1996b] Sameer A Nene, Shree K Nayar, and Hiroshi Murase. Columbia object image library (coil-20). In *Technical Report, Department of Computer Science, Columbia University CUCS-005-96*, 1996.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010.

[Shi *et al.*, 2019] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H.S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[Soviany *et al.*, 2022] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: a survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

[Sutter *et al.*, 2024] Thomas Sutter, Yang Meng, Andrea Agostini, Daphné Chopard, Norbert Fortin, Julia Vogt, Babak Shahbaba, and Stephan Mandt. Unity by diversity: improved representation learning for multimodal vaes. *Advances in Neural Information Processing Systems*, 37:74262–74297, 2024.

[Takahashi *et al.*, 2019] Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Variational autoencoder with implicit optimal priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5066–5073, 2019.

[Wang *et al.*, 2021] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

[Wang *et al.*, 2023] Hao Wang, Zhi-Qi Cheng, Jingdong Sun, Xin Yang, Xiao Wu, Hongyang Chen, and Yan Yang. Debunking free fusion myth: online multi-view anomaly detection with disentangled product-of-experts modeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3277–3286, New York, NY, USA, 2023. Association for Computing Machinery.

[Wu and Goodman, 2018] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[Xu *et al.*, 2019] Haowen Xu, Wenxiao Chen, Jinlin Lai, Zhihan Li, Youjian Zhao, and Dan Pei. On the necessity and effectiveness of learning the prior of variational auto-encoder. *arXiv preprint arXiv:1905.13452*, 2019.

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

[Xu *et al.*, 2022] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16051–16060, 2022.

[Yan *et al.*, 2023] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19863–19872, 2023.

[Zhang *et al.*, 2017] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, pages 1–15, 2017.

[Zhang *et al.*, 2024] Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, Yonghuai Liu, and Wei Zhang. Learning common semantics via optimal transport for contrastive multi-view clustering. *IEEE Transactions on Image Processing*, 33:4501–4515, 2024.

[Zou *et al.*, 2024] Guoliang Zou, Yangdong Ye, Tongji Chen, and Shizhe Hu. Learning dual enhanced representation for contrastive multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 8731–8739, New York, NY, USA, 2024. Association for Computing Machinery.