Multi-View Representation Learning via Total Correlation Objective

Hyeong.Joo Hwang, Geon-Hyeong Kim, Seunghoon Hong, Kee-Eung Kim
KAIST
{hjhwang, ghkim}@ai.kaist.ac.kr, {seunghoon.hong, kekim}@kaist.ac.kr

Abstract

Multi-View Representation Learning (MVRL) aims to discover a shared representation of observations from different views with the complex underlying correlation. In this paper, we propose a variational approach which casts MVRL as maximizing the amount of total correlation reduced by the representation, aiming to learn a shared latent representation that is informative yet succinct to capture the correlation among multiple views. To this end, we introduce a tractable surrogate objective function under the proposed framework, which allows our method to fuse and calibrate the observations in the representation space. From the information theoretic perspective, we show that our framework subsumes existing multi-view generative models. Lastly, we show that our approach straightforwardly extends to the Partial MVRL (PMVRL) setting, where the observations are missing without any regular pattern. We demonstrate the effectiveness of our approach in the multi-view translation and classification tasks, outperforming strong baseline methods.

1 Introduction

Multi-View Representation Learning (MVRL) aims to learn a shared representation of multiple observations from different types of views. In MVRL, it is important to encourage the shared representation to be complete enough to capture the correlation across views without losing view-specific information so that the learned representation can be readily applied to rich set of downstream tasks. For example, in sensor fusion for the multi-sensor system [51] and in clinical diagnosis based on patients' various types of medical records [47, 49], one would like to aggregate all the information from various observations in order to uncover the true underlying factors under the correlation. Although using views as many as possible seems to be always beneficial for the task of learning a good representation, it can make the problem itself harder depending on the complexity of correlations across all views and the scalability of handling a number of views.

MVRL becomes even more challenging when the model does not always have access to complete observations from all views for every data instance during training, which we call Partial Multi-View Representation Learning (PMVRL). PMVRL is closer to the practical setting since it is unrealistic to expect that observations from all different kinds of views are always available. For example, it is unlikely for all the sensors in a sensor-based system to have the same frequency to update their measurements or for all kinds of medical records to be available for any patient.

Considering those difficulties, any desirable MVRL methods are encouraged to satisfy three desiderata. The first one is *scalability* to the number of input views. The method should handle multiple observations in a way computationally scalable to arbitrary many views in both training and testing time. In addition, the method needs to be robust to *partial observability*. The method should be able to combine any combination of available observations in the representation space in test time. This is important for PMVRL because we have to handle observations arbitrarily missing not only in testing data but also in training data. Lastly, the method should discover *cross-view association*,

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

which can be learned by identifying both shared and view-specific factors of variation of each view in the representation space. Associating views correctly in the representation space allows the method to utilize any additional observations from different views to improve identification of the shared factors without determining any unobserved ones.

Combining multiple of Variational Auto-Encoder (VAE) [22] for each view[1], several generative models [30, 33, 34, 46] have been recently proposed to address MVRL. Since naive optimization of the evidence lower bound (ELBO) on the joint likelihood of multiple views does not address any of the desiderata, those methods impose different structural bias on the joint representation encoders whose strengths and weaknesses are complementary (see Section 2.4). Although they showed encouraging performance on multiple tasks such as predicting common attributes and inferring missing views, they often fail to simultaneously capture both the shared and view-specific factors of variation and turn out to poorly associate views in our experiments.

In this paper, we address the problem of MVRL with a principled approach grounded in information theory. Specifically, we formulate the representation learning task as maximizing the reduction in Total Correlation [12, 38, 39]. Based on our formulation, we derive a novel objective function that not only offers tractable optimization but also introduces multiple types of variational information bottlenecks which successfully associate views. We then show that our method naturally extends to the PMVRL setting via inverse variance weighting, a classical approach used in sensor fusion. We demonstrate the validity and effectiveness of our method in the multi-view translation and classification downstream tasks. Our contributions are three-fold:

1. Measuring the informativeness of the multi-view representation by Total Correlation, we propose a general information-theoretic framework to learn a complete representation, which encompasses existing multi-view generative models.
2. Under the proposed framework, we identify drawbacks of optimizing ELBO and derive a novel objective function that resolves them. Specifically, our method yields a representation that correctly associates views by capturing not only the common factors of variation but also the view-specific ones.
3. We conducted extensive evaluation with comparing methods and ours in translation and classification tasks in both MVRL and PMVRL settings, showing that our method is the most reliable method to obtain the latent representation agnostic to the downstream tasks.

2 Approach

Let $\vec{o} = \{o_v\}_{v=1}^{V}$ be the observation of a data instance composed of V different views, which is sampled from an unknown joint distribution $p_D(\vec{o})$, where we emphasize that this is a data distribution using the subscript D. Given these observations, MVRL is the task of learning a *complete* representation z across the views $\vec{o}$. Following [50], we define a complete representation as follows.

Definition 1 (Completeness for Multi-View Representation [50]) *A multi-view representation z is complete if each observation, i.e., o_v from $\vec{o}$, can be reconstructed from a mapping $f_v(\cdot)$, i.e., $o_v = f_v(z)$.*

The definition directly indicates that a complete representation describes all factors of variations in $\vec{o}$, since every view can be reconstructed solely from the complete representation. While MVRL considers complete observations for training data, PMVRL assumes otherwise, i.e. some views being missing in the training data. This poses a unique challenge of (1) learning to produce the complete representation with partial views (2) whose availability varies per instance in training and testing.

We first relate informativeness of the representation measured in terms of Total Correlation (TC) and the goal of MVRL (Section 2.1). We then show that the TC-based MVRL objective function encompasses existing multimodal generative models and analyze its limits based on a straightforward variational lower bound (Section 2.2). To resolve those limits, we derive an alternative variational lower bound which suits better for MVRL (Section 2.3). Lastly, we finalize our formulation by proposing the representation aggregation model that naturally extends to PMVRL (Section 2.4).

[1]We follow the conventional terminology in the related literature [11, 14, 24, 41, 44, 47, 48, 50, 53], but any choice among views, modalities [30, 33, 34, 46], and domains [19] can be suitable for our paper and our baseline methods, as long as there is a common latent representation that explains different views / modalities / domains.

2.1 Multi-View Representation Learning with Total Correlation

A complete representation Z should be informative enough to explain the correlation among V different views. Total correlation (TC) [42], defined as the Kullback-Leibler divergence of the joint distribution from the factored marginals, measures the amount of information shared among a finite set of random variables. In our MVRL context, TC is defined as

$$TC(\vec{O}) \triangleq D_{KL}\left[p_D\left(\vec{o}\right) \| \prod_{v=1}^{V} p_D\left(o_v\right)\right]. \tag{1}$$

We aim to find the encoder $p_\theta\left(z|\vec{o}\right)$ such that the knowledge of z would reduce TC as much as possible. This can be formulated by maximizing the objective

$$TC_\theta(\vec{O}; Z) \triangleq TC(\vec{O}) - TC_\theta(\vec{O} \mid Z), \tag{2}$$

where the conditional TC in the last term is given by

$$TC_\theta(\vec{O}|Z) \triangleq \mathbb{E}_{p_\theta(z)}\left[D_{KL}\left[p_\theta\left(\vec{o}|z\right) \| \prod_{v=1}^{V} p_\theta\left(o_v|z\right)\right]\right], \tag{3}$$

which is the expected Kullback-Leibler divergence of the joint conditional from the factored conditionals. The parameterized distributions in the above formula involve the encoder $p_\theta\left(z|\vec{o}\right)$ as follows: $p_\theta\left(z\right) = \int p_\theta\left(z|\vec{o}\right) p_D\left(\vec{o}\right) d\vec{o}$, $p_\theta\left(\vec{o}|z\right) = p_\theta\left(z|\vec{o}\right) p_D\left(\vec{o}\right)/p_\theta\left(z\right)$, and $p_\theta\left(o_v|z\right) = \int p_\theta\left(\vec{o}|z\right) d\vec{o}_{\setminus v}$.

Intuitively, minimization of Eq. (3) (*i.e.*, maximization of Eq. (2)) suits well for MVRL, since (1) any complete representation Z would minimize Eq. (3) (*e.g.*, $Z = \vec{o}$), and (2) it accords with the theoretical result that complete representation z should factorize the generative distribution [44, 50].

Further decomposition of (2) reveals that it encourages Z to encode the correlation across views [12], a desirable property for MVRL:

$$TC_\theta(\vec{O}; Z) = \sum_{v=1}^{V} I_\theta\left(O_v; Z\right) - I_\theta(\vec{O}; Z). \tag{4}$$

The first term in Eq. (4), Mutual Information (MI) between each observation and the representation, enforces the representation to be informative for every observation. On the other hand, the second term takes the role of Information Bottleneck (IB), which encourages the encoder to learn minimal sufficient representation [37]. Consequently, simultaneous optimization of both terms naturally allows the representation Z to capture correlations among views while being minimally sufficient. Note that when V=2, Eq. (4) coincides with Interaction Information [4, 19, 25, 36], which quantifies the amount of information shared among two views and their joint representation (see Section A.4).

2.2 Limitations of VAE models for multi-view data

Unfortunately, a direct optimization of MIs in Eq. (4) is intractable [2, 12]. A straightforward approach would be employing approximate distributions $q_\phi^v\left(o_v|z\right) \approx p_\theta\left(o_v|z\right)$ and $r\left(z\right) \approx p_\theta\left(z\right)$, and optimize a variational lower bound of Eq. (4) as follows (see Section A.1 in the supplementary material for full derivation):

$$TC_\theta(\vec{O}; Z) \geq \sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z|\vec{o})p_D(\vec{o})}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \underbrace{\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r(z)\right]\right]}_{\text{VIB}}, \tag{5}$$

where the entropy terms can be dropped from optimization since they are determined by the true data distribution $p_D\left(\vec{o}\right)$. Without the entropy terms, we note that Eq. (5) is essentially identical to evidence lower bound (ELBO) of the variational auto-encoder (VAE) models for multi-view data [30, 33, 34, 46] by switching the notations, for p for encoder and q for decoder. This notation switch follows the convention in [2, 12].

Although this lower bound contains the variational information bottleneck term that encourages the representation to be minimally sufficient for generalizing well even with small training data [2], it has the following fundamental limitations:

1. **Unbalanced representation**: If a subset of views is overwhelmingly informative enough to reconstruct the others, the encoder $p_\theta\left(z|\vec{o}\right)$ may learn to rely on those views while ignoring the rest, yielding a degenerate solution of Eq. (4) that fails cross-view association. This is problematic in MVRL since such views may not be available at test time.
2. **Missing views**: The encoder $p_\theta\left(z|\vec{o}\right)$ requires complete observations $\vec{o}$ not only in training but also in testing phases, while MVRL requires the model to encode incomplete observations $\tilde{o} \subseteq \vec{o}$ in test time. Furthermore, PMVRL requires to handle incomplete observations even in training.

In order to overcome these challenges, prior methods impose special structures (i.e. inductive hypotheses) for $p_\theta\left(z|\vec{o}\right)$, such as Product of Experts (PoE) [17, 46], Mixture of Experts (MoE) [30], or Mixture of Product of Experts (MoPoE) [34]. In our work, we present a more principled approach to this problem by deriving an alternative lower bound, described in the next section.

2.3 Conditional Variational Information Bottleneck

To resolve the first issue raised in the previous section, we start from Eq. (4) and reformulating it as follows:

$$\begin{aligned} TC_\theta(\vec{O}; Z) &= \sum_{v=1}^{V} \left[\frac{V-1}{V} I_\theta\left(O_v; Z\right) + \frac{1}{V} I_\theta\left(O_v; Z\right) - \frac{1}{V} I_\theta\left(\vec{O}; Z\right) \right] \\ &= \frac{1}{V} \sum_{v=1}^{V} \left[\left(V-1\right) I_\theta\left(O_v; Z\right) - I_\theta(\vec{O}_{\setminus v}; Z|O_v) \right], \end{aligned} \tag{6}$$

where the last equality is due to the chain rule of MI (see Section A.5 in the supplementary material for details). Interestingly, Eq. (6) transforms IB in Eq. (4) into multiple *conditional* MIs between the latent representation and $V-1$ other views given every view, each of which penalizes the extra information of the representation *not inferable* from the given view. Although Eq. (6) is essentially equal to Eq. (4), its conditional information constraints give us intuition to derive a new tractable lower bound on $TC_\theta(\vec{O}; Z)$ that regularizes unbalanced representation, which we present below.

Since the conditional MIs in Eq. (6) involve $p_\theta(z|o_v) = \int p_\theta(z|\vec{o}) p_D(\vec{o}_{\setminus v}|o_v) d\vec{o}_{\setminus v}$ which requires to compute intractable integration, we use the variational upper bound of those terms by introducing approximate distributions $r_\psi^v(z|o_v) \approx p_\theta\left(z|o_v\right)$ as follows (see Section A.2 in the supplementary material for the full derivation and analysis):

$$\begin{aligned} TC_\theta(\vec{O}; Z) \geq \frac{V-1}{V} \sum_{v=1}^{V} \left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z|\vec{o}) p_D(\vec{o})} \left[\ln q_\phi^v\left(o_v|z\right) \right] \right] \\ - \frac{1}{V} \sum_{v=1}^{V} \underbrace{\mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[p_\theta(z|\vec{o}) \| r_\psi^v(z|o_v) \right] \right]}_{\text{Conditional VIB}}. \end{aligned} \tag{7}$$

This lower bound is equipped with conditional VIBs which provide a number of benefits over Eq. (5) in handling challenges mentioned in the previous section. First, conditional VIBs, which upper bounds condtional MIs in Eq. (6) by introducing the view-specific encoder $r_\psi^v\left(z|o_v\right)$ for each view, regularize $p_\theta\left(z|\vec{o}\right)$ to encode representation *inferable* from $r_\psi^v\left(z|o_v\right)$ of every view. Consequently, the joint representation is enforced to be balanced rather than to be prone to uneven dependency on some subset of views. Second, each of them uses *forward* KL divergence $D_{KL}[p_\theta\left(z|\vec{o}\right) \| r_\psi^v\left(z|o_v\right)]$ to calibrate each encoder $r_\psi^v\left(z|o_v\right)$ to the joint encoder $p_\theta\left(z|\vec{o}\right)$, encouraging $r_\psi^v\left(z|o_v\right)$ to cover all the supports or modes of $p_\theta\left(z|\vec{o}\right)$. As a consequence, one can extract the representation z even when some views are missing in the observation. This property is critically important in (P)MVRL, where one needs to infer the complete representation from the partially available views without being overly confident on any unobserved factors. We remark that mmJSD [33] adopts *reverse* KL divergence, which is not ideal as we later demonstrate in the experiments.

Finally, although Eq. (7) has aforementioned desirable properties for MVRL, it is prone to overfitting when the size of training data is limited. This is because r_ψ^v can optimize the conditional VIB by

simply memorizing instead of learning to infer the representation of $p_\theta(z|\vec{o})$. In order to prevent overfitting, we found that VIB in Eq. (5) is an effective regularization as it favors the minimal sufficient encoding of the representation, which will be demonstrated in Section 4.2.1. Therefore, we formulate our objective function as a convex combination of Eq. (5) and Eq. (7) so that we regularize the training via VIB (see Section A.3 in the supplementary file for full derivation):

$$TC_\theta(\vec{O}; Z) \geq \frac{V-\alpha}{V} \sum_{v=1}^{V} \left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z|\vec{o})p_D(\vec{o})} \left[\ln q_\phi^v\left(o_v|z\right) \right] \right]$$
$$- \frac{\alpha}{V} \sum_{v=1}^{V} \underbrace{\mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[p_\theta(z|\vec{o}) \| r_\psi^v(z|o_v) \right] \right]}_{\text{Conditional VIB}} - (1-\alpha) \underbrace{\mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[p_\theta(z|\vec{o}) \| r(z) \right] \right]}_{\text{VIB}}, \quad (8)$$

where α is the hyperparmeter that trades off learning minimal sufficient representation in favor of calibrating r_ψ^v. For simplicity, we model the encoder, decoder, and approximate marginal distributions using the parameterized Gaussians with the diagonal covariance matrix, i.e. $r_\psi^v\left(z|o_v\right) = N\left(\mu_v, \sigma_v^2 I\right)$, $q_\phi^v\left(o_v|z\right) = N\left(\hat{\mu}_v, I\right)$, and $r\left(z\right) = N(0, \mathrm{I})$, respectively.

While the view-specific encoders r_ψ^v allow us to extract the representation from any available view individually, combining any subset of these representations (fusion) still remains as a problem. In addition, the use of joint-view encoder $p(z|\vec{o})$ makes our method limited to complete observations $\vec{o}$ during training, which needs to be addressed for PMVRL. In the next section, we show that these issues can be effectively resolved by a simple model design for $p_\theta\left(z|\vec{o}\right)$.

2.4 Models for Joint Representation Encoder

We review the models adopted by prior methods that make the joint representation encoder $p_\theta\left(z|\vec{o}\right)$ amenable to missing views and discuss their strengths and weaknesses.

PoE Product-of-Experts (PoE) [17] combines multiple probability distributions by their product. MVAE [46] models the joint representation encoder as a PoE, treating view-specific encoders as experts. The PoE can produce a sharper distribution as we increase the number of input views, thus an effective method for aggregating information across any subset of view-specific encoders. Assuming each of view-specific encoders as Gaussian distributions such that $r_\psi^v\left(z|o_v\right) = N\left(\mu_v, \sigma_v^2 I\right)$, the PoE joint encoder is obtained with computation linearly scales to the number of views by

$$p_\theta\left(z|\vec{o}\right) \triangleq N\left(\mu_p, \sigma_p^2 \mathbf{I}\right), \quad \text{where} \quad \mu_p \triangleq \frac{\sum_{v=1}^{V} \mu_v/\sigma_v^2}{\sum_{v=1}^{V} 1/\sigma_v^2} \quad \text{and} \quad \sigma_p^2 \triangleq \frac{1}{\sum_{v=1}^{V} 1/\sigma_v^2}. \quad (9)$$

Eq. (9) is also the formula of Inverse-Variance Weighted (IVW) method [6, 7], a classical method in statistics for aggregating multiple random variables, such as sensor fusion.

Unfortunately, a naive application of the PoE to the ELBO formulation (Eq.(5)) may fail to optimize the individual encoders, which is important in learning the balanced representation. MVAE [46], as an example, randomly samples subsets of views among 2^V combinations and jointly optimizes their ELBOs in order to ensure that all the view-specific encoders are optimized under PoE. However, such treatment may result in a precision miscalibration of view-specific encoders [30].

MoE The Mixture-of-Expert (MoE) takes an arithmetic mean of probability distributions, which is computationally scalable to the number of views as well. MMVAE [30] and mmJSD [33] adopt MoE of $r_\psi^v(z|o_v)$ as the model for the joint representation encoder. In MMVAE, the MoE is trained by pair-wise optimization in such a way that the latent representation from a view-specific encoder can reconstruct the observation in other views as well as its own view. However, this does not necessarily imply that the latent representation successfully aggregates the information across views. mmJSD addresses this issue by adopting a common learnable prior across views.

We remark that, in Eq. (8), modeling $p_\theta\left(z|\vec{o}\right)$ as MoE of $r_\psi^v(z|o_v)$ and setting $\alpha = 0$ yields the ELBO objective version of MMVAE. Assuming the same model for $p_\theta\left(z|\vec{o}\right)$ and $\alpha = 0$, and modeling $r(z)$ as the PoE of $r_\psi^v(z|o_v)$ yields the objective function of mmJSD. For tractable optimization of the KL term involving MoE, mmJSD derives a lower bound on the ELBO by decomposing it into multiple KL terms. This bound can be also obtained from Eq. (8) by setting $\alpha = 1$ and using the *reverse* KL for the conditional VIB terms. However, we empirically show that minimization of the reverse KL is not helpful with learning complete representation in Section 4.

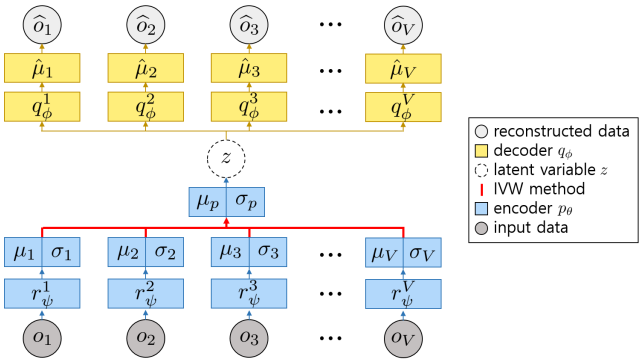


Figure 1: The architecture of Multi-View Total Correlation Auto-Encoder.

MoPoE The Mixture-of-Product-of-Experts (MoPoE) is a mixture of 2^V combination of PoE experts, which is used as the model for joint encoder in MoPoE-VAE [34]. Since the MoPoE joint encoder takes into account all possible combinations of views, it naturally learns to aggregate information across any given views while optimizing every view-specific encoder. Similar to mmJSD, MoPoE-VAE derives a lower bound on the objective to decompose the KL term into 2^V KL terms with analytic solutions. However, this would render the method intractable for tasks with many views.

Multi-View Total Correlation Auto-Encoder Since our model uses conditional VIBs that explicitly calibrate all the representations encoded by view-specific encoders, we can safely choose PoE as the model for the joint representation encoder without suffering from the precision miscalibration. We call our resulting model the Multi-View Total Correlation Auto-Encoder (MVTCAE), depicted in Figure 1. Thanks to PoE joint representation encoder, MVTCAE linearly scales to the number of input views. Furthermore, when training with partial observations, MVTCAE simply treats the covariance matrix of $q_\phi^v(o_v|z)$ to be $\infty\mathbf{I}$ for any missing observation $o_v \notin \tilde{o}$, so that it does not contribute to the reconstruction loss for missing views, similar to [50]. As a result, MVTCAE can be naturally extended to PMVRL. We show that MVTCAE successfully associates views in Section 4.

3 Related Work

Information-Theoretic Representation Learning Information Bottleneck (IB) [37] was introduced as a regularization method to obtain minimal sufficient encoding by constraining the amount of information captured by the latent variable from the observed variable. Deep Variational IB (VIB) [2] extended IB by parameterizing it with a neural network, which results in a simple yet effective method to achieve a representation that generalizes well. Furthermore, using VIB in unsupervised learning has been revealed to have close relationships among VIB, VAE [22] and β-VAE [16]. A number of follow-up works [27, 31] propose encouraging the encoders to learn representation invariant to any attribute given in advance. Similarly, a modified version of VIB was introduced [11] for learning a view-invariant representation across two views. While IB and VIB are concerned with computing MI, two similar but distinct generalizations of MI have been applied to learning disentangled representation, which are TC [42] and Interaction Information (II) [4, 25, 36]. The TC quantifies the dependency among all dimensions of the single latent variable, which motivated many works that learn disentangled representation in a single view [5, 10, 12, 20, 21]. II was used for disentangling shared representation from view-specific representation in cross-views [19].

Multi-View Representation Learning Canonical Correlation Analysis (CCA) [18] and its variants [1, 3, 41] are classical approaches for unsupervised cross-view representation learning. CCA projects two different views into one common latent space in a way that those two views are maximally correlated in the latent space. KCCA [1] uses kernels and DCCA [3] uses neural networks to learn the common representation. Similarly, DCCAE [41] trains autoencoders to obtain common representations. More recently, a number of notable MVRL methods have been proposed to support more than 2 views. DMF-MVC [53] extracts a common representation of multiple views through deep matrix factorization. MDcR [48] maps each view to a lower-dimensional space and applies the kernel matching to regularize the dependence across multiple views. ITML [9] learns a Mahalanobis distance function by Bregman optimization, whereas LMNN [43] learns a Mahalanobis distance metric to optimize the k-nearest neighbors classifier using labeled data. CPM-Nets [50] gives a formal definition of complete representation in PMVRL and proposes to learn it without encoders.

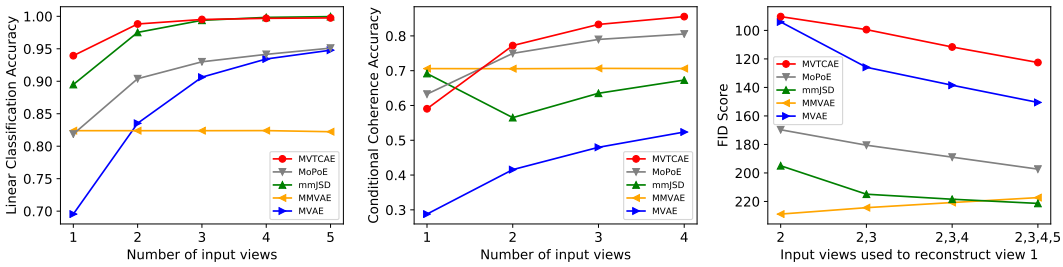


Figure 2: Performance evaluation on the Poly MNIST dataset.

Multi-View Generative Models (VAEs) MVAE [46] and its variants [30, 33, 34] are multi-view generative models that learn shared representation by maximizing the log-likelihood of joint views via latent variables. We compared these methods to ours in Section 2.2 and 2.4.

4 Experiments

4.1 Multi-View Representation Learning

To verify that our method successfully learns complete representation capturing both common factors and view-specific factors, we evaluate our method on the following two datasets used as evaluation benchmarks in the MVRL literature.

4.1.1 Multi-View Classification / Translation on PolyMNIST

We employ PolyMNIST dataset [34] composed of tuples with 5 different MNIST images, which have different backgrounds and writing style but share the same digit label. The background of each view is randomly cropped from one image which is not used by other views. Thus, the digit identity is the common factor of variation while the background and writing style are view-specific factors. There are 60K tuples of training samples and 10K of test samples. Although the digit ID should be observable in any view, it is hard to identify in some images depending on the background and the writing style. Therefore, aggregating information across views is essential for predicting the label.

Evaluation protocol To evaluate the learned representation, we follow the protocol in [34]. Specifically, after training all the models in an unsupervised manner using complete observations, we evaluate the learned representation with three different metrics, which are linear classification accuracy and conditional coherence accuracy. We also evaluate the quality of conditional generation with FID score [15]. We compare our method with various state-of-the-art multi-view generative models which are MVAE [46], MMVAE [30], mmJSD [33], and MoPoE-VAE [34].

To apply our method to classification, we fix the encoders and train a linear classifier to predict labels using the joint representation extracted from $p_\theta(z|\vec{o})$ feeding complete observations in training set. We then classify the representations of all subsets and compute the average classification accuracy over all subsets with the same subset size.

To measure the conditional coherence accuracy, we extract the representation of every subset of views using p_θ and generate views that are absent in the subset using q_ϕ. Those generated views are fed into the pretrained CNN-based classifier and see if the prediction from the classifier matches the label of the given subset. The results are averaged over all subsets with the same size.

Finally, we evaluate the sample quality of the first view images generated from different combination of input views ({2}, {2,3}, {2,3,4}, {2,3,4,5}) in terms of FID score. Since FID compares statistics of two sets (one is the set of samples generated by the models and the other is the first view images in the training set in our context), it takes into account not only the quality of generated samples but also the diversity of them. Thus, unlike two previous evaluation metrics, marking a lower (better) FID score requires the model to learn the representation that captures not only the common factors of variation but also view-specific factors to express diversity within the view.

Results The left plot in Figure 2 shows the result of linear classification. Using the PoE as a joint representation encoder same as MVAE, our method clearly outperforms all the baseline methods in the classification task, reaching 94% and 99% accuracy even when only one view and two views are given respectively. Comparing to MVAE, this result implies that conditional VIBs in our method calibrate every view-specific encoders according to the information shared across views so that each of them successfully captures the digit identity, resolving the issue of unbalanced representation.

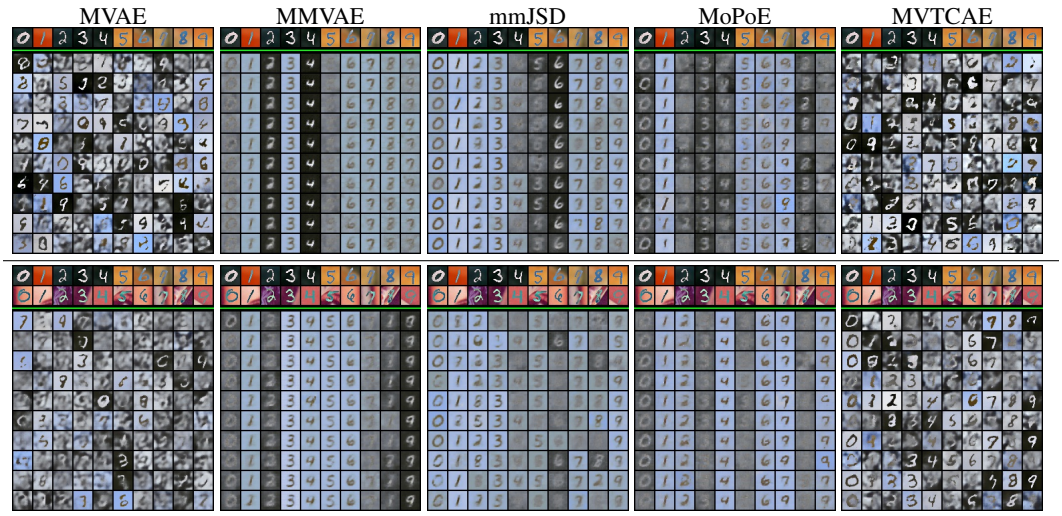


Figure 3: Conditionally generated images of the view 1 given images from the view 2 (top row) and images from the views 2 and 3 (bottom row).

On the other hand, while all the methods show monotonic improvement on its performance as the number of given views increases, the accuracy of MMVAE does not show any noticeable change. It is remarkable that our method even outperforms mmJSD and MoPoE-VAE with simpler aggregation model for the joint encoder, although the two prior methods use both PoE and MoE.

The middle plot in Figure 2 summarizes the result of conditional coherence. Our method outperforms all the baseline methods except when only one view is given. MMVAE and mmJSD fail to leverage additional input views, showing the performance staying flat or even degrading, due to the incapability of aggregating information inherent in MoE as we discussed in Section 2.4. In contrast, our method monotonically improves the coherence accuracy with more input views, implying that the view-specific encoders are well calibrated by conditional VIBs so that the joint representation aggregated by PoE produces accurate encoding of the digit identity.

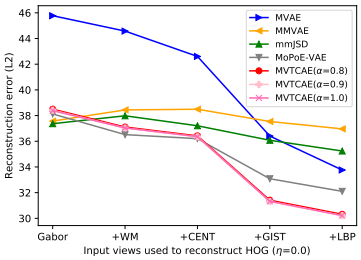
The right plot in Figure 2 shows the result of FID scores. Our method achieves the best performance in any combination of input views, which indicates that our method generates more diverse samples in better quality. Figure 3 presents qualitative results of conditional generation from baseline methods and ours. In each row, images above the green line are input observations (views {2} and {2,3} for top and bottom rows respectively) in the test set, whereas images below the line are generated images in view 1. The result shows that our method is successful in cross-view association, especially being much better than any comparing methods at (1) improving the identification of the shared factors using any additional views and (2) expressing the view-specific factors in the target view. Providing more views even improves the results of our method, which can be found in Section B.2.1. Compared to ours, MMVAE, mmJSD, and MoPoE hardly express view-specific diversity in the target view. We hypothesize that their joint representation encoders such as MoE or MoPoE are not sharp enough to discover all the view-specific factors of variation correctly. In contrast, although MVAE uses PoE joint encoder, it poorly preserves the shared factors due to the miscalibration of view-specific encoders. Considering that our method also uses PoE encoder, the performance gap between ours and MVAE shows the effectiveness of conditional VIBs.

4.1.2 Multi-View Translation on Caltech-101

We also evaluate our method on the multi-view dataset used in [24] where six visual features are extracted from images in Caltech-101 dataset. In this dataset, each image is associated with six features, which are of Gabor filter [29], Wavelet Moments (WM) [26], CENTRIST [45], Histogram of Oriented Gradients (HOG) [8], GIST [29], and Local Binary Pattern (LBP) [28]. We treat each feature as a view, varying from 40 up to 1984 dimensions. We remark that each feature can be considered as a lossy compression of the original image, where the extracted information from one view may not be necessarily inferrable for other views, thus making it nontrivial to learn the correlation across views.

Evaluation protocol Given the partial observations $\tilde{o}$ at test time, we extract the joint representation z by Eq. (9), and reconstruct the missing view $o_v \notin \tilde{o}$ using the output of the decoder $q_\theta^v(o_v|z)$

(a) Result of training with complete observations



(b) Result of training with partial observations

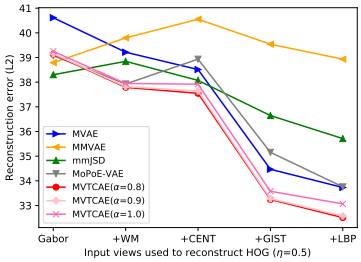


Figure 4: The multi-view translation performance of training with the complete and partial observations, measuring the reconstruction error of the HOG feature by incrementally adding features.

dedicated to view v. We then evaluate the L2 reconstruction error. We investigate the impact of missing views on the completeness of the representation. We train our model with complete observations, and apply the model to reconstruct the HOG feature[2] missing in the test time using partial observations $\tilde{o} \in \vec{o}_{\setminus \text{HOG}}$. The performance is averaged over the results of 10 independent runs. We compare to the same baseline methods employed in Section 4.1.

Results Figure 4a shows the impact of missing views at test time. We observe that our method effectively decreases the reconstruction error as the input views accumulates in the observation, only with slightly higher error in reconstruction with one view. Interestingly, we notice that the reconstruction error reduces significantly at some point (*i.e.*, when adding GIST feature to the observation). It implies that some views contain more information than the others, and our method is able to learn to utilize this property. On the other hand, MoPoE-VAE is less effective in utilizing GIST or LBP features, while mmJSD and MMVAE hardly show monotonic performance improvement. Remarkably, MVAE performs worst, although the joint encoder model is same as ours, i.e. PoE. As an ablation study, we explore the impact of weighting parameter α, which trades off between VIB and conditional VIB (Eq. (8)) in Section B.1.2 in the supplementary material.

4.2 Partial Multi-View Representation Learning

To verify that our method can be effectively generalized to the partial observation setting, we conduct experiments on two tasks: multi-view *translation* and *classification*. In both tasks, we aim to show that the representation learned by our method is complete enough to infer the missing views given the incomplete observations (translation task) and useful for downstream tasks (classification task). To simulate the partial observations in both tasks, we follow the protocol of Zhang et al. [50] and generate random view-missing patterns with missing rate $\eta = \sum_{v=1}^{V} U_v/(V \times S)$ $(0 \leq \eta < 1)$, where S is the number of entire samples and U_v is the number of samples missing in the v-th view.

4.2.1 Partial Multi-View Translation on Caltech-101

In this experiment, we investigate if our model can infer complete representations from partial observations. To quantify the completeness of the representation, we employ the multi-view translation task. The goal is to reconstruct a missing view using the representation extracted from other view(s). In this case, the reconstruction error serves as a direct measurement of the completeness.

Evaluation protocol Following the same procedures of Section 4.1.2, we repeat the same experiments but using the model trained with incomplete observation ($\eta = 0.5$). This experiment can demonstrate the robustness of our model to missing views in both inference and training time.

Results Figure 4b summarizes the experimental results using the incomplete training data ($\eta = 0.5$). We observe that our method exhibits similar trends with complete observation training data, exhibited by monotonic performance improvement in the number of available views, while other methods get negatively affected by additional views when trained with incomplete data (MMVAE, mmJSD, MoPoE-VAE). Interestingly, we observe that MVAE performs better when trained with incomplete data. This is because the missing views in the training data serve as a regularization similar to Dropout [32] or sub-sampled training paradigm [46]. Although MVAE shows monotonic improvement as ours, its performance is turned out to be not comparable to ours.

We also note that our method achieves the best performance with $\alpha = 0.8$ when trained with incomplete data, while the best is achieved at $\alpha = 0.9$ when trained with complete data. We suspect

[2]We choose the HOG since it has the largest dimension thus the reconstruction is most nontrivial.

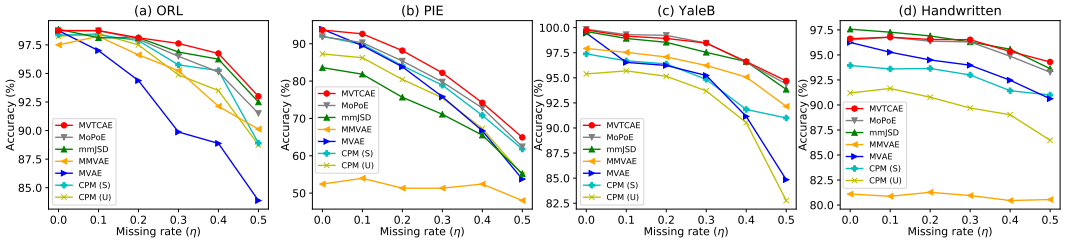


Figure 5: The partial multi-view classification performance under various view missing rate.

that this is because optimizing conditional VIB with partial observation becomes more difficult than with complete observation, and the model tends to solve it via memorization. In this case, the VIB can be useful to improve the generalization. To summarize, our method achieves the best performance for intermediate $0 < \alpha < 1$, which implies that simultaneous optimization of VIB and conditional VIBs in Eq. (8) is effective to solve MVRL and PMVRL. Qualitative results can be found in Section B.2.2 in the supplementary material.

4.2.2 Partial Multi-View Classification on Six Datasets [50]

The previous section suggests that our method is able to learn to calibrate representations across views, even when the training data is composed of incomplete observations. In this section, we evaluate the effectiveness of our approach as an unsupervised pre-training of the multi-view representation, investigating if the learned representation is useful for downstream tasks such as classification.

Datasets We evaluate our method on six feature-based image classification datasets used in multi-view learning [50], which are **ORL**,**PIE**, **YaleB**, **CUB**, **Animal** and **Handwritten**. Each dataset is associated with 2 and up to 6 visual features. Similar to the previous section, we treat each feature as a view of data. For all datasets, we follow the same preprocessing and training/test splits used in [50]. See Section D in the supplementary material for a comprehensive description of the datasets.

Evaluation protocol We follow [50] to evaluate our model trained with partial observations with various missing rates $\eta = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Similar to the experiment in Section 4.1, we first train our model in an unsupervised manner using the observable views. Then we fix the encoders, and train the classifier with labels using the learned representation. To isolate the impact of the additional classifier, we employ the simplest classifier, i.e. logistic regression with the learned representation as input. We report the performance by averaging the results from 10 independent runs.

Result Figure 5 summarizes the results under varying η on datasets ORL, PIE, YaleB, and Handwritten, which are with at least three views and thus in our primary interest. Due to the space limit, we present the results on CUB and Animal in Section B.1.3 in the supplementary material. In addition to the baseline methods compared in previous sections, we include two strong baselines, CPM-Nets(S) and CPM-Nets(U), which address the PMVRL with and without label information in the representation learning stage, respectively. Compared to the supervised baseline (CPM-Nets(S)), our method is clearly outperforming in all datasets, even though our model is trained in a purely unsupervised manner. Compared to the unsupervised baseline methods (CPM-Nets(U), MVAE, MMVAE, mmJSD, and MoPoE), our method achieves noticeable improvements, especially when the missing rate is reasonably high ($0.3 \leq \eta \leq 0.5$). It shows that our method is robust in learning the cross-view correlation under partial observations, and the learned representation is informative enough to be useful in downstream tasks, even though we do not use label information in the data or adopt a sophisticated aggregation model in the joint encoder.

5 Conclusion

We presented an information theoretic model for unsupervised partial multi-view representation learning. Based on Total Correlation (TC), we derived a novel variational lower bound that allows us to train the model that encodes complete latent representation from partial-view observations. Strictly trained in an unsupervised manner, we also demonstrated that the learned representation is highly effective in downstream tasks, such as multi-view classification and multi-view translation. Although we demonstrated that our method can even learn from partial multi-view data, it still has room for improvement such as learning from unaligned view data, and investigating more sophisticated representation aggregation models for the joint encoder, which we leave as future work.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation (NRF) of Korea (NRF-2019R1A2C1087634 red and NRF-2021R1C1C1012540), the Ministry of Science and Information communication Technology (MSIT) of Korea (IITP No. 2019-0-00075, IITP No. 2020-0-00940, IITP No. 2017-0-01779 (XAI), IITP No. 2021-0-00537, and IITP No. 2021-0-02068), the ETRI (Contract No. 21ZS1100), and Samsung Electronics.

References

[1] S. Akaho. A kernel method for canonical correlation analysis. *International Meeting of Psychometric Society*, 2001.

[2] A. Alemi, I. Fischer, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[3] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255. PMLR, 2013.

[4] A. Bell. The co-information lattice, 921–926. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Source Separation (ICA2003), Nara, Japan*, 2003.

[5] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

[6] W. G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1): 101–129, 1954.

[7] W. G. Cochran and S. P. Carroll. A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics*, 9(4):447–459, 1953.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, 2007.

[10] B. Esmaeili, H. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J.-W. van de Meent. Structured disentangled representations. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2525–2534. PMLR, 16–18 Apr 2019.

[11] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

[12] S. Gao, R. Brekelmans, G. Ver Steeg, and A. Galstyan. Auto-encoding total correlation explanation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1157–1166. PMLR, 2019.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[14] Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=OOsR8BzCnl5`.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[16] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2(5):6, 2017.

[17] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[18] H. Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

[19] H. Hwang, G.-H. Kim, S. Hong, and K.-E. Kim. Variational interaction information maximization for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[20] Y. Jeong and H. O. Song. Learning discrete and continuous factors of data via alternating disentanglement. In *International Conference on Machine Learning (ICML)*, 2019.

[21] H. Kim and A. Mnih. Disentangling by factorising. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[22] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[23] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[24] Y. Li, F. Nie, H. Huang, and J. Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[25] W. McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

[26] H. A. Moghaddam, T. T. Khajoie, A. H. Rouhi, and M. S. Tarzjan. Wavelet correlogram: a new approach for image indexing and retrieval. *Pattern Recognition*, 38(12):2506–2518, 2005.

[27] D. Moyer, S. Gao, R. Brekelmans, A. Galstyan, and G. Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pages 9084–9093, 2018.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[29] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[30] Y. Shi, N. Siddharth, B. Paige, and P. H. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 2019.

[31] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon. Learning controllable fair representations. *international conference on artificial intelligence and statistics*, 2018.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 (56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[33] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in Neural Information Processing Systems*, 2020.

[34] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized multimodal elbo. *International Conference on Learning Representations*, 2021.

[35] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. Cr-gan: Learning complete representations for multi-view generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 942–948. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/131. URL `https://doi.org/10.24963/ijcai.2018/131`.

[36] H. K. Ting. On the amount of information. *Theory of Probability & Its Applications*, 7(4): 439–447, 1962.

[37] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[38] G. Ver Steeg and A. Galstyan. Discovering structure in high-dimensional data through correlation explanation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 577–585. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/4f6ffe13a5d75b2d6a3923922b3922e5-Paper.pdf`.

[39] G. Ver Steeg and A. Galstyan. Maximally informative hierarchical representations of high-dimensional data. In *Artificial Intelligence and Statistics*, pages 1004–1012. PMLR, 2015.

[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[41] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015.

[42] S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[43] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[44] M. White, Y. Yu, X. Zhang, and D. Schuurmans. Convex multi-view subspace learning. In *Nips*, pages 1682–1690. Lake Tahoe, Nevada, 2012.

[45] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1489–1501, 2010.

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

[47] Y. Yuan, G. Xun, K. Jia, and A. Zhang. A multi-view deep learning framework for eeg seizure detection. *IEEE journal of biomedical and health informatics*, 23(1):83–94, 2018.

[48] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao. Flexible multi-view dimensionality co-reduction. *IEEE Transactions on Image Processing*, 26(2):648–659, 2016.

[49] C. Zhang, E. Adeli, T. Zhou, X. Chen, and D. Shen. Multi-layer multi-view classification for alzheimer's disease diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[50] C. Zhang, Z. Han, y. cui, H. Fu, J. T. Zhou, and Q. Hu. Cpm-nets: Cross partial multi-view networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 559–569. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/11b9842e0a271ff252c1903e7132cd68-Paper.pdf`.

[51] H. Zhang, T. S. Huang, N. M. Nasrabadi, and Y. Zhang. Heterogeneous multi-metric learning for multi-sensor fusion. In *14th International Conference on Information Fusion*, pages 1–8. IEEE, 2011.

[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[53] H. Zhao, Z. Ding, and Y. Fu. Multi-view clustering via deep matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Please read abstract and introduction
   (b) Did you describe the limitations of your work? [Yes] Please read conclusion
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Due to the space limit, we discuss it in the supplementary material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] we will submit the code as the supplementary file.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] we specify them in the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We ran 5$\sim$10 runs with different random seeds. Related information can be found in the supplementary file.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] included the supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In appropriate places
   (b) Did you mention the license of the assets? [N/A] Not applicable.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All the datasets we used are public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] None of datasets contain identifiable content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Not applicable.

Supplementary Material

Contents

A Theoretical Results

We begin with deriving Eq. (4) in detail since following subsections are based on it.

$$
\begin{aligned}
TC_\theta(\vec{O};Z) &= TC(\vec{O}) - TC_\theta(\vec{O}|Z) \\
&= D_{KL}\left(p_D\left(\vec{o}\right) \| \prod_{i=1}^{V} p_D\left(o_v\right)\right) - \mathbb{E}_{p_\theta(z)}\left[D_{KL}\left(p_\theta\left(\vec{o}|z\right) \| \prod_{i=1}^{V} p_\theta\left(o_v|z\right)\right)\right] \\
&= \sum_{v=1}^{V} H\left(O_v\right) - H(\vec{O}) - \sum_{v=1}^{V} H_\theta(O_v|Z) + H_\theta(\vec{O}|Z) \\
&= \sum_{v=1}^{V} H\left(O_v\right) - \sum_{v=1}^{V} H_\theta\left(O_v|Z\right) - H(\vec{O}) + H_\theta(\vec{O}|Z) \\
&= \sum_{v=1}^{V} I_\theta\left(O_v;Z\right) - I_\theta(\vec{O};Z), \qquad (10)
\end{aligned}
$$

where $p_\theta\left(z\right) = \int p_\theta\left(z|\vec{o}\right) p_D\left(\vec{o}\right) d\vec{o}$, $p_\theta\left(\vec{o}|z\right) = \frac{p_\theta(z|\vec{o}) p_D(\vec{o})}{p_\theta(z)}$, and $p_\theta\left(o_v|z\right) = \int p_\theta\left(\vec{o}|z\right) d\vec{o}_{\setminus v}$ are distributions involved with intractable integration w.r.t. the unknown density $p_D(\vec{o})$.
Thus, we derive three different variational lower bounds on $TC_\theta(\vec{O};Z)$ introduced in Section 2.1 and Section 2.3 below.

A.1 Lower Bound that introduces VIB (Eq. (5))

$$
\begin{aligned}
&TC_\theta(\vec{O};Z) = \sum_{v=1}^{V} I_\theta\left(O_v;Z\right) - I_\theta(\vec{O};Z) \\
&= \sum_{v=1}^{V}\left[\mathbb{E}_{p_\theta(z,o_v)}\left[\ln \frac{p_\theta\left(o_v|z\right)}{p_D(o_v)}\right]\right] - \mathbb{E}_{p_\theta(z,\vec{o})}\left[\ln \frac{p_\theta(z|\vec{o})}{p_\theta(z)}\right] \\
&= \sum_{v=1}^{V}\left[\mathbb{E}_{p_\theta(z,o_v)}\left[\ln \frac{p_\theta\left(o_v|z\right)}{p_D(o_v)} \cdot \frac{q_\phi^v\left(o_v|z\right)}{q_\phi^v\left(o_v|z\right)}\right]\right] - \mathbb{E}_{p_\theta(z,\vec{o})}\left[\ln \frac{p_\theta(z|\vec{o})}{p_\theta(z)} \cdot \frac{r(z)}{r(z)}\right] \\
&= \sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z,o_v)}\left[\ln q_\phi\left(o_v|z\right)\right]\right] - \mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o}) \| r(z)\right]\right] \\
&\quad + \sum_{v=1}^{V}\left[\mathbb{E}_{p_\theta(z)}\left[D_{KL}\left[p_\theta\left(o_v|z\right) \| q_\phi\left(o_v|z\right)\right]\right]\right] + D_{KL}\left[p_\theta(z) \| r(z)\right] \qquad (11) \\
&\geq \sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z,o_v)}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o}) \| r(z)\right]\right] \\
&= \sum_{v=1}^{V}\left[H\left(O_v\right) + \int\left(\int p_\theta(z|\vec{o}) p_D(\vec{o}) d\vec{o}_{\setminus v}\right) \ln q_\phi^v\left(o_v|z\right) do_v dz\right] - \mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o}) \| r(z)\right]\right] \\
&= \sum_{v=1}^{V}\left[H\left(O_v\right) + \int p_\theta(z|\vec{o}) p_D(\vec{o}) \ln q_\phi^v\left(o_v|z\right) d\vec{o} dz\right] - \mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o}) \| r(z)\right]\right] \\
&= \sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z|\vec{o}) p_D(\vec{o})}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o}) \| r(z)\right]\right], \qquad (12)
\end{aligned}
$$

where Eq. (11) is the gap between $TC_\theta(\vec{O};Z)$ and Eq. (12) (or Eq. (5)). Since $TC_\theta(\vec{O};Z)$ is upper bounded by $TC(\vec{O})$ which is a constant, maximization of Eq. (12) not only maximizes the original objective $TC_\theta(\vec{O};Z)$ but also minimizes Eq. (11), the gap between $TC_\theta(\vec{O};Z)$ and Eq. (12). This results in fitting $q_\phi^v\left(o_v|z\right) \approx p_\theta\left(o_v|z\right)$ and $r(z) \approx p_\theta(z)$[3].

[3] In practice, we fix $r\left(z\right) = N(0, \mathrm{I})$ for simplicity.

A.2 Lower Bound that introduces Conditional VIBs (Eq. (7))

$$
\begin{aligned}
TC_\theta(\vec{O};Z) &= \sum_{v=1}^{V} I_\theta\left(O_v;Z\right) - I_\theta(\vec{O};Z) \\
&= \sum_{v=1}^{V}\left[\frac{V-1}{V} I_\theta\left(O_v;Z\right) + \frac{1}{V} I_\theta\left(O_v;Z\right) - \frac{1}{V} I_\theta(\vec{O};Z)\right] \\
&= \frac{1}{V}\sum_{v=1}^{V}\left[(V-1)\, I_\theta\left(O_v;Z\right) - I_\theta(\vec{O}_{\setminus v};Z \mid O_v)\right] \qquad (13)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{V}\sum_{v=1}^{V}\left[(V-1)\,\mathbb{E}_{p_\theta(z,o_v)}\left[\ln\frac{p_\theta\left(o_v|z\right)}{p_D(o_v)}\right] - \mathbb{E}_{p_\theta(z,\vec{o})}\left[\ln\frac{p_\theta(z|\vec{o})}{p_\theta(z|o_v)}\right]\right] \\
&= \frac{1}{V}\sum_{v=1}^{V}\left[(V-1)\,\mathbb{E}_{p_\theta(z,o_v)}\left[\ln\frac{p_\theta\left(o_v|z\right)}{p_D(o_v)}\cdot\frac{q_\phi^v\left(o_v|z\right)}{q_\phi^v\left(o_v|z\right)}\right] - \mathbb{E}_{p_\theta(z,\vec{o})}\left[\ln\frac{p_\theta(z|\vec{o})}{p_\theta(z|o_v)}\cdot\frac{r_\psi^v(z|o_v)}{r_\psi^v(z|o_v)}\right]\right] \\
&= \frac{V-1}{V}\sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z,o_v)}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \frac{1}{V}\sum_{v=1}^{V}\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right] \\
&\quad + \frac{V-1}{V}\sum_{v=1}^{V}\left[\mathbb{E}_{p_\theta(z)}\left[D_{KL}\left[p_\theta\left(o_v|z\right)\|q_\phi\left(o_v|z\right)\right]\right]\right] + \frac{1}{V}\sum_{v=1}^{V}\left[\mathbb{E}_{p_D(o_v)}\left[D_{KL}\left[p_\theta(z|o_v)\|r_\psi^v(z|o_v)\right]\right]\right] \qquad (14)
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{V-1}{V}\sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z,o_v)}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] - \frac{1}{V}\sum_{v=1}^{V}\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right] \\
&= \frac{V-1}{V}\sum_{v=1}^{V}\left[H\left(O_v\right) + \int\left(\int p_\theta(z|\vec{o})p_D(\vec{o})d\vec{o}_{\setminus v}\right)\ln q_\phi^v\left(o_v|z\right)do_v dz\right] \\
&\quad - \frac{1}{V}\sum_{v=1}^{V}\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right] \\
&= \frac{V-1}{V}\sum_{v=1}^{V}\left[H\left(O_v\right) + \int p_\theta(z|\vec{o})p_D(\vec{o})\ln q_\phi^v\left(o_v|z\right)d\vec{o}dz\right] \\
&\quad - \frac{1}{V}\sum_{v=1}^{V}\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right] \\
&= \frac{V-1}{V}\sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z|\vec{o})p_D(\vec{o})}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] \\
&\quad - \frac{1}{V}\sum_{v=1}^{V}\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right], \qquad (15)
\end{aligned}
$$

where $p_\theta\left(z|o_v\right) = \int p_\theta(z|\vec{o})p_D(\vec{o})d\vec{o}_{\setminus v}$ is a distribution that requires intractable integration w.r.t. the unknown density $p_D(\vec{o})$. Note that the equality in Eq. (13) holds due to the chain rule for MI (see Section A.5). Similar to Eq. (12), maximization of Eq. (15) minimizes Eq. (14), the gap between Eq. (13) and Eq. (15). Thus, our variational optimization scheme fits not only $q_\phi^v\left(o_v|z\right) \approx p_\theta\left(o_v|z\right)$ but also $r_\psi^v(z|o_v) \approx p_\theta(z|o_v)$.

A.3 Convex Combination (Eq. (8))

$$
\begin{aligned}
TC_\theta(\vec{O};Z) &= (1-\alpha)\left(\sum_{v=1}^{V} I_\theta\left(O_v;Z\right) - I_\theta(\vec{O};Z)\right) \\
&\quad + \alpha\left(\frac{1}{V}\sum_{v=1}^{V}\left[(V-1)\, I_\theta\left(O_v;Z\right) - I_\theta(\vec{O}_{\setminus v};Z \mid O_v)\right]\right) \\
&= \frac{V(1-\alpha)+\alpha(V-1)}{V}\sum_{v=1}^{V} I_\theta\left(O_v;Z\right) - \frac{\alpha}{V}\sum_{v=1}^{V} I_\theta(\vec{O}_{\setminus v};Z \mid O_v) - (1-\alpha)\, I_\theta(\vec{O};Z) \\
&\geq \frac{V-\alpha}{V}\sum_{v=1}^{V}\left[H\left(O_v\right) + \mathbb{E}_{p_\theta(z|\vec{o})p_D(\vec{o})}\left[\ln q_\phi^v\left(o_v|z\right)\right]\right] \\
&\quad - \frac{\alpha}{V}\sum_{v=1}^{V}\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r_\psi^v(z|o_v)\right]\right] - (1-\alpha)\,\mathbb{E}_{p_D(\vec{o})}\left[D_{KL}\left[p_\theta(z|\vec{o})\|r(z)\right]\right], \quad (16)
\end{aligned}
$$

where $0 \leq \alpha \leq 1$.

A.4 Interaction Information and its Equivalence to $TC_\theta(\vec{O};Z)$ when $V=2$

When there are 2 views, Interaction Information (II) among O_1, O_2, and Z is defined as follows:

$$
\begin{aligned}
I_\theta\left(O_1;O_2;Z\right) &= I_\theta\left(O_1;Z\right) - I_\theta\left(O_1;Z \mid O_2\right) \\
&= I_\theta\left(O_2;Z\right) - I_\theta\left(O_2;Z \mid O_1\right) \\
&= I\left(O_1;O_2\right) - I_\theta\left(O_1;O_2 \mid Z\right)
\end{aligned}
$$

Applying the chain rule of MI (see Section A.5) to the first equality,
we can easily show the equivalence between $I_\theta\left(O_1;O_2;Z\right)$ and $TC_\theta(\vec{O};Z)$:

$$
\begin{aligned}
I_\theta\left(O_1;O_2;Z\right) &= I_\theta\left(O_1;Z\right) - I_\theta\left(O_1;Z \mid O_2\right) \\
&= I_\theta\left(O_1;Z\right) - \left(-I_\theta\left(O_2;Z\right) + I_\theta\left(O_1,O_2;Z\right)\right) = TC_\theta(\vec{O};Z) \quad (17)
\end{aligned}
$$

A.5 Chain Rule for Mutual Information

$$
\begin{aligned}
I_\theta\left(O_v;Z\right) - I_\theta(\vec{O};Z) &= \mathbb{E}_{p_\theta(z,o_v)}\left[\ln\frac{p_\theta\left(z|o_v\right)}{p_\theta(z)}\right] - \mathbb{E}_{p_\theta(z,\vec{o})}\left[\ln\frac{p_\theta(z|\vec{o})}{p_\theta(z)}\right] \\
&= \int\left(\int p_\theta(z|\vec{o})p_D(\vec{o})d\vec{o}_{\setminus v}\right)\ln\frac{p_\theta\left(z|o_v\right)}{p_\theta(z)}do_v dz \\
&\quad - \int p_\theta(z|\vec{o})p_D\left(\vec{o}\right)\ln\frac{p_\theta(z|\vec{o})}{p_\theta(z)}d\vec{o}dz \\
&= \int p_\theta(z|\vec{o})p_D\left(\vec{o}\right)\left(\ln\frac{p_\theta\left(z|o_v\right)}{p_\theta(z)} - \ln\frac{p_\theta(z|\vec{o})}{p_\theta(z)}\right)d\vec{o}dz \\
&= -\int p_\theta(z|\vec{o})p_D\left(\vec{o}\right)\ln\frac{p_\theta(z|\vec{o})}{p_\theta(z|o_v)}d\vec{o}dz = -\mathbb{E}_{p_\theta(z,\vec{o})}\left[\ln\frac{p_\theta(z|\vec{o}_{\setminus v},o_v)}{p_\theta(z|o_v)}\right] \\
&= -I_\theta(\vec{O}_{\setminus v};Z \mid O_v) \quad (18)
\end{aligned}
$$

A.6 Connection to Multi-View Information Bottleneck (MIB)

MIB [11] is proposed for learning view-invariant representation between two views. Although one can try to apply MIB to MVRL with more than 2 views by treating it as $\binom{V}{2}$ pair-wise representation learning, it combinatorially scales to the number of given views, making it infeasible to run with many views.

Interestingly, we observe that designing $p_\theta(z, \vec{o})$ as MoE of $r^v_\psi(z|o_v)$ relates conditional VIBs in Eq. (7) to the regularization terms used in MIB for discarding any view-specific information[4].

$$\sum_{v=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[p_\theta(z|\vec{o}) \| r^v_\psi(z|o_v) \right] \right] \tag{19}$$

$$= \sum_{v=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[p_\theta(z|\vec{o}) \| r^v_\psi(z|o_v) \right] \right] + \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[r^i_\psi(z|o_i) \| p_\theta(z|\vec{o}) \right] \right]$$
$$- \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[r^i_\psi(z|o_i) \| p_\theta(z|\vec{o}) \right] \right] \tag{20}$$

$$= \sum_{v=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[\int \left(\frac{1}{V} \sum_{i=1}^{V} r^i_\psi(z|o_i) \right) \ln \frac{p_\theta(z|\vec{o})}{r^v_\psi(z|o_v)} dz \right] + \sum_{v=1}^{V} \sum_{i=1}^{V} \left[\frac{1}{V} D_{KL} \left[r^i_\psi(z|o_i) \| p_\theta(z|\vec{o}) \right] \right]$$
$$- \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[r^i_\psi(z|o_i) \| p_\theta(z|\vec{o}) \right] \right]$$

$$= \sum_{v=1}^{V} \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[\frac{1}{V} \int r^i_\psi(z|o_i) \ln \frac{p_\theta(z|\vec{o})}{r^v_\psi(z|o_v)} dz + \frac{1}{V} \int r^i_\psi(z|o_i) \ln \frac{r^i_\psi(z|o_i)}{p_\theta(z|\vec{o})} dz \right]$$
$$- \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[r^i_\psi(z|o_i) \| p_\theta(z|\vec{o}) \right] \right]$$

$$= \sum_{v=1}^{V} \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[\frac{1}{V} \int r^i_\psi(z|o_i) \ln \frac{r^i_\psi(z|o_i)}{r^v_\psi(z|o_v)} dz \right] - \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[D_{KL} \left[r^i_\psi(z|o_i) \| p_\theta(z|\vec{o}) \right] \right]$$

$$\leq \sum_{v=1}^{V} \sum_{i=1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[\frac{1}{V} D_{KL} \left[r^i_\psi(z|o_i) \| r^v_\psi(z|o_v) \right] \right]$$

$$= \sum_{v=1}^{V-1} \sum_{i=v+1}^{V} \mathbb{E}_{p_D(\vec{o})} \left[\frac{2}{V} D_{SKL} \left[r^i_\psi(z|o_i) \| r^v_\psi(z|o_v) \right] \right], \tag{21}$$

where $D_{SKL} \left[r^i_\psi(z|o_i) \| r^v_\psi(z|o_v) \right] = \frac{1}{2} D_{KL} \left[r^i_\psi(z|o_i) \| r^v_\psi(z|o_v) \right] + \frac{1}{2} D_{KL} \left[r^v_\psi(z|o_v) \| r^i_\psi(z|o_i) \right]$. Remarkably, each of D_{SKL} terms in Eq. (21) is a regularization term used in MIB to discard any information not shared by two views, which encourages each of view-specific encoder to learn view-invariant representation only. Although Eq. (19) is a lower bound on Eq. (21), the gap Eq. (20) between Eq. (21) and Eq. (19) clearly shows that the optimal solutions of Eq. (21) and Eq. (19) have to be equal to:

$$r^1_\psi(z|o_1) = r^2_\psi(z|o_2) = ... = r^V_\psi(z|o_v)$$

Bearing in mind that our goal is to learn complete representation instead of view-invariant representation, Eq. (21) shows that MoE is not a good choice for the conditional VIBs.

B Comprehensive Experimental Results

In this section, we provide all the evaluations including any quantitative and qualitative results we possibly missed in the main text due to the space limit.

[4] view-specific information is called superfluous information in MIB [11].

B.1 Quantitative Results

We explicitly specify all the quantitative results we presented in Section 4. Some of additional results are included to make our overall experiments more comprehensive.

B.1.1 Results in multi-view classification / translation on PolyMNIST

Table 1 and 2 specify the numbers used to plot Figure 2.

Models	**Total input views (Linear Classification)**					**Total input views (Coherence)**			
(α)	1	2	3	4	5	1	2	3	4
MVAE	0.70	0.84	0.91	0.93	0.95	0.29	0.42	0.48	0.52
MMVAE	0.82	0.82	0.82	0.82	0.82	**0.71**	0.71	0.71	0.71
mmJSD	0.89	0.98	0.99	**1.0**	**1.0**	0.69	0.57	0.64	0.67
MoPoE	0.82	0.90	0.93	0.94	0.95	0.63	0.75	0.79	0.81
Ours (5/6)	**0.94**	**0.99**	**1.0**	**1.0**	**1.0**	0.59	**0.77**	**0.83**	**0.86**

Table 1: Comparisons on linaer classification and coherence accuracy. All the results are averaged over 5 independent runs. We omit the standard error which are less than 0.01 in most cases.

Models	**Input view(s)**			
(α)	View 2	Views 2,3	Views 2,3,4	Views 2,3,4,5
MVAE	94.06 ± 5.20	125.87 ± 5.97	138.46 ± 6.29	150.53 ± 6.58
MMVAE	228.86 ± 13.68	224.37 ± 14.28	220.76 ± 13.42	217.31 ± 11.89
mmJSD	194.96 ± 2.75	214.91 ± 3.69	218.44 ± 3.52	221.37 ± 3.64
MoPoE	169.70 ± 2.60	180.53 ± 2.11	188.92 ± 3.00	197.33 ± 3.56
Ours (5/6)	**90.32** ± 1.72	**99.44** ± 1.63	**111.64** ± 1.53	**122.51** ± 1.56

Table 2: Comparisons on FID scores averaged over 5 independent runs.

B.1.2 Ablation study in partial multi-view translation

To investigate the effect of α, we compare the performance of our method applying various settings of $\alpha = \{0.0, 0.7, 0.8, 0.9, 1.0\}$. The result is summarized in Table 3 below. In both cases of using complete ($\eta = 0$) and incomplete (0.5) observations, $\alpha \geq 0.7$ yields significant performance improvement comparing to $\alpha = 0.0$. It clearly shows that the conditional VIB ($\alpha = 1$) is very effective on calibrating the representation across views compared to VIB counterpart without cross-view calibration ($\alpha = 0$). Setting $\alpha = 0.9$ and $\alpha = 0.8$ shows the best performance in each case of $\eta = 0$ and $\eta = 0.5$ respectively, which implies that regularization using VIB can be also effective when observations are sparse.

Models	**Views used to reconstruct HOG** ($\eta = 0.0$)					**Views used to reconstruct HOG** ($\eta = 0.5$)				
(α)	Gabor	+WM	+CENT.	+GIST	+LBP	Gabor	+WM	+CENT.	+GIST	+LBP
MVAE	45.78	44.58	42.61	36.41	33.76	40.61	39.21	38.51	34.47	33.73
MMVAE	37.57	38.44	38.49	37.54	36.96	38.79	39.79	40.55	39.54	38.93
mmJSD	**37.37**	37.98	37.21	36.08	35.24	**38.30**	38.84	38.07	36.65	35.71
MoPoE	38.13	**36.52**	**36.20**	33.08	32.10	39.08	37.93	38.93	35.16	33.76
Ours (0.0)	51.27	45.68	42.95	36.05	33.41	40.51	39.22	38.52	34.49	33.79
Ours (0.7)	38.56	37.17	36.48	31.53	30.43	39.16	37.85	37.56	33.29	32.58
Ours (0.8)	38.50	37.11	36.42	31.41	30.32	39.13	**37.79**	**37.55**	**33.24**	**32.51**
Ours (0.9)	38.42	37.03	36.34	**31.31**	**30.21**	39.15	37.82	37.64	33.27	32.57
Ours (1.0)	38.38	37.04	36.36	31.33	30.22	39.25	37.95	37.92	33.58	33.07

Table 3: The translation performance trained with the complete dataset ($\eta = 0$, from the second to the sixth columns) and incomplete dataset ($\eta = 0.5$, from the seventh to the last columns). We measure the reconstruction error of the HOG by incrementally adding features, accumulated from the feature in the second and seventh columns. The results are the average performance of 10 independent runs. We omit the standard errors which are around 0.06 in most cases.

B.1.3 Results in partial multi-view classification on 6 datasets including CUB and Animal

In addition to ORL, PIE, YaleB, and Handwritten, Figure 6 shows the partial multi-view classification results on CUB and Animal which are datatsets composed of 2 views. The result shows that our method achieves performance competitive to the strong baseline methods on those 2-view datasets.

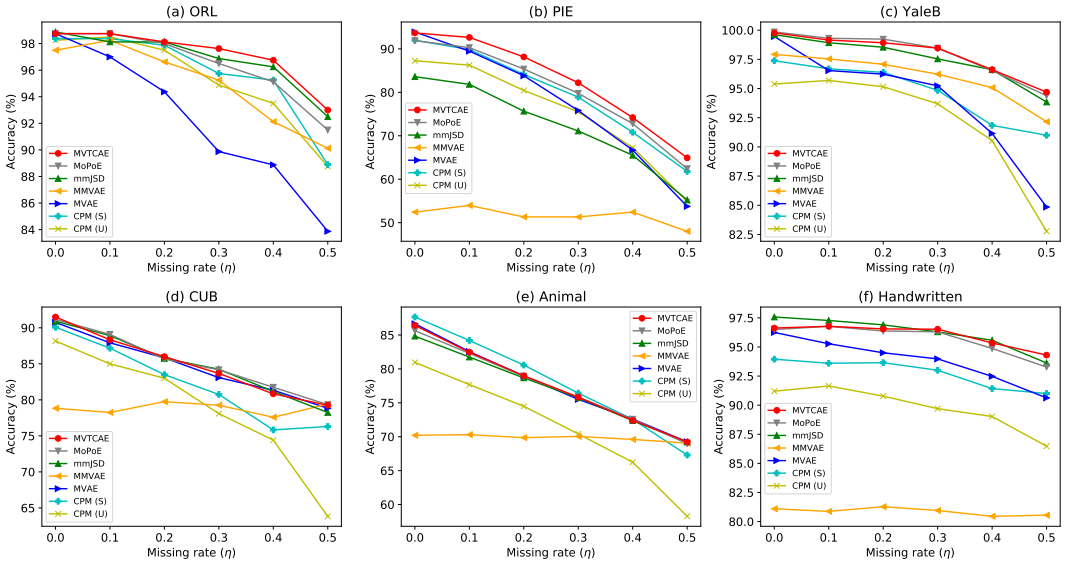


Figure 6: Classification performance on 6 datasets under various view missing rate.

B.1.4 Results in partial multi-view classification with additional baseline methods

We compare our method using incomplete dataset ($\eta = 0.5$) with additional state-of-the-art MVRL methods such as CCA [18], KCCA [1], DCCA [3], DCCAE [41], DMF-MVC [53], MDcR [48], ITML [9], LMNN [43], and CPM-Nets [50], as well as a naive baseline of concatenating all views (FeatCon). Table 4 summarizes the result. Please note that α is chosen in Figure 5 and Figure 6 according to the result in Table 4 and fixed across all settings of $\eta = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for each dataset.

		Datasets (# of views)					
Models	S/U	ORL (3)	PIE (3)	YaleB (3)	CUB (2)	Animal (2)	HW (6)
CCA	U	38.1	37.4	66.2	57.1	24.1	55.3
KCCA	U	42.4	33.8	67.8	57.6	23.4	56.7
DCCA	U	38.3	35.8	67.1	40.8	9.4	54.4
DCCAE	U	35.6	36.3	67.6	47.5	10.4	54.4
DMF	U	60.1	34.3	57.5	30.3	47.0	55.8
MDcR	U	65.1	23.1	58.0	70.0	61.7	55.4
FeatCon	U	66.3	36.3	59.8	70.8	61.9	87.1
ITML	S	76.3	36.6	81.2	70.2	56.0	73.1
LMNN	S	70.0	56.4	76.6	73.8	59.6	86.1
CPM (w/ class)	S	88.9	61.8	91.0	76.3	67.3	91.0
CPM(w/o class)	U	88.8 $\pm$ 0.9	54.9 $\pm$ 1.0	82.8 $\pm$ 1.3	63.8 $\pm$ 1.3	58.3 $\pm$ 0.2	86.5 $\pm$ 0.9
MVAE	U	83.9 $\pm$ 1.5	53.8 $\pm$ 0.9	84.8 $\pm$ 0.6	78.8 $\pm$ 0.8	**69.2** $\pm$ 0.3	90.6 $\pm$ 0.5
MMVAE	U	90.1 $\pm$ 0.9	48.0 $\pm$ 0.7	92.2 $\pm$ 0.8	**79.4** $\pm$ 0.9	69.0 $\pm$ 0.4	80.6 $\pm$ 0.4
mmJSD	U	92.5 $\pm$ 1.03	55.2 $\pm$ 0.9	93.8 $\pm$ 0.7	78.3 $\pm$ 1.0	**69.2** $\pm$ 0.4	93.6 $\pm$ 0.3
MoPoE	U	91.5 $\pm$ 0.6	62.4 $\pm$ 1.0	94.4 $\pm$ 0.5	79.3 $\pm$ 0.6	69.0 $\pm$ 0.3	93.3 $\pm$ 0.2
Ours ($\alpha = 0.8$)	U	92.6 $\pm$ 0.7	61.7 $\pm$ 0.9	94.2 $\pm$ 0.5	79.2 $\pm$ 0.6	**69.2** $\pm$ 0.3	93.2 $\pm$ 0.3
Ours ($\alpha = 0.9$)	U	**93.0** $\pm$ 0.7	**64.9** $\pm$ 0.9	94.0 $\pm$ 0.6	79.0 $\pm$ 0.6	**69.2** $\pm$ 0.3	93.6 $\pm$ 0.3
Ours ($\alpha = 1.0$)	U	92.8 $\pm$ 0.9	60.5 $\pm$ 1.1	**94.7** $\pm$ 0.7	78.6 $\pm$ 0.7	**69.2** $\pm$ 0.3	**94.3** $\pm$ 0.4

Table 4: Comparisons on classification accuracy (%) with missing rate $\eta = 0.5$. Each dataset is specified with the number of views inside of the parentheses in the second row. S stands for supervised learning and U stands for unsupervised learning in the second column. All the results are averaged over 10 independent runs.

The results when observations are complete ($\eta = 0$) are presented in Table 5 below.

Models	**Datasets** (# of views)					
	ORL (3)	PIE (3)	YaleB (3)	CUB (2)	Animal (2)	HW (6)
CPM (w/ class)	98.4 ± 0.4	92.0 ± 0.7	97.4 ± 0.5	90.1 ± 0.7	**87.7** ± 0.1	94.0 ± 0.4
CPM (w/o class)	98.3 ± 0.3	87.3 ± 1.7	95.4 ± 0.7	88.2 ± 1.1	81.0 ± 0.2	91.2 ± 0.4
MVAE	98.8 ± 0.3	93.9 ± 0.3	99.5 ± 0.3	90.8 ± 0.6	86.7 ± 0.3	96.3 ± 0.3
MMVAE	97.5 ± 0.4	52.4 ± 1.0	97.9 ± 0.4	78.8 ± 1.2	70.2 ± 0.4	81.1 ± 0.6
mmJSD	98.9 ± 0.2	83.6 ± 0.6	99.6 ± 0.1	90.9 ± 0.8	84.8 ± 0.4	**97.6** ± 0.2
MoPoE	98.8 ± 0.3	91.9 ± 0.4	**99.8** ± 0.1	91.2 ± 0.7	85.6 ± 0.4	96.5 ± 0.3
Ours ($\alpha = 0.8$)	98.9 ± 0.3	**94.9** ± 0.6	99.7 ± 0.1	91.5 ± 0.7	86.4 ± 0.3	96.7 ± 0.3
Ours ($\alpha = 0.9$)	98.8 ± 0.3	93.7 ± 0.4	**99.8** ± 0.2	91.5 ± 0.7	86.4 ± 0.3	97.0 ± 0.3
Ours ($\alpha = 1.0$)	98.9 ± 0.3	90.1 ± 0.5	**99.8** ± 0.2	**91.7** ± 0.7	86.3 ± 0.3	96.6 ± 0.3

Table 5: Comparisons on classification accuracy (%) with missing rate $\eta = 0$. All the results are averaged over 10 independent runs.

B.2 Qualitative Results

We provide comprehensive qualitative results on PolyMNIST along with some examples in the dataset.

B.2.1 Translation results on PolyMNIST [34] dataset

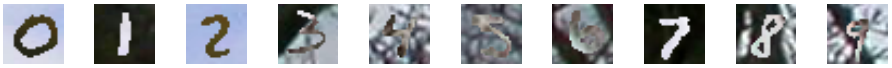
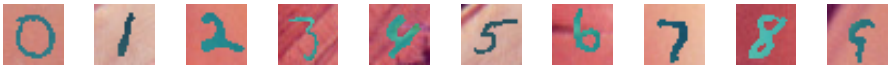
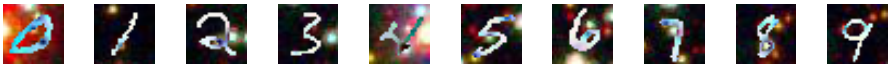
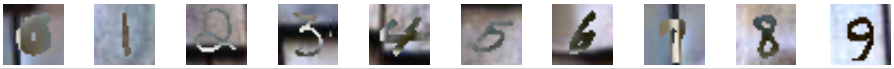
	Digit Identities									
View	0	1	2	3	4	5	6	7	8	9
1										
2										
3										
4										
5										

Table 6: Examples of samples in PolyMNIST.

Table 6 shows examples of PolyMNIST dataset, where each row is $0 \sim 9$ images in each view. Note that many images in view 1 are remarkably blurry as follows:

Figures 8, 9, 10, and 11 summarize the qualitative results of conditional generations of each model, where images above the green line are input observations from different view(s) ({2}, {2,3}, {2,3,4}, {2,3,4,5}) in the test set and images below the line are images in view 1 generated by models. Unlike our method, all the baseline methods expose at least one of following three issues:

Mode collapse in MMVAE, mmJSD, MoPoE Generated images in view 1 fail to show diversities in styles of backgrounds and digits, which can be observed by comparing rows in any figures.

Entangled representations in MMVAE, mmJSD, MoPoE Although styles of backgrounds and digits are view-specific factors of variation, comparison among the same columns in Figures 8, 9, 10, 11 shows that those styles get affected by additional observations from new views.

Discarded shared information in MVAE Comparing images above and below the green line in every figure clearly shows the failure in generating coherent samples whose digit identities are supposed to match to conditioned images. Furthermore, it is not obvious that the coherence is improved according to the increased number of given views.

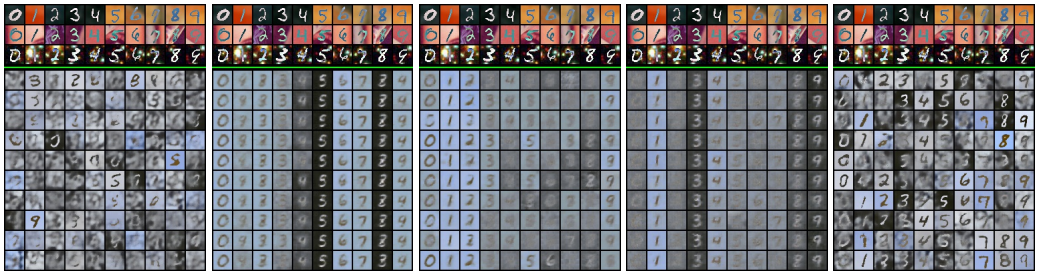
On the other hand, our method expresses view-specific style variations independent of conditioned views while showing better preservation of the digit identities as the number of given views increases.

MVAE MMVAE mmJSD MoPoE MVTCAE

Figure 8: Conditionally generated images of the view 1 given images from the view 2.

MVAE MMVAE mmJSD MoPoE MVTCAE

Figure 9: Conditionally generated images of the view 1 given images from the views 2 and 3.



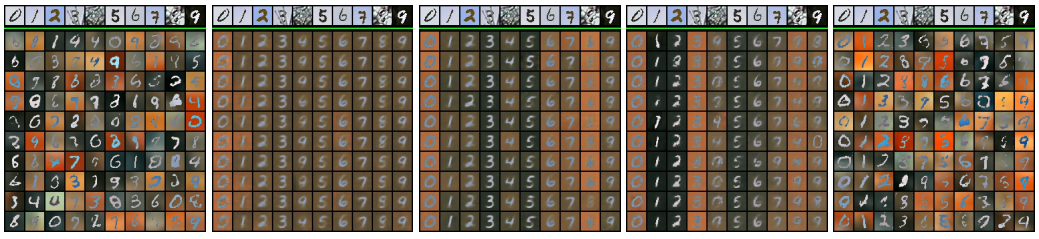
MVAE MMVAE mmJSD MoPoE MVTCAE

Figure 10: Conditionally generated images of the view 1 given images from the views 2, 3, and 4.

MVAE MMVAE mmJSD MoPoE MVTCAE

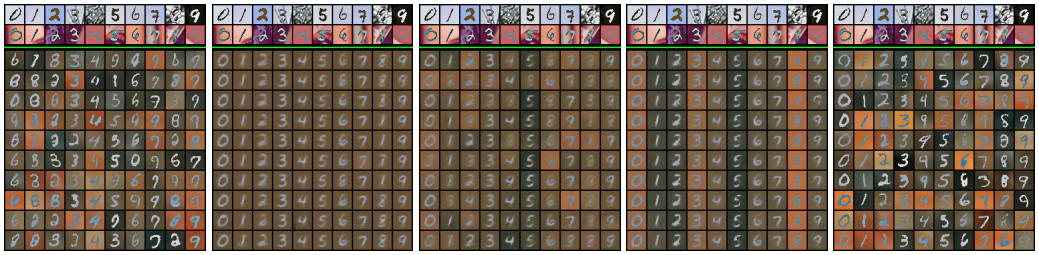
Figure 11: Conditionally generated images of the view 1 given images from the rest of views.

We present additional qualitative results in Figures 12, 13, 14, and 15 where images above the green line are conditioned observations from different view(s) ({1}, {1,3}, {1,3,4}, {1,3,4,5}) in the test set and images below the line are images in view 2 generated by models. Three issues we already identified in Figures 8, 9, 10, 11 are similarly observed.



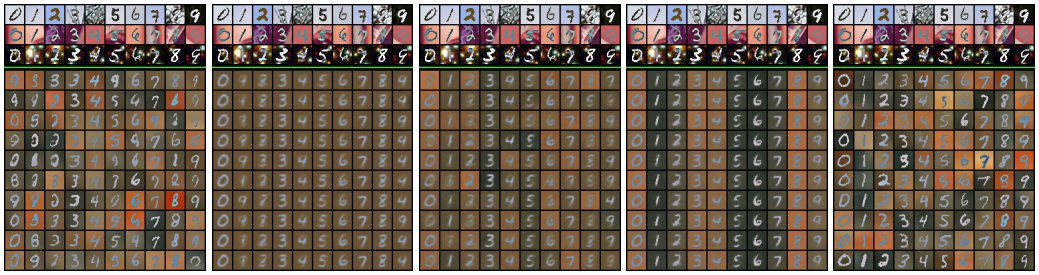
MVAE MMVAE mmJSD MoPoE MVTCAE

Figure 12: Conditionally generated images of the view 2 given images from the view 1.



MVAE MMVAE mmJSD MoPoE MVTCAE

Figure 13: Conditionally generated images of the view 2 given images from the views 1 and 3.



MVAE MMVAE mmJSD MoPoE MVTCAE

Figure 14: Conditionally generated images of the view 2 given images from the views 1, 3, and 4.

MVAE MMVAE mmJSD MoPoE MVTCAE

Figure 15: Conditionally generated images of the view 2 given images from the rest of views.

B.2.2 Translation results in Caltech-101 dataset trained with complete ($\eta = 0$) and incomplete observations ($\eta = 0.5$)

We present qualitative results in Caltech-101 dataset using complete and incomplete training data ($\eta = 0, 0.5$). In Figure 16 and Figure 17, HOG features reconstructed by our model trained with incomplete data ($\eta = 0.5$) show the comparable quality to the ones reconstructed by ours with complete data ($\eta = 0$), demonstrating the robustness of our method to partial observations. In Figure 16 and Figure 17, the labels of features are lamp, starfish, stop sign, motorbike, umbrella, scissors, airplane, butterfly, kangaroo, and watch.

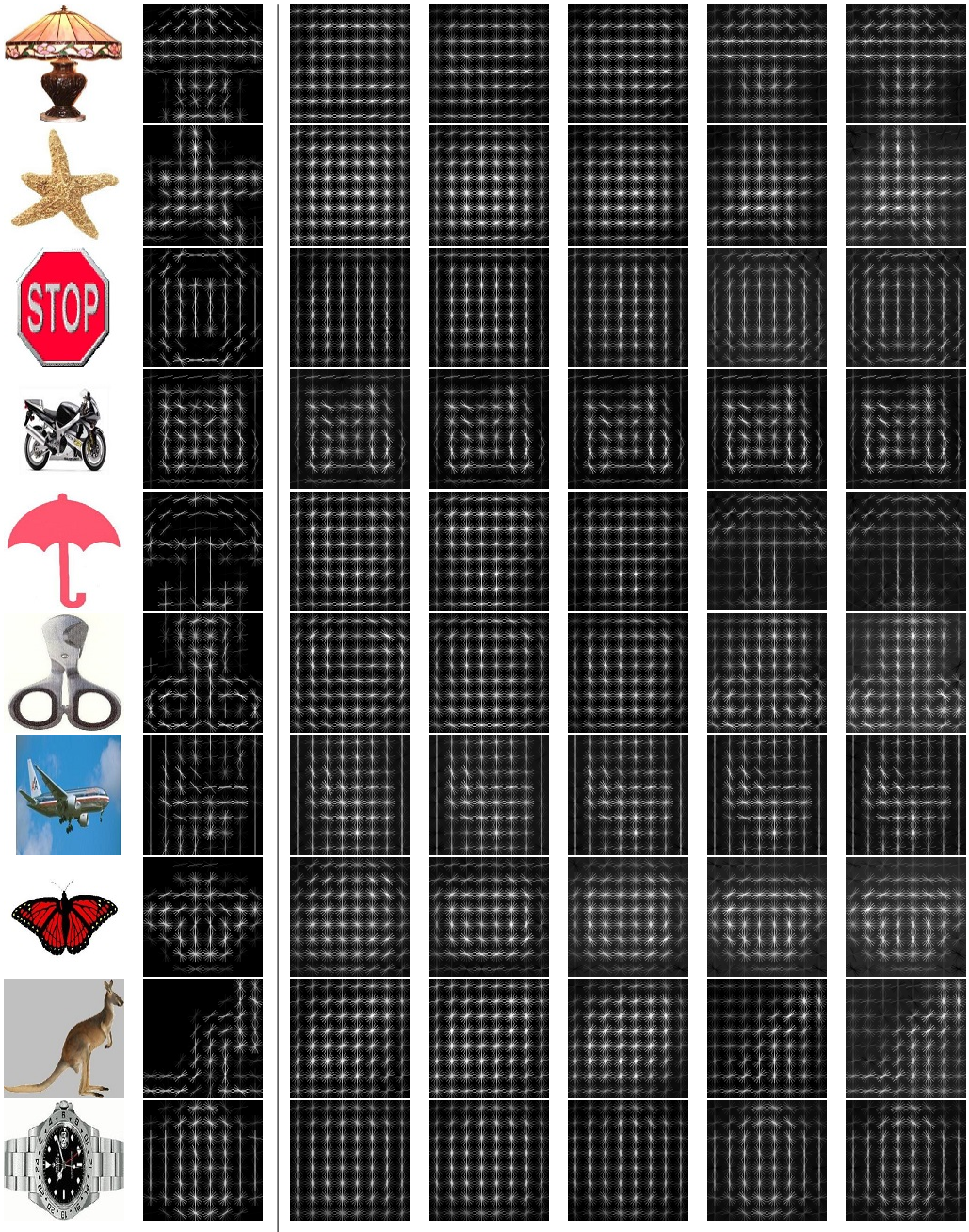


Figure 16: Qualitative results in multi-view translation using complete training data ($\eta = 0.0$). The HOG feature is reconstructed by incrementally adding features, accumulated from the left-most feature (i.e. Gabor).

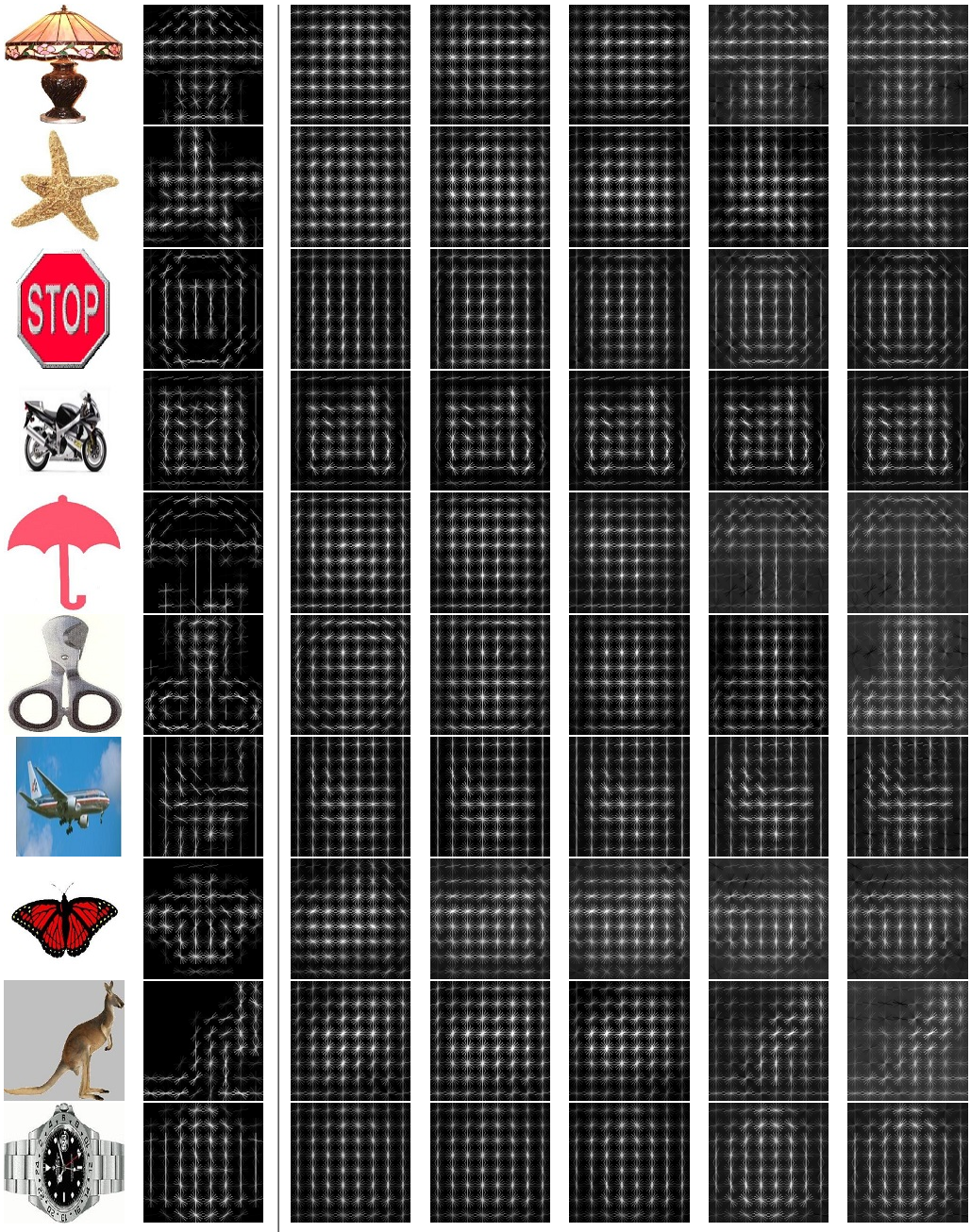


Figure 17: Qualitative results in multi-view translation using incomplete training data ($\eta = 0.5$). The HOG feature is reconstructed by incrementally adding features, accumulated from the left-most feature (i.e. Gabor).

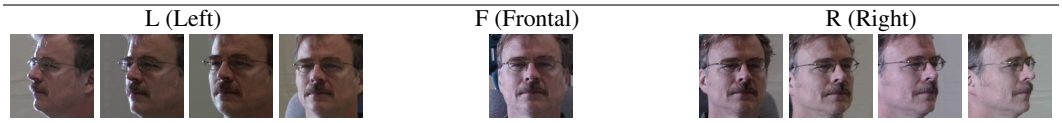
C New Experimental Results on Additional Datasets

To see in depth if our method generalizes well when views are only composed of raw observations, we conducted additional experiment with new datasets, which are Multi-PIE [13] and MNIST-SVHN [30]. Evaluating on those two datasets, we fixed $\alpha = 0.9$ for our method which was found to be reasonable in the previous experiments. Please note that **Multi-PIE** is the source from which **PIE** in Section 4.2.2 composed of 3 hand-crafted features are extracted, and thus they are different.

C.1 Additional Experimental Results on Multi-PIE in Pixels

To evaluate our method in multiple aspects, we follow the protocol similar to the one used in Section 4.1.1. Specifically, after training all the models in an unsupervised manner using complete observations, we evaluate the learned representation with 4 different metrics, which are linear classification accuracy, conditional coherence accuracy, sample generation quality, and sample diversity. We compare our method with MVAE [46], MMVAE [30], mmJSD [33], and MoPoE-VAE [34] same as Section 4.1.1. For every method, we searched KL coefficient (β) optimal across all metrics among $\{1, 2.5, 5, 10, 20\}$. Unlike ours and other baseline methods, we were not able to find the optimal β for mmJSD that makes the model work commonly well across all tasks. Thus, we report performance of mmJSD with two different settings of $\beta = 1, 20$. Other than β, we applied same hyperparameters such as epochs, dimensions of latent variable, and batch size to be 300, 128, and 16 respectively/ All the quantitative results below are averaged over 5 seeds (0$\sim$4), where as the qualitative results are from the single seed 0.

Dataset configuration Multi-PIE [13] is a dataset composed of 750K bust shot images of 337 human subjects with various facial expressions collected under the circumstance with 15 view points and 19 illumination conditions. Following [35], we extract 250 subjects with 9 poses (within $\pm 60^\circ$), 19 illuminations, and 2 facial expressions and assign the first 200 subjects to training set and the rest 50 for testing set. We group 9 poses into 3 views, where the first view is composed of images with 4 poses within -60°, the second view of images in 0°, and the third view of images in $+60^\circ$. We call first, second, and third views as L (Left), F (Frontal), and R (Right) respectively. For example:



Without applying any view-specific variation in view F, we choose variation in 19 illumination conditions and 2 facial expressions as two shared factors of variation across views while variation in 4 poses in each of views L and R is chosen as a view-specific factor. As a result, each subject owns 38 tuples of 3 images of L,F,R views sharing the illumiation conditions and facial expressions, where the image from each of views L and R is randomly chosen among 4 poses whenever the tuple is sampled.

Linear Classification To apply our method to classification, we fix the encoders and train two linear classifier to predict illumination condition and facial expression using the joint representation extracted from $p_\theta(z|\vec{o})$ feeding complete observations in training set. We count as positive prediction only the case two classifiers simultaneously yields correct predictions on both illumination condition and facial expression. We compute the average classification accuracy over all subsets with the same subset size.

Model (β)	Given 1	Given 2	Given 3 (All)
MVAE (1)	57.52 $\pm$ 0.83	63.22 $\pm$ 0.53	65.62 $\pm$ 0.65
MMVAE (1)	70.5 $\pm$ 0.33	70.61 $\pm$ 0.47	71.14 $\pm$ 0.44
mmJSD (1)	75.01 $\pm$ 0.56	77.61 $\pm$ 0.75	78.86 $\pm$ 0.74
mmJSD (20)	79.34 $\pm$ 0.66	**82.18** $\pm$ 0.76	**83.45** $\pm$ 0.93
MoPoE (1)	72.55 $\pm$ 0.89	74.03 $\pm$ 0.61	74.23 $\pm$ 0.43
MVTCAE (10)	**80.48** $\pm$ 0.67	81.87 $\pm$ 0.79	82.07 $\pm$ 0.81

Table 7: Joint classification accuracy of 19 illumination conditions and 2 facial expressions using the learned latent representation.

Table 7 summarizes the result of linear classification accuracy according to the number of input views. The result shows that mmJSD (20) and ours show the best performances whose error bars overlap, which implies that both methods can successfully extract the information shared across views.

Conditional coherence To measure the conditional coherence accuracy, we extract the representation of every subset of views using p_θ and generate views that are absent in the subset using q_ϕ. Those generated views are fed into the pretrained CNN-based classifier. We count as a correct prediction if the prediction on both the illumination condition and the facial expression from the classifier simultaneously matches two labels of the input view images. The results are averaged over all subsets with the same size.

	Target View			**# of Input Views**	
Model (β)	L	F	R	Given 1	Given 2
MVAE (1)	8.84 ± 0.69	39.7 ± 2.86	9.08 ± 0.29	18.09 ± 1.15	21.44 ± 1.15
MMVAE (1)	74.66 ± 0.68	85.1 ± 0.25	72.95 ± 0.89	77.54 ± 0.41	77.62 ± 0.44
mmJSD (1)	69.12 ± 1.97	83.45 ± 0.61	65.65 ± 2.07	73.7 ± 1.07	70.82 ± 1.03
mmJSD (20)	60.08 ± 0.47	75.22 ± 0.34	55.8 ± 1.39	68.77 ± 0.62	53.55 ± 0.55
MoPoE (1)	76.17 ± 0.32	85.37 ± 0.45	73.04 ± 0.8	77.36 ± 0.45	79.85 ± 0.52
MVTCAE (10)	**82.58** ± 0.6	**85.81** ± 0.15	**82.77** ± 0.5	**83.02** ± 0.28	**85.1** ± 0.26

Table 8: Joint coherence accuracy in the conditionally generated samples with respect to illumination conditions and facial expressions.

Table 8 summarizes the results of conditional coherence accuracy in two ways according to the target view and the number of input views. The result shows that our method outperforms all the comparing methods across all aspects. The results indicates that conditional VIBs in our method are very effective to identifying the shared factors of variation and improving preservation of them using additional input views. On the other hand, MVAE shows poor performance incomparable to any comparing methods, which implies that augmenting ELBO of each view to the ELBO of the joint views harms the preservation of shared factors of variation.

Sample quality To evaluate the sample quality of conditional generation, we reuse images generated for evaluating the conditional coherence by comparing them to the ground truth images in the target view paired with their input view images. We quantify similarities between those generated images and corresponding target images using LPIPS [52], which measures perceptual distance between two images. Considering that each of generated images in view L and R can have any of 4 different poses, we compute LPIPS distance between the generated images and each of 4 target images and count the minimum distance. The results are averaged over all subsets of input views with the same size.

	Target View			**# of Input Views**	
Model (β)	L	F	R	Given 1	Given 2
MVAE (1)	0.3262	0.1807	0.3180	0.2785	0.2679
MMVAE (1)	0.2953	0.1812	0.2931	0.2565	0.2566
mmJSD (1)	0.3207	0.1992	0.3180	0.2758	0.2863
mmJSD (20)	0.3595	0.2346	0.3564	0.3048	0.3409
MoPoE (1)	0.2868	0.1741	0.2855	0.2499	0.2466
MVTCAE (10)	**0.2202**	**0.1673**	**0.2211**	**0.2046**	**0.1995**

Table 9: LPIPS distance between generated samples and target images. Since views L and R have variation of 4 different poses as their own factors of variation, there are four candidate target images per generated sample if its target view is L or R. Thus, we count the minimum distance out of 4. Standard errors are omitted since they are negligibly small.

Table 9 summarizes the results of the sample quality evaluation in two ways according to the target view and the number of input views. The results shows that our method outperforms all the comparing methods across all aspects. Compared to MMVAE, mmJSD, and MoPoE-VAE, our method shows significant performance gap in the case target view is L or R while the gap is relatively small in the case the target view is F. This is because those methods are using MoE as their joint representation encoder that hardly expresses view-specific factors of variation, which results in generating blurry images collapsing to one pose in views L and R (see Figure 19). Although MVAE generates samples with variation in poses, those samples are not consistent to the given subject.

Sample diversity We measure the view-specific diversity in the generated samples by entropy. We extract the representation of subset of views $\{\text{L}, \text{F}, \text{R}, \text{LF}, \text{FR}\}$ using p_θ and generate 10 samples (per instance) in any of views L,R that are absent in the input subset using q_ϕ. Those 10 generated images are fed into another pretrained CNN-based classifier which predicts the pose among 4 candidates in the target view. To compute entropy, we first apply one-hot encoding to 10 predicted labels. Then we normalize those 10 encodings to make their sum to be 1 and compute entropy which stands for the diversity with respect to the pose. The results are averaged over all subsets with the same size.

	Target View		**# of Input Views**	
Model (β)	L	R	Given 1	Given 2
MVAE (1)	1.65 ± 0.01	1.59 ± 0.01	1.68 ± 0.01	1.48 ± 0.02
MMVAE (1)	0.05 ± 0.0	0.06 ± 0.01	0.05 ± 0.0	0.05 ± 0.0
mmJSD (1)	0.15 ± 0.02	0.16 ± 0.01	0.14 ± 0.01	0.19 ± 0.01
mmJSD (20)	0.42 ± 0.03	0.42 ± 0.02	0.37 ± 0.01	0.52 ± 0.01
MoPoE (1)	0.07 ± 0.01	0.07 ± 0.01	0.08 ± 0.01	0.05 ± 0.0
MVTCAE (10)	**1.70** ± 0.01	**1.66** ± 0.01	**1.72** ± 0.0	**1.6** ± 0.01

Table 10: Diversity of poses in the generated samples.

Table 10 summarizes the result of measuring diversity in the generated samples by their entropy. The result shows that our method significantly outperforms all the MoE-based methods (MMVAE, mmJSD, MoPoE-VAE) due to their issues on preserving view-specific factors as we discussed in Section 4.1.1. Our method even outperforms MVAE, which implies that conditional VIBs in our method are greatly effective to the cross-view association without introducing any side effects.

Qualitative results Lastly, Figure 19 presents 4 samples of conditionally generated samples in each of target views L and R feeding an image of the first subject in view F in the test set. The results show that our method (bottom row) generates samples in the finest quality with the best preservation of the subject's identity and the highest diverisity in pose, which is consistent with what we observed in the quantitative results above.

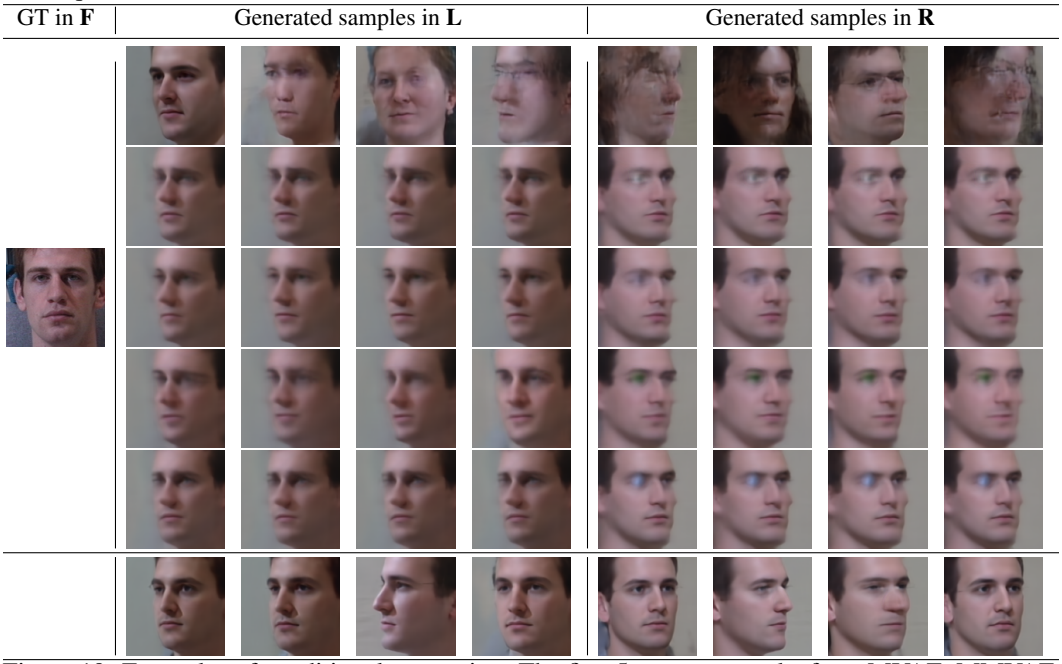


Figure 19: Examples of conditional generation. The first 5 rows are results from MVAE, MMVAE, mmJSD(1), mmJSD(20), and MoPoE. The bottom row is the result of our method.

Summary Showing state-of-the-art performance in the task of classification using the learned representation, our algorithm absolutely outperforms all the baseline methods in the translation tasks. It is remarkable that (1) even successfully preserving the information shared across views (observed in Table 7, 8), our method generate samples not only in the best quality (observed in Table 9 and Figure 19) but also in the highest diversity (observed in Table 10 and Figure 19).

C.2 Additional Experimental Results on MNIST-SVHN

The table below summarizes the result of the experiment in MNIST-SVHN (MS) dataset averaged over 5 runs with seeds $0 \sim 4$. The evaluation protocol, codebase, and hyperparameter settings follow the experiments in the MoPoE-VAE [34] on MNIST-SVHN-Text (MST) dataset, which augments to the original MS the text of digit IDs as the third modality. Please note that all the results below are from the baseline implementations in the MoPoE-VAE codebase whose VAE architectures dedicated to MNIST and SVHN are the same as the ones used by MMVAE. Lastly, we simply discarded the third modality.

Models	**Representation Classification (RC)**			**Coherent Generation (CG)**		
	M (MNIST)	S (SVHN)	MS	Joint	M→S	S→M
MVAE	87.38	58.23	87.37	42.98	56.65	35.63
MMVAE	72.83	60.89	66.89	42.45	26.76	**74.94**
mmJSD	88.58	**81.44**	93.81	12.73	22.33	65.44
MoPoE	82.48	70.62	87.96	44.95	21.23	72.38
MVTCAE	**93.48**	77.99	**94.97**	**46.71**	**81.09**	59.91

where representation classification (RC) is measured by the accuracy of the single linear classifier trained on the latent representation as input (z from M only, z from S only, and z from jointly M and S), and coherent generation (CG) is measured by the accuracy of the pretrained CNN classifier whose input is the image generated by each model (e.g. Joint is measured from MNIST and SVHN images generated from the same z sampled from the prior distribution, and M->S is measured by SVHN images generated from MNIST images).

Among 6 different evaluation results, our method outperforms baseline methods in 4 tasks (RC / MNIST and MS, CG / Joint and M->S) and performs competitive to the baselines in 1 task, RC / SVHN. Our method performs relatively poor only in CG / S->M (ranked 4th), the advantage of our method in M->S is much more noticeable. Comparing to strong baseline methods such as MMVAE, mmJSD, and MoPoE-VAE, our method shows more balanced performance in two different directions of CG, as measured in Joint, achieving the best average performance.

D Dataset Statistics

We present information on all the datasets we used in Section 4 in detail. Each feature is treated as one view in every dataset.

D.1 In Section 4.1.1

- **PolyMNIST** [34] is an image dataset composed of 5 views each of which is created by fusing each of MNIST images with a background, a 28x28x3 sized patch randomly cropped from one of five different images chosen by [34]. Each MNIST images is binarized, and colors of its background image is inverted at the locations where the digit of the MNIST is inserted. Some examples are showcased in Section B.2.1. License information of those 5 images used for backgrounds can be found in [34].

D.2 In Section 4.1.2 and 4.2.1

- **Caltech-101** [23] is a image dataset collected for object recognition task. Images in Caltech-101 are categorized as 101 different classes. From Caltech 101, six visual features are extracted and compiled as a multiview dataset by Li et al. [24], which are are 48 dimensional Gabor feature [29], 40 dimensional wavelet moments (WM), 254 dimensional CENTRIST [45] feature, 1984 dimensional HOG [8] feature, 512 dimensional GIST [29] feature, and 928 dimensional LBP [28] feature.

D.3 In Section 4.2.2

1. **ORL**[5] is a dataset composed of 400 facial images of 40 subjects. 4096 dimensional Intensity feature, 3304 dimensional LBP feature, and 6750 dimensional Gabor feature are extracted.
2. **PIE**[6] consists of 750K bust shot of 337 human subjects. A subset which contains 10 images for each of 68 people is collected, 680 images in total. We use 484 dimensional Intensity feature, 256 dimensional LBP feature, and 279 dimensional Gabor feature extracted from the subset.
3. **Yale Face Database B**[7] (YaleB) is a database which contains 5850 images of 10 subjects captured with 585 different illumination conditions (65 illumination conditions for 9 different poses). A subset which contains 650 images of 10 subjects is collected. We use 2500 dimensional Intensity feature, 3304 dimensional LBP feature, and 6750 dimensional Gabor feature extracted from the subset.
4. **CUB** [40] is a dataset consists of 11788 images of birds that belong to 200 different classes. A subset of 600 images that covers 10 categories are collected. 1024 dimensional GoogLeNet visual feature and 300 dimensional doc2vec feature are are extracted from the subset.
5. **Animal** is a dataset composed of 10158 images of animals distributed across 50 classes. Two different deep visual features are extracted, which are 4096 dimensional DECAF feature and 4096 dimensional VGG19 feature.
6. **Handwritten**[8] is a dataset that contains 2k handwritten digits of 0 to 9. Six features are generated, which are 76 dimensional Fourier coefficients of the character shapes feature, 216 dimensional profile correlations feature, 64 dimensional Karhunen-love coefficients feature, 240 dimensional (2×3) pixel averages feature, 47 dimensional Zernike moment feature, and 6 dimensional morphological feature.

Note that subsamples and features of ORL, PIE, YaleB, CUB, and Animal datasets are collected by Zhang et al. [50]. As a result, there are 3 features in ORL, PIE, YaleB and 2 features in CUB, Animal, whereas 6 features in Handwritten. Lastly, we followed the same preprocessing and training/test splits used in Zhang et al. [50] for all six datasets employed in Section 4.2.2.

[5] https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[6] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

[7] http://vision.ucsd.edu/~leekc/ExtYaleDatabase/Yale%20Face%20Database.htm

[8] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

E Implementation Details

We report implementation details in each experiment including the hyperparameters and the structures of encoders and decoders. We used two codebases, one is implemented in PyTorch by MoPoE-VAE[9] and the other is written in TensorFlow by CPM-Nets[10]. Except the joint representation encoder uniquely determined by each model, MVAE, MMVAE, mmJSD, MoPoE-VAE, and our method share the same network architectures and hyperparameter settings including the batch size, the size of encdoer/decoder and latent variables, coefficient of reconstruction (w) and KL regularizations (β) terms, and epochs. In the test phase, the representation fusion in each model is conducted using its own joint representation encoder (as identified in Section 2.4), except mmJSD which learns to fuse representations using its dynamic prior. For dimensionalities of inputs of encoders and outputs of decoders, please read Section D. Further information can be found in our official implementation[11].

E.1 In Section 4.1.1

Following MoPoE-VAE, we fixed $w = 1$ and $\beta = 2.5$ and ran for 300 epochs with 5 seeds ($0 \sim 4$). We also set network structures and dimension size of the latent variable (512) same as MoPoE-VAE. We fixed α, the only hyperparameter our method uniquely has, to be $\frac{5}{6}$ that equally weights the VIB and conditional VIBs.

E.2 In Section 4.1.2 and 4.2.1

Fixing $w = 200$ and $\beta = 1.0$, we ran each model with 10 seeds ($0 \sim 9$) for 10,000 epochs to ensure that all the methods are converged. As an ablation study, we evaluated our method with various settings of $\alpha = \{0.0, 0.7, 0.8, 0.9, 1.0\}$ as we reported in Section B.1.2, which results in $\alpha = 0.9$ and $\alpha = 0.8$ showing the best performance when $\eta = 0.0$ and 0.5 resepectively. We adopted following network architectures with 100-dimensional latent variables for all methods.

Dataset	Caltech 101
Network	Encoder $r_\psi^v(z\|o_v)$
Input	o_v
Layer 1	FC. 200. ReLU
Layer 2	$2\times$ FC. 100 $(\mu_v, \log\sigma_v^2)$
Network	Decoder $q_\phi^v(o_v\|z)$
Input	$z \sim p_\theta(z\|\vec{o})$
Layer 1	FC. 200. ReLU
Layer 2	FC. dim(o_v)

E.3 In Section 4.2.2

We ensured that structures and sizes of our decoders are same as the ones used in the official implementation of CPM-Nets. The only difference is the activation function being used. We used ReLU in the middle of two fully connected (FC) layers. We chose the structures of our view-specific encoders as the reverse of decoders, ensuring that the sizes of the latent variables we use are same as the ones used in CPM-Nets as well. We described how α is chosen in Section B.1.4. For MVAE, MMVAE, mmJSD, MoPoE, and ours, we applied the same encoder/decoder structures and ran for the same number of epochs with 10 seeds ($0 \sim 9$) per dataset to make fair comparison. We chose $w = 100$ and $\beta = 1.0$ for all datasets. Dimensions of the latent variable and the epoch per dataset are specified below.

	Datasets					
Hyperparamters	ORL	PIE	YaleB	CUB	Animal	Handwritten
Dimensions of z	256	150	128	128	512	128
Epochs	1,000	5,000	5,000	2,000	100	5,000

Table 11: Hyperparameters used in ORL, PIE, YaleB, CUB, Animal, and Handwritten datasets.

[9] `https://github.com/thomassutter/MoPoE`

[10] `https://github.com/hanmenghan/CPM_Nets`

[11] `https://github.com/gr8joo/MVTCAE`

F Computation Resources

We used 10 systems equipped with following devices.

CPU: Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz

Memory: 32 Gb.

GPU: TITAN Xp

G Societal Impact

Positively, our method could be used to reduce the number of sensors in multi-sensor system without losing sensor fusion accuracy, reducing carbon footprint and environmental waste due to redundant sensors. Negatively, we see the possibility that our method could be exploited in wrongful manner, such as Deepfake. Specifically, one might adopt our method to synthesize someone's image in the representation space and generate fake samples for fraudulent purposes. Similarly, our method can be utilized in synthesizing voice for impostors.