Distribution-Consistent Modal Recovering for Incomplete Multimodal Learning

Yuanzhi Wang, Zhen Cui*, Yong Li*
PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.
{yuanzhiwang, zhen.cui, yong.li}@njust.edu.cn

Abstract

*Recovering missing modality is popular in incomplete multimodal learning because it usually benefits downstream tasks. However, the existing methods often directly estimate missing modalities from the observed ones by deep neural networks, lacking consideration of the distribution gap between modalities, resulting in the inconsistency of distributions between the recovered and the true data. To mitigate this issue, in this work, we propose a novel recovery paradigm, Distribution-Consistent Modal Recovering (**DiCMoR**), to transfer the distributions from available modalities to missing modalities, which thus maintains the distribution consistency of recovered data. In particular, we design a class-specific flow based modality recovery method to transform cross-modal distributions on the condition of sample class, which could well predict a distribution-consistent space for missing modality by virtue of the invertibility and exact density estimation of normalizing flow. The generated data from the predicted distribution is integrated with available modalities for the task of classification. Experiments show that DiCMoR gains superior performances and is more robust than existing state-of-the-art methods under various missing patterns. Visualization results show that the distribution gaps between recovered modalities and missing modalities are mitigated. Codes are released at https://github.com/mdswyz/DiCMoR.*

1. Introduction

Multimodal machine learning dedicates to designing a strong model for understanding, reasoning, and learning by fusing multimodal data, such as language, acoustic, image, et al [15, 24]. Researchers have extensively exploited how to effectively encode the discriminative representations from different modalities [22, 33, 3, 28]. However, in real world scenarios, the well-trained model may be deployed when certain modalities are not available, e.g., language may be unavailable due to speech recognition errors; acoustic modality may be lost due to background noise or sensor sensing limitations; visual data may be unavailable due to lighting, occlusion, or social privacy security. In practice, the problem of missing modality inevitably degrades the multimodal understanding performance.

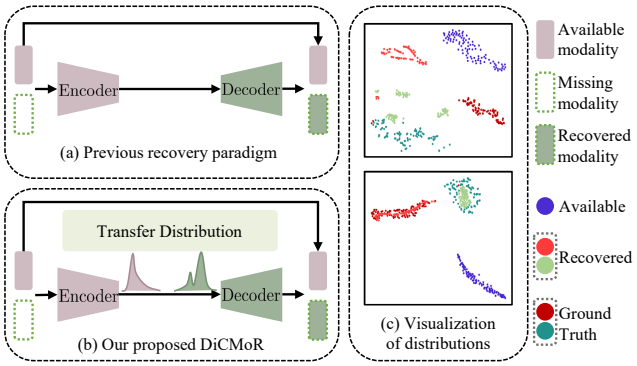

Figure 1. Two paradigms for missing modality recovery. (a) Typical paradigm usually exploits a well-crafted encoder and decoder modality recovery. (b) Our method transfers the distribution from the available to the missing modality, aiming to guarantee the distribution consistency. (c) Recovered and ground truth modality distribution visualization comparison between MCTN [20] (top) and our DiCMoR (bottom). DiCMoR exhibits a higher degree of proximity to the ground truth. More results can be found in Fig. 3.

To address this problem, a straightforward way is to conduct data recovery and then perform downstream tasks on the recovered data. As shown in Fig. 1 (a), this is a typical recovery paradigm that has been extensively studied by researchers [23, 20, 34, 16]. The core of this paradigm aims to design a well-crafted encoder and decoder, and then take as input the available modality so as to recover the missing modality. Tran *et al.* [23] proposed a cascaded residual autoencoder architecture to reconstruct the missing modalities; Lian *et al.* [16] leveraged graph neural networks to estimate the missing modalities from partially observed in-

* The corresponding authors.

put. However, this paradigm fails to fully consider the distribution gap caused the inherent modality heterogeneity, resulting in the inconsistent distribution between the *vanilla available* and the corresponding *recovered* modality.

In this work, we propose a novel framework to mitigate the above issues. As shown in Fig. 1 (b), different from the previous paradigm, we transfer the distribution from the available modalities to the missing modalities before decoding the missing data. The missing modality is then recovered under the estimated distribution. Thus, the key is to construct and learn a model that has the ability to perform transformations between cross-modal distributions.

To this end, we propose a distribution-consistent modal recovering (DiCMoR) method to complete those missing modalities for robust multimodal understanding. To transform cross-modal distributions, we introduce the modality-related flows and bridge different modalities within the embedded distribution space. To facilitate the distribution transfer, the invertible modality-specific normalizing flow is used for each modality to map the features of different modalities into the latent spaces with gaussian distributions, which reduces the distribution gap between the modalities. In the latent distribution space, the latent states of the missing modality could be sampled to feed into the inverse flow to faithfully estimate the original missing data. To increase the discriminability, the modality-related flows are built on condition of class labels to avoid the common collapse of different-class samples. In other words, we constrain the latent spaces from the same class but different modalities share the same class-specific Gaussian distribution, aiming to enhance the discriminative ability of the recovered modality. Owing to the favorable attributes inherent to flow-based models, such as their invertibility and capacity for precise density estimation, an assurance of distributional congruence emerges between the inferred data and the actual ground truth. Finally, the recovered modalities together with the available modalities could be jointly fed into multimodal fusion network for the downstream tasks. The contributions of this work are summarized as:

- We propose a novel missing modality recovery framework by transferring the distributions from the available modalities to the missing modalities, which reduces the distribution gap between the recovered data and the vanilla available data.
- We propose a cross-modal distribution transformation method by designing class-specific multimodal flows, which not only ensures the congruence of the distributions but also enhances the discriminative capacity.
- We experimentally verify the superiority of the method in various modality-missing patterns. Visualization results demonstrate that distribution gaps between recovered and missing modalities are obviously reduced.

2. Related Works

2.1. Incomplete multimodal learning

Learning from incomplete multimodal data is an essential research topic in machine learning, which allows trained models to be robust to inevitable modality-missing environments in real-world scenarios. Depending on whether to recover incomplete multimodal data, current methods can be divided into two categories: non-recovery methods [29, 11, 5] and recovery methods [19, 32, 20].

The non-recovery methods can be roughly classified into grouping strategy-based, correlation maximization-based, and knowledge distillation-based. The grouping strategy-based methods aim to split incomplete multimodal data into multiple complete subgroups and perform feature learning for each subgroup individually [29]. The correlation maximization-based methods aim to maximize the correlation between modalities and constrain different modalities to have correlated low-dimensional representations [11, 17]. The knowledge distillation-based methods learn a separate prediction model for each modality, and then distill knowledge between modalities, using the model of the partial modality for prediction in the inference phase [5, 18].

The basic principle of the recovery methods is to estimate and reconstruct the data of the missing modalities explicitly from the data of available modalities. The representative methods include zero-based recovery [19], average-based recovery [32], and deep learning-based recovery [20]. Among them, the zero and average-based recovery methods still cause a considerable gap between the recovered data and the true data because they do not utilize any supervision information. In contrast, deep learning-based methods can better estimate missing modalities by leveraging their powerful feature representation capabilities. For example, Tran *et al.* [23] proposed a cascaded residual autoencoder for the imputation of missing modalities. Pham *et al.* [20] and Zhao *et al.* [34] combined autoencoder with cycle consistency loss for modality reconstruction. Lian *et al.* [16] utilized graph neural networks to recover missing modalities to further improve the performance of downstream tasks. Our method has an essential difference from previous methods because we recover the missing modality from the perspective of data distribution, maintaining the distribution consistency and discriminability of recovered data.

2.2. Flow-based generative model

Normalizing flow is a classical generative probabilistic model, also known as a flow-based generative model, introduced by Dinh *et al.* [4] for efficient and exact density estimation. This model can be seen as a transformation of a known and tractable distribution $p_Z(\mathbf{z})$ (e.g., a Gaussian distribution) into an unknown and arbitrary distribution $p_U(\mathbf{u})$ by a sequence of invertible (i.e., bijective) and differ-

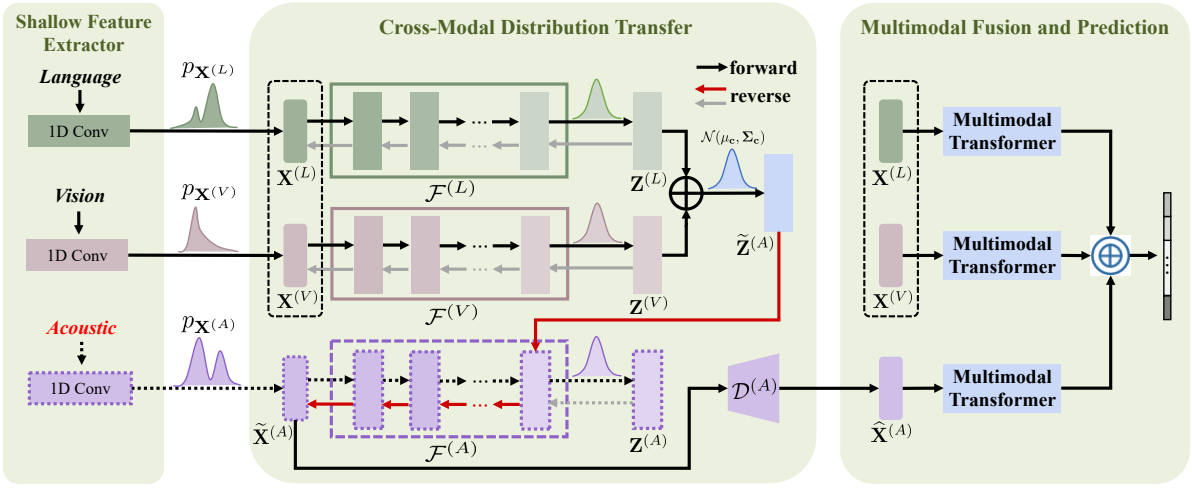

Figure 2. The framework of DiCMoR. Given the input incomplete multimodal data (here we assume acoustic modality is missing), DiCMoR encodes $\mathbf{X}^{(L)}$ and $\mathbf{X}^{(V)}$ via two exclusive shallow feature extractors (Sec. 3.3). In cross-modal distribution transfer, $\mathbf{X}^{(L)}$ and $\mathbf{X}^{(V)}$ are input to $\mathcal{F}^{(L)}$ and $\mathcal{F}^{(V)}$ to obtain latent state $\mathbf{Z}^{(L)} \sim \mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$ and $\mathbf{Z}^{(V)} \sim \mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$, respectively. $\mathbf{Z}^{(L)}$ and $\mathbf{Z}^{(V)}$ are subsequently averaged to sample and generate a latent acoustic state $\widetilde{\mathbf{Z}}^{(A)} \leftarrow (\mathbf{Z}^{(L)} + \mathbf{Z}^{(V)})/2 \sim \mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$. To mimic acoustic modality distribution, we inject $\widetilde{\mathbf{Z}}^{(A)}$ to $(\mathcal{F}^{(A)})^{-1}$ to generate $\widetilde{\mathbf{X}}^{(A)}$. Further, $\widetilde{\mathbf{X}}^{(A)}$ is fed into the acoustic-specific reconstruction module $\mathcal{D}^{(A)}$ to faithfully recover $\widehat{\mathbf{X}}^{(A)}$ (Sec. 3.4). Ultimately, the fusion of features from $\widehat{\mathbf{X}}^{(A)}$ with those of $\mathbf{X}^{(L)}$ and $\mathbf{X}^{(V)}$ is conducted for emotion prediction (Sec. 3.5).

entiable mappings, and vice versa.

The generative process is defined as $\mathbf{u} = \mathcal{G}(\mathbf{z})$, where $\mathbf{z} \sim p_Z(\mathbf{z})$ is a sample usually from standard MVG (multivariate Gaussian) distribution, $\mathbf{u} \sim p_U(\mathbf{u})$ is a true sample. As the $\mathbf{u}$ is observed during the training phase, the $\mathbf{z}$ can be obtained by $\mathbf{z} = \mathcal{F}(\mathbf{u}) = \mathcal{G}^{-1}(\mathbf{u})$, $\mathcal{F}$ is the inverse function of $\mathcal{G}$, and $\mathbf{z}$ is called a latent variable. $\mathcal{F}$ is composed of N invertible transformations: $\mathcal{F} = f_1 \circ f_2 \circ \cdots \circ f_N$, the transformation between $\mathbf{u}$ and $\mathbf{z}$ is $\mathbf{u} \overset{f_1}{\leftrightarrow} \mathbf{h}_1 \overset{f_2}{\leftrightarrow} \mathbf{h}_2 \cdots \overset{f_N}{\leftrightarrow} \mathbf{z}$, such a sequence of invertible transformations is called a normalizing flow. Then, the log-likelihood of any $\mathbf{u} \in U$ can be estimated by

$$\begin{aligned} \log p_U(\mathbf{u}) &= \log p_Z(\mathbf{z}) + \log \left|\det\left(\frac{\partial \mathbf{z}}{\partial \mathbf{u}}\right)\right| \\ &= \log p_Z(\mathbf{z}) + \sum_{i=1}^{N} \log \left|\det\left(\frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_{i-1}}\right)\right|, \end{aligned} \tag{1}$$

where $\log \left|\det(\partial \mathbf{h}_i / \partial \mathbf{h}_{i-1})\right|$ denotes the log-determinant of the Jacobian matrix $(\partial \mathbf{h}_i / \partial \mathbf{h}_{i-1})$ and this value is the change in log-density when going from $\mathbf{h}_{i-1}$ to $\mathbf{h}_i$ under transformation f_i. By maximizing $\log p_U(\mathbf{u})$, the function $\mathcal{F}$ can be optimized to transfer $p_U(\mathbf{u})$ into $p_Z(\mathbf{z})$, and the $p_U(\mathbf{u})$ can be transferred from $p_Z(\mathbf{z})$ by $\mathcal{F}^{-1}$ (i.e., $\mathcal{G}$).

Based on the above principles, various methods had been proposed. For example, Dinh *et al.* [4] proposed a real-valued non-volume preserving flow that combined multi-scale architecture with coupling layers. Kingma and Dhariwal [10] designed an invertible 1×1 convolution layer to construct normalizing flow. Further, many applications are use normalizing flow including image generation [7], video generation [12], audio generation [21], word embedding enhancement [13], anomaly detection [6], etc.

3. The Proposed Method

In this section, we first give the basic formulation to illustrate our main idea, then introduce the designed network, followed by three crucial modules.

3.1. Problem Formulation

Let a tuple $(X^{(1)}, X^{(2)}, \cdots, X^{(M)})$ denotes M heterogenous modalities of an example, where $X^{(m)}$ is the input of the m-th modality. In the complete case, all modalities are observed and easily fused to feed the downstream tasks. But in many limited scenes, some modalities could not be observed and need to be recovered for better fusion. For simplification, we introduce an indicator $\lambda \in \{0, 1\}$ to denote with $\lambda_m = 0$ if the m-th modality is missing, otherwise $\lambda_m = 1$. In the incomplete case, thus, the target is to recover those unobserved modalities $\mathcal{I}_{\text{miss}} = \{m | \lambda_m = 0\}$ based on those observed ones $\mathcal{I}_{\text{obs}} = \{m | \lambda_m = 1\}$. It is worth noting that the missing modalities need not to be consistent for the testing examples.

Our main idea is to transfer the distributions from observed modalities to missing ones through cross flows, and generate more confident prediction with high distribution consistency as shown in Fig. 1 (b). Concretely, we assume all modalities of one-class examples could be embedded into a latent space with Gaussian distributions. Due to the

powerful transfer ability, the normalizing flow is used as the embedding function, formally,

$$Z^{(m)} = \mathcal{F}^{(m)}(X^{(m)}) \sim \mathcal{N}(\mu_c, \Sigma_c), m \in \mathcal{I}_{\text{obs}}, \quad (2)$$

where $Z^{(m)}$ is the latent state of the m-th modality, $\mathcal{F}^{(m)}$ is the corresponding forward flow function, and c is the class label of the input sample $\mathcal{X}$. According to the known latent states, we infer each missing modality $\widetilde{X}^{(k)}(k \in \mathcal{I}_{\text{miss}})$ as follows:

$$\widetilde{Z}^{(k)} \leftarrow \psi(\{Z^{(m)} | m \in \mathcal{I}_{\text{obs}}\}) \sim \mathcal{N}(\mu_c, \Sigma_c), \quad (3)$$

$$\widetilde{X}^{(k)} = (\mathcal{F}^{(k)})^{-1}(\widetilde{Z}^{(k)}) \sim p_X(X^{(k)}), \quad (4)$$

where the estimated latent state $\widetilde{Z}^{(k)}$ is sampled from the Gaussian distribution of class c according to the observation on those existing modalities. In virtue of the invertibility of flow, we can infer the missing modality $\widetilde{X}^{(k)}$, abided by its original distribution $p_X(X^{(k)})$.

Although the estimated $\widetilde{X}^{(k)}$ follows the original distribution, it would deviate from the ground truth when the scatter degree of intra-class samples is large. Thus, we can further refine it with a light-weight decoder $\mathcal{D}$, i.e., $\widehat{X}^{(k)} = \mathcal{D}^{(k)}(\widetilde{X}^{(k)})$. At the training stage, we minimize the reconstruction error between $\widehat{X}^{(k)}$ and $X^{(k)}$. And, we find such a refinement step could improve the performance as shown in the experiment part.

3.2. Network Overview

The framework of our DiCMoR is illustrated in Fig. 2. It mainly consists of three parts: **shallow feature extractor**, **cross-modal distribution transfer (CMDT)**, and **multimodal fusion and prediction**. Considering the difference in the original dimensional space between modalities, we first extract multimodal shallow features and align the dimension of each modality to facilitate subsequent distribution transformation and modality recovery. To mitigate the distribution gap between recovered data and true data, we next build a CMDT network to learn the latent distribution space of each modality, and conduct cross-modal distribution transformation to estimate the distribution of missing modality, and finally recover the missing data by a decoder. To perform the classification task, in the end, the multimodal fusion and prediction part receives recovered complete multimodal data and uses multimodal transformers to fuse multimodal representation for label regression. The detail is introduced in the next subsections.

3.3. Shallow Feature Extractor

We consider three modalities: *language* (L), *visual* (V) and *acoustic* (A). Since the original dimensional spaces of the three modalities are often distinct, they are not suitable to be used directly for cross-modal transformations. To solve this problem, we design a shallow feature extractor that contains three independent temporal convolutional layers to extract the shallow features of three modalities and project them into the same dimensional space. Hence, the subsequent recovery task aims to estimate the shallow features of missing modalities from that of available modalities.

Given an input example of the class c, we can obtain the shallow features, $\mathcal{X} = \{\mathbf{X}^{(m)}\}, \mathbf{X}^{(m)} \in \mathbb{R}^{T \times d}$, where $m \in \{L, V, A\}$, T and d indicate the sequence length and the feature dimensionality. In the incomplete multimodal case, some modalities are missing either fixedly or randomly by guaranteeing at least one modality is available in $\mathcal{X}$. For the three modalities mentioned above, a total of seven missing combinational cases are included, which are reported in Tab. 1. For a convenient statement but without loss of generality, below we estimate the missing acoustic modality $\mathbf{X}^{(A)}$ from the other two modalities $\mathcal{X}_{\text{obs}} = \{\mathbf{X}^{(L)}, \mathbf{X}^{(V)}\}$, as shown in the exhibition of Fig. 2.

3.4. Cross-Modal Distribution Transfer

Let $\mathcal{F}^{(m)}$ denotes a normalizing flow model of modality m and $(\mathcal{F}^{(m)})^{-1}$ for its inverse transformation. Each normalizing flow model receives the shallow features $\mathbf{X}^{(m)}$ of that modality respectively and outputs multimodal latent states with the same Gaussian distribution, represented as $\mathbf{Z}^{(m)} = \mathcal{F}^{(m)}(\mathbf{X}^{(m)})$. At the same time, $\mathbf{Z}^{(m)}$ can be input to $(\mathcal{F}^{(m)})^{-1}$ to generate a sample $\widetilde{\mathbf{X}}^{(m)}$ with the true distribution, $\widetilde{\mathbf{X}}^{(m)} \sim p_{\mathbf{X}^{(m)}}$. Taking the example that the language and visual modalities are available, the acoustic modality is missing, as shown in Fig. 2, $\mathbf{X}^{(L)}$ and $\mathbf{X}^{(V)}$ can be input to $\mathcal{F}^{(L)}$ and $\mathcal{F}^{(V)}$ to obtain $\mathbf{Z}^{(L)}$ and $\mathbf{Z}^{(V)}$, respectively. As $\mathbf{Z}^{(L)} \sim \mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$ and $\mathbf{Z}^{(V)} \sim \mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$, we simply perform an average operation on them to sample a latent acoustic state $\widetilde{\mathbf{Z}}^{(A)} \leftarrow (\mathbf{Z}^{(L)} + \mathbf{Z}^{(V)})/2 \sim \mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$. Then it is injected to $(\mathcal{F}^{(A)})^{-1}$ to generate sample $\widetilde{\mathbf{X}}^{(A)}$ with the acoustic modality distribution. Formally,

$$\widetilde{\mathbf{X}}^{(A)} = (\mathcal{F}^{(A)})^{-1}([\mathcal{F}^{(L)}(\mathbf{X}^{(L)}) + \mathcal{F}^{(V)}(\mathbf{X}^{(V)})]/2). \quad (5)$$

Subsequently, $\widetilde{\mathbf{X}}^{(A)}$ is fed into the reconstruction module of acoustic modality to obtain the final recovered features: $\widehat{\mathbf{X}}^{(A)} = \mathcal{D}^{(A)}(\widetilde{\mathbf{X}}^{(A)})$, where $\mathcal{D}^{(A)}$ denotes the feature reconstruction module of the acoustic modality. We stack several residual channel attention blocks [27] to build a reconstruction module for each modality, where the 2D convolutional layers are replaced with 1D convolution to fit the temporal features. For any missing patterns, the set of recovered features can be denoted as $\widehat{\mathcal{X}}_{\text{miss}} = \{\widehat{\mathbf{X}}^{(m)} | m \in \mathcal{I}_{\text{miss}}\}$, thus the reconstruction loss $\mathcal{L}_{\text{rec}}$ is denoted as:

$$\mathcal{L}_{\text{rec}} = \sum_{m \in \mathcal{I}_{\text{miss}}} \|\widehat{\mathbf{X}}^{(m)} - \mathbf{X}^{(m)}\|_F^2. \quad (6)$$

Table 1. Comparison on fixed missing protocol. The values reported in each cell denote ACC_2/F1/ACC_7. **Bold** is the best.

Datasets	Available	DCCA [1]	DCCAE [26]	MCTN [20]	MMIN [34]	GCNet [16]	**DiCMoR (Ours)**
CMU-MOSI	$\{L\}$	73.6 / 73.8 / 30.2	76.4 / 76.5 / 28.3	79.1 / 79.2 / 41.0	83.8 / 83.8 / 41.6	83.7 / 83.6 / 42.3	**84.5 / 84.4 / 44.3**
	$\{V\}$	47.7 / 41.5 / 16.6	52.6 / 51.1 / 17.1	55.0 / 54.4 / 16.3	57.0 / 54.0 / 15.5	56.1 / 55.7 / 16.9	**62.2 / 60.2 / 20.9**
	$\{A\}$	50.5 / 46.1 / 16.3	48.8 / 42.1 / 16.9	56.1 / 54.5 / 16.5	55.3 / 51.5 / 15.5	56.1 / 54.5 / 16.6	**60.5 / 60.8 / 20.9**
	$\{L, V\}$	74.9 / 75.0 / 30.3	76.7 / 76.8 / 30.0	81.1 / 81.2 / 42.1	83.8 / 83.9 / 42.0	84.3 / 84.2 / 43.4	**85.5 / 85.4 / 45.2**
	$\{L, A\}$	74.7 / 74.8 / 29.7	77.0 / 77.0 / 30.2	81.0 / 81.0 / 43.2	84.0 / 84.0 / 42.3	84.5 / 84.4 / 43.4	**85.5 / 85.5 / 44.6**
	$\{V, A\}$	50.8 / 46.4 / 16.6	54.0 / 52.5 / 17.4	57.5 / 57.4 / 16.8	60.4 / 58.5 / 19.5	62.0 / 61.9 / 17.2	**64.0 / 63.5 / 21.9**
	$\{L, V, A\}$	75.3 / 75.4 / 30.5	77.3 / 77.4 / 31.2	81.4 / 81.5 / 43.4	84.6 / 84.4 / 44.8	85.2 / 85.1 / 44.9	**85.7 / 85.6 / 45.3**
	Average	63.9 / 61.9 / 20.0	66.1 / 64.8 / 24.4	70.2 / 69.9 / 31.3	72.7 / 71.4 / 31.6	73.1 / 72.8 / 32.1	**75.4 / 75.1 / 34.7**
CMU-MOSEI	$\{L\}$	78.5 / 78.7 / 46.7	79.7 / 79.5 / 47.0	82.6 / 82.8 / 50.2	82.3 / 82.4 / 51.4	83.0 / 83.2 / 51.2	**84.2 / 84.3 / 52.4**
	$\{V\}$	61.9 / 55.7 / 41.3	61.1 / 57.2 / 40.1	62.6 / 57.1 / 41.6	59.3 / 60.0 / 40.7	61.9 / 61.6 / 41.7	**63.6 / 63.6 / 42.0**
	$\{A\}$	62.0 / 50.2 / 41.1	61.4 / 53.8 / 40.9	62.7 / 54.5 / **41.4**	58.9 / 59.5 / 40.4	60.2 / 60.3 / 41.1	**62.9 / 60.4 / 41.4**
	$\{L, V\}$	80.3 / 79.7 / 46.6	80.4 / 80.4 / 47.1	83.2 / 83.2 / 50.4	83.8 / 83.4 / 51.2	84.3 / 84.4 / 51.1	**84.9 / 84.9 / 53.0**
	$\{L, A\}$	79.5 / 79.2 / 46.7	80.0 / 80.0 / 47.4	83.5 / 83.3 / 50.7	83.7 / 83.3 / 52.0	84.3 / 84.4 / 51.3	**85.0 / 84.9 / 52.7**
	$\{V, A\}$	63.4 / 56.9 / 41.5	62.7 / 59.2 / 41.6	63.7 / 62.7 / 42.1	63.5 / 61.9 / 41.8	64.1 / 57.2 / 42.0	**65.2 / 64.4 / 42.4**
	$\{L, V, A\}$	80.7 / 80.9 / 47.7	81.2 / 81.2 / 48.2	84.2 / 84.2 / 51.2	84.3 / 84.2 / 52.4	**85.2 / 85.1** / 51.5	85.1 / **85.1 / 53.4**
	Average	72.3 / 68.8 / 44.5	72.4 / 70.2 / 44.6	74.6 / 72.5 / 46.8	73.7 / 73.5 / 47.1	74.7 / 73.7 / 47.1	**75.8 / 75.4 / 48.2**

Class-specific flows. To optimize normalizing flows, the objective is typically to make all $\mathbf{Z}^{(m)}$ with the same standard MVG distribution (i.e., $\mathbf{Z}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$), but this may lead to loss of discriminability for different classes of samples. To address this issue, we introduce labels to adaptively learn class-specific Gaussian distributions, so that the latent states of different classes with different Gaussian distributions, thereby enhancing the discriminability. Given a sample $\mathcal{X}$ of class c, formally, according to Eq. (1), the loss function $\mathcal{L}_{\text{cdt}}$ of cross-modal distribution transfer can be defined as:

$$\mathcal{L}_{\text{cdt}} = -\sum_{m \in \mathcal{I}_{\text{obs}}} [\log p_{Z^{(m)}}(\mathbf{Z}^{(m)}|y=c) + \log|\det(\frac{\partial \mathbf{Z}^{(m)}}{\partial \mathbf{X}^{(m)}})|] \tag{7}$$

where the first term denotes the log-density of $\mathbf{Z}_m$ on the condition of its own category, and the second term denotes the log-determinant of normalizing flow model for modality m. In detail, $p_{Z^{(m)}}(\mathbf{Z}^{(m)}|y=c) \sim \mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$, thus the terms can be further formulated as:

$$\log p_{Z^{(m)}}(\mathbf{Z}^{(m)}) = \log(2\pi)^{-\frac{d}{2}}\det(\mathbf{\Sigma}_c)^{-\frac{1}{2}} - \frac{1}{2}(\mathbf{Z}^{(m)}-\mu_c)^T\mathbf{\Sigma}_c^{-1}(\mathbf{Z}^{(m)}-\mu_c), \tag{8}$$

$$\log|\det(\frac{\partial \mathbf{Z}^{(m)}}{\partial \mathbf{X}^{(m)}})| = \sum_{i=1}^{N}\log|\det(s_i^{(m)})|, \tag{9}$$

where $\log|\det(s_i^{(m)})|$ denotes the log-determinant of i^{th} affine coupling layer for normalizing flow of modality m, as used in [10].

Another question is how to learn $\{\mu_c, \mathbf{\Sigma}_c\}$ for difference classes. Let's take class c as an example, we first construct two tensors filled with the scalar value zero and two convolutional layers initialized with zero, and the two convolutional layers are denoted as $\text{ZeroConv}_{\mu}^{c}(\cdot)$ and $\text{ZeroConv}_{\Sigma}^{c}(\cdot)$. Then, two zero tensors are injected into two convolution layers to obtain the initialized mean μ_c and log covariance matrix $\log \mathbf{\Sigma}_c$ of the Gaussian distribution for class c respectively, represented as $\mu_c = \text{ZeroConv}_{\mu}^{c}(\mathbf{0})$ and $\log \mathbf{\Sigma}_c = \text{ZeroConv}_{\Sigma}^{c}(\mathbf{0})$. Therefore, in the initialization phase, the Gaussian distribution of class c is initialized to have a mean $\mu_c = \mathbf{0}$ and a covariance $\mathbf{\Sigma}_c = \mathbf{I}$. During the training process, μ_c and $\log \mathbf{\Sigma}_c$ of Gaussian distribution are changed additively as the bias of the convolution layers is updated, thereby adaptively fitting the class-specific Gaussian distributions based on different classes of training samples.

3.5. Multimodal Fusion and Prediction

The recovered data $\widehat{\mathcal{X}}_{\text{miss}} = \{\widehat{\mathbf{X}}^{(m)}|m \in \mathcal{I}_{\text{miss}}\}$ and the input available data $\mathcal{X}_{\text{obs}} = \{\mathbf{X}^{(m)}|m \in \mathcal{I}_{\text{obs}}\}$ are combined as the complete multimodal data for downstream multimodal fusion and prediction tasks. We employ multimodal transformers [24] to fuse multimodal features $\widehat{\mathcal{X}}_{\text{miss}} \cup \mathcal{X}_{\text{obs}}$, and the fused feature is used to predict results by fully connected layers. We integrate the above losses to reach the full optimization objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \beta(\mathcal{L}_{\text{cdt}} + \mathcal{L}_{\text{rec}}), \tag{10}$$

where $\mathcal{L}_{\text{task}}$ is the task-specific loss that is defined as mean absolute error in our experiments, β controls the importance of different loss functions. The entire optimization is implemented in an end-to-end manner, and the concrete training about the configuration of incomplete modalities could be found in the experiment part.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. To verify the effectiveness of DiCMoR, we conduct experiments on two standard multimodal video understanding datasets, including CMU-MOSI [30] and

Table 2. Comparison on random missing protocol. The values reported in each cell denote ACC_2/F1/ACC_7. **Bold** is the best.

Datasets	MR	DCCA [1]	DCCAE [26]	MCTN [20]	MMIN [34]	GCNet [16]	**DiCMoR (Ours)**
CMU-MOSI	0.0	75.3 / 75.4 / 30.5	77.3 / 77.4 / 31.2	81.4 / 81.5 / 43.4	84.6 / 84.4 / 44.8	85.2 / 85.1 / 44.9	**85.7 / 85.6 / 45.3**
	0.1	72.1 / 72.2 / 28.0	74.5 / 74.7 / 28.1	78.4 / 78.5 / 39.8	81.8 / 81.8 / 41.2	82.3 / 82.3 / 42.1	**83.9 / 83.9 / 43.6**
	0.2	69.3 / 69.1 / 26.8	71.8 / 71.9 / 27.6	75.6 / 75.7 / 38.5	79.0 / 79.1 / 38.9	79.4 / 79.5 / 40.0	**82.1 / 82.0 / 42.3**
	0.3	65.4 / 65.2 / 25.7	67.0 / 66.7 / 25.8	71.3 / 71.2 / 35.5	76.1 / 76.2 / 36.9	77.2 / 77.2 / 38.2	**80.4 / 80.2 / 40.6**
	0.4	62.8 / 62.0 / 24.2	63.6 / 62.8 / 24.2	68.0 / 67.6 / 32.9	71.7 / 71.6 / 34.9	74.3 / 74.4 / 36.6	**77.9 / 77.7 / 37.6**
	0.5	60.9 / 59.9 / 21.6	62.0 / 61.3 / 23.0	65.4 / 64.8 / 31.2	67.2 / 66.5 / 32.2	70.0 / 69.8 / 33.9	**76.7 / 76.4 / 36.4**
	0.6	58.6 / 57.3 / 21.2	59.6 / 58.5 / 20.9	63.8 / 62.5 / 29.7	64.9 / 64.0 / 29.1	67.7 / 66.7 / 29.8	**73.3 / 73.0 / 32.7**
	0.7	57.4 / 56.0 / 20.4	58.1 / 57.4 / 20.6	61.2 / 59.0 / 27.5	62.8 / 61.0 / 28.4	65.7 / 65.4 / 28.1	**71.1 / 70.8 / 30.0**
	Average	65.2 / 64.6 / 24.8	66.7 / 66.3 / 25.2	70.6 / 70.1 / 34.8	73.5 / 73.1 / 35.8	75.2 / 75.1 / 36.7	**78.9 / 78.7 / 38.5**
CMU-MOSEI	0.0	80.7 / 80.9 / 47.7	81.2 / 81.2 / 48.2	84.2 / 84.2 / 51.2	84.3 / 84.2 / 52.4	**85.2 / 85.1** / 51.5	85.1 / **85.1 / 53.4**
	0.1	77.4 / 77.3 / 46.2	78.4 / 78.3 / 46.9	81.8 / 81.6 / 49.8	81.9 / 81.3 / 50.6	82.3 / 82.1 / 51.2	**83.7 / 83.5 / 52.2**
	0.2	73.8 / 74.0 / 45.1	75.5 / 75.4 / 46.3	79.0 / 78.7 / 48.6	79.8 / 78.8 / 49.6	80.3 / 79.9 / 50.2	**81.8 / 81.5 / 51.4**
	0.3	71.1 / 71.2 / 43.6	72.3 / 72.2 / 45.6	76.9 / 76.2 / 47.4	77.2 / 75.5 / 48.1	77.5 / 76.8 / 49.2	**79.8 / 79.3 / 50.3**
	0.4	69.5 / 69.4 / 43.1	70.3 / 70.0 / 44.0	74.3 / 74.1 / 45.6	75.2 / 72.6 / 47.5	76.0 / 74.9 / 48.0	**78.7 / 77.4 / 48.8**
	0.5	67.5 / 65.4 / 42.5	69.2 / 66.4 / 43.3	73.6 / 72.6 / 45.1	73.9 / 70.7 / 46.7	74.9 / 73.2 / 46.7	**77.7 / 75.8 / 47.7**
	0.6	66.2 / 63.1 / 42.4	67.6 / 63.2 / 42.9	73.2 / 71.1 / 43.8	73.2 / 70.3 / 45.6	74.1 / 72.1 / 45.1	**76.7 / 73.7 / 46.8**
	0.7	65.6 / 61.0 / 42.1	66.6 / 62.6 / 42.5	72.7 / 70.5 / 43.6	73.1 / 69.5 / 44.8	73.2 / 70.4 / 44.5	**75.4 / 72.2 / 46.2**
	Average	71.5 / 70.3 / 44.1	72.6 / 71.2 / 45.0	77.0 / 76.1 / 46.9	77.3 / 75.4 / 48.2	77.9 / 76.8 / 48.3	**79.9 / 78.6 / 49.6**

CMU-MOSEI [31]. The above two datasets mainly focus on human multimodal sentiment analysis. **CMU-MOSI** consists of 2,199 short monologue video clips. Among the samples, 1,284, 229, and 686 samples are used as training, validation, and testing set. **CMU-MOSEI** contains 22,856 samples of movie review video clips from YouTube. According to the predetermined protocol, 16,326 samples are used for training, the remaining 1,871 and 4,659 samples are used for validation and testing. On the two datasets, we extract the language features via pre-trained BERT model [9] and obtain a 768-dimensional hidden state as the word features. For visual modality, each video frame was encoded via Facet [8] to represent the presence of the total 35 facial action units [14]. The acoustic modality was processed by COVAREP [2] to obtain the 74-dimensional features. Each sample in CMU-MOSI and CMU-MOSEI was labeled with a human sentiment score that ranges from -3 to 3, including *highly negative*, *negative*, *weakly negative*, *neutral*, *weakly positive*, *positive*, and *highly positive*. To make a comprehensive comparison, we evaluate the performance using the following metrics: 7-class accuracy (ACC_7), binary accuracy (ACC_2), and F1 score.

Implementation details. We investigate the performance of different methods on multimodal datasets with two missing protocols, including a fixed missing protocol and a random missing protocol. For the **fixed missing protocol**, we let the missing patterns be consistent for all samples (i.e., the available modalities are the same for all samples). Since the three modalities can produce seven different missing patterns, this protocol contains seven sets of experiments for seven different missing patterns. For the **random missing protocol**, the missing patterns are randomized for each sample (i.e., one or two modalities may be missing for each sample). Here, we use the missing rate (MR) to measure the overall missingness of the dataset. The MR is defined as $\text{MR} = 1 - \frac{\sum_{i=1}^{L} a_i}{L \times M}$, where a_i denotes the number of available modalities for i^{th} sample, L denotes the total number of samples, and M indicates the number of modalities. We also ensure that at least one modality is available for each sample, so $a_i \geq 1$ and $\text{MR} \leq \frac{M-1}{M}$. For three modalities, we choose the MR from $[0.0, 0.1, \cdots, 0.7]$, where 0.7 is an approximation of $\frac{M-1}{M}$ with the same meaning. We keep the same MR during training, validation, and testing phases, consistent with previous work [16]. The detailed neural network configurations are listed in the supplementary file, including the 1D temporal convolutions, the normalizing flow models, the reconstruction modules, and the multimodal transformers. The optimal setting for β is set to 0.1 via the performance on the validation set. We implemented all the experiments using PyTorch on a RTX 3090 GPU with 24GB memory. We set the training batch size as 16 and train DiCMoR for 50 epochs until convergence. We run each experiment five times and report the average values on the testing set.

4.2. Comparison with the state-of-the-arts

We compare DiCMoR with the current state-of-the-art incomplete multimodal learning methods, including two deep learning-based non-recovery methods with canonical correlation maximization: DCCA [1] and DCCAE [26], three deep learning-based recovery methods: MCTN [20], MMIN [34], GCNet [16]. Below, we report the quantitative and qualitative experimental results.

Quantitative results. Tab. 1 and Tab. 2 illustrate the comparison of fixed missing protocol and random missing protocol on two datasets, respectively. From these experimental results, we have the following observations:

1) On average, our method achieves the best perfor-

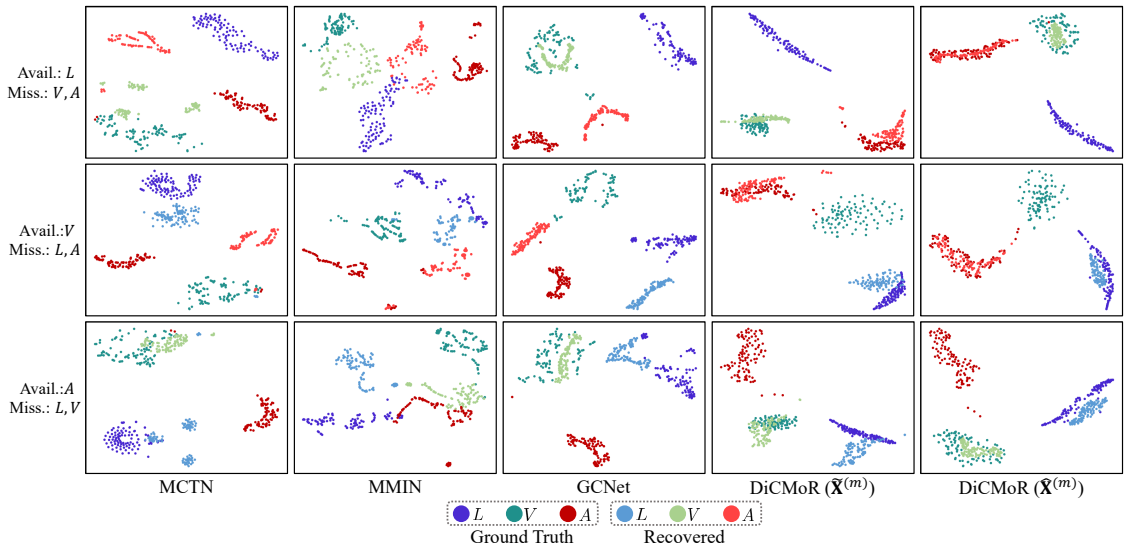

Figure 3. Visualization of recovered data and ground truth for different recovery methods under missing patterns with only one modality available. The distribution of data recovered by DiCMoR is much closer to ground truth than other methods.

mance on all datasets with two missing protocols. Compared with the non-recovery methods, the recovery methods including our DiCMoR obtain better performance. This is because the recovery methods can estimate and reconstruct missing modalities explicitly from the available modalities. Compared with the recovery methods [20, 34, 16], our proposed DiCMoR obtains consistent improvements, indicating the superiority of maintaining consistency of distribution between recovered data and true data. For further investigation, we will visualize the distribution of recovered data for different methods in the next part.

2) From the experimental results in Tab. 1, we can observe that the prediction performance of all methods tend to be better when the language modality is available, and this phenomenon also emerged in previous work [20, 34]. The potential reason is that language modality is a high-level semantic abstraction that contains more discriminative information. Therefore, it is crucial to enhance the performance of missing scenarios without language modality. Compared with these state-of-the-arts, our DiCMoR consistently obtains better results in missing scenarios without language modality. This suggests that the distribution transformed from the weak modality can improve the performance.

3) Experimental results in Tab. 2 show that the performance degradation of DiCMoR tends to be smaller than that of other recovery methods as the MR increases. Taking the results on the CMU-MOSI dataset as an example, as the MR increases from 0.0 to 0.7, the ACC_2 performance of other recovery methods declines $19.5\% \sim 21.8\%$ while our DiCMoR declines 14.6%. Notably, the performance gap between DiCMoR and other recovery methods becomes more obvious as the MR increases. For example, compared with the strongest baseline GCNet on the CMU-MOSI dataset,

Table 3. Ablation study of the key components in DiCMoR under average random missing protocol.

Dataset	$\mathcal{F}^{(m)}$	$\mathcal{D}^{(m)}$	ACC_2	F1	ACC_7
CMU-MOSI	✓	✓	**78.9**	**78.7**	**38.5**
	✓	×	76.1	75.8	36.8
	×	✓	75.8	75.6	36.8
	×	×	74.5	74.4	36.0
CMU-MOSEI	✓	✓	**79.9**	**78.6**	**49.6**
	✓	×	78.4	77.2	49.0
	×	✓	78.0	76.9	48.8
	×	×	76.5	76.0	48.5

the ACC_2 performance gap increases from 0.5% to 5.4% as the MR increases from 0.0 to 0.7. These results demonstrate that the DiCMoR improves prediction performance in missing scenarios, especially in severely missing cases.

Qualitative results. Fig. 3 visualize the distribution of recovered data and ground truth for different recovery methods when merely one modality remains, we randomly select 100 samples in the testing set from the CMU-MOSEI dataset. The features of the selected samples are projected into a 2D space by t-SNE [25]. From these results, we can observe that the distribution between true data and recovered data estimated by our DiCMoR is closer than other methods. The main reason is that DiCMoR explicitly models and learns the distribution space of different modalities and has ability to transfer the distribution across modalities.

4.3. Ablation study

Quantitative analysis. We evaluate the effects of key components for DiCMoR, including distribution transfer ($\mathcal{F}^{(m)}$) and feature reconstruction ($\mathcal{D}^{(m)}$). The results are illustrated in Tab. 3, we can draw the following conclu-

Table 4. Ablation study of $\mathcal{F}^{(m)}$ on MCTN under average random missing protocol.

Methods	CMU-MOSI			CMU-MOSEI		
	ACC_2	F1	ACC_7	ACC_2	F1	ACC_7
MCTN	70.6	70.1	34.8	77.0	76.1	46.9
MCTN w/ $\mathcal{F}^{(m)}$	**72.4**	**72.4**	**35.4**	**78.1**	**77.0**	**47.5**

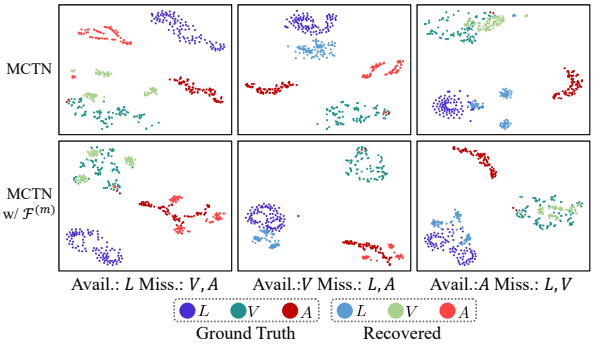

Figure 4. Visualization of recovered data and ground truth for MCTN and MCTN w/ $\mathcal{F}^{(m)}$. Obviously, our method can improve the distribution consistency of recovered data for MCTN.

sions: **1)** feature reconstruction with $\mathcal{D}^{(m)}$ or distribution transfer with $\mathcal{F}^{(m)}$ can improve the performance because both modules can explicitly recover data that can provide useful complementary information; **2)** combing distribution transfer with feature reconstruction brings further benefits, which proves that recovering data while maintaining consistent distribution is feasible and effective.

Besides, since the key difference between DiCMoR and other recovery methods is to a introduce distribution transfer network constructed by normalizing flow models $\mathcal{F}^{(m)}$, this network should be generalizable to other recovery methods. Thus, we apply this module to the classical MCTN [20] for further investigation. The results are shown in Tab. 4, which suggests that this module generalizing to other methods is feasible and effective. Furthermore, we visualize the distribution of data in Fig. 4, and we can observe that the distribution of recovered data of MCTN w/ $\mathcal{F}^{(m)}$ is much closer to ground truth.

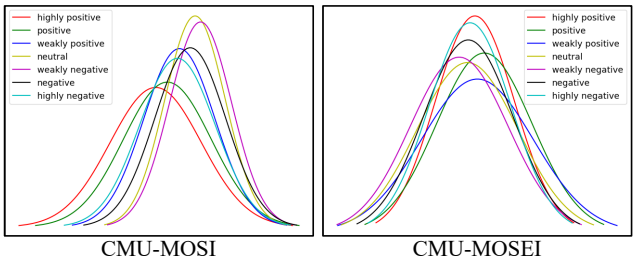

Figure 5. Visualization of class-specific Gaussian distributions.

Visualization of the class-specific Gaussian distributions and multimodal latent states. First, we visualize the class-specific Gaussian distributions learned by CMDT in Fig. 5, where we draw the one-dimensional Gaussian distribution curves by averaging the elements of μ_c and $\mathbf{\Sigma}_c$ respectively for each class. From this result, we can observe that the learned Gaussian distributions of different classes are basically scattered, which indicates that our proposed CMDT has the ability to adaptively learn Gaussian distributions with discrimination w.r.t different classes. *One question arises, can these learned Gaussian distributions help improve the discriminability of the latent states?* We further visualize the multimodal latent states of different classes in Fig. 6. Specifically, we randomly select 35 samples (five samples for each class) in the testing set of the CMU-MOSEI dataset. The features of the selected samples are projected into a 2D space by t-SNE [25]. We can discover that when using a standard Gaussian distribution (DiCMoR w/ $\mathcal{N}(\mathbf{0}, \mathbf{I})$) as the objective, different classes of multimodal latent states suffer from mode collapse due to their sharing a same Gaussian distribution space. In contrast, when using class-specific Gaussian distributions as the objective (DiCMoR w/ $\mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$), the latent states are distinguishable in seven classes, indicating that the learned Gaussian distributions can improve the discriminability of latent states.

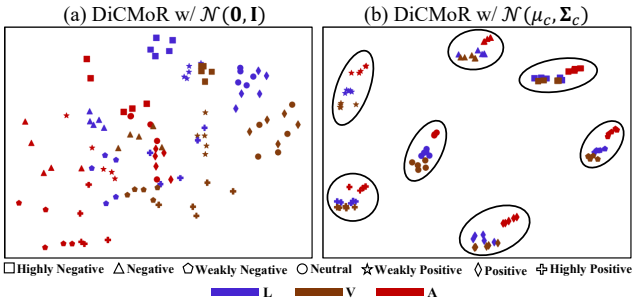

Figure 6. Visualization of the latent states. DiCMoR w/ $\mathcal{N}(\mu_c, \mathbf{\Sigma}_c)$ shows the promising 7-class separability in (b).

5. Conclusion and discussion

In this paper, we have proposed a recovery paradigm (DiCMoR) for incomplete multimodal learning that transfers the distributions from available modalities to missing modalities to maintain the distribution consistency. Further, we have designed a class-specific flow based modality recovery paradigm to transform cross-modal distributions. It facilitates cross-modal distribution transformation and enables the anticipation of a distribution-consistent latent space for the missing modality. The robustness and efficacy of DiCMoR are substantiated through extensive evaluations. It should be noted, however, that our method necessitates the availability of labels as a conditioning factor during its training phase to generate class-associated data. Under scenarios where unlabeled tasks are encountered, our method may be susceptible to a decline in performance.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62072244, 62102180), the fundamental research funds for the central universities (Grant No. 30919011232), the Natural Science Foundation of Shandong Province (Grant Nos. ZR2020LZH008, ZR2022LZH003), the Natural Science Foundation of Jiangsu Province (Grant No. BK20210329), Shuangchuang Program of Jiangsu Province(Grant No. JSSCBS20210210), and in part by State Key Laboratory of High-end Server & Storage Technology.

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 5, 6

[2] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014. 6

[3] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2, 3

[5] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 2

[6] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 3

[7] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730. PMLR, 2019. 3

[8] iMotions. Facial expression analysis. https://imotions.com/products/imotions-lab/modules/fea-facial-expression-analysis/, 2017. 6

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186, 2019. 6

[10] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 3, 5

[11] Samuel Kotz and Norman L Johnson. *Breakthroughs in statistics: methodology and distribution*. Springer Science & Business Media, 2012. 2

[12] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*, 2(5):3, 2019. 3

[13] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, 2020. 3

[14] Yong Li and Shiguang Shan. Contrastive learning of person-independent representations for facial action unit detection. *IEEE Transactions on Image Processing*, 32:3212–3225, 2023. 6

[15] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023. 1

[16] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 5, 6, 7

[17] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11174–11183, 2021. 2

[18] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 166–183, 2018. 2

[19] Srinivas Parthasarathy and Shiva Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404, 2020. 2

[20] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019. 1, 2, 5, 6, 7, 8

[21] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019. 3

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[23] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414, 2017. 1, 2

[24] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 1, 5

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7, 8

[26] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015. 5, 6

[27] Xiu-Shen Wei, Shu-Lin Xu, Hao Chen, Liang Xiao, and Yuxin Peng. Prototype-based classifier learning for long-tailed visual recognition. *Science China Information Sciences*, 65(6):160105, 2022. 4

[28] Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(3):303–315, 2017. 1

[29] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 2012. 2

[30] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 5

[31] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246, 2018. 6

[32] C Zhang, Y Cui, Z Han, JT Zhou, H Fu, and Q Hu. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2402–2415, 2022. 2

[33] Shiqing Zhang, Yijiao Yang, Chen Chen, Ruixin Liu, Xin Tao, Wenping Guo, Yicheng Xu, and Xiaoming Zhao. Multimodal emotion recognition based on audio and text by using hybrid attention networks. *Biomedical Signal Processing and Control*, 85:105052, 2023. 1

[34] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, 2021. 1, 2, 5, 6, 7