Variational Distillation for Multi-View Learning

Xudong Tian †, Zhizhong Zhang †, Cong Wang †, Wensheng Zhang †, Yanyun Qu †, Member, IEEE, Lizhuang Ma †, Member, IEEE, Zongze Wu †, Member, IEEE, Yuan Xie †, Member, IEEE, and Dacheng Tao †, Fellow, IEEE

Abstract—Information Bottleneck (IB) provides an information-theoretic principle for multi-view learning by revealing the various components contained in each viewpoint. This highlights the necessity to capture their distinct roles to achieve view-invariance and predictive representations but remains under-explored due to the technical intractability of modeling and organizing innumerable mutual information (MI) terms. Recent studies show that sufficiency and consistency play such key roles in multi-view representation learning, and could be preserved via a variational distillation framework. But when it generalizes to arbitrary viewpoints, such strategy fails as the mutual information terms of consistency become complicated. This paper presents Multi-View Variational Distillation (MV2D), tackling the above limitations for generalized multi-view learning. Uniquely, MV2D can recognize useful consistent information and prioritize diverse components by their generalization ability. This guides an analytical and scalable solution to achieving both sufficiency and consistency. Additionally, by rigorously reformulating the IB objective, MV2D tackles the difficulties in MI optimization and fully realizes the theoretical advantages of the information bottleneck principle. We extensively evaluate our model on diverse tasks to verify its effectiveness, where the considerable gains provide key insights into achieving generalized multi-view representations under a rigorous information-theoretic principle.

Index Terms—Multi-view learning, information bottleneck, mutual information, variational inference, knowledge distillation.

I. INTRODUCTION

As more and more real-world data are collected from diverse sources or obtained from different feature extractors, multi-view representation learning has become a key enabler in integrating various features (i.e., heterogeneous data or visual descriptors) from the same object, which greatly promotes the performance of existing machine learning system and thereby gains increasing attention. For example, in an auto-driving system, there usually exist camera sensors that assist LiDAR data to perceive the complex 3D visual world, and therefore enable us to take advantage of depth, texture, and color information provided by the multiple sensors. These Lidar-RGB perception methods have significantly improved the auto-driving performance.

To effectively explore multi-view data, many efforts have been devoted to learning a consistent representation for discriminative information mining, such as Canonical Correlation Analysis (CCA) [1], [2] or feature alignment [3], both of which primarily maximize the similarity between representations from different viewpoints, but are more prone to introduce non-predictive redundancy and even cause considerable loss of predictive information. Among various solutions, the information bottleneck (IB) provides an information-theoretic principle [4] for multi-view learning. It reveals that consistent information across views comprises heterogeneous components, each capturing distinct properties for representation learning (e.g., consensus and partially consistent terms as in Fig. 4(a)). However, literature on this subject [2], [3], [5] fails to differentiate between the distinct roles of these components. Instead, they naively treat them as equivalent. This results in a dilemma between preserving sample information (i.e., sufficiency) and achieving view-invariance (i.e., consistency), negatively impacting the performance and hindering real-world applications.

By fitting mutual information (MI), IB aims to maximize the correlation between representation and predictive information, while avoiding encoding task-irrelevant nuisances. This provides insights into resolving the aforementioned dilemma in multi-view learning. Yet, the practical use of IB remains...
a persistent challenge due to the notorious difficulty of estimating mutual information. Motivated by this, a plethora of trainable neural estimators [6], [7] have been presented, which, unfortunately, have poor scalability for complex networks and large-scale datasets. Another principal drawback of IB is that, its optimization objective is essentially a trade-off between having a concise representation and achieving good predictive power [4], [8], [9], which can further exacerbate the multi-view learning dilemma [10], [11].

Motivated by the preceding analysis, we propose a novel multi-view information bottleneck framework called Multi-View Variational Distillation (MV²D) to learn generalized representations across arbitrary viewpoints. In contrast to the prior works [12], [13], MV²D highlights the necessity to distinguish different compositions that constitute each viewpoint and addresses the critical challenges in modeling and organizing various information elements. This allows MV²D to identify the prioritization for each representation by automatically recognizing the useful consistent multi-view information and promises immunity against view-discrepancies and superior predictive performance. Moreover, by reformulating the objective of IB with variational inference, MV²D can fit mutual information without estimating it. This significantly alleviates the burden in optimization and enables the MV²D to fully realize the theoretical benefits of the information bottleneck principle. Additionally, we provide a detailed analysis of the distinctions between MV²D and existing cutting-edge variational information bottlenecks like Variational Self-Distillation (VSD), Variational Cross Distillation (VCD), Variational Mutual Distillation (VMD) [14], and illustrate the necessity and critical challenges in identifying the prioritization for different information elements.

We empirically investigate the proposed strategy via comprehensive experiments, and demonstrate its generalizability in various applications, including (i) Cross-modal person re-identification; (ii) Multi-view classification; (iii) LiDAR-RGB semantic segmentation. We show that our method substantially outperforms prior state-of-the-art techniques, especially on real-world datasets. Our main contributions are summarized as follows:

- We develop a novel information bottleneck strategy (MV²D) that tackles the intractability in differentiating constituents within each viewpoint by on their generalization ability. This guides an analytical and flexible solution for generalized multi-view learning and ensures both sufficiency and consistency.
- By rigorously reformulating the IB objective, we are able to purify the desirable information and filter out any nuisances irrelevant to the downstream task. This theoretical reformulation enables MV²D to fully unlock the potential of the IB principle, providing our model with maximum flexibility and scalability during training and inference.
- We extensively evaluate our strategy on diverse tasks, where the empirical results provide convincing evidence that MV²D offers both modeling and practical benefits beyond current SOTA techniques for multi-view learning.

II. RELATED WORK

A. Information Bottleneck

The seminal work of Information Bottleneck is from [4], [15], which introduces the general idea of using the information-theoretic objective to train a deep model. But, unfortunately, the optimization for IB relied on the iterative Blahut-Arimoto algorithm, which is infeasible over high-dimensional variables, e.g., deep neural networks (DNNs) [8]. On this basis, a series of explorations from theoretic study to the practical use of the IB principle have been witnessed.

Theoretic Study of IB: By using the variational inference and re-parameterization tricks, VIB [8] constructs a lower bound on IB objective (3), and enables the DNNs to handle the high-dimensional and continuous data under the guidance of IB without the restrictions that data must follow discrete or Gaussian distribution. Moreover, with the introduction of dual distortion [16], the dualIB framework is presented in [9], which shifts the research attention from training to prediction, leading to better stability compared with the classic IB structure. However, these methods fail to overcome the principal drawbacks of IB (i.e., dependency on MI estimation and the trade-off between prediction and compactness) and consequently have unsatisfactory applicability for real-world problems.

Practical Applications of IB: Typically, [17] applies (3) to the generative adversarial networks (GANs), with small modifications for a more robust generation process. [18] alleviates the overfitting of large-scale pre-trained language models in low-resource scenarios by directly adopting the original IB objective from [4], which suffers from the prediction-compactness trade-off and mutual information estimation as well. Other applications involving IB include decision-making system [19], speech processing [20], ensemble learning [21], neuroscience [22], and deep neural networks understanding [23]. However, they either heavily rely on the mutual information estimator, or reformulate the IB objective based on strong assumptions, resulting in inferior practicality.

Unlike previous studies, this work pioneers the prevention of MI estimation while realizing all theoretical advantages of the IB principle. The proposed variational distillation strategies ensure both sufficiency and compactness, thereby yielding representations that are both concise and predictive.

B. Representation Learning

The performance of machine learning methods is heavily dependent on the learned representations, which may entangle or hide different explanatory factors behind the data. Hence, a great deal of researches are devoted to designing data processing pipelines or transformations to attain representations that can support effective machine learning. Specifically, early works prefer feature engineering (we refer readers to [24] for comprehensive studies) to take advantage of human ingenuity and prior knowledge. After that, numerous deep models adopt
reconstruction-based [2] representation, which enforces similarity constraint between the input and reconstructed output; or contrastive methods [25], which learn representations by maximizing similarity between augmentations of the same data point and minimizing similarity between different data points.

Unfortunately, feature engineering is labor-intensive, and shows inferiority in extracting discriminative information in complex circumstances [24]. Moreover, reconstruction-based and contrastive methods highly rely on similarity maximization, which are more prone to overfitting or obtaining a trivial solution when encountering small-scale datasets. On the contrary, our approach enables the model to purify predictive information w.r.t. the given task while discarding those nuisances under an information-theoretic constraint. By comparison, our approach exhibits preferable flexibility and versatility in that it facilitates learning in various tasks without requiring large batch size or computation.

C. Multi-Modal/View Representation Learning

Multi-modal representation learning aims to build models that can process and relate information from multiple modalities, and its main difficulty is to explicitly measure the content similarity between the heterogeneous samples. Classical and deep multi-view representation learning methods can be roughly divided into joint representation [1], [26], alignment representation [3], [27], as well as shared and specific representation [3], [28]. The key idea of these methods is to establish a common representation space by exploring the semantic relationships among multi-view data. Please refer to [12], [13] for a comprehensive review.

Recently, another line of works (e.g., MIB [10] and MVIB [29]) extends the IB principle to the multi-view representation learning, which achieves promising results. Typically, MIB integrates the heterogeneous representations by introducing a variational bound of the traditional IB objective in [4]. However, it still requires explicit estimation of the mutual information and is restricted to the cross-view scenario. Consequently, it shows poor scalability for general cases and leaves the primary issue in multi-view learning unresolved. In contrast, this paper describes a general framework for obtaining robust representations by identifying the prioritization of different components in each viewpoint. As validated in Section V, MV2D can produce predictive yet concise representations, hence demonstrating superior generalization ability on various applications.

D. Knowledge Distillation

Knowledge distillation (KD) is a representative technique utilized for model compression and acceleration, which typically intends to learn a small student model from a large-scale teacher by minimizing a KL-divergence loss between their predictions. In addition to the student-teacher paradigm, recent developments have been extended to assistant-learning [30], mutual-learning [31], and self-learning [32] (please refer to [33] for more related works). Despite the wide use of KD, fundamental analysis of what information should be distilled is still lacking in the literature, remaining the mechanism of KD unclear.

By formulating a new knowledge distillation mechanism from the information-theoretic perspective, MV2D reveals the constituents and their prioritization enclosed in each variable and ensures purification to the target information while simultaneously eradicating those task-irrelevant nuisances under theoretically guaranteed.

E. Differences From the Preliminary Version

This study is a journal extension of the conference paper [14] with the following differences and improvements (see the brief summary in Table II):

- The motivations differ - our prior work [14] focuses on alleviating the IB optimization in single/cross-case. On the contrary, MV2D intends to retain and prioritize the innumerable MI terms within each view (see (15)). Note this is highly non-trivial as it promises view-invariance and superior predictive power on multi-view data, yet remains unresolved due to the technical intractability in (15).
- The formulations are different. Our prior work provides a specific solution to maximizing explicit and unitary objectives like \( I(y; z) \) and \( I(y; z_1; z_2) \). MV2D instead solves the (15) by recursively preserving and consolidating different constituents based on the frequency that they are shared across viewpoints. This aligns priority with generalization capability in an efficient, tractable manner, and thus enables preferable feasibility and scalability.
- We extensively evaluate MV2D across diverse multi-view applications and provide comprehensive ablations isolating each key component. The substantial performance gain across all applications verifies both the theoretical soundness and practical effectiveness of our modeling choice.

III. PRELIMINARY

In this section, we first provide a brief review of the IB principle [4] in the context of supervised learning. Then we introduce several variations of mutual information encompassed in our method. We summarize the basic notations and their descriptions in Table I.

A. Information Bottleneck Principle

Given data observations \( V \) and labels \( Y \), the goal of representation learning in the supervised context is to obtain an encoding \( Z \) which is maximally informative w.r.t \( Y \) (i.e., sufficiency in Fig. 1), and without any additional information about \( V \) (i.e., minimality in Fig. 1), measured by mutual information, i.e.,

\[
I(Z; Y) = \int p(z, y) \log \frac{p(z,y)}{p(z)p(y)} \, dz \, dy. \tag{1}
\]

Based on the information processing principle, we illustrate the definitions of sufficiency and minimality in Fig. 1, where the areas of the three circles represent \( H(v) \), \( H(y) \) and \( H(z) \), respectively. To encourage the encoding process to focus on the label information, IB was proposed to enforce an upper bound \( \bar{I} \) to the information flow from the observations \( V \) to the encoding...
**Table I**

| Notation | Meaning |
|----------|---------|
| \( x, y \) | object, and the corresponding ground-truth label observation, representation of \( x \) |
| \( v, x \) | observations collected from different viewpoints or the entire \( x_1, \ldots, x_n \) but excluding \( x_i \) |
| \( \{x_1, \ldots, x_n\} \) | multi-view representations, denoted as \( x_1, \ldots, x_n \) |
| \( I(v; z) \) | mutual information, denoted as MI for simplicity |
| \( I(v; z|y) \) | conditional MI, abbreviated as \( I_{v|y} \) in diagrams |
| \( I_{vz} \) | information exclusive to \( z \), i.e., \( I(y; z|x, z) \) |
| \( I_{vz} \) | information exclusive to \( z \), i.e., \( I(y; z|x, z) \) |
| \( I_{vz} \) | MI among \( y, z_1 \) and \( z_2 \), i.e., \( I(y; z_1; z_2) \) |
| \( I_{vz} \) | consistency of \( y, z_1, z_2, z_3, \) i.e., \( I(y; z_1; z_2; z_3) \) |
| \( E_{\theta}(z|v) \) | encoding process from \( v \) to \( z \) with parameter \( \theta \) |
| \( E_{\theta}(z|v) \) | expectation of \( z \) obtained from \( E_{\theta}(z|v) \) |
| \( H(\cdot) \) | Shannon entropy, conditional entropy |
| \( P_{\theta} \) | predictions conditioned on \( v \) and \( z \) |
| \( P_{\theta}(x|z, y) \) | prediction conditioned on \( x_{(1, \ldots, n)} \) |

**Table II**

| Method          | Target                  | Challenge                    |
|-----------------|-------------------------|------------------------------|
| VSD             | maximizing \( I(y; z) \) | reforming IB                 |
| VCD & VMD       | maximizing \( I(v; x_1; x_2) \) | modeling \& organizing countless MI terms |
| MV^2D           | retaining and prioritizing (for \( z_1 \)): \( \sum_{j \in \mathcal{V}_1} I(y; z_j|x_{(1, \ldots, n)}/s_j) \), \( \sum_{j,k \in \mathcal{V}_1} I(y; z_j; z_k|x_{(1, \ldots, n)}/s_{j,k}) \), \ldots, \( \sum_{m \in \mathcal{V}_1} I(y; z_{m_1}; \ldots; z_{m_n}|x_{m}) \), \( I(y; z_1; \ldots; z_n) \) | modeling \& organizing countless MI terms |

**Fig. 1.** Venn diagram visualization of entropies and mutual information for three variables \( v, z \) and \( y \).

Z, by maximizing the following objective:

\[
\max I(Z; Y) \quad \text{s.t. } I(Z; V) \leq \hat{I}.
\]  

Equation (2) implies that a compressed representation can improve the generalization ability by ignoring irrelevant distractors in the original input. By using a Lagrangian objective, IB allows the encoding to be maximally expressive about \( Y \) while being maximally compressive about \( X \) by:

\[
L_{IB} = I(Z; V) - \beta I(Z; Y),
\]

where \( \beta \) is the Lagrange multiplier. However, it has been shown that it is impossible to achieve both objectives in (3) practically [8], [10] due to the trade-off optimization between high compression and strong predictive power.

More significantly, estimating mutual information in high dimension imposes additional difficulties [6], [7] for optimizing IB.

**B. The Chain Rule of Mutual Information**

The chain rule [8], [10], [34] can be utilized to subdivide the mutual information into multiple terms, which are defined and visualized in Fig. 1 (e.g., \( I(v; y) \) can be decomposed to \( I(z; y) \) and \( I(v; z|y) \), and the latter one represents the shared information between \( v \) and \( z \) but excluding \( y \)). On this basis, numerous variants measuring statistical dependencies among different variables can be expressed, such as conditional mutual information [10], [34], interaction information [35], and total correlation [36].

However, in practice, estimating the mutual information with even the simplest form (e.g., \( I(v; z) \)) in high dimension can be particularly challenging, let alone other sophisticated variations.

**IV. METHOD**

We first start from a single-variable scenario [14] (Section IV-A) where \( v \in V \) is an observation of input data \( x \in X \) extracted from an encoder \( E(v|x) \). The challenge of optimizing an information bottleneck can be formulated as finding an extra encoding \( E(z|v) \) that preserves all label information contained in \( v \), while simultaneously discarding task-irrelevant distractors. To this end, we show the key roles of two characteristics of \( z \), i.e., sufficiency and consistency, and formulate a general solution to keeping both of them in typical multi-view scenarios (Section IV-C).

A. Single-View Case: Variational Self-Distillation

An information bottleneck is used to produce representations \( z \) for keeping all predictive information w.r.t label \( y \) while avoiding encoding task-irrelevant information. It is also known as sufficiency of \( z \) for \( y \), which is defined as:

**Definition 1. Sufficiency: Representations \( z \) are sufficient w.r.t. the label \( y \) if \( I(v; y) = I(z; y) \).**

Here \( v \) is the corresponding data observation containing all label information. Then, by factorizing the mutual information between \( v \) and \( z \), we can identify two terms:

\[
I(v; z) = I(z; y) + I(v; z|y),
\]

where \( I(z; y) \) denotes the amount of label information maintained in the representation \( z \), and \( I(v; z|y) \) represents the information shared between \( v \) and \( z \), excluding \( y \), i.e., task-irrelevant nuisances (see Fig. 1 for visualization). In the view of above, sufficiency requires: (i) maximizing \( I(z; y) \); and (ii) reducing \( I(v; z|y) \). To avoid such a min-max game, we utilize the data processing inequality \( I(z; y) \leq I(v; y) \) and have:

\[
I(v; z) \leq I(v; y) + I(v; z|y).
\]
Note (5) indicates approximating $I(z; y)$ to $I(v; y)$ complies with both (i) and (ii), hence we further reformulate the objective of sufficiency as:

$$\min I(v; y) - I(z; y).$$

(6)

However, due to the notorious difficulty in mutual information estimating, (6) is intractable to be directly optimized. To address this, we introduce the following theory:

**Theorem 1:** (6) can be achieved through minimizing the KL-divergence between the predicted distributions of a sufficient observation $v$ and its representation $z$. That is:

$$D_{KL} [P_v || P_z] = 0 \Rightarrow I(v; y) - I(z; y) = 0,$$

where $P_z = p(y|z)$, $P_v = p(y|v)$ represent the predicted distributions, and $D_{KL}$ denotes the KL-divergence.

Please refer to our previous work [14] for proofs. In the view of above, sufficiency of $z$ for $y$ could be acquired by the following objective:

$$L_{VSD} = \min_{\theta, \phi} E_{v \sim E_{\theta}(v|x)} \left[ E_{z \sim E_{\phi}(z|v)} \left[ D_{KL} [P_v || P_z] \right] \right].$$

(7)

Here, $\theta, \phi$ denote the parameters of the encoder and information bottleneck, respectively. More importantly, to prevent the trivial solution $v = z$ and degeneration of $I(v; y)$, we adopt a much lower dimension for $z$ and have the predictions $p(y|v)$ detached when computing (7). That allows us to seek a concise yet discriminative representation.

**Discussion:** Compared with other information bottlenecks [4], [8], (7) resolves the difficulty of mutual information optimization while maintaining all theoretical benefits of IB, which provides an effective learning strategy for single-view cases. On the other hand, it is essentially a self-distillation method that purifies the task-relevant knowledge, hence it is named Variational Self-Distillation (see Fig. 2 for the framework).

**B. Cross-View Case: Variational Cross Distillation and Variational Mutual Distillation**

To take a step forward, we next give a preliminary definition of consistency and introduce two strategies for dealing with typical cross-view issues.

Consider $v_1, v_2$ as two sufficient observations of $x$ from different viewpoints. Consistency entails the representations to be robust to view-discrepancy while maintaining sufficiency for the label $y$. That is, maximally preserving the predictive information accessible from both viewpoints. Thus, we have the following definition:

**Definition 2:** Consistency: $z_1$ and $z_2$ are view-consistent if $I(z_1; z_2; y) = I(v_1; v_2; y)$.

Apparently, consistency requires encoding $z_1$ and $z_2$ with the predictive information shared by both viewpoints. To achieve consistency, we first show the following factorization:

$$I(v_1; z_1) = I(v_1; z_1|z_2) + I(z_1; z_2).$$

(8)

Note $I(v_1; z_1|z_2)$ represents that, the information contained in $z_1$ is exclusive to $v_1$ and is not predictable by observing $z_2$, i.e., view-specific information, while $I(z_1; z_2)$ denotes the information shared by $z_1$ and $z_2$. Next, by further subdividing $I(z_1; z_2)$, we have:

$$I(z_1; z_2) = I(z_1; z_2|y) + I(z_1; z_2|\bar{y}),$$

(9)

where $I(z_2; z_2|y)$ denotes the task-irrelevant nuisances, and $I(z_1; z_2|\bar{y})$ is the predictive information shared by both views. Substituting (9) into (8), we show:

$$I(v_1; z_1) = I(v_1; z_1|z_2) + I(z_1; z_2|y) + I(z_1; z_2|\bar{y}),$$

(10)

reformulating consistency w.r.t. $z_1$ as (vice versa for $z_2$):

$$\min I(v_1; z_1|z_2) + I(z_1; z_2|y) - I_y,$$

where $I_y = I(z_1; z_2; y)$ denotes the consistent information. Moreover, $I_y = I_y^v$ and both of them are indistinguishable (see $I_{y, z_1, z_2}$ in Fig. 6(b)), which enables us to introduce the following theory to promote consistency in cross-view scenarios.

**Theorem 2:** Given two different sufficient observations $v_1, v_2$ of an object $x$, the corresponding representations $z_1$ and $z_2$ are view-consistent when:

$$D_{JS} [P_{z_1} || P_{z_2}] + D_{KL} [P_{v_1} || P_{z_1}] + D_{KL} [P_{v_2} || P_{z_2}] = 0,$$

where $D_{JS}$ and $D_{KL}$ denote the represents Jensen-Shannon divergence and KL-divergence, respectively. $P_{z_1} = p(y|z_1)$ is the predicted distribution (similar to the others).

Please refer to Appendix.B in our previous work [14] for detailed proofs. With the above strategy, consistency for $z_1$ and $z_2$ can now be achieved by:

$$L_{VMD} = \min_{\theta, \phi} E_{v_1, v_2 \sim E_{\theta}(v|x)} \left[ E_{z_1, z_2 \sim E_{\phi}(z|v)} \left[ D_{JS} [P_{z_1} || P_{z_2}] \right] \right],$$

$$L_{VCD} = \min_{\theta, \phi} E_{v_1, v_2 \sim E_{\theta}(v|x)} \left[ E_{z_1, z_2 \sim E_{\phi}(z|v)} \left[ D_{KL} [P_{v_1} || P_{z_1}] + D_{KL} [P_{v_2} || P_{z_2}] \right] \right].$$

(12)

Here $\theta$ and $\phi$ also parameterize the encoder and information bottleneck (see Fig. 3 for graphical illustration). Based on the symmetrically-crossing structure, the objectives in (12) are named Variational Mutual Distillation (i.e., VMD) and Variational Cross Distillation (i.e., VCD), respectively.

**Discussion:** In the cross-view case, consistency is explicitly defined as a unique mutual information term among the cross-view observations and label variable. However, the optimization
becomes complicated, as there exists plenty of partially overlapped consistent terms when generalizing to arbitrary viewpoints. To address this, we next introduce a generalized learning strategy to facilitate both sufficiency and consistency.

C. Generalized Variational Distillation for Multi-View Representation Learning

Considering the general case with arbitrary viewpoints, a critical challenge emerges where the consistent information is reconstituted with a variety of heterogeneous components. Given $z_i$ in Fig. 4(a) as an example, we have the following factorization to elucidate the information contained in the representation:

$$I(v_i; z_i) = I(v_i; z_i|y) + I(z_i; y).$$

(13)

Here, $z_i$ is the representation of the $i$th viewpoint $v_i$, and $I(v_i; z_i|y)$ denotes the label and task-irrelevant information, respectively. On this basis, $I(v_i; z_i)$ can be further factorized by decomposing $I(z_i; y)$:

$$I(v_i; z_i) = I(v_i; z_i|y) + I(z_i; y),$$

(14)

$$= I(v_i; z_i|y) + \sum_{j \in n} I(y; z_i|z_{1:n}/i) + I^c.$$  

(15)

Herein $I(y; z_i|z_{1:n}/i)$ denotes the task-irrelevant information exclusive to $z_i$, while $I^c$ is the consistent information and includes the following terms:

$$I^c = \sum_{j \in n} I(y; z_i; z_j|z_{1:n}/i,j).$$

(16)

where $y$ denotes the entire $\{z_1, \ldots, z_n\}$ excepting $z_i$ and $z_j$, thus $I(y; z_i; z_j|z_{1:n}/i,j)$ represents the label information shared between $z_i$ and $z_j$ but inaccessible to all other viewpoints (vice versa for other partially-consistent terms). $I(y; z_1; \ldots; z_n)$ is shared by $\{z_1, \ldots, z_n\}$ (consensus).

Unlike the explicitly and uniquely defined $I(z_1; z_2; y)$ in (10), the consistent information is now encoded with numerous components, and each of them demonstrates a disparate formulation. Hence solving (15) requires organizing countless multi-variant mutual information terms. Moreover, different constituents play a distinct role in gathering information and providing view-invariance which makes prior efforts either (i) indiscriminately capturing all of (15) prone to viewpoint discrepancies (Fig. 5(c)), or (ii) only collecting the consensus results in inferior prediction (Fig. 5(a)). This calls for a tractable and efficient way to not only retain the constituents in (15) but also prioritize them based on their diverse properties.

Motivated by this, we propose consolidating the overlapped consistent terms in $\{I_1^c, \ldots, I_n^c\}$ for each $z_i \in \{z_1, \ldots, z_n\}$. Such simplification allows us to differentiate the constituents in (15) based on how frequently they are shared across viewpoints. To elaborate, an initial solution can be formed as:

$$\max \sum_{i \in n} I^c_{i} \text{ consistent task-irrelevant}.$$  

(16)

Intuitively, (16) seeks to ensure $\{I_1^c, \ldots, I_n^c\}$ while filtering out the task-irrelevant nuisances. While ideal in concept, (16) remains practically intractable; e.g., each representation encompasses dozens of terms even when handling the most common benchmarks [37], [38].

Hence to alleviate this, we present the
following reformulation:

$$\min \sum_{i \in n} I(v_i; z_i | y) - I(z_i; y) + I(y; z_i | z_{1, \ldots, n} / i).$$ (17)

Note $I(y; z_i | z_{1, \ldots, n} / i)$ is against with view-consistency and should be eliminated as it can only provide view-specific bias. On this basis, (17) fulfills the same goal as (16) and introduces two key distinctions warrant emphasis: (i) It explicitly guarantees both sufficiency and compactness for each $z_i$; (ii) It ensures $\{f^c_i\}_{i=1}^n$ by instead solving the view-specific bias with the predictive information maximized.

To achieve (17), we present the following theory addressing the mutual information optimization.

**Theorem 3:** Given $n$ observations $\{v_1, v_2, \ldots, v_n\}$ from different viewpoints, the corresponding representations $\{z_1, z_2, \ldots, z_n\}$ are consistent iff:

$$\min_{i \in n} D_{KL}[P_{v_i} || P_{z_i}] + D_{KL}[P_{z_{1, \ldots, n}} || P_{z_{1, \ldots, n} / i}],$$

where $P_{z_{1, \ldots, n}} = p(y; z_{1, \ldots, n})$, $P_{z_{1, \ldots, n} / i} = p(y; z_{1, \ldots, n} / i)$, $P_{v_i} = p(y | v_i)$ and $P_{z_i} = p(y | z_i)$ are the predictions obtained with different variables.

Please refer to A.1-A.2 in the supplementary material for detailed proofs and formal assertions.

As depicted in Fig. 6(b), the consensus and partially consistent terms are weighted, namely prioritized, in accordance with their common inclusion across viewpoints. This confers two cardinal advantages: (i) It addresses the intractability in handling the innumerable components within $f^c_i$; (ii) It enables prioritization of distinct compositions by their generalization ability. Thereby satisfying the criteria delineated in (15).

In light of the preceding analysis, the general consistency can be defined as:

**Definition 3:** General Consistency: For any $z_i \in \{z_1, \ldots, z_n\}$, it is view-consistent iff: $I(v_i; z_i | y) + I(y; z_i | z_{1, \ldots, n} / i) = 0$.

This serves to preserve and prioritize different constituents in $\{f^c_i\}_{i=1}^n$ and maintains compatibility with the cross-view scenario. Formally, it can be achieved through:

$$L_{MV^2D} = \min_{\theta, \phi} \sum_{i \in n} E_{v_i \sim E_\theta (v_i | x) E_{z_{1, \ldots, n} / i} \sim E_\phi (z_{1, \ldots, n} / i)} [D_{KL}[P_{v_i} || P_{z_i}]] + \sum_{i \in n} E_{v_i \sim E_\theta (v_i | x) E_{z_{1, \ldots, n} / i} \sim E_\phi (z_{1, \ldots, n} / i)} [D_{KL}[P_{z_{1, \ldots, n}} || P_{z_{1, \ldots, n} / i}]].$$ (18)

The ensuing Multi-View Variational Distillation (MV$^2$D) framework is elaborated in Fig. 7, where $\theta$ and $\phi$ parameterize the encoder and information bottleneck, respectively.

**Discussion:** Unlike prior multi-view learning techniques [12], [13], MV$^2$D is the first to recognize and characterize the diversity encoded in consistent information. This key insight highlights the necessity to prioritize different compositions according to their generalization ability, and thus motivates an analytical solution to achieving both strong predictive power and robustness on multi-view data. On the other hand, MV$^2$D distinguishes itself among IB-based variants [10], [29], [35], [40] by (i) addressing the intractability in optimization for arbitrary input
viewpoints, (ii) and maintaining the merits of both sufficiency and consistency.

Furthermore, unlike most competitors limited to small-scale datasets [9], [28], [41], MV2D offers the highest flexibility and scalability, enabling strong performance on large-scale real-world applications (see Section V).

**Instantiation:** We next adopt the simplest example, namely 3-view case (Fig. 6(a)), to further elaborate our modeling strategy. Based on (15)–(14), we have:

\[
I(v_1; z_1) = I(v_1; z_1|y) + I(z_1; y),
\]

\[
= I(v_1; z_1|y) + I(z_1; y|z_2, z_3) + I^c_i,
\]

\[
I^c_i = \begin{cases} I(z_1; z_2; z_3; y) + I(z_1; z_2; y|z_3) + I(z_1; z_3; y|z_2) \\ \text{consensus} \\ \text{partially-consistent} \end{cases}
\]

(19)

On this basis, the consistent terms encoded in \( \{I^c_i\}_{i=1}^n \) can be unfolded as:

\[
I^c_1 = I(z_1; z_2; z_3; y) + I(z_1; z_2; y|z_3) + I(z_1; z_3; y|z_2),
\]

\[
I^c_2 = I(z_1; z_2; z_3; y) + I(z_2; z_1; y|z_3) + I(z_2; z_3; y|z_1),
\]

\[
I^c_3 = I(z_1; z_2; z_3; y) + I(z_3; z_1; y|z_2) + I(z_3; z_2; y|z_1). \quad (20)
\]

By ensuring \( \{I^c_i\}_{i=1}^n \), MV2D yields disparate replication quantities for each constituent, aligning their frequency to be shared across viewpoints with apt priorities - the consensus has the largest proportion (i.e., top priority), while partially consistent terms have decreasing proportions (as depicted in Fig. 6(b)).

**D. Distinctions Between Different Variations**

Prima facie, we have made things harder by decomposing the consistent information that is widely assumed to be unitary into a diversity of components. Yet, determining their priority is highly non-trivial: It enables one to attain all theoretical benefits from both sufficiency and consistency, and thus promises preferable feasibility in various tasks. However, this has received little attention in recent work [12], [13], including VCD & VMD [14] due to the technical obstacles in organizing countless mutual information terms.

In contrast, MV2D addresses such intractability by redesigning the objective, formulation and solution of prior variational distillation approaches, which confers several key differences to [14] and handles the imperative yet technically formidable challenges. Please refer to the B.1-B.2 in the supplementary material for detailed discussions distinguishing MV2D from our prior work [14] and elucidating their connection.

**V. APPLICATIONS**

In this section, we show the proposed variational distillation framework could be flexibly applied to various multimodal/multi-view representation learning tasks: (i) Visible-Infrared Person Re-identification; (ii) Multi-view Classification; (iii) LiDAR-Image Semantic Segmentation. The quantitative and qualitative results demonstrate the effectiveness of our approach in various circumstances (e.g., single-view, cross-view and multiple-view).

![Illustrative examples of the adopted viewpoints in 3-view experiments.](image)

**A. Cross-Modal Person Re-Identification**

We first evaluate our approach on Visible-Infrared Person Re-identification task. In this application, there are two kinds of images from different modals (i.e., infrared and visible), and the objective is to match the target person images among a gallery of images when given a query image from another modal. The key challenge of this task hence lies in the huge heterogeneous gap between the visible and infrared images, which requires both complementary and consistent information to facilitate cross-modal retrieval. To verify the effectiveness of MV2D framework, we also design an intermediate modality by transforming both kinds of images to a new uniform image representation (see Fig. 8).

1) **Evaluation Protocol and Benchmarks:** In this section, we introduce the adopted benchmark datasets and corresponding evaluation standards. We follow the popular protocol [48], [49] for evaluation, where both cumulative match characteristic (CMC) and mean average precision (mAP) are used.

**SYSU-MM01** [49] is collected from 6 cameras of both indoor and outdoor environments. It contains 287,628 visible images and 15,792 infrared images of 491 different persons in total, each of which is at least captured by two cameras. There are two search modes on SYSU-MM01, i.e., all-search mode and indoor-search mode, and the difference lies in whether the outdoor cameras are excluded from the gallery.

**RegDB** [50] is collected from two aligned cameras (one visible and one infrared) and it totally includes 412 identities, where each identity has 10 infrared images and 10 visible images. Following the experimental protocol in [50], we divide the dataset into training and test sets randomly, each of which includes non-overlapping 206 identities. We test our model in both visible-to-thermal and thermal-to-visible settings. The final reported results are averaged over 10 trials with different training/test splits.

2) **Implementation Details. Critical Architectures:** For both MM01 and RegDB, we deploy three parallel branches, each of which is composed of a ResNet50 backbone (i.e., encoder \( E_o \)) and an information bottleneck (i.e., \( E_o \): multi-layer perceptrons of 2 hidden ReLU units of size 1,024 and 512 respectively with an output of size \( 2 \times 256 \) that parameterizes mean and variance). For each branch, the backbone first encodes the input image to 2048-D feature (i.e., observation \( v \)), then it is forwarded to the information bottleneck to obtain the
| Settings | Visible2 Thermal | Thermal2 Visible |
|----------|-----------------|-----------------|
| Hi-CMD [42] | CVPR’20 | 70.9 | 66.0 | - | - |
| DDAG [43] | ECCV’20 | 69.3 | 63.5 | 68.1 | 61.8 |
| NTSS [44] | CVPR’21 | 80.5 | 72.1 | 77.9 | 69.8 |
| MCLNet [45] | CVPR’21 | 80.3 | 73.1 | 75.9 | 69.5 |
| CAMAlign [27] | ICCV’21 | 67.6 | 74.2 | 65.5 | 65.9 |
| AGW [48] | TPAMI’21 | 70.1 | 66.4 | 70.5 | 72.4 |
| ours (baseline) | - | 79.9 | 77.2 | 77.5 | 76.2 |
| ours (2-view) | - | 81.6 | 78.7 | 79.1 | 77.5 |
| ours (3-view) | - | 83.1 | 80.1 | 81.2 | 78.4 |

Note that all methods are measured by CMC and mAP on SYSU-MM01.

TABLE IV
Comparison with the state-of-the-arts on RegDB dataset under visible-thermal and thermal-visible settings

| Method | Venue | Visible | Rank-1 | Rank-5 | Rank-10 | Rank-20 | mAP |
|--------|-------|---------|--------|--------|--------|--------|-----|
| Hi-CMD [42] | CVPR’20 | 70.9 | 66.0 | - | - | - |
| DDAG [43] | ECCV’20 | 69.3 | 63.5 | 68.1 | 61.8 | - |
| NTSS [44] | CVPR’21 | 80.5 | 72.1 | 77.9 | 69.8 | - |
| MCLNet [45] | CVPR’21 | 80.3 | 73.1 | 75.9 | 69.5 | - |
| CAMAlign [27] | ICCV’21 | 67.6 | 74.2 | 65.5 | 65.9 | - |
| AGW [48] | TPAMI’21 | 70.1 | 66.4 | 70.5 | 72.4 | - |
| ours (baseline) | - | 79.9 | 77.2 | 77.5 | 76.2 | - |
| ours (2-view) | - | 81.6 | 78.7 | 79.1 | 77.5 | - |
| ours (3-view) | - | 83.1 | 80.1 | 81.2 | 78.4 | - |

TABLE V
Accuracy of our method when using different training strategies. Note: All experiments are conducted on SYSU-MM01 under all-search single-shot mode

| Settings | R1 | R10 | R20 | mAP |
|----------|----|-----|-----|-----|
| 1 $E_S$ | 53.19 | 88.19 | 94.69 | 49.16 |
| 2 $E_S$ + CIB | 38.02 | 74.68 | 83.85 | 37.94 |
| 3 $E_S$ + BS | 56.95 | 93.11 | 97.66 | 57.01 |
| 4 $E_S$ + BS + VCD + VMD | 61.16 | 94.61 | 97.97 | 60.76 |
| 5 $E_{1/V}$ | 57.61 | 93.68 | 97.68 | 56.24 |
| 6 $E_{S1/V}$ + CIB | 41.65 | 79.65 | 88.77 | 41.69 |
| 7 $E_{1/V}$ + BS1/V | 62.26 | 95.08 | 98.79 | 59.27 |
| 8 $E_{1/V}$ + BS1/V + VSD | 69.33 | 95.71 | 98.63 | 66.27 |
| 9 $E_{S1/V}$ | 58.60 | 93.59 | 97.83 | 57.35 |
| 10 $E_{S1/V}$ + CIB | 43.81 | 82.91 | 92.82 | 41.77 |
| 11 $E_{S1/V}$ + BS1/V | 64.15 | 94.42 | 98.68 | 61.74 |
| 12 $E_{S1/V}$ + BS1/V + MV2D | 70.02 | 96.17 | 98.76 | 66.70 |

1 Some results are compared for completeness, as conventional IB does not explicitly enforce any constraints to the observation.

TABLE VI
Computational cost of different methods

| Method | Enc | IB | MIE | Time | Params |
|--------|-----|----|-----|------|-------|
| Baseline | √ | | | 1.0x | 1.0x |
| Ours | √ | √ | | 1.09x | 1.15x |
| CIB | √ | √ | | 1.26x | 1.35x |
IB for training. “VSD”, “VMD”, “VCD” and “MV2D” denote our approaches. Based on Table V, we have the following observations:

i) Information bottleneck architecture can improve the performance in single-view, cross-view, and triple-view cases (see 3rd, 7th and 11th row in Table V).

ii) Conventional IB strategy exhibits no advantages in learning predictive information (see 2nd, 6th and 10th row in Table V). We conjecture it is because, by explicitly reducing $I (v; z)$, conventional IB may not recognize label information from task-irrelevant distractors, and probably discard all of them. On the other hand, estimation of mutual information in high dimensions is difficult, especially when involving multi-modal data and latent variables in our setting, which leads to a sharp drop in performance.

iii) Our approach provides remarkable improvement under various settings (see 4th, 8th and 12th row in Table V). In single-view case, by maximally preserving predictive information while simultaneously reducing superfluous details, VSD outperforms the conventional IB by 27.68%@Rank-1 and 24.58%@mAP (comparing 8th with 6th row in Table V). In cross-view case, VCD and VMD achieve 23.14%@Rank-1 and 22.82%@mAP improvement against the conventional IB (see 4th and 2nd row in Table V), demonstrating huge advantages as well.

Sufficiency & Consistency: For better illustration, we plot the 2D projection of the representations by using t-SNE on Figs. 9 and 10, where we compare our approach with the conventional IB. In particular, $z_{sp}$ and $z_{sh}$ denote the representations obtained from the modal-specific branches and the modal-shared branch respectively, and the superscripts $I$ and $V$ indicate the corresponding inputs are infrared or visible. Based on Figs. 9 and 10, we have:

i) As shown in Fig. 9(e)–(h), the embedding space of conventional IB is mixed, demonstrating the inferior predictive power and severe redundancy of the learned representation. On the contrary, our method shows an evident boost to the discriminative ability with clear class boundaries (see Fig. 9(a)–(d)).

ii) From Fig. 9(g), (h), and 10(b), we observe the conventional IB is quite vulnerable and sensitive to modal changes, where we can find the discrepant embedding spaces from different modals. Such phenomenon is not surprising since conventional IB cannot explicitly distinguish modal-consistent/specific information. By contrast, the embedding space of $z_{sh}$ and $z_{sp}$ obtained from our method appears to coincide with each other (see Fig. 9(a), (d), and 10(a)), implying that we can learn a consistent representation.

Complexity: We also compare the extra computational and memory costs brought by our method and conventional IB. As shown in Table VI, “Enc” denotes the encoder, i.e., backbone network, “IB” and “MIE” represents the information bottleneck architecture and mutual information estimator. Clearly, our approach avoids explicit calculation to (3), and thus implements the IB principle with negligible cost. On the other hand, our method also demonstrates robustness and flexibility to the variance of batch size (see Fig. 12(c)), preventing the reliance on large storage costs.

B. Multi-View Classification

Multi-view classification aims to optimally integrate various representations from different visual views to improve classification accuracy. “Multi-view” in this context means each object is described by different descriptors, and hence there exists a pre-extracted feature set including heterogeneous features with tremendous diversity and complementary information. This scenario is therefore satisfactory to validate the effectiveness of our MV2D (multiple views cases), which allows the network to promote both sufficiency and view-consistency.

1) Evaluation Protocol and Benchmarks: Following [1], [2], [28], the adopted multi-view benchmark datasets are split into three parts (i.e., 70%/20%/10%) for training, validation and testing, respectively. Classification accuracy is utilized as the prominent evaluation metric for conducting comparisons with state-of-the-art techniques.

Caltech-101/20 [37] consists of 101 categories of images. Following [28], [55], we select the widely used 2386 images of
20 classes and 9144 images of 102 classes (101 object categories and an additional background class), respectively, denoted as Caltech-20 and Caltech-101. This dataset provides 6 kinds of pre-extracted features for each image, i.e., 48-D Gabor, 40-D Wavelet moments, 254-D CENTRIST, 1984-D HOG, 512-D GIST, and 928-D LBP.

AWA [39] is composed of 30,475 images of 50 different animals with 6 heterogeneous pre-extracted features for each image. Specifically, they are 2688-D Color Histogram, 2000-D Local Self-Similarity, 252-D Pyramid HOG, 2000-D SIFT, 2000-D color SIFT, and 2000-D SURF.

NUSOBI [56] is a subset of NUS-WIDE and contains 31 object categories and 30,000 images in total. It has 5 types of low-dimensional features extracted from all images, including 64-D color histogram, 225-D block-wise color moments, 144-D color correlogram, 73-D edge direction histogram, and 128-D wavelet texture.

 Reuters [57] is a document dataset collected from 5 different languages. It contains 18,758 documents in total, all of which are uniformly categorized into 6 classes. Note different languages can be seen as different views, that is, English (21,531-D), French (24,892-D), German (34,251-D), Italian (15,506-D) and Spanish (11,547-D).

Hand [38] consists of features of handwritten numerals extracted from a collection of Dutch utility maps, 200 patterns per category (a total of 2000 patterns). These digits are represented in terms of 6 feature sets, containing 76-D Fou, 216-D Fac, 64-D Kar, 240 Pix, 47-Dzer, and 6-D Mor.

2) Implementation Details, Critical Architectures: We choose MvNNcor [28] as our baseline, which is composed of two parts, i.e., a set of neural networks \( \{ f_i \}_{i=1}^{M} \) (i.e., the encoder \( E_i \) in Fig. 7), and an auxiliary module \( \{ f_w \} \). Formally, \( M \) denotes the total number of viewpoints and each \( f_i \) is a fully-connected network consisting of \( d_i \) input units and two hidden layers with 512 and 256 units equipped with ReLU activation function. To implement MV\(^2\)D, we append an information bottleneck architecture to each \( f_i \) referred to Fig. 7, where we omit \( f_w \) for simplicity (detailed descriptions w.r.t our multi-view classification framework can be found in D.1 in the supplementary material).

Training: We follow the same experimental configurations in [28], where all experiments are optimized by Adam with \( \beta_1 = 0.5 \) and \( \beta_2 = 0.9 \). The learning rate is initialized with \( 10^{-3} \) and decays 20 times at 30th and 60th epoch. All networks are trained from scratch with a batch size of 64 and are updated with 160 epochs in total. The training objective includes three terms, i.e., classification loss, rank loss, and \( \text{(18)} \).

3) Experimental Results: Comparison: Tables VII and VIII summarize the quantitative results on multi-view classification. The baseline is developed from MvNNcor with an additional material classification framework can be found in D.1 in the supplementary material).

Analysis on Batch Size: Compared with other generalized representation learning techniques [64] which may require large batch size, MV\(^2\)D prevents such limitation and reaches the optimum within acceptable storage cost (see Fig. 12(d) for illustration).

Sufficiency & Consistency: We also plot the 2D projection of representations by using t-SNE on Fig. 11, where we compare the representations obtained from our approach and conventional IB. By observing the scatter, we have: (i) the embedding space information, which neutralizes sensitivity to view-changes. On the other hand, we also observe our information-theoretic constraint drives the deep models [28], [41], [53] to learn sufficient and consistent representations, by achieving stronger performance without requiring complex designs.

Ablation Study: Based on the Tables VII and VIII, we can draw similar conclusions in the multi-view case: (i) the IB architecture can improve the performance although it introduces additional parameters; (ii) conventional IB strategy still has no benefits in promoting the accuracy under multi-view cases; (iii) MV\(^2\)D can evidently boost the performance on all datasets but excluding Hand [38]. The reason might be the dimension of feature in this dataset is significantly low (e.g., 6-D Mor), which can hardly include rich sources of information. Such phenomenon also reveals the shortcomings of MV\(^2\)D, i.e., incapability to choose the optimal dimension, and becoming mediocre when handling low-dimensional situations.

Analysis on Feature Dimension of IB: As is shown in Fig. 12, the accuracy first climbs to a peak with the increase of output dimension of IB, and then degrades. We deduce there are two reasons accounting for this phenomenon: (i) necessary information would be inevitably discarded if the dimension is extremely reduced, which can be concluded from our inferior performance on Hand dataset; (ii) compact representations are usually beneficial for the downstream tasks.

Analysis on Batch Size: Compared with other generalized representation learning techniques [64] which may require large batch size, MV\(^2\)D prevents such limitation and reaches the optimum within acceptable storage cost (see Fig. 12(d) for illustration).
produced by CIB appears to lack discrimination, where we can spot obvious overlapping within each class and indistinguishable boundaries between different categories; (ii) by contrast, almost all the clusters obtained by MV²D are around a respective centroid, suggesting the sufficiency and view-consistency information are better preserved.

C. LiDAR-RGB Semantic Segmentation

In this section, we further evaluate the variational distillation framework on LiDAR-RGB semantic segmentation, which, in practice, is a typical cross-modal learning problem. It is a fundamental task for scene perception and understanding, which aims to predict a dense label map by fusing complementary information from both LiDAR and RGB sensors. Thus, it is also quite suitable to evaluate the variational distillation framework in such a scalable and complex representation learning problem.

1) Evaluation Protocol and Benchmarks: To evaluate the proposed method, we follow the official protocol [68], [69] to leverage mean intersection-over-union (mIoU) as the evaluation metric. For a given class \( i \), IoU is formulated as: 

\[
\text{IoU}_i = \frac{TP_i}{TP_i + FP_i + FN_i},
\]

where \( TP_i, FP_i, FN_i \) represent true positive, false positive, and false negative predictions for the \( i \)th class and the mIoU is the mean value of IoU over all classes.

nuScenes [68] collects 1000 scenes of 20 s duration with 32 beams LiDAR sensor. The number of total frames is 40,000, and are split into 28,130 training frames and 6,019 validation frames. After merging similar classes and removing rare classes, a total of 16 classes for the LiDAR semantic segmentation are remained. Unlike SemanticKITTI, which provides only the

TABLE IX

| Methods          | mIoU(%) |
|------------------|---------|
| RangNet++ [58]   | 65.5    |
| SPVCNN [59]      | 67.8    |
| PolarNet [60]    | 71.0    |
| Cylinder3D [61]  | 76.1    |
| AF25SN [62]      | 78.3    |
| PMF [63]         | 76.9    |
| ours (baseline)  | 77.2    |
| ours             | 78.5    |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
images of the front-view camera, nuScenes has 6 cameras for different views of LiDAR.

SemanticKITTI [69] is a large-scale driving scene dataset for point cloud segmentation, which provides 43,000 scans with point-wise semantic annotation. This dataset consists of 22 sequences in total, splitting sequences 00 to 10 as the training set (where sequence 08 is used as the validation set), and sequences 11 to 21 as a test set. 19 classes are used for training and evaluation after ignoring and merging the classes with very few points or with different moving statuses.

2) Implementation Details. Critical Architectures: Our framework is mainly composed of two sub-networks to handle different modals (i.e., image and LiDAR), and each of which includes a backbone network (i.e., $E_\theta$) and an information bottleneck to implement MV$^2$D. In addition, we also adopt a LI-fusion module [70] to enhance the LiDAR point representation by incorporating image features at multiple scales. The image sub-network is implemented by SwiftNet [71] pretrained on ImageNet, and extracts the semantic information with a set of convolutional operations. We adopt SPVCNN [59] as the Point Cloud sub-network, which outputs the final representation for segmentation. Details and graphical illustration of our framework can be found in E.1 in the supplementary material.

Training: All experiments are optimized by SGD with Nesterov, where weight decay and momentum are set to $1 \times 10^{-4}$ and 0.9, respectively. The learning rate starts at $2 \times 10^{-1}$ and adopts the warm-up with a cosine scheduler. We train our model for 40 epochs in total with batch size fixed to 8, and we conduct all experiments on NVIDIA RTX A6000 GPUs. In addition to the widely adopted cross-entropy, the training objective also consists of multi-class focal loss [72], Lovász-softmax loss [73] and our MV$^2$D.

3) Experimental Results. Quantitative Analysis: Tables IX and X shows the comparison on the validation set of nuScenes and SemanticKITTI. We can draw the following conclusion: Our approach evidently boosts the performance and outperforms the baseline and other competitors in terms of mIoU on both benchmark datasets. More specifically, the proposed variational distillation framework exceeds the SPVCNN [59] (our point cloud baseline) by a large margin, and it also
provides a visual illustration of the produced labeling map. Obviously, VCD can better facilitate the fusion of complementary information and thus attains preferable segmentation results. By comparison, we observe some categories that are hardly recognized are ignored by the baseline, which shows IB can handle the huge modal-discrepancy.

VI. CONCLUSION

In this work, we provide an analytical solution to fitting mutual information by using variational inference, rather than designing a sophisticated estimator. On this basis, we reformulate the objective of IB and propose a generalized variational distillation framework, which enables us to jointly preserve the sufficiency of representations and get rid of task-irrelevant distractors. Multi-View Variational Distillation (MV²D) and its special cases, i.e., Variational Cross-Distillation (VCD), and Variational Mutual-Distillation (VMD), can produce view-consistent representations among multiple heterogeneous data observations. The future works would include learning an adaptive method to determine the output dimension of IB. Also, broader multi-view applications such as medical and text would be studied.

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in Proc. Int. Conf. Mach. Learn., 2013, pp. 1247–1255.
[2] W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multi-view representation learning,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1083–1092.
[3] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, “Multi-view discriminant analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 1, pp. 188–194, Jan. 2016.
[4] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” in Proc. Allerton Conf. Commun., Control Comput., 1999, pp. 368–377.
[5] S. Huang et al., “Multi-view subspace clustering on topological manifold,” in Proc. Adv. Neural Inf. Process. Syst., 2022, pp. 25883–25894.
[6] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, “On variational bounds of mutual information,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5171–5180.
[7] M. I. Belghazi et al., “Mutual information neural estimation,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 531–540.
[8] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” in Proc. Int. Conf. Learn. Representations, 2017.
[9] Z. Piran, R. Shwartz-Ziv, and N. Tishby, “The dual information bottleneck method,” 2020, arXiv: 2006.04941.
[10] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, “Learning robust representations via multi-view information bottleneck,” in Proc. Int. Conf. Learn. Representations, 2020.
[11] Z. Wan, C. Zhang, P. Zhu, and Q. Hu, “Multi-view information-bottleneck representation learning,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 10085–10092.
[12] Y. Li, M. Yang, and Z. Zhang, “A survey of multi-view representation learning,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 10, pp. 1863–1883, Oct. 2019.
[13] C. Xu, D. Tao, and C. Xu, “A survey on multi-view learning,” 2013, arXiv: 1304.5634.
[14] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, “ Farrell to mutual information: Variational distillation for cross-modal person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 1522–1531.
[15] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in Proc. IEEE Inf. Theory Workshop, 2015, pp. 1–5.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
2000. He joined the Institute of Software, CAS, in 2000. He is currently a professor of machine learning and computer vision. His work interests include image processing, computer vision, machine learning, and pattern recognition.

Zhizhong Zhang received the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2020. He is currently an associate scholar with the School of Computer Science and Software Engineering, East China Normal University. His research interests include image processing, computer vision, machine learning, and pattern recognition.

Cong Wang received the MS degree in Software System Design from University of Science And Technology of China (USTC), in 2019. He is currently a research engineer with Distributed and Parallel Software Laboratory, 2012 Labs, Huawei Technologies. His research interests include distributed parallelism, computer architecture, and computer vision. His work is mainly the development and design of the AI framework MindSpore.

Wensheng Zhang received the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2000. He joined the Institute of Software, CAS, in 2001. He is currently a professor of machine learning and data mining and the director of the Research and Development Department, Institute of Automation, CAS. His research interests include computer vision, pattern recognition, and artificial intelligence.
Yanyun Qu (Member, IEEE) received the PhD degree in pattern recognition and intelligent systems from the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China, in 2006. She is currently a professor in Computer Science Department, in School of Informatics, Xiamen University, Xiamen, China. She has published more than 90 papers in major international journals and conferences. Her current research interests include image processing, computer vision, machine learning, and pattern recognition.

Lizhuang Ma (Member, IEEE) received the BS and PhD degrees from the Zhejiang University, China in 1985 and 1991, respectively. He is now a distinguished professor, with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China and the School of Computer Science and Technology, East China Normal University, China. He was a visiting professor with the Frounhofer IGD, Darmstadt, Germany, in 1998, and a visiting professor with the Center for Advanced Media Technology, Nanyang Technological University, Singapore from 1999 to 2000. His research interests include computer vision, computer aided geometric design, computer graphics, scientific data visualization, computer animation, digital media technology, and theory and applications for computer graphics, CAD/CAM. He serves as the reviewer of IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, CVPR, AAAI, etc.

Zongze Wu (Member, IEEE) received the PhD degree in pattern reorganization and intelligent systems from Xi'an Jiaotong University, in 2005, Xi'an, China. He is currently a professor with the College of Mechatronics and Control Engineering, Shenzhen University, and also with the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China. He is the author and coauthor of 70 SCI-indexed papers, which include IEEE TAC, Automatica, IEEE TCYB, IEEE/CAA Journal of Automatica Sinica, IEEE TIP, IEEE TCSVT, Pattern Recognition, and Signal Processing.

Yuan Xie (Member, IEEE) received the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2013. He is currently a full professor with the School of Computer Science and Technology, East China Normal University, Shanghai, China. His research interests include image processing, computer vision, machine learning and pattern recognition. He has published around 90 papers in major international journals and conferences including the International Journal of Computer Vision, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, TNNLS, TCYB, and NIPS, ICML, CVPR, ECCV, ICCV, etc. He also has served as a reviewer for more than 15 journals and conferences. He received the National Science Fund for Excellent Young Scholars 2022.

Dacheng Tao (Fellow, IEEE) is the Inaugural Director of the JD Explore Academy and a senior vice president of JD.com. He is also an advisor and chief scientist of the digital science institute in the University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and more than 200 publications in prestigious journals and proceedings at leading conferences. He received the 2015 Australian Scopus-Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a fellow of the Australian Academy of Science, AAAS, ACM, and IEEE.