Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-Identification

Xudong Tian[1], Zhizhong Zhang[1](✉), Shaohui Lin[1], Yanyun Qu[2], Yuan Xie[1](✉), Lizhuang Ma[1,3]
[1]East China Normal University, [2]Xiamen University, [3]Shanghai Jiao Tong University
txd51194501066@gmail, zzzhang@cs.ecnu.edu.cn, shaohuilin007@gmail.com,
yyqu@xmu.edu.cn, xieyuan8589@foxmail.com, ma-lz@cs.sjtu.edu.cn

Abstract

The Information Bottleneck (IB) provides an information theoretic principle for representation learning, by retaining all information relevant for predicting label while minimizing the redundancy. Though IB principle has been applied to a wide range of applications, its optimization remains a challenging problem which heavily relies on the accurate estimation of mutual information. In this paper, we present a new strategy, Variational Self-Distillation (VSD), which provides a scalable, flexible and analytic solution to essentially fitting the mutual information but without explicitly estimating it. Under rigorously theoretical guarantee, VSD enables the IB to grasp the intrinsic correlation between representation and label for supervised training. Furthermore, by extending VSD to multi-view learning, we introduce two other strategies, Variational Cross-Distillation (VCD) and Variational Mutual-Learning (VML), which significantly improve the robustness of representation to view-changes by eliminating view-specific and task-irrelevant information. To verify our theoretically grounded strategies, we apply our approaches to cross-modal person Re-ID, and conduct extensive experiments, where the superior performance against state-of-the-art methods are demonstrated. Our intriguing findings highlight the need to rethink the way to estimate mutual information.

1. Introduction

The Information Bottleneck (IB) [35] has made remarkable progress in the development of modern machine perception systems such as computer vision [6], speech processing [21], neuroscience [30] and natural language processing [18]. It is essentially an information-theoretic principle that transforms raw observation into a, typically lower-dimensional, representation and this principle is naturally extended to representation learning or understanding Deep Neural Networks (DNNs) [31, 24, 9].

By fitting mutual information (MI), IB allows the learned representation to preserve complex intrinsic correlation structures over high dimensional data and contain the information relevant with downstream task [35]. However, despite successful applications, there is a significant drawback in conventional IB hindering its further development (*i.e.,* estimation of mutual information).

In practice, mutual information is a fundamental quantity for measuring the statistical dependencies [2] between random variables, but its estimation remains a challenging problem. To address this, traditional approaches [7, 5, 33, 16, 23, 15, 8, 14, 32, 22] mainly resort to non-parametric estimator under very limited problem setting where the probability distributions are known or the variables are discrete [29, 2]. To overcome this constraint, some works [13, 6, 29] adopt the trainable parametric neural estimators involving reparameterization, sampling, estimation of posterior distribution [2], which, unfortunately, practically has very poor scalability. Besides, the estimation of posterior distribution would become intractable when the network is complicated.

Another obvious drawback of the conventional IB is that, the optimization of IB is essentially a tradeoff between having a concise representation and one with good predictive power, which makes it impossible to achieve both high compression and accurate prediction [35, 1, 28, 38]. Consequently, the optimization of conventional IB becomes tricky, and its robustness is also seriously compromised due to the mentioned reasons.

In this paper, we propose a new strategy for the information bottleneck named as Variational Self-Distillation (VSD), which enables us to preserve sufficient task-relevant information while simultaneously discarding task-irrelevant distractors. We should emphasize here that our approach essentially **fits the mutual information but without explicitly estimating it**. To achieve this, we use variational inference to provide a theoretical analysis which obtains an analytical solution to VSD. Different from traditional methods that attempt to develop estimators for mutual information, our method avoids all complicated designs and allows

the network to grasp the intrinsic correlation between the data and label with theoretical guarantee.

Furthermore, by extending VSD to multi-view learning, we propose Variational Cross-Distillation (VCD) and Variational Mutual-Learning (VML), the strategies that improve the robustness of information bottleneck to view-changes. VCD and VML eliminate the view-specific and task-irrelevant information without relying on any strong prior assumptions. More importantly, we implement VSD, VCD and VML in the form of training losses and they can benefit from each other, boosting the performance. As a result, two key characteristics of representation learning (*i.e.,* sufficiency and consistency) are kept by our approach.

To verify our theoretically grounded strategies, we apply our approaches to cross-modal person re-identification[1], a cross-modality pedestrian image matching task. Extensive experiments conducted on the widely adopted benchmark datasets demonstrate the effectiveness, robustness and impressive performance of our approaches against state-of-the-arts methods. Our main contributions are summarized as follows:

- We design a new information bottleneck strategy (VSD) for representation learning. By using variational inference to reconstruct the objective of IB, we can preserve sufficient label information while simultaneously getting rid of task-irrelevant details.
- A scalable, flexible and analytical solution to fitting mutual information is presented through rigorous theoretical analysis, which fundamentally tackle the difficulty of estimation of mutual information.
- We extend our approach to multi-view representation learning, and it significantly improve the robustness to view-changes by eliminating the view-specific and task-irrelevant information.

2. Related Work and Preliminaries

The seminal work is from [35], which introduces the IB principle. On this basis, [1, 6, 27] either reformulate the training objective or extend the IB principle, remarkably facilitating its application.

In contrast to all of the above, our work is the first to provide an analytical solution to fitting the mutual information without estimating it. The proposed VSD can better preserve task-relevant information, while simultaneously getting rid of task-irrelevant nuisances. Furthermore, we extend VSD to multi-view setting and propose VCD and VML with significantly promoted robustness to view-changes.

To better illustrate, we provide a brief review of the IB principle [35] in the context of supervised learning. Given data observations V and labels Y, the goal of representation learning is to obtain an encoding Z which is maximally informative to Y, measured by mutual information:

$$I(Z;Y) = \int p(z,y) \log \frac{p(z,y)}{p(z)p(y)} dzdy. \tag{1}$$

To encourage the encoding process to focus on the label information, IB was proposed to enforce an upper bound I_c to the information flow from the observations V to the encoding Z, by maximizing the following objective:

$$\max I(Z;Y) \; s.t. \; I(Z;V) \leq I_c. \tag{2}$$

Eq. (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 Z to be maximally expressive about Y while being maximally compressive about X by:

$$\mathcal{L}_{IB} = I(Z;V) - \beta I(Z;Y), \tag{3}$$

where β is the Lagrange multiplier. However, it has been shown that it is impossible to achieve both objectives in Eq. (3) practically [6, 1] due to the trade-off optimization between high compression and high mutual information.

More significantly, estimating mutual information in high dimension imposes additional difficulties [26, 2, 29] for optimizing IB. As a consequence, it inevitably introduces irrelevant distractors and discards some predictive cues in the encoding process. Next, we show how we design a new strategy to deal with these issues, and extend it to multi-view representation learning.

3. Method

Let $v \in V$ be 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**) based on the information theory, and design two variational information bottlenecks to keep both characteristics.

To be specific, we propose a Variational Self-Distillation (VSD) approach, which allows the information bottleneck to preserve sufficiency of representation z, where the amount of label information is unchanged after the encoding process. In the design of VSD, we further find it can be extended to multi-view task, and propose Variational Cross-Distillation (VCD) and Variational Mutual-Learning (VML) approaches based on the consistency of representations, both of which are able to eliminate the sensitivity of view-changes and improve the generalization ability for

[1]We don't explicitly distinguish multi-view and multi-modal throughout this paper.

multi-view learning. More importantly, the proposed VSD, VCD and VML can benefit from each other, and essentially fit the mutual information in the high dimension without explicitly estimating it through the theoretical analysis.

3.1. 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:

$$I(z;y) = I(v;y), \tag{4}$$

where v is an observation containing all label information. By factorizing the mutual information between v and z, we can identify two terms [6]:

$$I(v;z) = I(z;y) + I(v;z|y), \tag{5}$$

where $I(z;y)$ denotes the amount of label information maintained in the representation z, and $I(v;z|y)$ represents the irrelevant information encoded in z regarding given task [6], *i.e.*, superfluous information. Thus sufficiency of z for y is formulated as maximizing $I(z;y)$ and simultaneously minimizing $I(v;z|y)$. Based on the data processing inequality $I(z;y) \leq I(v;y)$, we have:

$$I(v;z) \leq I(v;y) + I(v;z|y). \tag{6}$$

The first term in the right of Eq. (6) indicates that preserving sufficiency undergoes two sub-processes: maximizing $I(v;y)$, and forcing $I(z;y)$ to approximate $I(v;y)$.

In this view, sufficiency of z for y is re-formulated as three sub-optimizations: maximizing $I(v;y)$, minimizing $I(v;y)-I(z;y)$ and minimizing $I(v;z|y)$. Obviously, maximizing the first term $I(v;y)$ is strictly consistent with the specific task and the last two terms are equivalent. Hence the optimization is simplified to:

$$\min I(v;y) - I(z;y). \tag{7}$$

However, it is hard to conduct min-max game in Eq. (5), due to the notorious difficulty in estimating mutual information in high dimension, especially when involving latent variable optimization. To deal with this issue, we introduce the following theory:

Theorem 1. *Minimizing Eq. (7) is equivalent to minimizing the subtraction of conditional entropy $H(y|z)$ and $H(y|v)$. That is:*

$$\min I(v;y) - I(z;y) \iff \min H(y|z) - H(y|v),$$

where $H(y|z) := -\int p(z)dz \int p(y|z)\log p(y|z)dy$.

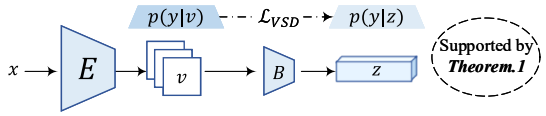


Figure 1: Illustration of VSD, in which E and B denote the encoder and the information bottleneck, respectively.

More specifically, given a sufficient observation v for y, we have following Corollary:

Corollary 1. *If the KL-divergence between the predicted distributions of a sufficient observation v and the representation z equals to 0, then z is sufficient for y as well i.e.,*

$$D_{KL}[\mathbb{P}_v||\mathbb{P}_z] = 0 \implies H(y|z) - H(y|v) = 0,$$

where $\mathbb{P}_z = p(y|z)$, $\mathbb{P}_v = p(y|v)$ represent the predicted distributions, and D_{KL} denotes the KL-divergence.

Detailed proof and formal assertions can be found in supplementary material. As a consequence, sufficiency of z for y could be achieved by the following objective:

$$\mathcal{L}_{VSD} = \min_{\theta,\phi} \mathbb{E}_{v\sim E_\theta(v|x)} \left[\mathbb{E}_{z\sim E_\phi(z|v)} \left[D_{KL}[\mathbb{P}_v \| \mathbb{P}_z]\right]\right], \tag{8}$$

where θ, ϕ stand for the parameters of the encoder and information bottleneck, respectively. On the other hand, based on Eq. (6) and Eq. (5), the minimization of $I(v;y) - I(z;y)$ is equivalent to reducing $I(v;z|y)$, indicating that Eq. (8) also enables IB to eliminate irrelevant distractors. In this perspective, our approach is essentially a self-distillation method that purifies the task-relevant knowledge. More importantly, through using the variational inference, we reformulate the objective of IB, and provide a theoretical analysis which obtains an analytical solution to fitting mutual information in high dimension. Hence we name our strategy as Variational Self-Distillation, *i.e.*, VSD.

Discussion. Compared with other self-distillation methods like [46], one primary advantage of our approach is that VSD is able to retrieve those useful but probably discarded information while simultaneously avoiding task-irrelevant information under theoretically guaranteed. Different from explicitly reducing $I(v;z)$, we iteratively perform VSD to make the representation sufficient for the task. Ideally, when we have $I(v;y) = I(z;y)$, we could achieve the sufficient representation with minimized superfluous information, *i.e.*, optimal representation.

3.2. Variational Cross-Distillation and Variational Mutual-Learning

As more and more real-world data are collected from diverse sources or obtained from different feature extractors, multi-view representation learning has gained increasing attention. In this section, we show VSD could be flexibly extended to multi-view learning.

Consider v_1 and v_2 as two observations of x from different viewpoints. Assuming that both v_1 and v_2 are sufficient for label y, thus any representation z containing all the information accessible from both views would also contain the necessary label information. More importantly, if z only captures the cues that are accessible from both v_1 and v_2, it would eliminate the view-specific details and is robust to view-changes [6].

Motivated by this, we define **consistency** *w.r.t* z_1, z_2 obtained from an information bottleneck as:

Definition 1. Consistency: *z_1 and z_2 are view-consistent if and only if* $I(z_1; y) = I(v_1 v_2; y) = I(z_2; y)$.

Intuitively, z_1 and z_2 are view-consistent only when they possess equivalent amount of predictive information. Analogous to Eq. (5), we first factorize the mutual information between the observation v_1 and representation z_1 to clearly reveal the essence of consistency:

$$I(v_1; z_1) = \underbrace{I(v_1; z_1|v_2)}_{\text{view-specific}} + \underbrace{I(z_1; v_2)}_{\text{consistent}}. \tag{9}$$

Suggested by [6], $I(v_1; z_1|v_2)$ represents that, the information contained in z_1 is unique to v_1 and is not predictable by observing v_2, *i.e.*, view-specific information, and $I(z_1; v_2)$ denotes the information shared by z_1 and v_2, which is named as view-consistent information.

To obtain view-consistent representation with minimal view-specific details, we need to jointly minimize $I(v_1; z_1|v_2)$ and maximize $I(z_1; v_2)$. On the one hand, to reduce view-specific information and note that y is constant, we can use the following equation to approximate the upper bound of $I(v_1; z_1|v_2)$ (proofs could be found in supplementary material).

$$\min_{\theta,\phi} \mathbb{E}_{v_1,v_2\sim E_\theta(v|x)} \mathbb{E}_{z_1,z_2\sim E_\phi(z|v)} \left[D_{KL}[\mathbb{P}_{z_1} \| \mathbb{P}_{z_2}]\right], \tag{10}$$

where $\mathbb{P}_{z_1} = p(y|z_1)$ and $\mathbb{P}_{z_2} = p(y|z_2)$ denote the predicted distributions. Similarly, we have the following objective to reduce $I(v_2; z_2|v_1)$:

$$\min_{\theta,\phi} \mathbb{E}_{v_1,v_2\sim E_\theta(v|x)} \mathbb{E}_{z_1,z_2\sim E_\phi(z|v)} \left[D_{KL}[\mathbb{P}_{z_2} \| \mathbb{P}_{z_1}]\right]. \tag{11}$$

By combining Eq. (10) and Eq. (11), we introduce the following objective to minimize the view-specific information for both z_1 and z_2:

$$\mathcal{L}_{VML} = \min_{\theta,\phi} \mathbb{E}_{v_1,v_2\sim E_\theta(v|x)} \mathbb{E}_{z_1,z_2\sim E_\phi(z|v)} \left[D_{JS}[\mathbb{P}_{z_1} \| \mathbb{P}_{z_2}]\right], \tag{12}$$

in which D_{JS} represents Jensen-Shannon divergence. In practice, Eq. (12) forces z_1, z_2 to learn from each other, thus we name it variational mutual-learning.

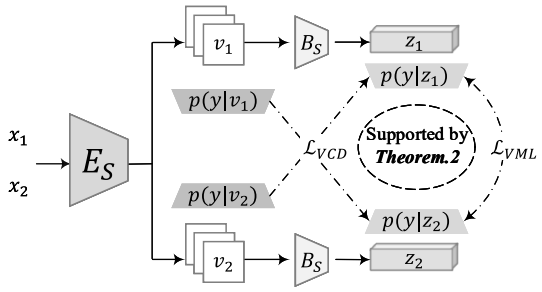


Figure 2: Illustration of $\mathcal{L}_{VCD}$ and $\mathcal{L}_{VML}$, where the subscripts are used to denote different views and E_S, B_S refer to the parameter-shared encoder and information bottleneck.

On the other hand, by using the chain rule to subdivide $I(z_1; v_2)$ into two components [6], we have:

$$I(z_1; v_2) = \underbrace{I(v_2; z_1|y)}_{\text{superfluous}} + \underbrace{I(z_1; y)}_{\text{predictive}}. \tag{13}$$

Combining with Eq. (9), we have:

$$I(v_1; z_1) = \underbrace{I(v_1; z_1|v_2)}_{\text{view-specific}} + \underbrace{I(v_2; z_1|y)}_{\text{superfluous}} + \underbrace{I(z_1; y)}_{\text{predictive}}. \tag{14}$$

Eq. (14) implies that the view-consistent information also includes superfluous information. Therefore, based on the above analysis, we give the following theorem to clarify view-consistency:

Theorem 2. *Given two different sufficient observations v_1, v_2 of an input x, the corresponding representations z_1 and z_2 are view-consistent when the following conditions are satisfied::*

$$I(v_1; z_1|v_2) + I(v_2; z_1|y) \leq 0,$$
$$I(v_2; z_2|v_1) + I(v_1; z_2|y) \leq 0.$$

The proof of Theorem 2 can be found in supplementary material.

According to Theorem 1 and Corollary 1, the following objective can be introduced to promote consistency between z_1 and z_2:

$$\mathcal{L}_{VCD} = \min_{\theta,\phi} \mathbb{E}_{v_1,v_2\sim E_\theta(v|x)} \mathbb{E}_{z_1,z_2\sim E_\phi(z|v)} \left[D_{KL}[\mathbb{P}_{v_2} \| \mathbb{P}_{z_1}]\right], \tag{15}$$

where $\mathbb{P}_{z_1} = p(y|z_1)$ and $\mathbb{P}_{v_2} = p(y|v_2)$ denote the predicted distributions. Based on Theorem 1 and Corollary 1, Eq. (15) enables the representation z_1 to preserve predictive cues, while simultaneously eliminating the superfluous information contained in $I(z_1; v_2)$ (vice versa for z_2 and $I(z_2; v_1)$), and is named as variational cross-distillation.

Discussion. Notice that MIB [6] is also a multi-view information bottleneck approach. However, there are three

fundamental differences between ours and MIB: 1) our strategy essentially fits the mutual information without estimating it through the variational inference. 2) our method does not rely on the strong assumption proposed in [6] that each view provides the same task-relevant information. Instead, we explore both the complementarity and consistency of multiple views for representation learning. 3) MIB is essentially an unsupervised method and hence it keeps all consistent information in various views due to the lack of label supervision. However, with the predictive information, our method is able to discard superfluous information contained in consistent representation, and hence show improved robustness.

3.3. Multi-Modal Person Re-ID

In this section, we show how we apply VSD, VCD and VML to multi-modal learning (*i.e.*, Multi-Modal Person Re-ID). In this context, there are two kinds of images from different modals (*i.e.*, infrared images x_I and visible images x_V). The essential objective of Multi-Modal Person Re-ID is to match the target person from a gallery of images from another modal.

In particular, we use two parallel modal-specific branches equipped with VSD to handle the images from a specific modal. Besides, as shown in Fig. 3, a modal-shared branch trained with VCD and VML is deployed to produce modal-consistent representations. To facilitate Re-ID learning, we also adopt some commonly-used strategies in the Re-ID community. Thus the overall loss is given as:

$$\mathcal{L}_{train} = \mathcal{L}_{ReID} + \beta \cdot (\mathcal{L}_{VSD} + \mathcal{L}_{VCD} + \mathcal{L}_{VML}). \quad (16)$$

More specifically, $\mathcal{L}_{ReID}$ can be further divided into the following terms,

$$\mathcal{L}_{ReID} = \mathcal{L}_{cls} + \mathcal{L}_{metric} + \alpha \cdot \mathcal{L}_{DML}, \quad (17)$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{metric}$, $\mathcal{L}_{DML}$ denote the classification loss with label smooth [34], metric constraint [39] and deep mutual learning loss [47].

4. Experiments

In this section, we conduct a series of experiments to present a comprehensive evaluation of the proposed methods. To promote the culture of reproducible research, source codes (implemented in MindSpore) will be released later.

4.1. Experimental settings

Dataset: The proposed method is evaluated on two benchmarks, *i.e.*, SYSU-MM01 [41] and RegDB [25]. Specifically, SYSU-MM01 is a widely adopted benchmark dataset for infrared-visible person re-identification. It is collected from 6 cameras of both indoor and outdoor environments. It contains 287,628 visible images and 15,792

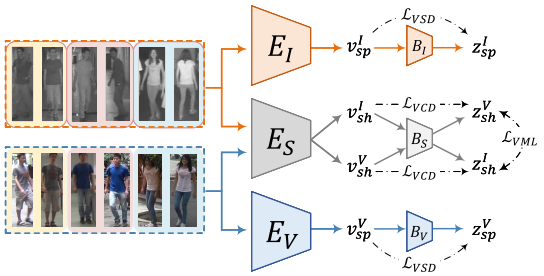


Figure 3: Network architecture for Multi-Modal Re-ID. $E_{I/S/V}$ and $B_{I/S/V}$ represent the encoder (ResNet-50) and information bottleneck (multi-layer perceptrons), respectively. v and z denote the observations and representations from encoder and information bottleneck, respectively

infrared images of 491 different persons in total, each of which is at least captured by two cameras. RegDB 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.

Evaluation Metric: On both datasets, we follow the most popular protocol [44, 19, 17, 37] for evaluation, where cumulative match characteristic (CMC) and mean average precision (mAP) are used. On SYSU-MM01, there are two search modes, *i.e.*, all-search mode and indoor-search mode. For all-search mode, the gallery consists of all RGB images (captured by CAM1, CAM2, CAM4, CAM5) and the probe consists of all infrared images (captured by CAM3, CAM6). For indoor-search mode, the difference is the gallery only contains images captured from indoor scene (excluding CAM4 and CAM5). On both search modes, we evaluate our model under the single-shot setting [44, 19]. On RegDB, following [45], we report the average result by randomly splitting of training and testing set 10 times.

4.2. Implementation Details

Critical Architectures. As visualized in Fig. 3, we deploy three parallel branches, each of which is composed of a ResNet50 backbone [12] and an information bottleneck. Notice that we drop the last fully-connected layer in the backbone and modify the stride as 1 in last block. Following recent multi-modal Re-ID works [19, 3, 17, 44], we adopt the strong baseline [20] with some training tricks, *i.e.*, warm up [10] (linear scheme for first 10 epochs) and label smooth [34]. The information bottleneck is implemented with multi-layer perceptrons of 2 hidden ReLU units of size 1,024 and 512 respectively with an output of size 2×256 that parametrizes mean and variance for the two Gaussian posteriors for the conventional IB principle.

Training. All experiments are optimized by Adam optimizer with an initial learning rate of 2.6×10^{-4}. The learning rate decays 10 times at 200 epochs and we totally train 300 epochs. Horizontal flip and normalization are utilized to augment the training images, where the images are re-

Table 1: Performance of the proposed method compared with state-of-the-arts. Note that all methods are measured by CMC and mAP on SYSU-MM01 under single-shot mode.

Settings			All Search				Indoor Search			
Type	Method	Venue	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Network Design	Zero-Pad [41]	ICCV'17	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
Metric Design	TONE [43]	AAAI'18	12.52	50.72	68.60	14.42	20.82	68.86	84.46	26.38
Metric Design	HCML [43]	AAAI'18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
Metric Design	BDTR [45]	IJCAI'18	17.01	55.43	71.96	19.66	-	-	-	-
Network Design	cmGAN [4]	IJCAI'18	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
Metric Design	D-HSME [11]	AAAI'18	20.68	32.74	77.95	23.12	-	-	-	-
Generative	D^2LR [40]	CVPR'19	28.9	70.6	82.4	29.2	-	-	-	-
Metric Design	MAC [42]	MM''19	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
Generative	AlignGAN [36]	ICCV'19	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
Generative	X-modal [17]	AAAI'20	49.92	89.79	95.96	50.73	-	-	-	-
Generative	JSIA-ReID [37]	AAAI'20	38.1	80.7	89.9	36.9	52.9	43.8	86.2	94.2
Network Design	cm-SSFT [19]	CVPR'20	52.4	-	-	52.1	-	-	-	-
Network Design	Hi-CMD [3]	CVPR'20	34.94	77.58	-	35.94	-	-	-	-
Network Design	DDAG [44]	ECCV'20	**54.75**	**90.39**	**95.81**	**53.02**	**61.02**	**94.06**	**98.41**	**67.98**
Representation	ours	-	**60.02**	**94.18**	**98.14**	**58.80**	**66.05**	**96.59**	**99.38**	**72.98**

Table 2: Comparison with the state-of-the-arts on RegDB dataset under visible-thermal and thermal-visible settings.

Settings		Visible to Thermal		Thermal to Visible	
Method	Venue	Rank-1	mAP	Rank-1	mAP
Zero-Pad [41]	ICCV'17	17.8	18.9	16.6	17.8
HCML [43]	AAAI'18	24.4	20.1	21.7	22.2
BDTR [45]	IJCAI'18	33.6	32.8	32.9	32.0
D-HSME [11]	AAAI'18	50.8	47.0	50.2	46.2
D^2LR [40]	CVPR'19	43.4	44.1	-	-
MAC [42]	MM''19	36.4	37.0	36.2	36.6
AlignGAN [36]	ICCV'19	57.9	53.6	56.3	53.4
X-modal [17]	AAAI'20	62.2	60.2	-	-
JSIA-ReID [37]	AAAI'20	48.5	49.3	48.1	48.9
cm-SSFT [19]	CVPR'20	62.2	63.0	-	-
Hi-CMD [3]	CVPR'20	**70.9**	**66.0**	-	-
DDAG [44]	ECCV'20	69.3	63.5	**68.1**	**61.8**
ours	-	**73.2**	**71.6**	**71.8**	**70.1**

sized to 256×128. The batch size is set to 64 for all experiments, in which it contains 16 different identities, and each identity includes 2 RGB images and 2 IR images. For the hyper-parameters, α and β in Eq. (17) and Eq. (16) are fixed to 8, 2 for all experiments. In order to facilitate the comparison between the conventional IB and the proposed methods, Jensen-Shannon I_{JS} [13, 29, 6, 2] is introduced in the estimation of mutual information, which has shown better performance.

4.3. Comparison with State-of-the-art Methods

In this section, we demonstrate the effectiveness of our approach against state-of-the-art methods. The comparison mainly consists of three categories, *i.e.*, multi-branch network, metric and adversary learning based methods.

As shown in Tab. 1 and Tab. 2, we are able to draw the following conclusions: 1) With recent achievements of generative adversarial network (GAN), generative methods [36, 37, 17, 40], have dominated this community, compared with metric methods [45, 11, 42]. However, despite the promising performance, their success partly attributes to sophisticated designs, (*e.g.*, image/modality generation, auxiliary branches, local feature learning, attention block) and therefore result in better performance. 2) Our approach significantly outperforms all the competitors on both datasets. More importantly, we do not require any modality generation process [40, 36, 37, 17], which implicitly enlarges the training set. Our optimization is quite efficient since the estimation of mutual information are avoided. The complexity comparison would be illustrated in ablation study. 3) Our method is essentially a representation learning method, and is most relevant to the metric methods. However, our method is based on the information theory to extract predictive information for representation learning, rather than the distance between samples. As a result, our method complements with the metric learning and is able to further boost their performance. To our knowledge, this is the first work to explore the informative representation.

4.4. Ablation Study

In this subsection, we conduct ablation study to show the effectiveness of each component in our approach. For a fair comparison, we follow the same architecture to analyze the influence of sufficiency and consistency for cross-modal person Re-ID.

We first clarify various settings in this study. As shown in Tab. 3, "$E_{S/I/V}$" denotes that whether we use model-shared/infrared/visible branch, where we use $\mathcal{L}_{ReID}$ to train each branch. "$B_{S/I/V}$" denotes that in each branch whether we utilize the information bottleneck architecture. "CIBS" denotes we adopt conventional IB for training. "VSD", "VML" and "VCD" denote our approaches. It is noteworthy that the results reported in this table are achieved by using observation v or representation z alone.

Main Results. As illustrated in Tab. 3, we have the following observations:

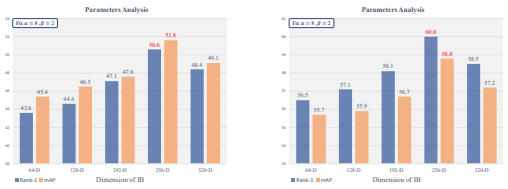
1) Add model-specific branch and concatenate the features could bring obvious performance improvement (see 1st-3rd row in Tab. 3), which achieves 48.82@Rank-1 and 49.95@mAP. This three-branch network can be considered as a new strong baseline.

2) By using information bottleneck, the performance is further boosted (see 4th-6th row in Tab. 3). With such

Table 3: Accuracy of the representation z and observation v when using different training strategies. Note that we evaluate the model on SYSU-MM01 under all-search single-shot mode.

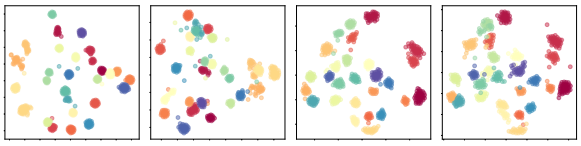
	Settings							Accuracy of v		Accuracy of z	
	E_S	E_I and E_V	B_S	B_I and B_V	CIBS	VSD	VCD+VML	Rank-1	mAP	Rank-1	mAP
1	√	-	-	-	-	-	-	44.95	46.27	-	-
2	-	√	-	-	-	-	-	47.25	48.31	-	-
3	√	√	-	-	-	-	-	48.82	49.95	-	-
4	√	-	√	-	-	-	-	47.20	47.03	38.33	41.81
5	-	√	-	√	-	-	-	53.27	51.92	41.15	43.60
6	√	√	√	√	-	-	-	54.85	53.97	42.04	44.30
7 †	√	-	√	-	√	-	-	8.55	11.01	24.34	28.01
8 †	-	√	-	√	√	-	-	7.76	9.77	28.69	32.42
9 †	√	√	√	√	√	-	-	8.79	11.23	30.43	33.67
10	√	-	√	-	-		√	49.22	48.74	41.48	43.02
11	-	√	-	√	-	√	-	**59.62**	**57.99**	**50.11**	**51.24**
12	√	√	√	√	-	√	√	**60.02**	**58.80**	**50.62**	**51.55**

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

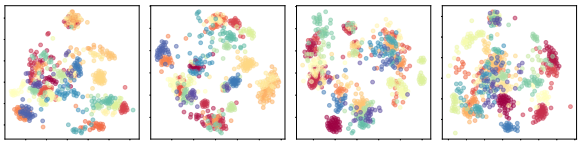


(a) Accuracy of z when the dimension of IB varies. (b) Accuracy of v when the dimension of IB varies.

Figure 4: Evaluation of different dimension of the information bottleneck. The evaluation is conducted on SYSU-MM01 under all-search and single-shot mode.



(a) z^V_{sp} (VSD) (b) z^I_{sp} (VSD) (c) z^V_{sh} (VCD) (d) z^I_{sh} (VCD)



(e) z^V_{sp} (CIB) (f) z^I_{sp} (CIB) (g) z^V_{sh} (CIB) (h) z^I_{sh} (CIB)

Figure 5: 2D Projection of the embedding space by using t-SNE. The results are obtained from our method and conventional IB on SYSU-MM01 test set. Different colors are used to represent different person IDs.

a simple modification, we arrive at 54.85@Rank-1 and 53.97@mAP, which beats all state-of-the-art methods.

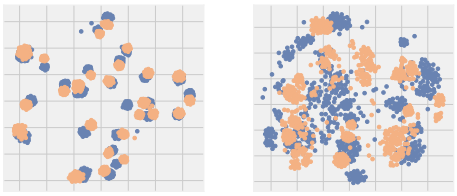
3) It seems that conventional IB has no advantage in promoting the learning of label information (see the 7$^{\text{th}}$-9$^{\text{th}}$ row in Tab. 3). We conjecture it is because that by explicitly reducing $I(v;z)$, conventional IB may not distinguish predictive and superfluous information and probably discard all of them. On the other hand, estimation of mutual information in high dimension appears to be difficult, leading to inaccurate results, especially when involving multi-modal data and latent variables in our setting.

4) Our approach provides remarkable improvement compared with baseline method (see the 10$^{\text{th}}$-12$^{\text{th}}$ row in Tab. 3). By learning predictive cues via the encoding process, it allows both observation and representation to enhance their discriminative ability. However, it is practically impossible to discard all task-irrelevant information without losing any predictive cues, which leads to the performance gap between v and z. In practice, the proposed variational distillation methods could bring about 8.58% and 5.17% increments of Rank-1 accuracy for representation z and observation v, respectively (compare the 12$^{\text{th}}$ with the 6$^{\text{th}}$ row). Furthermore, comparing with the conventional IB (see the 7$^{\text{th}}$-12$^{\text{th}}$ row), a dramatic accuracy boost of 20.19%@Rank-1 and 17.88%@mAP could be observed.

Sufficiency. We also find that the performance gap between z_{sp} and v_{sp} (refer to Fig. 3) is evidently reduced when using our approach (compare 5$^{\text{th}}$ with 11$^{\text{th}}$ row in Tab. 3). This phenomenon shows that sufficiency of the representation could be better preserved by VSD. To approximate the minimal sufficient representation, we evaluate VSD by using different dimensions of IB. As is shown in Fig. 4, we have:

1) When the dimension increases, the accuracy of representation first climbs to a peak, and then degrades. Apparently, if the channel number is extremely reduced, z can not preserve necessary information, while redundant channels are also prone to introduce distractors. Both of them are easy to compromise the accuracy of z.

2) Different dimensions of z affect the accuracy of the observation. We deduce that the encoder may be implicitly forced to concentrate on the information contained in the representation, and other useful cues are somewhat ignored.
3) The performance gap between Z and V decreases along with the increase of the dimension of z. Clearly, the representation is able to hold more useful information with more



(a) z_{sh} (VCD) (b) z_{sh} (CIB)

Figure 6: 2D projection of the joint embedding spaces of z^I_{sh} and z^V_{sh} obtained by using t-SNE on SYSU-MM01.

channels, thus the discrepancy between them is reduced.

We plot the 2D projection of z_{sp} by using t-SNE on Fig. 5, where we compare the representations obtained from our approach and conventional IB. As is illustrated in Fig. 5(e) and Fig. 5(f), the embedding space is mixed in conventional IB, such that the discriminative ability of learned model is naturally limited. On the contrary, our model, which is shown in Fig. 5(a) and Fig. 5(b), is able to distinguish different pedestrians with the proposed VSD loss, hence we get a substantial improvement whether we use z or v.

Consistency: As illustrated in Fig. 5, we project the representations obtained from the modal-shared branch onto two principal components using t-SNE. Compared with conventional IB, we make the following observations:

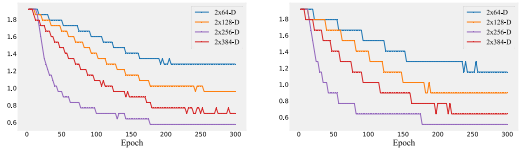
1) From Fig. 5(c) and Fig. 5(d), we observe that the embedding space corresponding to z^I_{sh} and z^V_{sh} appears to coincide with each other, indicating that most of the view-specific information is eliminated by the proposed method. More importantly, almost all the clusters are distinct and concentrate around a respective centroid, which contrasts with Fig. 5(g) and Fig. 5(h). This demonstrates the superiority of our approach to the conventional IB. 2) With the comparison between Fig. 5(g) and Fig. 5(h), we observe that in the embedding space, images of different modals are dramatically discrepant with each other. Such phenomenon is not surprising since conventional IB does not explicitly distinguish view-consistent/specific information.

To better reveal the elimination of view-specific information, we plot the 2D projection of the joint embedding space of z^I_{sh} and z^V_{sh} in Fig. 6, where the orange and blue circles are used to denote the representation of images of two modals, *i.e.*, x_I and x_V, respectively. As illustrated in Fig. 6(a), clusters of the same identity from different modals are distinct and concentrate around a respective centroid, which contrasts to Fig. 6(b). This demonstrates the effectiveness of our approach to improve the robustness to view/modal-changes and reduce the discrepancy across modals.

Complexity: We also compare the extra computational and memory cost brought by our method and conventional IB. As shown in Table 4, "**Enc**" denotes the encoder, *i.e.*, ResNet-50. "**IB**" denotes the information bottleneck architecture. "**MIE**" denotes the mutual information estimator. Note that the IB is essentially a MLP network with three fully-connected layers. Hence it inevitably introduces extra parameters and computational cost. However, since we use the variation inference to avoid the mutual information estimation, this cost is almost negligible (1.09x training time and 3.04M additional parameters), compared with the heavy mutual information estimator.

Method	Enc	IB	MIE	Time	Params
Re-ID Baseline	√			1.0x	26.37M
Ours	√	√		1.09x	29.41M
Conventional IB	√	√	√	1.26x	35.71M

Table 4: Computational cost of different methods.



(a) $\mathcal{L}_{VSD}$ (b) $\mathcal{L}_{VCD}$

Figure 7: Evaluation of the discrepancy between $I(v;y)$ and $I(z;y)$, when the dimension of the representation varies. Note the experiments are conduct on SYSU-MM01 dataset under all-search single-shot mode.

Mutual Information Fitting: Fig. 7 reveals the process of the mutual information fitting between $I(z;y)$ and $I(v;y)$. The amount of predictive information contained in the representation gradually approximates the information contained in observation along the training procedure. Meanwhile, the discrepancy between $I(z;y)$ and $I(v;y)$ varies with the dimension of z. Both phenomenons are consistent with our previous observations.

5. Conclusion

In this work, we provide theoretical analysis and obtain an analytical solution to fitting the mutual information by using variational inference, which fundamentally tackle the historical problem. On this basis, we propose Variational Self-Distillation (VSD) and reformulate the objective of IB, which enables us to jointly preserve sufficiency of the representation and get rid of those task-irrelevant distractors. Furthermore, we extend VSD to multi-view representation learning and propose Variational Cross-Distillation (VCD) and Variational Mutual-Learning (VML) to produce view-consistent representations, which are robust to view-changes and result in superior accuracy.

ACKNOWLEDGMENT. This work is supported by the National Natural Science Foundation of China 61772524, 61876161, 61972157; National Key Research and Development Program of China (No. 2019YFB2103103); Natural Science Foundation of Shanghai (20ZR1417700); CAAI-Huawei Mind-Spore Open Fund; the Research Program of Zhejiang Lab (No.2019KD0AC02).

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. 1, 2

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, 2018. 1, 2, 6

[3] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, 2020. 5, 6

[4] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, 2018. 6

[5] Georges A. Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. In *IEEE Trans. Inf. Theory*, 1999. 1

[6] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020. 1, 2, 3, 4, 5, 6

[7] Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. In *Physical review A*, 1986. 1

[8] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *AISTATS*, 2015. 1

[9] Ziv Goldfeld, Ewout van den Berg, Kristjan H. Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *arXiv:1810.05728*, 2018. 1

[10] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. In *arXiv:1706.02677*, 2017. 5

[11] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, 2019. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[13] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 1, 6

[14] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabás Póczos, Larry A. Wasserman, and James M. Robins. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *NIPS*, 2015. 1

[15] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. In *Physical review E*, 2004. 1

[16] Nojun Kwak and Chong-Ho Choi. Input feature selection by mutual information based on parzen window. In *IEEE Trans. Pattern Anal. Mach. Intell*, 2002. 1

[17] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, 2020. 5, 6

[18] Xiang Lisa Li and Jason Eisner. Specializing word embeddings (for parsing) by information bottleneck. In *IJCAI*, 2020. 1

[19] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, 2020. 5, 6

[20] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, 2019. 5

[21] Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *ICCV*, 2019. 1

[22] Kevin R. Moon, Kumar Sricharan, and Alfred O. Hero III. Ensemble estimation of mutual information. In *ISIT*, 2017. 1

[23] Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. In *Physical Review E*, 1995. 1

[24] Charlie Nash, Nate Kushman, and Christopher K. I. Williams. Inverting supervised representations with autoregressive neural density models. In *arXiv:1806.00400*, 2018. 1

[25] Dat Tien Nguyen, Hyung Gil Hong, Ki-Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. In *Sensors*, 2017. 5

[26] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. In *IEEE T INFORM THEORY*, 2010. 2

[27] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *ICLR*, 2019. 2

[28] Zoe Piran, Ravid Shwartz-Ziv, and Naftali Tishby. The dual information bottleneck. In *arXiv:2006.04641*, 2020. 1

[29] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, , and George Tucker. On variational bounds of mutual information. In *ICML*, 2019. 1, 2, 6

[30] Elad Schneidman, Noam Slonim, Naftali Tishby, Rob R. de Ruyter van Steveninck, and William Bialek. Analyzing neural codes using the information bottleneck method. In *NIPS*, 2001. 1

[31] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. In *arXiv:1703.00810*, 2017. 1

[32] Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *NIPS*, 2016. 1

[33] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *FSDM*, 2008. 1

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the in-

ception architecture for computer vision. In *CVPR*, 2016. 5

[35] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *arXiv:physics/0004057*, 2000. 1, 2

[36] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, , and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, 2019. 6

[37] Guan'an Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zengguang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, 2020. 5, 6

[38] Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. Deep multi-view information bottleneck. In *SDM*, 2019. 1

[39] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil Martin Robertson. Ranked list loss for deep metric learning. In *CVPR*, 2019. 5

[40] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, 2019. 6

[41] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, 2017. 5, 6

[42] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *ACM Multimedia*, 2019. 6

[43] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018. 6

[44] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 2020. 5, 6

[45] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, 2018. 5, 6

[46] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, 2019. 3

[47] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 5