A Survey on Information Bottleneck

Shizhe Hu, *Member, IEEE*, Zhengzheng Lou, *Member, IEEE*, Xiaoqiang Yan, and Yangdong Ye, *Member, IEEE*

(Survey Paper)

***Abstract*—This survey is for the remembrance of one of the creators of the information bottleneck theory, Prof. Naftali Tishby, passing away at the age of 68 on August, 2021. Information bottleneck (IB), a novel information theoretic approach for pattern analysis and representation learning, has gained widespread popularity since its birth in 1999. It provides an elegant balance between data compression and information preservation, and improves its prediction or representation ability accordingly. This survey summarizes both the theoretical progress and practical applications on IB over the past 20-plus years, where its basic theory, optimization, extensive models and task-oriented algorithms are systematically explored. Existing IB methods are roughly divided into two parts: traditional and deep IB, where the former contains the IBs optimized by traditional machine learning analysis techniques without involving any neural networks, and the latter includes the IBs involving the interpretation, optimization and improvement of deep neural works (DNNs). Specifically, based on the technique taxonomy, traditional IBs are further classified into three categories: *Basic*, *Informative* and *Propagating IB*; While the deep IBs, based on the taxonomy of problem settings, contain *Debate: Understanding DNNs with IB*, *Optimizing DNNs Using IB*, and *DNN-based IB* methods. Furthermore, some potential issues deserving future research are discussed. This survey attempts to draw a more complete picture of IB, from which the subsequent studies can benefit.**

***Index Terms*—Information bottleneck, survey, information theory, pattern analysis, representation learning, information compression.**

I. Introduction

LEARNING, essentially, is the ability to obtain compact data representations. In fact, for human, the fundamental goal of learning is trying to find the compact, simple, meaningful understandings about the seemingly chaotic and complicated world. This is also the case for machine learning, particularly the pattern analysis and representation learning. With the arrival of Big Data era, how to find a compact yet effective representations of data in a proper manner has become imperative and challenging. In general, there are two rules required to obey. For one thing, the input data samples should be significantly compressed; For another, the learned data patterns or features, depending on specific tasks, should be effectively discovered.

The canonical rate distortion theory is the exemplar of such information compression. It forces the trade-off between lossy compression and the corresponding distortion, where the former is to seek an optimal compact representation of the source data so that not too much information about the input loses, and the latter is realized by defining a distortion function to describe how distorted the input and the compact data are. However, the distortion function itself has to be defined before, which is the main shortcoming of this theory. By introducing an additional relevant variable, a new information-theoretic paradigm for information compression—Information Bottleneck (IB) theory, which simultaneously considers data compression and information preservation and thus obtains informative representations, was initially proposed by Naftali Tishby, Fernando C. Pereira and William Bialek [1] in 1999.

Although Tishby et al. proposed the IB method and corresponding iterative optimization algorithm, there are still lots of issues needed to be resolved. By following the pioneering work, Tishby's PhD student Noam Slonim conducted more outstanding works containing theoretical analysis and empirical study, including exploration on more effective optimization algorithms, convergence and computational analysis, extension of IB to multivariate version, and investigation on real-world applications, where all of the above were further well organized into his doctoral dissertation—*The Information Bottleneck: Theory and Applications* [2] in 2002.

In fact, IB theory, which formulates an information compression model for learning patterns or representations, has two philosophical connotations—commonness compression and individuality preservation. On one hand, in IB, similar inputs are compressed/mapped into probability distributions of "bottleneck" variables, which commonly correspond with particular semantic explanation, such as the discovered actions of jogging, kicking or throwing in human action recognition. On the other hand, IB distinguishes different input samples and maximally preserves the individuality. Therefore, IB has been successfully applied into numerous practical tasks and applications, especially the pattern recognition field. For the example in document clustering [3], IB performs better than many classic methods, e.g., k-means method, and is even comparable to the supervised Naive Bayes classifier.

Additionally, the balance between compression and preservation will produce discriminative and representative

Manuscript received 8 June 2023; accepted 5 February 2024. Date of publication 15 February 2024; date of current version 2 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62206254 and Grant 62176239 and in part by China Postdoctoral Science Foundation under Grant 2023M743186. Recommended for acceptance by M. Kloft. *(Corresponding author: Yangdong Ye.)*

The authors are with the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China (e-mail: ieshizhehu@zzu.edu.cn; zzlou@zzu.edu.cn; iexqyan@zzu.edu.cn; ieydye@zzu.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2024.3366349

results to some extent. When IB theory is used to explore how DNNs work and explain why the deep learning methods work so well [4], the DNN's layered structure is regarded as a Markov chain of intermediate representations between the input samples layer and output label layer. Then, DNNs can be analyzed by IB where each layer maximally preserves the relevant information with the output label, while minimizing the information with its previous representation layer, so that each layer can be guaranteed to be the most discriminative representation about output and meanwhile the most representative one about the input. In other word, each layer is the approximate minimal sufficient statistics [5] of the input.

Due to the elegant trade-off between information compression and preservation, IB has become a powerful tool in representation learning and pattern analysis, and been widely used in the communities of natural language processing, data mining, pattern recognition, neural optimization, computer vision, signal processing, machine learning and medical analysis. Specifically, the detailed applications of IB contain text classification [6], sentence summarization [7], time series prediction [8], word embeddings specialization [9], topic link detection [10], community detection [11], image classification [12], image clustering [13], object category discovery [14], image segmentation [15], salient region discovery [16], fake digital image identification [17], image retrieval [18], acoustic event detection [19], speech recognition [20], discrete amino acid profile alignment [21], feature detection of fMRI data [22], gene ontology subsets identification [23], sentiment analysis [24], robotic system controlling [25], human action recognition [26], action clustering [27], video search reranking [28], and so forth.

Researchers from different fields, e.g., computer scientists, biologists and mathematicians, have made great endeavors to study IB principle, and contributed a large number of remarkable results since the creation of IB in 1999. Therefore, it is imperative to conduct a systematic survey on IB, thus motivating more promising works in the future. It is observed that there exists several IB surveys [29], [30], [31], but they only focus on specific aspects about IB, failing to investigate it from a systematic and comprehensive perspective. Facing this issue, we, in this survey, systematically explore the IB about its basic theory, optimization, extensive models and task-oriented algorithms over the past 20-plus years. We accordingly propose new taxonomies to well organize these works. Besides, this survey summarizes both the theoretical progress and practical applications on IB, where these works appear in several top conferences and journals, such as CVPR, NeurIPS, ICCV, ICLR, ICML, ECCV, ACM MM, AAAI, IJCAI, ACM SIGKDD, ACM SIGIR, IEEE TPAMI, TIP, TKDE, TNNLS, TCYB, ACM TKDD, Information Fusion and Pattern Recognition. We attempt to draw a more complete picture of IB, from which the subsequent studies can benefit.

Existing IB methods are roughly divided into two parts, as shown in Fig. 1, including:

- *Traditional IB*, which has been developing from 1999, and contains the IB methods optimized by traditional machine learning analysis techniques without involving any neural networks,

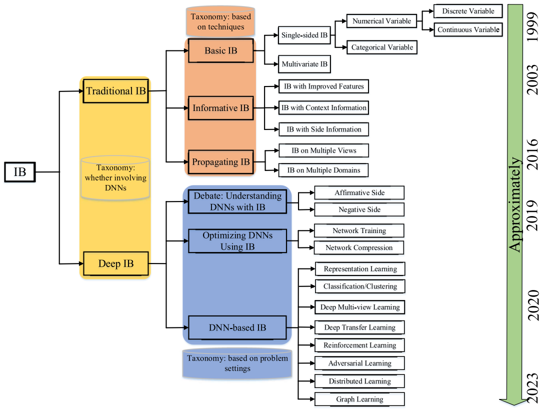


Fig. 1. Organization of IB models and algorithms.

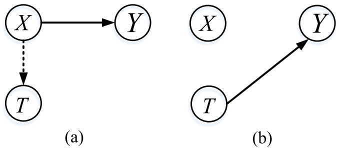


Fig. 2. IB theory. (a) Data compression. The dotted line denotes the compression process from source X to T, while the solid line denotes the joint probability distribution between X and Y. (b) Relevancy preservation. The solid line denotes the relevant information which the compact variable T preserves about Y.

- *Deep IB*, which has been developing from 2015, and includes the IB methods involving the interpretation, optimization and improvement of DNNs.

Traditional IB consists of basic optimization algorithms and extensive models, as well as the pragmatic issue of IB with regard to multi-source heterogeneous data. Based on the technique taxonomy, it is further categorized into three subclasses.

- *Basic IB*, which involves typical single-sided and multivariate IB optimization algorithms;
- *Informative IB*, which contains IB methods with improved features, context information, or side information;
- *Propagating IB*, which explores and propagates the correlated information among multi-source heterogeneous data, e.g., multi-view data or multi-domain data, for learning useful patterns.

With the prosperous development of deep learning, recently researchers are trying to find the intimate connections between DNNs and IB, or combine both of them to enjoy the best of both worlds. Based on the taxonomy of problem settings, deep IBs are further categorized into three subclasses.

- *Debate: Understanding DNNs with IB*, which explores the possibilities of IB theory in understanding the working mechanism of the popular black-box—DNNs;
- *Optimizing DNNs Using IB*, which optimizes the DNNs, such as network training or compression, using IB theory;
- *DNN-based IB*, which involves designing new DNN-based IB methods to improve the learning performance of IB theory on the DNNs.

The remaining parts of this survey are organized as follows. In the next section, we briefly define the basic concept of mutual information and give a general overview about IB theory. Then, we elaborate the details of the traditional and deep IB methods. Afterwards, the relationships or differences of IB with other problems or methods are discussed. Finally, we conclude this survey and discuss some potential open issues deserving future research.

II. The Information Bottleneck Theory

A. *Concepts and Definitions of Information Bottleneck Theory*

In this section, we briefly review the IB theory, where the basic concept of mutual information is first defined.

Definition 1 (MI): Mutual Information (MI) is a kind of quantity measuring the information one random variable contains about another. For any two discrete variables X and Y, the MI between them is defined by

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{1}$$

Similarly, for continuous variables, the MI is defined by

$$I(X;Y) = \int_x \int_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy, \tag{2}$$

where $p(x,y)$ indicates the joint probability distribution, $p(x)$ and $p(y)$ indicates the marginal distributions of variable x and y respectively.

Then, we shortly introduce the Rate Distortion (RD) theory, since IB theory is a special case of it. RD theory provides a general framework for studying optimal lossy compression, which has been widely applied into data compression, pattern classification and feature selection. Given source data with a distortion measure, RD theory forces the trade-off between lossy compression and the corresponding distortion. 1) Lossy compression is to seek an optimal compact representation T of the source data X so that not too much information about X loses, where the compression can be formally defined by a mapping probability distribution $p(t|x)$ from element $x \in X$ to its compact compression $t \in T$. A natural way to measure how much information a t holds is using the above quantity of MI between X and T, i.e., $I(T;X)$. Thus, higher $I(T;X)$ means a less compact representation, while lower value means more compact T about X. 2) Distortion function $d(x,t)$ is user-defined to describe how distorted x and t are, e.g., Hamming distance or squared error. Now, the RD theory is formulated by

$$R(D) = \min_{p(t|x):E(D) \leq D^*} I(X;T), \tag{3}$$

where $E(D) = \sum_{x,t} p(x)p(t|x)d(x,t)$ denotes the expected distortion given the mapping distribution $p(t|x)$, and D^* indicates the threshold value given in advance.

Essentially, defining the distortion function $d(x,t)$ is to find the relevant information in source data X. However, $d(x,t)$ itself has to be defined before, which is difficult to adapt for practical applications and thus becomes the main shortcoming of the above RD theory. Additionally, the threshold value D^* has to be given before, and is usually unknown without any prior knowledge. Rather than a generalization, IB theory can be seen as a special case of RD theory with a special loss function and the Lagrangian formulation of the constrained optimization. The IB theory is proposed by introducing an another relevant variable Y and a trade-off parameter β, as shown in Fig. 2. Specifically, IB maximally compresses the source variable to the compact one while preserving the relevant information about another variable as much as possible, which can be formulated by

$$\mathcal{F}_{\min}[p(t|x)] = I(T;X) - \beta I(T;Y), \tag{4}$$

where X, T and Y denote the source, compact and relevant variables respectively. $I(T;X)$ measures the compression from variable X to T, $I(T;Y)$ measures how much information the compact variable T preserves about Y. β is the trade-off between the compression and preservation. $p(t|x)$ denotes the mapping from x to t.

TABLE I
Relationship Between RD and IB Theory

Theory	Rate Distortion (RD)	Information Bottleneck (IB)
Goal	Under constraint of D^*, obtaining compact T	Maximally preserving information about Y, obtaining compact T
Distance Metric	User-defined	$d(x,t) = D_{KL}[p(y|x)\|\|p(y|t)]$
Limitation	$d(x,t)$ and D^* have to be given before	Relevant variable Y and parameter β have to be given before
Applications	Data communications, pattern recognition, computer vision, data mining and so on	

To make a clear understanding on the relationship between RD and IB theory, we analyze and show them in Table I from different aspects. One uniqueness of IB theory compared to RD theory lies in the conversion of the choice of distortion function to the choice of additional relevant variable. Actually, in many practical scenarios it is easier to select a relevant variable determining what is relevant [1]. For example, it might be the labels for images in image classification, or the relevant features of data in unsupervised feature selection.

B. *Basic Scientific Problems Solved by IB Theory*

Different from other machine learning methods, e.g., non-negative matrix factorization, spectral analysis, or k-means, IB is an information-theoretic learning-based method, which can basically solve the feature learning and pattern recognition problems. For better understanding, we illustrate them with simple specific examples shown in Fig. 3.

Feature learning problem. We elaborate this problem from both supervised and unsupervised view.

1) On the supervised aspect: We further classify it into traditional and deep learning framework from the perspective of optimization manner. *i) Traditional framework:* In Fig. 3(a), the input matrix X is compressed by IB into a compact representation T, which maximally captures the related mutual information about the output label space Y. The typical IB-related works on this problem have large-margin multi-view IB [32] and IB learning using privileged information [33]. *ii) Deep framework:* In Fig. 3(b), IB works in a similar way but optimizes under deep neural networks, such as the deep variational IB [34].

2) On the unsupervised aspect: IB aims to compress the original high-dimensional feature representation (X,Y) into a compact yet discriminative low-dimensional one (X,T) by maintaining the related characteristics with the source data variable X as much as possible. Thus, the formulation of IB

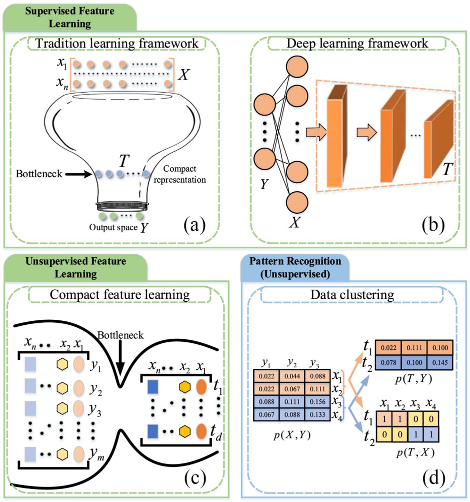


Fig. 3. Basic scientific problems solved by IB theory.

turns to $\mathcal{F}_{\min} = I(T;Y) - \beta I(T;X)$. In Fig. 3(c), given a feature matrix $(X,Y)_{n*m}$ where n indicates the number of samples and m the feature dimensionality, IB compresses the column-wise feature variable Y into T, leading to a descriptive and discriminative feature representation $(X,T)_{n*d}$, such as the interactive IB [35] and multi-task IB co-clustering [36].

Pattern recognition problem: Pattern recognition generally contains supervised classification and unsupervised clustering.

1) On the supervised aspect: It can be realized by combining the above supervised feature learning and a classifier, e.g., SVM or softMax.

2) On the unsupervised aspect: IB seeks to learn a compact grouping $p(t|x)$ by compressing data variable X to T while maximally preserving the relevant information the variable T about Y. In Fig. 3(d), given input data matrix represented by joint probability distribution $p(X,Y)$, where X denotes the discrete source variable and Y the discrete feature variable, IB groups the sample $\{x_1, x_2\}$ into the first cluster t_1, thus the corresponding data assignment $p(t_1|x_1) = p(t_1|x_2) = 1$, and centroid distribution $p(y|t_1) = [\sum_{i=1}^{2} p(y|x_i)]/2$. The remaining two samples are grouped into the second cluster t_2, and the related calculations are similar to the above. The typical IB-related methods on this problem are sequential IB [3] and multi-feature IB [37].

By solving the above basic scientific problems, IB can deal with more complex or downstream problems, such as classification, segmentation, reinforcement learning, transfer learning, distributed learning, and multi-view learning.

C. Our Insights Into the Philosophical Principles of IB Theory

In this section, we provide some insights about the philosophical principles of IB theory in the following aspects.

What is most effective is always simple: Given various kinds of complex input data, IB theory attempts to extract the significant factors from the complicated input by preserving the most discriminative information while eliminating the noisy and redundant information, thus making it much simpler. However, these discriminative information, though simple, is quite effective in many fields, especially feature representation learning and pattern recognition.

What really affect our understanding about the whole world are not the majority but the minority: Massive properties of large amounts of data from the world may not help us to better understand the whole world, and always impose a substantial burden. For instance, people usually remember or recognize objects or places by only using a few key features or properties. Similarly, IB theory tries to compress them into a small amount of compact, representative yet distinguishable information, which contains only the minor key and fundamental properties about the whole input.

The highly-efficient learning is not memorizing knowledge but compact patterns abstraction: Recall that students at younger age usually learn new things or abilities by rote memorization. This may not work and they will forget most of them as time goes by. In contrast, the highly-efficient learning is to abstract the underlying patterns, concepts or rules and then gradually absorb them. Similarly, IB theory learns the compact and relevant patterns of the input data while "forgetting"(i.e., dropping) the irrelevant parts about the data. Thus, the generalization ability of IB is improved.

III. Traditional Information Bottleneck

A. Basic IB

This part mainly contains some classical basic IB models and their improved variants, categorized into single-sided and multivariate IB methods. We introduce several representative models for each category, and give the remaining methods a brief overview in Table II.

1) Single-Sided IB: Single-sided IB methods involve only one source, compact or relevant variable, which play a fundamental and important role in developing other IB models. In the following, we introduce the single-sided IB methods on numerical (discrete and continuous variable) and categorical variable.

IB on discrete variable: There are mainly three basic IB optimization methods with different strategies, including iterative IB (iIB) [1], agglomerative IB (aIB) [38] and sequential IB (sIB) [3], followed with their variants.

1) Iterative IB: iIB adopts alternating iterative optimization technique to update the involved probability distributions. Since the objective function of IB is non-convex, iIB can only converge to locally optimal solution. As a result, the self-consistent equations can be derived as follows:

$$\begin{cases} p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \\ p(y|t) = \frac{1}{p(t)} \sum_x p(x,y)p(t|x) \\ p(t) = \sum_{x,y} p(x,y,t) = \sum_x p(x)p(t|x) \end{cases} \tag{5}$$

where $D_{KL}(A||B) = E_A(\log \frac{A}{B})$ denotes the *Kullback-Leibler* divergence [5], $Z(x,\beta)$ denotes a normalization function. For the case of clustering, these three distributions have specific meanings: $p(t|x)$, $p(y|t)$ and $p(t)$ denote the cluster mapping distribution of X to T, the clustering centroid and the clustering marginal distribution, respectively.

2) Agglomerative IB: In some situations, we may prefer to discover a hierarchical solutions of input samples. To this

TABLE II
SUMMARY OF SOME BASIC IB METHODS

Category	Method	Variable Type	Contribution
Single-sided IB	AsIB [43]	Discrete	Adapting minimum description length principle to sIB for automatically determining the number of clusters
	AisIB [44]	Discrete	An improved sIB method based on simulated annealing strategy to promote the clustering efficiency
	DSIB [45]	Discrete	A data selection model based on sIB method to alleviate the influence of unclear data patterns for clustering
	SmGIB [46]	Continuous	A sparse meta-Gaussian IB to tackle the dimensionality reduction problem
	CDsIB [47]	Categorical	An improved sIB by extending data attributes and conducting binarization for categorical data analysis
	Distributed SIB [48]	Categorical	A distributed version of [49] to solve the large-scale categorical data clustering
Multivariate IB	Mf-MIB [50]	-	A novel multivariate IB method for unsupervised video categorization
	D-MIB [51]		A distributed MIB for adapting more complex parallel related compression systems

end, aIB [38] was designed to greedily perform agglomerative clustering from $|X|$ data clusters to one cluster in a hierarchical bottom-up manner. Specifically, it begins with the initialization of a fine-grained data partition $T = X$, i.e., each input in X is mapped into a single cluster. Then, pairwise values were iteratively merged into a new single value with a minimal cost until no values can be merged, i.e., $|T| = 1$. Thus, the final resulting hierarchical tree showed a set of different clustering solutions at different "layers". aIB has been successfully applied into human action recognition [26]. However, the time complexity tends to $O(|X|^3|Y|)$, which may be impractical for large datasets.

3) Sequential IB: To reduce the complexity of aIB, sIB [3] was proposed by sequentially merging each data point to the best cluster with minimal cost until it converges to a stable solution, which was quite similar to the optimization process of the popular k-means. Specifically, the input value in X is first randomly initialized to the real number of data labels, denoted by c. Then, for each step, each data point x is sequentially drawn from its original cluster and become a new singleton cluster $\{x\}$. To make the number of clusters c unchanged, $\{x\}$ is now merged into a new cluster with minimal cost. After several iterations, e.g., r, the merger process stops and the sequential procedure converges to a stable solution. For the "draw and merge" process, the merger cost has to be computed, and the time complexity takes $O(|T||Y|)$. Thus, the overall complexity takes $O(r|X||T||Y|)$. Since $r|T| \ll |X|^2$, the complexity of sIB has improved a lot compared to that of aIB.

4) Variants of aIB and sIB: To improve the efficiency of aIB and sIB methods, one may develop new strategies to decrease the cost of computation for adapting large datasets. *i)* For aIB, Fulkerson et al. [39] and Wang et al. [40] designed the fast aIB and fast approximate aIB methods for dictionary learning and distributional word clustering respectively, and achieved promising performance. *ii)* For sIB, a finite sIB [41] method was presented to adapt sIB for finite data, e.g., large-scale sparse dataset. Additionally, Yuan and Ye [42] proposed a mutation-based iterative sIB method, which first randomly chose a subset of cluster labels obtained by sIB and performed mutation on them, and then iteratively optimized the results by genetic algorithms. Although the preceding variants have improved a lot, they can only be applied into solving problems with discrete variables.

IB on continuous variable: For the continuous IB problem, Chechik et al. [52] proposed a Gaussian IB method. Specifically, given two joint Gaussian variables (X, Y) with dimensionality n_x and n_y, we denote $\sum_x$ and $\sum_y$ the covariance matrix of X and Y respectively, and $\sum_{xy}$ the cross-covariance matrix. It aimed to compress X to the compact variable $T \in R^{n_x}$ by a stochastic transformation, and meanwhile maintained information about Y as much as possible. By presenting the noisy linear transformation from variables X to T with joint Gaussian distribution as $T = AX + \xi$ where $\xi \sim N(0, \Sigma_\xi)$ was a Gaussian independent of the variable X, we formulate it as

$$\min_{A,\sum_\xi} \mathcal{F} = I(T;X) - \beta I(T;Y), \tag{6}$$

where Σ_ξ indicates the noise covariance, and T follows normal distribution $T \sim N(0, \sum_t)$ with $\sum_t = A\sum_x A^T + \sum_\xi$.

Although Gaussian IB has been successfully applied into practical fields, e.g., speaker recognition [53], it has its shortcoming of estimating the joint covariance matrix of (X, Y) when the marginal probability distributions of both high-dimensional variables are not independent. To address this issue, Rey and Roth [54] introduced a copula function to discover the dependency structure between the marginal distribution of random variables and then proposed a meta-Gaussian IB to obtain the optimal copula of variables X and T, given the copula of X and Y.

IB on categorical variable: Given numerical variables, e.g., discrete or continuous variable, lots of well-defined distance or similarity measures can be adopted. However, it would be quite difficult when confronted with categorical data, e.g., movie data with descriptive attributes of actor, director and genre. To overcome this issue, Andritsos et al. [49] proposed a scalable IB method, especially for categorical clustering problem, which adopted mutual information as distance measure for categorical tuples and attributes values, and then recast data clustering as the compression from one variable to another compact one, while preserving maximal information between the data clusters and attribute values.

Analysis: The above single-sided IB methods have provided elegant basic optimization framework for many follow-up works. However, one of the major disadvantages is that these methods can not be directly adopted for solving the IB models in complex (e.g., multivariate) scenarios.

2) Multivariate IB: Multivariate information bottleneck (MIB) [55] is a general extended version of the above single-sided IB theory to multivariate settings, i.e., involving multiple source, compact or relevant variables. MIB has remedied the major limitation of single-sided IB and has been successfully applied into many fields, such as unsupervised human action video categorization [50] and alternative clustering [56]. In the following we give the detailed information on MIB, followed by some typical variants.

To measure the amount of information among multiple variables, MIB utilizes a new concept of multi-information, which is an extension of the classical concept of mutual information and is formulated by

$$\mathcal{I}(X_1,\ldots,X_n) = D_{KL}[P(X_1,\ldots,X_n) \parallel P(X_i)\cdots P(X_n)], \tag{7}$$

where $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ denote a set of random variables. The above multi-information defines the similarity between distribution $P(X_1, \ldots, X_n)$ and its factored marginal distributions.

To further organize different involved variables, MIB resorts to the popular Bayesian network, a directed A-cyclic graph G where $\mathbf{X}$ denotes the vertices and $\mathbf{Pa}^G_{X_i}$ the parents of each X_i. By assuming that the above distribution satisfies $P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i|\mathbf{Pa}^G_{X_i})$ so that the distribution P is consistent with G ($P \models G$), we rewrite the above multi-information as

$$\mathcal{I}^G \equiv \mathcal{I}(X_1, \ldots, X_n) = \sum_i I(X_i; \mathbf{Pa}^G_{X_i}). \tag{8}$$

In (8), it clearly says that the multi-information among variables equals to the sum of mutual information between each X_i and its parent. Based on the above definition, MIB theory is formulated by two Bayesian networks, G_{in} and G_{out}. G_{in} denotes the data compression from a set of source variables $\mathbf{Pa}^{G_{\text{in}}}_{T_i} \subseteq \mathbf{X}$ to each compact variable in $\mathbf{T} = \{T_1, \ldots, T_m\}$, indicating different partitions of different subsets of $\mathbf{X}$; G_{out} denotes the relevant knowledge to be maintained. MIB attempts to minimize the information in G_{in} while maximizing the information contained in G_{out}, leading to the following formulation

$$\begin{aligned}\mathcal{F}_{\max}&\left[P\left(T_1|\mathbf{Pa}^{G_{\text{in}}}_{T_1}\right), \ldots, P\left(T_m|\mathbf{Pa}^{G_{\text{in}}}_{T_m}\right)\right]\\ &= \mathcal{I}^{G_{\text{out}}} - \beta^{-1}\mathcal{I}^{G_{\text{in}}},\end{aligned} \tag{9}$$

where $\beta \in (0, +\infty)$ denotes the Lagrange multiplier trading off the data compression in G_{in} and the relevancy preservation in G_{out}, and $P(T_i|\mathbf{Pa}^{G_{\text{in}}}_{T_i})$ denotes the partition of $\mathbf{X}$.

Similar to agglomerative IB [38], Slonim et al. [57] provided a general agglomerative MIB framework for solving the optimization problem, which considered multiple hierarchical systems of solutions and exhibited promising performance in text clustering and gene expression analysis. Note that, by setting $\mathbf{X} = \{X, Y\}$ and $\mathbf{T} = \{T\}$, the multivariate IB degrades into the simple form of the single-sided IB principle, i.e., $\mathcal{F}_{\max}[p(t|x)] = I(T;Y) - \beta^{-1}I(T;X)$, as shown in Fig. 2.

Analysis: MIB and its variants benefit from the close collaboration and proper balance of multiple variables. However, such dynamic mechanism may bring inefficient computation, especially on the high-dimensional noisy features. In addition, how to ensure the specific form of G_{in} and G_{out} is quite tough. For instance, in $\mathbf{T} = \{T_1, \ldots, T_m\}$, how to set the value of m is a model selection problem and definitely difficult.

B. Informative IB

As discussed in the above section, Basic IB can only utilize the original input information, failing to integrate more effective and informative knowledge, e.g., context or side information. To mitigate this issue, it is important to introduce an additional informative term on T and/or Y to improve the learning ability.

The informative IB models aim to learn improved features or incorporate additional information into IB framework, which can be unified using the general objective function

$$\mathcal{F}_{\max} = -\beta^{-1}I(T;X) + \Phi(T, Y), \tag{10}$$

where $\Phi(T, Y)$ is the informative term to learn or employ informative features for accurate pattern recognition accuracy, β is the balance parameter trading off the data compression and information preservation.

According to different forms of the term $\Phi(T, Y)$, Informative IB algorithms are divided into three subclasses: 1) IB with improved features; 2) IB with context information; 3) IB with side information.

1) IB With Improved Features: Feature representation is significant for applying IB on practical tasks, e.g., classification, and thus high-quality features can naturally improve the model performance, especially the high-dimensional data clustering [35]. The goal of his line of works is to learn discriminative features to improve IB framework, and they are roughly categorized into the word clusters [6], [58], [59] and co-clustering [60], [61] improved methods.

Word clusters: Slonim and Tishby [6], [58] addressed the document clustering [58] and classification [6] problem by first clustering the words feature Y to word clusters $\tilde{Y}$ and then cluster the input documents with the "word clusters" representation, where the clustering was implemented by the agglomerative IB [38]. This method is called "double clustering", and the primary advantage of doing so is that more compact or discriminative features $\tilde{Y}$ can be extracted from the original high-dimensional sparse or noisy features Y, thus benefiting the subsequent clustering. Following this work, an extended iterative version of it [59] was naturally presented for further improvement.

Co-clustering: Essentially, the above methods work in a sequential manner, failing to fully preserving the close relationships among data, features, and their clusters. Wang et al. [60] and Liu et al. [61] proposed the IB co-clustering method and its possibilistic fuzzy version respectively, which simultaneously performed the row-wise data clustering and column-wise feature clustering, and worked in an alternate, mutually beneficial and win-win fashion. For instance, the objective function of [60] can be formulated as follows

$$\begin{aligned}\mathcal{F}_{\max}[p(\tilde{x}|x), p(\tilde{y}|y)] &= [I(\tilde{X};Y) + I(\tilde{Y};X) + I(\tilde{X};\tilde{Y})]\\ &\quad - \beta^{-1}[I(\tilde{X};X) + I(\tilde{Y};Y)].\end{aligned} \tag{11}$$

where $\tilde{X}$ and $\tilde{Y}$ denote the compact variable of data variable X and feature variable Y respectively, $p(\tilde{x}|x)$ and $p(\tilde{y}|y)$ denote the data and feature partition respectively. Compared to the related information-theoretic co-clustering method [62], these two IB co-clustering methods compress the data and feature into resulting clusters while preserving more relevant information, so that more accurate clustering results can be achieved.

Analysis: The major advantage of these IB methods lies in the learning of discriminative features from original input features, thus benefiting the subsequent tasks. However, they still suffer from two limitations: 1) Dimensionality of the learned improved features is difficult to determine without given any prior knowledge, which remains an open research question. 2) The informativeness is ensured by only its own features, while the additional auxiliary information, e.g., context, is ignored.

2) IB With Context Information: Context information, e.g., density [63], [64], [65] or local consistency [66], [67], often reflects the local characteristic of data, and thus plays an important role in improving the learning performance. We here show some typical ones in two categories.

Density information: Instead of computing the information loss between pairwise instances, Ren et al. [63], [64] introduced the concept of density-based chains to calculate the information loss among neighbors of each input distance, and then incorporated it into the basic agglomerative IB to achieve a hierarchical clustering solution. Further, an extension of the above method to a sequential optimization version [65] was designed and worked more efficiently.

Local consistency: To deal with the image clustering problem, Hu et al. [67] presented a content-context IB method to integrate both intrinsic content information, e.g., appearance feature, and rich context information, e.g., cross-image distance or similarity, for accurate image clusters discovery. The objective function is shown as follows

$$\mathcal{F}_{\max} = -\beta^{-1} I(T;X) + [\alpha I(T;Y) + (1-\alpha) I(T;\hat{Y})], \tag{12}$$

where $\beta \in (0, +\infty)$ trades off the data compression and information preservation, and $\alpha \in [0, 1]$ balances the contribution of the content (Y) and context ($\hat{Y}$) information.

Analysis: The success of above IB methods fully verifies the valuable contribution of context information in improving the IB learning ability. In the future, more types of context are worthy of investigation and exploration, such as the semantic, space or scale context in computer vision.

3) IB With Side Information: Side information generally contains two kinds of meanings, including negative irrelevant information [68], [69], [70] and helpful auxiliary knowledge [71]. Incorporating these side information into IB methods has been proven to significantly improve the performance of some target-oriented tasks, such as feature dimensionality reduction and data classification/clustering. Here we elaborate on some representative works.

Negative irrelevant information: Chechik and Tishby [68] tried to extract the relevant discriminative aspects from the complex data with multiple conflicting structures by designing an IB with side information method for text categorization. It compressed the input variable X to a compact one T while maximally preserving positive information with relevant variable Y^+ and removing the negative information about irrelevant variable Y^-. Following [68], [69], a conditional IB [70] was presented to maximally maintain the conditional mutual information with the relevant variable when given side information in advance, and its objective function is formulated by

$$\mathcal{F}_{\max} = -\beta^{-1} I(T;X) + I(T;Y^+|Y^-). \tag{13}$$

The authors also successfully applied the above method to the field of information retrieval and document mining so that non-redundant clustering results can be found when given a known classification as side information.

Helpful auxiliary knowledge: Smieja and Geiger [71] regarded pairwise constraints provided by user, i.e., must-link and cannot-link where two points are constrained in the same or different clusters, as partition-level side information and then proposed a cross-entropy clustering method with IB constraint for semi-supervised clustering. One of the main strengths is that it does not give the number of clusters as input and all clusters in the partition-level side information.

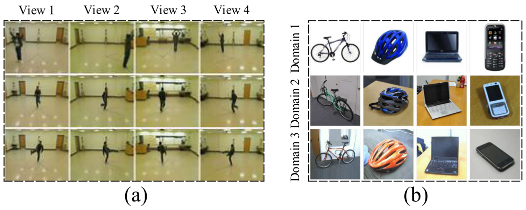


Fig. 4. Examples of multi-source heterogeneous data. (a) Multi-view human action videos from the WVU action recognition dataset[1] where waving hand, jogging, and kicking actions in three rows are captured across four views. (b) Multi-domain images from the Office-31 dataset[2] where the similar objects captured from three different domains, i.e., Amazon, DSLR, and Webcam.

Analysis: Compared to the previous two parts (i.e., subsection III-B1 and III-B2) focusing on learning more positive information, the difference is that IB methods in this part also consider additional negative irrelevant information. However, all of the above methods are still limited to single-source data processing, leading to poor generalization to more complex scenario.

C. Propagating IB

Different from the above informative IB methods that can only handle the single-source data with the same data distribution, propagating IB attempts to learn the underlying data patterns by exploring and propagating the correlations across multiple sources, e.g., views or domains. It well handles the complex multi-source heterogeneous data (typical examples in Fig. 4) in a supervised or unsupervised manner, particularly classification or clustering respectively.

In the past decade, many emerging propagating IB methods have been successfully applied into multi-view applications such as multi-language document clustering, multi-feature image classification or multi-view human action video learning, and also the multi-domain applications such as multi-task classification where the multi-domain data [13] describe the data from multiple domains with different data distributions but similar feature space. We formally define the propagating IB as follows.

Definition 2 (Propagating IB): Given multi-source heterogeneous data $\mathcal{X}$ from m sources, denoted by variable set $\mathbf{X}$, we denote its feature space $\mathcal{Y}$ by the relevant variable set $\mathbf{Y} = \{Y^i\}_{i=1}^m$ and its label space $\mathcal{L}$ by the label set $\mathbf{L}$. Then, m corresponding joint probability distributions $\{p(\mathbf{X}, Y^i)\}_{i=1}^m$ is constructed by pre-processing models, e.g., the popular Bag-of-Words model. Propagating IB aims to extract the compact patterns $\mathbf{T}$ from the input data $\mathbf{X}$. On one hand, for the intra-source, $\mathbf{T}$ maximally preserves the relevant information with variable set $\mathbf{Y}$ or the label set $\mathbf{L}$. On the other hand, for the cross-source, correlated information from the feature level $\mathbf{Y}$ or the pattern level $\mathbf{T}$ is explored and propagated among sources so that more accurate data patterns can be discovered.

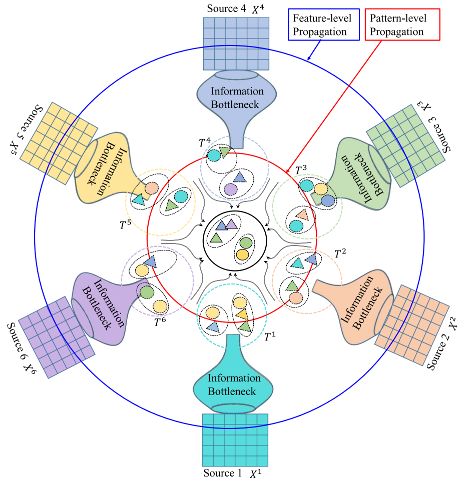


Fig. 5. Illustration of the propagating IB framework. For multi-domain data, six partition results $\{T^i\}_{i=1}^6$ for each source are obtained with the feature-level or pattern-level propagation or both. For multi-view data, the final consistent partition result (denoted by the core black circle) is reached by combining the six partition results with the feature-level or pattern-level propagation or both.

Note that, given multi-view data, the variable sets $\mathbf{X}$, $\mathbf{L}$ and $\mathbf{T}$ only have one element respectively, i.e., $\mathbf{X} = X$, $\mathbf{L} = L$ and $\mathbf{T} = T$; While given multi-domain data, the variable sets $\mathbf{X}$, $\mathbf{L}$ and $\mathbf{T}$ have m elements respectively, i.e., $\mathbf{X} = \{X^i\}_{i=1}^m$, $\mathbf{L} = \{L^i\}_{i=1}^m$ and $\mathbf{T} = \{T^i\}_{i=1}^m$. To further show the propagating IB on both kinds of datasets, we give an illustration on Fig. 5.

Based on the above definition of propagating IB, we further give the unified objective function for supervised learning:

$$F_{\max} = F_{intra}(\mathbf{X}, \mathbf{L}, \mathbf{T}) + F_{inter}\left(\mathbf{X} \bigotimes \mathbf{T}\right), \tag{14}$$

where $\mathbf{X}$ denotes the input multi-source heterogeneous feature variable, $\mathbf{L}$ denotes the training data labels, $\mathbf{T}$ denotes the data partitions, F_{intra} denotes the intra-source objective function, and F_{inter} the cross-source one. $\bigotimes$ denotes "or" or "and", which means the correlation exploration is feature-level or pattern-level or both.

The unified objective function of propagating IB for unsupervised learning is defined as follows:

$$F_{\max} = F_{intra}(\mathbf{X}, \mathbf{Y}, \mathbf{T}) + F_{inter}\left(\mathbf{Y} \bigotimes \mathbf{T}\right), \tag{15}$$

where $\mathbf{X}$ denotes the input multi-source heterogeneous data variable, $\mathbf{Y}$ denotes the feature representation, $\mathbf{T}$ denotes the data partitions, F_{intra} denotes the intra-source objective function, and F_{inter} the cross-source one. $\bigotimes$ denotes "or" or "and", which means the correlation exploration is feature-level or pattern-level or both.

According to different formula of F_{intra} and F_{inter}, propagating IB methods are categorized into the following subclasses and the differences are:

- IB on multiple views (Multi-view propagating IB): 1) focus on multi-view data where the same sample or person is depicted by different views; 2) propagate the correlations among views to learn a good consistent classification or clustering result.
- IB on multiple domains (Multi-task propagating IB): 1) handle more challenging multi-domain data where similar but different samples are captured from various domains; 2) learn a classification or clustering result for each domain on which the learning tasks mutually benefit each other to improve individual task performance.

1) IB on Multiple Views: In this case, multi-view propagating IB methods are further classified into the following three categories in terms of the typical problems or applications. In each part, we present some representative works in detail and summarize the remaining methods in Table III.

Typical multi-view learning: LMIB [32] aims to solve multi-view classification by learning a shared feature subspace among views and then presenting a margin maximization approach to maintain enough discrimination for data classification, so that the related information can be propagated among views, and the accuracy and complexity of the model were properly balanced. The detailed objective function is described as follows:

$$\min \mathcal{F} = \sum_{i}^{m} I(X^i; T) + \frac{C_1}{2}||w||^2 + \frac{C_2}{n}\sum_{j}^{n} h_j, \tag{16}$$

where T and X^i denote the compact shared representation and the input matrix of the i-th view respectively, $C_1, C_2 \geq 0$ are two constants, w and h_j are the weight vector and the hinge loss respectively.

However, supervised learning always involves tagging for numerous training data samples, which is time-consuming and labor-expensive. Hence, it is imperative to design unsupervised IB methods for multi-view clustering. Hu et al. [81] defined a discrimination-compression rate to learn each view weight and then applied them on the discriminative feature representation of each view obtained by the proposed multi-view IB co-clustering method. However, it can only discover the feature relation among views, ignoring the underlying cluster-level view correlations. To fix it, a DMIB [82] method is designed to further explore both the feature-level and cluster-level correlations among views by learning a shared feature space and the shared mutual information among the clustering result of each view respectively.

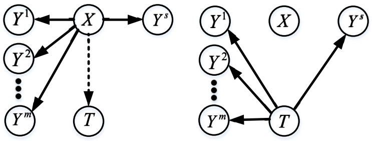
Multi-view text/image/video learning: To date, lots of multi-view IB methods have been designed and successfully applied into many practical fields.

1) Text: Gao et al. [83] proposed a multi-view IB method for web document clustering by introducing a compatible constraint to maximize the clustering agreement of multiple views. One shortcoming of this method is that the abundant complementary information among views is ignored. To mitigate this, a parallel IB [84] is presented for multilingual document clustering, with rich complementary language information maintained.

2) Image: Unsupervised multi-view image categorization has recently attracted much attention in the computer vision community, and we here show some typical methods devoted to this area. Lou et al. [37] fully employed the complementary information of multi-feature images in a weighted manner, where each kind of visual feature is regarded as a view. The detailed objective function is $\mathcal{F}_{\max} = \sum_i^m w^i[I(T; Y^i) - \beta^{-1} I(T; X)]$, where w^i denotes the i-th view weight. Unfortunately, the weights are given in advance, which are difficult to be determined without any prior knowledge. To solve this issue and meanwhile consider

TABLE III
SUMMARY OF SOME MULTI-VIEW PROPAGATING IB METHODS

Category	Method	Learning	Application	Contribution
Typical Multi-view Learning	MEIB [72]	Supervised	-	A multi-view IB method optimized by matrix-based Renyi's α-order entropy
	EMvIB [73]	Supervised		Two efficient multi-view IB methods for reducing the complexity
	NRMVIB [74]	Unsupervised		Non-redundant multi-view IB by preserving relevant and partition information
	DWMVIB [75]	Unsupervised		Learning both content and context representations with dual-weighted multi-view IB
	PMIB [76]	Unsupervised		A cross-view weighted parameter-free multi-view IB clustering method
Multi-view Text/Image/Video Learning	CMIB [77]	Unsupervised	Image	Exploring the correlations in visual contexts of images by contextual multivariate IB
	MVC2IB [78]	Unsupervised	Image	Two auto-weighted multi-view IB methods for Image clustering
	SPIB [79]	Unsupervised	Image	Clustering by integrating intra-view private and cross-view shared information
	HDTC [80]	Unsupervised	Image	Learning visual-textual information for heterogeneous dual-task clustering
	MTIBCC [36]	Unsupervised	Video	Multi-task IB co-clustering method for unsupervised multi-view action categorization



(a) Information compression (b) Information preservation

Fig. 6. Joint IB method. (a) The solid lines denote the joint distributions between X and m view-specific feature variables $\{Y^i\}_{i=1}^m$ and one view-correlated feature variable Y^s. The dotted line denotes the compression from input variable X to compact variable T. (b) The solid lines denote the compact variable T jointly maintains the information about view-specific and view-correlated variables.

the cluster importance in each view, a CWMVIB method [85] is designed for further improving the clustering quality.

3) Video: To handle the complex visual recognition problem, Motiian et al. [33] extended the IB method to learn a multi-view visual classifier for the testing data without auxiliary data view. It has exhibited promising potential on various visual datasets, but can be only suitable for two views, i.e., main and auxiliary view. The authors in [86] considered the more general multi-view video clustering by presenting a Bag-of-Shared-Words model to find the shared visual words among different human action views to propagate the view-correlated information, which was further integrated with the view-specific information by a proposed joint IB method (as shown in Fig. 6) to boost the action clustering quality. The objective function is formulated as follows:

$$\mathcal{F} = \sum_{i}^{m}[I(Y^i;T) - \beta^{-1}I(X;T)] + \gamma I(Y^s;T), \tag{17}$$

where Y^s denotes the shared feature representation, γ denotes the trade-off between the view-specific and view-correlated information.

Joint multi-view and ensemble clustering: To overcome the limitations of multi-view and ensemble clustering, a synergetic IB [87], [88] is presented for simultaneous running of them by fully utilizing the various features and the basic clusterings. Further, a MI-based measurement is designed to discover the correlations between the features and clusterings, which is claimed to generally benefit the related fields, e.g., transfer learning [88].

Analysis: Although the above IB methods have exhibited remarkable performance in multi-view learning, they are limited to the setting where samples in different views are exactly the same and the one-to-one alignment is strictly required. The aim of further managing broader situations may benefit from the following multi-task propagating IB methods.

2) IB on Multiple Domains: In Big Data era, multi-domain data [89] have been increasingly active on the Internet. This motivates the development of multi-task propagating IB methods in recent years, which have been successfully applied into lots of downstream tasks, e.g., video retrieval and segmentation [90].

Motiian et al. [91] extended IB to deal with the challenging unsupervised domain adaptation problem without auxiliary data view available during testing. The auxiliary data view is only given in source domain, which is trained, along with the main view in source domain and the unlabelled data in the target view, for learning visual classifier. However, it has two limitations: 1) It can only be used for two-domain case, i.e., source and target domain. 2) It works in a supervised manner and needs a large number of labelled training data, which is quite time-consuming and labor-expensive in practice. Hence, it is natural to adopt multi-task clustering paradigm. Yan et al. [27] and Zhang et al. [90] proposed multi-task IB methods to learn and propagate the shared information among action clustering tasks for improving the individual task performance. Take the work in [27] as an example, the detailed objective function is shown as follows:

$$\mathcal{F} = \sum_{i}^{m}[I(Y^i;T^i) - \beta^{-1}I(X^i;T^i)] + \sum_{s=1}^{m}\lambda_s \cdot \sum_{t=1,t\neq s}^{m} I(T^s;T^t). \tag{18}$$

where T^i denotes the compact variable of the i-th task, λ_s balances the influence from other tasks on the s-th task and $I(T^s;T^t)$ measures the correlations between pairwise tasks.

The preceding methods can be widely applied for cross-view and real-world action recognition. For multi-domain image data, e.g., images from various online shopping websites, a correlation propagation based multi-task IB [13] is designed to learn the related images between pairwise tasks and explore the positive correlations among tasks, from which each task can benefit a lot.

Analysis: The methods in this part are the typical multi-task propagating IB attempts, and have also shown some nice properties in the correlation exploration mechanism. These may inspire deeper research on more visual recognition problems, e.g., the efficient and interpretable learning.

D. Summary and Discussion

Although most aforementioned traditional IB methods have gained impressive achievements in machine learning, pattern recognition and the corresponding applications, they still face

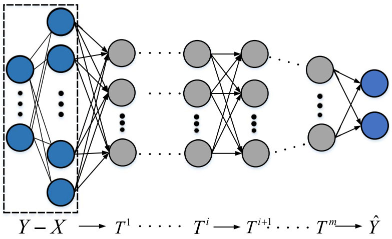

$Y - X \longrightarrow T^1 \cdots\cdots T^i \longrightarrow T^{i+1} \cdots\cdots T^m \longrightarrow \hat{Y}$

Fig. 7. Markov chain of intermediate representations $\{T^i\}_{i=1}^m$ in the DNNs layers referred from [4]. Given the mutual information between the input X and the desired output Y, the IB bound helps to find the optimal intermediate representations $\{T^i\}_{i=1}^m$ on the DNNs layers, through which the final predicted output $\hat{Y}$ can be obtained after training.

two main weaknesses. 1) Large-scale or/and high-dimensional setting. Since large-scale or high-dimensional data may lead to "big" pre-processed input matrix, the computational complexity will thus be very high, and further the running efficiency can not be ensured. In the future, more research on efficient learning is worth investigating. 2) Incomplete dataset. Multi-source heterogeneous data sometimes may be incomplete since some data samples are damaged or missing across sources. To alleviate this problem, effective incomplete propagating IB methods for multi-view or multi-task learning require more future research.

IV. Deep Information Bottleneck

A. Debate: Understanding DNNs With IB

In the past few decades, deep learning based methods have achieved remarkable performance and successfully applied into numerous practical fields. However, how the DNNs work and why the deep learning methods work so well remain mysterious. Recently, several researchers have tried to explore the working mechanism of DNNs using IB theory, and have acquired some stage achievements.

1) Affirmative Side:

> *The most important part of learning is actually forgetting.*
>
> -Naftali Tishby

As a pioneer work, Tishby and Zaslavsky [4] proposed to analyze DNNs in an IB framework for a better understanding of working mechanism. The authors took the general supervised feedforward DNNs (Fig. 7) as an example for analysis. By quantifying the DNNs with mutual information among input/output and each hidden layer, the authors computed the optimal IB limits on DNNs and finite sample generalization bounds, and then pointed out the connections between the DNN's layered structure and the structural phase transitions (a series of critical points) among IB information plane (Fig. 8). In Fig. 7, DNN's layered structure is treated as a Markov chain of intermediate representations $\{T^i\}_{i=1}^m$, forming the approximate sufficient statistics [5] of the input. Then, DNNs can be analyzed by IB principle where each layer maximally preserves the relevant information with the output, i.e., $I(T^i, Y)$, while minimizing the information with its previous layer, i.e., $I(T^i; T^{i-1})$, thus each layer is guaranteed to be most relevant about output and most concise about the input.

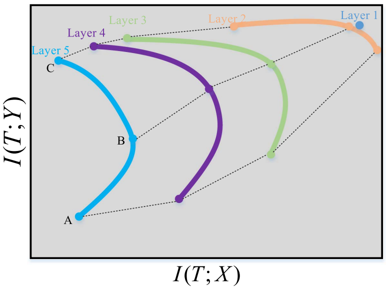


Fig. 8. IB information plane depicting the MI values that each layer T^i maintains on the input X and output label Y. It shows the increasing training epochs and mutual information quantification from A to C on each layer of the DNNs. Specifically, A → B denotes the ERM/fitting phase, taking a few hundred epochs, which focuses on training/fitting labels and increases $I(T;Y)$. B → C denotes the compression phase, taking much longer epochs, which forgets irrelevant information and compresses X while preserving relevance of label Y, and decreases $I(T;X)$. Note that it converges to a balance state of accuracy and compression at point C in this figure.

Following the work [4] formulating the goal of DNNs as sequentially optimizing the IB trade-off between compression and prediction on each layer, Schwartz-Ziv and Tishby [92] deeply explored the working mechanism of DNNs and provided some new insights via visualizing DNNs with a potent lens of IB information plane (Fig. 8). Their findings mainly contain: *1) Two dynamic phases:* Stochastic gradient decent (with no batch normalization) generally used for optimizing DNNs has two dynamic phases in the information plane: empirical error minimization (ERM) and compression, as shown in Fig. 8. ERM phase is short and the mutual information between the layers and the input/output, i.e., $I(T^i; X)$ and $I(T^i; Y)$, is increasing, which is also called fitting phase. Compression phase begins when the mean and standard deviation of stochastic gradient undertake a phase transition, where the signal-to-noise ratio decreases significantly. This phase lasts much longer, and the mutual information between the layers and the input, i.e., $I(T^i; X)$, is decreasing. And this phase is further claimed to lead to the remarkable generalization ability of DNNs. *2) Benefit of hidden layers:* Experiments showed adding hidden layers to DNNs can reduce the training epochs, and the compression phase of each layer would be shortened if the training started from a previous compressed layer. *3) IB bound of layers:* The converged hidden layers stayed or close to theoretical IB bound with different trade-off β.

The findings in [92] moved a further step for analyzing DNNs, mainly including the learning dynamics, compression effect and generalization abilities. Following these, Bang et al. [93], more recently, proposed a variational IB modelled by DNNs, for interpretable machine learning community. It provided a brief yet comprehensive explanation containing a cognitive chunk of features, e.g., word or a crowd of pixels, which maximally preserved informative knowledge about the decision of a black-box system while compressing the input as much as possible.

Analysis: The above works suggest that IB theory offered a new insight into theoretically explaining the working mechanism of the black box–DNNs. However, the discoveries were made on the experiments with several specific settings, and more general scenarios could be further considered.

2) Negative Side: Although the authors in [4], [92] have made some exciting findings, it is unclear whether these findings still hold for general network settings. To this end, more investigations have been made recently, where some typical ones are elaborated below.

M.Saxe et al. [94] mainly focused on three specific points in general cases, including the existence of two dynamic phases, the causal link of compression phase to generalization ability of DNNs, and the relation between stochasticity in training and compression phase. After conducting extended experimental analysis, the authors offered some different discoveries. *1) Nonlinearity and compression phase:* Compression phase occurs when using double-sided saturating nonlinearities (e.g., tanh) with mutual information estimated by binning activations or adding noise, but fails if adopting the single-sided ones (e.g., ReLU) and linear activation functions. *2) Link between compression and generalization:* Various experiments suggested that there is no causal link between the compression, or say forgetting, and the generalization performance of DNNs, which is also confirmed in [95], [96]. *3) Relation between stochasticity and compression:* The compression phase, if exists, does not result from the stochasticity, because the compression also occurs when using full batch gradient descent. *4) Temporal relation between two phases.* It is found that the fitting and compression phases, if exist, coincide instead of being subsequent when the input X contains a subset of both task-relevant and task-irrelevant information. In conclusion, these findings emphasized the vital role of noise assumptions in analyzing the DNNs with information theory, and showed that the IB theory may be of no generality in explaining the success of deep learning.

The above new insights motivated many other followup works [95], [97], [98], [99]. The authors in [95] claimed that the compression phase can still occur in spite of using non double-sided saturating activation in particular case. Moreover, Li and Liu [97] applied the IB theory on convolutional neural networks and found there is no compression even using double-sided saturating nonlinearities, which is partly consistent with the finding in [94] that compression phase does not always emerge in the training process. To further investigate this issue, Chelombiev et al. [98] developed a more robust adaptive mutual information estimation and found that the saturation of the activation was not necessary for compression. Concerning the relation between compression and generalization, they suggested that only compression of the last layer in networks is closely related with generalization. When the output Y is deterministic about the input X, Kolchinsky et al. [99] showed that there is no strict IB trade-off between compression and prediction among hidden layers in DNNs when using multi-layer classifiers with zero probability of error, which differs from the claim in [92].

Analysis: The above works have achieved significant progress in analysing DNNs using IB, and many new impressive findings were showed over general settings with different network types, activation functions or optimization methods. Still, many issues concerning the IB theory in analyzing the DNNs remain confusing, warranting further investigation.

B. Optimizing DNNs Using IB

Observing the ability of IB theory in understanding the working mechanism of DNNs, many researchers try to further use IB to help optimizing DNNs for improving their learning performance, such as network training and network compression.

1) Network Training: Network training is to appropriately tune the weight parameters of user-designed neural network architecture, which plays a fundamental role in obtaining promising learning performance. Recent years have witnessed the success of integrating IB principle [12], [100], [101], [102], [103] in boosting the training performance. Some typical works are introduced below.

Inspired by the works [4], [92], Cheng et al. [12] presented an IB information plane-based evaluation framework to help understanding the capability and improving the classification accuracy of the DNNs, e.g., CNNs. Instead of using both IB loss on intermediate layers and training loss on output layer, Elad et al. [102] proposed to only adopt layer-by-layer IB loss without any classification loss, where mutual information neural estimator [104] was modified for quantification. To adapt the IB for training more challenging generative normalizing flows using invertible networks, an IB-based method [103] was developed and achieved a competitive accuracy in comparison to standard classifiers.

To better train the DNNs, many variants [105], [106], [107] of IB theory have been designed and demonstrated promising learning performance. For instance, a Hilbert-Schmidt independence criterion bottleneck [106] was presented to improve IB on training DNNs. It did not need any training loss and backpropagation, which alleviated vanishing gradients and thus were specifically suitable for extremely deep networks. To further adapt IB for optimization in real-world applications, Kim and Bansal [107] interpreted the self-driving DNNs to enhance transparency by proposing an attentional bottleneck, which integrated visual attention and the IB theory together.

Analysis: Actually, the improvement on network training with IB benefits from the nice properties of compression and preservation in IB theory. However, training on frequently-used large-scale networks always involves the tuning of redundant parameters, thus leading to extra storage space and inefficient computation.

2) Network Compression: To further accelerate the computation process and reduce the storage cost, network compression came into being. It is really crucial in many practical scenarios, such as the deployment of deep models on mobile devices with limited resources. Several IB-based works [108], [109], [110] have made contributions to this field, where some typical ones are presented below.

Borrowing from the idea in [1], [4], a new variational IB network [108] was proposed to compress large DNNs for reducing the time and memory requirements, which was quite significant for practical deployment. It focused on redundancy reduction among adjacent layers by pruning neurons with a latent sparse regularization so that informative neurons can be effectively preserved. For the compression of the complex DNNs, e.g., recurrent neural networks, Srivastava et al. [109]

TABLE IV
SUMMARY OF SOME DNN-BASED IB METHODS

Category	Method	Learning	Contribution
Representation Learning	Robust IB [124]	Supervised	A robust IB balancing the robustness (by Fisher information) and accuracy (by MI, estimation error)
	Deficiency IB [125]	Supervised	Minimizing deficiency by a variational deficiency IB method
	Hashing VIB [126]	Supervised	A simple deep hashing variational IB to learn effective binary representations of large-scale data
	Dual IB [127]	Supervised	A dual IB method and its variational version by introducing another predicted label variable
	IB-VAE [128]	Unsupervised	An IB modeled VAE for disentangled 3D facial expression learning
Classification/ Clustering	Nonlinear IB [129]	Supervised	Nonlinear encoding/decoding mapping with nonlinear IB
	Distributed VIB [130]	Supervised	A generalized deep distributed VIB for improving the running efficiency
	DI-VIB [131]	Supervised	Diversity inducing IB for generating effective ensembles of neural networks
Deep Multi-view Learning	Disentangled VIB [132]	Supervised	Learning decomposed view-specific and view-shared representations by a disentangled VIB
	Incomplete VIB [133]	Supervised	A incomplete VIB for incomplete multi-view data with missing samples
	MfIB-Net [134]	Supervised	A multi-feature deep IB network for breast cancer classification
	MIB-Net [135]	Supervised	Learning minimal sufficient unimodal and multimodal representations by multi-modal IB network
	SIB-Net [136]	Unsupervised	A self-supervised deep IB network for multi-view subspace clustering
Deep Transfer Learning	LR-VIB [137]	Supervised	A novel VIB generalizing to low-resource domains
	IIB [138]	Supervised	Domain generalization with an Invariant IB
	CDRIB [139]	Supervised	A novel VIB for cross-domain recommendation
	IMIB-Net [140]	Supervised	An interpretable multi-task IB network for breast tumor classification and segmentation
Graph Learning	HIBPool [141]	Supervised	Improve graph pooling functions with IB by balancing the expressiveness and robustness
	SGIB [142]	Unsupervised	Self-supervised graph IB by adapting symmetry and asymmetry of graphs to contrastive learning
	HGIB [143]	Unsupervised	A heterogeneous graph IB focusing on the consensus hypothesis in GNNs

first employed a variational IB-based method for pruning and further integrated it with a group-lasso regularization technique to facilitate compression.

Analysis: The above IB-based models generally incorporate the strategies of network pruning and parameter quantization into IB framework for effective network compression. In the future, more ways, e.g., by integrating knowledge distillation strategy, are worthy of exploration.

C. Dnn-Based Ib

Thanks to the powerful information compression and preservation of the IB principle, IB has been successfully applied into addressing various challenging problems in DNNs, such as natural language processing [9], [111], [112], [113] and visual recognition [16], [114], [115], [116], [117]. In the following, we show some representative works in the commonly-seen areas, while the remaining methods are listed in Table IV.

1) Representation Learning: Representation learning aims to learn informative deep representations of input data, which usually provides the basic yet important support for the subsequent practical tasks, such as data classification. Generally, learning representations using DNN-based IB methods can be categorized into two classes, including supervised and unsupervised representation learning methods.

Supervised learning: Since the semantic information from the input data label is properly preserved, representations learned in a supervised fashion are often discriminative, which could be effective in application to the downstream tasks. Achille and Soatto [118] proposed an information dropout method, a generalization of dropout method in deep learning, which aimed to build a bridge by the IB principle between the learning of optimal representations and dropout methods for avoiding overfitting. However, it failed to consider the sparsity of the learned representations. Facing this problem, Wieczorek et al. [119] incorporated a copula transformation to the deep IB model by restoring the invariant properties and then proposed a deep copula IB method to make best use of the power of deep latent variable models in representation learning, so that sparse and disentangled latent features can be learned.

Different from the above, a prediction-oriented decodable IB [120] was proposed to discover optimal (i.e., minimal but sufficient) representations, which can ensure satisfactory testing performance and improve the generalization ability. Although the preceding IB methods can well learn useful representations, it is difficult for them to optimize the IB Lagrangian trading off the compression and prediction terms. Pan et al. [121] presented a supervised disentangled IB method to maximally compress the source data while not damaging the prediction performance. It worked by bringing another variable S to capture the target Y-irrelevant aspect in addition to encoding the target Y-relevant aspect to T, and in this wonderful disentangling manner the optimization of the IB Lagrangian was avoided.

Unsupervised learning: Although many remarkable works have been devoted to the supervised representation learning, the annotation process in these methods is labor-intensive and time-consuming, making it difficult to acquire enough training data labels in limited costs. To tackle this problem, Yamada et al. [122] proposed an unsupervised generative factorizing variational autoencoders (VAE) model based on IB principle for disentangled representation learning of sequential data. It can disentangle representations among challenging dynamic factors of the same sequential data time dependency, and further showed its effectiveness on speech and video scenarios. In real-world applications, speech information is usually represented by multiple components, e.g., content, pitch and rhythm, so it is significant to learn disentangled representations for various practical usage. To address this issue, an encoder-decoder structure-based triple IB [123] was carefully designed for speech decomposition, where "triple" denotes there exist three IB-based encoders.

Analysis: Compared to the traditional hand-crafted techniques, DNN-based IB models usually have the following advantages: 1) The learned features are more deep, abstract and robust. 2) External factors, e.g., illumination and pose, have slight impact on the learning process. 3) It works automatically without user intervention. However, deep models also have some disadvantages: 1) Overly dependent on the network architecture. 2) Requiring large-scale training samples from different scenarios and scales. In the future, the combination of traditional feature learning methods and deep IB-based methods may be a hot research topic.

2) Classification/Clustering: Pattern recognition aims to partition the input samples into specific classes depending on their feature representations, which usually contains two types, supervised classification and unsupervised clustering. In the past few decades, it has been successfully adopted to document/speech/remote-sensing image recognition, medical diagnosis, autonomous driving and so forth. Below we introduce some typical deep pattern recognition models.

Supervised classification: One of the most successful applications of DNN-based IB methods is the supervised classification. The aim of these methods is to learn informative compressed DNN representations which maximally preserve relevance with the input data label. For example, Amjad and Geiger [144] utilized IB principle to learn useful representations for network-based classification. They mainly focused on two problems in the optimization process: 1) IB functional is either infinite or partial constant for deterministic DNNs. 2) It is hard to capture the property of robustness and simplicity of the learned DNNs' intermediate representations due to the invariance of the IB functional under bijections.

Another important line is about the deep variational IB (DVIB) [34] method and its improved variants. The pioneer work of DVIB, a variational approximation of IB, aims to parameterize IB with a DNN and then utilize the popular re-parameterization trick for training. It tried to find an encoding T by compressing the input X, formulated by a parametric encoder $p(t|x;\theta)$, where T is informative about the label Y as much as possible. DVIB is formulated by the following:

$$\max E_{p(t|x)}[\log q(y|t) - \beta KL(p(t|x)||r(t))], \tag{19}$$

where $q(y|t)$ and $r(t)$ are the variational approximation to $p(y|t)$ and $p(t)$ respectively. Note that the widely-used VAE [145] can be seen as a special case of the unsupervised version of the DVIB function with β set as 1.0.

Further, a case study of DVIB [146] was conducted to show its ability in improving classification calibration and discovering out-of-distribution data samples. Attribution methods have attracted lots of attention since it can select the most relevant inputs benefiting for decision-making models by assigning a relevance score to each input, e.g., pixels of an image. To solve this problem, Schulz et al. [147] proposed an IB attribution method to learn informative input regions by quantifying the amount of information each input contained, and it also improved the model interpretability and the reliability of the final results.

Semi-supervised classification: In addition to the above supervised classification methods, semi-supervised classification problem solved in an IB framework has raised concerns in the very recent years. Voloshynovskiy et al. [148] designed a semi-supervised VIB model with handworked and learnable priors imposed on the latent space of the classifier, and also showed that several existing state-of-the-art semi-supervised models, e.g., VAE (M1+M2) [149], can be seen as its special cases. Similar to the attribution problem [147], a semi-supervised IB approach [16] is proposed to learn image attention masks targeting for classification. It focused on detecting the meaningful regions/features closely related to the image label and removing the "unwanted" noisy features, and also showed its promising potential in application to medical image annotation.

Clustering: Few IB-based works have been done for the completely unsupervised deep clustering. For instance, Ugur et al. [150] developed a VIB-GMM method combining VIB and Gaussian mixture model (GMM) for generative clustering. This method employed the VIB to trade off the accuracy and regularization of the model and modeled the learned latent space as a mixture of Gaussians. It is worth noting that VIB-GMM can be treated as a generalization of the supervised DVIB method [34] and the unsupervised variational deep embedding clustering method [151].

Analysis: The above supervised methods are the most active research interest in this field. However, in this Big Data era, we believe that deep clustering will be a hugely important trend with the rapid growth of deep learning. In our view, the incorporation of self-supervised learning (e.g., contrastive learning) or pseudo-label-guided learning will probably lead to improved deep IB clustering models.

3) Deep Multi-View Learning: Data with multiple views, intuitively, contain more complementary or/and consistent information than those with single view, which is beneficial for representation learning or downstream tasks. However, views sometimes may carry noisy or irrelevant information, inhibiting the model's generalization ability. Fortunately, IB principle provides an effective min-max strategy to learn task-oriented, robust or disentangled representations.

Supervised learning: Wang et al. [152] designed a simple yet effective multi-view IB for accurate representation learning in a supervised fashion. It attempted to minimize the MI between original m input view representation $\{V_1, V_2, \ldots, V_m\}$ and the learned joint representation of each view $\mathbf{T} = \{T_1, T_2, \ldots, T_m\}$ and meanwhile maximize the learned joint representation $\mathbf{T}$ and data label Y, formulated by

$$\max_{T,T_1,T_2,\ldots,T_m} I(Y;\mathbf{T}) - \sum_{i}^{m} \alpha_i I(V_i;T_i), \tag{20}$$

where T_i denotes the latent representation of the i-th view V_i, and α_i denotes the parameter regularizing the MI between V_i and T_i. Then, the variational inference method was adopted for its optimization. Similarly, Wang et al. [153] proposed a dual-view IB method to learn consistent discriminative joint representations among views for solving the downstream task of multi-view logo classification.

Unsupervised learning: However, the above methods can be only suitable for the supervised representation learning, and may probably not work well in the case where insufficient training data are provided. To circumvent this limitation, Federici et al. [154] presented a robust multi-view IB method for unsupervised representation learning by eliminating the information not shared by view representations. To this end, it maximized the MI between view-specific representations and meanwhile discarded the information not shared among them, thus ensuring robustness. However, this method only considers two views and has difficulty in extending to multiple views. Facing this, a collaborative multi-view IB [155] was proposed for flexible unsupervised representation learning. It took the advantages of the IB principle to integrate both view-common and view-specific representations, so that the comprehensive complementary and

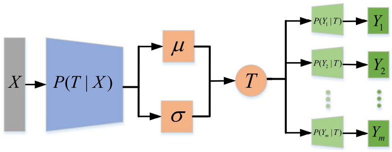


Fig. 9. Multi-task VIB model referred from [156]. X and $\{Y_i\}_{i=1}^m$ denote the input and output of the model respectively. $P(T|X)$ and $P(Y_i|T)$ denote the encoder and decoder respectively. μ, σ and T are the mean, standard deviation, and latent variable respectively.

consistent information among views can be fully learned and employed while the worthless information was discarded.

Analysis: Deep multi-view IB models enjoy the properties of comprehensive correlation exploration and exploitation. However, the above IB methods have the following shortcomings: 1) These methods only consider the view correlations, ignoring the quantization and interaction among views. 2) The resulting multi-view results lack of analysis on the trustworthiness and interpretability, and may not be suitable for safety-critical applications, such as multi-modal medical diagnosis or autonomous driving.

4) Deep Transfer Learning: Transfer learning [157], as a new learning paradigm, aims to transfer useful knowledge from source to target domain, where both domains hold different data distributions or feature spaces. During the past decade, it has achieved great success in many areas, such as image classification, image segmentation and recommendation system. Several methods, more recently, have been devoted to this community.

For the instances of applications in unsupervised domain adaptation where target domain has no labels, Luo et al. [158] proposed a significance-aware IB adversarial network for this adaptation problem in semantic segmentation, i.e., assigning a label to each pixel of an image. It first utilized IB to eliminate the task-independent factors among domain features to obtain purified semantic features, which then helped to promote the following adversarial adaptation. Besides, a variational bottleneck domain adaptation method [159] was presented to learn task-relevant information for source and target domains and meanwhile disregard the task-irrelevant factors to enhance the feature transferability.

However, the above two methods can not generalize to low-resource domains [137] or unseen domains [160] (e.g., generalizing a cup classifier trained on image to that on sketches). Du et al. [160] tried to deal with the challenging unseen domain generalization problem by introducing a meta variational IB for bridging the gaps between domains to learn domain-invariant features and meanwhile maximize the classification accuracy. However, these methods can only be applied into transfer learning problems with two domains, thus motivating multi-task VIB models [140], [156]. For instance, a multi-task VIB was proposed [156] to solve the difficult multi-task learning problem involving knowledge transfer among multiple domains. It aimed to maximally compress the input to a latent representation with an encoder while predicting the output for multiple tasks in parallel with multiple decoders, shown in Fig. 9.

Analysis: Although the above methods have a broader range of applications compared to other deep IB-based methods, these methods still face two main weaknesses: 1) Few works touched the negative transfer problem when source and target domain are not very similar. 2) They fail to accurately measure the domain relations, thus how much knowledge should be transferred is uncertain.

5) Reinforcement Learning: Motivated by the new skill learning process of humans, reinforcement learning (RL) utilizes the previous experience and mistakes to find the optimal strategy in order to maximize the future reward. Robotics, game industry, autonomous driving and machine reasoning are the major successes of RL. Recently, remarkable achievements [161], [162], [163], [164], [165], [166], [167] have been made on the RL with IB methods, and some typical ones are shown as follows.

Igl et al. [161] focused on the generalization of learned policies to new setting or environments and designed a new regularization technique by proposing the selective noise injection and the IB actor-critic, where the former can enable the agent to observe the future rewards and stabilize the training process, and the latter allowed the agent to learn more general and robust features, thus ensuring generalization ability. Additionally, the authors in [163] divided the input of the RL into two parts, including standard and privileged input, and then proposed a variational bandwidth bottleneck method to only compress the privileged part to ensure the relevant standard input can be fully utilized for deciding, just like one only focuses on the position and speed of a car (“standard input”) and ignores its color or shape (“privileged input”) when crossing the road. Finally, to address the more challenging multi-agent [164], [165] or multi-view RL [166] problem, several notable works have been carefully designed. For example, Wang et al. [164] obtained low-entropy and informative transmitted messages by presenting a variational IB based regularization method to learn an informative communication protocols across agents.

Analysis: The major advantage of above RL methods lies on the strong adaptability to the corresponding interactive environment. However, most of them can only work well in the current environment, which reveals poor robustness to new environments.

6) Adversarial Learning: Adversarial learning works by mutual promotion of adversarial attack and defense, thus resulting in improved version of each one. A typical example is the Generative Adversarial Networks (GANs) [168], which adopts generator (generate samples as real as possible) and discriminator (judge whether the sample is fake or real) networks for adversarial training. To improve the training performance, many IB-based methods have been proposed and successfully applied into many fields, such as representation learning and safe artificial intelligence [169].

This line of works [169], [170], [171], [172], [173] generally incorporate or adapt IB theory into adversarial models (e.g., GANs) for better learning ability. Peng et al. [170] proposed a variational discriminator bottleneck method to impose a constraint on the MI between input and the discriminator’s internal representation, and further demonstrated this information flow constraint can benefit the imitation learning and GANs. To improve the existing InfoGAN model [174], an unsupervised IB-GAN [171], shown in Fig. 10, was presented for disentangled and interpretable representation learning. The difference

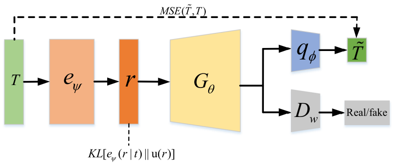


Fig. 10. IB-GAN model referred from [171]. The encoder $e_\psi(r|t)$ and KL-divergence loss are derived from IB principle, where z and r denote the input latent code and intermediate representation respectively. G_θ and D_w denote the generator and discriminator respectively. Note that the encoder $q_\phi(r|t)$ and $u(r)$ are defined as Gaussian. MSE is short for the mean square error loss.

between them is that, an intermediate layer of generator (i.e., a learnable latent distribution) in IB-GAN was employed to constrain the MI between the generated data and input, which was further trained with the generator in an end-to-end manner. Therefore, IB-GAN can be seen as a general form of the Info-GAN model. More recently, by not making any assumption on the distribution of intermediate representations, an adversarial IB [173] was explicitly modelled and optimized in an effective min-max framework.

Analysis: It is surprising to see that the unsupervised adversarial IB-based methods can obtain such promising performance. However, they suffer from the following limitations: 1) most of them solved the image tasks, without considering more data types, e.g., text, audio or video. 2) The adversarial training is unstable, sometimes leading to model collapse.

7) Distributed Learning: Distributed learning is to train a global learning model by collaborative learning of multiple learners, each of which only can use a subset of or an independent noisy version of the whole dataset. This learning paradigm is imperative in many scenarios where data are collected from multiple sensors, nodes or sources. Recently, there emerged an increasing line of works [130], [175], [176] on the distributed version of IB theory, with lots of ramifications and practical applications, such as edge inference, medical imaging, and Internet of Things environment.

Aguerri and Zaidi [175] first proposed the distributed IB method where multiple inputs $\{X^i\}_{i=1}^m$ are separately processed into different corresponding compact representations $\{T^i\}_{i=1}^m$, maximally maintaining the relevant information about true label Y. Then, they developed two distributed IB models with the discrete and vector Gaussian sources. It is notable that Murphy and Bassett [177] successfully applied the distributed IB [175] for the explanatory structure of complex systems, e.g., applied mathematics. For generalization to deep learning, the authors in [130] designed a deep variational distributed IB, and further extended the evidence lower bound to the distributed version for obtaining a new variational bound. Afterwards, they [176] made a comprehensive elaboration on the (deep) distributed IB, including the discrete and Gaussian models, variational bound, and rich theoretical and empirical analysis. Since a typical characteristic of learning in networks is the distributiveness of data among different sensors or nodes, the distributed IB [178], [179] was generalized to the networks that can be modelled by directed graphs, in which the learning was conducted distributively in both the training and testing phase, called "in-network learning".

Analysis: Distributed IB methods enjoy the major strengths of fast training process, decentralized storage and data privacy protection. By deeply analyzed the working mechanism of distributed IB, in our view, these methods could be further improved by multi-view/feature/modal/task IB methods in the future.

8) Graph Learning: Recently, deep graph learning, e.g., graph neural networks (GNNs), has attracted extensive attention and shown impressive performance by fully exploiting the informative nodes and graph structure. However, the graph learning is susceptible to the noise, leading to degraded performance. Fortunately, this goal of information preservation and noise elimination just coincides with the idea of IB principle.

Here we list some typical methods. The pioneer work is the graph IB (GIB) method [180], aiming to learn minimal sufficient representation by trading off its expressiveness and robustness. In addition to node features, GIB also regularized the graph structure and further designed two instantiated methods, which showed their superiority on the adversarial attacks. The objective function is formulated by

$$\min_{p(T|D)} GIB_\beta = -I(Y;T) + \beta I(D;T), \tag{21}$$

where T and Y denote the representation and the target label, respectively, and $D = (X, A)$ denotes the input graph data, where X and A indicate the node features and graph structure, respectively.

Following up with GIB, many variants of works [141], [142], [143], [181], [182], [183] have emerged very recently. Sun et al. [181] proposed a VIB guided framework to solve the challenging low-quality (e.g., noisy or incomplete) graph representation learning problem, and further gained superior performance on classification tasks. Different from the preceding methods, Yu et al. [182], [183] focused on subgraph recognition problem and introduced a graph IB framework and its improved version [183] to learn compressed yet informative subgraph of the original one. The frameworks were general and easy to be adapted into different GNNs, and the learned IB-subgraph was further demonstrated to be useful in graph denoising and classification.

Analysis: Above IB-based methods have shown unique advantages in the graph compression, subgraph learning and heterogeneous graph learning. In the future, more interesting graph-related tasks, e.g., link prediction, large-scale or incomplete graph learning, deserve future investigation.

D. Summary and Discussion

Throughout the DNN-based IB works in the above fields, it is clearly seen that, although numerous supervised or semi-supervised deep IB methods have been proposed and obtained promising performance in the past decade, few deep IB methods are designed for unsupervised learning, e.g., clustering. Actually, it is difficult to ensure the labeled training data is enough for obtaining a good model, since "enough is never quite enough"—both the accuracy of the label reflecting from the data and the diversity of the data are needed to be considered. Hence, instead of finding more training data for deep IB methods, exploring unsupervised deep IB methods is probably a bright way.

Generally, traditional IB works by first building input feature matrix with hand-crafted feature extraction technique and then

TABLE V
PROS AND CONS OF TRADITIONAL AND DEEP IB

IB	Pros	Cons
Traditional IB	Transparent training Better generalization ability Less training parameters	Hand-crafted feature input Suboptimal performance Poor adaptability for large-scale data
Deep IB	Representation learning ability Remarkable performance End-to-end optimization	Unclear "black box" Rely on numerous training data Massive network parameters

mapping the input to output space. However, the difficulty of constructing hand-crafted feature for various data types (e.g., text, image and video) may limit its development. In contrast, deep IB usually conducts end-to-end mapping from original data to output, which enables learning powerful feature representation. We summarize the pros and cons on both kinds of methods in Table V.

Despite the mentioned limitations, deep IB, in our view, is still promising since it remedies major weaknesses of traditional IB methods. More works are expected for developing its potential. Here we also leave a question–"Will or not deep IB methods totally replace the traditional IB methods in the near future?".

V. RELATION TO OTHER PROBLEMS/METHODS

In this part, we briefly discuss some other problems or methods closely related to IB theory.

Relation to Source Coding Problem: Source coding is to map the source data to encoded data so that the source information can be fully recovered or recovered under certain distortion, thus the important role of which is for data compression. Similarly, IB theory can also be used for information compression, and is closely related to the source coding. Specifically, Rate Distortion (RD) theory stands in the center of source coding, which seeks the minimal rate under the pre-set distortion between source and compressed data, formulated in the (3). Rather than a generalization, IB theory can be seen as a special case of RD theory with a special loss function and the Lagrangian formulation of the constrained optimization. Additionally, IB problem is also related to the indirect source coding problem [184] with logarithmic loss distortion measure [185], and the Wyner-Ziv coding problem with common reconstruction [186].

Relation to the Wyner-Ahlswede-Korner Problem: Given two memoryless variable X and Y, both of them are respectively encoded in the rates R_X and R_Y in an independent manner. Then, one decoder tried to utilize the two compressed representations to losslessly reconstruct Y. The above problem was challenging and has ever separately investigated by the authors in [187], [188], thus termed Wyner-Ahlswede-Korner (WAK) problem, which is formulated by $\hat{R}_Y(R) = \min_{P(T|X):I(T;X)\leq R} H(Y|T)$, where $R_X = R$. Discovering the minimal rate R_Y needing for reconstructing Y is also to optimize the formulation $\max_{P(T|X):I(T;X)\leq R} I(T;Y) = H(Y) - \hat{R}_Y(R)$. Thus, it is equivalent for the optimization of the IB problem and the WAK problem [31].

Relation to Cloud Radio Access Network Model with Oblivious Processing: Cloud radio access network (CRAN) defines the information transmitted from the user input to a distant receiver output through relay stations which link with the receiver by finite-capacity error-free links [189]. While the oblivious processing means the relays work with no information about the inputs' codebooks. Here, the IB theory with the user input, relays and receiver set as variables X, T and Y respectively, is closely related to the CRAN model with oblivious processing where the relays are independent conditionally on the inputs [185].

VI. CONCLUSION

In brief, as an information-theoretic and pattern analysis technique, IB improves its prediction ability due to the elegant balance between data compression and preservation. The results described in this survey clearly demonstrate the effectiveness and superiority of IB in supervised/semi-supervised/unsupervised/reinforcement learning and many practical applications. Although remarkable progress has been made on IB, there are still lots of potential problems deserving our immediate research, such as:

Global optimum: Most existing traditional IB methods displayed in the Section III can only gain local optimal value, thus designing optimization methods for learning global solution is quite imperative. Essentially, seeking global optimum is NP-hard. Hence, it turns to how to solve this NP-hard problem, and some ideas from the combinatorial or approximate optimization techniques may be borrowed.

Determination of the trade-off parameter β: The parameter β in the IB principle is used to balance the compression and preservation, which has a significant impact on the final results. Fortunately, Wu et al. [190] analyzed the relationships between β and learnability of the supervised IB model, and then gave a theoretical guidance to select a proper β. However, no systematic technique has been designed for learning the optimal β in an unsupervised fashion to date, impeding the progress of IB in many practical applications.

Explainable IB: The core of explainable machine learning is to consider and balance both model prediction accuracy and explainability, which is very useful in many areas, e.g., financial and medical industries. Exploring explainable mechanism for IB is conducive to improving its reliability and security. However, none explainable IB methods have been developed so far.

Efficient IB: Although IB has demonstrated its effectiveness in many communities, its working efficiency is often not satisfactory compared to other models, e.g., k-means, spectral or non-negative matrix factorization methods. Additionally, with the arrival of Big Data era, designing IB methods for large-scale data becomes a more urgent issue.

IB with deep graph learning: With the emerging of deep graph learning in more recent years, the interpretability of the black-box of GNNs has not been touched. In the light of the IB principle for interpreting DNNs [4], [92], the possibility of explaining GNNs with IB principle is worth investigation. Additionally, more deep graph-based IB methods like GIB [180], [182] are also worth studying.

Unexplored yet promising applications of IB: Although IB has been successfully applied into lots of fields due to its effectiveness, there are still many unexplored yet promising applications that IB could be considered, such as data privacy protection and financial risk analysis.

REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Annu. Allerton Conf. Communnication Control Comput.*, 1999, pp. 368–377.

[2] N. Slonim, "The information bottleneck: Theory and applications," Ph.D Thesis, Hebrew, Hebrew Univ., Israel, 2002.

[3] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2002, pp. 129–136.

[4] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop*, 2015, pp. 1–5.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Hoboken, NJ, USA: Wiley, 2006.

[6] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *Proc. 23rd Eur. Colloq. Inf. Retrieval Res.*, 2001.

[7] P. West, A. Holtzman, J. Buys, and Y. Choi, "BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle," in *Proc. Conf. Empir. Methods Natural Lang. Joint Conf. Natural Lang. Process.*, 2019, pp. 3750–3759.

[8] D. Xu and F. Fekri, "Time series prediction via recurrent neural networks with the information bottleneck principle," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wirel. Commun.*, 2018, pp. 1–5.

[9] X. L. Li and J. Eisner, "Specializing word embeddings (for parsing) by information bottleneck," in *Proc. Conf. Empir. Methods Natural Lang. Joint Conf. Natural Lang. Process.*, 2019, pp. 2744–2754.

[10] Y. Yang, P. Liu, S. Fei, and C. Zhang, "A topic link detection method based on improved information bottleneck theory," *Acta Automatica Sinica*, vol. 40, no. 3, pp. 471–479, 2014.

[11] H. Shen, X. Cheng, H. Chen, and Y. Liu, "Information bottleneck based community detection in network," *Chin. J. Comput.*, vol. 31, no. 4, pp. 677–686, 2008.

[12] H. Cheng, D. Lian, S. Gao, and Y. Geng, "Evaluating capability of deep neural networks for image classification via information plane," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 181–195.

[13] S. Hu, X. Yan, and Y. Ye, "Multi-task image clustering through correlation propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 3, pp. 1113–1127, Mar. 2021.

[14] Z. Lou, Y. Ye, and D. Liu, "Unsupervised object category discovery via information bottleneck method," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 863–866.

[15] A. Bardera, J. Rigau, I. Boada, M. Feixas, and M. Sbert, "Image segmentation using information bottleneck method," *IEEE TIP*, vol. 18, no. 7, pp. 1601–1612, Jul. 2009.

[16] A. Zhmoginov, I. Fischer, and M. Sandler, "Information-bottleneck approach to salient region discovery," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2020, pp. 531–546.

[17] A. Ghosh, Z. Zhong, S. Cruz, S. Veeravasarapu, T. E. Boult, and M. Singh, "To beta or not to beta: Information bottleneck for digital image forensics," 2019, *arXiv: 1908.03864*.

[18] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 487–493.

[19] Y. Li et al., "Unsupervised detection of acoustic events using information bottleneck principle," *Digit. Signal Process.*, vol. 63, pp. 123–134, 2017.

[20] R. M. Hecht, E. Noor, G. Dobry, Y. Zigel, A. Bar-Hillel, and N. Tishby, "Effective model representation by information bottleneck principle," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 8, pp. 1755–1759, Aug. 2013.

[21] S. O'Rourke, G. Chechik, R. Friedman, and E. Eskin, "Discrete profile alignment via constrained information bottleneck," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 1009–1016.

[22] B. Thirion and O. D. Faugeras, "Feature detection in fMRI data: The information bottleneck approach," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, vol. 2879, 2003, pp. 83–91.

[23] B. Jin and X. Lu, "Identifying informative subsets of the gene ontology with information bottleneck methods," *Bioinformatics*, vol. 26, no. 19, pp. 2445–2451, 2010.

[24] W. Du, S. Tan, X. Cheng, and X. Yun, "Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2010, pp. 111–120.

[25] V. Pacelli and A. Majumdar, "Task-driven estimation and control via information bottlenecks," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2061–2067.

[26] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[27] X. Yan, S. Hu, and Y. Ye, "Multi-task clustering of human actions by sharing information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4049–4057.

[28] W. H. Hsu, L. S. Kennedy, and S. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 35–44.

[29] H. Hafez-Kolahi and S. Kasaei, "Information bottleneck and its applications in deep learning," 2019, *arXiv: 1904.03743*.

[30] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[31] A. Zaidi, I. E. Aguerri, and S. S. (Shitz), "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 2, 2020, Art. no. 151.

[32] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[33] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Information bottleneck learning using privileged information for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1496–1505.

[34] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2017.

[35] S. Hu, R. Wang, and Y. Ye, "Interactive information bottleneck for high-dimensional co-occurrence data clustering," *Appl. Soft Comput.*, vol. 111, 2021, Art. no. 107837.

[36] X. Yan, Z. Lou, S. Hu, and Y. Ye, "Multi-task information bottleneck co-clustering for unsupervised cross-view human action categorization," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 2, pp. 15:1–15:23, 2020.

[37] Z. Lou, Y. Ye, and X. Yan, "The multi-feature information bottleneck with application to unsupervised image categorization," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1508–1515.

[38] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 1999, pp. 617–623.

[39] B. Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 179–192.

[40] L. Wang, J. Zhang, L. Zhou, and W. Li, "A fast approximate AIB algorithm for distributional word clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 556–563.

[41] J. Peltonen, J. Sinkkonen, and S. Kaski, "Sequential information bottleneck for finite data," in *Proc. Int. Conf. Mach. Learn.*, 2004.

[42] H. Yuan and Y. Ye, "Iterative sib algorithm," *Pattern Recognit. Lett.*, vol. 32, no. 4, pp. 606–614, 2011.

[43] Y. Ye, D. Liu, L. Jia, and G. Li, "An sIB algorithm for automatically determining parameter," *Chin. J. Comput.*, vol. 30, pp. 969–978, 2007.

[44] Y. Ye, J. Zhang, and D. Liu, "An improved sequential IB algorithm for document clustering," *Pattern Recognit. Artif. Intell.*, vol. 21, pp. 417–423, 2008.

[45] Z. Lou, C. Yang, and Y. Ye, "An IB algorithm based on data selection model," *Acta Electron.a Sinica*, vol. 42, pp. 1839–1846, 2014.

[46] M. Rey, V. Roth, and T. J. Fuchs, "Sparse meta-Gaussian information bottleneck," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 910–918.

[47] Y. Ye, X. He, and L. Jia, "CD-sIB: A kind of SIB algorithm orient to categorical data," *Acta Electronica Sinica*, vol. 37, pp. 2165–2172, 2009.

[48] N. Tagasovska and P. Andritsos, "Distributed clustering of categorical data using the information bottleneck framework," *Inf. Syst.*, vol. 72, pp. 161–178, 2017.

[49] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik, "LIMBO: Scalable clustering of categorical data," in *Proc. Int. Conf. Extending Database Technol.*, 2004, pp. 123–146.

[50] X. Yan, Y. Ye, and Z. Lou, "Unsupervised video categorization based on multivariate information bottleneck method," *Knowl.-Based Syst.*, vol. 84, pp. 34–45, 2015.

[51] S. Hassanpour, D. Wübben, and A. Dekorsy, "A novel approach to distributed quantization via multivariate information bottleneck method," in *Proc. IEEE Glob. Commun. Conf.*, 2019, pp. 1–6.

[52] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 1213–1220.

[53] R. M. Hecht, E. Noor, and N. Tishby, "Speaker recognition by Gaussian information bottleneck," in *Proc. INTERSPEECH*, 2009, pp. 1567–1570.

[54] M. Rey and V. Roth, "Meta-Gaussian information bottleneck," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1925–1933.

[55] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby, “Multivariate information bottleneck,” in *Proc. Conf. Uncertainty Artif. Intell.*, 2001, pp. 152–161.

[56] Y. Ye, R. Liu, and Z. Lou, “Incorporating side information into multivariate information bottleneck for generating alternative clusterings,” *Pattern Recognit. Lett.*, vol. 51, pp. 70–78, 2015.

[57] N. Slonim, N. Friedman, and N. Tishby, “Agglomerative multivariate information bottleneck,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2001, pp. 929–936.

[58] N. Slonim and N. Tishby, “Document clustering using word clusters via the information bottleneck method,” in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2000, pp. 208–215.

[59] R. El-Yaniv and O. Souroujon, “Iterative double clustering for unsupervised and semi-supervised learning,” in *Proc. Eur. Conf. Mach. Learn.*, 2001, pp. 121–132.

[60] P. Wang, C. Domeniconi, and K. B. Laskey, “Information bottleneck co-clustering,” in *Proc. Text Mining Workshop*, 2010.

[61] Y. Liu, T. Duan, X. Wan, and H. Chao, “A mixed co-clustering algorithm based on information bottleneck,” *J. Inf. Process. Syst.*, vol. 13, no. 6, pp. 1467–1486, 2017.

[62] I. S. Dhillon, S. Mallela, and D. S. Modha, “Information-theoretic co-clustering,” in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 89–98.

[63] Y. Ren, Y. Ye, and G. Li, “The density-based agglomerative information bottleneck,” in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2008, pp. 333–344.

[64] Y. Ren, Y. Ye, and G. Li, “The density connectivity information bottleneck,” in *Proc. 9th Int. Conf. Young Comput. Scientists*, 2008, pp. 1783–1788.

[65] Y. Ye, Y. Ren, and G. Li, “Using local density information to improve IB algorithms,” *Pattern Recognit. Lett.*, vol. 32, no. 2, pp. 310–320, 2011.

[66] Z. Lou, Y. Ye, and Z. Zhu, “Information bottleneck with local consistency,” in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2012, pp. 285–296.

[67] S. Hu, Z. Hou, Z. Lou, and Y. Ye, “Content versus context: How about ”walking hand-in-hand” for image clustering?,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 3707–3711.

[68] G. Chechik and N. Tishby, “Extracting relevant structures with side information,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, pp. 857–864.

[69] A. Globerson, G. Chechik, and N. Tishby, “Sufficient dimensionality reduction with irrelevance statistics,” in *Proc. Conf. Uncertainty Artif. Intell.*, 2003, pp. 281–288.

[70] D. G. Department and D. Gondek, “Conditional information bottleneck clustering,” in *Proc. IEEE 3rd Int. Conf. Data Mining Workshop Clustering Large Data Sets*, 2003, pp. 36–42.

[71] M. Smieja and B. C. Geiger, “Semi-supervised cross-entropy clustering with information bottleneck constraint,” *Inf. Sci.*, vol. 421, pp. 254–271, 2017.

[72] Q. Zhang, S. Yu, J. Xin, and B. Chen, “Multi-view information bottleneck without variational approximation,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 4318–4322.

[73] T. Huang, A. E. Gamal, and H. E. Gamal, “Efficient multi-view learning based on the information bottleneck paradigm,” in *Proc. 56th Asilomar Conf. Signals Syst. Comput.*, 2022, pp. 1393–1397.

[74] Z. Lou, Y. Ye, and R. Liu, “Non-redundant multi-view clustering based on information bottleneck,” *J. Comput. Res. Develop.*, vol. 50, pp. 1865–1875, 2013.

[75] S. Hu, Z. Lou, R. Wang, X. Yan, and Y. Ye, “Dual-weighted multi-view clustering,” *Chin. J. Comput.*, vol. 43, pp. 1708–1720, 2020.

[76] S. Hu et al., “A parameter-free multi-view information bottleneck clustering method by cross-view weighting,” in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3792–3800.

[77] X. Yan, Y. Ye, X. Qiu, M. Manic, and H. Yu, “CMIB: Unsupervised image object categorization in multiple visual contexts,” *IEEE Trans. Ind. Inform.*, vol. 16, no. 6, pp. 3974–3986, Jun. 2020.

[78] S. Hu, B. Wang, Z. Lou, and Y. Ye, “Multi-view content-context information bottleneck for image clustering,” *Expert Syst. Appl.*, vol. 183, 2021, Art. no. 115374.

[79] X. Yan and Y. Ye, “Cross-media clustering by share and private information maximization,” *J. Comput. Res. Develop.*, vol. 56, pp. 1370–1382, 2019.

[80] X. Yan, Y. Mao, S. Hu, and Y. Ye, “Heterogeneous dual-task clustering with visual-textual information,” in *Proc. SIAM Int. Conf. Data Mining*, 2020, pp. 658–666.

[81] S. Hu, X. Yan, and Y. Ye, “Dynamic auto-weighted multi-view co-clustering,” *Pattern Recognit.*, vol. 99, 2020, Art. no. 107101.

[82] S. Hu, Z. Shi, and Y. Ye, “DMIB: Dual-correlated multivariate information bottleneck for multiview clustering,” *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4260–4274, Jun. 2022.

[83] Y. Gao, S. Gu, J. Li, and Z. Liao, “The multi-view information bottleneck clustering,” in *Proc. Adv. Databases: Concepts, Syst. Appl.*, 2007, pp. 912–917.

[84] X. Yan, Y. Lu, Z. Lou, and Y. Ye, “Multilingual documents clustering algorithm based on parallel information bottleneck,” *Pattern Recognit. Artif. Intell.*, vol. 30, pp. 559–568, 2017.

[85] S. Hu, Z. Lou, and Y. Ye, “View-wise versus cluster-wise weight: Which is better for multi-view clustering?,” *IEEE Trans. Image Process.*, vol. 31, pp. 58–71, 2022.

[86] S. Hu, X. Yan, and Y. Ye, “Joint specific and correlated information exploration for multi-view action clustering,” *Inf. Sci.*, vol. 524, pp. 148–164, 2020.

[87] X. Yan, Y. Ye, and X. Qiu, “Unsupervised human action categorization with consensus information bottleneck method,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2245–2251.

[88] X. Yan, Y. Ye, X. Qiu, and H. Yu, “Synergetic information bottleneck for joint multi-view and ensemble clustering,” *Inf. Fusion*, vol. 56, pp. 15–27, 2020.

[89] F. Zhuang, P. Luo, H. Xiong, Y. Xiong, Q. He, and Z. Shi, “Cross-domain learning from multiple sources: A consensus regularization perspective,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 12, pp. 1664–1678, Dec. 2010.

[90] M. Zhang, X. Yan, S. Hu, and Y. Ye, “An information maximization multi-task clustering method for egocentric temporal segmentation,” *Appl. Soft Comput.*, vol. 94, 2020, Art. no. 106425.

[91] S. Motiian and G. Doretto, “Information bottleneck domain adaptation with privileged information for visual recognition,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–647.

[92] R. Shwartz-Ziv and N. Tishby, “Opening the black box of deep neural networks via information,” 2017, *arXiv: 1703.00810*.

[93] S. Bang, P. Xie, H. Lee, W. Wu, and E. P. Xing, “Explaining a black-box by using a deep variational information bottleneck approach,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11 396–11 404.

[94] A. M. Saxe et al., “On the information bottleneck theory of deep learning,” in *Proc. Int. Conf. Learn. Representations*, 2018.

[95] M. Gabrié et al., “Entropy and mutual information in models of deep neural networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1826–1836.

[96] Z. Goldfeld, E. van den Berg, K. H. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, “Estimating information flow in deep neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2299–2308.

[97] J. Li and D. Liu, “Information bottleneck methods on convolutional neural networks,” 2019, *arXiv: 1911.03722*.

[98] I. Chelombiev, C. J. Houghton, and C. O’Donnell, “Adaptive estimators show information compression in deep neural networks,” in *Proc. Int. Conf. Learn. Representations*, 2019.

[99] A. Kolchinsky, B. D. Tracey, and S. V. Kuyk, “Caveats for information bottleneck in deterministic scenarios,” in *Proc. Int. Conf. Learn. Representations*, 2019.

[100] T. T. Nguyen and J. Choi, “Layer-wise learning of stochastic neural networks with information bottleneck,” 2017, *arXiv: 1712.01272*.

[101] T. T. Nguyen and J. Choi, “Markov information bottleneck to improve information flow in stochastic neural networks,” *Entropy*, vol. 21, no. 10, 2019, Art. no. 976.

[102] A. Elad, D. Haviv, Y. Blau, and T. Michaeli, “Direct validation of the information bottleneck principle for deep nets,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 758–762.

[103] L. Ardizzone, R. Mackowiak, C. Rother, and U. Köthe, “Training normalizing flows with the information bottleneck for competitive generative classification,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 7828–7840.

[104] M. I. Belghazi et al., “Mutual information neural estimation,” in *Proc. Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 530–539.

[105] A. Kirsch, C. Lyle, and Y. Gal, “Scalable training with information bottleneck objectives,” in *Proc. Int. Conf. Mach. Learn. Workshops*, 2020.

[106] K. W. Ma, J. P. Lewis, and W. B. Kleijn, “The HSIC bottleneck: Deep learning without back-propagation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5085–5092.

[107] J. Kim and M. Bansal, “Attentional bottleneck: Towards an interpretable deep driving network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 1310–1313.

[108] B. Dai, C. Zhu, B. Guo, and D. P. Wipf, "Compressing neural networks using the variational information bottleneck," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1143–1152.

[109] A. Srivastava, O. Dutta, J. Gupta, S. Agarwal, and P. AP, "A variational information bottleneck based method to compress sequential networks for human action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2744–2753.

[110] Y. Zhuo and L. Jiang, "Neural network hybrid compression method based on information bottleneck," *Application Res. Comput.*, vol. 38, pp. 1463–1467, 2021.

[111] J. Wu, X. Li, X. Ao, Y. Meng, F. Wu, and J. Li, "Improving robustness and generality of NLP models using disentangled representations," 2020, *arXiv: 2009.09587*.

[112] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer, "An information bottleneck approach for controlling conciseness in rationale extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1938–1952.

[113] Z.-Q. Yu, Z.-T. Yu, Y.-X. Huang, J.-J. Guo, and S.X. Gao, "Improving semi-supervised neural machine translation with variational information bottleneck," *Acta Autom. Sinica*, vol. 48, no. 7, pp. 1678–1689, 2022, doi: 10.16383/j.aas.c190477.

[114] J. Lambert, O. Sener, and S. Savarese, "Deep learning under privileged information using heteroscedastic dropout," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8886–8895.

[115] S. Rezaeifar, M. Diephuis, B. Razeghi, D. Ullmann, O. Taran, and S. Voloshynovskiy, "Distributed semi-private image classification based on information-bottleneck principle," in *Proc. Eur. Signal Process. Conf.*, 2020, pp. 755–759.

[116] A. Zhang, Y. Gao, Y. Niu, W. Liu, and Y. Zhou, "Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 598–607.

[117] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3404–3413.

[118] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2897–2905, Dec. 2018.

[119] A. Wieczorek, M. Wieser, D. Murezzan, and V. Roth, "Learning sparse latent representations with the deep copula information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2018.

[120] Y. Dubois, D. Kiela, D. J. Schwab, and R. Vedantam, "Learning optimal representations with the decodable information bottleneck," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020.

[121] Z. Pan, L. Niu, J. Zhang, and L. Zhang, "Disentangled information bottleneck," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 9285–9293.

[122] M. Yamada, H. Kim, K. Miyoshi, T. Iwata, and H. Yamakawa, "Disentangled representations for sequence data using information bottleneck principle," in *Proc. Asian Conf. Mach. Learn.*, 2020, pp. 305–320.

[123] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7836–7846.

[124] A. Pensia, V. Jog, and P. Loh, "Extracting robust and accurate features via a robust information bottleneck," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 131–144, May 2020.

[125] P. K. Banerjee and G. Montúfar, "The variational deficiency bottleneck," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.

[126] Y. Shen, J. Qin, J. Chen, L. Liu, F. Zhu, and Z. Shen, "Embarrassingly simple binary representation learning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 2883–2892.

[127] Z. Piran, R. Shwartz-Ziv, and N. Tishby, "The dual information bottleneck," *arXiv: 2006.04641*, 2020.

[128] H. Sun, N. E. Pears, and Y. Gu, "Information bottlenecked variational autoencoder for disentangled 3D facial expression modelling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2334–2343.

[129] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, 2019, Art. no. 1181.

[130] A. Zaidi and I. E. Aguerri, "Distributed deep variational information bottleneck," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wirel. Commun.*, 2020, pp. 1–5.

[131] S. Sinha, H. Bharadhwaj, A. Goyal, H. Larochelle, A. Garg, and F. Shkurti, "DIBS: Diversity inducing information bottleneck in model ensembles," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 9666–9674.

[132] F. Bao, "Disentangled variational information bottleneck for multiview representation learning," 2021, *arXiv:2105.07599*.

[133] C. Lee and M. van der Schaar, "A variational information bottleneck approach to multi-omics data integration," in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, 2021, pp. 1513–1521.

[134] J. Song et al., "Multi-feature deep information bottleneck network for breast cancer classification in contrast enhanced spectral mammography," *Pattern Recognit.*, vol. 131, 2022, Art. no. 108858.

[135] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Trans. Multimedia*, vol. 25, pp. 4121–4134, 2022.

[136] S. Wang, C. Li, Y. Li, Y. Yuan, and G. Wang, "Self-supervised information bottleneck for deep multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 32, pp. 1555–1567, 2023.

[137] R. K. Mahabadi, Y. Belinkov, and J. Henderson, "Variational information bottleneck for effective low-resource fine-tuning," in *Proc. Int. Conf. Learn. Representations*, 2021.

[138] B. Li et al., "Invariant information bottleneck for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 7399–7407.

[139] J. Cao, J. Sheng, X. Cong, T. Liu, and B. Wang, "Cross-domain recommendation to cold-start users via variational information bottleneck," in *Proc. IEEE 38th Int. Conf. Data Eng.*, 2022, pp. 2209–2223.

[140] J. Wang et al., "Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation," *Med. Image Anal.*, vol. 83, 2023, Art. no. 102687.

[141] K. K. Roy, A. Roy, A. K. M. Mahbubur, M. A. Amin, and A. A. Ali, "Structure-aware hierarchical graph pooling using information bottleneck," in *Proc. Int. Joint Conf. Neural Netw.*, Shenzhen, China, 2021, pp. 1–8, doi: 10.1109/IJCNN52387.2021.9533778.

[142] J. Gu, Z. Zheng, W. Zhou, Y. Zhang, Z. Lu, and L. Yang, "Self-supervised graph representation learning via information bottleneck," *Symmetry*, vol. 14, no. 4, 2022, Art. no. 657.

[143] L. Yang et al., "Heterogeneous graph information bottleneck," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1638–1645.

[144] R. A. Amjad and B. C. Geiger, "Learning representations for neural network-based classification using the information bottleneck principle," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2225–2239, Sep. 2020.

[145] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[146] A. A. Alemi, I. Fischer, and J. V. Dillon, "Uncertainty in the variational information bottleneck," 2018, *arXiv: 1807.00906*.

[147] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *Proc. Int. Conf. Learn. Representations*, 2020.

[148] S. Voloshynovskiy, O. Taran, M. Kondah, T. Holotyak, and D. J. Rezende, "Variational information bottleneck for semi-supervised classification," *Entropy*, vol. 22, no. 9, 2020, Art. no. 943.

[149] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[150] Y. Ugur, G. Arvanitakis, and A. Zaidi, "Variational information bottleneck for unsupervised clustering: Deep Gaussian mixture embedding," *Entropy*, vol. 22, no. 2, 2020, Art. no. 213.

[151] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 1965–1972.

[152] Q. Wang, C. Boudreau, Q. Luo, P. Tan, and J. Zhou, "Deep multi-view information bottleneck," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 37–45.

[153] J. Wang, Y. Zheng, J. Song, and S. Hou, "Cross-view representation learning for multi-view logo classification with information bottleneck," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4680–4688.

[154] 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.

[155] Z. Wan, C. Zhang, P. Zhu, and Q. Hu, "Multi-view information-bottleneck representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10 085–10 092.

[156] W. Qian, B. Chen, and F. Gechter, "Multi-task variational information bottleneck," 2020, *arXiv: 2007.00339*.

[157] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[158] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6777–6786.

[159] Y. Song et al., "Improving unsupervised domain adaptation with variational information bottleneck," in *Proc. Eur. Conf. on Artif. Intell.*, 2020, pp. 1499–1506.

[160] Y. Du et al., “Learning to learn with variational information bottleneck for domain generalization,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 200–216.

[161] M. Igl et al., “Generalization in reinforcement learning with selective noise injection and information bottleneck,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13 956–13 968.

[162] A. Goyal et al., “Infobot: Transfer and exploration via the information bottleneck,” in *Proc. Int. Conf. Learn. Representations*, 2019.

[163] A. Goyal, Y. Bengio, M. Botvinick, and S. Levine, “The variational bandwidth bottleneck: Stochastic evaluation on an information budget,” in *Proc. Int. Conf. Learn. Representations*, 2020.

[164] R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich, “Learning efficient multi-agent communication: An information bottleneck approach,” in *Proc. Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 9908–9918.

[165] Y. Jin, S. Wei, J. Yuan, and X. Zhang, “Information-bottleneck-based behavior representation learning for multi-agent reinforcement learning,” in *Proc. IEEE Int. Conf. Auton. Syst.*, 2021, pp. 1–5.

[166] J. Fan and W. Li, “DRIBO: Robust deep reinforcement learning via multi-view information bottleneck,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 6074–6102.

[167] G. Xiang, S. Dian, S. Du, and Z. Lv, “Variational information bottleneck regularized deep reinforcement learning for efficient robotic skill adaptation,” *Sensors*, vol. 23, no. 2, 2023, Art. no. 762.

[168] I. J. Goodfellow et al., “Generative adversarial nets,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[169] J. Kim and S. Cho, “Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck,” in *Proc. AAAI Conf. Artif. Intell. Workshop*, 2020, pp. 105–112.

[170] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, “Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow,” in *Proc. Int. Conf. Learn. Representations*, 2019.

[171] I. Jeon, W. Lee, M. Pyeon, and G. Kim, “IB-GAN: Disentangled representation learning with information bottleneck generative adversarial networks,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 7926–7934.

[172] M. Xu, T. Zhang, Z. Li, and D. Zhang, “InfoAT: Improving adversarial training using the information bottleneck principle,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 1255–1264, Jan. 2024.

[173] P. Zhai and S. Zhang, “Adversarial information bottleneck,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 221–230, Jan. 2024.

[174] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[175] A. I. Estella and A. Zaidi, “Distributed information bottleneck method for discrete and Gaussian sources,” in *Proc. Int. Zurich Seminar Inf. Commun.*, 2018, pp. 35–39.

[176] I. E. Aguerri and A. Zaidi, “Distributed variational representation learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 120–138, Jan. 2021.

[177] K. A. Murphy and D. S. Bassett, “The distributed information bottleneck reveals the explanatory structure of complex systems,” 2022, *arXiv: 2204.07576*.

[178] M. Moldoveanu and A. Zaidi, “In-network learning for distributed training and inference in networks,” in *Proc. IEEE Globecom 2021 Workshops*, 2021, pp. 1–6.

[179] M. Moldoveanu and A. Zaidi, “On in-network learning. A comparative study with federated and split learning,” in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wirel. Commun.*, 2021, pp. 221–225.

[180] T. Wu, H. Ren, P. Li, and J. Leskovec, “Graph information bottleneck,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020.

[181] Q. Sun et al., “Graph structure learning with variational information bottleneck,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 4165–4174.

[182] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He, “Recognizing predictive substructures with subgraph information bottleneck,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1650–1663, Mar. 2024.

[183] J. Yu, J. Cao, and R. He, “Improving subgraph recognition with variational graph information bottleneck,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19 374–19 383.

[184] H. S. Witsenhausen, “Indirect rate distortion problems,” *IEEE Trans. Inf. Theory*, vol. 26, no. 5, pp. 518–521, Sep. 1980.

[185] A. Zaidi, I. E. Aguerri, G. Caire, and S. S. Shitz, “Uplink oblivious cloud radio access networks: An information theoretic overview,” in *Proc. Inf. Theory Appl. Workshop*, 2018, pp. 1–9.

[186] Y. Steinberg, “Coding and common reconstruction,” *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4995–5010, Nov. 2009.

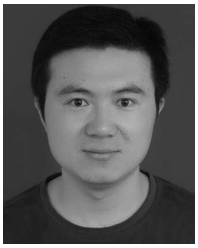
[187] A. D. Wyner, “On source coding with side information at the decoder,” *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.

[188] R. Ahlswede and J. Körner, “Source coding with side information and a converse for degraded broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, Nov. 1975.

[189] I. E. Aguerri, A. Zaidi, G. Caire, and S. S. Shitz, “On the capacity of cloud radio access networks with oblivious relaying,” *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4575–4596, Jul. 2019.

[190] T. Wu, I. S. Fischer, I. L. Chuang, and M. Tegmark, “Learnability for the information bottleneck,” *Entropy*, vol. 21, no. 10, 2019, Art. no. 924.

Shizhe Hu (Member, IEEE) received the BE and PhD degrees in software engineering from Zhengzhou University, China, in 2015 and 2021, respectively. He is an associate research fellow with the School of Computer and Artificial Intelligence with Zhengzhou University. His main research interests include information bottleneck theory, computer vision, and pattern recognition. He has published several papers in peer-reviewed prestigious journals and conferences, such as *IEEE Transactions on Image Processing*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Cybernetics*, CVPR, and ACM MM. He serves as the reviewer for many journals and conferences, such as *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Neural Networks and Learning Systems*, CVPR, and AAAI. More details can be found at https://shizhehu.github.io/.



Zhengzheng Lou, (Member, IEEE) received the PhD degree from the School of Information Engineering, Zhengzhou University, in 2014. He is an associate professor with the Zhengzhou University of School of Computer and Artificial Intelligence. His main research interests include pattern recognition, data mining and information bottleneck theory. He has published several papers in peer-reviewed prestigious journals and conferences, such as *IEEE Transactions on Image Processing* and ACMMM.

Xiaoqiang Yan received the PhD degree from the School of Information Engineering, Zhengzhou University, in 2018. He worked one year as a visiting scholar with the University of Portsmouth, Portsmouth, U.K., in 2017. He is an associate professor with the School of Computer and Artificial Intelligence, Zhengzhou University. He has more than 20 publications in top conferences and journals, such as *IEEE Transactions on Neural Networks and Learning Systems* and IJCAI.

Yangdong Ye (Member, IEEE) received the PhD degree from China Academy of Railway Sciences. He is a professor with the Zhengzhou University of School of Computer and Artificial Intelligence. He worked one year as a senior visiting scholar with Deakin University, Australia. He has wide research interests, mainly including information bottleneck, pattern recognition, knowledge engineering and intelligent system. He has published some papers in peer-reviewed prestigious journals and conferences, such as *IEEE Transactions on Image Processing*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Industrial Informatics*, *Information Fusion*, *Neural Networks*, *Pattern Recognition*, CVPR, AAAI, IJCAI and ACM MM. He serves as the reviewer for many journals and conferences, such as *IEEE Transactions on Neural Networks and Learning Systems*, AAAI and IJCAI.