Deep and Structured Robust Information Theoretic Learning for Image Analysis

Yue Deng, Feng Bao, Xuesong Deng, Ruiping Wang, *Member, IEEE*, Youyong Kong,
and Qionghai Dai, *Senior Member, IEEE*

***Abstract*— This paper presents a robust information theoretic (RIT) model to reduce the uncertainties, i.e., missing and noisy labels, in general discriminative data representation tasks. The fundamental pursuit of our model is to simultaneously learn a transformation function and a discriminative classifier that maximize the mutual information of data and their labels in the latent space. In this general paradigm, we, respectively, discuss three types of the RIT implementations with linear subspace embedding, deep transformation, and structured sparse learning. In practice, the RIT and deep RIT are exploited to solve the image categorization task whose performances will be verified on various benchmark data sets. The structured sparse RIT is further applied to a medical image analysis task for brain magnetic resonance image segmentation that allows group-level feature selections on the brain tissues.**

***Index Terms*— Data embedding, mutual information, deep learning, structured-sparse learning, image classification, brain MRI segmentation.**

I. Introduction

DATA transformation is perhaps the most prevalent and effective approach to be adopted when dealing with real-world image of high dimensionality. Transforming high-dimensional image into a latent space is plausible due to its two prominent advantages in data compression and feature learning. In this paper, we will focus on the discriminative data transformation approaches that incorporate labels into the learning phase. While this task-driven feature learning topic has been discussed in some previous works, there are three important issues that are hardly addressed, or at least not simultaneously considered.

First, in typical embedding framework, the representation and classification functions are trained sequentially. Such training procedures convey no classifier information into the feature learning part. It makes much sense if one can train the classifier and transformation simultaneously to encourage the most suitable features for a particular task. Secondly, for real world problems, the acquisition of labels are very expensive. In cases of insufficient labels, the discriminative learning results cannot capture the whole structures of the dataset. Thirdly, even though plenty of labeled data are available, in some cases, their labels are not definitely reliable. The noisy labels may potentially cause bias in both the features and the classifiers.

To address the aforementioned three challenges, in this paper, we propose a robust information theoretic embedding (RIT) algorithm by exploiting the mutual information as the discriminative criteria. Different from previous works, it simultaneously learns an transformation function and a probabilistic classifier to classify the points in the latent space. The incorporated probabilistic classifier, *i.e.* a multinomial logistic regression [1] (*a.k.a.* soft-max function), does not only encourage the class margins but also defines the probability density function (PDF). Such well-defined PDF facilitate the calculations of the conditional entropy [2] in the latent space that is helpful to detect noisy labels.

The aforementioned RIT learning framework provides a general paradigm for feature learning. It seamlessly works with different data transformation functions to address diverse machine learning tasks. In this paper, as most subspace models, we first consider the most intuitive and basic implementation of RIT with the linear transformation. A toy illustration of this linear RIT model and its corresponding robust embedding results are provided in Fig.1. Apparently, when there are unsupervised and noisy labeled samples, RIT significantly outperforms other methods from both the visualization effects and quantitative evaluations. In addition to this basic version, two other types of sophisticated data transforming strategies will be also considered in the RIT paradigm.

In the first extension, the deep learning (DL) concept [3] is incorporated into the RIT framework. Unlike linear subspace model, DL performs nonlinear transformation on the raw data by a deep neural network. The advantages of deep RIT are mainly concluded as two points. First, it adopts the deep structure to hierarchically transform information from

Manuscript received February 6, 2015; revised July 7, 2015 and June 6, 2016; accepted June 26, 2016. Date of publication July 7, 2016; date of current version July 21, 2016. This work was supported by the National Natural Science Foundation of China under Grant 61327902, Grant 61120106003, and Grant 61379083. The work of Y. Kong was supported by the National Science Foundation of Jiangsu Province, China, under Grant BK20150650. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kiyoharu Aizawa. *(Yue Deng and Feng Bao contributed equally to this work.)*

Y. Deng is with the Tsinghua National Laboratory for Information Science and Technology, Automation Department, Tsinghua University, Beijing 100084, China, and also with the San Francisco Medical Center, School of Medicine, University of California at San Francisco, San Francisco, CA 94158 USA (e-mail: yuedeng.thu@gmail.com).

F. Bao and Q. Dai are with the Tsinghua National Laboratory for Information Science and Technology, Automation Department, Tsinghua University, Beijing 100084, China (e-mail: bf14@mails.tsinghua.edu.cn; qhdai@tsinghua.edu.cn).

X. Deng and R. Wang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: dxuesong7@gmail.com; wangruiping@ict.ac.cn).

Y. Kong is with the School of Computer Science and Engineering, Southeast University, Nanjing 210000, China (e-mail: kongyouyong@seu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2588330

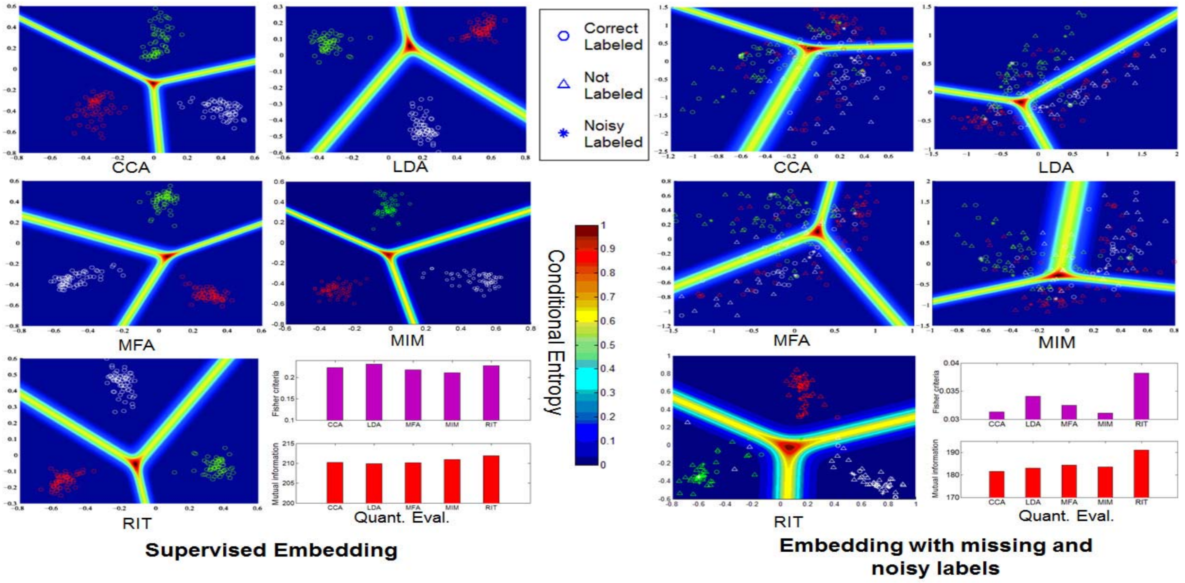


Fig. 1. The embedding results of the faces from three categories in Yale-B dataset. Different colors represent different classes. The left panel are the embedding results for data that are all correctly supervised and the right panel are embedding results with missing labels and incorrect labels. The last figure in each panel quantitatively evaluates the class separability in the embedding space by calculating both the mutual information and the Fisher criteria. For embedding results in the right panel, we use the semi-supervised version of the algorithm if it can be extended with a Laplacian regularization. The background colors represent the conditional entropy for each point in a 2D space. To calculate the conditional entropy for other embedding methods, we first project the data into the latent space and then fit a multinomial logistic regression in the latent space.

layer to layer that is proven to be more effective than the shallow functions. Besides, deep RIT takes the advantage the information theoretic quantity as the learning objective, which could potentially reduce the label uncertainties in the training set. The performances of this deep RIT model will be verified on some image datasets, *e.g.* ImageNet [4] for image categorization.

In the second extension, we enhance the robustness of RIT by introducing the prevalent structured sparse norms into it. Structured sparse norm does not encourage entry-level sparseness as conventional ℓ_1-type sparse problems. Instead, it enhances the sparsity on a group of variables. Such plausible mechanism allows RIT to select more reasonable feature groups in the subspace that sheds light on feature learning sides. In a nutshell, the structured sparse RIT (SS-RIT) exhibits two significant advantages: 1) employing the information theoretic approaches to reduce the uncertainties among labels and 2) incorporating structured sparsity-induced norms for group-level feature learning. We will apply the SS-RIT to a challenging task of brain magnetic resonance image (MRI) segmentation where both label and feature uncertainty occur.

In sum, the contributions of this paper are mainly summarized as two-folds:

- We present an information theoretic learning framework that is able to conduct feature learning and classification jointly. The RIT model is robust to the uncertainties in the training data and could achieve reliable performances even though with missing and noisy labels.
- The proposed RIT is a flexible feature transformation framework that seamlessly works with different types of feature transformations, *e.g.* deep and structured learning, to cope with diverse practical problems.

The remaining of the paper is organized as follows: Section II reviews related works on discriminative feature learning. The proposed RIT learning framework is introduced in Section III and its detailed implementations with various feature learning functions are discussed in Section IV. Section V evaluates the performances linear RIT, Deep RIT and sparse structured RIT on different tasks. The paper is concluded in Section VI.

II. Related Works

Subspace models are widely used in the machine learning field for data representation. Statistic methods, such as Principle Component Analysis (PCA) [5] , Linear Discriminant Analysis (LDA) [6], Canonical Correlation Analysis (CCA) [7] are early attempts. Manifold learning [8], [9] and its variants [10], [11] find projections that optimally preserves the graph distances of high dimensional data. The graph structure enables the exploration of various nonlinear graph-based similarities, *e.g.* commute time [11], to describe the intrinsic relationship among data. Marginal Fisher Analysis (MFA) is the discriminative manifold learning method that extends Fisher discriminant criteria into a manifold space [9]. Discriminative Locality Alignment (DLA) [12] combines the locality similarity metric and the marginal sample weighting strategy. DKA more reasonably utilizes the local information of data and thus leads to more robust performance. Generalized Multiview Analysis (GMA) [13] seeks an optimal subspace by solving a quadratic constrained quadratic program (QCQP) over different subspaces (*e.g.* PCA, LDA, MFA). The promises of these generalized models have been witnessed in a number of benchmark datasets.

Unlike subspace based method, dictionary-based models do not impose orthogonal restrictions on the projections, allowing more flexibility to adapt the representation to the data. Within the dictionary learning framework, some priors can be placed on the dictionary to encourage the desired data structure. Widely used priors include discriminative structure [14], sparse [15]–[17] and structured sparse [18]. In [19], discriminative dictionaries are generated by incorporating a discriminative function into the task-driven framework.

Deep learning is an emerging technique which has wide influences in the big-data industries [3], [20], [21]. It tackles a fundamental problem in machine learning: how to generate informative features in a task-driven manner. While the deep learning has been extensively used to reduce the noises in the raw data, less efforts have been devoted to handle the noises in the labels. In fact, label uncertainty is meanwhile a critical problem that needs careful considerations. Different from existing works, in this paper, we introduce a new learning objective into the typical DL framework from a novel perspective of information theory.

In the community of machine learning, the information theoretic quantities have been used for data representation [22], clustering [23] and feature selection [24]. For data embedding, Mutual Information Maximization (MIM) has been proposed to extract features in a discriminative manner [25]. Its formulation resembles the Fisher discrimination but defines the discriminant via information quantity. Although both [25] and our RIT model share one common thing in utilizing mutual information as the discriminative criteria, the two algorithms are quite different. In MIM, all the probability density function are estimated via the nonparametric way. Therefore, MIM is a fully supervised embedding method and is sensitive to the quality of labels. RIT provides a more flexible way to interpret the probabilities with a probabilistic classifier which can be easily extended to semi-supervised version and is robust to noises in the given labels. Our RIT model was inspired by the RIM work in [23] on information theoretic function design. However, RIM just considers discriminative clustering in the original data space without any feature learning mechanism involved. The major concern of RIT is about feature learning. It considers generating more reasonable feature representations to enhance the discriminative structure in a transformed space. In detail, we will consider three data transformation functions in this work including subspace, deep and structured sparse transformations.

III. Robust Information Theoretic Learning

In this part, we will introduce the robust information theoretic embedding (RIT) model and its solutions.

A. Model

For a flexible description, we adopt a probabilistic framework to address the task of discriminative learning. In probability theory and information theory, the mutual information is a quantity that measures the mutual dependence of the two random variables. It measures how much knowing one of these variables reduces uncertainty about the other. Therefore, in our formulation, the mutual information serves as the basic discriminant criteria to measure the class separability in the transformed space.

For the ease of illustration, we define $\mathbf{x}_i \in \mathbb{R}^n$ as the original data obtained in real world and $\mathbf{y}_i = g(\mathbf{x}_i) \in \mathbb{R}^m, m < n$ as the corresponding point of $\mathbf{x}_i$ in the latent space. $g(\cdot)$ is a transformation. $l_i = k$ means that the i^{th} point belongs to the k^{th} class, $k = 1 \dots C$. In Shannon's information theory, the mutual information of latent points and labels, *i.e.*, $I(\mathbf{L}, \mathbf{Y})$, can be expressed in the following form,

$$\begin{aligned} I(\mathbf{L}, \mathbf{Y}) &= H(\mathbf{L}) - H(\mathbf{L}|\mathbf{Y}) \\ &= -\int p(l) \log p(l) dl + \iint p(l, \mathbf{y}) \log p(l|\mathbf{y}) d\mathbf{y} dl \end{aligned} \tag{1}$$

where $H(\cdot)$ denotes the entropy.

As shown in (1), the mutual information can be expanded as the summation of two entropy terms. The conditional entropy $H(\mathbf{L}|\mathbf{Y})$ reveals the total uncertainty of labels by observing the latent features. Therefore, it should be minimized. As indicated in [23] and [26], this conditional entropy implicitly represents the margins between different classes. A small conditional entropy corresponds to a large margin. Besides, the entropy $H(\mathbf{L})$ encodes the label distribution which is always maximized in semi-supervised learning to avoid label bias on some specific classes.

The estimation of information theoretic quantities depends on probability density function (PDF) of transformed data and labels. In the RIT model, we assume that the data in the latent space can be well separated by a probabilistic classifier. In machine learning, one extensively used probabilistic classifier is the multinomial logistic regression (MNL). Without the loss of generality, in this paper, we exploit the MNL in the RIT formulation. We assume there are C classes in total and get C pairs of $\theta_j = (\mathbf{w}_j, b_j)$ in the parameter space of the MNL. It is worth noting here that l is not the supervised label. In fact, it is the label assigned by the MNL in the latent space. With the MNL, the conditional probability $p(l_i = k|\mathbf{y}_i)$ is explicitly defined,

$$p_{ik} = p(l_i = k|\mathbf{y}_i) = \frac{\exp(\mathbf{w}_k^T \mathbf{y}_i + b_k)}{\sum\limits_{j=1}^{C} \exp(\mathbf{w}_j^T \mathbf{y}_i + b_j)}. \tag{2}$$

According to (1), it is obvious that this kind of implicit labels will be integrated out in the calculation of the mutual information. Of course, there is another kind of supervised labels, *i.e.*, l^s, which are explicitly provided by the user. Accordingly, we treat data as two kinds regarding whether their labels are explicitly given or not. For training, we assume there are N feature vectors in total, *i.e.* $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2 \dots \mathbf{y}_N\}$. Among these N data points, we get t supervised features, *i.e.* $\mathbf{Y}^s = \{\mathbf{y}_1^s, \mathbf{y}_2^s \dots \mathbf{y}_t^s\} \subset \mathcal{S}$ with their labels explicitly provided as $\mathbf{L}^s = \{l_1^s, l_2^s \dots l_t^s\}$.

We define $\mathbf{X} \in \mathbb{R}^{n \times N}$ are the original feature and $\mathbf{Y} \in \mathbb{R}^{m \times N}, m < n$ are the points in the latent space. $g(\cdot)$ is the mapping or data transformation function. In this part, we consider the most widely used linear transformation, *i.e.*, $\mathbf{Y} = \mathbf{\Omega X}, \mathbf{\Omega} \in \mathbb{R}^{m \times n}$. More general data transformations will be discussed in Section IV. Accordingly, we give the general form of the robust information theoretic embedding (RIT) model,

$$\begin{aligned} &\min \quad -I(\mathbf{L},\mathbf{Y}) - \lambda\mathcal{C}(\mathbf{L}^s,\mathbf{Y}^s) \\ &s.t. \;\; \mathbf{Y} = g(\mathbf{X}) \end{aligned} \tag{3}$$

In the objective function of (3), the first term is the mutual information of all the data no matter whether they are supervised or not. The second term $\mathcal{C}(\mathbf{L}^s,\mathbf{Y}^s)$ is the regularization by penalizing the loss of the probabilistic classifier which only involves supervised data and their labels. We will discuss its expression and effectiveness in the next subsection in Eq.5. Till now, why our RIT model naturally handles semi-supervised embedding tasks is self-evidently. It utilizes all the samples (both supervised or not) in the mutual information term and the supervised information are further penalized in the second term.

B. RIT With Noisy Labels

RIT is meanwhile very robust to the noises in the supervised labels. This desired advantage owes to two points. Firstly, the objective of RIT does not only over-fit the losses of supervised data. In addition to the logit loss $\mathcal{C}(\cdot)$ in (3), RIT simultaneously seeks for a balance to maximize the mutual information term, which does not rely on the supervised label. It is conceivable that an over-fitted logistic machine may achieve a good score on the loss of $\mathcal{C}(\mathbf{L}^s,\mathbf{Y}^s)$. However, such a bad logistic classifier learned from noisy labels may not achieve a good score on the mutual information. Therefore, the mutual information term alleviates the disturbance of the noisy labels.

Secondly, the MNL itself could also contribute to alleviating the noisy labels. In detail, $p(l_i^s|\mathbf{x}_i^s)$ exactly reveals the uncertainty of the supervised labels by observing the features. It is conceivable that a well trained MNL could not fit all the data perfectly. Particularly, it hardly fits the outliers in the training set. Therefore, the noisy labeled data generally exhibit small conditional probability implying that they cannot be well explained by the current MNL. Accordingly, following the idea in [27], we can define the weight $\phi_i = p(l_i^s|\mathbf{x}_i)$ for the i^{th} supervised sample and incorporate this quantity to design a weighted MNL,

$$\mathcal{L}(\mathbf{L}^s,\mathbf{Y}^s) = \prod_{i=1}^{t} (p_{ik})^{\phi_i}, \tag{4}$$

The function $\mathcal{L}$ is not a likelihood in the usual sense; but it has much general meaning to alleviate the disturbances of outliers in the training set. With the weighted likelihood, we get its log-likelihood expression and the cost function is subject to the following equation, *i.e.*

$$\mathcal{C}(\mathbf{L}^s,\mathbf{Y}^s) = \sum_{i=1}^{t} C_i = \sum_{i=1}^{t} \phi_i \log p_{ik}. \tag{5}$$

From the log-likelihood, obviously, when ϕ_i is small, the i^{th} sample contributes less to the global cost. C_i is the loss that the i^{th} supervised sample contributed to the global objective. On the contrary, a large weight enhances the effectiveness of the i^{th} sample to the optimization. Therefore, to make a robust embedding, it is plausible if we can denote small weights to the samples whose labels are wrongly supervised. Fortunately, within the probabilistic framework, it is possible to define such kind of weight by the conditional probability returned by the MNL.

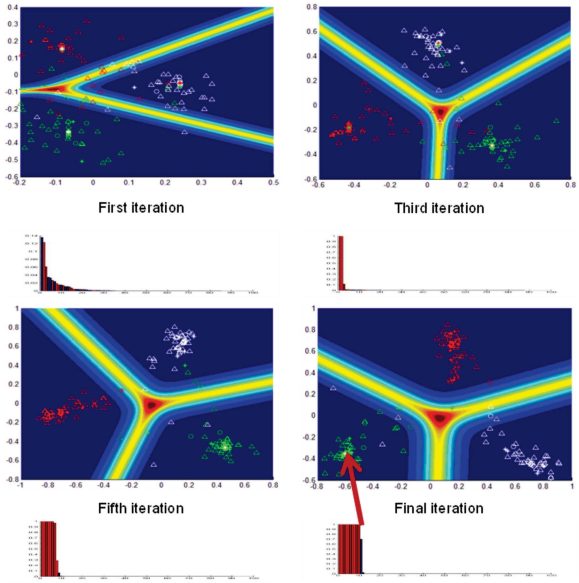


Fig. 2. The embedding results and the conditional probability for supervised data in different iterations of RIT optimization for the toy demo discussed in Figure 1. In each subfigure, the histograms report $1 - p(l_i^s|\mathbf{x}_i^s)$ of each labeled point which are arranged in a decreasing order. The red bars indicate the noisy supervised data. Blue bars denote the samples whose labels are correctly supervised.

The weight is updated along with the processing of the whole RIT optimizations. Till now, the cost $\mathcal{C}(\mathbf{L}^s,\mathbf{Y}^s)$ exactly corresponds to the general losses used in weighed logistic regression. In the optimization, these weights are dynamically updated and the whole optimization is cast to a sequence of reweighted programming. The details of the optimization and the weight updating procedures are provided in Fig.2.

From Fig.2, it is obvious that with the processing of the iterations, the noisy labeled data are automatically identified by our algorithm (see the red bars). According to Eq.4, it is apparent that these noisy data may contribute little to fit the logistic regression and their effectiveness to discrimination are only represented in the mutual information term that does not rely on the supervised labels. However, our algorithm cannot perfectly alleviate all the disturbances of noisy labels, it is found from the last subfigure in Fig.2 that one noisy point (indicated by the arrow) is still embedded to a wrong place.

C. RIT Subspace Model

We show the optimization of the RIT model in this part. With p_{ik} defined in (2) and $p_k = \frac{1}{N}\sum_{i=1}^{N} p_{ik}$, we give the empirically estimation of the mutual information term that,

$$I(\mathbf{L},\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} I_i = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}\{p_{ik}[\log p_k - \log p_{ik}]\} \tag{6}$$

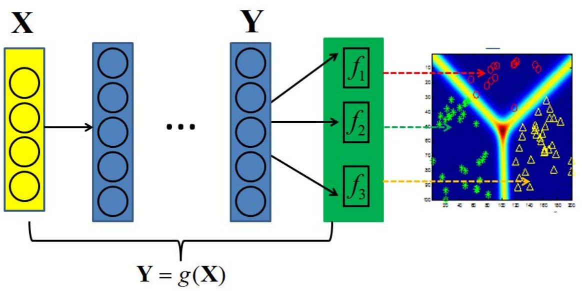


Fig. 3. The schematic illustration of the deep RIT (DRIT) model by exploiting deep neural network as the data transformation function $g(\cdot)$. In the classification layer (green layer) of the DNN, the classifier assigns different points to its corresponding category by maximizing the information theoretic term in (3) as the learning objective.

The term I_i defines the loss that the i^{th} sample contributed to the mutual information [23]. For term $\mathcal{C}(\cdot)$, we choose it to be the log-weighed-likelihood of the losses of MNL. As discussed previously, the optimization involves two variables, *i.e.* the data transformation $\mathbf{\Omega}$ and the MNL parameters $\theta_i = (\mathbf{w}_i, b_i), i = 1..C$.

The gradient of the two terms with respect to $\mathbf{\Omega}, \mathbf{w}$ and b can all be derived through chain rule. For example, $\frac{\partial I}{\partial \Omega} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}[\log\frac{p_k}{p_{ik}}+1]\frac{\partial p_{ik}}{\partial \Omega}$ and $\frac{\partial \mathcal{C}}{\partial \Omega} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_{il_i^s}}\frac{\partial p_{il_i^s}}{\partial \Omega}$. This chain rule is also applied to the derivatives for $\mathbf{w}$ and b. The only modification is to change the partial derivative of $\mathbf{\Omega}$ to be the partial derivatives with $\mathbf{w}$ and b, respectively.

After getting the derivatives, the whole RIT optimization is solved in an alternating framework. We denote the objective in (3) as $f(\mathbf{X}, \mathbf{L}^s|\mathbf{\Omega}, \theta, \Phi)$, where Φ is the weight matrix. The updating rule is provided by $\theta^{k+1} = \arg\min f(\mathbf{X}, \mathbf{L}^s|\mathbf{\Omega}^k, \theta, \Phi^k)$ and $\mathbf{\Omega}^{k+1} = \arg\min f(\mathbf{X}, \mathbf{L}^s|\mathbf{\Omega}, \theta^k, \Phi^k)$. Both the updating of $\mathbf{\Omega}$ and θ depend on the gradient descent method and we use the L-BFGS quasi-Newton optimization algorithm[1] to get a fast and robust convergence.

IV. RIT Extensions

In the previous part, the general paradigm of RIT subspace model with linear transformation has been discussed. In fact, RIT is a robust information theoretic feature learning framework that works friendly with many kinds of data transformation functions. As extensions, we will introduce other two types of prevalent data transformation strategies into the RIT framework from the perspectives of deep learning and structured sparse learning.

A. Deep RIT

In this part, we show how to incorporate deep learning concepts into RIT to improve the performances of feature learning. A schematic summarization of the deep RIT (DRIT) model has been provided in Fig.3 which is mainly composed of two parts of feature transformation (blue layers) and task-driven learning parts (green layer).

At a glance, the deep neural network (DNN) plays the role of a mapping function $g(\cdot)$ in Eq.3 that transforms the input data/image (yellow layer $\mathbf{X}$) into a high-level representation $\mathbf{Y}$. We follow the general principle in the field to define the activations of the neural network. In details, the j^{th} node of the $(l)^{th}$ layer is connected to the nodes on the $(l-1)^{th}$ layer with parameters $\theta^{(lj)} = \{\mathbf{w}^{(lj)}, \mathbf{b}^{(lj)}\}$, *i.e.*

$$o^{(lj)} = \gamma\,(a^{(lj)}), a^{(lj)} = \mathbf{w}^{(lj)}\mathbf{o}^{(l-1)} + b^{(lj)}, \tag{7}$$

where $\gamma(\cdot)$ is a nonlinear mapping; $\mathbf{w}^{(lj)}$ and $b^{(lj)}$ are the weights and bias. The number of hidden layers of this deep representation part can be large. In the image analysis tasks, the dimensions of input data are very high and the convolutional operations are always used in the first coupe of layers of the DNN. In this part, we do not prefer to specify the detailed DNN structure because DRIT works well with many types of deep configurations. The specific DNN setting that is used in this work for the image categorization will be discussed in the corresponding experimental part.

We remark on the differences between DNN transformation and the linear transformation in the RIT subspace model. In the previous linear model, only a mapping matrix $\mathbf{\Omega}$ is optimized while this deep transformation involves millions of hidden parameters. Such a large amount of parameters allow hierarchical transformations that could potentially increase the chances to get a better representation $\mathbf{Y}$ on the last representation layer. On the other hand, deep training also imposes great computational burdens and requires tons of training samples.

After obtaining the latent representation $\mathbf{Y}$ of the DNN, a multinomial logistic regression layer (green layer) is connected to it for data classification. In Fig.3, for simplicity, only three classes and their corresponding classifiers $f_1, \ldots f_3$ are shown. In typical DNN training, the objective function always directly minimizes the logistic loss to make minimal prediction error on the training set. In the DRIT model, we further utilize the information theoretic quantity defined in (3) to reduce the uncertainties in the supervised labels. Such objective generalizes all the nice properties of the RIT learning to this deep learning framework. More importantly, similar as the methods in Section III-B, DRIT also exhibits the plausible mechanism to detect wrongly supervised labels in the training set which will be experimentally discussed later.

While DRIT exploits the RIT as the learning objective, typical back-propagation (BP) algorithm [28] still easily applies to solve it. To note, we have denoted two terms, *i.e.* $\mathcal{C}(\cdot)$ and $I(\cdot)$ in the RIT objective. These two terms should be simultaneously considered in the BP process to adjust parameters in DNN. In general, the gradient for the parameter in DNN can be determined by the following additive formulation.

$$\begin{aligned}\frac{\partial C}{\partial \theta^{(li)}} &= \sum\nolimits_i \delta(i \in \mathcal{S})\,\underbrace{(\frac{\partial C_i}{\partial o_i^{(lj)}})}_{BP}\frac{\partial o_i^{(lj)}}{\partial a_i^{(lj)}}\frac{\partial a_i^{(lj)}}{\partial \theta^{(lj)}} \\ &\quad + \sum\nolimits_i \underbrace{(\frac{\partial I_i}{\partial o_i^{(lj)}})}_{BP}\frac{\partial o_i^{(lj)}}{\partial a_i^{(lj)}}\frac{\partial a_i^{(lj)}}{\partial \theta^{(lj)}}. \end{aligned}\tag{8}$$

[1] We used the public optimization package "minFunc" at http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html.

The above gradient for parameter $\theta^{(lj)}$ is easily verified according to the chain rule. The terms in the brackets come from error back propagation (BP) and the remaining terms out of the brackets are easily calculated with the matrix-form derivations. In the above formulation, we have used $\mathcal{S}$ to denote the set of supervised data points and $\delta(\cdot)$ is the indicator function which is 1 iff. $i \in \mathcal{S}$ and 0 otherwise. I_i and C_i have been defined in (1) and (5), respectively. In practical training, we adopt the stochastic gradient descent strategy to update the parameters with parallel computing [29].

B. Structured Sparse RIT

In many practical problems, the data themselves exhibit certain structure [30]. In this part, we will show how to effectively exploits this feature structure information to improve the performance of RIT. From a general view, the linear RIT model is comparable with a regression problem with each dimension in the latent space sharing its own regression vector. To support this claim, we recall the linear transformation in the RIT model that for any point $\mathbf{x}_i \in \mathbb{R}^n$, there exists a mapping $\mathbf{y}_i = \mathbf{\Omega}\mathbf{x}_i$. We know that $\mathbf{y}_i = [y_{i1}, y_{i2} \dots y_{im}] \in \mathbb{R}^m$. Consequently, it is straightforward to get a regression-type formulation that $y_{iu} = \omega_u \mathbf{x}_i, \forall i = 1 \dots N$ where $\omega_u \in \mathbb{R}^{1\times n}$ is the u^{th} row of $\mathbf{\Omega}$, *i.e.* $\mathbf{\Omega} = [\omega_1^T \dots \omega_u^T \dots \omega_m^T]^T$.

From the discussions aforementioned, it is apparent that all the attributes in the feature vector $\mathbf{x}$ will contribute to the final embedding. In statistic, it has been widely investigated that such a dense regression is not the optimal one in most cases. In machine learning, one prevalent approach is to place sparse priors on the regression parameter to further improve the prediction accuracy of the model. Intuitively, sparse learning assumes that only a portion of factors in the original feature vector contribute to the learning process. The incorporation of the sparse norms facilitate automatic feature selection and alleviate the over-fitting problem for training data to a large extend.

However, in many cases, only constraining the sparseness of the factors does not seem appropriate because the considered factors are not only expected to be sparse but also to have a certain structure [18]. Therefore, structured sparsity-inducing norms are now drawing more and more attentions in the community of machine learning, *e.g.* in grouped lasso. Therefore, it is nontrivial to incorporate structured regularization into the RIT model for the sake of better data interpretation.

Before introducing structured sparse RIT (SS-RIT), we will first introduce some sparse norms that play very critical roles in sparse learning. First, we define the general ℓ_p norm for a vector $\mathbf{a} \in \mathbb{R}^r$ as $\|\mathbf{a}\|_p = (\sum_i^r |a_i|^p)^{\frac{1}{p}}$. $\|\mathbf{a}\|_0$ denotes the ℓ_0 norm that counts the number of non-zero elements in $\mathbf{a}$. However, ℓ_0 norm is discrete and is analytically intractable. Therefore, its convex envelope, *i.e.* ℓ_1 norm is extensively used as a convex surrogate for sparse learning [31]. Based on the ℓ_1 norm, we will introduce the ℓ_1/ℓ_2 norm for structured sparse learning.

For the ease of explanation, we divide r dimensions of $\mathbf{a}$ into $|\mathcal{G}|$ overlapping groups, *i.e.* $G = 1, \dots |\mathcal{G}|$, which implies that one attribute a_k can be assigned to different groups. We define $d_j^G > 0$ as the weight for the j^{th} variable in the G^{th} group. $d_j^G = 0$ means that the j^{th} attribute is excluded from the G^{th} group. Accordingly, the structured sparsity inducing norm [32] can be defined as,

$$\Psi_{\mathcal{G}}(\mathbf{a}) = \sum_{G\in\mathcal{G}} \left\{ \sum_{j\in G} (d_j^G)^2 |a_j|^2 \right\} = \sum_{G\in\mathcal{G}} \|d_G \circ \mathbf{a}\|_2, \quad (9)$$

The operator $\circ$ is the component-wise product. The norm in (9) is called ℓ_1/ℓ_2 norm because it encourages sparse selections at the group level and, in each group, the variables are densely penalized by a ℓ_2 norm.

Moreover, as indicated in [18], ℓ_1 norm is a specific case of (9) when $\mathcal{G}$ is the set of all singletons and with all the weights setting to 1. Accordingly, we present the general formulation of the structured sparse RIT model (SS-RIT) in (10) and the sparse RIT with ℓ_1 regularization is only a specific case of SS-RIT.

$$\min f(\mathbf{X}, \mathbf{L}^s|\mathbf{\Omega}, \theta) + \mu \sum_{u=1}^{m} \Psi_{\mathcal{G}}(\omega_u) \quad (10)$$

In the above formulation, the first term is the loss of the RIT model in (3) and the second term is the structured sparse regularization of the projecting vector. ω_u is the u^{th} row of the linear transformation $\mathbf{\Omega}$. With such a norm, it is apparent that each dimension y_u in the latent space is only associated with a number of attributes of $\mathbf{x}$ in the selected groups.

The optimization of SS-RIT is almost the same as the the solutions to the RIT model which depends on the alternation between $\mathbf{\Omega}$ and θ. For the logistic parameters, they are irrelative to the added structured norms and thus the updating rule for RIT still applies to it without any change. However, in SS-RIT, the gradient of $\mathbf{\Omega}$ now involves an extra term, *ı.e.* $\Psi_{\mathcal{G}}(\omega_u)$. To handle this structured sparsity-inducing norm, we follow the work in [18] and [32] to introduce the variational variable η and solve the SS-RIT in a reweighted manner.

As following the result in [32], we have the following lemma,

Lemma 1: For any matrix $\mathbf{x}$, its ℓ_1 norm is equivalent to the following problem with a variational variable $\mathbf{z}$,

$$2 \parallel \mathbf{x} \parallel_{\mathbf{1}} = \min_{\mathbf{z}\in\mathbf{R}^{\mathbf{p}}} \sum_{\mathbf{z}=\mathbf{1}}^{\mathbf{p}} \frac{\mathbf{x}_{\mathbf{j}}^{\mathbf{2}}}{\mathbf{z}_{\mathbf{j}}} + \parallel \mathbf{z} \parallel_{\mathbf{1}}, \quad (11)$$

whose minimum is uniquely obtained for $z_j = |x_j|$.

Following (11), by defining η_u^G as the variational variable, $2\sum_{u=1}^m \Psi_{\mathcal{G}}(\omega_u)$ can be reformulated into the following variational form,

$$2\sum_{u=1}^{m} \Psi_{\mathcal{G}}(\omega_u) = \min_{(\eta_u^G)_{G\in\mathcal{G}}} \sum_{u=1}^{m} [\parallel (\eta_u^G)_{G\in\mathcal{G}} \parallel_1 + \sum_{G\in\mathcal{G}} \parallel \omega_u \circ d^G \parallel_2^2 (\eta_u^G)^{-1}], \quad (12)$$

By merging the variational variable and the ℓ_2 norm together, Eq. 12 can be rewritten in turn as

$$\min_{(\eta_u^G)_{G\in\mathcal{G}}} \sum_{u=1}^{m} \omega_u Diag(\zeta)\omega_u^T + \parallel (\eta_u^G)_{G\in\mathcal{G}} \parallel_1,$$

TABLE I

THE CLASSIFICATION RESULTS ON DIFFERENT DATASETS WITH DIFFERENT DATA EMBEDDING ALGORITHMS AND CLASSIFIERS (ACCURACY ± STD %)

	Nearest Neighbor Classifier			SVM			Logistic Regression		
Methods	Yale-B	Scene-15	COIL-100	Yale-B	Scene-15	COIL-100	Yale-B	Scene15	COIL-100
Raw	67.6±2.8	55.1±2.4	62.2±2.1	89.6±2.1	75.1±1.5	81.7±2.2	88.7±2.6	74.6±2.3	83.4±1.9
PCA	78.6±1.7	62.1±2.6	73.2±1.5	88.5±1.7	71.5±2.5	80.6±1.3	87.3±1.6	70.6±2.4	81.1±1.5
LDA	89.6±1.3	65.6±2.1	75.7±1.6	92.7±1.2	74.2±2.1	82.9±1.4	92.2±1.3	75.3±2.1	83.1±1.3
CCA	90.3±1.6	66.4±2.3	73.8±1.7	92.3±1.3	75.9±2.2	83.7±1.5	92.7±1.4	76.1±2.2	83.8±1.4
MFA	91.7±1.4	68.2±2.2	74.1±1.5	92.9±1.3	76.8±2.3	83.3±1.6	93.1±1.5	76.3±2.2	83.3±1.5
MIM	90.7±1.7	66.1±2.7	76.1±1.5	93.3±1.6	75.6±2.4	83.9±1.7	92.5±1.7	76.7±2.3	82.7±1.8
TSC	89.9±1.5	67.2±2.4	77.6±1.7	94.5±1.2	76.5±2.5	85.3±1.4	93.8±1.3	77.8±2.4	85.8±1.6
DLA	91.2±1.3	67.3±1.9	75.1±1.3	93.7±1.2	76.7±2.4	85.7±1.7	93.7±1.3	78.3±2.5	84.2±1.5
GMLDA	89.2±1.1	68.7±2.1	78.1±1.9	92.1±1.1	76.6±2.3	86.7±2.1	92.7±1.2	78.3±2.4	88.2±1.7
GMMFA	89.5±1.1	67.9±2.3	77.9±2.3	91.8±1.4	77.3±2.4	87.1±2.1	93.4±1.3	78.7±2.6	87.2±1.4
RIT	91.3±1.6	68.9±2.6	77.2±1.5	93.9±1.3	77.3±2.3	86.2±1.5	94.3±1.2	79.8±2.4	87.3±1.3

where the j^{th} element in vector ζ is $\zeta_j^t = \sum_{G\in\mathcal{G}}(d_j^G)^2(\eta_u^G)^{-1}$, $j = 1\dots n$. $Diag(\cdot)$ is an operation to write a vector into a diagonal matrix. Till now, the SS-RIT optimization in (10) is subject to the following variational optimization,

$$\min f(\mathbf{X}, \mathbf{L}^s|\mathbf{\Omega}, \theta) + \frac{\mu}{2}\left\{\sum_{u=1}^{m}\omega_u Diag(\zeta^t)\omega_u^T + \parallel(\eta_u^G)_{G\in\mathcal{G}}\parallel_1\right\}. \tag{13}$$

The problem in (13) is not jointly convex and led themselves well to simple alternating optimization scheme between $\mathbf{\Omega}$,$(\eta_u^G)_{G\in\mathcal{G}}$ and θ.

The updating rules of $\mathbf{\Omega}$ and θ is trivial following the gradient descend method. The updating rules of of the variational variable $\{(\eta_u^G)_{G\in\mathcal{G}}\}$ is given in lemma 1. In practice, $\{(\eta_u^G)_{G\in\mathcal{G}}\}$ is provided by

$$\{(\eta_u^G)_{G\in\mathcal{G}}\}^{k+1} \leftarrow \max\{\|\omega_u^k \circ d^G\|_2, \epsilon\}, \tag{14}$$

where $\epsilon \ll 1$ to avoiding numerical instability near zero. $\omega_u^k \in \mathbf{\Omega}^k$ is the u^{th} row of the optimal $\mathbf{\Omega}^k$ obtained in the k^{th} iteration. Till now, the whole SS-RIT optimization can be solved following the steps of alternating optimization. The whole optimization is regarded as converged when $\frac{\|\mathbf{\Omega}^{k+1}-\mathbf{\Omega}^k\|_F^2}{\|\mathbf{\Omega}^k\|_F^2} < 10^{-3}$.

V. RIT FOR IMAGE CATEGORIZATION

A. RIT Subspace Model

In this part, we investigate the performances of RIT on three benchmark image datasets including Yale-B face dataset [33], fifteen-scene dataset [34] and the COIL-100 dataset [35]. In Yale-B face dataset, we simply use the cropped images in [33] and [36] and resize them to 32 × 32 pixels. This dataset now has 38 individuals and around 64 near frontal images under different illuminations per individual. Fifteen scene dataset contains images from fifteen categories including both indoor and outdoor pictures. The COIL dataset contains the images of 100 objects from multi-views [35].

In the Yale-B dataset, we use the gray-scale pixel values on the raw face images to generate the feature vector. For the scene and object dataset, we follow the bag-of-feature method to extract visual features. In a nutshell, to describe an image, we use a grid-based method to extract the dense SIFT features. The dense SIIF features [37] are extracted on 16 × 16 pixel patches sampled every 8 pixels. To generate features for fifteen scene and COIL dataset, the local sift features are assigned to a codebook with 1024 codewords by the kernel assignment [38] and lead to a final feature vector of $\mathbb{R}^{1024}$. For multi-view models, we also considered the gist feature [39] as another view.

For comparison purpose, we pit RIT against many prevalent subspace models including statistic methods (*e.g.* PCA [40], LDA [5] , CCA [7]), DLA [12] and GMA [13] with LDA and MFA (termed as GMLDA and GMMFA), graph-based methods (*e.g.* MFA [9]), information theoretic learning (MIM [25]) and task driven sparse coding (TSC) with logistic regression as the objective [19]. For the ease of computational efficiency, before discriminative embedding, the original large feature vectors are pre-processed by PCA to a low dimensional subspace where 90% energy are preserved. In the implementation of RIT model, we fit $\lambda = 0.1$ and the learning procedures are regarded as converged when the changes of the objective is less than 10^{-4}. For multiple class categorization task, we follow the idea in [19] to train the model with the one-versus-all strategy. The experimental validations are divided into three parts, *i.e.* supervised embedding, semi-supervised embedding and embedding with noisy labels.

In the first test, we investigate the performance of RIT in a definitely supervised fashion. For training purpose, 30 samples per class in Yale-B dataset, 100 samples in each class of fifteen scene dataset and 40 images in COIL dataset are randomly selected as training samples. The rest images are used for testing and the experiments are repeated for 10 times. After data embedding, for data classification, we test three classifiers including Nearest Neighbors Classifier, Support Vector Machine (SVM) and logistic regression. The best classification results of data embedding methods with different classifiers are reported in Table I. In Table I, the first row reports the classification accuracies on raw data as a comparison baseline. It is interesting to note, although less data dimensions are used in the latent space, the classification accuracies are even improved. This improvement owes to feature learning mechanism of discriminative data embedding [7], [25].

We consider the performances of RIT model to conduct semi-supervised discriminative embedding. As stated in Section III, RIT naturally embraces the unlabeled samples

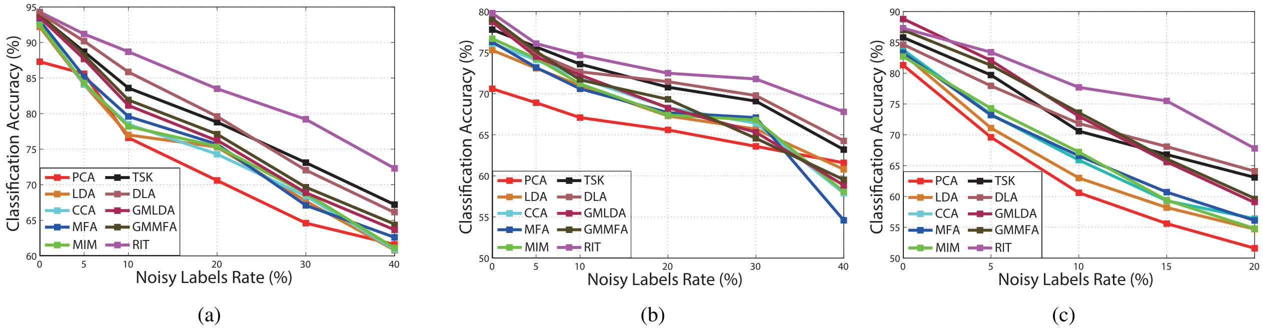


Fig. 4. The classification accuracy of different data embedding method with noisy labels. (a) Yale-B. (b) Fifteen Scene. (c) COIL-100.

TABLE II

THE CLASSIFICATION RESULTS OF SEMI-SUPERVISED DISCRIMINATIVE EMBEDDING (ACCURACY ± STD%)

Methods	Yale	Fifteen Scene	COIL-100
SDA	94.6±1.3	77.3±1.7	84.2±1.4
MIM+SDA	93.8±1.0	77.7±1.5	84.3±1.0
LapMIM	94.2±1.1	78.8±1.4	87.6±1.1
TSC	94.1±0.6	78.3±1.0	86.5±0.7
SDLA	95.5±1.2	80.2±0.9	88.2±1.2
RIT	96.3±0.9	81.6±1.3	89.7±0.9

into the model with the mutual information maximization. For comparisons, we compare RIT with other semi-supervised discriminative embedding methods. The first competitor is semi-supervised discriminant analysis (SDA) [6] that resembles LDA but places a Laplacian term to encourage unlabeled points staying very close to the similar labeled points. The MIM model is a fully supervised model and we propose two possible ways to extend MIM to a semi-supervised version. First, we can initialize nonconvex MIM with the optimal projections learned by SDA (SDA+MIM). Besides, it is also possible to extend MIM to a semi-supervised version by incorporating a Laplacian term (LapMIM). Task driven sparse coding (TSC) [19] is straight-forward to be extended to the semi-supervised by using the labeled data in the classifier and keeping all the unlabeled data in the reconstruction term. DLA can also incorporate the unlabeled samples in the alignment stage [12] and lead to the semi-supervised DLA (SDLA).

The results of semi-supervised learning results are reported in Table II by using the same feature and training samples as in Table I. The classifier used in this test is the Logistic Regression. From the results, it is interesting to find the performances of discriminative embedding are further improved with some unlabeled points. By comparing the results with the supervised embedding results in Table I, it is noted that semi-supervised-based embedding results exhibit smaller standard deviation.

From the experiments presented above, we find that among all the semi-supervised embedding methods, RIT achieves the best performances. According to previous discussions, other semi-supervised methods generally utilize a Laplacian term to regularize the unlabeled samples which shed no light on the discriminative side. RIT model directly enhances the discrimination of unsupervised points by optimizing the mutual information. Moreover, another significant advantage of RIT model is its flexibility in handling both supervised and semi-supervised embedding tasks. Other discriminative embedding models almost need extra modifications, *e.g.* adding another term into the objective, to adjust themselves to the semi-supervised version. Fortunately, RIT does not require any modifications in the model which is only determined by the training data type (supervised or semi-supervised) fed to it.

Consequently, we further consider a very challenging task that noisy labels are involved in the discriminative learning. To conduct the experiments, we randomly select a number of samples from training set and their labels are wrongly denoted. For each noisy level, the experiments are repeated for 10 times and the average classification accuracy on different datasets are reported in Fig.4. We compare the RIT model with other benchmark data embedding and representation methods with the same training samples and noisy labels.

From the results, obviously, the performances of different data embedding methods gradually drop along with the noisy labels rate increasing. Fortunately, our RIT model is the most stable one to the noisy labels. Meanwhile, TSC achieves relatively good performances on this test. This is because TSC does not only address the discriminative objective and meanwhile considers optimal signal reconstruction. GMA also attains comparable results when the noisy label rate is small. Its performance further decreases along with the increases of noisy label number. For other discriminative embedding methods, *e.g.* LDA,CCA and MIM, their performances drop significantly with the increases of noisy labels. From the aforementioned discussions, we know RIT is the most robust one when compared to other discriminative embedding methods.

B. Deep RIT for Image Categorization

In this part, we evaluate the performances of DRIT model on three datasets. The first two are the fifteen scene and COIL-100 datasets which have been introduced and discussed in the previous part. In addition, a large-scale dataset, *i.e.* ImageNet [4] will also be used here. ImageNet task requires categorizing more than 100*k* testing images into 1000 classes which is quite challenging.

We choose the famous convolutional deep neural network (CDNN) proposed in [41] as deep learning part because it has been widely regarded as the benchmark configuration in

TABLE III

THE IMAGE CATEGORIZATION ACCURACY VIA DEEP LEARNING

	Fifteen Scene	COIL 100	ImageNet
HDNN	82.6±1.0	88.2±0.8	56.8
CDNN	82.3± 0.8	88.6±0.7	57.1
DRIT	83.4± 0.7	89.7±0.5	58.3
CDNN+SVM	82.9± 1.0	89.1±0.9	57.2
HDNN+SVM	83.2±1.1	87.7±0.9	57.3
DRIT+SVM	84.1± 1.1	90.4±0.9	58.6

the field. In details, CDNN is composed of eight hidden layers: 5 convolutional layers (with pooling and ReLU [42] nonlinear transformation) and 3 fully connected layers. CDNN directly use the logistic regression as the final layer while the proposed DRIT makes use of RIT function as the objective function.

We follow a standard protocol [29], [41] to train the DNN by normalizing all the input images into the size of 224 × 224. The mean value on each RGB channel is subtracted from the original image. The DNN here involves more than 60 million hidden parameters to be learned from the data that requires a huge amount of training data. ImageNet dataset provides a sufficient amount of 1.2 million training images for deep learning. However, the fifteen scene and COIL datasets only contain limited number of training samples. Accordingly, on these two datasets, we follow the same idea in [43] to use the ImageNet training results to initialize the DNN. Then, the training images from these two datasets are used to fine-tune the parameters via back propagation. In practice, a random set of 100 and 40 images in each class of fifteen scene and COIL datasets are selected as the training samples. Both the average accuracy and standard deviations on these two datasets are reported in Table.III. ImageNet provides the training/testing list and only the accuracy on it is recorded.

We report the image categorization results of different deep learning methods in Table.III. We further consider the DNN with Hinge loss (HDNN) in the task layer. Hinge loss is also known as max margin loss which explicitly penalizes the margins of different classes. In addition to the categorization results with the logistic regression, the SVM classifier is also tested here. For SVM method, the values on the last representation hidden layer (before the categorization layer) are used to train a linear SVM classifier. From the results, it is apparent that deep learning strategy significantly outperforms the linear methods in Table I. We have also tried the bag-of-feature methods (with 1000 codewords) on the ImageNet dataset and then classify them with linear RIT. Unfortunately, the accuracy is only around 26% which is far away from the deep learning results implying only deep learning could make reasonable predictions on this challenging ImageNet task. This is reasonable because DL is built upon millions of parameters while linear subspace model is only configured in a shallow framework. When comparing different deep learning results in Table III, the advantages of DRIT model are self-evident. DRIT improves the accuracy for 1.2 points than typical CDNN on the ImageNet dataset. The improvements are also verified from the results with the SVM classifier where DRIT wins CDNN for 1.4% on ImageNet. The similar experimental findings are also made on the other two datasets.

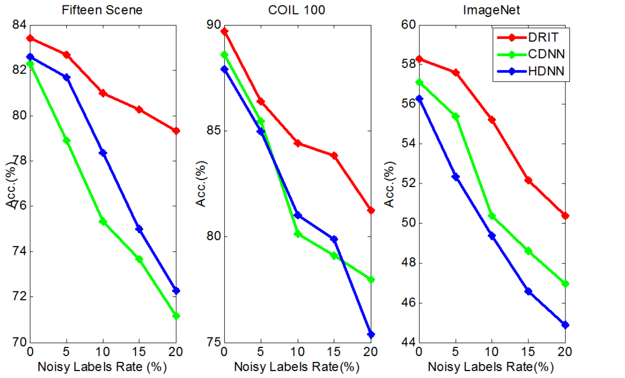


Fig. 5. The categorization accuracy of deep learning methods with different noisy label rates.

The previous experiments on deep learning verify that the RIT is a better objective than logistic and hinge losses for general deep learning tasks. The advantages of RIT can be further highlighted on its robustness in reducing label ambiguity as discussed in Section III-B. In this part, we further investigate DRIT's performances in treating label noises. We randomly select $p\%$ of training samples and denote definitely wrong labels to them. These wrongly supervised samples are mixed with ground truth training samples to conduct DNN learning. The noisy label rates are varied from 5% to 20% and, at each noisy rate level, the experiments are repeated for 10 times with the average accuracies reported in Fig.5. From the deep learning investigations in Table III, it is found that SVM classifier achieves similar performances as logistic regression. Accordingly, in this part, only the deep learning results with logistic classifier are reported. By analyzing the results, we have observed the curves of DRIT suffer less drop than CDNN and HDNN. The results suggest that DRIT could cope with the noises in the training set much better leading to a relatively reliable curve with increases of the noisy label rates. The mechanism why RIT could alleviate this kind of noisy labels has been discussed in Section III-B.

Till now, two implementations of RIT with linear transformation and deep transformation have been discussed and verified. From experimental comparison, it is concluded that deep RIT could always achieve much higher classification accuracy than the linear RIT model. However, the nature of DNN training requires sufficient training samples and heavy computational complexity. Therefore, for some challenging tasks, *e.g.* ImageNet, DRIT is strongly recommended due to its advancements in performances. On some small-scale dataset, e.g. COIL-100, the flexible linear RIT itself could already achieve very sound performances. Meanwhile, we have noticed that RIT and DRIT both exhibits the robustness in treating noisy labels in the training set. Both of them performs more reliable than other linear and deep approaches with noisy labels.

C. SS-RIT for Brain MRI Segmentation

The discussions in the last section successful verify that RIT model is robust to label uncertainty. In many practical image analysis tasks, feature uncertainty is another critical

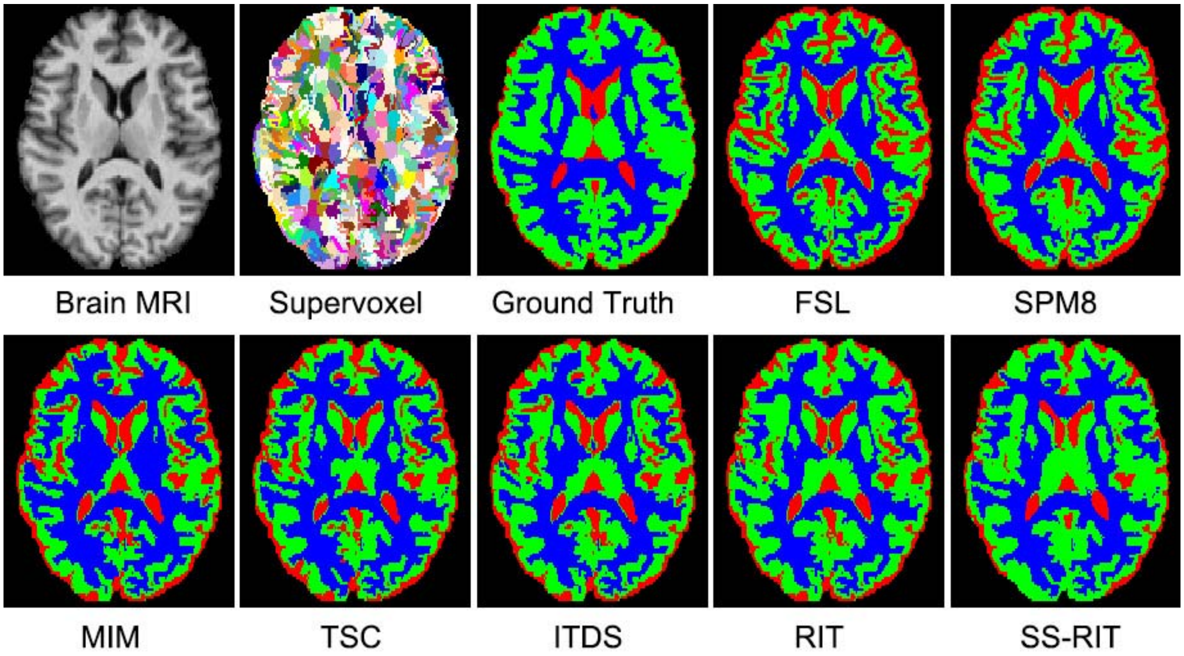


Fig. 6. Segmentation results of one brain MRI from the IBSR dataset.

issue that should be comprehensively considered. In this part, the discussions are extended on a practical task, *i.e.* MRI segmentation, where both the label and feature uncertainties simultaneously happen. In general, MRI segmentation is an important clinical task [44], that requires assigning brain tissues into white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) [44].

By analyzing the problem, the challenges mainly stem from two effects, *i.e.* partial volume and bias field effects [45]. In [46], it is revealed that partial volume effect and bias filed effect may respectively lead to the label uncertainty and feature uncertainty. The label uncertainty issue can be potentially well addressed by the RIT model whose robustness in conquering label ambiguity has been verified in the previous part. The remaining challenge here is to overcome the feature uncertainty. To cope with the bias effect, in the paper, we introduce a new strategy to conducting MRI segmentation on the super-voxel level. In details, the homogenous local regions on the brain tissue are grouped together into a super-voxel according to the method in [47]. An instance of super-voxels-level segmentation generated by the SLIC [47] algorithm has been provided in Fig.6.

Then, on each super-voxel, multiple feature descriptors can be generated from different views that provides comprehensive quantitative summarizations of the coherent structure on the tissue. In this paper, we divide the visual descriptors into four groups as intensity, texture, SIFT and HOG features. The intensity feature is extracted by computing the intensity histogram with 64 bin. Local binary pattern [48] is exploited as a texture descriptor and can be summarized in a 36-dimensional feature vector. SIFT [49] is calculated on each super-voxel and leads to a vector in 128 dimensions. Finally, 31-dimension histogram of oriented gradient [50] is extracted. In total, a 259-dimensional feature vector is generated for each super-voxel.

The feature extraction strategy on super-voxel depicts the tissue content by descriptors from multiple views, avoiding the description biases from a single view. However, a natural question consequently raised here: which type of visual descriptors and their combinations are most suitable to the brain segmentation tasks? The structured sparse norm discussed in Section IV-B well solves this issue by enabling group-level feature selection/combination in subspace. Different from typical ℓ_1-norm-based sparse feature learning, structured sparse norm pays particular attentions to the physical structure of each feature group. It encourages the sparsity only at the group level avoiding destroying the original structure in a feature group. It thus makes more reasonable high-level summarizations of the original data [18] by keeping their inherent information content.

In summary, SS-RIT is a plausible paradigm in coping with the challenging brain MRI segmentation. First, the label uncertainty from partial volume effects is solved by the RIT model. Moreover, after combining the structured sparse norm into RIT, the SS-RIT naturally exhibits the group-level feature learning mechanism when generating the projection matrix. Finally, SS-RIT simultaneously performs data embedding and classification in the joint framework, which is flexible and effective in enhancing the discriminative structure of the data points in the latent space.

The experiments were conducted on two widely used datasets from Internet Brain Segmentation Repository (IBSR) [51] and BrainWeb database [52]. The IBSR dataset consists 18 real images with a size of $256 \times 256 \times 128$ voxels. BrainWeb dataset consists of 18 images with a size of $181 \times 217 \times 181$ voxels. All of these images

TABLE IV
PERFORMANCE OF DIFFERENT SEGMENTATION METHODS ON IBSR AND BRAINWEB DATASETS

	IBSR			BrainWeb		
Methods	CSF	GM	WM	CSF	GM	WM
FSL	0.53	0.76	0.87	0.85	0.88	0.90
SPM8	0.55	0.80	0.86	0.86	0.89	0.91
MIM	0.52	0.79	0.80	0.83	0.85	0.86
TSC	0.58	0.80	0.84	0.88	0.90	0.90
ITDS	0.60	0.81	0.86	0.90	0.90	0.92
RIT	0.63	0.83	0.87	0.92	0.92	0.93
SS-RIT	0.68	0.86	0.88	0.93	0.93	0.95

are provided with ground truth segmentations for quantitative evaluations.

We exploit the structured-sparse RIT here to make predictions on the super-voxel level. To note, RIT is implemented on the semi-supervised version and, on each image, a random set of 150 super-voxels on each MRI are labeled with their ground truth label. In medical image segmentations, rather than the prediction accuracy, the dice similarity coefficients (DSC) is instead used as a criteria to assay a method's performance [53].[2]

For comparison purpose, the SS-RIT is pit against other leading methods for MRI segmentation including brain tissue segmentation algorithms in FMRIB Software Library (FSL) [54] and Statistical Parametric Mapping 8 (SPM8) package [55]. These two tools perform voxel-vise segmentation and have been widely regarded as benchmark methods in the neuroimaging community. Further, on the super-voxel level, the SS-RIT model is compared with other subspace models including MIM and TSC. MIM is also an information theoretic embedding method and TSC achieves much robust results according to previous tests. Finally, the structured sparse model is compared with RIT and Information Theoretic Discriminative Segmentation (ITDS) [46]. The major differences between ITDS and SS-RIT is the former exploits the ℓ_1 sparse norm for feature selection in the original space while the later conducts group-level feature learning in subspaces.

The segmentation results are visualized in Fig.6. Each panel illustrates the brain tissue segmentation result of the same brain volume selected from the IBSR dataset. The color of red, green and blue voxels represent the tissues of CSF, GM and WM, respectively. By comparing the segmentations to the ground truth, the advantage of SS-RIT segmentation over other six approaches is apparent. In particular, the SS-RIT segmentation shows particularly better delineations of CSF and GM tissues.

The quantitative evaluations on two datasets are reported in Table IV. A higher value of DSC represents a better correspondence to the ground truth. From the results, it is noted that the super-voxel level segmentations (last five rows in Table IV.) are much better than the voxel level segmentations (FSL and SPM8). Among all the subspace segmentation methods, the performance of RIT is much robust than MIM and TSC. This is because RIT conquers the uncertainties in the supervised labels and thus better captures the coherent discriminative structures in MRI data.

Further, we will discuss the advantages of exploiting structured-sparse norm which can be verified by comparing SS-RIT with RIT (no feature selection) and ITDS (sparse selection). Within the same experimental setting, learning group-level feature transformations (structured subspace learning) generally outperforms other two methods on DSC values. In addition, SS-RIT also achieves the lowest standard deviation for three types of tissues than others. This serves as an evidence to demonstrate that structured sparse learning does not only improve the accuracy but also enhances the robustness.

VI. CONCLUSION

This work introduces an information theoretic method that successfully alleviates both label and feature uncertainties in general data representation tasks. The main advantages of the proposed method are represented from its flexibility and robustness. In the view of flexibility, RIT model works friendly with different types of feature transformation functions to conduct information theoretic learning. In this paper, we have implemented linear RIT, deep RIT and structured-sparse RIT models to address different image analysis tasks. The RIT framework generally improves other similar methods in the field. In the view of robustness, RIT is proven to be much effective to reduce the ambiguities in the training samples. Both the linear and deep RIT achieve much reliable performances in spite of label noises in the training samples. Moreover, its structured extension well addresses the partial volume effect (label uncertainty) and bias field effect (feature uncertainty) in MRI of brain tissue and thus achieves much better segmentation results than other state-of-the-arts.

REFERENCES

[1] D. Böhning, "Multinomial logistic regression algorithm," *Ann. Inst. Statist. Math.*, vol. 44, no. 1, pp. 197–200, Mar. 1992.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2006.

[3] L. G. S. Giraldo, and J. C. Principe. (2013). "Rate-distortion auto-encoders." [Online]. Avaliable: http://arxiv.org/abs/1312.7381

[4] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2014.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[6] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE 11th Int. Conf. Comput. Vis. (ICCV)*, Oct. 2007, pp. 1–7.

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[8] Y. Deng, Y. Li, Y. Qian, X. Ji, and Q. Dai, "Visual words assignment via information-theoretic manifold embedding," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1924–1937, Oct. 2014.

[9] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[10] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao, "Maximal linear embedding for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1776–1792, Sep. 2011.

[11] Y. Deng, Q. Dai, R. Wang, and Z. Zhang, "Commute time guided transformation for feature extraction," *Comput. Vis. Image Understand.*, vol. 116, no. 4, pp. 473–483, Apr. 2012.

[2] DSC is calculated from true positive (TP), false positive (FP) and false negative (FN) rates as, $DSC = \frac{2\times TP}{2\times TP+FP+FN}$.

[12] T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *Proc. 10th Eur. Conf. Comput. Vis.*, vol. 5302. Marseille, France, 2008, pp. 725–738.

[13] A. Sharma, A. Kumar, H. Daume, III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2160–2167.

[14] Y. Deng, Y. Zhao, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "Discriminant kernel assignment for image coding," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2547941.

[15] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19. 2007, p. 801.

[16] Y. Deng, Y. Kong, F. Bao, and Q. Dai, "Sparse coding-inspired optimal trading system for HFT industry," *IEEE Trans. Ind. Informat.*, vol. 11, no. 2, pp. 467–475, Apr. 2015.

[17] Y. Deng, Q. Dai, and Z. Zhang, "Graph Laplace for occluded face completion and recognition," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2329–2338, Aug. 2011.

[18] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proc. Artif. Int. and Statist.*, 2010, pp. 366–373.

[19] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[20] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, "Deep direct reinforcement learning for financial signal representation and trading," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2522401.

[21] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Trans. Fuzzy Syst.*, to be published, doi: 10.1109/TFUZZ.2016.2574915.

[22] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.

[23] R. G. Gomes, A. Krause, and P. Perona, "Discriminative clustering by regularized information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23. 2010, pp. 775–783.

[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[25] K. Torkkola, "Feature extraction by non parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, Jan. 2003.

[26] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17. 2005, pp. 1–8.

[27] M. A. Newton and A. E. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 56, no. 1, pp. 3–48, 1994.

[28] B. L. Cun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[29] J. Donahue *et al.*. (2013). "DeCAF: A deep convolutional activation feature for generic visual recognition." [Online]. Available: https://arxiv.org/abs/1310.1531

[30] Y. Deng, Y. Liu, Q. Dai, Z. Zhang, and Y. Wang, "Noisy depth maps fusion for multiview stereo via matrix completion," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 566–582, Sep. 2012.

[31] Y. Deng, Q. Dai, R. Liu, Z. Zhang, and S. Hu, "Low-rank structure learning via nonconvex heuristic recovery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 3, pp. 383–396, Mar. 2013.

[32] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, Jan. 2011.

[33] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[34] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2005, pp. 524–531.

[35] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

[36] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 886–893.

[38] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[39] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May 2001.

[40] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Autom. Control*, vol. 26, no. 1, pp. 17–32, Feb. 1981.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[42] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 8609–8613.

[43] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[44] L. He and N. A. Parikh, "Automated detection of white matter signal abnormality using T2 relaxometry: Application to brain segmentation on term MRI in very preterm infants," *NeuroImage*, vol. 64, pp. 328–340, Jan. 2013.

[45] H. Greenspan, A. Ruf, and J. Goldberger, "Constrained Gaussian mixture model framework for automatic segmentation of MR brain images," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1233–1245, Sep. 2006.

[46] Y. Kong, Y. Deng, and Q. Dai, "Discriminative clustering and feature selection for brain MRI segmentation," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 573–577, May 2015.

[47] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[48] S. Liao, M. W. K. Law, and A. C. S. Chung, "Dominant local binary patterns for texture classification," *IEEE Trans. Image Process.*, vol. 18, no. 5, pp. 1107–1118, May 2009.

[49] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[50] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[51] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 153–163, Feb. 2012.

[52] R. K.-S. Kwan, A. C. Evans, and G. B. Pike, "MRI simulation-based evaluation of image-processing and classification methods," *IEEE Trans. Med. Imag.*, vol. 18, no. 11, pp. 1085–1097, Nov. 1999.

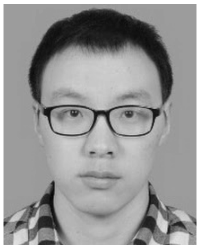
[53] B. Dogdas, D. W. Shattuck, and R. M. Leahy, "Segmentation of skull and scalp in 3-D human MRI using mathematical morphology," *Human Brain Mapping*, vol. 26, no. 4, pp. 273–285, Dec. 2005.

[54] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

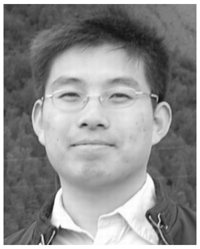
[55] J. Ashburner and K. J. Friston, "Unified Segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.

Yue Deng received the B.E. degree (Hons.) in automatic control from Southeast University, Nanjing, China, in 2008, and the Ph.D. degree (Hons.) in control science and engineering from the Department of Automation, Tsinghua University, Beijing, China, in 2013. He is currently a Post-Doctoral Fellow with the School of Pharmacy, University of California, San Francisco Medical Center, CA, USA. His current research interests include machine learning, signal processing, and computational biology.

Feng Bao received the B.E. degree in electronics and information engineering from Xidian University, Xi'an, China, in 2014. He is currently pursuing the M.S. degree with the department of automation, Tsinghua University, Beijing, China. His current research interests include machine learning and computational biology.



Xuesong Deng received the B.E. degree in electronic information engineering from the Beijing University of Posts and Telecommunications, and the M.S. degree in computer science and technology from the Institute of Computing Technology, Chinese Academy of Sciences, in 2013 and 2016, respectively. He was a Data Mining Engineer with Baidu from 2016.



Ruiping Wang (S'08–M'11) received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2010.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, from 2010 to 2012. He was a Research Associate with the Computer Vision Laboratory, Institute for Advanced Computer Studies, the University of Maryland, College Park, from 2010 to 2011. He has been a Faculty Member with the Institute of Computing Technology, Chinese Academy of Sciences, since 2012, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning.

Youyong Kong received the B.S. (Hons.) and M.S. degrees in computer science and engineering from Southeast University, China, in 2008 and 2011, respectively, and the Ph.D. degree in imaging and diagnostic radiology from the Chinese University of Hong Kong, Hong Kong, in 2014. He is currently an Assistant Professor with the College of Computer Science and Engineering, Southeast University. His research interests include machine learning, medical image processing, and analysis.

Qionghai Dai (SM'05) received the B.S. degree in mathematics from Shanxi Normal University, Xian, China, in 1987, and the M.E. and Ph.D. degrees in computer science and automation from Northeastern University, Shenyang, China, in 1994 and 1996, respectively. He has been the Faculty Member of Tsinghua University, since 1997. He is currently a Cheung Kong Professor with Tsinghua University and the Director of the Broadband Networks and Digital Media Laboratory. His current research interests include signal processing and computer vision.