Semi-Supervised Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport

Yang Yang, Zhao-Yang Fu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang

Abstract—Complex objects are usually with multiple labels, and can be represented by multiple modal representations, e.g., the complex articles contain text and image information as well as multiple annotations. Previous methods assume that the homogeneous multi-modal data are consistent, while in real applications, the raw data are disordered, e.g., the article constitutes with variable number of inconsistent text and image instances. Therefore, Multi-modal Multi-instance Multi-label (M3) learning provides a framework for handling such task and has exhibited excellent performance. However, M3 learning is facing two main challenges: 1) how to effectively utilize label correlation and 2) how to take advantage of multi-modal learning to process unlabeled instances. To solve these problems, we first propose a novel Multi-modal Multi-instance Multi-label Deep Network (M3DN), which considers M3 learning in an end-to-end multi-modal deep network and utilizes consistency principle among different modal bag-level predictions. Based on the M3DN, we learn the latent ground label metric with the optimal transport. Moreover, we introduce the extrinsic unlabeled multi-modal multi-instance data, and propose the M3DNS, which considers the instance-level auto-encoder for single modality and modified bag-level optimal transport to strengthen the consistency among modalities. Thereby M3DNS can better predict label and exploit label correlation simultaneously. Experiments on benchmark datasets and real world WKG Game-Hub dataset validate the effectiveness of the proposed methods.

Index Terms—Semi-supervised learning, multi-modal multi-instance multi-label learning, modal consistency, optimal transport

1 Introduction

With the development of data collection techniques, objects can always be represented by multiple modal features, e.g., in the forum of famous mobile game " Strike of Kings", the articles are with image and content information, and they belong to multiple categories if they are observed from different aspects, e.g., an article belongs to "Wukong Sun" (Game Heroes) as well as "golden cudgel" (Game Equipment) from the images, while it can be categorized as "game strategy", "producer name" from contents and so on. The major challenge for addressing such problem is how to jointly model multiple types of heterogeneities in a mutually beneficial way. To solve this problem, multi-modal multi-label learning approaches utilize multiple modal information, and require modal-based classifiers to generate similar predictions, e.g., Huang et al. proposed a multi-label conditional restricted boltzmann machine, which uses multiple modalities to obtain shared representations under the supervision [1]; Yang et al. learned a novel graph-based model to learn both label and feature heterogeneities [2]. However, a real-world object may contain variable number of inconsistent multi-modal instances, e.g., the article usually contains multiple images and content paragraphs, in which each image or content paragraph can be regarded as an instance, yet the relationships between the images and contents have not been marked as shown in Fig. 1.

Therefore, several Multi-modal Multi-instance Multi-label methods have been proposed. Nguyen et al. proposed M3LDA with a visual-label part, a textual-label part, and a label topic part, in which the topic decided by visual information and the topic decided by textual information should be consistent [3]; Nguyen et al. developed a multi-modal MIML framework based on hierarchical Bayesian network [4]. Nevertheless, there are two drawbacks of the existing M3 models. In detail, previous approaches rarely consider the correlations among labels, besides, M3 methods are all supervised methods, which violate the intuition of multi-modal learning using unsupervised data.

Thus, considering the label correlation, Yang and He studied a hierarchical multi-latent space, which can leverage the task relatedness, modal consistency and the label correlation simultaneously to improve the learning performance [5]; Huang and Zhou proposed the ML-LOC approach which allows label correlation to be exploited locally [6]; Frogner et al. developed a loss function with ground metric for multi-label learning, which is based on the wasserstein distance [7]. Previous works mainly assumed that there exists some prior knowledge such as label similarity matrix or the ground metric [7], [8]. In reality, semantic information among labels is indirect or complicated, thus the confidence of the label similarity matrix or ground metric is weak. On the other hand, considering the labeling cost, there are many unlabeled instances. The most important advantage of multi-modal methods is that they use unlabeled data, e.g., co-training [9] style methods utilized the complementary principle to label

- *Y. Yang, Z.-Y. Fu, D.-C. Zhan, and Y. Jiang are with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. E-mail: {yangy, fuzy, zhandc, jiangy}@lamda.nju.edu.cn.*
- *Z.-B. Liu is with Tencent WXG, Shenzhen, Guangdong 518057, China. E-mail: lewiszbliu@tencent.com.*

Manuscript received 6 Oct. 2018; revised 20 July 2019; accepted 29 July 2019. Date of publication 2 Aug. 2019; date of current version 11 Jan. 2021. (Corresponding author: Yang Yang.)
Recommended for acceptance by L. B. Holder.
Digital Object Identifier no. 10.1109/TKDE.2019.2932666

Fig. 1. An illustration of the M3 (Multi-Modal Multi-instance Multi-label) Data in an article of WKG Game-Hub. Each article is with context bag and image bag, each bag contains variable number of instances (context paragraphs/images), while each article has multiple label representations. It is notable that different modalities are heterogeneous, i.e., there have no congruent relationships between the articles and images.

unlabeled data for each other; co-regularize [10] style methods exploited unlabeled multi-modal data with consistency principle. Meanwhile, it is notable that previous proposed M3 based methods are hard to adopt the unlabeled instances. Therefore, another issue is how to bypass the limitation of M3 style methods by using unlabeled multi-modal instances.

In this work, aiming at learning the label prediction and exploring label correlation with semi-supervised M3 data simultaneously, we proposed a novel general Multi-modal Multi-instance Multi-label Deep Network, which models the independent deep network for each modality, and imposes the modal consistency on bag-level prediction. To better consider the label correlation, M3DN first adopts Optimal Transport (OT) [11] distance to measure the quality of prediction. The adoption provides a more meaningful measure in multi-label tasks by capturing the geometric information of the underlying label space. The raw data may not calculate the raw ground metric confidently, thus we cast the label correlation exploration as a latent ground metric learning problem. Moreover, considering the unlabeled data information, we propose the semi-supervised M3DN (M3DNS). M3DNS utilizes the instance-level auto-encoder to build the single modal network, and considers the bag-level consistency among different unlabeled modal predictions with the modified OT theory. Consequently, M3DNS could automatically learn the predictors from different modalities and the latent shared ground metric.

The main contributions of this paper are summarized in the following points:

- We propose a novel Multi-modal Multi-instance Multi-label Deep Network (M3DN), which models the deep independent network for each modality, and imposes the modal consistency on bag-level prediction;
- We consider label correlation exploration as a latent ground metric learning problem between different modalities, rather than a fix ground metric using prior raw knowledge;
- We utilize the extrinsic unlabeled data, by considering instance-level auto-encoder, and the bag-level consistency among different unlabeled modal predictions with the modified OT metric;
- We achieve superior performances on real-world applications, comprehensively evaluate on the performance and obtain consistently superior performances stably.

Section 2 summarizes related work, our approaches are presented in Section 3. Section 4 reports our experiments. Finally, Section 5 gives the conclusion.

2 Related Work

The exploitation of multi-modal multi-instance multi-label learning has attracted much attention recently. In this paper, our method concentrates on deep multi-label classification for semi-supervised inconsistent multi-modal multi-instance data, and considers the label correlation using optimal transport technique. Therefore, our work is related to M3 learning and the optimal transport.

Multi-modal learning deals with data from multiple modalities, i.e., multiple feature sets. The goals are to improve performance and reduce the sample complexity. Meanwhile, multi-modal multi-label learning has been well studied, e.g., Fang and Zhang proposed a multi-modal multi-label learning method based on the large margin framework [12]. Yang et al. modeled both the modal consistency and the label correlation in a graph-based framework [13]. The basic assumption behind these methods is that multi-modal data is consistent. However, in real applications, the multi-modal data are always heterogeneous on the instance-level, e.g., articles have variable number of inconsistent images and text paragraphs, videos have variable length of inconsistent audio and image frames. Articles and videos only have consistency on the bag level, rather than instance level. Thus, multi-modal multi-instance multi-label learning is proposed recently. Nguyen et al. developed a multi-modal MIML framework based on hierarchical Bayesian network [4]; Feng and Zhou exploited deep neural network to generate instance representation for MIML and it can be extended to multi-modal scenario. Nevertheless, previous approaches rarely consider the confidence of label correlation. More importantly, the current M3 approaches are supervised, which obviously lose the advantage of multi-modal learning for processing unlabeled data.

Considering the label correlation, several multi-label learning methods are proposed [15], [16], [17]. Recently, Optimal Transport (OT) [11] is developed to measure the difference between two distributions based on given ground metric, and it has been widely used in computer vision and image processing fields, e.g., Qian et al. proposed a novel method that exploits knowledge in both data manifold and feature correlation [18]; Courty et al. proposed a regularized unsupervised optimal transportation model to perform the alignment of the representations [19]. However, previous works mainly assumed that prior knowledge for cost matrix already exists, and ignored deficiency of information or domain knowledge. Thus, Cuturi and Avis, Zhao and Zhou suggested to formulate the cost metric learning problem with the side information [20], [21]. On the other hand, existing M3 methods are almost supervised methods, while multi-modal methods aim to utilize the complementary [9] or consistency [10] principle using the unlabeled instance. Thereby how to take unlabeled data into consideration becomes a challenge.

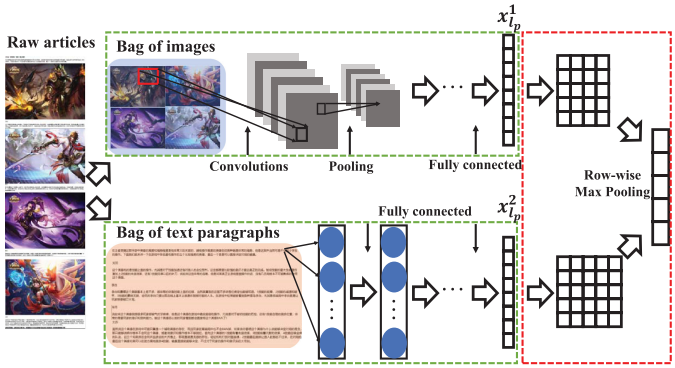


Fig. 2. The flowchart of the M3DN, the raw articles can be divided into two homogeneous modal bag with variable number of heterogeneous instances, i.e., the image bag with four images and content bag with five text paragraphs. The instances of different modalities can be calculated with different deep networks, and finally represented as $\mathbf{x}_{l_p}^1$ or $\mathbf{x}_{l_p}^2$, the output features are fully connected with the labels, and we can get the bag-concept layer for different modalities. Eventually, we can acquire the final prediction by mean-max pooling the bag-concept layer of different modalities.

3 Proposed Method

3.1 Notation

In the multi-instance extension of the multi-modal multi-label framework, we are given N bags of instances, let $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_l}\}$ denotes the label set, $\mathbf{y}_i \in \mathbb{R}^L$ is the label vector of $i-$th bag, where $y_{i,j} = 1$ denotes positive class, and $y_{i,j} = 0$ otherwise. On the other hand, suppose we are given K modalities, without any loss of generality, we consider two modalities in our paper, i.e., images and contents. Let $\mathcal{D} = \{([\mathbf{X}_1^1, \mathbf{X}_1^2], \mathbf{y}_1), ([\mathbf{X}_2^1, \mathbf{X}_2^2], \mathbf{y}_2), \ldots, ([\mathbf{X}_{N_l}^1, \mathbf{X}_{N_l}^2], \mathbf{y}_{N_l}), ([\mathbf{X}_{N_l+1}^1, \mathbf{X}_{N_l+1}^2]), \ldots, ([\mathbf{X}_{N_l+N_u}^1, \mathbf{X}_{N_l+N_u}^2])\}$ represents the training dataset, where N_l/N_u denotes the number of labelled/unlabelled instances. $\mathbf{X}_i^1 = \{\mathbf{x}_{i,1}^1, \mathbf{x}_{i,2}^1, \ldots, \mathbf{x}_{i,m_i}^1\}$ denotes the bag representation of m_i instances of $\mathbf{X}_i^1$, similarly, $\mathbf{X}_i^2 = \{\mathbf{x}_{i,1}^2, \mathbf{x}_{i,2}^2, \ldots, \mathbf{x}_{i,n_i}^2\}$ is the bag representation of n_i instances of $\mathbf{X}_i^2$, it is notable that bags of different modalities may contain variable number of instances.

The goal is to generate a learner to annotate new bags based on its inputs $\mathbf{X}^1, \mathbf{X}^2$, e.g., annotate a new complex article with its images and contents.

3.2 Optimal Transport

Traditionally, several measurements such as Kullback-Leibler divergences, Hellinger and total variation, have been utilized to measure the similarity between two distributions. However, these measurements play little effect when the probability space has geometrical structures. On the other hand, Optimal transport [11], also known as Wasserstein distance or earth mover distance [22], defines a reasonable distance between two probability distribution over the metric space. Intuitively, the Wasserstein distance is the minimum cost of transporting the pile of one distribution into the pile of another distribution, which formulates the problem of learning the ground metric as minimizing the difference between two polyhedral convex functions over a convex set of distance matrices. Therefore, the Wasserstein distance is more powerful in such situations by considering the pairwise cost.

Definition 1 (Transport Polytope). *For two probability vectors r and c in the simplex $\sum_L$, $U(r, c)$ is the transport polytope of r and c, namely the polyhedral set of $L \times L$ matrices,*

$$U(r, c) = \{P \in \mathbb{R}_+^{L \times L} | P\mathbf{1}_L = r, P^\top \mathbf{1}_L = c\}.$$

Definition 2 (Optimal Transport). *Given a $L \times L$ cost matrix M, the total cost of mapping from r to c using a transport matrix (or coupling probability) P can be quantified as $\langle P, M \rangle$. The optimal transport (OT) problem is defined as,*

$$d_M(r, c) = \min_{P \in U(r,c)} \langle P, M \rangle.$$

When M belongs to the cone of metric matrices $\mathbb{M}$, the value of $d_M(r, c)$ is a distance [11] between r and c, parameterized by M. In that case, assuming implicitly that M is fixed and only r and c vary, we will refer to the optimal transport distance between r and c. It is notable that $d_M(r, c)$ is the cost of the optimal plan for transporting the predicted mass distribution r to match the target distribution c. The penalty increases when more mass is transported over longer distances, according to the ground metric M.

Theorem 1. *d_M defined in Definition 2 is a distance on $\sum_L$ whenever M is a metric matrix [11].*

3.3 Multi-Modal Multi-Instance Multi-Label Deep Network (M3DN)

Multi-modal Multi-instance Multi-label (M3) learning provides a framework for handling the complex objects, and we propose a novel M3 based parallel deep network (M3DN). Based on the M3DN, we can bypass the limitation of initial label correlation metric using the Optimal Transport (OT) theory, and further take advantage of unlabeled data considering the modal consistency. In this section, we propose the Multi-Modal Multi-instance Multi-label Deep Network (M3DN) framework. M3DN models deep networks for different modalities and imposes the modal consistency.

The raw articles contain variable number of heterogeneous multi-modal information, i.e., when no corresponding relationships exist among each the contents and images, it is difficult to utilize the consistency principle with previous multi-modal methods. Thus, we turn to utilize the consistency among the bags of different modalities, rather than the instance-level. Specifically, raw articles can be divided into two modal bags of heterogeneous instances, i.e., the image bag with 4 images and content bag with 5 text paragraphs as shown in Fig. 2, while only the homogeneous bags share the same multiple labels. Each instance $\mathbf{x}^1(\mathbf{x}^2)$ in different modal bag can be calculated among several layers and can be finally represented as $\mathbf{x}_{l_{p1}}(\mathbf{x}_{l_{p2}})$.

Without any loss of generality, we use the convolutional neural network for images and the fully connected networks for text. Then, the output features are fully connected with the bag-concept layer. All parameters including deep network facts and fully connected weights can be organized as $\Theta_1 = \{\theta_{l_1}, \theta_{l_2}, \ldots, \theta_{l_{p1-1}}, W_1\}(\Theta_2 = \{\theta_{l_1}, \theta_{l_2}, \ldots, \theta_{l_{p2-1}}, W_2\})$. Concretely, once the label predictions of the instances for a bag $\mathbf{X}_i^v$ are obtained, we propose a fully connected 2D layer (bag-concept layer) with the size of $m_i(n_i) \times L$ as shown in Fig. 3, in which each column represents corresponding

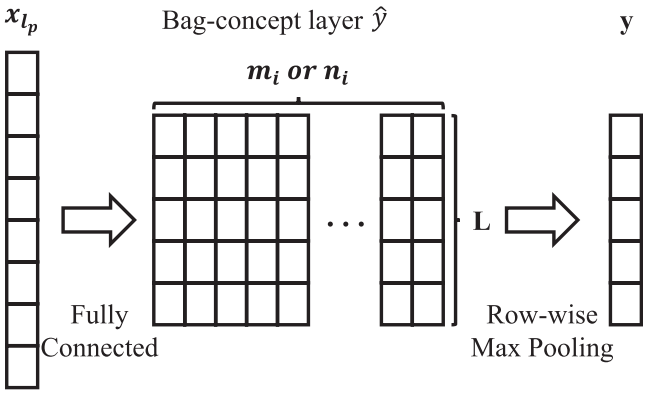


Fig. 3. The schematic of the bag-concept layer. We can acquire the bag-concept layer with the output feature representations of a bag of instances, in which each column represents corresponding prediction of each instance. Eventually, the final label prediction is calculated by row-wise max pooling.

prediction of each instance in the image/content bag. Formally, for a given bag of instances $\mathbf{X}_i^v$, the (k,j)th node in the 2D bag-concept layer represents the prediction score between the instance $\mathbf{x}_{i,j}^v$ and the $k-$th label. Therefore, the j-column has the following form of activation:

$$\hat{\mathbf{y}}_j^v = g(W_v \mathbf{x}_{i,j}^v + b_v). \tag{1}$$

Here, $g(\cdot)$ can be any convex activation function, and we use softmax function here. In the bag-concept layer, we utilize the row-wise max pooling: $f_v(i) = max(\hat{\mathbf{y}}_{i,\cdot})$. The final prediction value is: $f = \frac{f_1+f_2}{2}$.

3.4 Explore Label Correlation

However, fully connection to the label output rarely considers the relationship among labels. Recently, Optimal Transport (OT) theory [11] is used in multi-label learning, which captures geometric information of the underlying label space. According to the Definitions 2 and 1, the loss function implied in the parallel network structure can be formulated without any loss of generality as:

$$\begin{aligned} &\min_{P_v \in U(f(X_i^v),\mathbf{y}_i)} \sum_{v=1}^{2}\sum_{i=1}^{N} \langle P_v, M\rangle \\ s.t.\quad & U(f(\mathbf{X}_i^v),\mathbf{y}_i) = \{P_v \in \mathbb{R}_+^{L\times L} | P_v \mathbf{1}_L = f(X_i^v), P_v^\top \mathbf{1}_L = \mathbf{y}_i\}, \end{aligned} \tag{2}$$

where M is the shared latent cost matrix. However, this method requires prior knowledge to construct the cost matrix M. However, in reality, indirect or incomplete information among labels leads to weak cost matrix M and poor classification performance.

Therefore, we can define the process of learning cost metric as an optimization problem. Optimizing the cost metric directly is difficult and it consumes $O(L^2)$ constraints. Thus, [20], [21] proposed to formulate the cost metric learning problem with the side information, i.e., the label similarity matrix S as [21], and [20] has proved that the cost metric matrix M, which computes corresponding optimal transport distance d_M between pairs of labels, agrees with the side information. More precisely, this criterion favors matrix M, in which the distance $d_M(r;c)$ is small for pairs of similar histograms r and c (corresponding $S(r;c)$ is large) and large for pairs of dissimilar histograms (corresponding $S(r;c)$ is small). Consequently, optimizing M can be turned to optimize the S. Finally, the goal of M3DN can be turned to learn label predictor and explore label correlation simultaneously.

In detail, we first introduce the connection between nonlinear transformation and pseudo-metric:

Definition 3. *With the nonlinear transformation $\emptyset(\cdot)$, the euclidean distance after the transformation can be denoted as:*

$$D_\emptyset(r,c) = \|\emptyset(r) - \emptyset(c)\|_2.$$

And [23] proved that $D_\emptyset$ satisfies all properties of a well-defined pseudo-metric in the original input space.

Theorem 2. *For a pseudo-metric M defined in Definition 3 and histograms $r, c \in \sum_L$, the function $(r,c) \rightarrow \mathbf{1}_{r\neq c} d_M(r,c)$ satisfies all four distance axioms, i.e., non-negativity, symmetry, definiteness and sub-additivity (triangle inequality) as in [20].*

Thus, M can be turned to learn the kernel S defined by the non-linear transformation $\emptyset(\cdot)$:

$$S_{ij} = S(\mathbf{y}_i, \mathbf{y}_j) = \emptyset(\mathbf{y}_i)^\top \emptyset(\mathbf{y}_j), \tag{3}$$

where the $\mathbf{y}_i$ represents the label vector of $i-$th instance. Besides, it is notable that the cost matrix M is computed as $M_{ij} = D_\emptyset^2(\mathbf{y}_i, \mathbf{y}_j)$, while the kernel S is defined as Eq. (3). Thus, the relation between M and S can be derived as:

$$M_{ij} = S_{ii} + S_{jj} - 2S_{ij}. \tag{4}$$

The non-linear mapping preserves pseudo metric properties in Definition 3, therefore it only needs a projection to positive semi-definite matrix cone when learning the kernel matrix S. Thus, we can avoid the projection to metric space which is complicated and costly. Therefore, we propose to conduct the label predictions and label correlation exploration simultaneously based on substituted optimal transport, the combination of Eqs. (4) and (2) can be reformulated as:

$$\begin{aligned} &\min_{S, P_v \in U(f(X_i^v),\mathbf{y}_i)} \sum_{v=1}^{2}\sum_{i=1}^{N} \langle P_v, M\rangle + \lambda_1 r(S, S_0) \\ s.t.\quad & U(f(\mathbf{X}_i^v),\mathbf{y}_i) = \{P_v \in \mathbb{R}_+^{L\times L} | P_v \mathbf{1}_L = f(X_i^v), P_v^\top \mathbf{1}_L = \mathbf{y}_i\} \\ & S \in \mathcal{S}_+, \quad M_{ij} = S_{ii} + S_{jj} - 2S_{ij}, \end{aligned} \tag{5}$$

where λ_1 is a trade-off parameter, $\mathcal{S}_+$ denotes the set of positive semi-definite matrix. We adopt OT distance as the loss between prediction and groundtruth, and then incorporate the ground metric learning by kernel biased regularization in 2nd term, where $\lambda_1 r(S, S_0)$ can be any convex regularization. The regularizer $\mathcal{S}_+ \times \mathcal{S}_+ \rightarrow \mathcal{R}_+$ allows us to exploit prior knowledge on the kernelized similar matrix, encoded by a reference matrix S_0. Since typically no strong prior knowledge is available, we use $S_0 = \mathcal{Y}' \times \mathcal{Y}$. Following common practice [24], we utilize the asymmetric Burg divergence, which yields:

$$r(S, S_0) = \mathrm{tr}(SS_0^{-1}) - logdet(SS_0^{-1}) - p.$$

where p is the balance parameter, and we set as 1 in our experiments.

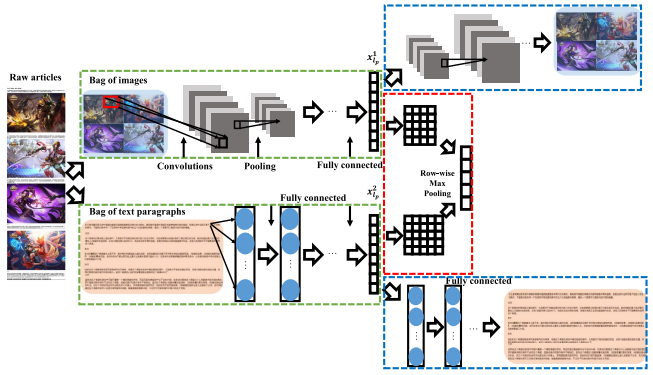


Fig. 4. The flowchart of the M3DNS consider unlabeled data. Similar to M3DN, the raw articles can be divided into two homogeneous modal bags with variable number of heterogeneous instances. The instances of different modalities can be calculated with different deep networks, and finally represented as $\mathbf{x}_{l_p}^1$ or $\mathbf{x}_{l_p}^2$. The output features of labeled data are fully connected with the labels, while we add decoder networks for each modality to process the unlabeled data. On the other hand, we can get bag representations of all data from the bag-concept layer for different modalities. Eventually, we can acquire the final predictions of different modalities and calculate the semi-supervised loss.

3.5 Consider Unsupervised Data

M3DN provides a framework for handling complex multi-modal multi-instance multi-label objects, and it considers the label correlation as an optimization problem in Eq. (8). The limitation of manual labeling is that, in real application, it leaves over large number of unlabeled data. In other words, unlabeled data is readily available, while labeled data tends to be of smaller size. The basic intuition of multi-modal learning is to utilize the complement or consistent information of unlabeled data, to get better performance. Yet M3DN leaves the unlabeled data without consideration, and this obviously loses the advantage of multi-modal learning. Consequently, how to extend M3DN to semi-supervised scenario is an urgent problem.

To consider the extrinsic consistency, i.e., the unlabeled information of different modalities, we propose a semi-supervised M3DN (M3DNS) methods for learning each modal predictors. Different from previous co-regularize style methods using instance-level consistency principle, M3 learning only has bag-level consistency among different modalities, rather than instance-level consistency. Thus, there exist two challenges in using unlabeled data in M3 learning: 1) how to utilize different modal instance-level unlabeled data; 2) how to utilize different modal bag-level consistency of unlabeled data.

To solve this problem, M3DNS utilizes the instance-level unlabeled instances with auto-encoder and bag-level unlabeled instances with modified OT. As shown in Fig. 4, since different modal bags include various number of instances, and the correspondences among different modal instances are unknown, we turn to utilize the auto-encoder based networks to reconstruct the input instances for different modalities, which can build more robust encoder networks. On the one hand, bag-level correspondences are known, thereby for the bag-level unlabeled data, we utilize modified OT consistency term to constraint different modalities.

Specifically, each modal ordinal network can be replaced by auto-encoder (AE) network, which minimizes the reconstruction error of all the instances, i.e., auto-encoder CNN for image modality and auto-encoder fully connected network for content modality. Without any loss of generality, AE can be formulated as square loss:

$$AE(\mathbf{x}_k) = \min_{\Theta_{f_v},\Theta_{r_v}} \sum_{i=N_l+1}^{N_u} \|\mathbf{x}_i{}^v - r_v(f_v(x_i{}^v))\|_F^2, \tag{6}$$

where $\Theta_{f_v}, \Theta_{r_v}$ are the weight parameters of encoder network f_v and decoder network r_v of the $v-$th modality.

On the other hand, Eq. (2) only utilizes the supervised information, while neglect the unlabeled modal bag-level correspondences. Thus, with the unlabeled information, Eq. (2) can be reformulated as:

$$\begin{aligned} \min_{P_v\in U,\hat{P}\in\hat{U}} & \sum_{v=1}^{2}\sum_{i=1}^{N_l}\langle P_v, M\rangle + \sum_{i=1}^{N_u}\langle \hat{P}, M\rangle \\ s.t.\quad & U = \{P_v \in \mathbb{R}_+^{L\times L} | P_v\mathbf{1}_L = f(X_i^v), P_v^\top \mathbf{1}_L = \mathbf{y}_i\} \\ & \hat{U} = \{\hat{P} \in \mathbb{R}_+^{L\times L} | \hat{P}\mathbf{1}_L = f(X_i^1), \hat{P}^\top \mathbf{1}_L = f(X_i^2)\}, \end{aligned} \tag{7}$$

where $\hat{P}$ is the pseudo transport matrix (or coupling probability) for unlabeled data. The extra unlabeled modal predictions can be regarded as the pseudo labels in $\hat{P}$ for constructing more discriminative predictors. In detail, when learning one modal predictor, the predictions of other modalities can act as the pseudo label, which can assist learning more discriminative predictors with unlabeled data. Thus M3DNS can well utilize the bag-level consistency among different modalities. Therefore, M3DNS can acquire more robust ground metric M, which potentially utilizes the consistency between different modal bags.

As a result, with the unlabeled information, we can combine the Eq. (7) and (6). The semi-supervised M3DN method (M3DNS) can be given as:

$$\begin{aligned} \min_{P_v\in U,\hat{P}\in\hat{U}} & \sum_{v=1}^{2}\sum_{i=1}^{N_l}\langle P_v, M\rangle + \sum_{i=N_l+1}^{N_u} AE(x_i^v) + \sum_{i=1}^{N_u}\langle \hat{P}, M\rangle \\ & + \lambda_1 r(S, S_0) \\ s.t.\quad & U = \{P_v \in \mathbb{R}_+^{L\times L} | P_v\mathbf{1}_L = f(X_i^v), P_v^\top \mathbf{1}_L = \mathbf{y}_i\} \\ & \hat{U} = \{\hat{P} \in \mathbb{R}_+^{L\times L} | \hat{P}\mathbf{1}_L = f(X_i^1), \hat{P}^\top \mathbf{1}_L = f(X_i^2)\} \\ & S \in \mathcal{S}_+, \quad M_{ij} = S_{ii} + S_{jj} - 2S_{ij}. \end{aligned} \tag{8}$$

3.6 Optimization

The $\hat{P}$ is similar with the P when considering the extra modal predictions as the pseudo label. Thus, we analyze the optimization of the Eqs. (5) and (8) has similar solution. In detail, The 1st term in Eq. (5) involves the product of predictors f and cost matrix S, which makes the formulation not joint convex. Consequently, the formulation cannot be optimized easily. We provide the optimization process below:

Fix S, Optimize f_1, f_2. When updating f_1, f_2 with a fixed S, the 2nd term of Eq. (5) is irrelevant to f_1, f_2, and the Eq. (5) can be reformulated as follows:

$$\begin{aligned} & \min_{P_v\in U(f(X_i^v),\mathbf{y}_i)} \sum_{v=1}^{2}\sum_{i=1}^{N}\langle P_v, M\rangle \\ s.t.\quad & U(f(\mathbf{X}_i^v), \mathbf{y}_i) = \{P_v \in \mathbb{R}_+^{L\times L} | P_v\mathbf{1}_L = f(X_i^v), P_v^\top \mathbf{1}_L = \mathbf{y}_i\}. \end{aligned} \tag{9}$$

The empirical risk minimization function of Eq. (9) can be optimized by stochastic gradient descent. However, it requires to evaluate the descent direction for the loss, with respect to the predictor f. Computing the exact subgradient is quite costly, it needs to solve a linear program with $O(L^2)$ constraints, which are with high expense with the L (the label dimension) increase.

Algorithm 1. The Pseudo Code of Learning the Predictors

Input:

- Sampled Batch Dataset: $\{[X_i^1, X_i^2], \mathbf{y}\}_{i=1}^n$, kernelized similar matric S^t, current mapping f_1, f_2
- Parameter: λ

Output:

- Gradient of the target mapping: $\partial L/\partial f_1, \partial L/\partial f_2$

Calculate $M \leftarrow$ Eq. (4)
Initialize $K = exp(-\lambda M - 1)$, $\nabla \leftarrow \mathbf{0}$
for $v = 1, 2$ **do**
  for $i = 1, 2, \ldots, n$ **do**
    $u_i^v \leftarrow \mathbf{1}$
    while u_i^v not converged **do**
      $u_i^v \leftarrow f_v(\mathbf{x}_i^v) \oslash (K(\mathbf{y}_i^v \oslash K^\top u_i^v))$
    end while
    $\nabla^{f_v} \leftarrow \nabla^{f_v} + \frac{log u_i^v}{\lambda} - \frac{log u_i^{v\top}\mathbf{1}}{\lambda L} \cdot \mathbf{1}$
  end for
end for

Similar to [7], the loss is a linear program, and the subgradient can be computed using Lagrange duality. Therefore, we use primal-dual approach to compute the gradient by solving the dual LP problem. From [25], we know that the dual optimal α is, in fact, the subgradient of the loss of training sample $(\mathbf{X}^v, \mathbf{y})$ with respect to its first argument f_v. However, it is costly to compute the exact loss directly. In [26], Sinkhorn relaxation is adopted as the entropy regularization to smooth the transport objective, which results in a strictly convex problem that can be solved through Sinkhorn matrix scaling algorithm, at a speed that is faster than that of transport solvers [26].

For a given training bag of instances $([\mathbf{X}^1, \mathbf{X}^2], \mathbf{y})$, the dual LP of Eq. (9) is:

$$d_M(f_v(\mathbf{X}^v), y) = \max_{\alpha,\beta \in C_M} \alpha^\top f(\mathbf{X}_i^v) + \beta \mathbf{y}, \tag{10}$$

where $C_M = \{\alpha, \beta \in \mathbb{R}^L : \alpha_i + \beta_j < M_{i,j}\}$.

Definition 4 (Sinkhorn Distance). *Given a $L \times L$ cost matrix M, and histograms $(r, c) \in \sum_L$. The Sinkhorn distance is defined as:*

$$\begin{aligned} d_M^\lambda(r, c) &= \min_{P^\lambda \in U(r,c)} \langle P^\lambda, M \rangle \\ P^\lambda &= \arg \min_{P \in U(f(X_i^v), \mathbf{y}_i)} \langle P, M \rangle - \frac{1}{\lambda} H(P), \end{aligned} \tag{11}$$

where $H(P) = -\sum_{i=1}^L \sum_{j=1}^L p_{ij} log p_{ij}$ is the entropy of P, and $\lambda > 0$ is entropic regularization coefficient.

Based on the Sinkhorn theorem, we conclude that the transportation matrix can be written in the form of $P^* = diag(u)Kdiag(v)$, where $K = exp(-\lambda M - 1)$ is the element-wise exponential of $\lambda M - 1$. Besides, $u = exp(\lambda\alpha)$ and $v = exp(\lambda\beta)$.

Therefore, we adopt the well-known Sinkhorn-Knopp algorithm, which is used in [20], [26] to update the target mapping f_v given the ground metric. f_v can be defined as Eq. (1). The detailed procedure is summarized in Algorithm 1, then with the help of Back Propagation technique, gradient descent could be adopted to update the network parameters.

Fix f_1, f_2, Optimize S.

When updating S with the fixed f_1, f_2, the sub-problem can be rewritten as following:

$$\begin{aligned} &\min_S \sum_{v=1}^2 \sum_{i=1}^N \langle P, M \rangle + \lambda_1 r(S, S_0) \\ &s.t. \quad K \in \mathcal{S}_+, \quad M_{ij} = S_{ii} + S_{jj} - 2S_{ij}. \end{aligned} \tag{12}$$

This sub-problem has closed-form solution. The differential can be formulated as:

$$S = (\bar{P} + S_0^{-1} - p)^{-1}, \tag{13}$$

where

$$\bar{P} = \begin{cases} -2P_{ij}, & when \quad i \neq j, \\ \sum_{k \neq i}^L (P_{ik} + P_{ki}), & when \quad i = j \end{cases}.$$

Then, we project S back to positive semi-definite cone as:

$$S = \mathbf{P}roj(S) = Umax(\sigma, 0)U^\top, \tag{14}$$

where *Proj* is a projection operator, U and σ correspond to the eigenvectors and eigenvalues of S. The whole procedure is summarized in Algorithm 2.

Algorithm 2. The Pseudo Code of M3DN

Input:

- Dataset: $\mathcal{D} = \{[X_i^1, X_i^2], \mathbf{y}\}_{i=1}^N$
- Parameter: λ_1, λ
- $maxIter$: T, learning rate: $\{\alpha_t\}_{t=1}^T$

Output:

- Classifiers: f_1, f_2
- Label similar matric: S, M

Initialize $S_0 \leftarrow \mathcal{Y}' \times \mathcal{Y}$
while true **do**
  Create Batch: Randomly pick up n examples from $\mathcal{D}$ without replacement
  Calculate $S^{t+1} \leftarrow$ Eqs. (13) and (14)
  Calculate $\partial L/\partial f_1^t, \partial L/\partial f_2^t \leftarrow$ Algorithm 1
  Weight Propagation step: Obtain the derivative $\partial f_1^t/\partial \mathbf{\Theta}_1$, $\partial f_2^t/\partial \mathbf{\Theta}_2$;
  Update parameters $\mathbf{\Theta}_1, \mathbf{\Theta}_2$
  $Func_{obj}^{t+1} \leftarrow$ calculate obj. value in Eq. (5) with F^{t+1}
  if $\|Func_{obj}^{t+1} - Func_{obj}^t\| \leq \epsilon$ or $t \geq T$ **then**
    Break;
  end if
end while

TABLE 1
Comparison Results (Mean ± std.) of M3DN/M3DNS with Compared Methods on Benchmark Datasets

Methods	Coverage ↓				Macro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3LDA	12.345±.214	11.620±.042	47.400±.622	6.670±.205	.532±.015	.526±.003	.507±.015	.509±.012
MIMLmix	17.114±1.024	15.720±.543	64.130±1.121	14.167±1.140	.472±.018	.554±.096	.471±.019	.493±.020
CS3G	8.168±.137	7.153±.178	50.138±2.146	8.028±.907	.837±.007	.817±.006	.717±.011	.530±.022
DeepMIML	9.242±.331	8.931±.421	27.358±.654	8.369±.119	.766±.035	.795±.022	.827±0.006	.823±.005
M3MIML	11.760±1.121	9.125±.553	42.420±.2.696	5.210±.920	.687±.087	.724±.033	.650±.032	.649±.084
MIMLfast	12.155±.913	12.711±.315	41.048±.831	8.634±.028	.524±.050	.485±.009	.506±.010	.522±.008
SLEEC	9.568±.222	9.494±.105	47.502±.448	7.390±.275	.706±.007	.675±.007	.661±.014	.620±.006
Tram	7.959±.187	8.156±.163	28.417±.945	9.934±.026	.780±.009	.746±.007	.776±.011	.493±.007
ECC	14.818±.086	14.229±.258	47.124±.675	7.941±.194	.532±.013	.484±.009	.630±.023	.634±.009
ML-KNN	10.379±.115	9.523±.072	27.568±.066	4.610±.062	.591±.008	.723±.006	.823±.003	.736±.008
RankSVM	11.439±.196	11.941±.078	37.300±.835	8.292±.054	.512±.019	.499±.009	.521±.033	.501±.001
ML-SVM	11.311±.158	11.755±.270	39.258±.294	7.890±.020	.503±.010	.502±.010	.497±.016	.561±.001
M3DN	7.502±.129	6.936±.065	26.921±.320	4.599±.050	.822 ±.009	.798±.002	.811±.004	.826±.006
M3DNS	**3.947±.307**	**4.214±.202**	**6.119±.262**	**2.764±.071**	**.892±.004**	**.876±.003**	**.838±.003**	**.898±.008**

Methods	Ranking Loss ↓				Example AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3LDA	.301±.009	.377±.002	.247±.001	.257±.006	.707±.008	.630±.005	.770±.006	.652±.009
MIMLmix	.609±.036	.675±.012	.609±.040	.583±.081	.391±.036	.325±.012	.391±.040	.417±.082
CS3G	.118±.005	.155±.005	.202±.009	.170±.032	.881±.005	.835±.005	.798±.009	.642±.032
DeepMIML	.149±.012	.166±.017	.089±.002	.164±.007	.791±.044	.834±.017	.911±.002	.835±.007
M3MIML	.271±.053	.250±.011	.191±.016	.284±.030	.729±.053	.751±.011	.811±.017	.717±.031
MIMLfast	.275±.033	.435±.021	.194±.006	.430±.009	.724±.033	.626±.013	.811±.005	.646±.009
SLEEC	.316±.009	.413.006	.455±.005	.512±.008	.843±.003	.761±.005	.796±.002	.713±.008
Tram	.132±.004	.203±.007	.117±.004	.456±.004	.867±.004	.797±.007	.883±.005	.591±.001
ECC	.804±.024	.928±.013	.461±.009	.617±.020	.642±.005	.529±.012	.775±.005	.697±.013
ML-KNN	.235±.005	.264±.004	.097±.002	.176±.003	.764±.005	.736±.004	.903±.001	.824±.003
RankSVM	.236±.006	.344±.001	.199±.098	.323±.008	.763±.006	.656±.001	.801±.098	.677±.001
ML-SVM	.232±.005	.337±.009	.179±.004	.314±.002	.768±.005	.662±.009	.822±.004	.686±.002
M3DN	.108±.003	.151±.002	**.085±.002**	**.117±.002**	.891±.003	.850±.003	**.915±.003**	**.883±.001**
M3DNS	**.108±.001**	**.142±.002**	.112±.003	.119±.003	**.899±.004**	**.858±.005**	.898±.008	.881±.006

Methods	Average Precision ↑				Micro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3LDA	.371±.005	.311±.007	.399±.007	.338±.005	.693±.006	.609±.002	.773±.005	.657±.008
MIMLmix	.207±.038	.183±.008	.213±.041	.167±.020	.436±.024	.438±.060	.434±.026	.472±.015
CS3G	**.749±.008**	.622±.006	.542±.012	.597±.031	.867±.005	.827±.006	.738±.007	.557±.021
DeepMIML	.621±.027	.619±.025	.633±.005	.583±.008	.835±.009	.802±.017	.914±.002	.852±.003
M3MIML	.423±.056	.490±.020	.446±.030	.443±.076	.745±.034	.707±.017	.816±.020	.762±.020
MIMLfast	.432±.064	.339±.013	.413±.005	.365±.021	.712±.022	.540±.010	.745±.012	.630±.005
SLEEC	.608±.006	.473±.010	.565±.003	.392±.007	.824±.004	.736±.005	.795±.002	.701±.005
Tram	.653±.011	.523±.008	.494±.007	.336±.002	.842±.003	.782±.007	.883±.006	.554±.002
ECC	.416±.012	.278±.011	.462±.007	.438±.014	.646±.004	.514±.008	.779±.005	.702±.009
ML-KNN	.398±.006	.403±.010	.585±.002	.439±.006	.752±.005	.729±.003	.905±.002	.817±.004
RankSVM	.467±.005	.364±.004	.427±.066	.401±.001	.748±.005	.649±.004	.791±.093	.680±.003
ML-SVM	.466±.006	.367±.006	.441±.007	.443±.007	.753±.004	.656±.009	.825±.004	.724±.001
M3DN	.719±.006	.634±.003	.680±.005	**.691±.001**	**.876±.003**	.834±.001	**.918±.002**	.877±.003
M3DNS	.698±.002	**.637±.007**	**.691±.004**	.634±.003	.858±.003	**.863±.004**	.877±.006	**.878±.005**

Eq. (8) can be easily optimized as M3DN with GCD method. Without any loss of generality, in semi-supervised scenario, the extra modal prediction $f(X^{3-i})$ can be regarded as the pseudo label similar to the $\mathbf{y}$ in the supervised term when updating f_1, f_2. S can be updated in similar form, where

$$\bar{P} = \begin{cases} -2(P_{ij} + \hat{P}_{ij}), & when \quad i \neq j, \\ \sum_{k \neq i}^{L} (P_{ik} + P_{ki} + \hat{P}_{ik} + \hat{P}_{ki}), & when \quad i = j \end{cases}.$$

4 Experiments

4.1 Datasets and Configurations

M3DN/M3DNS can learn more discriminative multi-modal feature representation on bag level for supervised/semi-supervised multi-label classification, while considering the label correlation among different labels. Thus, in this section, we provide empirical investigations and performance comparisons of M3DN on multi-label classification and label correlation. Without any loss of generality, we experiment on 4 public real-world datasets, i.e., FLICKR25K [27], IAPR TC-12 [28], MS-COCO [29] and NUS-WIDE [30]. Besides, we experiment on 1 real-world complex article dataset, i.e., WKG Game-Hub. *FLICKR25K*: consists of 25,000 images collected from Flickr website, and each image is associated with several textual tags. The text for each instance is represented as a 1386-dimensional bag-of-words vector. Each point is manually annotated with 24 labels. We select 23,600 image-text pairs that belong to the 10 most frequent concepts; *IAPR TC-12*: consists of 20,000 image-text pairs which annotate 255 labels. The text for each point is represented as a

TABLE 2
Comparison Results (mean ± std.) of M3DN/M3DNS with Compared Methods on WKG Game-Hub Dataset

Methods	Content Modality					
	Coverage ↓ ($\times 10^2$)	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
M3LDA	.466±.020	.470±.015	1.000±1.000	.360±.056	.098±.001	.381±.036
MIMLmix	.334±.003	.507±.002	.445±.006	.539±.001	.111±.001	.540±.003
CS3G	.362±.002	.593±.001	.340±.003	.659±.003	.371±.002	.614±.007
DeepMIML	.341±.010	.533±.018	.415±.027	.186±.025	.600±.030	.634±.014
M3MIML	N/A	N/A	N/A	N/A	N/A	N/A
MIMLfast	.363±.040	.496±.050	.414±.056	.585±.056	.162±.033	.567±.040
M3DN	.258±.006	.761±.016	.276±.008	.723±.008	.329±.002	.753±.007
M3DNS	**.246±.002**	**.763±.001**	**.255±.002**	**.744±.002**	**.332±.001**	**.763±.001**
Methods	Image Modality					
	Coverage ↓ ($\times 10^2$)	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
M3LDA	.466±.010	.455±.054	1.000±.000	.359±.019	.098±.001	.384±.030
MIMLmix	.329±.002	.502±.003	.427±.005	.557±.001	.114±.001	.560±.002
CS3G	.395±.004	.545±.001	.405±.003	.595±.003	.304±.003	.563±.006
DeepMIML	.383±.006	.512±.002	.515±.009	.484±.009	.121±.001	.488±.018
M3MIML	N/A	N/A	N/A	N/A	N/A	N/A
MIMLfast	.402±.070	.512±.061	.433±.059	.566±.059	.170±.037	.547±.058
M3DN	.175±.001	.896±.001	.210±.002	.789±.002	.402±.001	.586±.000
M3DNS	**.164±.001**	**.910±.003**	**.196±.001**	**.803±.001**	**.407±.000**	**.869±.000**
Methods	Overall					
	Coverage ↓ ($\times 10^2$)	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
M3LDA	.466±.008	.468±.026	1.000±.000	.359±.030	.098±.001	.383±.017
MIMLmix	.358±.003	.504±.002	.488±.007	.496±.001	.101±.001	.519±.003
CS3G	.361±.004	.589±.003	.346±.004	.653±.004	.365±.001	.612±.004
DeepMIML	.362±.005	.518±.002	.488±.008	.512±.008	.125±.001	.524±.018
M3MIML	N/A	N/A	N/A	N/A	N/A	N/A
MIMLfast	.393±.060	.509±.064	.430±.052	.596±.052	.170±.036	.549±.054
SLEEC	.603±.013	.518±.004	.756±.007	.493±.005	.150±.006	.583±.006
Tram	.712±.005	.429±.008	.109±.010	.545±.003	.164±.008	.464±.006
ECC	.622±.017	.630±.002	.632±.009	.530±.017	.198±.002	.592±.011
ML-KNN	.675±.020	.712±.006	.175±.003	.802±.015	.265±.004	.814±.001
RankSVM	N/A	N/A	N/A	N/A	N/A	N/A
ML-SVM	.742±.023	.561±.002	.223±.009	.782±.008	.234±.003	.793±.002
M3DN	.163±.003	.924±.002	.190±.004	.809±.004	.401±.003	.866±.003
M3DNS	**.149±.002**	**.933±.001**	**.180±.009**	**.828±.003**	**.409±.001**	**.880±.001**

Six commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger/smaller the better of the criterion.

2912-dimensional bag-of-words vector; *NUS-WIDE*: contains 260,648 web images, and images are associated with textual tags where each point is annotated with 81 concept labels. We select 195,834 image-text pairs that belong to the 21 most frequent concepts. The text for each point is represented as a 1000-dimensional bag-of-words vector; *MS-COCO*: contains 82,783 training, 40,504 validation image-text pairs which belong to 91 categories. We select 38,000 image-text pairs that belong to the 20 most frequent concepts. The text for each point is represented as a 2912-dimensional bag-of-words vector; *WKG Game-Hub*: consists of 13,750 articles collected from the Game-Hub of " Strike of Kings" with 1744 concept labels. We select 11,000 image-text pairs that belong to the 54 most frequent concepts. Each article contains several images and content paragraphs, and the text for each point is represented as a 300-dimensional w2v vector.

We run each compared method 30 times for all datasets, and then randomly select 70 percent for training and the remaining are for test. For all the training examples, we randomly choose 30 percent as the labeled data, and the other 70 percent as unlabeled ones as [31]. For the 4 benchmark datasets, each image is divided into 10 regions using [32] as image bag, while the corresponding text tags are also separated into several independent tags as text bag. For the WKG Game-Hub dataset, each article is denoted as an image bag and a content bag. The deep network for image encoder is implemented the same as Resnet-18 [33]. We run the following experiments with the implementation of an environment on NVIDIA K80 GPUs server, and our model can be trained around 290 images per second with a single K80 GPGPU. In the training phase, the parameters λ_1 is selected by 5-fold cross validation from $\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$ with further splitting on only the training datasets, i.e., there is no overlap between the test set and the validation set for parameter picking up. Empirically, when the variation between the objective values of Eq. (13) is

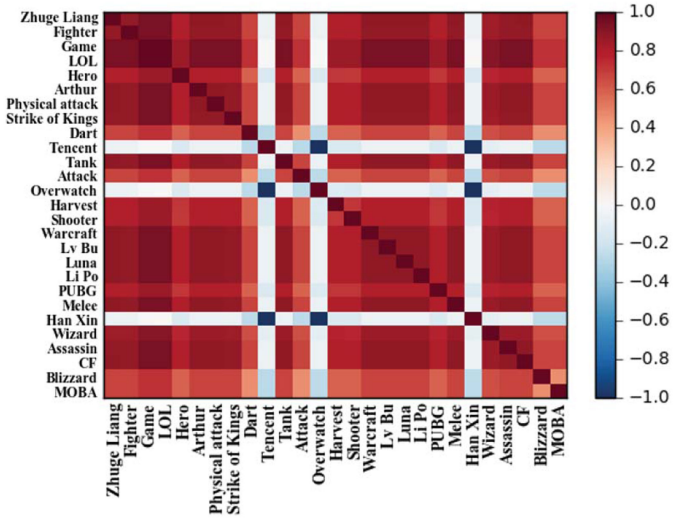


Fig. 5. Illustration of learned label correlations for different datasets, and the value has been scaled in [−1,1]. Red color indicates a positive correlation, and blue one indicates a negative correlation.

less than 10^{-6} in iteration, we treat M3DN or M3DNS converged.

4.2 Compared Methods

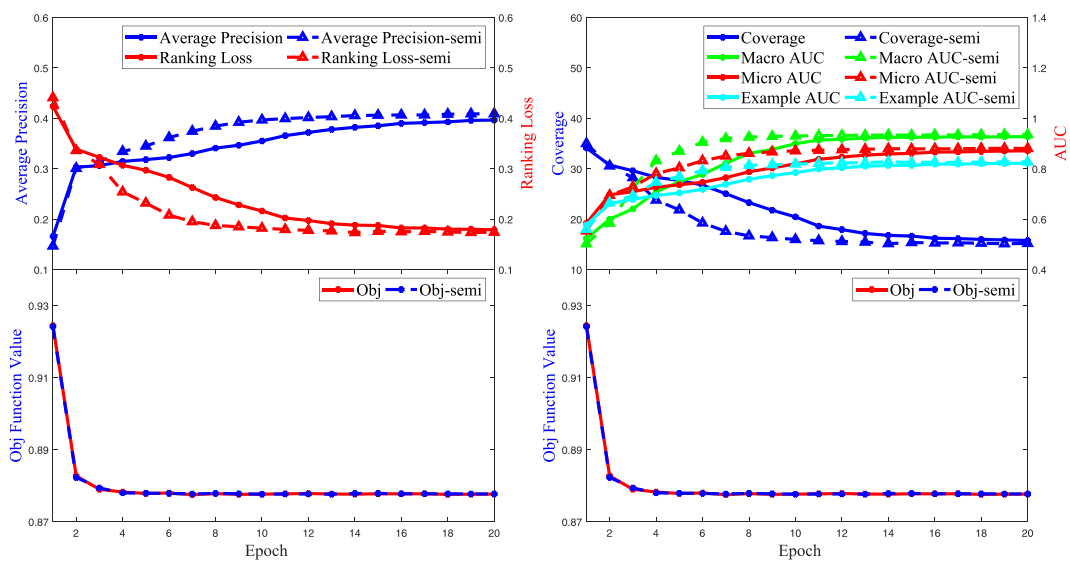
In our experiments, first, we compare our methods with multi-modal multi-instance multi-label methods, i.e., M3LDA [3], MIMLmix [4]. Besides, M3DN can be degenerated into different settings, we also compare with multi-modal multi-label methods, i.e., CS3G [34]; multi-instance multi-label methods, i.e., DeepMIML [14], M3MIML [35], MIMLfast [36]. Moreover, we compare our methods with multi-label methods, i.e., SLEEC [37], Tram [38], ECC [39], ML-KNN [40], RankSVM [41], ML-SVM [42]. Specifically, for multi-modal multi-label methods, we calculate the average of all instances' representations as the bag-level feature representation. In the multi-instance multi-label methods, all modalities of a dataset are concatenated together as a single modal input. As to the multi-label learners, we first calculate bag-level feature representation for different modalities independently, then we concatenate all modalities together as a single modal input. As to the semi-supervised scenario, considering that existing M3 methods are supervised methods, we compare our methods with semi-supervised multi-modal multi-label methods, i.e., CS3G [34]; and semi-supervised multi-label methods, i.e., Tram [38], COINS [17], iMLU [43].

4.3 Benchmark Comparisons

M3DN is compared with other methods on 4 benchmark datasets to demonstrate the abilities. Results of compared methods and M3DN/M3DNS on 6 commonly used criteria are listed in Table 1. The best performance for each criterion is bolded. ↑ / ↓ indicates that the larger/smaller, the better of the criterion. From the results, it is obvious that our M3DN/M3DNS approaches can achieve the best or second performance on most datasets with different performance measures. Therefore the M3DN/M3DNS approach are highly competitive multi-modal multi-label learning methods.

4.4 Complex Article Classification

In this subsection, M3DN approach is tested on the real-world complex article classification problem, i.e., WKG Game-Hub dataset. There are 13,570 articles in collection, with image and text modalities to promote classification. Specifically, each article contains variable number of images and text paragraphs. Thus, each article can be divided into both image bag and text bag. Comparison results (independent modalities and overall) against compared methods are listed in Table 2, where notation "N/A" means the method cannot give a result in 60 hours. We use the same 6 measurement criteria as in previous subsection, i.e., Coverage, Ranking Loss, Average Precision, Macro AUC, example AUC and Micro AUC. It is notable that multi-label methods concatenate all of the modal features, which have no independent modal classification performance. The results show that on both of the independent modalities and overall prediction, our M3DN



(a) M3DN (b) M3DNS

Fig. 6. Objective function value convergence and corresponding classification performance (Coverage, Ranking Loss, Average Precision, Macro AUC, example AUC, and Micro AUC) versus number of iterations of M3DN and M3DNS.

Fig. 7. Sample test complex articles predictions of the WKG Game-Hub. Left is the image bag, the middle are label predictions, and the right is the context bag.

TABLE 3
Semi-Supervised Comparison Results (Mean ± std.) of M3DNS with Compared Methods on Four Benchmark Datasets

Methods	Coverage ↓				Macro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
CS3G	10.346±.227	7.545±.056	6.968±.060	9.819±.931	.844±.006	.798±.002	.699±.006	.662±.077
Tram	6.857±.645	5.793±.359	55.059±1.888	9.359±.223	.827±.001	.805±.001	**.891±.001**	.890±.045
COINS	22.940±5.082	20.598±4.513	25.839±10.629	20.126±4.072	.891±.004	.863±.006	.814±.014	.873±.017
iMLU	23.411±1.160	23.401±8.939	26.462±5.548	21.030±4.844	.880±.009	.835±.003	.812±.004	.835±.048
M3DNS	**3.947±.307**	**4.214±.202**	**6.119±.262**	**2.764±.071**	**.892±.004**	**.876±.003**	.838±.003	**.898±.008**

Methods	Ranking Loss ↓				Example AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
CS3G	.109±.003	.120±.001	.168±.001	.196±.070	.890±.003	.879±.001	.831±.001	.803±.070
Tram	.108±.002	.119±.001	.183±.001	.183±.076	.893±.002	**.880±.001**	.816±.001	.816±.076
COINS	.150±.009	.171±.002	.305±.008	.297±.028	.849±.009	.828±.002	.694±.008	.702±.028
iMLU	.167±.007	.242±.014	.344±.013	.346±.015	.832±.007	.757±.014	.655±.013	.653±.015
M3DNS	**.108±.001**	**.142±.002**	**.112±.003**	**.119±.003**	**.899±.004**	.858±.005	**.898±.008**	**.881±.006**

Methods	Average Precision ↑				Micro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
CS3G	.671±.003	**.678±.001**	.661±.003	.586±.083	.860±.007	.820±.002	.769±.003	.724±.084
Tram	.670±.006	.507±.004	.348±.003	.318±.091	**.910±.001**	.859±.001	.874±.001	.868±.057
COINS	.570±.007	.419±.007	.258±.033	.216±.016	.884±.007	.852±.003	.788±.018	.856±.025
iMLU	.538±.015	.325±.016	.220±.043	.187±.015	.860±.015	.793±.007	.760±.013	.798±.078
M3DNS	**.698±.002**	.637±.007	**.691±.004**	**.634±.003**	.858±.003	**.863±.004**	**.877±.006**	**.878±.005**

Six commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger/smaller the better of the criterion.

TABLE 4
Semi-Supervised Comparison Results (Mean ± std.) of M3DNS with Compared Methods on WKG Game-Hub Dataset

Methods	Coverage ↓ ($\times 10^3$)	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
CS3G	.326±.002	.683±.021	.187±.014	.812±.014	.404±.057	.728±.026
Tram	1.731±.083	.854±.031	.190±.024	.809±.024	.245±.046	.852±.024
COINS	.186±.021	.782±.087	.252±.029	.747±.029	.195±.037	.783±.072
iMLU	.225±.027	.786±.070	.288±.033	.711±.030	.169±.026	.763±.010
M3DNS	**.149±.002**	**.933±.001**	**.180±.009**	**.828±.003**	**.409±.001**	**.880±.001**

Six commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger/smaller, the better of the criterion.

TABLE 5
Ablation Study Results (Mean ± std.) of M3DNS on Four Benchmark Datasets

Methods	Coverage ↓				Macro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3DNS-F	8.678±.002	6.875±.010	9.280±.003	11.042±.009	**.896±.000**	.868±.000	.829±.002	.858±.001
M3DNS-M	8.889±.010	6.964±.003	9.764±.001	11.043±.005	.885±.001	.862±.000	.757±.001	.843±.000
M3DNS-MP	4.039±.021	5.047±.038	.8.708±.028	3.230±.003	.874±.000	.860±.000	.779±.001	.837±.001
M3DNS	**3.947±.307**	**4.214±.202**	**6.119±.262**	**2.764±.071**	.892±.004	**.876±.003**	**.838±.003**	**.898±.008**
Methods	Ranking Loss ↓				Example AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3DNS-F	**.074±.000**	.146±.000	.134±.001	.184±.000	.825±.000	.804±.000	.866±.001	.816±.000
M3DNS-M	.109±.001	.149±.000	.150±.000	.132±.000	.783±.001	.696±.000	.686±.000	.540±.001
M3DNS-MP	.106±.000	.145±.001	.150±.001	.190±.001	.818±.000	.790±.001	.848±.000	.810±.001
M3DNS	.108±.001	**.142±.002**	**.112±.003**	**.119±.003**	**.899±.004**	**.858±.005**	**.898±.008**	**.881±.006**
Methods	Average Precision ↑				Micro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3DNS-F	.693±.000	.592±.000	**.693±.000**	.624±.000	**.917±.000**	.863±.002	.868±.003	.877±.000
M3DNS-M	.614±.002	.588±.000	.639±.001	.610±.000	.819±.001	.790±.000	.850±.003	.814±.001
M3DNS-MP	.681±.000	.582±.001	.684±.001	.616±.001	.809±.000	.791±.000	.846±.001	.807±.002
M3DNS	**.698±.002**	**.637±.007**	.691±.004	**.634±.003**	.858±.003	**.863±.004**	**.877±.006**	**.878±.005**

Six commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger/smaller the better of the criterion.

and M3DNS approaches can get the best results over all criteria. The statistics validates the effectiveness of our method when solving the complex article classification problem.

4.5 Label Correlations Exploration

Since M3DN can learn label correlation explicitly, in this subsection, we examine effectiveness of M3DN in label correlations exploration. Due to page limitation, the exploration is conducted on the real-world dataset WKG Game-Hug. We randomly sampled 27 labels, with the learned ground metric shown in Fig. 5, and scaled the original value in cost matrix into $[-1, 1]$. Red color indicates a positive correlation, and blue indicates a negative correlation. We can see that the learned pairwise cost accords with intuitions. Taking a few examples, the cost between Overwatcha and Tencent indicates a very small correlation, and this is reasonable as the game Overwatch has no correlation with Tencent. While the cost between (Zhuge Liang, Wizard) indicates a very strong correlation, since Zhuge Liang belongs to the wizard role in the game.

4.6 Empirical Investigation on Convergence

To investigate the convergence of M3DN iterations empirically, we record the objective function value, i.e., the value of Eq. (5) and the different criteria of classification performance of M3DN/M3DNS in each epoch. Due to page limits, results on WKG Game-Hug dataset are plotted in Fig. 6. It clearly reveals that the objective function value decreases as the iterations increase, and all of the classification performance is stable after several iterations in Fig. 6. Moreover, these additional experiment results indicate that our M3DN/M3DNS can converge fast, i.e., M3DN converges after 10 epoches.

4.7 Empirical Illustrative Examples

Fig. 7 shows 6 illustrative examples of the classification results on WKG Game-Hub dataset. Qualitatively, illustration of the predictions clearly discovers the modal-instance-label relation on the test set. E.g., the first example shows that the article has separated three images and four content paragraphs. We can predict the Zhuge liang, battlefront labels from both the images and contents, and acquire the master, cooperation labels form the context.

5 Conclusion

This paper focuses on the issues of complex objects classification with semi-supervised M3 information, and extends our preliminary research [44]. Complex objects, i.e., the articles, the videos, etc, can always be represented by multi-modal

TABLE 6
Ablation Study Results (Mean ± std.) of M3DNS on WKG Game-Hub Dataset

Methods	Coverage ↓ ($\times 10^3$)	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
M3DNS-F	.279±.003	.821±.000	.183±.001	.822±.000	.345±.000	.872±.000
M3DNS-M	.287±.041	.840±.000	.182±.001	.823±.000	.379±.001	.870±.002
M3DNS-MP	.286±.008	.818±.000	.190±.001	.817±.001	.333±.000	.869±.002
M3DNS	**.149±.002**	**.933±.001**	**.180±.009**	**.828±.003**	**.409±.001**	**.880±.001**

Six commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger/smaller, the better of the criterion.

TABLE 7
Missing Modal Comparison Results (Mean ± std.) of M3DNS on Four Benchmark Datasets

Methods	Coverage ↓				Macro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
0%	**3.947±.307**	**4.214±.202**	**6.119±.262**	**2.764±.071**	**.892±.004**	**.876±.003**	**.838±.003**	**.898±.008**
10%	4.012±.013	5.017±.015	6.443±.002	2.815±.018	.891±.000	.858±.001	.822±.000	.865±.001
30%	4.033±.009	5.604±.013	6.324±.007	2.834±.010	.888±.001	.870±.001	.817±.001	.866±.000
50%	4.080±.003	5.862±.000	6.496±.004	3.381±.002	.887±.000	.862±.004	.812±.000	.834±.001
70%	4.180±.021	5.840±.002	6.378±.005	3.213±.001	.880±.000	.861±.000	.806±.001	.846±.000
90%	4.485±.004	5.897±.001	6.816±.017	3.615±.004	.869±.000	.856±.000	.781±.000	.820±.001
Methods	Ranking Loss ↓				Example AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
0%	**.108±.001**	**.142±.002**	**.112±.003**	**.119±.003**	**.899±.004**	**.858±.005**	**.898±.008**	**.881±.006**
10%	.178±.000	.159±.000	.140±.000	.178±.000	.892±.000	.840±.000	.859±.000	.871±.000
30%	.180±.000	.150±.001	.138±.000	.178±.000	.879±.000	.849±.000	.861±.001	.871±.000
50%	.181±.000	.157±.000	.143±.000	.192±.000	.878±.001	.842±.000	.856±.000	.857±.000
70%	.185±.001	.155±.000	.139±.000	.187±.001	.874±.001	.844±.000	.854±.000	.862±.004
90%	.190±.002	.159±.001	.156±.000	.199±.000	.869±.000	.839±.001	.843±.000	.850±.000
Methods	Average Precision ↑				Micro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
0%	**.698±.002**	**.637±.007**	**.691±.004**	**.634±.003**	.858±.003	**.863±.004**	**.877±.006**	**.878±.005**
10%	.689±.000	.631±.000	.684±.000	.631±.000	.817±.000	.845±.000	.860±.000	.870±.000
30%	.678±.000	.635±.000	.686±.002	.631±.000	.812±.000	.855±.002	.862±.001	.869±.000
50%	.678±.000	.628±.000	.679±.001	.598±.000	.815±.000	.849±.000	.857±.000	.853±.000
70%	.666±.001	.629±.000	.680±.000	.593±.000	.808±.001	.848±.000	.862±.000	.858±.000
90%	.659±.000	.610±.000	.663±.001	.590±.000	.802±.000	.846±.000	.842±.000	.846±.000

Six commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger/smaller the better of the criterion.

multi-instance information, with multiple labels. However, we usually only have bag-level consistency among different modalities. Therefore, Multi-modal Multi-instance Multi-label (M3) learning provides a framework for handling such task. Meanwhile, previous M3 methods rarely consider label correlation and unlabeled data. In this paper, we propose a novel Multi-modal Multi-instance Multi-label Deep Network (M3DN) framework, and exploit label correlation based on the Optimal Transport (OT) theory. Moreover, considering unlabel information, M3DNS utilizes the instance-label and bag-level unlabel information for more excellent performance. Experiments on the real world benchmark datasets and special complex article dataset WKG Game-Hub validate effectiveness of the proposed methods. Meanwhile, how to extend to multiple modalities is an interesting future work.

Appendix A
Semi-Supervised Classification

M3DNS takes unlabeled instances into consideration, i.e., using auto-encoder for single modal network, and consistency among different modalities for joint predictions. Thus, in this section, we provide empirical investigations and performance comparisons of M3DNS with several state-of-the-art semi-supervised methods. The introduction to data configuration and comparison methods are in Sections 4.1 and 4.2. The results are recorded in Tables 3 and 4. The results indicate that M3DNS approach can achieve the best or second performance on most datasets with different performance measures, thus M3DNS can make better use of unlabeled data.

Appendix B
Ablation Study

In order to explore the impact of different operators in the network structure, we conduct more experiments. In detail, 1) in order to verify different pooling methods to get bag-level prediction, we compare max pooling with mean pooling, denoted as M3DNS-M with mean pooling; 2) based on the better bag-level pooling method, we compare average prediction with max prediction to evaluate different ensemble methods for final predictions, denoted as M3DNS-MP with max operator; 3) based on the better pooling method and

TABLE 8
Missing Modal Comparison Results (Mean ± std.) of M3DNS on WKG Game-Hub Dataset

Methods	Coverage ↓ ($\times 10^3$)	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
0%	**.149±.002**	**.933±.001**	**.180±.009**	**.828±.003**	**.409±.001**	**.880±.001**
10%	.264±.007	.844±.000	.183±.000	.776±.000	.379±.000	.877±.000
30%	.273±.003	.830±.000	.191±.000	.768±.001	.363±.000	.868±.000
50%	.276±.013	.825±.000	.193±.000	.766±.000	.350±.000	.866±.000
70%	.284±.002	.812±.000	.201±.000	.758±.000	.336±.000	.859±.000
90%	.299±.008	.802±.000	.207±.000	.752±.000	.329±.001	.848±.000

Six commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger/smaller, the better of the criterion.

prediction operator, we fix the ground metric as the initial value without any change to explore the advantage of learning ground metric, denoted as M3DNS-F. The results are recorded in Tables 5 and 6. It is notable that M3DNS is with max pooling, mean prediction operator. The results reveal that max pooling are always better than the mean pooling in getting bag-level prediction. This is because there are often only a few positive examples in the bag that can represent the prediction of this bag, yet mean pooling will bring a lot of noise on the contrast. This phenomenon is also consistent with the assumption of multi-instance learning. Furthermore, the results reveal that mean prediction operator is always better than the max operator, which is also according with the ensemble learning methods. An interesting thing is that, though M3DNS is better than M3DNS-F on most datasets, it is worse on one dataset, i.e., FLICKR25K. This result shows that learning ground metric is not definitely effective. Considering the noise data, it may affect the learning of ground metric. Thus, how to modify the learning process or design a suitable initialization method could be an interesting future work.

Appendix C
Comparison with Missing Modality

Specifically, in order to explore the impact of modal missing scenario, we conduct more experiments. Following [45], in each split, we randomly select 10 to 90 percent of examples, with 20 percent as interval, for homogeneous examples with complete modality. And the remaining are incomplete instances. The results are recorded in Tables 7 and 8. It shows that M3DNS achieves competitive results when comparing the results in Tables 1, 2, 5, and 6 with missing modalities, and the performance of M3DNS increases faster than compared methods as incomplete ratio decreases.

Acknowledgments

This research was supported by the National Key R&D Program of China (2018YFB1004300), NSFC (61773198, 61632004, 61751306), NSFC-NRF Joint Research Project under Grant 61861146001, and Collaborative Innovation Center of Novel Software Technology and Industrialization, Postgraduate Research & Practice Innovation Program of Jiangsu province (KYCX18-0045).

References

[1] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, Nov. 2015.
[2] P. Yang, H. Yang, H. Fu, D. Zhou, J. Ye, T. Lappas, and J. He, "Jointly modeling label and feature heterogeneity in medical informatics," *ACM Trans. Knowl. Discovery Data*, vol. 10, no. 4, pp. 39:1–39:25, 2016.
[3] C. Nguyen, D. Zhan, and Z. Zhou, "Multi-modal image annotation with multi-instance multi-label LDA," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1558–1564.
[4] C. Nguyen, X. Wang, J. Liu, and Z. Zhou, "Labeling complicated objects: Multi-view multi-instance multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 2013–2019.
[5] P. Yang and J. He, "Model multiple heterogeneity via hierarchical multi-latent space learning," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1375–1384.
[6] S. Huang and Z. Zhou, "Multi-label learning by exploiting label correlations locally," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 949–955.
[7] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. A. Poggio, "Learning with a Wasserstein loss," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2053–2061.
[8] A. Rolet, M. Cuturi, and G. Peyre, "Fast dictionary learning with a smoothed Wasserstein loss," in *Proc. 19th Int. Conf. Artif. Intell. Statistics*, 2016, pp. 630–638.
[9] A. Blum and T. M. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.
[10] U. Brefeld, T. Gartner, T. Scheffer, and S. Wrobel, "Efficient co-regularised least squares regression," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 137–144.
[11] C. Villani, *Optimal Transport: Old and New*, vol. 338, Berlin, Germany: Springer, 2008.
[12] Z. Fang and Z. M. Zhang, "Simultaneously combining multi-view multi-label learning with maximum margin classification," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 864–869.
[13] P. Yang, J. He, H. Yang, and H. Fu, "Learning from label and feature heterogeneity," in *Proc. IEEE Int. Conf. Data Mining*, 2014, pp. 1079–1084.
[14] J. Feng and Z. Zhou, "Deep MIML network," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1884–1890.
[15] W. Bi and J. T. Kwok, "Multilabel classification with label correlations and missing labels," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1680–1686.
[16] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[17] W. Zhan and M. Zhang, "Inductive semi-supervised multi-label learning with co-training," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1305–1314.
[18] W. Qian, B. Hong, D. Cai, X. He, and X. Li, "Non-negative matrix factorization with sinkhorn distance," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1960–1966.
[19] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853–1865, Sep. 2017.
[20] M. Cuturi and D. Avis, "Ground metric learning," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 533–564, 2014.
[21] P. Zhao and Z.-H. Zhou, "Label distribution learning by optimal transport," in *Proc. AAAI Conf. Artif. Intell.*, 2018.
[22] R. Yossi, L. Guibas, and C. Tomasi, "The earth mover's distance multi-dimensional scaling and color-based image retrieval," in *Proc. ARPA Image Understanding Workshop*, 1997, pp. 661–668.

[23] D. Kedem, S. Tyree, K. Q. Weinberger, F. Sha, and G. R. G. Lanckriet, "Non-linear metric learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 2582–2590.
[24] J. Hoffman, E. Rodner, J. Donahue, B. Kulis, and K. Saenko, "Asymmetric and category invariant feature transformations for domain adaptation," *Int. J. Comput. Vis.*, vol. 109, no. 1–2, pp. 28–41, 2014.
[25] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, vol. 6. Belmont, MA, USA: Athena Scientific, 1997.
[26] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2292–2300.
[27] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.
[28] H. J. Escalante, C. A. Hernandez, J. A. Gonzalez, A. Lopez-Lopez, M. Montes-y-Gomez, E. F. Morales, L. E. Sucar, L. V. Pineda, and M. Grubinger, "The segmented and annotated IAPR TC-12 benchmark," *Comput. Vis. Image Understanding*, vol. 114, no. 4, pp. 419–428, 2010.
[29] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[30] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, Art. no. 48.
[31] M. Zhang, Y. Li, X. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview," *Frontiers Comput. Sci.*, vol. 12, no. 2, pp. 191–202, 2018.
[32] R. B. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385*, 2015.
[34] H. Ye, D. Zhan, X. Li, Z. Huang, and Y. Jiang, "College student scholarships and subsidies granting: A multi-modal multi-label approach," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, pp. 559–568.
[35] M. Zhang and Z. Zhou, "M3MIML: A maximum margin method for multi-instance multi-label learning," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 688–697.
[36] S. Huang, W. Gao, and Z. Zhou, "Fast multi-instance multi-label learning," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1868–1874.
[37] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 730–738.
[38] X. Kong, M. K. Ng, and Z. Zhou, "Transductive multilabel learning via label set propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 704–719, Mar. 2013.
[39] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, 2011.
[40] M. Zhang and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[41] T. Joachims, "Optimizing search engines using click through data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 133–142.
[42] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, 2004.
[43] L. Wu and M. Zhang, "Multi-label classification with unlabeled data: An inductive approach," in *Proc. Asian Conf. Mach. Learn.*, 2013, pp. 197–212.
[44] Y. Yang, Y. Wu, D. Zhan, Z. Liu, and Y. Jiang, "Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 2594–2603.
[45] S. Li, Y. Jiang, and Z. Zhou, "Partial multi-view clustering," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1968–1974.

Yang Yang is working towards the PhD degree with the National Key Lab for Novel Software Technology, the Department of Computer Science & Technology, Nanjing University, China. His research interests include machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining.

Zhao-Yang Fu is working towards the MSc degree with the National Key Lab for Novel Software Technology, the Department of Computer Science & Technology, Nanjing University, China. His research interests include machine learning and data mining, including multi-modal learning.

De-Chuan Zhan received the PhD degree in computer science, Nanjing University, China, in 2010. At the same year, he became a faculty member in the Department of Computer Science and Technology at Nanjing University, China. He is currently an associate professor with the Department of Computer Science and Technology at Nanjing University. His research interests include machine learning, data mining and mobile intelligence. He has published more than 20 papers in leading international journal/conferences. He serves as an editorial board member of *IDA* and *IJAPR*, and serves as SPC/PC in leading conferences such as IJCAI, AAAI, ICML, NIPS, etc.

Zhi-Bin Liu received the BS degree in automatic control engineering from Central South University, Changsha, China, in 2004, and the MS and PhD degrees in control science and engineering from Tsinghua Universtiy, Beijing, China, in 2010. His research interests include big data minning, machine learning, AI, NLP, computer vision, information fusion, etc.

Yuan Jiang received the PhD degree in computer science from Nanjing University, China, in 2004. At the same year, she became a faculty member in the Department of Computer Science & Technology at Nanjing University, China, and is currently a professor. She was selected in the Program for New Century Excellent Talents in the University, Ministry of Education in 2009. Her research interests are mainly in artificial intelligence, machine learning, and data mining. She has published more than 50 papers in leading international/national journals and conferences.