This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Deep Semisupervised Multiview Learning With Increasing Views

Peng Hu, Xi Peng, *Member, IEEE*, Hongyuan Zhu, *Member, IEEE*, Liangli Zhen, Jie Lin, Huaibai Yan, and Dezhong Peng

***Abstract*—In this article, we study two challenging problems in semisupervised cross-view learning. On the one hand, most existing methods assume that the samples in all views have a pairwise relationship, that is, it is necessary to capture or establish the correspondence of different views at the sample level. Such an assumption is easily isolated even in the semisupervised setting wherein only a few samples have labels that could be used to establish the correspondence. On the other hand, almost all existing multiview methods, including semisupervised ones, usually train a model using a fixed dataset, which cannot handle the data of increasing views. In practice, the view number will increase when new sensors are deployed. To address the above two challenges, we propose a novel method that employs multiple independent semisupervised view-specific networks (ISVNs) to learn representation for multiple views in a view-decoupling fashion. The advantages of our method are two-fold. Thanks to our specifically designed autoencoder and pseudolabel learning paradigm, our method shows an effective way to utilize both the labeled and unlabeled data while relaxing the data assumption of the pairwise relationship, that is, correspondence. Furthermore, with our view decoupling strategy, the proposed ISVNs could be separately trained, thus efficiently handling the data of increasing views without retraining the entire model. To the best of our knowledge, our ISVN could be one of the first attempts to make handling increasing views in the semisupervised setting possible, as well as an effective solution to the noncorresponding problem. To verify the effectiveness and efficiency of our method, we conduct comprehensive experiments by comparing 13 state-of-the-art approaches on four multiview datasets in terms of retrieval and classification.**

***Index Terms*—Cross-view retrieval, heterogeneous recognition, latent common space, semisupervised multiview learning.**

I. Introduction

MULTIVIEW learning aims to learn a common space shared by different views, which has shown promising performance to facilitate the downstream tasks, such as multiview clustering [1]–[4], classification [5]–[7], and retrieval [8]–[10]. Among them, cross-view retrieval and classification are two interesting topics due to their flexibility in real-world applications. More specifically, cross-view retrieval/classification aims to flexibly retrieve/recognize semantically relevant samples of one view (i.e., query/probe set) from another view (i.e., database/gallery set). The key of retrieval and classification is to measure the similarity between the query/probe and database/gallery, so that the common representations across different views (e.g., image, text, etc.) are learned. To the end, a variety of approaches has been proposed with different problem settings, for example, unsupervised [11]–[14]; supervised [6], [15]–[17]; and semisupervised [18], [19] methods. In this article, we mainly focus on a semisupervised setting since it could maximally exploit all available training data in practice, for example, a few cost-prohibitive labeled data and a large number of unlabeled data.

To leverage all available labeled and unlabeled multiview data, a number of semisupervised approaches have been proposed to learn discriminative representations shared by different views [18], [20], [21], and their major difference is the choice of different strategies in using unlabeled data. More specifically, these methods usually take one of the following two choices: 1) maximizing the pairwise correlations between cross-view samples [21]–[24] or 2) preserving the intrinsic information in the common space by using a Laplacian regularizer [18], [19]. Although these methods have achieved promising performance, almost all of these methods have faced the following limitations. To be specific, the methods with the first choice implicitly require the unlabeled data subsets to satisfy the pairwise constraint, which is easily isolated in practice as shown in Fig. 1. In other words, the multiview data are probably not pairwise/aligned due to view missing, independent

Manuscript received October 13, 2020; revised March 4, 2021 and June 3, 2021; accepted June 14, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62102274, Grant 61971296, Grant U19A2078, and Grant 61836011; in part by the fellowship of China Postdoctoral Science Foundation under Grant 2021M692270; in part by the Sichuan Science and Technology Planning Project under Grant 2020YFH0186, Grant 2021YFG0317, and Grant 2021YFG0301; in part by the Fundamental Research Funds for the Central Universities under Grant YJ201949 and Grant 1082204112616; and in part by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funds under Project A18A2b0046 and Project A1892b0026. This article was recommended by Associate Editor S. Das. *(Corresponding author: Dezhong Peng.)*

Peng Hu is with the College of Computer Science, Sichuan University, Chengdu 610065, China, and also with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore.

Xi Peng is with the College of Computer Science, Sichuan University, Chengdu 610065, China.

Hongyuan Zhu and Jie Lin are with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore.

Liangli Zhen is with the Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore.

Huaibai Yan is with Chengdu Sefon Software Company, Ltd., Chengdu 610041, China.

Dezhong Peng is with the College of Computer Science, Sichuan University, Chengdu 610065, China, also with Peng Cheng Laboratory, Shenzhen 518052, China, and also with the College of Computer and Information Science, Southwest University, Chongqing 400715, China (e-mail: pengdz@scu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3093626.

Digital Object Identifier 10.1109/TCYB.2021.3093626

2168-2267 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

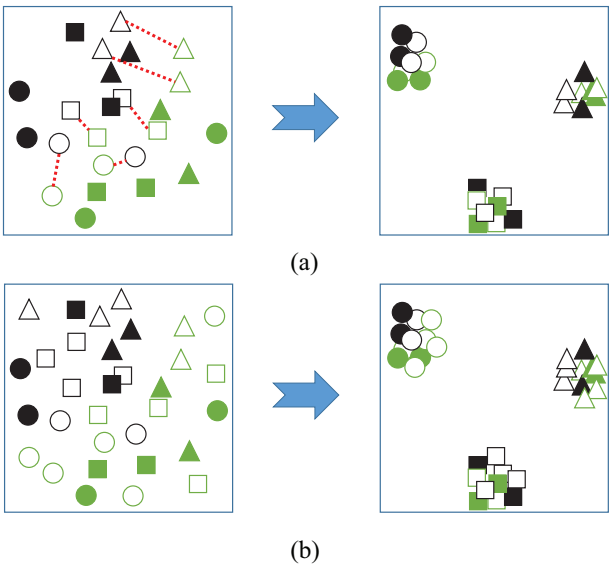


Fig. 1. Difference between (a) pairwise-constraint semisupervised methods and (b) generalized ones. In the figure, red-dotted lines represent the pairwise constraints between two views; solid items represent labeled samples; hollow items denote unlabeled points; different colors represent different views; and different shapes denote different categories. In brief, most traditional semisupervised methods require the unlabeled data to satisfy the pairwise constraint. However, this constraint is often hardly satisfied in practice due to view missing, independently sampling, and so on. In contrast, generalized semisupervised methods directly capture the intrinsic information from the unlabeled data without the pairwise constraint, thus embracing more flexibility in practice.

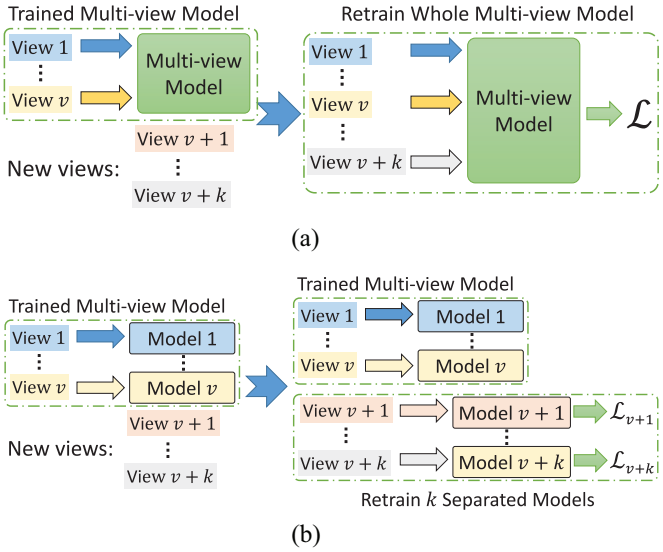


Fig. 2. Difference between (a) existing joint multiview learning and (b) our independent multiview learning. In brief, the traditional methods use all views to learn the common space. They are difficult to handle increasing views since their models are optimized depending on all views. Thus, they should retrain the whole model to handle new views, which is inefficient with abandoning the trained model. In contrast, our method independently trains the k view-specific models for the k new views, thus efficiently handling increasing views.

sampling, spatial–temporal connection breaking, and so on. Once there is no available pairwise relationship between unlabeled cross-view data, the methods would fail to achieve a desirable result. Although the second kind of method can avoid the pairwise correspondence by directly enforcing a Laplacian regularizer on the unlabeled data, they have suffered from very high computational complexity as the graph Laplacian is time- and memory-prohibitive. Furthermore, almost all existing methods, including those mentioned above, cannot handle the increasing views because they often jointly learn the view-specific transformations on all views to narrow the view gap. Such a joint learning paradigm requires retraining the entire model once some new views are coming, as shown in Fig. 2(a). In the real world, the view number probably increases by deploying new sensors or applications.

To address the aforementioned problems, we propose a novel multiview learning method, called the independent semisupervised view-specific network (ISVN). The proposed method consists of multiple ISVNs, which aim to learn representation for different views. All ISVNs work in an independent manner so that the newly observed views could be handled without retraining the trained model. For one new view, we only need to stack and train a new ISVN for it. The independent working mechanism of ISVN is derived from our view decoupling strategy, which leverages the unified semantic information shared diverse views to alleviate the cross-view discrepancy instead of the cross-view relationship. Specifically, for the labeled samples $\mathcal{X}^i$, the ith ISVN aims at projecting $\mathcal{X}^i$ into the common discriminant space that is predefined by a fixed shared linear classifier $\mathbf{W}$ as shown in Fig. 3. Thanks to the unified semantic space and the shared classifier, each ISVN could independently project the corresponding view to the common space by only utilizing the labeled data of its own view. Alternatively, for the unlabeled samples $\mathcal{U}^i$, the intrinsic and discriminative information is excavated from $\mathcal{U}^i$ to improve the ith ISVN with the reconstruction and pseudolabel (PL) learning as shown in Fig. 3. Thus, each ISVN could be independently trained on only its corresponding labeled and unlabeled data without any interview constraints.

Different from existing semisupervised methods, our ISVN utilizes a reconstruction criterion and a PL strategy to exploit the intrinsic and discriminative information from the unpaired unlabeled data. Thanks to the reconstruction criterion, our ISVN maximizes the mutual information between the learned representations and the view-specific inputs to smoothly capture the data manifold instead of the memory-intensive graphic regularizer. By learning from the PL strategy, the fixed classifier can be more confident in its classification on the unlabeled data. Thus, the discrimination between distinct classes can be learned from the labeled and unlabeled data. Another difference from the existing joint learning methods is that our ISVN can separately train each view-specific network on its corresponding view without sharing any constraints and trainable cross-view parameters as shown in Fig. 2(b). Due to the separate training strategy, the proposed method can efficiently and flexibly tackle new views and a large number of views.

The differences with existing semisupervised multiview learning approaches are given as follows. First, different from the aforementioned first kind of methods [21]–[24], our method could use the unlabeled data, while avoiding establishing the pairwise correspondence of views as shown in Fig. 1. Second, different from the second kind of methods [18], [19],

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

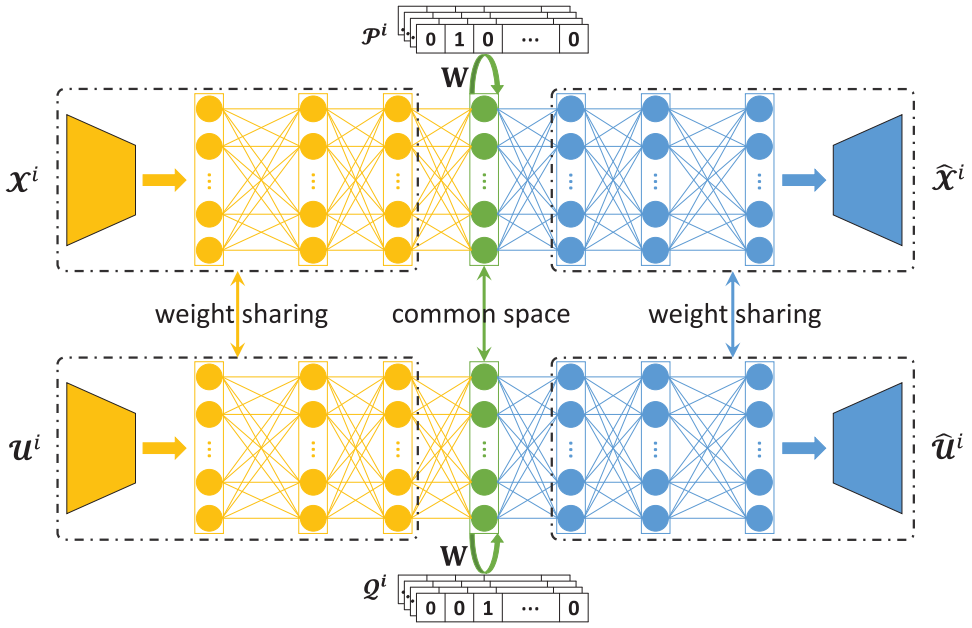


Fig. 3. Pipeline of our ISVN for the ith view. For v views, v ISVNs are constructed to separately learn the common representations on each corresponding view. In the figure, $\mathcal{X}^i = \{\mathbf{x}_k^i\}_{k=1}^{N_i}$ represents the set of N_i labeled samples from the ith view and $\mathcal{P}^i = \{\mathbf{p}_k^i\}_{k=1}^{N_i}$ is the set of the corresponding labels. $\mathcal{U}^i = \{\mathbf{u}_k^i\}_{k=1}^{M_i}$ denotes the set of M_i unlabeled samples in the ith view. $\mathcal{Q}^i = \{\mathbf{q}_k^i\}_{k=1}^{M_i}$ is the set of the PLs predicted by the network on $\mathcal{U}^i$. $\mathbf{W}$ is the associated matrix with the fixed unified linear classifier used to define the common space. $\hat{\mathcal{X}}^i$ and $\hat{\mathcal{U}}^i$ are the reconstructed samples from $\mathcal{X}^i$ and $\mathcal{U}^i$, respectively. Since distinct views share the unified semantic information, the labeled data $\mathcal{X}^i$ could be used to bridge the different views to learn common representations without the cross-view pairwise relationship. The unlabeled data $\mathcal{U}^i$ are exploited to extract the intrinsic and discriminative information from $\mathcal{U}^i$ into the common space through reconstruction (i.e., $\hat{\mathcal{X}}^i$ and $\hat{\mathcal{U}}^i$) and PL (i.e., $\mathcal{Q}^i$) learning. Thus, all views could be separately projected into the common space without any interview constraints, and could easily and efficiently handle new views.

our method employs a specifically designed autoencoder (AE) rather than graph Laplacian to exploit the unlabeled data, thus enjoying higher computational efficiency. Third, our method could efficiently handle the newly coming views without retraining the entire model as shown in Fig. 2. To the best of our knowledge, fewer efforts have been devoted to developing a semisupervised multiview neural network that could handle the increasing views. The main contributions of this article could be summarized as follows.

1) A novel ISVN is proposed to decouple the joint multiview semisupervised learning, enabling toseparately train each view-specific network. Thus, our ISVN could be efficient and flexible to handle the data of increasing views.
2) A novel semi-supervised training strategy is proposed to efficiently exploit the unlabeled unpair multiview data to encapsulate the inhered data information in the common representations. Thus, our method could be efficient and effective in tackling unlabeled multiview data without pairwise constraints.

II. Related Work

Multiview learning has been the source of more attention from academic and industry communities, such as multiview clustering [1], [25], [26], [27]; cross-modal retrieval [28], [29]; etc. Multiview clustering aims at exploring clusterings to organize the multiview data into meaningful groups [1], [25], [26], [30]. Specifically, the individuality and commonality of multiview data are creatively utilized to generate high-quality and diverse clusterings in [25]. In this work, we mainly focus on multiview learning for cross-view retrieval and classification, which could be roughly classified into unsupervised, supervised, and semisupervised categories based on the availability of category label information. In this section, we will briefly introduce some related multiview learning methods according to the three aspects.

A. Unsupervised Multiview Learning

Unsupervised multiview learning aims at projecting multiview data into a common space wherein all views are statistically correlated. One pioneer is canonical correlation analysis (CCA), which maximizes the cross-view correlations to learn two linear transformations so that different views are projected into a latent common space. Similarly, Sharma and Jacobs [12] proposed a partial least squares method (PLSs), which linearly projects two views into a latent common space by maximizing their covariance. One major limitation of CCA and PLS might be that they can only handle the biview data, and it is costly to extend them to the multiview case. To overcome this limitation, some works were proposed, which often maximize the sum of correlations between all pairwise views. For example, multiview CCA (MCCA) [11], [31] proposes learning v view-specific linear transformations for v views. To excavate the intrinsic structure in the relative low-dimensional space, Ma *et al.* [32] proposed an unsupervised non-negative matrix factorization-based method by introducing manifold regularizations into the multiview discriminant analysis. To tackle missing samples in the training set, Lampert and Krömer [33] presented a dimensionality reduction method that

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

is able to work with weakly paired data and also robust to partially missing data. Zhu *et al.* [34] proposed a novel cross-modal hashing approach to enable scalable training data with a linear time complexity to the training dataset. In [35], a multimodal graph regularized smooth matrix factorization hashing (MSFH) approach is proposed to counter quantization loss caused by relaxing hash codes for unsupervised cross-modal retrieval. Moreover, some variants of CCA are proposed with kernel trick to make handling linear inseparable data possible, such as Kernel CCA (KCCA) [36] and kernel nonlinear orthogonal iterations (KNOIs) [37]. However, it lacks a golden criterion to choose a suitable kernel function, and the performance of these approaches heavily depends on the used kernel [38]. To tackle the problem, some deep multiview methods have been proposed [39]–[41].

B. Supervised Multiview Learning

Different from the unsupervised setting, supervised methods assume that all training data are well annotated. With the semantic information rooted in the class label, a variety of methods has been proposed to learn a latent common discriminant space wherein the samples from the same class are compacted; meanwhile, the ones from different classes are enforced to be scattered [6], [42]. Moreover, Ma *et al.* [43] and Sun *et al.* [44] proposed incorporating the label into CCA to learn a single discriminant space by minimizing the between-class correlation and maximizing the within-class correlation across different views. Hou *et al.* [45], [46] studied how to prevent the new views from worsening performance and presented a stable method to guarantee that the performance does not become worse with more views. To efficiently explore the complementary properties from multiple different feature domains, Zhang *et al.* [47] presented a simultaneous spectral-spatial feature selection and extraction algorithm to project the spectral-spatial feature into a common feature space. Mandal *et al.* [48] proposed a simple hashing framework that is able to handle different scenarios wherein multimodal data may be associated with a single label or multilabel with or without pairwise correspondence. Yu *et al.* [28] proposed a flexible cross-modal hashing approach (FlexCMH) to learn effective binary codes from weakly paired multimodal data, which is a challenge scenario wherein correspondence across different modalities is partially (or even completely) missed. To capture high nonlinearity across different views, some recent works proposed employing a deep neural network to project different views into a latent common discriminant space with the semantic information [49], [50]. In very recent, motivated by the success of generative adversarial nets [51] in modeling data distribution, adversarial learning was introduced to model the joint distribution over all views with the semantic information to learn common discriminative representations [29], [42].

C. Semisupervised Multiview Learning

Although the supervised methods have achieved promising performance, their performance fully relies on sufficient labeled multiview data. It is time and cost-prohibitive to obtain an amount of well-annotated multiview data, and even impossible for some applications that need careful expert labeling, e.g., medicine. Thus, it is highly encouraged to design semisupervised multiview methods to exploit the discrimination from a few labeled multiview data and a large number of unlabeled data. To the end, a number of algorithms have been proposed, and one most popular solution is combining the supervised objective and a Laplacian regularizer [18]–[20]. For instance, Chen *et al.* [20] proposed a dimensionality reduction method, which performs CCA on a few paired data and utilizes both the local discriminative information of labeled data and the global structural information of unlabeled data to compensate for the limited pairedness. Moreover, to address the incompleteness issue of pairwise correspondence and category labels in multiview, multiinstance, and multilabel learning (M3L), Xing *et al.* introduced a weakly supervised M3L approach (WSM3L) based on multimodal dictionary learning in [52]. By considering the nonlinearity of data, some deep semisupervised methods are recently proposed, which project different views into a common space in a progressive way [21], [53]. Although these methods have achieved promising performance in their specific settings, the view-dependent training strategy will hinder their efficiency and flexibility from handling increasing views.

III. Proposed Algorithm

In this section, we introduce our method, which consists of v independent view-specific ISVNs for v views. The ith network pipeline for the ith view is shown in Fig. 3. Since all the view-specific ISVNs are independent of each other, thanks to our view-decoupling approach, we will introduce one view-specific ISVN in the following example without loss of generality.

A. Problem Formulation

Let $\mathcal{X}^i = \{\mathbf{x}_k^i\}_{k=1}^{N_i}$ be the set of N_i labeled samples from the ith view with the corresponding one-hot labels $\mathcal{P}^i = \{\mathbf{p}_k^i\}_{k=1}^{N_i}$ sampled from c classes, and let $\mathcal{U}^i = \{\mathbf{u}_k^i\}_{k=1}^{M_i}$ be a set of M_i unlabeled samples from the ith view. Although our method belongs to the semisupervised family, we consider one more challenging situation, that is, only an extremely small portion of data is labeled and the unlabeled data does not satisfy the pairwise constraint. Note that most existing works [22]–[24] implicitly assumed the unlabeled data are with well-established correspondence. In contrast, we assume that one does not know whether two cross-view points belong to the same subject or not. Clearly, such a setting is more difficult and practical in the real world.

The key to solving the above challenge is to decouple the joint cross-view learning paradigm. More specifically, these existing methods often learn a common space wherein the cross-view data points of the same subject are with the similar even the same representation. To learn the common space, the correspondence of points is necessary so that the cross-view data points of the same subject are known and the joint optimization could work. Once we do not adopt the joint cross-view learning paradigm, it is unnecessary to establish the view

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

correspondence on the unlabeled data. As another benefit of such a view-decoupling paradigm, one could handle the newly coming view.

To achieve the aforementioned goal, different from the joint multiview learning methods, the objective functions and networks of all views should be individual with each other. Thus, the overall objective function of our ISVN for the ith view could be formulated as follows:

$$\begin{aligned}\left\{\Theta_i^*, \Phi_i^*\right\} &= \arg\min_{\Theta_i, \Phi_i} \mathcal{L}\left(\mathcal{X}^i, \mathcal{P}^i, \mathcal{U}^i, \mathcal{Q}^i\right) \\ &= \arg\min_{\Theta_i, \Phi_i} \left(\mathcal{L}\left(\mathcal{X}^i, \mathcal{P}^i\right) + \beta \mathcal{L}\left(\mathcal{U}^i, \mathcal{Q}^i\right)\right) \end{aligned} \tag{1}$$

where $\beta > 0$ balances the contributions of the labeled and unlabeled data, Θ_i and Φ_i are parameter sets of the encoder and decoder, $\mathcal{Q}^i$ is the PLs obtained by Eq. (4), and $\mathcal{L}$ is a unified loss function for both labeled and unlabeled data, which are elaborated in the following sections.

B. Independent Semisupervised View-Specific Network

To achieve the aforementioned goal, we employ an untrained orthogonal matrix $\mathbf{W}$ to replace the popular learning-based common space. Note that if $\mathbf{W}$ is orthogonal, the within-class similarity will be maximized and the between-class similarity will be minimized according to [54] and [55]. To be specific, for a given labeled/unlabeled data point, we feed it through the encoder to obtain the feature $\mathbf{y}_k^i$, that is

$$\mathbf{y}_k^i = f\left(\mathbf{s}_k^i\right), \ \ \mathbf{s} \in \{\mathbf{x}, \mathbf{u}\} \tag{2}$$

where $f(\cdot)$ denotes the encoder.

After that, we project $\mathbf{y}_k^i$ into the label space via $\mathbf{W}^T\mathbf{y}_k^i$. Here, one could observe that $\mathbf{W}$ performs like a classifier with an untrained orthogonal matrix. We fix it due to the following reason. To be specific, if $\mathbf{W}$ is trainable, we will still face the view-coupling problem, that is, it is necessary to train $\mathbf{W}$ using all data, thus disabling handling the increasing view problem.

After projecting $\mathbf{y}_k^i$ into the label space, we aim at minimizing

$$\begin{aligned}\mathcal{L}_c\left(\mathcal{S}^i, \mathcal{O}^i\right) &= \frac{1}{N_i}\sum_{k=1}^{N_i} \eta \left\|\mathbf{W}^T\mathbf{y}_k^i - \mathbf{o}_k^i\right\|_2 \\ &= \frac{1}{N_i}\sum_{k=1}^{N_i} \eta \left\|\mathbf{W}^T f\left(\mathbf{x}_k^i\right) - \mathbf{o}_k^i\right\|_2 \end{aligned} \tag{3}$$

where $\mathbf{p}_k^i$ denotes the label of the kth sample of the ith view, $\eta = 1/\sum_{l=1}^{c} p_{kl}^i$ is a penalty parameter, $\|\cdot\|_2$ is the ℓ_2-norm operator, $\mathcal{S} \in \{\mathcal{X}, \mathcal{U}\}$, $\mathcal{O} \in \{\mathcal{P}, \mathcal{Q}\}$, $\mathbf{o} \in \{\mathbf{p}, \mathbf{q}\}$, $\mathcal{Q}^i = \{\mathbf{q}_k^i\}_{k=1}^{M_i}$, and $\mathbf{q}_k^i$ is the PL of $\mathbf{u}_k^i$, which will be elaborated in the following. Here, η is used to weaken the indeterminate predicted PLs for unlabeled data. In short, when data are labeled, $\eta = 1$. Otherwise, η will be with real value. For the unlabeled data point $\mathbf{u}_k^i$, we assign a PL to it via $\mathbf{q}_k^i = h(\mathbf{W}^T f(\mathbf{u}_k^i))$, $h(\cdot)$ is a sharpening function as follows:

$$h(\mathbf{z})_i = \begin{cases} 1, & \frac{z_i}{\max(\mathbf{z})} > \gamma \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Algorithm 1 Optimization Procedure of ISVN for the ith View

Input: The labeled training samples $\mathcal{X}^i = \left\{\mathbf{x}_k^i\right\}_{k=1}^{N_i}$ and the unlabeled points $\mathcal{U}^i = \left\{\mathbf{u}_k^i\right\}_{k=1}^{M_i}$ from the i-th view, objective dimensionality d, batch size N_b, positive balance parameters α and β, and learning rate τ.

1: **while** not converge **do**
2: Randomly sample N_b points from $\mathcal{X}^i$ and $\mathcal{U}^i$ to construct a labeled min-batch $\mathcal{X}_b$ and $\mathcal{P}_b$, and an unlabeled min-batch $\mathcal{U}_b$ of the i-th view, respectively.
3: Compute the pseudo-labels $\mathcal{Q}_b$ for $\mathcal{U}_b$ by Eq. (4).
4: Calculate the loss $\mathcal{L}(\mathcal{X}_b, \mathcal{P}_b, \mathcal{U}_b, \mathcal{Q}_b)$ for $\mathcal{X}_b$ and $\mathcal{U}_b$ by Eq. (1).
5: Update the network parameters Θ_i and Φ_i by minimizing the obtained loss with descending their stochastic gradient:
$\Theta_i = \Theta_i - \tau \frac{\partial \mathcal{L}(\mathcal{X}^i, \mathcal{P}^i, \mathcal{U}^i, \mathcal{Q}^i)}{\partial \Theta_i}$
$\Phi_i = \Phi_i - \tau \frac{\partial \mathcal{L}(\mathcal{X}^i, \mathcal{P}^i, \mathcal{U}^i, \mathcal{Q}^i)}{\partial \Phi_i}$
6: **end while**

Output: The optimized ISVN model for the i-th view.

where $0 < \gamma < 1$ is a positive threshold, and $\max(\mathbf{z})$ returns the maximum element of the vector $\mathbf{z}$.

Besides the discrimination loss defined in the above formulation, our loss function also consists of a reconstruction term as follows:

$$\begin{aligned}\mathcal{L}_r\left(\mathcal{S}^i\right) &= \frac{1}{N_i}\sum_{k=1}^{N_i} \left\|\hat{\mathbf{s}}_k^i - \mathbf{s}_k^i\right\|_2 \\ &= \frac{1}{N_i}\sum_{k=1}^{N_i} \left\|g\left(f\left(\mathbf{s}_k^i\right)\right) - \mathbf{s}_k^i\right\|_2 \end{aligned} \tag{5}$$

where $\mathbf{s} \in \{\mathbf{x}, \mathbf{u}\}$, $\hat{\mathbf{s}}_k^i$ is the reconstruction of the given data point $\mathbf{s}_k^i$, namely, the output of decoder $g(\cdot)$.

Putting the discriminative loss $\mathcal{L}_c$ and view-specific reconstruction loss $\mathcal{L}_r$ together, we have

$$\mathcal{L}\left(\mathcal{S}^i, \mathcal{O}^i\right) = (1-\alpha)\mathcal{L}_c\left(\mathcal{S}^i, \mathcal{O}^i\right) + \alpha \mathcal{L}_r\left(\mathcal{S}^i\right) \tag{6}$$

where α is a balanced parameter for classification and reconstruction losses. Therefore, we obtain the unified loss function for both labeled and unlabeled data, which can be applied on (1) to train the models.

In summary, each view-specific network could be separately trained by optimizing the objective function (1). The optimization process for the ith view can be summarized as Algorithm 1.

IV. Experiments

To evaluate the effectiveness of the proposed method for cross-view retrieval and recognition, we conduct experiments on four benchmark multiview datasets, that is: 1) XMediaNet [56]; 2) NUS-WIDE [57]; 3) INRIA-Websearch [58]; and 4) MNIST-SVHN [59], [60]. We not only compare our ISVN with 11 state-of-the-art methods but also conduct ablation study to investigate the

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
GENERAL STATISTICS OF THE FOUR DATASETS USED IN THE EXPERIMENTS, WHERE "*/*/*" IN THE "#SAMPLE" COLUMN STANDS FOR THE NUMBER OF TRAINING/VALIDATION/TEST SUBSETS

Dataset	#Class	View	#Sample	Data
XMediaNet	200	Image	32,000/4,000/4,000	4,096D VGG
		Text	32,000/4,000/4,000	3,00D Doc2Vec
NUS-WIDE	10	Image	42,941/5,000/23,661	4,096D VGG
		Text	42,941/5,000/23,661	3,00D Doc2Vec
INRIA-Websearch	100	Image	9,000/1,332/4,366	4,096D AlexNet
		Text	9,000/1,332/4,366	1,000D LDA
MNIST-SVHN	10	MNIST	50,000/10,000/10,000	28×28D pixels
		SVHN	63,257/10,000/26,032	$32 \times 32 \times 3$D pixels

contribution of each component of ISVN. Besides, we also carry out experiments of parameter analysis and the efficacy analysis for new views.

A. Experiment Settings

We compare our method with 13 state-of-the-art methods, including: 1) MCCA [11]; 2) PLS [12]; 3) DCCA [39]; 4) DCCAE [14]; 5) GMLDA [7]; 6) MvDA [55]; 7) MvDA-VC [6]; 8) ACMR [42]; 9) FGCrossNet [61]; 10) JRL [19]; 11) GSS-SL [18]; 12) DSCMR [62]; and 13) SMLN [21]. Among these methods, MCCA, PLS, GMLDA, MvDA, and MvDA-VC are shallow models, of which the optimal dimensionality of the common space is searched into the range of [10:250] using the validation subset of each data collection. For the evaluated deep models, we adopt the values recommended by the corresponding authors.

For a fair comparison, all methods adopt the same features as shown in Table I. In other words, the parameters of the feature extractors (e.g., VGGNet [63], AlexNet [64], Doc2vec [65], etc.) are fixed during training even though our ISVN can be trained in an end-to-end manner. Moreover, we employ v extra four-layer neural networks to extract features from XMediaNet, NUS-WIDE, and INRIA-Websearch, that is, $d_i \rightarrow 4096 \rightarrow 4096 \rightarrow d$ for the ith view, where d_i is the input dimensionality of the ith view and d is the objective dimensionality of the common space. For the MNIST-SVHN dataset, the raw data are used as inputs to train each ISVN in an end-to-end manner. The CNN network architecture of our ISVN is implemented according to MNIST-SVHN CycleGAN transfer.[1] For the other methods, an image of this dataset is reshaped as an input vector. In the inference process, the outputs of each ISVN are the common representations of the corresponding input samples. N_b and d are set to 16 and 1024 for all datasets, respectively. The learning rate γ is set 0.0001 for all views.

The ADAM [66] optimizer is employed to optimize our each ISVN with the maximal epoch of 200. The optimized ISVN model of the last epoch (i.e., the 200th epoch) is used as the inference model to compute the common representation for a testing sample of the corresponding view. Note that for supervised and unsupervised multiview methods, only the labeled samples can be used to train their models since any two views of unlabeled multiview data are unpaired. On the other hand, the semisupervised methods without pairwise restriction could fully employ both labeled and unlabeled multiview data, that is, JRL, GSS-SL, and our ISVN.

[1] https://github.com/yunjey/mnist-svhn-transfer

B. Evaluation Metric

We evaluate the effectiveness of our method in two tasks, that is: 1) cross-view retrieval and 2) classification. Specifically, we adopt the mean average precision (mAP) to evaluate the cross-view retrieval performance on the XMediaNet, NUS-WIDE, and INRIA-Websearch datasets. Like [8] and [42], we calculate the mAP scores on the ranked lists of the retrieved results for two distinct tasks to evaluate the performance, namely: 1) retrieving relevant text instances using an image query (Img2Txt) and 2) retrieving relevant image samples using a text query (Txt2Img). Noticed, the mAP is a widely used performance evaluation criterion for cross-view retrieval [42], which is the mean value of average precision (AP) scores for each query. For the ith query, the corresponding AP is calculated as follows:

$$\bar{P}_i = \frac{1}{R(n)} \sum_{k=1}^{n} \frac{R(k)}{k} \times P(k) \tag{7}$$

where n is the number of retrieving samples, and $R(k)$ counts the number of the relevant instances in the top k returned results. $P(k)$ is a Boolean function, which is equal to 1 if the returned result of the rank k is a relevant point, and zero otherwise. With AP, the mAP score could be computed as follows:

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} \bar{P}_i. \tag{8}$$

From the above formulation, the mAP score simultaneously considers the precision as well as the rank of the returned retrieval results. It should be noted that the mAP scores of all experiments are computed on the all returned results following [42]. Besides mAP, the precision–recall curves are illustrated to visually investigate the performance of our ISVN and its counterparts.

In addition, the top-1 classification accuracy is reported to evaluate the performance of cross-view classification on the MNIST-SVHN dataset. The top-k classification accuracy can be calculated through

$$\mathrm{ACC}(k) = \frac{1}{n} \sum_{l=1}^{n} Q(l, k) \tag{9}$$

where n is the number of probes. $Q(l, k)$ is an indicator function that is equal to 1 if the top k returned instances from the gallery have the same class as the kth probe, and zero otherwise. Similar to cross-view retrieval, the top-1 classification accuracy is computed for some different cross-view pairwise tasks, that is, the samples from a view are adopted as the gallery set while the ones from another view are used as the probes. For example, M2S (S2M) denotes that the testing set of MNIST (SVHN) is used as the gallery set while the testing set of SVHN (MNIST) is used as the probe set. Note that the similarity between two points is computed through cosine distance in all experiments.

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
PERFORMANCE COMPARISON IN TERMS OF MAP SCORES ON THE XMEDIANET DATASET. THE HIGHEST SCORE IS SHOWN IN *Boldface*

Method	400 labels			1000 labels			2000 labels			3000 labels			4000 labels		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
MCCA* [11]	0.030	0.035	0.032	0.057	0.063	0.060	0.114	0.124	0.119	0.148	0.162	0.155	0.174	0.189	0.181
PLS* [12]	0.043	0.048	0.046	0.068	0.070	0.069	0.099	0.093	0.096	0.113	0.103	0.108	0.125	0.115	0.120
DCCA* [39]	0.011	0.011	0.011	0.017	0.019	0.018	0.027	0.030	0.028	0.037	0.040	0.039	0.044	0.048	0.046
DCCAE* [14]	0.012	0.013	0.012	0.017	0.019	0.018	0.025	0.028	0.027	0.031	0.034	0.033	0.037	0.041	0.039
GMLDA† [7]	0.016	0.021	0.018	0.095	0.094	0.095	0.158	0.153	0.155	0.195	0.187	0.191	0.218	0.207	0.213
MvDA† [6]	0.029	0.036	0.033	0.099	0.108	0.104	0.170	0.179	0.175	0.213	0.221	0.217	0.241	0.247	0.244
MvDA-VC† [6]	0.049	0.059	0.054	0.118	0.126	0.122	0.191	0.197	0.194	0.229	0.232	0.230	0.254	0.258	0.256
JRL† [19]	0.060	0.069	0.064	0.103	0.109	0.106	0.206	0.196	0.201	0.239	0.223	0.231	0.253	0.237	0.245
GSS-SL† [18]	0.030	0.035	0.032	0.130	0.141	0.135	0.220	0.229	0.225	0.271	0.276	0.273	0.304	0.306	0.305
ACMR† [42]	0.022	0.025	0.023	0.040	0.042	0.041	0.059	0.064	0.061	0.065	0.084	0.075	0.096	0.111	0.104
DSCMR† [62]	0.118	0.157	0.138	0.112	0.156	0.134	0.256	0.284	0.270	0.194	0.227	0.211	0.255	0.283	0.269
FGCrossNet† [61]	0.043	0.057	0.050	0.084	0.108	0.096	0.157	0.190	0.174	0.212	0.249	0.230	0.262	0.289	0.275
SMLN† [21]	0.053	0.062	0.058	0.144	0.112	0.128	0.257	0.196	0.226	0.304	0.260	0.282	0.341	0.300	0.321
JRL‡ [19]	0.065	0.072	0.069	0.144	0.147	0.146	0.217	0.208	0.213	0.256	0.241	0.249	0.282	0.265	0.274
GSS-SL‡ [18]	0.059	0.067	0.063	0.133	0.143	0.138	0.215	0.220	0.217	0.256	0.261	0.258	0.285	0.286	0.286
ISVN	**0.157**	**0.189**	**0.173**	**0.321**	**0.357**	**0.339**	**0.418**	**0.458**	**0.438**	**0.464**	**0.497**	**0.481**	**0.500**	**0.532**	**0.516**

$*$, † and ‡ indicate that the corresponding methods run in unsupervised, supervised and semi-supervised settings, respectively.

C. Datasets

In this section, we elaborate the tested datasets, that is, XMediaNet [56], [67], NUS-WIDE [57], INRIA-Websearch [58], and MNIST-SVHN [59], [60]. The general statistics and basic information of the datasets are summarized in Table I and some details are given as follows.

1) *XMediaNet [56], [67]* is a large-scale multiview dataset, which includes 40 000 images, 40 000 texts, 10 000 audio, 2000 3-D models, and 10 000 videos, of which each sample is classified into 200 nonoverlap categories. In this article, the experiments are conducted on the image and text data of the dataset, where the images are collected from Flickr and the text sentences are selected from the Wikipedia articles. We evenly split the dataset to three sets following [56], [68] as shown in Table I.
2) *NUS-WIDE [57]* consists of about 270 000 images distributed over 81 categories. In this dataset, one sample may belong to multiple classes. In the experiments, we only choose the samples from the ten categories with the largest quantity, and each sample belongs to a single class by following [69]. Besides, the obtained samples are split into three subsets as shown in Table I.
3) *INRIA-Websearch [58]* consists of 71 478 images and 71 478 text descriptions (sentences or tags). All the samples of this dataset are from 353 classes. In our experiments, we use the subset of INRIA-Websearch provided by Wei *et al.* [70]. In the dataset, 14 698 samples of 100 largest classes are used by removing the irrelevant image–text pairs. We adopt the training and testing data partitions used in [70]. We also use the provided features, that is, 4096-dimensional CNN visual feature for image and 1006-dimensional LDA feature for text. Furthermore, we further split the training set as a new training set and a validation set as shown in Table I.
4) *MNIST-SVHN [59], [60]* is an union of MNIST [59] and SVHN [60], which is used to evaluate the performance of the proposed method for cross-view classification. In the experiments, we combine the two datasets as a multiview dataset with a training set (including 60 000 MNIST training images and 73 257 SVHN training images) and a testing set (including 10 000 MNIST testing images and 26 032 SVHN testing images). We further randomly split the training set into two subsets as shown in Table I.

D. Comparisons With the State of the Art

In this section, we evaluate the effectiveness of our ISVN by comparing with 11 cross-view methods on four widely used benchmark datasets. To comprehensively investigate the semisupervised performance of our method, we conduct five different semisupervised settings on each dataset, that is, learning from 400, 1000, 2000, 3000, and 4000 labeled data for each view.

1) Cross-View Retrieval: We applied the XMediaNet, NUS-WIDE, and INRIA-Websearch datasets for cross-view retrieval. The experimental results on these datasets are shown in Tables II–IV, respectively. As shown in these tables, our ISVN achieves the best performance comparing with 13 state-of-the-art methods on the three datasets. From Tables II–IV, we could draw the following observations.

1) The performance of unsupervised, supervised, and semisupervised methods could be improved by increasing the data amount. It indicates that more data can be used to improve the performance in the training stage.
2) The existing graph-based semisupervised methods could improve their performance using the unlabeled data in some cases, such as the results on XMediaNet. However, in most cases, the Laplacian regularizer cannot work well to exploit the intrinsic information from the labeled and unlabeled data. Another disadvantage of these graph-based methods is the heavy cost of constructing the Laplacian matrix. In contrast, our method can capture the intrinsic information in the data in a batch-by-batch manner, and thus easily handling large-scale datasets.

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
PERFORMANCE COMPARISON IN TERMS OF MAP SCORES ON THE NUS-WIDE DATASET. THE HIGHEST SCORE IS SHOWN IN *Boldface*

Method	400 labels			1000 labels			2000 labels			3000 labels			4000 labels		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
MCCA* [11]	0.410	0.407	0.409	0.386	0.381	0.383	0.337	0.331	0.334	0.289	0.279	0.284	0.229	0.223	0.226
PLS* [12]	0.477	0.439	0.458	0.491	0.458	0.475	0.526	0.511	0.518	0.540	0.528	0.534	0.560	0.561	0.560
DCCA* [39]	0.347	0.353	0.350	0.356	0.366	0.361	0.403	0.413	0.408	0.403	0.408	0.406	0.423	0.432	0.428
DCCAE* [14]	0.349	0.355	0.352	0.391	0.397	0.394	0.423	0.427	0.425	0.449	0.453	0.451	0.458	0.466	0.462
GMLDA† [7]	0.450	0.430	0.440	0.516	0.483	0.499	0.535	0.504	0.519	0.547	0.516	0.532	0.552	0.523	0.537
MvDA† [6]	0.456	0.475	0.465	0.541	0.553	0.547	0.601	0.608	0.604	0.604	0.616	0.610	0.600	0.612	0.606
MvDA-VC† [6]	0.510	0.537	0.523	0.573	0.588	0.581	0.596	0.609	0.603	0.597	0.614	0.606	0.603	0.619	0.611
JRL† [19]	0.553	0.556	0.554	0.584	0.592	0.588	0.601	0.611	0.606	0.608	0.619	0.614	0.614	0.625	0.619
GSS-SL† [18]	0.583	0.604	0.593	0.610	0.629	0.619	0.622	0.639	0.630	0.626	0.644	0.635	0.629	0.647	0.638
ACMR† [42]	0.545	0.579	0.562	0.596	0.600	0.598	0.598	0.617	0.607	0.629	0.623	0.626	0.622	0.623	0.623
DSCMR† [62]	0.565	0.611	0.588	0.605	0.628	0.616	0.632	0.634	0.633	0.643	0.639	0.634	0.642	0.655	0.649
FGCrossNet† [61]	0.577	0.564	0.571	0.609	0.607	0.608	0.624	0.613	0.619	0.631	0.634	0.633	0.637	0.644	0.641
SMLN† [21]	0.580	0.533	0.557	0.604	0.592	0.598	0.605	0.582	0.593	0.612	0.601	0.607	0.625	0.619	0.622
JRL‡ [19]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
GSS-SL‡ [18]	0.483	0.518	0.500	0.531	0.576	0.553	0.556	0.599	0.577	0.570	0.610	0.590	0.579	0.616	0.598
ISVN	**0.600**	**0.625**	**0.612**	**0.629**	**0.649**	**0.639**	**0.655**	**0.663**	**0.654**	**0.649**	**0.664**	**0.657**	**0.655**	**0.672**	**0.663**

∗, † and ‡ indicate that the corresponding methods run in unsupervised, supervised and semi-supervised settings, respectively. / denotes out-of-memory error.

TABLE IV
PERFORMANCE COMPARISON IN TERMS OF MAP SCORES ON THE INRIA-WEBSEARCH DATASET. THE HIGHEST SCORE IS SHOWN IN *Boldface*

Method	400 labels			1000 labels			2000 labels			3000 labels			4000 labels		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
MCCA* [11]	0.065	0.077	0.071	0.116	0.123	0.120	0.189	0.191	0.190	0.208	0.208	0.208	0.154	0.152	0.153
PLS* [12]	0.062	0.072	0.067	0.123	0.131	0.127	0.214	0.221	0.218	0.272	0.279	0.276	0.307	0.315	0.311
DCCA* [39]	0.086	0.077	0.081	0.113	0.103	0.108	0.090	0.080	0.085	0.081	0.071	0.076	0.092	0.081	0.087
DCCAE* [14]	0.080	0.071	0.076	0.094	0.086	0.090	0.073	0.063	0.068	0.068	0.061	0.064	0.078	0.071	0.074
GMLDA† [7]	0.128	0.143	0.136	0.226	0.243	0.235	0.310	0.320	0.315	0.345	0.352	0.348	0.366	0.375	0.370
MvDA† [6]	0.131	0.152	0.141	0.236	0.248	0.242	0.323	0.330	0.327	0.357	0.365	0.361	0.379	0.389	0.384
MvDA-VC† [6]	0.116	0.141	0.128	0.223	0.240	0.232	0.318	0.327	0.322	0.354	0.364	0.359	0.377	0.387	0.382
JRL† [19]	0.181	0.223	0.202	0.313	0.359	0.336	0.367	0.381	0.374	0.448	0.468	0.458	0.472	0.496	0.484
GSS-SL† [18]	0.103	0.133	0.118	0.263	0.273	0.268	0.383	0.388	0.386	0.429	0.437	0.433	0.459	0.470	0.465
ACMR† [42]	0.158	0.156	0.157	0.235	0.238	0.237	0.307	0.309	0.308	0.329	0.336	0.332	0.353	0.362	0.357
DSCMR† [62]	0.247	0.251	0.249	0.368	0.373	0.371	0.434	0.448	0.441	0.467	0.484	0.476	0.491	0.509	0.500
FGCrossNet† [61]	0.085	0.075	0.080	0.164	0.151	0.157	0.255	0.250	0.253	0.306	0.307	0.306	0.362	0.364	0.363
SMLN† [21]	0.200	0.223	0.212	0.348	0.353	0.350	0.431	0.440	0.435	0.463	0.465	0.464	0.481	0.489	0.485
JRL‡ [19]	0.132	0.172	0.152	0.311	0.326	0.318	0.401	0.419	0.410	0.438	0.458	0.448	0.465	0.489	0.477
GSS-SL‡ [18]	0.129	0.147	0.138	0.252	0.270	0.261	0.337	0.357	0.347	0.383	0.403	0.393	0.413	0.439	0.426
ISVN	**0.294**	**0.302**	**0.298**	**0.398**	**0.402**	**0.400**	**0.457**	**0.465**	**0.461**	**0.483**	**0.487**	**0.485**	**0.505**	**0.511**	**0.508**

∗, † and ‡ indicate that the corresponding methods run in unsupervised, supervised and semi-supervised settings, respectively. / denotes out-of-memory error.

3) Most supervised cross-view methods are superior to the unsupervised ones, which indicates the importance of the labeled data in cross-view retrieval. This observation also can be obtained, for example, more labeled data, better performance.
4) Although DNN can capture highly nonlinear information in the dataset, and the limited available labeled data hinder the performance of both unsupervised and supervised methods, that is, DCCA, DCCAE, ACMR, and FGCrossNet. Even if the unsupervised multiview methods do not need the labels, they still need to establish the pairwise relationship between cross-view samples. In other words, they cannot utilize the unpaired unlabeled data in our settings. Thus, these data-driven deep methods cannot achieve promising performance when training data are insufficient.
5) The proposed method achieves the best performance compared with its counterparts in the cross-view retrieval tasks. One possible reason is that our ISVN can exploit the intrinsic and discriminative information from both labeled and unlabeled data with the autoencoder and PL confidence strategy.

In addition to the mAP scores, we plot the precision–recall curves to visually show the effectiveness of the proposed method and its counterparts. Fig. 4 demonstrates the precision–recall curves of all tested methods on the XMediaNet, NUS-WIDE, and INRIA-Websearch datasets, respectively. From the figures, one could see that our ISVN is superior to all the compared methods, which is consistent with the above observations about the mAP scores.

2) Cross-View Classification: The cross-view classification tasks are conducted on the MNIST-SVHN dataset. The

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
PERFORMANCE COMPARISON IN TERMS OF CROSS-VIEW TOP-1 CLASSIFICATION ON THE MNIST-SVHN DATASET. THE HIGHEST SCORE IS SHOWN IN *Boldface*

Method	400 labels			1000 labels			2000 labels			3000 labels			4000 labels		
	M2S	S2M	Avg.	M2S	M2S	Avg.	M2S	M2S	Avg.	M2S	M2S	Avg.	M2S	M2S	Avg.
MCCA* [11]	0.105	0.110	0.108	0.102	0.104	0.103	0.102	0.103	0.102	0.102	0.103	0.102	0.101	0.103	0.102
PLS* [12]	0.102	0.103	0.103	0.102	0.104	0.103	0.103	0.103	0.103	0.103	0.104	0.103	0.102	0.103	0.103
DCCA* [39]	0.104	0.119	0.112	0.103	0.124	0.114	0.102	0.124	0.113	0.106	0.122	0.114	0.103	0.117	0.110
DCCAE* [14]	0.102	0.121	0.112	0.104	0.123	0.113	0.108	0.130	0.119	0.105	0.121	0.113	0.107	0.121	0.114
GMLDA† [7]	0.103	0.115	0.109	0.104	0.138	0.121	0.104	0.165	0.134	0.102	0.174	0.138	0.104	0.155	0.129
MvDA† [6]	0.110	0.144	0.127	0.112	0.161	0.137	0.115	0.167	0.141	0.116	0.161	0.139	0.115	0.148	0.132
MvDA-VC† [6]	0.113	0.152	0.132	0.112	0.159	0.135	0.116	0.161	0.138	0.116	0.156	0.136	0.118	0.161	0.139
JRL† [19]	0.101	0.102	0.101	0.101	0.102	0.101	0.101	0.102	0.101	0.101	0.102	0.101	0.101	0.102	0.101
GSS-SL† [18]	0.114	0.165	0.139	0.117	0.177	0.147	0.121	0.183	0.152	0.123	0.184	0.153	0.124	0.186	0.155
ACMR† [42]	0.109	0.121	0.115	0.127	0.170	0.148	0.248	0.097	0.173	0.283	0.181	0.232	0.322	0.097	0.209
FGCrossNet† [61]	0.615	0.311	0.463	0.689	0.356	0.523	0.723	0.449	0.586	0.809	0.489	0.649	0.873	0.511	0.692
SMLN† [21]	0.267	0.117	0.192	0.692	0.137	0.414	0.857	0.523	0.690	0.898	0.581	0.740	0.928	0.596	0.762
JRL‡ [19]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
GSS-SL‡ [18]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
ISVN$_{FC}$	**0.675**	**0.375**	**0.525**	**0.891**	**0.507**	**0.699**	**0.900**	**0.551**	**0.725**	**0.940**	**0.606**	**0.773**	**0.946**	**0.621**	**0.783**
ISVN$_{CNN}$	**0.952**	**0.604**	**0.778**	**0.974**	**0.788**	**0.881**	**0.983**	**0.840**	**0.911**	**0.984**	**0.864**	**0.924**	**0.985**	**0.869**	**0.927**

∗, † and ‡ indicate that the corresponding methods run in unsupervised, supervised and semi-supervised settings, respectively. / denotes out-of-memory error. ISVN$_{FC}$ and ISVN$_{CNN}$ are the fully-connected and convolutional versions of our ISVN, respectively.

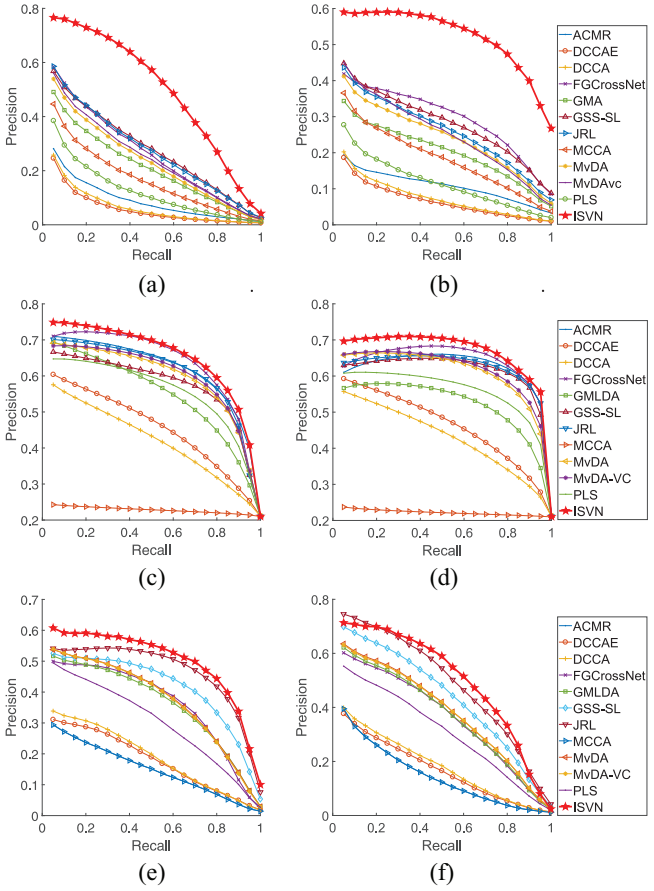


Fig. 4. Precision–recall curves on XMediaNet, NUS-WIDE, and INRIA-Websearch. (a) Img2Txt on XMediaNet. (b) Txt2Img on XMediaNet. (c) Img2Txt on NUS-WIDE. (d) Txt2Img on NUS-WIDE. (e) Img2Txt on INRIA-Websearch. (f) Txt2Img on INRIA-Websearch.

experimental results are shown in Table V. From the table, one could observe that our ISVN achieves the best performance again comparing with other multiview methods. In brief:

1) like the result on cross-view retrieval, most supervised methods outperform the unsupervised methods. Thus, semantic information is also much important for cross-view classification;
2) the graph-based semisupervised multiview methods (i.e., GSS-SL and JRL) take high memory to compute the Laplacian similarity matrices, which hinders them from handling the large-scale multiview dataset, e.g., MNIST-SVHN. In contrast, our method shows efficiency in handling large-scale datasets;
3) our method remarkably outperforms the traditional and deep multiview methods. Especially, our ISVN improves the average top-1 recognition accuracy from 0.318 to 0.778 with only 400 labeled data. Moreover, the CNN architecture (i.e., ISVN$_{\text{CNN}}$) shows better result than the fully connected network (i.e., ISVN$_{FC}$).

E. Ablation Study

To investigate the contributions of different components of our model, we carry out ablation study using the following three variants.

1) *ISVN ~~(AE+PL)~~* removes both the reconstruction objective and the unlabeled data from the proposed method, that is, ISVN with $\mathcal{L}_c(\mathcal{X}^i, \mathcal{P}^i)$ only.
2) *ISVN(~~AE~~)* is a variant of ISVN without considering the reconstruction loss $\mathcal{L}_r$, that is, ISVN without $\mathcal{L}_r$.
3) *ISVN(~~PL~~)* does not adopt the unlabeled data to predict PLs to back boost the performance, that is, ISVN without $\mathcal{L}(\mathcal{U}^i, \mathcal{Q}^i)$.

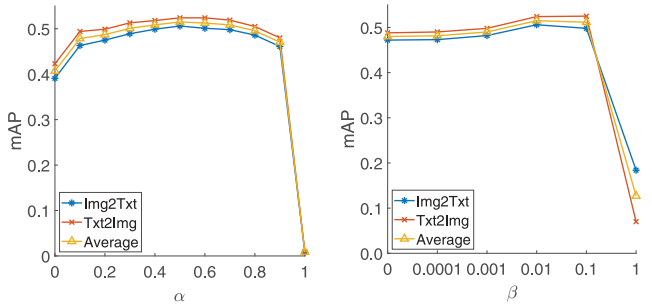
The difference between these variants and our ISVN is only the loss functions, and the other factors (e.g., network architectures) are the same with full ISVN. We compare these variants with our ISVN in terms of cross-view retrieval and classification and report the experimental results in Table VI. From the table, one could see that all the components (i.e.,

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VI
ABLATION STUDY ON DIFFERENT DATASETS. ~~X~~ DENOTES TRAINING ISVN WITHOUT X, AND X COULD BE AE AND PL. THIS TABLE SHOWS THE EXPERIMENTAL RESULTS OF CROSS-VIEW RETRIEVAL ON XMEDIANET AND NUS-WIDE, AND OF CROSS-VIEW CLASSIFICATION ON MNIST-SVHN. THE HIGHEST SCORE IS SHOWN IN *Boldface*

Dataset	Method	400 labels			1000 labels			2000 labels			3000 labels			4000 labels		
		Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
XMediaNet	ISVN (~~AE+PL~~)	0.135	0.178	0.156	0.248	0.298	0.273	0.177	0.203	0.190	0.161	0.171	0.166	0.189	0.208	0.198
	ISVN (~~AE~~)	0.154	0.176	0.165	0.242	0.307	0.275	0.184	0.208	0.196	0.211	0.228	0.219	0.359	0.391	0.375
	ISVN (~~PL~~)	0.130	0.171	0.150	0.262	0.306	0.284	0.377	0.415	0.396	0.429	0.457	0.443	0.455	0.486	0.470
	ISVN	**0.157**	**0.195**	**0.176**	**0.321**	**0.357**	**0.339**	**0.418**	**0.458**	**0.438**	**0.464**	**0.497**	**0.481**	**0.512**	**0.532**	**0.522**
		Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
NUS-WIDE	ISVN (~~AE+PL~~)	0.581	0.599	0.590	0.618	0.622	0.620	0.624	0.628	0.626	0.619	0.621	0.620	0.615	0.635	0.625
	ISVN (~~AE~~)	0.596	0.610	0.603	0.614	0.622	0.618	0.625	0.637	0.631	0.625	0.636	0.631	0.629	0.637	0.633
	ISVN (~~PL~~)	0.589	0.608	0.598	0.622	0.637	0.630	0.633	0.648	0.641	0.640	0.658	0.649	0.647	0.664	0.655
	ISVN	**0.600**	**0.625**	**0.612**	**0.629**	**0.649**	**0.639**	**0.655**	**0.663**	**0.654**	**0.649**	**0.664**	**0.657**	**0.655**	**0.672**	**0.663**
		Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
INRIA-Websearch	ISVN (~~AE+PL~~)	0.273	0.278	0.275	0.373	0.382	0.377	0.435	0.443	0.439	0.451	0.466	0.458	0.460	0.474	0.467
	ISVN (~~AE~~)	0.283	0.288	0.286	0.383	0.376	0.380	0.433	0.438	0.436	0.464	0.465	0.464	0.477	0.485	0.481
	ISVN (~~PL~~)	0.270	0.272	0.271	0.378	0.382	0.380	0.447	0.455	0.451	0.474	0.484	0.479	0.492	0.506	0.499
	ISVN	**0.294**	**0.302**	**0.298**	**0.398**	**0.402**	**0.400**	**0.457**	**0.465**	**0.461**	**0.483**	**0.487**	**0.485**	**0.498**	**0.508**	**0.503**
		M2S	S2M	Avg.	M2S	S2M	Avg.	M2S	S2M	Avg.	M2S	S2M	Avg.	M2S	S2M	Avg.
MNIST-SVHN	ISVN (~~AE+PL~~)	0.862	0.515	0.688	0.934	0.685	0.810	0.952	0.765	0.859	0.963	0.807	0.885	0.970	0.819	0.894
	ISVN (~~AE~~)	0.920	**0.612**	0.766	0.963	**0.793**	0.878	0.970	0.823	0.896	0.977	0.849	0.913	0.971	0.860	0.916
	ISVN (~~PL~~)	0.896	0.500	0.698	0.950	0.703	0.827	0.967	0.772	0.869	0.976	0.820	0.898	0.981	0.841	0.911
	ISVN	**0.952**	0.604	**0.778**	**0.974**	0.788	**0.881**	**0.983**	**0.840**	**0.911**	**0.984**	**0.864**	**0.924**	**0.985**	**0.869**	**0.927**



(a) (b)

Fig. 5. Cross-view retrieval performance of our proposed method in terms of mAP versus different values of α and β on the XMediaNet dataset, respectively. (a) mAP versus. α. (b) mAP versus. β.

reconstruction and PL) contribute to the final performance of our ISVN, indicating that all the terms should be simultaneously optimized to achieve better performance. Furthermore, the reconstruction loss is more important than the PL loss in improving the cross-view retrieval performance, whereas the PL loss is more important in cross-view classification.

F. Parameter Analysis

In this section, we investigate the impact of the parameters α and β. In order to analyze the influence of the parameters, Fig. 5(a) and (b) is drawn to show the cross-view retrieval performance versus their different values. From Fig. 5, one could observe that both classification and reconstruction losses contribute to the cross-view retrieval. Without reconstruction loss (i.e., $\alpha = 0$) or classification loss (i.e., $\alpha = 1$), the performance of cross-view retrieval will be degraded. Obviously, the classification loss is very important for our ISVN because it bridges the heterogeneous gap. Namely, the proposed method cannot work without the classification loss (i.e., $\alpha = 1$), which is consistent with the experimental results. Similarly, from Fig. 5(b), one can see that the PL classification loss [i.e., $\mathcal{L}_c(\mathcal{U}^i, \mathcal{Q}^i)$ in (4)] also contributes to the final performance. Furthermore, if the PL and real-label classification loss have equal weight, the unlabeled data will confuse the unified classifier and degrade the performance at $\beta = 1$. From both Fig. 5(a) and (b), it can be seen that our method is robust to the parameters with a reasonable range. In our experiments, the values of these parameters are determined on the corresponding validation set of the benchmark databases.

TABLE VII
EFFICIENCY COMPARISON IN TERMS OF GPU MEMORY (MIB) USAGE AND AVERAGE TRAINING TIME (S) OF 5 RUNS FOR NEW VIEWS (MNIST-M [71] AND USPS [72]) ON MNIST-SVHN

New View / Method	MNIST-M		USPS		MNIST-M & USPS	
	Memory	Time	Memory	Time	Memory	Time
Baseline	1019	700.44	1019	691.54	1159	761.21
$ISVN_s$	885	191.54	885	187.71	885	376.80
$ISVN_p$	885	191.54	885	187.71	885×2	192.59

G. Efficiency Analysis for New Views

In this section, we evaluate the efficiency of the proposed method for handling newly coming views. To investigate the advantage of the proposed independent training strategy, we developed and assessed two variants of our ISVN on two TITAN RTX GPUs. The variants are as follows.

1) *$ISVN_p$* parallelly trains all view-specific networks for the new views on different GPU devices.
2) *$ISVN_s$* serially trains each view-specific network for the corresponding view. In other words, the new views are trained in a one-by-one manner.

The baseline is the traditional joint multiview methods [21], [29], [49], [50], which have to merge the new and existing views as a new dataset to retrain the enire model. For a fair comparison, all the methods use the same settings to evaluate their efficiency, that is, the batch size is 32 and the training epoch is 20. The experimental results are shown in Table VII. From the table, one could see that our ISVN is

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VIII
PERFORMANCE COMPARISON IN TERMS OF AVERAGE CROSS-VIEW TOP-1 CLASSIFICATION ACCURACY FOR NEW VIEWS ON THE MNIST-SVHN DATASET WITH 400 LABELS. THE HIGHEST SCORE IS SHOWN IN *Boldface*

New View	MCCA* [11]	GMLDA† [7]	MvDA† [6]	MvDA-VC† [6]	SMLN† [21]	ISVN$_{FC}$	ISVN$_{CNN}$
+MNIST-M	0.108	0.120	0.158	0.167	0.390	**0.503**	**0.754**
+USPS	0.147	0.194	0.283	0.317	0.533	**0.632**	**0.827**
+MNIST-M+USPS	0.132	0.164	0.279	0.309	0.500	**0.602**	**0.816**

$*$, † and ‡ indicate that the corresponding methods run in unsupervised, supervised and semi-supervised settings, respectively. ISVN$_{FC}$ and ISVN$_{CNN}$ are the fully-connected and convolutional versions of our ISVN, respectively.

much flexible and scalable with higher efficiency to handle the new views (i.e., MNIST-M [71] and USPS [72]) than the joint learning strategy. For the low resource device, the serial version (ISVN$_s$) could remarkably reduce the GPU memory cost. Moreover, the parallel variant (ISVN$_p$) could significantly reduce the training time if there are sufficient computing resources. Furthermore, we also give the average cross-view accuracy to evaluate the performance of ISVN for new views in Table VIII. From the experimental results, we could see that our method also achieves the best performance under separately handling new views.

V. CONCLUSION

In this article, we proposed a semisupervised multiview approach that employs multiple ISVNs to learn common representations for multiple views. One major advantage of our method is the capacity to handle increasing views, thanks to your view-decoupling paradigm. In short, we employ an untrained matrix to project the latent view-specific features into the label space so that our method does not need all views to learn the common space. With such a view-decoupling paradigm, our ISVN could be separately trained and deployed, thus enjoying stability in data size and view numbers. We conduct comprehensive experiments on four multiview datasets to verify the effectiveness and efficiency of the proposed approach. In the future, we plan to investigate how to transfer knowledge from external databases to further boost the cross-modal retrieval and classification performance of our method.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and anonymous reviewers for their constructive comments and valuable suggestions that remarkably improve this work.

REFERENCES

[1] C. Zhang *et al.*, "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 86–99, Jan. 2020.

[2] C. Xu, Z. Guan, W. Zhao, Y. Niu, Q. Wang, and Z. Wang, "Deep multi-view concept learning," in *Proc. Int. Joint Conf. Artif. Intell. Org.*, 2018, pp. 2898–2904.

[3] T. Zhou, C. Zhang, C. Gong, H. Bhaskar, and J. Yang, "Multiview latent space learning with feature redundancy minimization," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1655–1668, Apr. 2020.

[4] C. Zhang, H. Fu, J. Wang, W. Li, X. Cao, and Q. Hu, "Tensorized multi-view subspace representation learning," *Int. J. Comput. Vis.*, vol. 128, no. 8, pp. 2344–2361, 2020.

[5] Z. Guan, L. Zhang, J. Peng, and J. Fan, "Multi-view concept learning for data representation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3016–3028, Nov. 2015.

[6] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

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

[8] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, pp. 3893–3903, 2018.

[9] P. Hu, L. Zhen, D. Peng, and P. Liu, "Scalable deep multimodal learning for cross-modal retrieval," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 635–644.

[10] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2400–2413, Jun. 2020.

[11] J. Rupnik and J. Shawe-Taylor, "Multiview canonical correlation analysis," in *Proc. Conf. Data Min. Data Warehouses (SiKDD)*, 2010, pp. 1–4.

[12] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 593–600.

[13] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, Aug. 2018.

[14] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[15] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[16] X. Xu, K. Lin, L. Gao, H. Lu, H. T. Shen, and X. Li, "Learning cross-modal common representations by private-shared subspaces separation," *IEEE Trans. Cybern.*, early access, Aug. 11, 2020. [Online]. Available: doi.org/10.1109/TCYB.2020.3009004

[17] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, "Learning cross-modal retrieval with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5403–5413.

[18] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, Jan. 2018.

[19] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.

[20] X. Chen, S. Chen, H. Xue, and X. Zhou, "A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data," *Pattern Recognit.*, vol. 45, no. 5, pp. 2005–2018, 2012.

[21] P. Hu, H. Zhu, X. Peng, and J. Lin, "Semi-supervised multi-modal learning with balanced spectral decomposition," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, New York, NY, USA, Feb. 2020, pp. 99–106.

[22] X. Wang, W. Zhu, and C. Liu, "Semi-supervised deep quantization for cross-modal search," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1730–1739.

[23] J. Zhang, Y. Peng, and M. Yuan, "SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 489–502, Feb. 2020.

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[24] X. Wang, P. Hu, P. Liu, and D. Peng, "Deep semisupervised class- and correlation-collapsed cross-view learning," *IEEE Trans. Cybern.*, early access, May 4, 2020, doi: 10.1109/TCYB.2020.2984489.

[25] S. Yao, G. Yu, J. Wang, C. Domeniconi, and X. Zhang, "Multi-view multiple clustering," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 4121–4127.

[26] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "COMIC: Multi-view clustering without parameter selection," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 5092–5101.

[27] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1134–1143.

[28] G. Yu, X. Liu, J. Wang, C. Domeniconi, and X. Zhang, "Flexible cross-modal hashing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 14, 2020, doi: 10.1109/TNNLS.2020.3027729.

[29] P. Hu, D. Peng, X. Wang, and Y. Xiang, "Multimodal adversarial network for cross-modal retrieval," *Knowl. Based Syst.*, vol. 180, pp. 38–50, Sep. 2019.

[30] Z. Huang, P. Hu, J. T. Zhou, J. Lv, and X. Peng, "Partially view-aligned clustering," in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Assoc., Inc., Virtual, 2020.

[31] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, pp. 293–305, 2002.

[32] J. Ma, Y. Zhang, and L. Zhang, "Discriminative subspace matrix factorization for multiview data clustering," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107676.

[33] C. H. Lampert and O. Krömer, "Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 566–579.

[34] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 143–152.

[35] Y. Fang, H. Zhang, and Y. Ren, "Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing," *Knowl. Based Syst.*, vol. 171, pp. 69–80, May 2019.

[36] D. Lopez-Paz, S. Sra, A. J. Smola, Z. Ghahramani, and B. Schölkopf, "Randomized nonlinear component analysis," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1359–1367.

[37] W. Wang and K. Livescu, "Large-scale approximate kernel canonical correlation analysis," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[38] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured AutoEncoders for subspace clustering," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, Oct. 2018.

[39] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[40] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[41] X. Wang, D. Peng, P. Hu, and Y. Sang, "Adversarial correlated autoencoder for unsupervised multi-view representation learning," *Knowl. Based Syst.*, vol. 168, pp. 109–120, Mar. 2019.

[42] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Multimedia Conf.*, 2017, pp. 154–162.

[43] Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant analysis in correlation similarity measure space," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 577–584.

[44] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *Proc. 8th IEEE Int. Conf. Data Min. (ICDM)*, Pisa, Italy, 2008, pp. 1043–1048.

[45] C. Hou, L.-L. Zeng, and D. Hu, "Secure classification with augmented features," 2017. [Online]. Available: arXiv:1711.00239.

[46] C. Hou, L.-L. Zeng, and D. Hu, "Safe classification with augmented features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2176–2192, Sep. 2019.

[47] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018.

[48] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, pp. 102–112, 2019.

[49] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4847–4855.

[50] P. Hu, D. Peng, Y. Sang, and Y. Xiang, "Multi-view linear discriminant analysis network," *IEEE Trans. Image Process.*, vol. 28, pp. 5352–5365, 2019.

[51] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2014, pp. 2672–2680.

[52] Y. Xing, G. Yu, J. Wang, C. Domeniconi, and X. Zhang, "Weakly-supervised multi-view multi-instance multi-label learning," in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2020, pp. 3124–3130.

[53] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1276–1288, May 2019.

[54] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Hum. Genet.*, vol. 7, no. 2, pp. 179–188, 1936.

[55] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.

[56] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, pp. 5585–5599, 2018.

[57] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from national university of Singapore," in *Proc. ACM Conf. Image Video Retrieval (CIVR)*, Santorini, Greece, Jul. 2009, pp. 1–9.

[58] J. Krapac, M. Allan, J. Verbeek, and F. Juried, "Improving web image search results using query-relative classifiers," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, 2010, pp. 1094–1101.

[59] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[60] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011.

[61] X. He, Y. Peng, and L. Xie, "A new benchmark and approach for fine-grained cross-media retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1740–1748.

[62] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 10394–10403.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: arXiv:1409.1556.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2012, pp. 1097–1105.

[65] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," in *Proc. 1st Workshop Represent. Learn. NLP*, 2016, pp. 78–86. [Online]. Available: http://aclweb.org/anthology/W16-1609

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[67] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[68] Y. Peng, J. Qi, and Y. Yuan, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1, pp. 1–24, 2019.

[69] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.

[70] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[71] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.

[72] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

Authorized licensed use limited to: ASTAR. Downloaded on February 10,2022 at 15:41:02 UTC from IEEE Xplore. Restrictions apply.