Self-supervised Correlation Learning for Cross-Modal Retrieval

Yaxin Liu, Jianlong Wu, *Member, IEEE*, Leigang Qu, Tian Gan, Jianhua Yin, Liqiang Nie, *Senior Member, IEEE*

***Abstract*—Cross-modal retrieval aims to retrieve relevant data from another modality when given a query of one modality. Although most existing methods that rely on the label information of multimedia data have achieved promising results, the performance benefiting from labeled data comes at a high cost since labeling data often requires enormous labor resources, especially on large-scale multimedia datasets. Therefore, unsupervised cross-modal learning is of crucial importance in real-world applications. In this paper, we propose a novel unsupervised cross-modal retrieval method, named Self-supervised Correlation Learning (SCL), which takes full advantage of large amounts of unlabeled data to learn discriminative and modality-invariant representations. Since unsupervised learning lacks the supervision of category labels, we incorporate the knowledge from the input as a supervisory signal by maximizing the mutual information between the input and the output of different modality-specific projectors. Besides, for the purpose of learning discriminative representations, we exploit unsupervised contrastive learning to model the relationship among intra- and inter-modality instances, which makes similar samples closer and pushes dissimilar samples apart. Moreover, to further eliminate the modality gap, we use a weight-sharing scheme and minimize the modality-invariant loss in the joint representation space. Beyond that, we also extend the proposed method to the semi-supervised setting. Extensive experiments conducted on three widely-used benchmark datasets demonstrate that our method achieves competitive results compared with current state-of-the-art cross-modal retrieval approaches.**

***Index Terms*—Cross-Modal Retrieval; Self-supervised Contrastive Learning; Mutual Information Estimation**

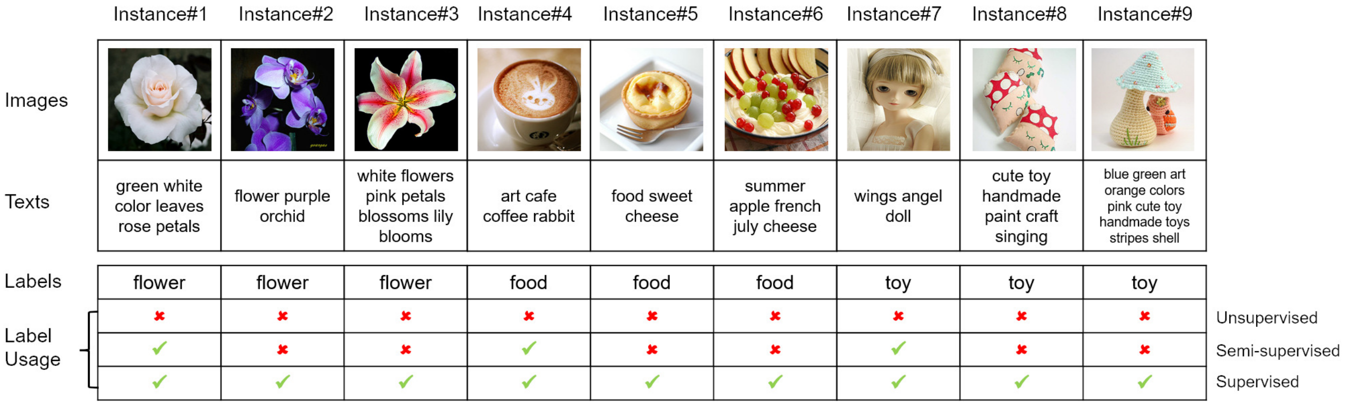
I. Introduction

THE explosive growth of multimedia data on various platforms and increasing demands for multimedia applications have attracted a surge of research interest in multimodal learning, including multimodal representation, translation, alignment, and fusion [1]. Over the past decades, cross-modal retrieval has been a popular research topic that can retrieve the relevant samples with the same semantic category from one modality when given a query of another modality. Although supervised cross-modal methods have made significant progress, they suffer from poor performance in the cases of label deficiency. In practical application scenarios, it is often difficult to label large-scale datasets, thus limiting the performance of these methods. On the current multimedia platform, paired multimedia data is readily accessible; therefore, exploiting the abundant unlabeled data is promising in broader applications. Towards this end, to alleviate the reliance on labeled data, this paper investigates the unsupervised representation learning for cross-modal retrieval to learn valid and discriminative representations from a large amount of unlabeled data.

Because of diverse distributions and feature representations in different modalities, the critical challenge for cross-modal tasks lies in how to bridge the heterogeneity gap. Currently, the mainstream cross-modal retrieval methods focus on common space learning [2]. They follow the idea that multimodal data sharing the same semantics have latent correlations, which makes it possible to construct a common space [3]. By mapping features of different modalities into a common space, it becomes much easier to measure distances among features of different modalities directly. Several pioneer efforts have been dedicated to exploring the unsupervised learning paradigms [4], [5], [6], [7], [8], [9], including traditional approaches, and deep learning based approaches. For instance, one typical kind of methods endeavors to learn multiple modality-specific transformations by maximizing the correlations between different modalities, such as Canonical Correlation Analysis (CCA) [4], Kernel Canonical Correlation Analysis (KCCA) [10], and Deep Canonical Correlation Analysis (DCCA) [7]. They evolve from a traditional linear model to a nonlinear model with the kernel trick and then to the model based on Deep Neural Networks (DNN). Although DNN-based methods can overcome the shortcomings of traditional methods, they still possess the following three disadvantages: 1) The media gap among different modalities still exists, which degrades the performance for cross-modal retrieval. 2) Due to the lack of supervision signal except for label information, they are incapable of learning high-level semantic features for multimedia data. Furthermore, 3) they do not fully harness the intra- and inter-modality correlations. As a result, the learned representations are not discriminative enough. To address the above issues, we propose a new unsupervised cross-modal method, called self-supervised correlation learning, which also follows the paradigm of latent common space learning. Firstly, in order to bridge the media gap, we use a weight-sharing scheme and minimize the modality-invariant loss in the joint representation space. In addition, we design a cross-modal mutual information loss, which provides a supervisory signal for semantic feature learning and reduces the media gap by

This work is supported by the National Natural Science Foundation of China (grant nos. 62006140, 62172261, 62176137, and U1936203), and the Shandong Provincial Natural Science Foundation (grant nos. ZR2019JQ23, ZR2019QF001 and ZR2020QF106). *(Corresponding author: Jianlong Wu.)*

Y. Liu, J. Wu, L. Qu, T. Gan, J. Yin, and L. Nie are with the School of Computer Science and Technology, Shandong University, Qingdao, 266237, China (e-mail: liuyaxin0429@gmail.com; jlwu1992@sdu.edu.cn; leigangqu@gmail.com; gantian@sdu.edu.cn; jhyin@sdu.edu.cn; nieliqiang@gmail.com).



	Instance#1	Instance#2	Instance#3	Instance#4	Instance#5	Instance#6	Instance#7	Instance#8	Instance#9	
Images										
Texts	green white color leaves rose petals	flower purple orchid	white flowers pink petals blossoms lily blooms	art cafe coffee rabbit	food sweet cheese	summer apple french july cheese	wings angel doll	cute toy handmade paint craft singing	blue green art orange colors pink cute toy handmade toys stripes shell	
Labels	flower	flower	flower	food	food	food	toy	toy	toy	
Label Usage	✗	✗	✗	✗	✗	✗	✗	✗	✗	Unsupervised
	✓	✗	✗	✓	✗	✗	✓	✗	✗	Semi-supervised
	✓	✓	✓	✓	✓	✓	✓	✓	✓	Supervised

Fig. 1: Definition of unsupervised learning, semi-supervised learning, and supervised learning. An image-text pair is regarded as an instance. All unlabeled data are used for unsupervised learning, a small proportion of labeled data and a large amount of unlabeled data are used for semi-supervised learning (1/3 labels used in the figure) while all labeled data are exploited for supervised learning.

maximizing the cross-modal mutual information between the input and the output of different projectors. Furthermore, for the purpose of learning more discriminative representations, we resort to unsupervised contrastive learning to model the relationship among intra- and inter-modality instances. Extensive experiments conducted on three benchmark datasets have demonstrated that our method can significantly improve the performance towards unsupervised cross-modal retrieval.

The main contributions of this paper are summarized as follows:

- We present a novel unsupervised cross-modal retrieval method, named self-supervised correlation learning, which exploits unsupervised contrastive learning to learn more discriminative representations by modeling the correlations among intra- and inter-modality instances. Especially, it is unnecessary to adopt a complicated sampling strategy. To the best of our knowledge, we are the first to incorporate unsupervised contrastive learning into cross-modal retrieval.
- We design a cross-modal mutual information loss to preserve the cross-modal correlation structure by maximizing the mutual information between the input and the output of different transformations. Meanwhile, we use a weight-sharing scheme and minimize the modality-invariant loss to further eliminate the modality gap.
- Extensive experiments on three widely-used benchmark datasets show that our proposed method achieves competitive performance compared with several state-of-the-art cross-modal retrieval approaches.

The organization of this article is as follows. In Section II, we review the related work in cross-modal retrieval. In Section III, we elaborate on our proposed method. In Section IV, we describe the datasets, evaluation metrics, experiments results, and ablation study. In Section V, we conclude this paper.

II. Related Work

According to whether to exploit class labels, existing cross-modal methods [11], [12], [13] can be divided into three categories: unsupervised, semi-supervised, and supervised algorithms. We will briefly introduce these related work in this section. The definition of unsupervised learning, semi-supervised learning, and supervised learning are displayed in Figure 1.

A. *Unsupervised Cross-Modal Methods*

The goal of unsupervised cross-modal retrieval is to retrieve relevant samples of one modality given a query of another modality under the condition of no label information. One of the traditional unsupervised methods is the well-known CCA [4], [10], [6], which maximizes the cross-modal correlations by learning two linear transformation functions. Another classical method is Partial Least Square (PLS) [5], [14], which is similar to CCA to maximize the covariance of two modalities by two linear projections. Furthermore, to model the relationship of multi-modal data, Multiset CCA (MCCA) [6] learns a shared space by maximizing correlations among all possible pairwise modalities. To improve these linear methods, KCCA [10] applies the kernel method to CCA, which maximizes the correlations in Reproducing Kernel Hilbert Spaces (RKHS) by two nonlinear transformations. However, the predefined kernel limits the performance and is intractable to choose. Afterwards, with the prosperity of deep neural networks, the shortcomings of kernel trick can be tackled by projecting data into a high-level common subspace. For example, DCCA [7], [15] extends CCA to a deep cross-modal model, which learns two complex nonlinear transformations to project two modalities into a latent common subspace, where the resulting representations are highly linearly correlated. Inspired by both DCCA and autoencoder networks, Deep Canonically Correlated Autoencoders (DCCAE) [9] improves DCCA by adding an autoencoder regularization term. However, these methods focus solely on the correlation between coupled cross-modal items (e.g. an image and its corresponding text descriptions) while ignoring the intra- and inter-modality correlation. Recently, Zheng *et al.* [16] propose a dual-path

CNN model with instance loss for visual-textual embedding learning. Chen *et al.* [17] integrate the information theory and adversarial learning to reduce the semantic gap. These methods usually adopt triplet loss to associate the intra- and inter-modality similarity, which require semi-hard negative mining and neglect the semantic information in the original features.

B. Semi-supervised Cross-Modal Methods

Semi-supervised methods [18], [19], [20], [21], [22], [23], [24] exploit a small proportion of labeled cross-modal data and a large amount of unlabeled cross-modal data to learn a latent discriminative common subspace for cross-modal retrieval. Considering that previous methods usually focus on modeling the pairwise correlation or semantic information separately, Joint Representation Learning (JRL) [18] jointly explores the correlation and semantic information in a unified optimization framework, which uses the sparse and graph regularization to extract useful clues in both labeled and unlabeled data. However, it models different media types into different graphs separately while neglecting the cross-media correlations among various modalities. To make full use of cross-media semantic relationship, Semi-Supervised Unified Patch Graph ($\mathrm{S^2UPG}$) [19] regularizes the cross-media feature learning by simultaneously modeling multimedia instances and their patches in one joint graph. In addition, Generalized Semi-supervised Structured Subspace Learning (GSS-SL) [21] expoits a label graph as linkage to optimize multiple transformations, making the graphs of different modalities consistent with the same label graph. However, these methods need to compute the graph matrix based on the whole training dataset, which suffers from high computation and space complexity, and thus is inapplicable for large-scale datasets. To overcome this issue, Semi-supervised Multimodal Learning Network (SMLN) [24] correlates different modalities by constructing a similarity matrix based on labeled and unlabeled multimedia data and designs a novel eigenvalue-based loss function to exactly balance all eigenvalues, which can be trained in a batch-by-batch manner and handle large-scale databases.

C. Supervised Cross-Modal Methods

Supervised cross-modal methods exploit label information to learn the common space, achieving considerable performance in cross-modal retrieval. Over the past few years, a great deal of supervised methods have sprung up [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]. With the guidance of semantic labels, they are able to learn a common discriminative subspace, in which the intra-class variation is minimized while the inter-class one is maximized. For example, Adversarial Cross-Modal Retrieval (ACMR) [26] seeks an effective common subspace based on adversarial training by learning modality-invariant representations during the minimax game. In specific, it imposes a triplet constraint to minimize the gap of all items from different modalities of the same class while maximizing the distances among semantically different samples. Deep Supervised Cross-Modal Retrieval (DSCMR) [30] minimizes the discrimination loss in both the label space and the common representation space to supervise the model learning discriminative features. Scalable Deep Multimodal Learning (SDML) [31] predefines a common subspace restricted to label space by a fixed orthogonal matrix. In this way, it can train different modality-specific networks independently and be scalable to the variable number of modalities.

Although the existing supervised cross-modal methods have achieved excellent performance, they highly depend on the label information that requires much human labor. Compared with supervised methods, unsupervised ones completely exploit self-supervised information, which can learn more general semantic features.

D. Contrastive Learning Methods

Contrastive learning is first proposed by Hadsell *et al.* [36] for dimension reduction, which minimizes the distance between similar pairs and enforces a margin between the distance of dissimilar pairs. To better distinguish between positive and negative samples, triplet loss [37] minimizes the distance between an anchor and a positive point and maximizes the distance between an anchor and a negative point. N-pair loss [38] generalizes the triplet loss by comparing more than one negative example. In recent years, self-supervised learning has received widespread attention because of its ability to learn effective representations from a large amount of unlabeled data, and many studies [39] based on contrastive learning have been produced. For example, Wu *et al.* [40] propose to learn feature representations via non-parametric instance discrimination, which uses noise-contrastive estimation to approximate the similarity to all instances and maintains a memory bank to store the feature representations. Oord *et al.* [41] propose InfoNCE loss and relates it to the maximization of mutual information between latent representations. MoCo [42] and SimCLR [43] adapt a dual-branch structure to maximize the agreement of two augmentation views. For negative sampling, instead of a memory bank, MoCo uses a momentum-updated encoder to obtain negative samples, and SimCLR employs the samples in the current batch with a large batch size.

III. Proposed method

In this section, we detail our proposed framework as shown in Figure 2. First of all, we present the problem formulation and the modality-invariant loss in Section III-A. Then, we introduce the intrinsic structure preservation module and the contrastive discrimination function in Section III-B and Section III-C, respectively. Afterwards, we optimize the whole objective function as shown in Algorithm 1 in Section III-D. In the end, we extend the proposed method to the semi-supervised setting in Section III-E.

A. Problem Formulation

Without losing generality, we focus on cross-modal representation learning for bimodal data, specially for images and texts. Assuming that there is a cross-modal dataset,

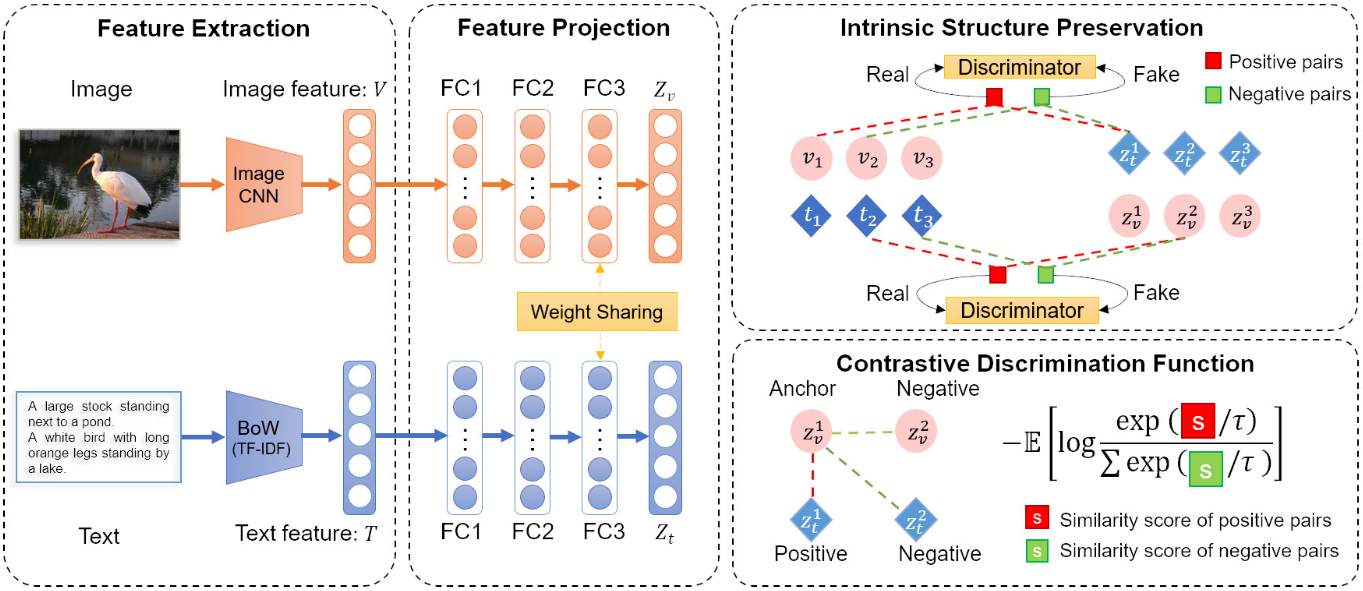


Fig. 2: The general framework of our proposed SCL. It consists of four modules: feature extraction, feature projection, intrinsic structure preservation, and contrastive discriminative function. The intrinsic structure preservation module correlates the original feature space and common representation space. And the contrastive discrimination function models the relationship among intra- and inter-modality pairs in the common space.

represented as $\mathcal{D} = \{\mathcal{D}_{tr}, \mathcal{D}_{va}, \mathcal{D}_{te}\}$, where $\mathcal{D}_{tr}$, $\mathcal{D}_{va}$, and $\mathcal{D}_{te}$ denote the training, validation, and testing sets, respectively. Specifically, for the training set, $\mathcal{D}_{tr} = \{\mathcal{V}_{tr}, \mathcal{T}_{tr}\}$, where $\mathcal{V}_{tr} = \{\boldsymbol{v}_i\}_{i=1}^{n_{tr}}$ and $\mathcal{T}_{tr} = \{\boldsymbol{t}_i\}_{i=1}^{n_{tr}}$. $\boldsymbol{v}_i \in \mathbb{R}^{d_v}$ is the image feature vector and $\boldsymbol{t}_i \in \mathbb{R}^{d_t}$ is the text feature vector. Here, d_v and d_t are the dimensions of image and text features, respectively. n_{tr} is the number of image-text pairs in the training set. Note that, because it is time-consuming and labor-intensive to annotate the large-scale data in practice, there are no labels in the training set. For the testing set, $\mathcal{D}_{te} = \{\mathcal{V}_{te}, \mathcal{T}_{te}, \mathcal{Y}_{te}\}$, where $\mathcal{Y}_{te} = \{\boldsymbol{y}_i\}_{i=1}^{n_{te}}$ is provided for evaluation metrics. $\boldsymbol{y}_i \in \mathbb{R}^c$ is the label of the ith sample, where c is the class number of the dataset. For the single-label data, $\boldsymbol{y}_i$ is a one-hot vector, while for the multilabel data, $\boldsymbol{y}_i$ consists of more than one nonzero value. n_{te} is the number of image-text pairs in the testing set. The validation set is denoted similarly to the testing set. Our goal is to retrieve the corresponding samples with the same semantic category from another modality when given a query from one modality.

It is well-known that the main challenge of cross-modal retrieval is the media gap between different modalities. To overcome it, we employ two modality-specific non-linear transformations to project original features into a common subspace, where we can measure the similarity between features of different modalities as in single-modality scenarios. We denote the projected representation as $\boldsymbol{z}_v = f_v(\boldsymbol{v}; \theta_v)$ and $\boldsymbol{z}_t = f_t(\boldsymbol{t}; \theta_t)$ for image and text, respectively, where θ_v and θ_t are respectively the trainable parameters of image projector f_v and text projector f_t. $\boldsymbol{z}_v$ and $\boldsymbol{z}_t$ have the same dimension as d_z.

Following the assumption that multimedia data share high-level semantic representations, to ensure that two subnetworks learn a latent common representation, we enforce them to share the weights of the last layers. Besides, a modality-invariant loss is applied to maintain the pairwise correlation in the common space. By this means, it not only encodes the consistent information in the original features of different modalities through two nonlinear projectors but also reduces the scale of parameters of the network, subsequently alleviating the training difficulty.

The modality-invariant loss is defined as:

$$\mathcal{L}_{CM} = \frac{1}{N}\left\|\boldsymbol{Z}_v - \boldsymbol{Z}_t\right\|_2, \qquad (1)$$

where $\boldsymbol{Z}_v$ and $\boldsymbol{Z}_t$ are the projected representations in a minibatch, and each row of them denotes the common representation of an image and a text, respectively.

B. Intrinsic Structure Preservation

As mentioned before, due to the lack of supervision of label information, previous unsupervised methods are incapable of learning high-level semantic features for cross-modal retrieval. Thus, one of our challenges is to figure out how to introduce a supervisory signal to guide the model to learn useful representations. In fact, there is a large amount of semantic information in the finely extracted features, which can be regarded as the supervision information. Based on that, we construct the intrinsic correlations between the original features and the common representations by maximizing the Mutual Information (MI) between the input and the output of modality-specific projectors. In this way, the learned representations in the common space are consistent with the original features, which preserves the semantic structure of cross-modal data.

Our mutual information maximization framework is presented in the upper right of Figure 2. Considering that there exists the media gap between different modalities, we expect to model the cross-modal correlation between the original features and the common representations and meanwhile bridge the heterogeneous gap. Therefore, in addition to the intra-modal mutual information losses, we design the cross-modal mutual information losses between two modality-specific projectors, which simultaneously maximizes the mutual information between the original image features and the projected text representations, as well as between the original text features and the projected image representations.

Following [44], [45], we train a discriminator T_ω to distinguish between samples coming from the joint $\mathbb{P}_{VZ_t}$ and the product of marginals $\mathbb{P}_V \times \mathbb{P}_{Z_t}$. Similarly, we train another discriminator T_ϕ to distinguish between samples coming from the joint $\mathbb{P}_{TZ_v}$ and the product of marginals $\mathbb{P}_T \times \mathbb{P}_{Z_v}$. If samples are from the same instance, they are positive pairs, otherwise, they are negative pairs. Referring to [41], we employ the infoNCE [46] estimator as a lower-bound on MI to maximize it.

The cross-modal mutual information losses are defined as follows:

$$\mathcal{L}^{vt}_{MI} = \mathbb{E}_{\mathbb{P}_{VZ_t}}[T_{\theta_t,\omega}(\boldsymbol{v}, \boldsymbol{z}_t) - \log \sum_{\mathbb{P}_V \times \mathbb{P}_{Z_t}} e^{T_{\theta_t,\omega}(\boldsymbol{v},\boldsymbol{z}_t)}], \quad (2)$$

$$\mathcal{L}^{tv}_{MI} = \mathbb{E}_{\mathbb{P}_{TZ_v}}[T_{\theta_v,\phi}(\boldsymbol{t}, \boldsymbol{z}_v) - \log \sum_{\mathbb{P}_T \times \mathbb{P}_{Z_v}} e^{T_{\theta_v,\phi}(\boldsymbol{t},\boldsymbol{z}_v)}], \quad (3)$$

and the intra-modal mutual information losses are defined as follows:

$$\mathcal{L}^{vv}_{MI} = \mathbb{E}_{\mathbb{P}_{VZ_v}}[T_{\theta_v,\omega}(\boldsymbol{v}, \boldsymbol{z}_v) - \log \sum_{\mathbb{P}_V \times \mathbb{P}_{Z_v}} e^{T_{\theta_v,\omega}(\boldsymbol{v},\boldsymbol{z}_v)}], \quad (4)$$

$$\mathcal{L}^{tt}_{MI} = \mathbb{E}_{\mathbb{P}_{TZ_t}}[T_{\theta_t,\phi}(\boldsymbol{t}, \boldsymbol{z}_t) - \log \sum_{\mathbb{P}_T \times \mathbb{P}_{Z_t}} e^{T_{\theta_t,\phi}(\boldsymbol{t},\boldsymbol{z}_t)}], \quad (5)$$

where θ_v and θ_t are the parameters of image and text projectors, respectively. Because the projector and mutual information estimator require similar computations, we share layers of these functions as [45], so that $T_{\theta_t,\omega} = D_\omega \circ g \circ (\boldsymbol{v}, f_t(\boldsymbol{t}))$, $T_{\theta_v,\phi} = D_\phi \circ g \circ (\boldsymbol{t}, f_v(\boldsymbol{v}))$, $T_{\theta_v,\omega} = D_\omega \circ g \circ (\boldsymbol{v}, f_v(\boldsymbol{v}))$, and $T_{\theta_t,\phi} = D_\phi \circ g \circ (\boldsymbol{t}, f_t(\boldsymbol{t}))$, where g is a combination function. In detail, we first feed the original images or texts features into two fully connected layers, where the number of units is 1024 and 512. Then, we concatenate its output with the projector output. After that, we pass this to three fully connected layers $(1024 \rightarrow 512 \rightarrow 512 \rightarrow 1)$ to get the score of the discriminator. Each hidden layer is followed by a ReLU activation. In addition, to simplify the discriminator, we share layers of two discriminators with the same computations.

Finally, we combine the mutual information loss as:

$$\mathcal{L}_{MI} = \mathcal{L}^{vt}_{MI} + \mathcal{L}^{tv}_{MI} + \mathcal{L}^{vv}_{MI} + \mathcal{L}^{tt}_{MI}. \quad (6)$$

Algorithm 1 The optimization procedure of our proposed SCL

```
Input: The training dataset of image feature $\mathcal{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n_{tr}}\}$ and text feature $\mathcal{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{n_{tr}}\}$, the batch size $N$, the learning rate $\eta$, the hyper-parameters $\alpha$ and $\beta$, and the temperature parameter $\tau$.
Output: The optimal weight paramters $\theta_i$, $\theta_t$, $\omega$ and $\phi$.
while not converge do
    Randomly select $N$ samples of each modality to contrust a minibatch.
    Learn the latent representations $\boldsymbol{z}_v$, $\boldsymbol{z}_t$ for the samples in the minibatch by forward-propagation.
    Compute the modality-invariant loss $\mathcal{L}_{CM}$ according to Eq. (1).
    Compute the mutual information loss $\mathcal{L}_{MI}$ according to Eq. (6).
    Compute the contrastive loss $\mathcal{L}_{CD}$ according to Eq. (7).
    Compute the overall loss $\mathcal{L}$ according to Eq. (8) and update parameters $\theta_v$, $\theta_t$, $\omega$ and $\phi$ by stochastic gradient descent as follows:
        $\theta_v \leftarrow \theta_v - \eta\frac{\partial\mathcal{L}}{\partial\theta_v}$; $\theta_t \leftarrow \theta_t - \eta\frac{\partial\mathcal{L}}{\partial\theta_t}$;
        $\omega \leftarrow \omega - \eta\frac{\partial\mathcal{L}_{MI}}{\partial\omega}$; $\phi \leftarrow \phi - \eta\frac{\partial\mathcal{L}_{MI}}{\partial\phi}$.
end while
```

C. Contrastive Discrimination Function

As discussed above, most existing unsupervised methods as [7], [15] only focus on the pairwise correlation whereas neglecting the correlation of dissimilar samples, which is insufficient to generate discriminative representations. Currently, supervised methods [26], [27] have achieved promising performance for cross-modal retrieval by utilizing the semantic category labels in the following two aspects: 1) They employ the classification task to exploit the semantic information in each modality for learning intra-modal discriminative representations. And 2) they model the correlations between similar and dissimilar instances in different modalities, which minimizes the distance between the image-text pairs with the same label and maximizes the distance between pairs with different labels. However, the performance conferred by the labeled data suffers from a significant cost since labeling data often requires enormous human labor resources.

To overcome these limitations, we propose to model the intra- and inter-modality correlation in an unsupervised contrastive learning framework [41], [40], [42], [43] and seek a common subspace, where similar samples are closer and dissimilar samples are far away. Different from previous work of contrastive learning of visual representations, we introduce an unsupervised contrastive learning paradigm in the field of cross-modal retrieval. Significantly, it does not depend on the category labels to guide the procedure of representation learning, and no complicated sampling mechanism is required.

The idea behind contrastive learning is to learn a discriminative function to distinguish samples from different distributions, i.e., correctly select a positive sample x out of the set $\mathcal{S} = \{\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$ that contains k negative samples. In order to bridge the media gap and learn discriminative representations, we narrow the distance between pairwise instances and extend the distance between unpaired instances. The sampling strategy of our defined contrastive loss is described in the following. Firstly, we randomly select N image-text pairs to construct a minibatch. Then, the contrastive loss is computed across all the image-text pairs derived from the minibatch, resulting in $2N$ data points. As demonstrated in the bottom right of Figure 2, given a sample of an image as the anchor, we view its corresponding pairwise text representation as the positive sample. Moreover, instead of sampling negative examples complicatedly, we regard the representations of the other $2(N-1)$ data points within a minibatch as the negative samples similar to [47]. Similarly, given a text example as the anchor, the positive and negative samples are defined in the same way.

Finally, we integrate the contrastive loss on all positive pairs in a minibatch as:

$$\mathcal{L}_{CD} = -\frac{1}{2N}\sum_{i=1}^{2N}\log\frac{\exp\left(\cos(\boldsymbol{z}_a^i, \boldsymbol{z}_p)/\tau\right)}{\sum_{j=1}^{2(N-1)}\exp\left(\cos(\boldsymbol{z}_a^i, \boldsymbol{z}_n^j)/\tau\right)}, \quad (7)$$

where $\boldsymbol{z}_a$ represents the anchor sample, $\boldsymbol{z}_p$ indicates the positive sample, and $\boldsymbol{z}_n$ denotes one of the negative samples. τ is an adjusted temperature parameter and $\cos(\boldsymbol{z}_i, \boldsymbol{z}_j) = (\boldsymbol{z}_i \cdot \boldsymbol{z}_j)/(\|\boldsymbol{z}_i\|_2 \cdot \|\boldsymbol{z}_j\|_2)$ is the cosine similarity between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$.

D. Optimization and Training Strategy

The overview framework for our SCL is shown in Figure 2. It is composed of four different modules, i.e., feature extraction, feature projection, intrinsic structure preservation, and contrastive discriminative function. The overall objective of our method is as follows:

$$\mathcal{L} = \mathcal{L}_{CM} + \alpha \cdot \mathcal{L}_{MI} + \beta \cdot \mathcal{L}_{CD}, \quad (8)$$

where $\mathcal{L}_{CM}$ is the modality-invariant loss, $\mathcal{L}_{MI}$ is the mutual information loss for correlation structure preservation, and $\mathcal{L}_{CD}$ is the contrastive loss for contrastive discrimination function. The above three losses are minimized simultaneously to eliminate the media gap among different modalities and learn discriminative representations in the common space. α and β are the trade-off hyper-parameters for modality-invariant loss, mutual information loss, and contrastive loss. We summarize the proposed method in Algorithm 1.

E. Extension to the Semi-supervised Situation

To fully exploit the labeled and unlabeled data as well as achieve a better trade-off between annotations and performance, we extend our SCL to the semi-supervised setting from two aspects: label prediction and class-aware contrastive discrimination.

1) Label Prediction: In order to learn intra-modal discriminative representations, a classifier is employed to predict the semantic labels of items for each modality. Given the joint embeddings $\boldsymbol{z}_v$ and $\boldsymbol{z}_t$, the formulation of the classification objective is as follows:

$$\boldsymbol{p}_v^i = \boldsymbol{softmax}(\boldsymbol{W}^\top \boldsymbol{z}_v^i + \boldsymbol{b}), \quad (9)$$

$$\boldsymbol{p}_t^i = \boldsymbol{softmax}(\boldsymbol{W}^\top \boldsymbol{z}_t^i + \boldsymbol{b}), \quad (10)$$

$$\mathcal{L}_C = -\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i \cdot (\log \boldsymbol{p}_v^i + \log \boldsymbol{p}_t^i)), \quad (11)$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weight and bias of the last fully connected layer, respectively. Besides, $\boldsymbol{p}_v$ and $\boldsymbol{p}_t$ denote the predicted probability of semantic categories of image and text modalities.

2) Class-aware Contrastive Discrimination: Instead of the instance-aware contrastive discrimination function, the class-aware one takes the label information into account and measures the intra- and inter-class discrepancies across modalities. To be specific, the intra-class discrepancy is minimized to compact the feature representations of samples within a class, whereas the inter-class discrepancy is maximized to push away the representations of samples from different classes. Like Eq. (7), we formulate the class-aware contrastive loss as follows:

$$\mathcal{L}_{CCD} = -\frac{1}{2N}\sum_{i=1}^{2N}\log\frac{\sum_j \exp\left(\cos(\boldsymbol{z}_a^i, \boldsymbol{z}_p^j)/\tau\right)}{\sum_k \exp\left(\cos(\boldsymbol{z}_a^i, \boldsymbol{z}_n^k)/\tau\right)}, \quad (12)$$

where $\boldsymbol{z}_a$ represents the anchor sample, $\boldsymbol{z}_p$ and $\boldsymbol{z}_n$ indicate one of the positive samples and the negative samples, respectively. Different from Section III-C, all instances in a minibatch with the same class as anchor are treated as the positive samples while the others are considered as the negative samples. τ is an adjusted temperature parameter.

3) Objective Function: The loss function for our proposed semi-supervised SCL consists of two terms, a supervised loss $\mathcal{L}_S$ and an unsupervised loss $\mathcal{L}_U$. Let us define $\mathcal{X} = \{\boldsymbol{x}_i : i \in (1, \ldots, N)\}$ as a batch of N labeled instances, where $\boldsymbol{x}_i = (\boldsymbol{v}_i, \boldsymbol{t}_i, \boldsymbol{y}_i)$ are image-text pairs with one-hot labels and let $\mathcal{U} = \{\boldsymbol{u}_i : i \in (1, \ldots, N)\}$ be a batch of N unlabeled examples, where $\boldsymbol{u}_i = (\boldsymbol{v}_i, \boldsymbol{t}_i)$ are image-text pairs without labels. For labeled and unlabeled data, the supervised and unsupervised losses are computed respectively as follows:

$$\mathcal{L}_S = \mathcal{L}_C(\boldsymbol{x}_i) + \gamma \cdot \mathcal{L}_{CCD}(\boldsymbol{x}_i), \quad (13)$$

$$\mathcal{L}_U = \mathcal{L}_{CM}(\boldsymbol{u}_i) + \alpha \cdot \mathcal{L}_{MI}(\boldsymbol{u}_i) + \beta \cdot \mathcal{L}_{CD}(\boldsymbol{u}_i). \quad (14)$$

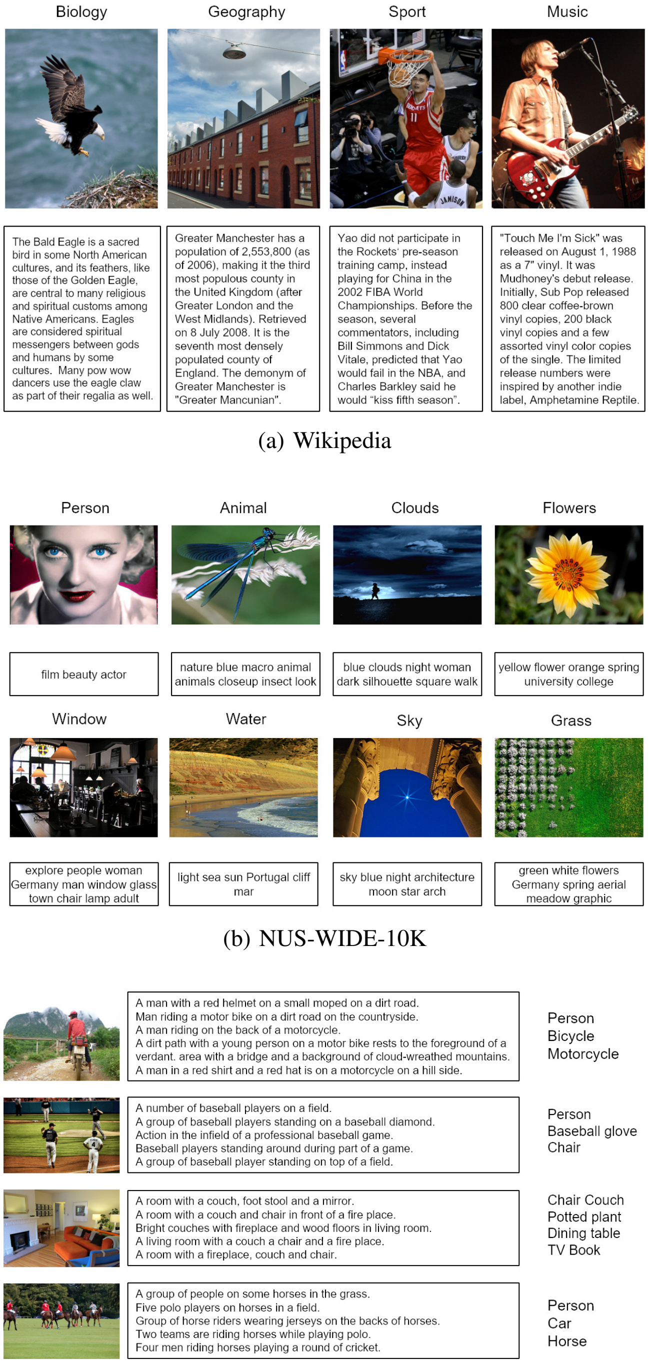
The overall objective function is summarized as:

$$\mathcal{L}_{SS} = \mathcal{L}_S + \lambda \cdot \mathcal{L}_U. \quad (15)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3152086, IEEE Transactions on Multimedia

TABLE I: General statistics of the three datasets used in the experiments, where "*/*/*" in the "Instances" column stands for the number of training/validation/testing subsets.

Dataset	Instances	Labels	Image feature	Text feature
Wikipedia	2,173/231/462	10	4,096D VGG	5,000D BoW
NUS-WIDE-10K	8,000/1,000/1,000	10	4,096D VGG	1,000D BoW
MS-COCO	82,081/5,000/5,000	80	4,096D VGG	3,000D BoW



The Bald Eagle is a sacred bird in some North American cultures, and its feathers, like those of the Golden Eagle, are central to many religious and spiritual customs among Native Americans. Eagles are considered spiritual messengers between gods and humans by some cultures. Many pow wow dancers use the eagle claw as part of their regalia as well.

Greater Manchester has a population of 2,553,800 (as of 2006), making it the third most populous county in the United Kingdom (after Greater London and the West Midlands). Retrieved on 8 July 2008. It is the seventh most densely populated county of England. The demonym of Greater Manchester is "Greater Mancunian".

Yao did not participate in the Rockets' pre-season training camp, instead playing for China in the 2002 FIBA World Championships. Before the season, several commentators, including Bill Simmons and Dick Vitale, predicted that Yao would fail in the NBA, and Charles Barkley said he would "kiss fifth season".

"Touch Me I'm Sick" was released on August 1, 1988 as a 7" vinyl. It was Mudhoney's debut release. Initially, Sub Pop released 800 clear coffee-brown vinyl copies, 200 black vinyl copies and a few assorted vinyl color copies of the single. The limited release numbers were inspired by another indie label, Amphetamine Reptile.

(a) Wikipedia

Person
Animal
Clouds
Flowers
film beauty actor
nature blue macro animal animals closeup insect look
blue clouds night woman dark silhouette square walk
yellow flower orange spring university college
Window
Water
Sky
Grass
explore people woman Germany man window glass town chair lamp adult
light sea sun Portugal cliff mar
sky blue night architecture moon star arch
green white flowers Germany spring aerial meadow graphic

(b) NUS-WIDE-10K

A man with a red helmet on a small moped on a dirt road.
Man riding a motor bike on a dirt road on the countryside.
A man riding on the back of a motorcycle.
A dirt path with a young person on a motor bike rests to the foreground of a verdant. area with a bridge and a background of cloud-wreathed mountains.
A man in a red shirt and a red hat is on a motorcycle on a hill side.

A number of baseball players on a field.
A group of baseball players standing on a baseball diamond.
Action in the infield of a professional baseball game.
Baseball players standing around during part of a game.
A group of baseball player standing on top of a field.

A room with a couch, foot stool and a mirror.
A room with a couch and chair in front of a fire place.
Bright couches with fireplace and wood floors in living room.
A living room with a couch a chair and a fire place.
A room with a fireplace, couch and chair.

A group of people on some horses in the grass.
Five polo players on horses in a field.
Group of horse riders wearing jerseys on the backs of horses.
Two teams are riding horses while playing polo.
Four men riding horses playing a round of cricket.

(c) MS-COCO

Fig. 3: Examples of the image, text, and label (s) for the Wikipedia dataset, the NUS-WIDE-10K dataset, and the MS-COCO dataset.

IV. Experiments

To verify the effectiveness of our model, we conducted experiments on three widely-used cross-modal benchmark datasets: the Wikipedia dataset [48], the NUS-WIDE-10k dataset [49], and the MS-COCO dataset [50]. In the experiments reported below, we first compared our proposed method with several state-of-the-art methods. Then we conducted additional evaluations to investigate the effectiveness of each component of our method.

A. Experimental Setup

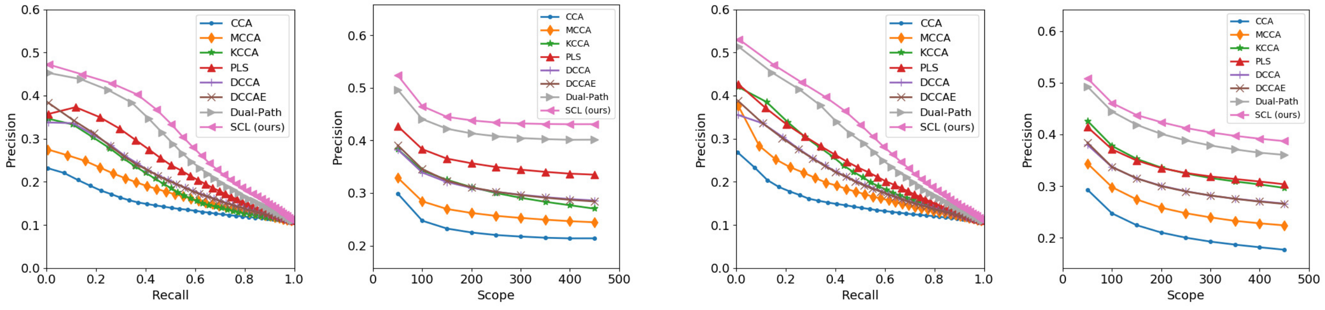
1) Datasets and Features: Three cross-modal datasets are adopted in our experiments, including Wikipedia, NUS-WIDE-10K, and MS-COCO. The statistics of three datasets are summarized in Table I. Some samples of the image, text, and label (s) for Wikipedia, NUS-WIDE-10K, and MS-COCO are shown in Figure 3.

- The Wikipedia dataset [48] is the most widely-used dataset for cross-modal retrieval. It contains 2,866 image-text pairs with 10 semantic categories. Each pair has an image and the corresponding text. Following [8], it is split into three partitions: 2,173 pairs for training, 231 pairs for validation, and 462 pairs for testing.
- The NUS-WIDE-10K dataset is a subset of the NUS-WIDE dataset [49] with 10,000 image-text pairs. Each pair selected from one of the 10 largest categories consists of an image and a unique tag as the class label. Following [8], this dataset is divided into three subsets: training set with 8,000 pairs, validation set with 1,000 pairs, and testing set with 1,000 pairs.
- The MS-COCO dataset [50] contains 123,287 images and their annotated sentences with their labels categorized into 80 classes. Note that each image usually contains several objects. After pruning images without category information, MS-COCO consists of 82,081 training images and 40,173 validation images, each of which is associated with five sentences. In our experiments, we selected 82,081 pairs for training, 5,000 pairs for validation, and 5,000 pairs for testing.

The image features in our experiments are extracted from the fc7 layer of the 19-layer VGGNet pre-trained on ImageNet. The text features are represented by a Bag of Words (BoW) vector with the TF-IDF weighting scheme. The feature dimensions of diverse datasets are presented in Table I.

2) Implementation Details: In this work, we employed two fully-connected layers to project the original modality-specific features into a common space. Each layer is followed by a ReLU [51] activation. The numbers of the hidden units for the two layers are 1,024 and 512. The input dimension is set according to the feature size of specific modalities in different datasets, e.g., 4,096 for image features and 5,000 for text features in the Wikipedia dataset. Furthermore, we used the l2-normalization after the last layer of each projector.

We set the learning rate to 0.0001 and trained the models for 200 epochs with the Adam [52] in all our experiments. After fixing the value of batch size N at 256 and the temperature



(a) Image2Text

(b) Text2Image

Fig. 4: Precision-recall curves and precision-scope curves for the image-query-texts and text-query-images experiments on the Wikipedia dataset.

parameter τ at 0.5, we tuned the parameters α and β on each dataset using grid search. The analysis of α and β is shown in Figure 5. Similarly, we experimented on parameter sensitivity of N and τ when setting α and β to 0.01 and 1.0, respectively. The retrieval performance is displayed in Figure 6.

The proposed method is trained on one Nvidia TITAN Xp GPU in Pytorch. At inference time, we calculated the cosine similarity among cross-modal samples directly in the common space.

3) Evaluation Metric and Compared Methods: To evaluate the performance of the methods, we performed cross-modal retrieval tasks, i.e., retrieving one modality with query of another modality, such as retrieving text by image query (Image2Text) and retrieving image by text query (Text2Image). The evaluation metric we adopted is the mean Average Precision (mAP) calculated on all returned results following [31], where cosine similarity is used to measure the distances of representations. In addition, we also plotted precision-recall curves and precision-scope curves for additional comparisons.

To verify the effectiveness of our method, we compared it with 13 state-of-the-art cross-modal retrieval methods including four traditional cross-modal methods, namely CCA [4], MCCA [6], KCCA [10], and PLS [5], and nine DNN-based cross-modal methods, namely DCCA [7], DCCAE [9], JRL [18], GSS-SL [21], ACMR [26], DSCMR [30], SDML [31], Dual-Path [16] and SMLN [24]. Concretely, CCA, MCCA, KCCA, PLS, DCCA, DCCAE and Dual-Path are unsupervised methods; ACMR, DSCMR, and SDML are supervised methods; JRL, GSS-SL, and SMLN are semi-supervised methods. For a fair comparison, all the compared methods adopt the same image and text features as our approach.

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

In this section, we evaluated the effectiveness of our proposed method by comparing with 13 state-of-the-art cross-modal methods on two cross-modal retrieval tasks, i.e., Image2Text and Text2Image, on three cross-modal datasets, i.e., Wikipedia, NUS-WIDE-10K, and MS-COCO under unsupervised and semi-supervised settings.

1) Comparison Under Unsupervised Settings: We compared the unsupervised and supervised methods under the unsupervised settings, where only unlabeled data are used to train the models. The average results are shown in Table II.

From this table, we can observe that our proposed method achieves the best performance on all datasets compared to traditional and DNN-based methods. As we can see from Table II, our SCL outperforms the best competitor (i.e., Dual-Path) by 2.9%, 0.6%, and 1.3% for image-query-texts, 2.6%, 0.7%, and 1.4% for text-query-images, and 2.8%, 0.7%, and 1.4% for average on three datasets, respectively. This shows the effectiveness of modeling the correlations between pairwise instances, and among intra- and inter-modality samples.

Furthermore, we can see that the performance of supervised methods is even worse than most of unsupervised methods under the circumstances where no label information is offered, which proves that the supervised methods highly rely on the labeled data while ignoring the correlation in a large amount of unlabeled data.

In addition to the evaluation terms of the mAP score, we also drew precision-recall curves and precision-scope curves for additional comparisons. The precision-recall and precision-scope curves of the image-query-texts and text-query-images on the Wikipedia dataset are respectively plotted in Figure 4 (a) and Figure 4 (b). The scope (i.e., the top K retrieved samples) of the precision-scope varies from $K = 50$ to 500. As we can see from these curves, our SCL gets the best results compared with the other 7 state-of-the-art unsupervised cross-modal retrieval methods.

2) Comparison Under Semi-supervised Settings: In this section, we compared our unsupervised and semi-supervised SCL with the state-of-the-art supervised and semi-supervised methods under semi-supervised settings of 5%, 10%, and 30% labels, where the corresponding percentage of labeled data and all the unlabeled data are used to train. The average results are shown in Table III.

As illustrated in Table III, the supervised methods perform terribly with very few labeled data, and the performance gradually improves as the proportion of labeled data increases. It indicates that models can learn high-level semantic features from the labeled data with the supervision of label information.

In contrast, our SCL$_{us}$ exceeds existing supervised and semi-supervised methods trained with 5% and 10% labeled

TABLE II: Performance comparison in terms of mAP scores on Wikipedia, NUS-WIDE-10K and MS-COCO, where the subscript "us", and "s" denote "unsupervised" and "supervised", respectively. And "-" means no repeated result available yet due to high computation complexity and memory cost. The best results on each dataset are highlighted in bold and the second best results are underlined.

Method	Wikipedia			NUS-WIDE-10K			MS-COCO		
	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average
CCA$_{us}$ [4]	0.214	0.176	0.195	0.352	0.346	0.349	0.509	0.507	0.508
MCCA$_{us}$ [6]	0.244	0.223	0.234	0.331	0.328	0.330	0.557	0.554	0.555
KCCA$_{us}$ [10]	0.270	0.295	0.283	0.424	0.334	0.379	-	-	-
PLS$_{us}$ [5]	0.335	0.302	0.318	0.444	0.424	0.434	0.634	0.632	0.633
DCCA$_{us}$ [7]	0.285	0.265	0.275	0.377	0.367	0.372	0.535	0.523	0.529
DCCAE$_{us}$ [9]	0.283	0.265	0.274	0.375	0.369	0.372	0.567	0.557	0.562
ACMR$_{s}$ [26]	0.167	0.155	0.161	0.305	0.329	0.317	0.347	0.425	0.386
DSCMR$_{s}$ [30]	0.125	0.126	0.125	0.311	0.302	0.306	0.355	0.387	0.371
SDML$_{s}$ [31]	0.142	0.136	0.139	0.289	0.288	0.288	0.333	0.336	0.335
Dual-Path$_{us}$ [16]	0.402	0.360	0.381	0.502	0.462	0.482	0.655	0.654	0.655
SCL$_{us}$ (ours)	**0.431**	**0.386**	**0.409**	**0.508**	**0.469**	**0.489**	**0.668**	**0.668**	**0.668**

TABLE III: Performance comparison under semi-supervised settings of 5%, 10%, and 30% labeled data on the Wikipedia dataset, where the subscript "ss" and "s" denote "semi-supervised" and "supervised", respectively. The best results are highlighted in bold and the second best results are underlined.

Method	5% Labeled Data			10% Labeled Data			30% Labeled Data		
	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average
JRL$_{ss}$ [18]	0.229	0.225	0.227	0.297	0.280	0.289	0.377	0.34	0.359
GSS-SL$_{ss}$ [21]	0.258	0.244	0.251	0.307	0.274	0.291	0.345	0.306	0.326
ACMR$_{s}$ [26]	0.347	0.294	0.321	0.389	0.338	0.364	0.454	0.398	0.426
DSCMR$_{s}$ [30]	0.360	0.310	0.335	0.410	0.361	0.385	0.473	0.408	0.441
SDML$_{s}$ [31]	0.370	0.314	0.342	0.409	0.362	0.386	0.480	0.416	0.448
SMLN$_{ss}$ [24]	0.389	0.359	0.374	0.407	0.362	0.385	0.459	0.413	0.436
SCL$_{us}$ (ours)	0.431	0.386	0.409	0.431	0.386	0.409	0.431	0.386	0.409
SCL$_{ss}$ (ours)	**0.460**	**0.406**	**0.433**	**0.478**	**0.428**	**0.453**	**0.504**	**0.442**	**0.473**

TABLE IV: Generalization on unseen NUS-WIDE-10K testing set using different pre-trained models on MIRFlickr dataset, where the subscript "us", and "s" denote "unsupervised" and "supervised", respectively. The best results are highlighted in bold and the second best results are underlined.

Method	Image2Text	Text2Image	Average
CCA$_{us}$ [4]	0.247	0.241	0.244
MCCA$_{us}$ [6]	0.253	0.257	0.255
DCCA$_{us}$ [7]	0.232	0.238	0.235
DCCAE$_{us}$ [9]	0.227	0.233	0.230
PLS$_{us}$ [5]	0.350	0.361	0.356
ACMR$_{s}$ [26]	0.363	0.312	0.338
DSCMR$_{s}$ [30]	0.360	0.369	0.364
SDML$_{s}$ [31]	0.369	0.374	0.372
Dual-Path$_{us}$ [16]	0.394	0.407	0.401
SCL$_{us}$ (ours)	**0.407**	**0.421**	**0.414**

data. Meanwhile, our SCL$_{ss}$ significantly outperforms its competitors with different proportions (e.g., 5%, 10%, 30%) of labeled data, which shows that our method can make good use of the self-supervised information including the correlation between single modality feature space and common representation space as well as the contrastive discriminative correlation in the unlabeled data, thereby alleviating the dependence on the labeled data.

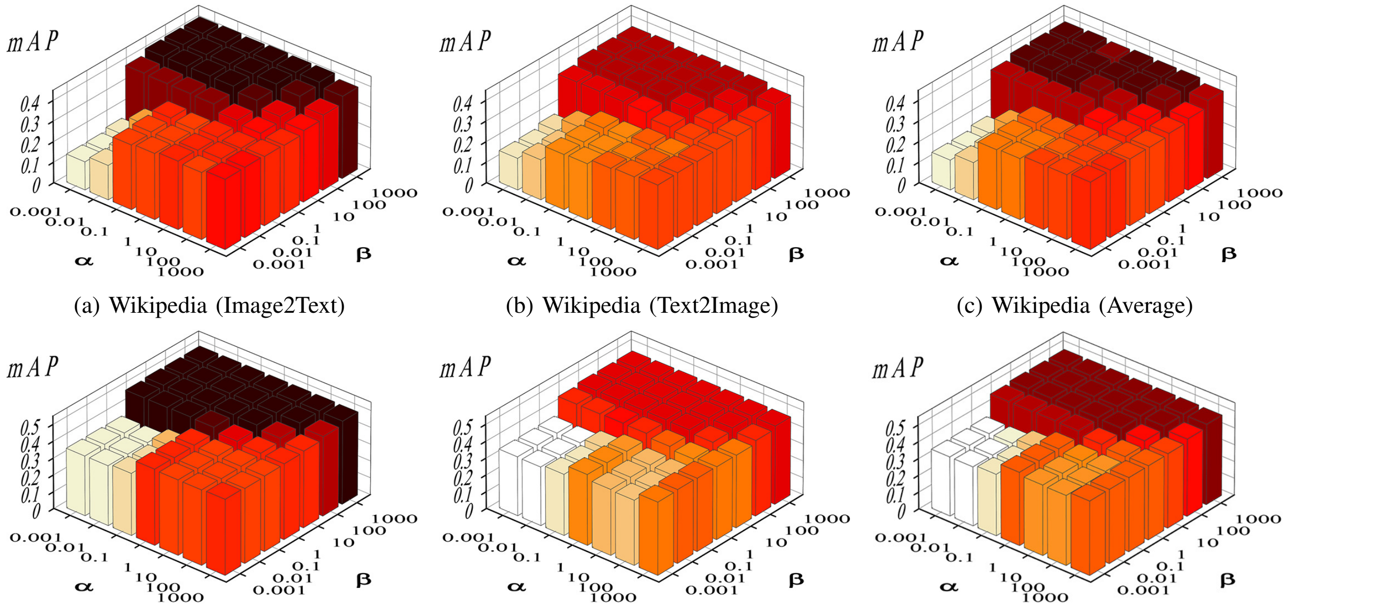
C. Generalization on Unseen Dataset

Since it is important for the model to generalize to out-of-domain data, we further conducted experiments to evaluate the generalization ability of our method. The evaluation settings are as follows. We first pre-trained a model on one dataset and then measured its performance on another dataset that is unseen in the training. Specifically, we utilized the MIRFlickr dataset [53] for training and tested models on the NUS-WIDE-10K test split. The results are shown in Table IV. From this table, we can see that our model is superior not only to existing unsupervised methods but also to existing supervised methods. This is because our model does not rely on the category label and is better able to distinguish between similar and dissimilar samples, thus being more generalized to unseen datasets.

D. Ablation Study

In this section, we performed the ablation study to investigate the effectiveness of each component of our method, including modality-invariant loss, contrastive discriminative loss, cross-modal mutual information loss, and weight-sharing scheme with the following four different variants.

- SCL-1 is a variant of our proposed SCL, which only uses Weight-Sharing (WS) strategy and minimizes the modality-invariant loss on paired instances.



(a) Wikipedia (Image2Text)

(b) Wikipedia (Text2Image)

(c) Wikipedia (Average)

(d) NUS-WIDE-10K (Image2Text)

(e) NUS-WIDE-10K (Text2Image)

(f) NUS-WIDE-10K (Average)

Fig. 5: Retrieval performance of SCL with different values of trade-off parameters α and β on Wikipedia and NUS-WIDE-10K.

- SCL-2 is a variant of our SCL without cross-modal mutual information loss.
- SCL-3 is a variant without the contrastive discrimination function.
- SCL-4 is a variant that simultaneously optimizes $\mathcal{L}_{CM}$, $\mathcal{L}_{MI}$, and $\mathcal{L}_{CD}$ but without the weight-sharing scheme.

For a fair comparison, all variants employ the same basic network and parameter settings as our method. The performance comparison of our proposed SCL and its variants on the MS-COCO dataset is shown in Table V. The difference between SCL-1 and SCL-2 is whether to use the contrastive discrimination loss. We can see that SCL-2 significantly surpasses the SCL-1, which indicates the effectiveness of modeling the correlation among intra- and inter-modality samples. SCL-3 adds the cross-modal mutual information loss on the basis of SCL-1, which achieves better performance than SCL-1 by maximizing the mutual information between the input and the output of different modality-specific projectors. Under the constraint of this self-supervised information, the network is prevented from learning arbitrary representations. Next, comparing SCL-4 with our full SCL, the performance decreases by 1.7%, which explains the effectiveness of sharing the weights of two subnetworks. In the end, compared to the other variants, we observed that our proposed SCL achieves the best performance, indicating that simultaneously optimizing the $\mathcal{L}_{CM}$, $\mathcal{L}_{MI}$, and $\mathcal{L}_{CD}$ in our model performs better than optimizing only one of them. To sum up, the contribution of each component enables our SCL to learn discriminative and modality-invariant representations as well as preserve the intrinsic structure of multimedia data.

TABLE V: Ablation study on the contributions of CM, CD, MI loss and WS strategy on the MS-COCO dataset.

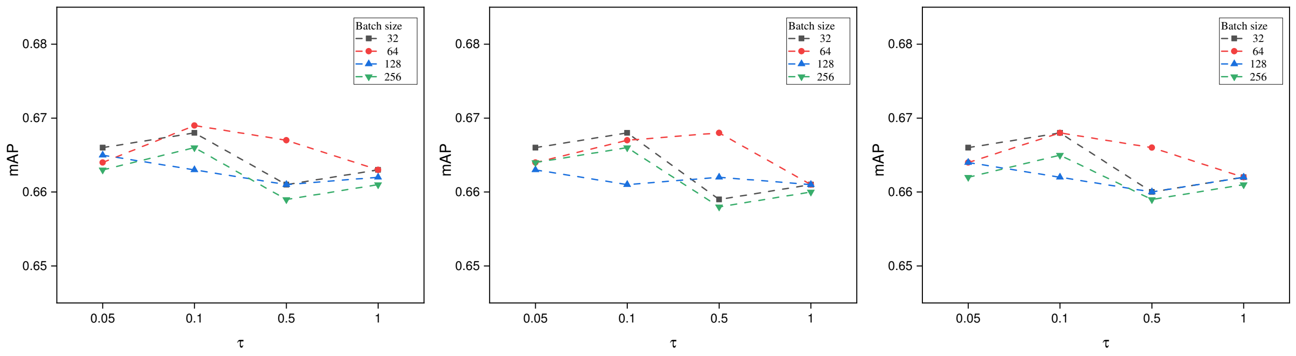
Method	Loss and Strategy				Metrics		
	CM	CD	MI	WS	Img2Txt	Txt2Img	Avg.
SCL-1	✓			✓	0.355	0.387	0.371
SCL-2	✓	✓		✓	0.654	0.650	0.652
SCL-3	✓		✓	✓	0.634	0.632	0.633
SCL-4	✓	✓	✓		0.645	0.647	0.646
SCL	✓	✓	✓	✓	**0.668**	**0.668**	**0.668**

E. Parametric Sensitivity Analysis

In this section, we carried out experiments to explore the impact of the parameter τ, the batch size N and the trade-off parameters α and β. To analyze how the parameters α and β affect the retrieval performance, we fixed the temperature parameter τ and the batch size N and tuned the parameters α and β by searching the grid of [0.001, 0.01, 0.1, 1, 10, 100, 1000] on the validation set. The results of α and β on Wikipedia and NUS-WIDE-10K are shown in Figure 5. We can observe that the results are stable when α varies in the range of [0.001, 0.01, 0.1, 1] and β varies in the range of [1, 10, 100, 1000]. In a similar way, we tested the importance of the batch size N and temperature τ with different values when fixing the parameters α and β. Figure 6 shows that our SCL is not sensitive to N and τ with the regularization effect of intrisic structure preservation.

F. Feature Visualization

In Figure 7, we visualized the learned representaions for the testing data on Wikipedia by using t-SNE [54]. The results of the image samples represented by 4,096-dimensional (VGGNet) features and the text samples represented by 5,000-dimensional (BoW) features are displayed in Figure

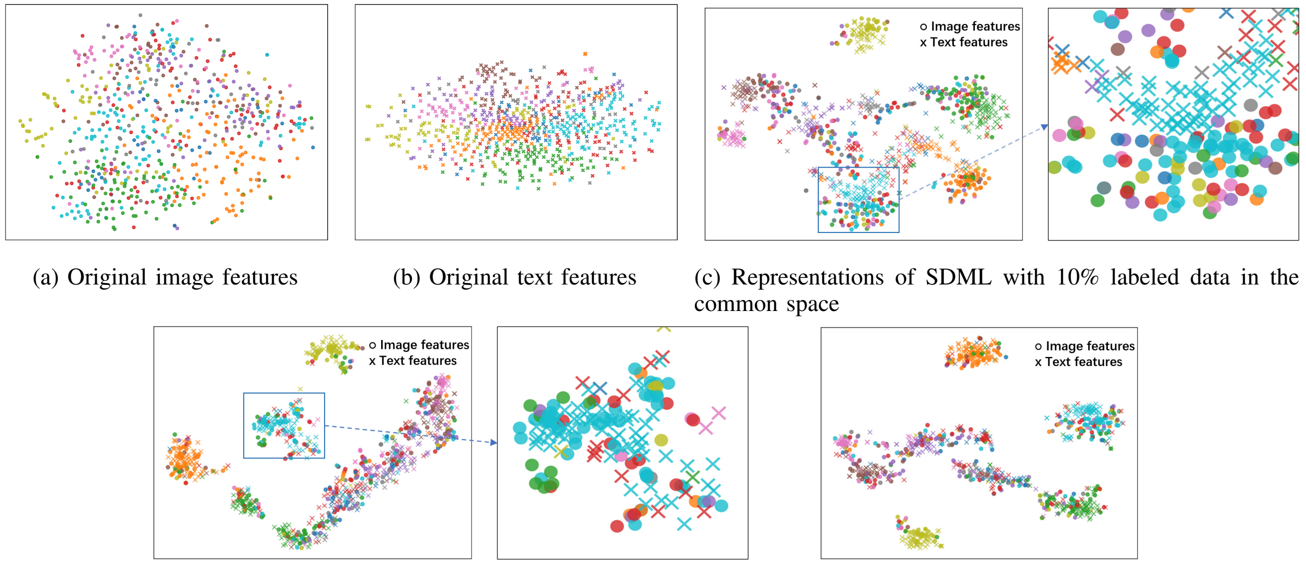


(a) Image2Text

(b) Text2Image

(c) Average

Fig. 6: Retrieval performance of SCL with different values of temperature τ and batch size N on the MS-COCO dataset.



(a) Original image features

(b) Original text features

(c) Representations of SDML with 10% labeled data in the common space

(d) Representations of SCL without labeled data in the common space

(e) Representations of SCL with 10% labeled data in the common space

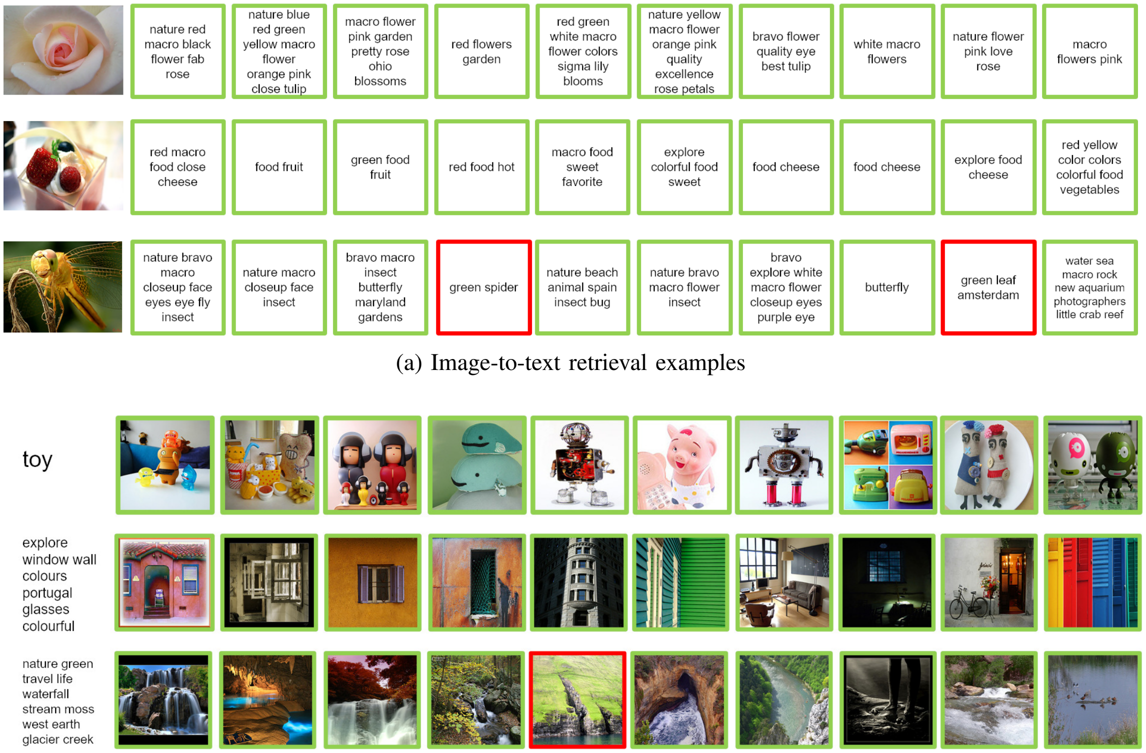
Fig. 7: t-SNE visulization for the testing data on the Wikipedia dataset. The circles denote the image features, and the x symbols describe the text features. The features of the same semantic category are marked with the same color.

7 (a) and Figure 7 (b), respectively. Figure 7 (c) shows the image and text representations of SDML trained with 10% labeled data in the common space. Figure 7 (d) and Figure 7 (e) display the learned common representations of our SCL without labeled data and with 10% labeled data, respectively. From these figures, we can see that our SCL with contrastive discrimination function is able to model the relationship among intra- and inter-modality samples, which effectively clusters similar samples with the same semantic category into one group and separates the dissimilar samples with different classes. In addition, compared with the representations obtained by the supervised method SDML, which is trained using 10% labeled data, we can find that there exist clear boundaries between image and text representations in Figure 7 (c) while the image and text modality are completely mixed together in Figure 7 (d). It indicates that our method can effectively eliminate the modality gap so as to obtain the modality-agnostic representations. Moreover, comparing Figure 7 (e) with Figure 7 (c), we can see that the representation of SCL trained with 10% labeled data is more discriminative and modality-invariant than the representations of SDML with 10% labeled data, which shows the superiority of our method.

G. *Qualitative Results*

We displayed several examples on the NUS-WIDE-10K test split for image-to-text retrieval and text-to-image retrieval in Figure 8. In the first two lines of Figure 8 (a) and (b), we can see that our model can successfully retrieve relevant images or texts. In Figure 8 (a), although false cases indicate the different classes from the query, they have similar semantics. In Figure 8 (b), our method can retrieve by single or multiple tags. The retrieval results are highly relevant to the query.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3152086, IEEE Transactions on Multimedia


nature red macro black flower fab rose
nature blue red green yellow macro flower orange pink close tulip
macro flower pink garden pretty rose ohio blossoms
red flowers garden
red green white macro flower colors sigma lily blooms
nature yellow macro flower orange pink quality excellence rose petals
bravo flower quality eye best tulip
white macro flowers
nature flower pink love rose
macro flowers pink
red macro food close cheese
food fruit
green food fruit
red food hot
macro food sweet favorite
explore colorful food sweet
food cheese
food cheese
explore food cheese
red yellow color colors colorful food vegetables
nature bravo macro closeup face eyes eye fly insect
nature macro closeup face insect
bravo macro insect butterfly maryland gardens
green spider
nature beach animal spain insect bug
nature bravo macro flower insect
bravo explore white macro flower closeup eyes purple eye
butterfly
green leaf amsterdam
water sea macro rock new aquarium photographers little crab reef
toy
explore window wall colours portugal glasses colourful
nature green travel life waterfall stream moss west earth glacier creek

(a) Image-to-text retrieval examples

(b) Text-to-image retrieval examples

Fig. 8: Image-to-text and text-to-image retrieval results on NUS-WIDE-10K. The matched images are annotated in green boxes, and the false ones are in red.

Although the false case belongs in a different category, it can be searched according to its semantics.

V. Conclusion and Future Work

In this paper, we propose a new unsupervised cross-modal retrieval method. It aims to retrieve relevant samples from different modalities without label information. To achieve this goal, we need to overcome the following three difficulties. One is to narrow the media gap among different modalities, the second is to introduce a supervisory signal to preserve the cross-modal correlation structure of a mass of unlabeled data, and the third is to learn discriminative representations where similar samples are close while dissimilar samples are separated. Therefore, on the one hand, we incorporate the prior knowledge in the input as a supervisory signal to guide the learning procedure by maximizing the mutual information between the original modality-specific features and the common representations. On the other hand, we exploit the contrastive discrimination function to model the correlation among intra- and inter-modality instances to learn discriminative representations. Moreover, to further eliminate the media gap, we use a weight-sharing scheme and minimize the modality-invariant loss in the common space. The extensive experiments conducted on three benchmark datasets demonstrate that our proposed method can effectively address these problems and achieve competitive results compared with 13 state-of-the-art approaches. In the future, we will explore more self-supervised frameworks and applications in cross-modal scenarios. Besides, we plan to incorporate the hashing module into our model to achieve efficient retrieval.

References

[1] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[2] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus, “Probabilistic embeddings for cross-modal retrieval,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 8415–8424.

[3] Y. Peng, X. Huang, and Y. Zhao, “An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2372–2385, 2018.

[4] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[5] R. Rosipal and N. Krämer, “Overview and recent advances in partial least squares,” in *International Statistical and Optimization Perspectives Workshop” Subspace, Latent Structure and Feature Selection”*. Springer, 2005, pp. 34–51.

[6] J. Rupnik and J. Shawe-Taylor, “Multi-view canonical correlation analysis,” in *Proceedings of the Conference on Data Mining and Data Warehouses*, 2010, pp. 1–4.

[7] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in *Proceedings of the International Conference on Machine Learning*. PMLR, 2013, pp. 2284–2292.

[8] F. Feng, X. Wang, and R. Li, “Cross-modal retrieval with correspondence autoencoder,” in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 7–16.

[9] W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multi-view representation learning,” in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 1083–1092.

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

[11] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, “Triplet-based deep hashing network for cross-modal retrieval,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018.

[12] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, “Progressive cross-modal semantic network for zero-shot sketch-based image retrieval,” *IEEE Transactions on Image Processing*, vol. 29, pp. 8892–8902, 2020.

[13] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, “Multi-task consistency-preserving adversarial hashing for cross-modal retrieval,” *IEEE Transactions on Image Processing*, vol. 29, pp. 3626–3637, 2020.

[14] A. Sharma and D. W. Jacobs, “Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*. IEEE, 2011, pp. 593–600.

[15] F. Yan and K. Mikolajczyk, “Deep correlation for matching images and text,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 441–450.

[16] Y.-d. Shen, “Dual-path convolutional image-text embeddings with instance loss,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–23, 2020.

[17] W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, “Integrating information theory and adversarial learning for cross-modal retrieval,” *Pattern Recognition*, vol. 117, p. 107983, 2021.

[18] X. Zhai, Y. Peng, and J. Xiao, “Learning cross-media joint representation with sparse and semisupervised regularization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2014.

[19] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, “Semi-supervised cross-media feature learning with unified patch graph regularization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 583–596, 2016.

[20] L. Zhang, B. Ma, J. He, G. Li, Q. Huang, and Q. Tian, “Adaptively unified semi-supervised learning for cross-modal retrieval,” in *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI Press, 2017, pp. 3406–3412.

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

[22] F. Zou, X. Bai, C. Luan, K. Li, Y. Wang, and H. Ling, “Semi-supervised cross-modal learning for cross modal retrieval and image annotation,” *World Wide Web*, vol. 22, no. 2, pp. 825–841, 2019.

[23] D. Mandal, P. Rao, and S. Biswas, “Semi-supervised cross-modal retrieval with label prediction,” *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2345–2353, 2020.

[24] P. Hu, H. Zhu, X. Peng, and J. Lin, “Semi-supervised multi-modal learning with balanced spectral decomposition,” in *Proceedings of The AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 99–106.

[25] Y. Peng, X. Huang, and J. Qi, “Cross-media shared representation by hierarchical learning with multiple deep networks,” in *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 3846–3853.

[26] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, “Adversarial cross-modal retrieval,” in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2017, pp. 154–162.

[27] Y. Peng, J. Qi, X. Huang, and Y. Yuan, “CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network,” *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2017.

[28] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, “Deep adversarial metric learning for cross-modal retrieval,” *World Wide Web*, vol. 22, no. 2, pp. 657–672, 2019.

[29] Y. Peng and J. Qi, “CM-GANs : Cross-modal generative adversarial networks,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1, pp. 1–24, 2019.

[30] L. Zhen, P. Hu, X. Wang, and D. Peng, “Deep supervised cross-modal retrieval,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 10 386–10 395.

[31] P. Hu, D. Peng, L. Zhen, and P. Liu, “Scalable deep multimodal learning for cross-modal retrieval,” in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2019, pp. 635–644.

[32] X. He, Y. Peng, and L. Xie, “A new benchmark and approach for fine-grained cross-media retrieval,” in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2019, pp. 1740–1748.

[33] P. Kang, Z. Lin, Z. Yang, X. Fang, Q. Li, and W. Liu, “Deep semantic space with intra-class low-rank constraint for cross-modal retrieval,” in *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 2019, pp. 226–234.

[34] X. Wang, P. Hu, L. Zhen, and D. Peng, “Drsl: Deep relational similarity learning for cross-modal retrieval,” *Information Sciences*, vol. 546, pp. 298–311, 2021.

[35] S. Qian, D. Xue, H. Zhang, Q. Fang, and C. Xu, “Dual adversarial graph neural networks for multi-label cross-modal retrieval,” in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2021, pp. 2440–2448.

[36] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, pp. 1735–1742.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 815–823.

[38] K. Sohn, “Improved deep metric learning with multi-class n-pair loss objective,” in *Proceedings of the International Conference on Neural Information Processing Systems*. NIPS, 2016, pp. 1849–1857.

[39] S. Cheng, B. Zhong, G. Li, X. Liu, Z. Tang, X. Li, and J. Wang, “Learning to filter: Siamese relation network for robust tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 4421–4431.

[40] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 3733–3742.

[41] A. Van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 9729–9738.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[44] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, “Mutual information neural estimation,” in *Proceedings of the International Conference on Machine Learning*. PMLR, 2018, pp. 531–540.

[45] R. Devon Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization,” in *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2019.

[46] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics*. JMLR, 2010, pp. 297–304.

[47] T. Chen, Y. Sun, Y. Shi, and L. Hong, “On sampling strategies for neural network-based collaborative filtering,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 767–776.

[48] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, “A new approach to cross-modal multimedia retrieval,” in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2010, pp. 251–260.

[49] T. S. 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 *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2009, pp. 1–9.

[50] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[51] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the International Conference on Machine Learning*. Omnipress, 2010, pp. 807–814.

[52] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2015.

[53] M. J. Huiskes and M. S. Lew, “The mir flickr retrieval evaluation,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*. ACM, 2008, pp. 39–43.

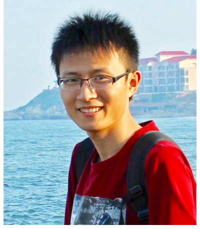
[54] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, pp. 2579–2605, 2008.

Yaxin Liu received her B.E. degree from Qingdao University in 2019. She is currently a master with the School of Computing Science and Technology, Shandong University. Her research interests include cross-modal retrieval and semi-supervised learning.

Jianlong Wu (Member, IEEE) received his B.E. and Ph.D. degrees from Huazhong University of Science and Technology in 2014 and Peking University in 2019, respectively. He is currently an assistant professor with the School of Computer Science and Technology, Shandong University. His research interests lie primarily in computer vision and machine learning, especially weakly supervised learning. He has published more than 20 research papers in top journals and conferences, such as TIP, ICML, NeurIPS, and ICCV. He serves as a Senior Program Committee Member of IJCAI 2021, an area chair of ICPR 2020, and a reviewer for many top journals and conferences, including TPAMI, IJCV, ICML, and ICCV.

Leigang Qu received the B.E. degree from Ocean University of China in 2019. He is currently a master with the School of Computing Science and Technology, Shandong University. His research interests include multimedia computing and information retrieval.

Tian Gan is currently an Associate Professor with the School of Computer Science and Technology, Shandong University. She received her B.S. from East China Normal University in 2010, and the Ph.D. degree from National University of Singapore, Singapore, in 2015. She was a Research Scientist in Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR). Her research interests include social media marketing, video understanding, and multimedia computing.



Jianhua Yin is currently an Associate Professor with the School of Computer Science and Technology, Shandong University. He received his B.E. degree from Xidian University in 2012, and the Ph.D. degree from Tsinghua University in 2017. His research interests include data mining and machine learning.

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University and the Ph.D. degree from the National University of Singapore (NUS). After Ph.D. degree, he continued his research in NUS as a Research Fellow for three years. He is currently a Professor with Shandong University and the Dean with Shandong AI institute. He has coauthored more than 200 articles and four books, received more than 11,000 Google Scholar citations as of August 2021. His research interests lie primarily in multimedia computing and information retrieval. He received many awards, such as ACM MM and SIGIR Best Paper Honorable Mention in 2019, SIGMM Rising Star in 2020, TR35 China 2020, DAMO Academy Young Fellow in 2020, and SIGIR Best Student Paper in 2021. He is an AE of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, ACM ToMM, and Information Sciences. Meanwhile, he is the Area Chair of ACM MM 2018-2021.