MCCN: Multimodal Coordinated Clustering Network for Large-Scale Cross-modal Retrieval

Zhixiong Zeng, Ying Sun, Wenji Mao*
SKL-MCCS, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
{zengzhixiong2018,sunying2019,wenji.mao}@ia.ac.cn

ABSTRACT

Cross-modal retrieval is an important multimedia research area which aims to take one type of data as the query to retrieve relevant data of another type. Most of the existing methods follow the paradigm of pair-wise learning and class-level learning to generate a common embedding space, where the similarity of heterogeneous multimodal samples can be calculated. However, in contrast to large-scale cross-modal retrieval applications which often need to tackle multiple modalities, previous studies on cross-modal retrieval mainly focus on two modalities (*i.e.*, text-image or text-video). In addition, for large-scale cross-modal retrieval with modality diversity, another important problem is that the available training data are considerably modality-imbalanced. In this paper, we focus on the challenging problem of *modality-imbalanced cross-modal retrieval*, and propose a Multimodal Coordinated Clustering Network (MCCN) which consists of two modules, Multimodal Coordinated Embedding (MCE) module to alleviate the imbalanced training data and Multimodal Contrastive Clustering (MCC) module to tackle the imbalanced optimization. The MCE module develops a data-driven approach to coordinate multiple modalities via multimodal semantic graph for the generation of modality-balanced training samples. The MCC module learns class prototypes as anchors to preserve the pair-wise and class-level similarities across modalities for intra-class compactness and inter-class separation, and further introduces intra-class and inter-class margins to enhance optimization flexibility. We conduct experiments on the benchmark multimodal datasets to verify the effectiveness of our proposed method.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**.

KEYWORDS

Cross-modal retrieval; multimodal coordinated embedding; multimodal contrastive clustering; prototype learning

*Corresponding author

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475670

ACM Reference Format:
Zhixiong Zeng, Ying Sun, Wenji Mao. 2021. MCCN: Multimodal Coordinated Clustering Network for Large-Scale Cross-modal Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475670

1 INTRODUCTION

Recent years have seen a rapidly growing trend of multimodal data describing the same topic in social media, including image, text, video, audio and 3D model. Cross-modal retrieval task as a core area of information retrieval aims to take one type of data as the query to retrieve relevant data of another type [47]. It is a fundamental research task and has a number of practical applications in domains such as image retrieval [42], image caption [33], video recommendation [38], automatic story generation [17] and so forth. However, it is well known that the inconsistent representation and distribution of distinct modalities causing the heterogeneous gap across modalities, which makes cross-modal similarity cannot be directly computed [8]. Therefore, the main challenge of cross-modal retrieval is how to measure the similarity between the samples from different modalities.

A common approach to bridge the heterogeneous gap is to find a common embedding space by learning modality-specific transformations, where the cross-modal similarity can be directly compared [11, 34]. To preserve the multimodal semantic structure in the common embedding space, most of the existing cross-modal retrieval methods follow the paradigm of pair-wise learning and class-level learning. The former usually leverages the triplet ranking loss function to optimize the similarity of paired samples (from the same or different modalities), which encourages the similarity of the related pairs larger than the unrelated ones [24, 34, 39]. The latter basically learns to classify each training sample to its target class with a classification loss [9, 11, 26, 34], which essentially learns a semantic partition of the common embedding space. Recent work focuses on introducing adversarial training process to learn modality invariant embeddings with a modality classifier and achieves superior performances [24, 39].

However, most existing methods on cross-modal retrieval only consider two modalities, usually text and image (or text and video) modalities. For large-scale cross-modal retrieval applications, it is often needed to handle multiple modalities (*e.g.*, image, text, video, audio and 3D model). In addition, for large-scale cross-modal retrieval with multiple modalities, another important problem is that the available training data are considerably *modality-imbalanced*, due to the discrepancies between different modalities in the difficulty of data collection and labor annotation. The modality-imbalanced

training data inevitably lead to insufficient learning over multiple modalities. Although several methods have been proposed to improve the modality-imbalance problem in the context of two modalities, most of them concentrate on introducing new constraints [6, 11] or parameters [45] to generate modality-balanced training samples, which would lead to slow convergence and poor performance when extended to multiple modalities.

In this paper, we focus on the challenging problem of *modality-imbalanced cross-modal retrieval* in large-scale applications with multiple modalities. In addition to the issue of alleviating modality-imbalanced training data for effective learning over multiple modalities, another research issue is the imbalanced optimization. Due to the huge heterogeneous modality gap between samples from different modalities, previous work mainly optimizes the similarity of paired samples from different modalities (*i.e.*, heterogeneous similarity) yet ignores the paired samples from the same modality (*i.e.*, homogeneous similarity). The imbalanced optimization will be further exacerbated when extended to multiple modalities, and result in worse preservation of the semantic structure for multimodal data.

To tackle the above challenges, we propose a Multimodal Coordinated Clustering Network (MCCN) for modality-imbalanced cross-modal retrieval problem with multiple modalities. The MCCN consists of two modules, Multimodal Coordinated Embedding module (MCE) and Multimodal Contrastive Clustering module (MCC). The MCE module employs a data-driven approach to coordinate multiple modalities for generating modality-balanced training samples, based on the intuition that multimodal samples belonging to the same category usually share the same semantic content but follow different distributions. It randomly walks on a predefined multimodal semantic graph to stochastically make transitions between embeddings over different modalities to generate modality-balanced samples with consistent labels. Benefited from the MCE module, the independent modality-specific networks are forced to explicitly consider aligning multimodal samples, which is essential to the deep understanding of multimodal content. Inspired by prototype learning, the MCC module takes class prototypes as anchors to integrate the semantic similarities of heterogeneous and homogeneous samples. It then preserves the pair-wise similarity and class-level similarity across multiple modalities for intra-class compactness and inter-class separation. To this end, we present a novel contrastive clustering loss to jointly learn the common embedding space and class prototypes, and further introduce an intra-class margin and an inter-class margin to enhance the optimization flexibility.

The main contributions of our work are as follows:

- For large-scale retrieval task with multiple modalities, we identify the important problem of modality-imbalanced cross-modal retrieval, and propose a novel multimodal coordinated clustering network MCCN to tackle this problem.
- To alleviate the imbalanced training data, we present a multimodal coordinated embedding module MCE to stochastically transit between embeddings over different modalities with consistent labels.
- To tackle the imbalanced optimization, we propose a multimodal contrastive clustering module MCC that jointly learns class prototypes to minimize the intra-class variations and meanwhile maximize the inter-class variations.
- We conduct experiments on the benchmark multimodal datasets with multiple modalities, and experimental results demonstrate the effectiveness of our method.

2 RELATED WORK

2.1 Cross-modal Retrieval

The key challenge of cross-modal retrieval is to bridge the heterogeneity gap and find a common embedding space in which the semantic similarity of multimodal samples can be compared. The typical methods can be divided into two main categories, traditional methods and deep learning based methods. Traditional methods use statistical correlation analysis to transform multimodal samples to common representations by maximizing pair-wise correlations, in which the representative methods are CCA [7] and its extensions [1, 5, 37]. Another class of traditional methods utilizes graph regularization to learn the statistical correlations between multimodal samples. The representative work [46] proposes the JRL model which employs both graph regularization and semi-supervised information to perform joint representation learning.

Deep learning based methods take advantage of the powerful representation capabilities of deep neural networks to learn the common representation for multimodal samples, which mainly focus on learning modality-specific transformations to find a common embedding space for cross-modal retrieval. To preserve the multimodal semantic associations in the common embedding space, most of the existing cross-modal retrieval methods follow the paradigm of preserving pair-wise similarity and class-level similarity. Angrew *et al.* [1] uses two modality-specific transformations to nonlinearly project image and text into a latent common embedding space, where the projected embeddings are highly correlated. In [3], Feng *et al.* propose a cross-modal auto-encoder that optimize pair-wise constraints at different levels to preserve the semantic associations. As one item in one modality may correspond to more than one semantically related items in another modality, Wang *et al.* [34] propose a triplet constraint that optimize coupled samples belonging to the same category across different modalities. The triplet constraint can effectively minimize the gap among the representations of all semantically related samples from different modalities, and some new variants have been proposed in subsequent cross-modal retrieval methods [24, 39, 47]. To gain discriminative embedding space, some methods further exploits label information to capture class-level similarity structure of multimodal samples [22, 39, 47].

The aforementioned methods mainly focus on cross-modal retrieval with two modalities (typically image and text), ignoring the growing need of large-scale cross-modal retrieval with multiple modality data involving image, text, audio, video and 3D model. Although the above methods can be directly extended to multiple modalities, they will inevitably encounter the modality-imbalanced training data and imbalanced optimization issues. Some early methods for learning embeddings of multiple modalities have been proposed based on linear projections, which jointly learn multiple view-specific transformations to maximize cross-modal similarity in a non-pairwise manner [13, 29]. To capture the complex nonlinear correlations, some deep lerning based methods [35, 40]

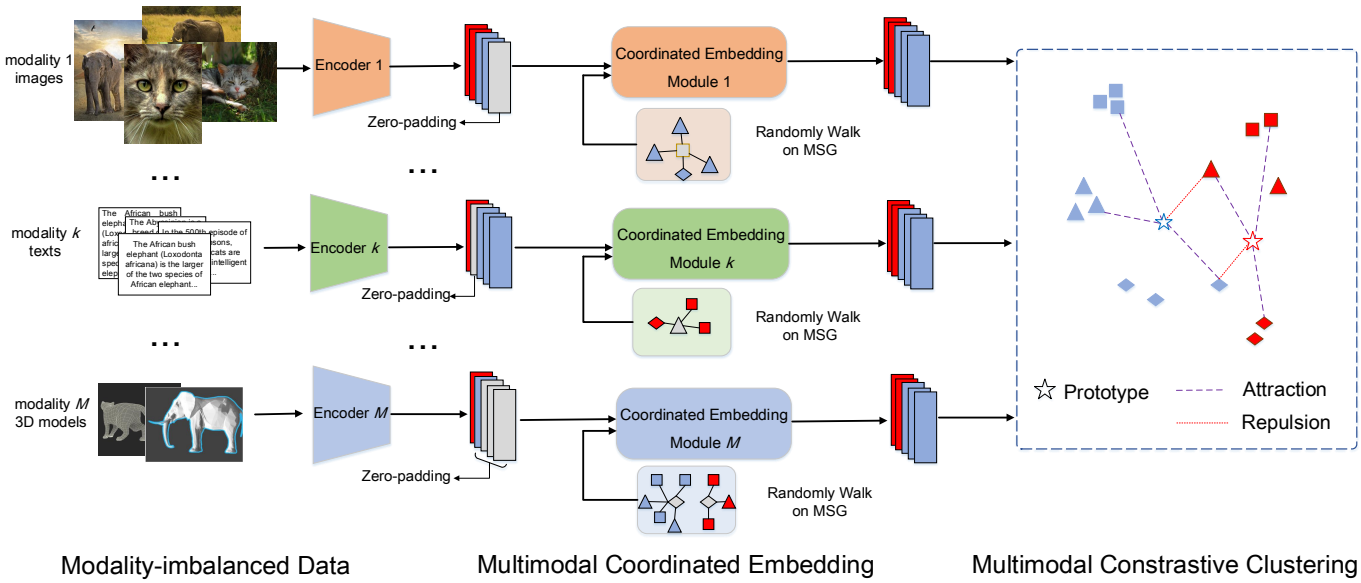


Figure 1: Overall architecture of the proposed MCCN. It first pads the modality-imbalanced training data to a fixed length, and further connects samples belonging to the same category from different modalities to generate multimodal semantic graph (MSG). Then the MCE module randomly walks on the predefined MSG to make transition between padded embeddings and real embeddings with consistent semantic label, which generates modality-balanced training data to avoid insufficient learning. Finally, the MCC module jointly learns class prototypes to minimize the intra-class variations as well to maximize inter-class variations in the common embedding space, which preserves cross-modal semantic correlations for effective retrieval. As an example, the square represents the embedding of image modality, the triangle represents the embedding of text modality, and the rhombus represents the embedding of 3D model modality. Different colors denote different semantic categories.

have been proposed to learn the optimal latent subspace shared by multiple modalities. More recent work [8] employs separate modality-specific sub-networks to preserve the semantic discrimination in a predefined embedding space, yet it fails to take the cross-modal similarity into consideration.

In essence, these methods straightforwardly extend embedding space learning methods from two modalities to multiple modalities, without considering the modality-imbalanced training data and imbalanced optimization issues. In our work, we identify these two important issues in modality-imbalanced cross-modal retrieval for large-scale applications, and propose the multimodal coordinated embedding module and multimodal contrastive clustering module to alleviate these non-trivial issues.

2.2 Prototype Learning

Prototype learning is a classical and representative method in pattern recognition and machine learning [43]. The earliest prototype learning method is the k-nearest-neighbor (K-NN), which calculates the distance from all samples and obtains the nearest k samples as the classification basis. To reduce the heavy burden of computation requirement and storage space of K-NN, many prototype learning methods have been proposed to select or synthesize prototypes that better represent the class distributions [10, 30]. These methods are also effective to improve the classification accuracy, in which Learning Vector Quantization (LVQ) [15] is a representative method that offers intuitive and simple, yet powerful learning capacity in supervised learning. The LVQ has been widely studied and has many variants, which can be divided into two main categories according to the updating methods of prototypes. One category concentrates on designing suitable updating conditions and rules to learn the prototypes, while the performance is limited by the initialization of prototypes and the selection of informative patterns [4, 18]. The other category learns the prototypes in a parameter optimization way, by defining loss functions with regard to the prototypes and learning the prototypes through optimizing the loss functions [43]. A detailed review and evaluation of the prototype based learning methods can be found in [19].

As many prototype learning algorithms based on loss minimization are promising to give better performance [19], in this work, we focus on learning class prototypes through optimizing the loss functions. We propose a novel multimodal contrastive clustering loss to jointly update multimodal embeddings and class prototypes, which aims to minimize intra-class variations and meanwhile maximize inter-class variations for multimodal samples. More importantly, the class prototypes integrate the distribution information of heterogeneous and homogeneous samples, thus can effectively alleviate the modality-imbalanced optimization caused by heterogeneous sample pairs. To the best of our knowledge, this is the first work on combining the prototype learning with multimodal embedding

space learning to perform large-scale cross-modal retrieval over multiple modalities.

3 PROPOSED METHOD

In this section, we will first describe the problem formulation of the cross-modal retrieval task with imbalanced multiple modalities. Figure 1 illustrates the overall architecture of our proposed method MCCN. We then will introduce the proposed multimodal coordinated embedding module (MCE) that randomly walks on a predefined multimodal semantic graph (MSG) to generate modality-balanced training data with consistent labels. Finally, we will describe the proposed multimodal contrastive clustering module (MCC) that jointly learns class prototype to minimize the intra-class variations as well to maximize the inter-class variations across multiple modalities.

3.1 Problem Formulation

Suppose that we have a collection of data from M different modalities, with $\{x_i^1, x_i^2, ..., x_i^M\}$ representing the same underlying content or objects. For example, the image, text, video, audio and 3D model are often used to describe the same topic. For each modality, we can denote the i-th sample of the z-th modality as x_i^z, and denote the set containing all the n_z samples of the z-th modality as $\Phi^z = \{x_1^z, x_2^z, ..., x_{n_z}^z\}$. The corresponding label vector of the z-th modality is represented as $Y^z = [y_1^z, y_2^z, ..., y_{n_z}^z]$, and $y_i^z \in \{1, 2, ..., C\}$ is the index of category, where C is the number of semantic categories. We then utilize a modality-specific encoder to extract the original embedding, denote as:

$$u_i^z = \mathbf{E}_z(x_i^z, \zeta_z) \in \mathcal{R}^{d_z} \tag{1}$$

where u_i^z denotes the original embedding in the modality-specific embedding space, $\mathbf{E}_z$ is the encoder of the $z-$th modality (*e.g.*, Bert model for text modality encoding, and VGG-net for image modality encoding), d_z is the dimension of the $z-$th embedding space, and ζ_z is the parameters pre-trained on other datasets. Here we fine-tune the image encoder pre-trained on ImageNet and the text encoder pre-trained on Wiki Corpus. Therefore, the original embedding matrix for $z-$th modality can be denoted as:

$$U^z = \{u_1^z, u_2^z, ..., u_{n_z}^z\} \in \mathcal{R}^{d_z \times n_z} \tag{2}$$

3.2 Multimodal Coordinated Embedding (MCE)

The goal of cross-modal retrieval is to learn a modality-specific transformation function for each modality to project them into a common embedding space. Since the training samples are considerably modality-imbalanced over multiple modalities, we propose a multimodal coordinated embedding module (MCE) to generate modality-balanced embeddings, which utilizes a predefined multimodal semantic graph to stochastically make transitions between embeddings over different modalities. Without losing generality, let us assume that modality k has sufficient samples in each category, which is usually text modality due to the convenience of data collection. Following [20], we first utilize the zero-padding operation to pad the multimodal features into a unified length, denoted as:

$$\overline{U}^z = \{\overline{u}_i^z\} = \{U^z, Q^z\} \in \mathcal{R}^{d_z \times n_k} \tag{3}$$

where $Q^z = \{q_l^z\} \in \mathcal{R}^{d_z \times (n_k - n_z)}$ is a all-zero matrix. The embeddings of U^z are real embeddings containing distribution information over different modalities, and Q^z consists of padding embeddings that aims to generate balanced multimodal samples with consistent semantic labels. Similarly, we pad the label vector Y^z as:

$$\overline{Y}^z = [\overline{y}_i^z] = [Y^z, H^z] \in \mathcal{R}^{n_k} \tag{4}$$

where $H^z = Y^k - Y^z$.

Then we employ the semantic labels $[\overline{Y}^z]_{z=1}^M$ to construct multimodal semantic graph that can be used to create stochastic multimodal training samples with consistent labels over multiple modalities. Specifically, two embeddings are connected with an edge in the multimodal semantic graph, if and only if they belong to the same category and come from different modalities. The connection matrix can be defined as:

$$e_{ij}^{z_1 z_2} = \begin{cases} 1, & \text{if } \overline{y}_i^{z_1} = \overline{y}_j^{z_2} \text{ and } z_1 \neq z_2 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Note that we do not connect samples of the same modality, which is motivated by the fact that embeddings of samples belonging to the same category and the same modality follow the same distribution [36, 40]. To avoid the transition between the padding embeddings, we set the edge between the padding embeddings to 0.

To generate transition probability from embeddings of the multimodal semantic graph, inspired by [41], we use a random walk with random restart and self-loop to fill out the transition probability table. We can define the transition probability $p(\overline{u}_i^{z_1}, \overline{u}_j^{z_2})$ as probability of transition from embedding $\overline{u}_i^{z_1}$ to embedding $\overline{u}_j^{z_2}$. To encourage the transition within padding embeddings and real embeddings belonging to the same category across different modalities, when $\overline{u}_i^{z_1}$ is connected with $\overline{u}_j^{z_2}$ but not with $\overline{u}_l^{z_3}$ in the multimodal semantic graph, we can set the ratio of $p(\overline{u}_i^{z_1}, \overline{u}_j^{z_2})$ and $p(\overline{u}_i^{z_1}, \overline{u}_l^{z_3})$ to be a constant greater than 1. Mathematically, we have:

$$\overline{u}_i^{z_1} \sim \overline{u}_j^{z_2}, \overline{u}_i^{z_1} \nsim \overline{u}_l^{z_3} \rightarrow p(\overline{u}_i^{z_1}, \overline{u}_j^{z_2}) / p(\overline{u}_i^{z_1}, \overline{u}_l^{z_3}) = \epsilon \tag{6}$$

where $\epsilon > 1$ and is a tuning parameter, $\sim$ denotes connect and $\nsim$ denotes not connect. The larger the value of ϵ, the more likely the padding embedding will be replaced by the real embedding belonging to the same category and from other modalities. We also have:

$$p(\overline{u}_i^{z_1}, \overline{u}_i^{z_1}) = \epsilon_0 \tag{7}$$

where ϵ_0 is the self-loop probability. Since the sum of transition probabilities from one embedding to other arbitrary embeddings is 1, *i.e.*, $\sum_{u \in \overline{U}} p(\overline{u}_i^{z_1}, u | \epsilon, \epsilon_0) = 1$, we can derive transition probabilities between any two embeddings to fill out the transition probability table $\mathcal{T}$. Based on the transition probability table $\mathcal{T}$, we can define a transition function Υ to generate the modality-balanced embeddings, denote as:

$$\widehat{U}^z = \Upsilon(\overline{U}^z, \mathcal{T}) \tag{8}$$

Finally, we can calculate the multimodal coordinated embeddings as:

$$H^z = f_z(\widehat{U}^z, \theta_z) \in \mathcal{R}^{d \times n} \tag{9}$$

where $H^z = \{h_1^z, h_2^z, ..., h_n^z\}$ is the learned embeddings of z–th modality in the common embedding space, θ_z is the trainable parameters of the z-th modality-specific transformation function, and d is the dimension of the common embedding space.

3.3 Multimodal Contrastive Clustering (MCC)

To preserve the semantic relationships of multimodal samples in the common embedding space, previous work mainly construct paired samples (can be from the same or different modalities) to minimize the inter-class similarity and maximize the intra-class similarity [24, 34, 39, 47]. However, these methods of constructing pair-wise constraints simultaneously optimizes the heterogeneous and homogeneous sample pairs, which may suffer modality-imbalanced optimization due to the modality gap. We find inspiration from prototype learning [43], which provide a gradually evolved class prototype to jointly learn the intra-class cluster and inter-class separation of multiple modalities. More importantly, the class prototypes integrate the distribution information of heterogeneous and homogeneous samples, which can effectively alleviate the modality-imbalanced optimization problem. For each class $c \in \{1, 2, ..., C\}$, we maintain a prototype vector $o_c \in \mathcal{R}^d$ in the common embedding space by calculating the mean of embeddings corresponding to the c-th category from all modalities. The prototype vector o_c is jointly trained by the prototype-based loss function. A natural way to jointly train multimodal embeddings and class prototype vector is to minimize the distance from the embedding to the corresponding prototype while maximizing the distance from the embedding to wrong prototypes, which is usually modeled as a contrastive clustering problem [12].

For embeddings of z-th modality, let us assume that there are K positive Euclidean distances from embeddings H^z to their corresponding prototype and L negative Euclidean distances from embeddings to wrong prototypes. We represent these positive distances as $\{d_\alpha^z\}(\alpha = 1, 2, ..., K)$ and negative distances $\{\hat{d}_\beta^z\}(\beta = 1, 2, ..., L)$, respectively. To minimize each positive distance d_α^z ae well as to maximize each d_β^z, we define the multimodal contrastive clustering loss as follows:

$$\mathcal{L} = \sum_{z=1}^{M}(log[1 + \sum_{\beta=1}^{L}\sum_{\alpha=1}^{K} exp(\lambda(d_\alpha^z - \hat{d}_\beta^z + \Delta))]) \tag{10}$$

where λ is a scale factor and Δ is a margin for class separation. Here we utilize the same scale factor and margin for each modality to reduce parameters. It iterates every distance to minimize the intra-class variations and maximize the inter-class variations. It aims to optimize $d_\alpha^z \to 0$ and $\hat{d}_\beta^z \to \Delta$.

To enhance the optimization flexibility, inspired by [32], we introduce the intra-class margin and inter-class margin, thus the final learning objective of MCCN can be formulated as:

$$\begin{aligned}\mathcal{L} &= \sum_{z=1}^{M}(log[1 + \sum_{\beta=1}^{L}\sum_{\alpha=1}^{K} exp(\lambda((d_\alpha^z - \Delta_p) - (\hat{d}_\beta^z - \Delta_n)))]) \\ &= \sum_{z=1}^{M}(log[1 + \sum_{\alpha=1}^{K} exp(\lambda(d_\alpha^z - \Delta_p)) \sum_{\beta=1}^{L} exp(-\lambda(\hat{d}_\beta^z - \Delta_n))])\end{aligned} \tag{11}$$

where Δ_p denotes the intra-class margin and Δ_n denotes the inter-class margin. For simplicity, we set $\Delta_p = m$ and $\Delta_n = 1 - m$, thus only preserve a single parameter m. We can see that our loss expects the positive distance $d_\alpha^z < m$ the negative distance $d_\beta^z > 1 - m$. We will evaluate the impact of λ and m in the experiment.

4 EXPERIMENT

In the experiments, we first compare our proposed MCCN with representative methods by performing cross-modal retrieval over multiple modalities. We then conduct cross-modal retrieval experiments specifically for image and text to further evaluate the performance of the proposed MCCN. Moreover, we conduct ablation studies to evaluate the impact of each component in MCCN. Finally, we conduct a detailed parameter analysis on the hyper-parameters of our method.

4.1 Experimental Setup

4.1.1 Datasets and Features. To verify the effectiveness of our proposed MCCN, we conduct experiments on two modality-imbalanced multimodal datasets, namely XMedia [26, 46] and XMediaNet [23, 25]. The Xmedia dataset is the benchmark dataset for cross-modal retrieval with multiple modalities. It contains samples of 5 modalities from 20 different semantic categories. The XmediaNet is a large-scale multimodal dataset with 200 semantic categories, we test the most common image, text and video modalities with imbalanced data. In additional, to evaluate the effectiveness of MCCN on conventional image-text retrieval, we also carry out experiments on benchmark Wikipedia [28] and Pascal [27] datasets. For Xmedia and XmediaNet datasets, the feature files are provided by the authors. For Wikipedia and Pascal datasets with two modalities, we use the pre-trained BERT [2] model to extract embeddings from text, and use the pre-trained VGG-19 [31] to extract embeddings from image. The detailed statistics of the four datasets are summarized in Table

Table 1: General statistics of the four datasets used in our experiments, where '/' in the second column denotes the number of train/valid/test set. The split of XMedia dataset strictly follows that in [8], and the split of other datasets strictly follows that in [47].

Dataset	Label	Modality	Instance	Feature
XMedia	20	Image	4000/500/500	4,096D VGG
		Text	4000/500/500	3,000D BoW
		Video	969/87/87	4,096D CNN
		Audio	800/100/100	29D MFCC
		3D-model	400/50/50	4,700d LightF
XMediaNet	200	Image	32000/4000/4000	4,096D VGG
		Text	32000/4000/4000	3,000D BoW
		Video	8000/1000/1000	4,096D CNN
Wikipedia	10	Image	2173/231/462	4,096D VGG
		Text		768d BERT
Pascal	20	Image	800/100/100	4,096D VGG
		Text		768d BERT

Table 2: Performance Comparison with representative methods on XMedia dataset.

Method	Query	Image				Text				Video				Audio				3D				Avg
	Target	Text	Video	Audio	3D	Image	Video	Audio	3D	Image	Text	Audio	3D	Image	Text	Video	3D	Image	Text	Video	Audio	
MvDA		0.795	0.516	0.482	0.548	0.786	0.537	0.506	0.582	0.483	0.477	0.295	0.387	0.473	0.480	0.354	0.425	0.533	0.547	0.412	0.358	0.499
JFSSL		0.859	0.582	0.526	0.568	0.867	0.608	0.551	0.610	0.527	0.509	0.326	0.412	0.511	0.553	0.396	0.463	0.584	0.571	0.443	0.406	0.544
JLSLR		0.864	0.604	0.541	0.583	0.874	0.614	0.563	0.636	0.541	0.536	0.338	0.430	0.556	0.572	0.421	0.486	0.602	0.620	0.472	0.412	0.563
SDML		0.887	0.645	0.585	0.655	0.904	0.674	0.604	0.696	0.586	0.587	0.379	0.488	0.575	0.608	0.425	0.517	0.667	0.674	0.501	0.416	0.604
MCCN		0.908	0.654	0.614	0.665	0.915	0.677	0.627	0.711	0.601	0.592	0.393	0.499	0.586	0.621	0.442	0.526	0.675	0.686	0.514	0.425	0.617

1. To ensure fair comparison, all the compared methods adopt the same features used in our method.

4.1.2 Evaluation Metrics. The evaluation results of all the experiments are presented in terms of the mean average precision (MAP), which is a standard performance evaluation criterion in cross-modal retrieval research [9, 34, 47]. Specifically, we compute the MAP scores on the ranked lists of the retrieved results for multiple cross-modal retrieval tasks, *e.g.*, retrieving text, video, audio, 3D instances using image queries. The cosine distance is adopted to measure the similarity of features. To calculate the MAP, we first evaluate the average precision (AP) of a set of R retrieved items by:

$$AP = \frac{1}{T}\sum_{r=1}^{R} P_r \times \delta(r) \tag{12}$$

where T is the number of relevant items in the retrieved set, $P(r)$ represents the precision of the top r retrieved items, and $\delta(r)$ is an indicator function, whose value is 1 if the r-th retrieved item is relevant (here relevant means belonging to the category of the query). The MAP can be calculated by averaging the AP values.

4.1.3 Implementation Details. For image modality, we first resize it into 224*224 and utilize pretrained VGG-19 [31] to extract a 4096-dimensional feature vector from the fc7 layer as the original image feature. For text, we use a pretrained BERT to extract 768-dimensional text feature. Similar to the BERT paper [2], we take the embedding associated with [CLS] to represent the whole sentence. Other variants such as XLNET [44] and ALBERT [16] could also be used to extract text feature. It is notable that all the compared methods adopt the same image feature and text feature for fair comparison. The value of ϵ in Equation (6) is set to be 1000. The self-loop probability ϵ_0 is set to be 0 for padding embeddings and set to be 1 for real embeddings, since every padding embedding should be replaced by the real embedding. After multimodal coordinated embedding, we employ two fully-connected layers with the Rectified Linear Unit (ReLU) [21] active function for each modality to project them into a common embedding space. The numbers of the hidden units for the two layers are 2048 and 1024, respectively. The weights of the second fully-layers of the two sub-networks are shared to learn the correlation of the two different modalities. The entire network optimized by Adam update rule [14]. We set the initial learning rate as 10^{-4}, the batch size 128. Regarding the parameter λ and m in Equation (10), we will analyze the impact of parameter setting in Figure 2. The best reported results of MCCN are obtained by the optimal values of λ and m per dataset.

4.2 Cross-modal Retrieval on Multiple Modalities

In this section, we first conduct cross-modal retrieval experiments on multiple modalities. Specifically, we conduct experiments on the XMedia dataset involving five modalities (*i.e.*, image, text, video, audio and 3D model), and on the large-scale XmediaNet dataset involving three common modalities (*i.e.*, image, text, video). Note that both of the two multimodal datasets are considerably imbalanced in different modalities.

4.2.1 Baselines. We compare our proposed MCCN with the recently proposed cross-modal retrieval methods that learn embeddings of multiple modalities, namely MvDA [13], JFSSL [35], JLSLR [40] and SDML [8]. Table 2 and Table 3 report the MAP scores of our proposed MCCN method and the baseline methods on Xmedia and XmediaNet, respectively.

4.2.2 Compared with Baselines. From the experimental results on the Xmedia dataset, we can see that our proposed MCCN method achieves the best performance in all cases. Specifically, our MCCN outperforms the previous best model SDML [8], achieving the average improvement 1.3% in terms of the MAP score. This demonstrates the benefit of our model in tackling modality-imbalanced training data and modality-imbalanced optimization. We can also see that the performances of the graph-based regression models JFSSL and JLSLR are significantly better than MvDA, showing the advantage of constructing the graph structure of multimodal samples to preserve the semantic relationships. Moreover, the superiority of SDML over JFSSL and JLSLR demonstrates the effectiveness of preserving the semantic discrimination in the predefined embedding space for multiple modalities.

For the results on the large-scale XmediaNet dataset, our MCCN still outperforms the SOTA model SDML, achieving the average improvement 1.7%. This indicates the obvious advantage of our model in handling large-scale cross-modal retrieval task. Compared to the results on the Xmedia dataset, we can find that our MCCN and the baseline methods all encounter clear performance drop, due to the dramatic increase of the number of semantic categories and still the number of instances in each category remains the same in XmediaNet. In general, the above experimental results verify the

Table 3: Performance Comparison with representative methods on XMediaNet dataset.

Method	Query	Image		Text		Video		Avg
	Target	Text	Video	Image	Video	Image	Text	
MvDA		0.520	0.354	0.514	0.254	0.309	0.252	0.367
JFSSL		0.556	0.372	0.562	0.289	0.350	0.296	0.404
JLSLR		0.571	0.403	0.585	0.302	0.356	0.315	0.422
SDML		0.604	0.412	0.607	0.346	0.387	0.317	0.446
MCCN		0.621	0.415	0.626	0.361	0.415	0.339	0.463

effectiveness of our method for modality-imbalanced cross-modal retrieval with multiple modalities.

4.3 Cross-modal Retrieval on Image and Text

In this section, we conduct cross-modal retrieval experiments specifically for image and text to further evaluate the performance of the proposed MCCN. Note that images and texts often co-occurs on social media, making it significantly less difficult to collect paired image and text datasets than other modalities like audio, video and 3D model. Therefore, multimodal datasets of image and text usually contain modality-balanced training data. We will show the superiority of the multimodal contrastive clustering method proposed by MCCN on these datasets compared with representative cross-modal retrieval methods, including latest adversarial training methods.

4.3.1 Baselines. For comparison, We compare MCCN with seven representative baseline methods, including two traditional methods, namely DCCA [1] and JRL [46], and five deep learning based methods, namely ACMR [34], SDML [8], DSCMR [47], DAVAE [11], and MS^{2}GAN [39]. Table 4 reports the MAP scores of our proposed MCCN method and the comparative methods. Here we denote "Image query Text" as "I2T", and denote "Text query Image" as "T2I".

4.3.2 Compared with Baselines. From the results, we can see that the proposed MCCN achieves the best results on all of the three datasets. Specifically, our MCCN outperforms the previous best model, *i.e.*, MS^{2}GAN [39], with improvements 1.1%, 1.4% and 1.1% in terms of the average MAP scores on Wikipedia, Pascal, and XMediaNet datasets, respectively. The superiority of our MCCN indicates the advantage of multimodal contrastive clustering in preserving the similarity relationships between heterogeneous and homogeneous multimodal samples. We also noticed that the existing cross-modal retrieval methods can better overcome the heterogeneous modality gap by introducing adversarial training in the common embedding space, as we can see that MS^{2}GAN has achieved the past SOTA performance. It is worth noting that, compared to introducing more complex adversarial training, our proposed MCCN uses a lightweight model to achieve the best results.

4.4 Ablation Study

To evaluate the performance of each component used in our MCCN, we conduct a detailed ablation study on various variants of our model. The ablation results for cross-modal retrieval with multiple modalities are given in Table 5, and the ablation results for image and text retrieval are given in Table 6.

Table 4: Performance comparison of conventional image and text retrieval on three widely-used benchmark datasets.

Method	Wikipedia			Pascal			XMediaNet*		
	I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg
DCCA	0.518	0.455	0.486	0.603	0.624	0.618	0.653	0.659	0.656
JRL	0.516	0.460	0.488	0.587	0.582	0.585	0.623	0.620	0.622
ACMR	0.535	0.476	0.505	0.671	0.674	0.672	0.704	0.695	0.699
SDML	0.528	0.466	0.497	0.677	0.682	0.680	0.719	0.715	0.717
DSCMR	0.541	0.472	0.506	0.685	0.695	0.690	0.725	0.727	0.726
DAVAE	0.540	0.474	0.507	0.687	0.683	0.685	0.727	0.731	0.729
MS^{2}GAN	0.544	0.475	0.509	0.690	0.697	0.693	0.732	0.730	0.731
MCCN	0.552	0.487	0.520	0.705	0.712	0.707	0.741	0.743	0.742

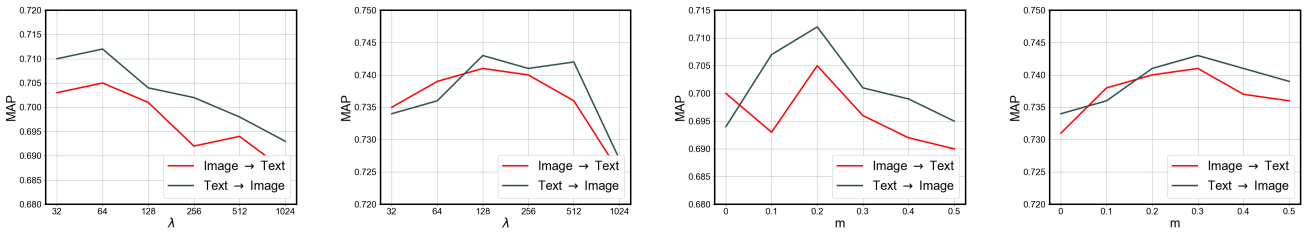
* We replace the text features of the original dataset with Bert features, and thus the performances of both our method and the baselines are significantly improved.

For cross-modal retrieval on the XMedia dataset involving 5 modalities, we fully evaluated the impact of multimodal coordinated embedding module (MCE) and multimodal contrastive clustering module (MCC). In general, we find these variants underperform full MCCN, which means that all of the components utilized in our MCCN contribute to the final retrieval accuracy. We firstly remove the intra-class margin Δ_p and the inter-class margin Δ_n respectively to analyze the impact of optimization flexibility. We can see that the performance of our model has a obvious decline after removing Δ_p and Δ_n, which shows that the intra-class and inter class margins can improve the semantic partition of labeled multimodal samples and thus enhance the cross-modal retrieval performance. Comparing these two variants, we found that removing inter-class margin Δ_n performs worse than removing intra-class margin Δ_p. It indicates that inter-class relationship plays a more important role in improving the quality of common embedding space, which has been largely ignored in previous work. Then, we remove the entire MCC module and utilize the typical pair-wise constraints as [47] to preserve cross-modal associations, and the model gets the worst performance as we expected. This demonstrates the effectiveness of our contrastive clustering module in handling the imbalanced optimization between heterogeneous and homogeneous similarities. Finally, we remove the MCE module and only utilize MCC module to learn common embedding space with modality-imbalanced training data. It can be seen that the performance is seriously descended, showing that our proposed data-driven MCE module can effectively alleviate the modality-imbalanced problem.

For image and text retrieval, we carry out ablation studies on Pascal dataset. Since multimodal datasets of image and text usually contain modality-balanced training data, we only evaluate each component in MCC module in ablation experiments. From Table 6, we can see that the complete MCCN achieves the best performance,

Table 5: Ablation results on the XMedia dataset.

Variant	Query	Image				Text				Video				Audio				3D				Avg
	Target	Text	Video	Audio	3D	Image	Video	Audio	3D	Image	Text	Audio	3D	Image	Text	Video	3D	Image	Text	Video	Audio	
MCCN		0.908	0.654	0.614	0.665	0.915	0.677	0.627	0.711	0.601	0.592	0.393	0.499	0.586	0.621	0.442	0.526	0.675	0.686	0.514	0.425	0.617
MCCN w/o Δ_p		0.904	0.641	0.603	0.652	0.907	0.665	0.622	0.701	0.596	0.587	0.391	0.476	0.577	0.614	0.435	0.519	0.654	0.671	0.504	0.417	0.607
MCCN w/o Δ_n		0.901	0.646	0.605	0.648	0.904	0.659	0.617	0.694	0.589	0.582	0.384	0.471	0.579	0.613	0.438	0.511	0.663	0.667	0.502	0.409	0.604
MCCN w/o MCC		0.889	0.620	0.592	0.632	0.888	0.638	0.603	0.681	0.574	0.566	0.381	0.463	0.562	0.600	0.428	0.509	0.646	0.657	0.488	0.398	0.591
MCCN w/o MCE		0.905	0.631	0.594	0.642	0.912	0.656	0.613	0.693	0.595	0.590	0.370	0.453	0.581	0.617	0.420	0.501	0.664	0.679	0.486	0.397	0.600



(a) Pascal (b) XMediaNet (c) Pascal (d) XMediaNet

Figure 2: Parameter analysis of λ and m.

Table 6: Ablation results on the Pascal dataset.

Variant	Img2Text	Text2Img	Avg	△Avg
MCCN	0.705	0.712	0.707	-
MCCN w/o Δ_p	0.673	0.707	0.690	-0.017
MCCN w/o Δ_n	0.683	0.701	0.692	-0.015
MCCN w/o MCC	0.670	0.694	0.687	-0.020

indicating that each component is beneficial for image and text retrieval. Then, we separately remove intra-class margin Δ_p and the inter-class margin Δ_n, and the MCCN encounters different degrees of performance degradation. The most obvious decline comes from removing the entire MCC and replacing it with the typical pair-wise constraints, since it optimizes heterogeneous similarities and homogeneous similarities indiscriminately and thus leading to modality-imbalanced optimization.

4.5 Parameter Analysis

The parameters of our method are analyzed in this section. The objective function contains two parameters λ and m, we evaluate their influences on Pascal and XmediaNet datasets. We vary λ from 32 to 1024, and show the impact of different values of λ in Figure 2(a) and (b). From the figures we can see that the MAP first increase with the growth of λ on both datasets, and then begins a slow decline after λ surpasses a threshold. The best parameter setting of λ are 64 and 128 on the two datasets. Regarding the parameter m, we vary it from 0 to 0.5 and show its impact in Figure 2(c) and (d). We can see that the change of m has a small performance change on XMediaNet, while encounters a large change on the pascal dataset, which may be caused by the size of the dataset. The best parameter setting of m are 0.2 and 0.3 on Pascal and XMediaNet dataset, respectively.

5 CONCLUSION

In this paper, we identify the important issues of modality-imbalanced cross-modal retrieval for large-scale applications with multiple modalities. We propose a novel multimodal coordinated clustering network MCCN to tackle these issues which consists of two modules. The multimodal coordinated embedding module employs a data-driven approach to coordinate multiple modalities for generating modality-balanced training samples. The multimodal contrastive clustering module jointly optimizes the pair-wise similarity and class-level similarity across multiple modalities for preserving multimodal semantic associations. Experimental results on the benchmark datasets demonstrate the effectiveness of our proposed method.

ACKNOWLEDGEMENTS

This work was supported in part by the Ministry of Science & Technology of China under Grants #2020AAA0108401 and #2020AAA0108405, and NSFC Grants #11832001 and #71621002.

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*. 1247–1255.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 7–16.

[4] Shlomo Geva and Joaquin Sitte. 1991. Adaptive nearest neighbor pattern classification. *IEEE Transactions on Neural Networks* 2, 2 (1991), 318–322.

[5] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision* 106, 2 (2014), 210–233.

[6] Jun Guo and Wenwu Zhu. 2019. Collective affinity learning for partial cross-modal hashing. *IEEE Transactions on Image Processing* 29 (2019), 1344–1355.

[7] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12 (2004), 2639–2664.

[8] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable deep multi-modal learning for cross-modal retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 635–644.

[9] Xin Huang and Yuxin Peng. 2018. Deep cross-media knowledge transfer. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*. 8837–8846.

[10] Xiao-Bo Jin, Cheng-Lin Liu, and Xinwen Hou. 2010. Regularized margin-based conditional log-likelihood loss for prototype learning. *Pattern Recognition* 43, 7 (2010), 2428–2438.

[11] Mengmeng Jing, Jingjing Li, Lei Zhu, Ke Lu, Yang Yang, and Zi Huang. 2020. Incomplete cross-modal retrieval with dual-aligned variational autoencoders. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3283–3291.

[12] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. 2021. Towards open world object detection. *arXiv preprint arXiv:2103.02603* (2021).

[13] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2015. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 1 (2015), 188–194.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Teuvo Kohonen. 1998. The self-organizing map. *Neurocomputing* 21, 1-3 (1998), 1–6.

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[17] Jiacheng Li, Siliang Tang, Juncheng Li, Jun Xiao, Fei Wu, Shiliang Pu, and Yueting Zhuang. 2020. Topic adaptation and prototype encoding for few-shot visual storytelling. *arXiv preprint arXiv:2008.04504* (2020).

[18] Cheng-Lin Liu, In-Jung Eim, and Jin Hyung Kim. 1997. High accuracy handwritten Chinese character recognition by improved feature matching method. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*. 1033–1037.

[19] Cheng-Lin Liu and Masaki Nakagawa. 2001. Evaluation of prototype learning algorithms for nearest-neighbor classifier in application to handwritten character recognition. *Pattern Recognition* 34, 3 (2001), 601–615.

[20] Xin Liu, Yiu-ming Cheung, Zhikai Hu, Yi He, and Bineng Zhong. 2021. Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval. *IEEE Transactions on Emerging Topics in Computational Intelligence* 5, 4 (2021), 607–619.

[21] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*.

[22] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-media shared representation by hierarchical learning with multiple deep networks.. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 3846–3853.

[23] Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2017. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (2017), 2372–2385.

[24] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1 (2019), 1–24.

[25] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. 2018. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing* 27, 11 (2018), 5585–5599.

[26] Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang. 2015. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 3 (2015), 583–596.

[27] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 139–147.

[28] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*. 251–260.

[29] Jan Rupnik and John Shawe-Taylor. 2010. Multi-view canonical correlation analysis. In *Proceedings of the 2010 Conference on Data Mining and Data Warehouses*. 1–4.

[30] Atsushi Sato and Keiji Yamada. 1996. Generalized learning vector quantization. In *Proceedings of the 10th Conference on Neural Information Processing Systems*. 423–429.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[32] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition*. 6398–6407.

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[34] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*. 154–162.

[35] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. 2016. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 10 (2016), 2010–2023.

[36] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* (2016).

[37] Weiran Wang and Karen Livescu. 2015. Large-scale approximate kernel canonical correlation analysis. *arXiv preprint arXiv:1511.04773* (2015).

[38] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.

[39] Fei Wu, Xiao-Yuan Jing, Zhiyong Wu, Yimu Ji, Xiwei Dong, Xiaokai Luo, Qinghua Huang, and Ruchuan Wang. 2020. Modality-specific and shared generative adversarial network for cross-modal retrieval. *Pattern Recognition* 104 (2020), 107335.

[40] Jianlong Wu, Zhouchen Lin, and Hongbin Zha. 2017. Joint latent subspace learning and regression for cross-modal retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 917–920.

[41] Liwei Wu, Shuqing Li, Cho-Jui Hsieh, and James Sharpnack. 2019. Stochastic shared embeddings: Data-driven regularization of embedding layers. *arXiv preprint arXiv:1905.10630* (2019).

[42] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. 2014. Supervised hashing for image retrieval via image representation learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*. 2156–2162.

[43] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. 2018. Robust classification with convolutional prototype learning. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*. 3474–3482.

[44] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*. 5753–5763.

[45] Zhixiong Zeng, Shuai Wang, Nan Xu, and Wenji Mao. 2021. PAN: Prototype-based Adaptive Network for Robust Cross-modal Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1125–1134.

[46] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2013. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2013), 965–978.

[47] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep Supervised Cross-Modal Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10394–10403.