PAN: Prototype-based Adaptive Network for Robust Cross-modal Retrieval

Zhixiong Zeng, Shuai Wang, Nan Xu, 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,wangshuai2017,xunan2015,wenji.mao}@ia.ac.cn

ABSTRACT

In practical applications of cross-modal retrieval, test queries of the retrieval system may vary greatly and come from unknown category. Meanwhile, due to the cost and difficulty of data collection as well as other issues, the available data for cross-modal retrieval are often imbalanced over different modalities. In this paper, we address two important issues to increase the robustness of cross-modal retrieval system for real-world applications: handling test queries from unknown category and modality-imbalanced training data. The first issue has not been addressed by existing methods and the second issue was not well addressed in the related research. To tackle the above issues, we take the advantage of prototype learning, and propose a prototype-based adaptive network (PAN) for robust cross-modal retrieval. Our method leverages a unified prototype to represent each semantic category across modalities, which provides discriminative information of different categories and takes unified prototypes as anchors to learn cross-modal representations adaptively. Moreover, we propose a novel prototype propagation strategy to reconstruct balanced representations which preserves the semantic consistency and modality heterogeneity. Experimental results on the benchmark datasets demonstrate the effectiveness of our method compared to the SOTA methods, and further robustness tests show the superiority of our method in solving the above issues.

CCS CONCEPTS

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

KEYWORDS

Cross-modal retrieval; prototype learning; unknown category; modality imbalance; robustness

ACM Reference Format:
Zhixiong Zeng, Shuai Wang, Nan Xu, 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 (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462867

*Corresponding author

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462867

1 INTRODUCTION

With the rapid growth of multi-modal data in social media, cross-modal retrieval (*e.g.* image and text retrieval), which aims to take one type of data as the query to retrieve relevant data of another type [48], is in great demand. It has become a core multi-modal research area and had a number of applications in domains such as image retrieval [42], image caption [33], video recommendation [37], automatic story generation [19] and so forth.

Since features of different modalities usually have inconsistent distributions and representations, the key challenge of cross-modal retrieval is to bridge the gap of modality heterogeneity, that is, developing computational means to assess the semantic similarity of samples across modalities. A typical approach to bridge the heterogeneity gap is representation learning, which projects cross-modal data into a common representation space to directly compute their similarity. Early work mainly uses statistical correlation analysis to convert cross-modal data to the common representation [8, 11, 36], while recent work leverages the representation capabilities of deep neural networks (DNN) to learn the complex nonlinear cross-modal associations [1, 6]. To gain more effective common representation, recent research further exploits label information and adversarial learning to preserve intra-modality discrimination and inter-modality invariance for the retrieval task, and achieves superior performances [25, 26, 34, 38, 48].

Despite the success of DNN based methods, two main drawbacks in the existing methods hinder the practical applications of cross-modal retrieval. First, in practice, the content of a query image/text may vary greatly, and the retrieval system often needs to handle queries with unknown class labels [20]. However, existing methods cannot correctly identify test queries from *unknown category*, as they require the queries in the retrieval system fall into some predefined categories. This is due to the fact that most existing methods [25–27, 34, 38] utilize classification layer activated by softmax to learn semantically discriminative representations. It essentially learns a partition of the whole representation space, and thus the samples from unknown category are still predicted to some known categories with high confidence [45]. Recently, several zero-shot cross-modal retrieval methods have been proposed to tackle the inconsistency between training and test categories [2, 3, 21], in which the semantic categories of the test set are *unseen* in the training set. Since these methods take the word embeddings of class categories as external knowledge to enhance knowledge transfer, they actually assume the (unseen) categories of the test set are *known*. Therefore, to increase the robustness of the retrieval system

for practical applications, a vital research issue is to handle test queries from unknown category, which has not been addressed in previous research on cross-modal retrieval.

Second, the mainstream cross-modal retrieval methods rely on the balanced multi-modal data, that is, under the assumption where there is a sample in one modality, there is a corresponding sample in the other modality with the same label [14]. However, this assumption of modality-balanced data cannot always be satisfied in real-world applications, due to the discrepancies caused by different modalities in the difficulty and cost of data collection and other issues such as content preference or privacy. In reality, the available data between text, image and audio, for example, are all considerably *modality-imbalanced*. As the imbalanced data is incomplete and insufficient to learn the invariant representations for different modalities, the heterogeneity gap cannot be well bridged in this situation. Recently, several hashing based methods have been proposed to address the imbalanced data for cross-modal retrieval [10, 40], which mainly utilize the *semantic consistency* between similar samples (*i.e.*, belonging to the same category) from different modalities to reconstruct equivalent numbers of samples. More recent work [14] investigates the imbalanced problem in the framework of deep generative models, but is still based on the semantic consistency to reconstruct incomplete samples. Unfortunately, while addressing the imbalanced problem in cross-modal retrieval, these methods fail to consider the *modality heterogeneity* in the reconstruction process, and thus cannot supply informative sample pairs to overcome the modality heterogeneity gap. Therefore, to increase the robustness of the retrieval system for practical applications, another important research issue is to effectively handle modality-imbalanced training data, which has not been well addressed in previous research.

To tackle the above issues and increase the robustness of cross-modal retrieval for real-world applications, in this paper, we propose a novel Prototype-based Adaptive Network (PAN) for cross-modal retrieval. As prototypes can be viewed as the representatives of each discriminative category in the common representation space, it provides valuable clues to discriminate category information and bridge cross-modal associations. Thus the central idea of our method is the leverage of a unified prototype to represent each semantic category across modalities, which provides both the discriminative information to differentiate unknown category from known ones and the commonalities of multi-modal samples belonging to the same category to imply the semantic consistency. To this end, we develop a prototype-based representation learning method to interactively learn the common representations across modalities and the unified prototypes for each category. We take prototypes as anchors to adaptively learn invariance loss and discrimination loss, and the learned prototypes can provide inference clues for differentiating test queries of unknown category. Furthermore, to tackle the modality-imbalanced problem, we propose a prototype propagation strategy to fuse semantically consistent prototype and the nearest neighbors of the another modality for heterogeneity.

The main contributions of our work are as follows:

- We identify two important issues to increase the robustness of cross-modal retrieval system in real-world applications and propose a prototype-based adaptive network PAN for robust cross-modal retrieval.
- We develop a prototype-based representation learning method to jointly learn the common representations across modalities and the unified prototypes for differentiating test queries of unknown category.
- We propose a prototype propagation strategy to reconstruct balanced representations for each modality, which can preserve modality heterogeneity and semantic consistency simultaneously.
- We conduct extensive experiments on several benchmark datasets and the results demonstrate the effectiveness of our method compared to the SOTA methods. Further robustness tests show the superiority of our method in handling unknown category and modality-imbalanced data.

2 RELATED WORK

2.1 Cross-modal Retrieval

The key challenge of cross-modal retrieval is to bridge the modality gap and learn a common representation space in which the semantic similarity of items across modalities can be compared. The typical methods can be divided into two main groups, shallow learning methods and deep learning methods. Shallow learning methods use statistical correlation analysis to transform multi-modal data to the common representation by maximizing pairwise correlations, in which the representative methods are CCA [11] and its extensions [8, 36]. Deep learning methods take advantage of the powerful representation capabilities of deep neural networks to learn the common representation for multi-modal data, and optimize pairwise constraints at different levels to preserve the semantic associations [6, 23].

To gain a more effective representation space, recent research further exploits label information to capture the underlying semantic structure of multi-modal samples [18, 25, 34, 38, 48], including intra-modality discrimination and inter-modality invariance. Inspired by the great success of adversarial learning [9], several adversarial cross-modal retrieval methods have been proposed with the utilization of a modality classifier to further learn modality-invariant representations [26, 34, 38], achieving promising performance on cross-modal retrieval task.

2.2 Cross-modal Retrieval with Unseen Category Queries

Conventional cross-modal retrieval methods are based on the assumption that test categories remain the same with training categories, which means the queries of a retrieval system fall into some pre-defined categories. In practice, the content of a query image/text may vary extensively and come from unknown category [20], which bring about a great challenge to the practical applications of cross-modal retrieval. Because of the discrepancy between real-world applications and the above assumption, some zero-shot cross-modal retrieval methods have been proposed to tackle this challenge, which attempt to construct a test set that does not have overlapping categories with the training set to perform zero-shot learning[16]. Chi *et al.* [2] propose the first zero-shot cross-modal retrieval method, which exploits the word embeddings of both seen and unseen categories to enhance knowledge transfer to unseen categories. Xu *et al.* [43] leverage the word embedding

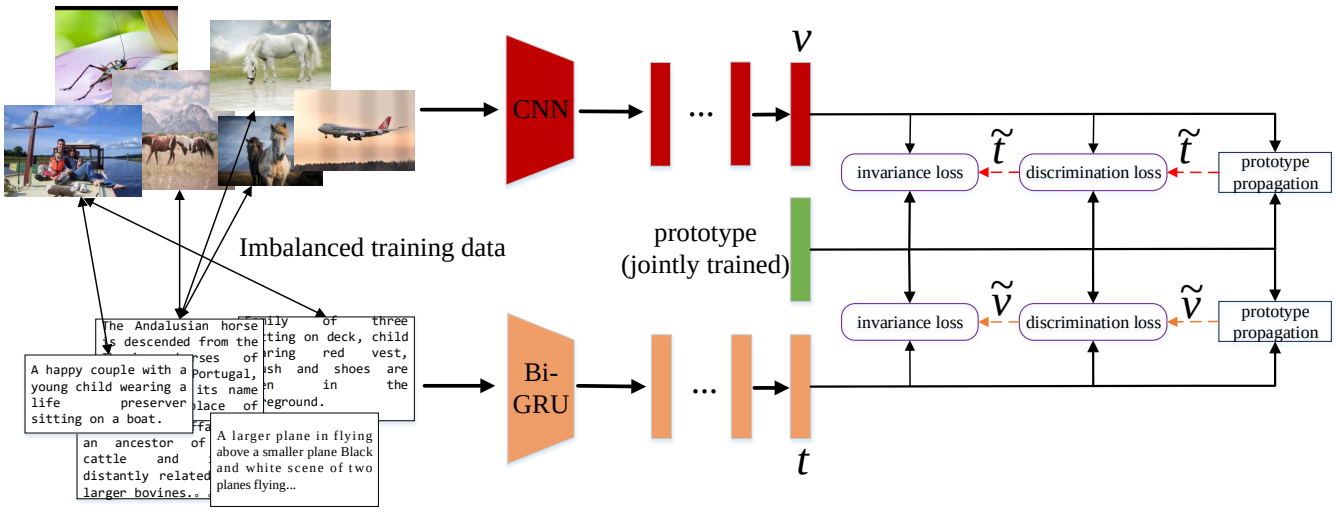


Figure 1: Overall architecture of the proposed model PAN. It first utilizes a number of fully-connected layers to project the original image and text features into a common representation space separately, and then leverages a unified prototype to preserve the modality invariance and semantic discrimination. We further present a prototype propagation stategy to reconstruct text samples from excessive images, or reconstruct image samples from excessive texts, so as to obtain modality-balanced training data. The prototypes are jointly trained and can provide inference clues to differentiate unknown category.

labels as guidance to supervise the semantic feature learning and enhance knowledge transfer to unseen categories. Similar approach has been followed in other work [3, 44]. However, similar to most zero-shot learning methods [17, 41], they are actually based on the assumption that the (unseen) categories of the test set are known, which still can not satisfy the requirement in real-world applications. Our proposed model is among the first to handle test queried from unknown category for robust cross-model retrieval in real-world applications. It take the advantage of prototypes to provide inference clus that differentiating test queries of unknown category, thus it is robust to open world queries.

2.3 Imbalanced Cross-modal Retrieval

Most existing cross-modal retrieval methods focus on dealing with the balanced multi-modal data, that is, for a sample in one modality, there is a corresponding sample in the other modality with the same label. However, the assumption of modality-balanced data cannot always be satisfied in real-world applications. In reality, the available data for training cross-modal retrieval system are all considerably modality-imbalanced. A straight forward strategy to tackle the modality-imbalanced problem is removing the samples only with incomplete modalities, yet the model trained will clearly lose information and introduce extra noises [46]. Several cross-modal hashing methods have been proposed to process modality-imbalanced data, which mainly utilize the semantic consistency to reconstruct equivalent numbers of samples for each modality [10, 40]. Wu *et al.* [40] employs adversarial training scheme to learn a couple of hash functions enabling translation between modalities, and regenerate the missing modality sample from the available modality one. Guo *et al.* [10] propose a collective affinity learning method and derive a probabilistic model to collectively reconstruct the affinities of the missing modality sample via the available modality one. However, hashing based methods address the modality-imbalanced problem on the Hamming space, which focus on retrieval efficiency rather than accuracy. More recently, Jing *et al.* [14] investigates the modality-imbalanced problem in the framework of deep generative models in the real-valued space. They propose to reconstruct incomplete multi-modal samples by dual-aligned variational autoencoders. Unfortunately, these methods fail to consider the modality heterogeneity in the reconstruction process, and thus cannot supply informative sample pairs to overcome the modality heterogeneity gap. Our method propose a prototype propagation strategy to fuse semantically consistent prototype and the nearest neighbors of the another modality for heterogeneity, thus it is more robust to modality-imbalanced data.

Next, we shall first give the problem formulation of cross-modal retrieval, and then elaborate on our proposed prototype-based adaptive network for robust cross-modal retrieval in detail.

3 PROBLEM FORMULATION

Without losing generality, we focus on cross-modal retrieval for image and text. For conventional cross-modal retrieval, the datasets are collections of n instances of image-text pairs, denoted as $\Psi_P = \{(x_k^v, x_k^t)\}_{k=1}^n$, where x_k^v is the input image sample and x_k^t is the input text sample. Each pair (x_k^v, x_k^t) is assigned a semantic label $y_k \in \{1, 2, ..., C\}$, where C is the number of semantic categories. Considering the modality-imbalanced training data, we assume that there is a collection of n_1 instances of image samples $\Psi_V = \{x_i^v\}_{i=1}^{n_1}$, which can be devided into $\{\Psi_V^1, \Psi_V^2, ..., \Psi_V^C\}$, where Ψ_V^r represents the subset of images belonging to the r-th category, and n_1^r denotes

the length of Ψ_V^r. Similarly, the text collection of n_2 instances $\Psi_T = \{x_j^t\}_{j=1}^{n_2}$, which can be devided into $\{\Psi_T^1, \Psi_T^2, ..., \Psi_T^C\}$, where Ψ_T^r represents the subset of text belonging to the r-th category, and n_2^r denotes the length of Ψ_T^r.

The imbalance of image modality and text modality on the r-th category can be defined as $n_1^r \neq n_2^r$. Without losing generality, we assume $n_1^r > n_2^r$, thus we can define an excess set $\Omega^r = \{v_1^r, v_2^r, ..., v_{\delta^r}^r\}$, where $\delta^r = n_1^r - n_2^r$. As the samples belonging to the same category are related to each other, the excess set Ω^r is randomly selected from Ψ_V^r. The straight forward strategy to avoid the modality-imbalanced problem is to discard the excess set, but it will further lose information and introduce extra noise. Therefore, our goal is to reconstruct a text set of the same size from the excess set, denoted as $\widetilde{\Psi_T^r}$.

4 PROPOSED METHOD

Figure 1 illustrates the overall architecture of our proposed model PAN. First, we learn some modality-specific transformation functions to project original image and text features into a common representation space. Then, we leverage a unified prototype for each category as anchor to jointly calculate the invariance loss and discrimination loss, so that the modality invariance and semantic discrimination can be preserved in the common space. Finally, we present a prototype propagation strategy to reconstruct modality-balanced data with the semantic consistency and modality heterogeneity when the input training data are modality-imbalanced. In the testing phase, we utilize the learned prototypes of known categories to provide inference clues for differentiating test queries of unknown category.

4.1 Prototype-based Representation Learning

Since features of different modalities usually have heterogeneous representations, the semantic similarity of items across different modalities cannot be directly computed. Therefore, we utilize two sub-networks to learn modality-specific transformation functions to project image and text features into a common representation space, which can be formulated as:

$$v_i = f_V(CNN(x_i^v); \theta_V), \quad x_i^v \in \Psi_V \tag{1}$$

$$t_j = f_T(GRU(x_j^t); \theta_T), \quad x_j^t \in \Psi_T \tag{2}$$

where $f_V(\cdot)$ and $f_T(\cdot)$ are modality-specific transformation functions for image and text, θ_V and θ_T are the trainable parameters of the two functions, $v_i \in \mathbb{R}^d$ and $t_j \in \mathbb{R}^d$ are the projected features in the common representation space, and d is the feature dimension of the common representations.

To effectively perform cross-modal retrieval, we further preserve the modality invariance and semantic discrimination so that the items belonging to the same category across different modalities are closest in the common representation space. Thus, we propose a novel prototype-based representation learning method to jointly learn modality-invariant and semantic-discriminative representations. Since items from different modalities share the same semantic content [35, 39], the unified prototypes for both image and text modality can be denoted as:

$$M = \{m_c | c = 1, 2, ..., C\} \tag{3}$$

where c denotes the index of the categories. Note that we randomly generate a set of prototypes in the initialization phase, and then jointly learn prototypes and common representations during the training phase.

Inspired by [45], given the feature representation v_i (or t_j) in the common space, we classify it to the category whose prototype is closest to v_i in all prototypes. Specifically, we calculate the distance between v_i and m_c to measure the probability of v_i belonging to the prototype m_c, which can be formulated as:

$$p(v_i \in m_c) \propto -d(v_i, m_c) \tag{4}$$

where $d(\cdot)$ denotes the Euclidean distance function. To satisfy the non-negative and sum-to-one properties of the probability, we further define the probability $p(v_i \in m_c)$ as:

$$p(v_i \in m_c) = \frac{e^{-\gamma d(v_i, m_c)}}{\sum_{k=1}^{C} e^{-\gamma d(v_i, m_k)}} \tag{5}$$

where γ is a hyper-parameter to control the hardness of probability assignment. To preserve the semantic discrimination in the common representation space, we define the distance based cross entropy (CE) loss as:

$$l(v_i, M) = \sum_{c=1}^{C} \mathbf{1}\{y_i = c\} log(p(v_i \in m_c)) \tag{6}$$

Similarly, the distance based cross-entropy loss for text item can be formulated as:

$$l(t_j, M) = \sum_{c=1}^{C} \mathbf{1}\{y_j = c\} log(p(t_j \in m_c)) \tag{7}$$

To preserve the modality invariance, we utilize a prototype-based invariant loss to bridge cross-modal similarity relationships, by pulling the image and text representations closer to their corresponding prototypes, which is formulated as:

$$\begin{aligned} pl(v_i, M) &= \|v_i - m_y^i\|_2^2 \\ pl(t_j, M) &= \|t_j - m_y^j\|_2^2 \end{aligned} \tag{8}$$

where m_y^i and m_y^j are the corresponding prototypes of y_i and y_j. Prototype-based invariance loss is suitable for cross-modal retrieval, because: (1) the modality invariance can be preserved by collectively approaching to the corresponding prototype, which can effectively reduce storage space and computation compared pairwise constraints commonly utilized of previous work [34, 38, 48]; (2) the invariant loss can further improve the intra-class compactness and inter-class separability, which is beneficial for learning semantically discriminative representations.

4.2 Prototype Propagation for Imbalanced Data Alignment

We define the modality-imbalanced problem by introducing an excess set $\Omega^r = \{v_1^r, v_2^r, ..., v_{\delta^r}^r\}$, here we assume $n_1^r > n_2^r$. A straight forward strategy to avoid the imbalanced problem is to discard the excess set Ω^r, but it will further lose information and introduce extra noise. Hence our goal is to reconstruct a text set $\widetilde{\Psi_T^r} = \{\widetilde{t_1^r}, \widetilde{t_2^r}, ...\widetilde{t_{\delta^r}^r}\}$ of the same size from Ω^r, which can preserve the semantic consistency and modality heterogeneity for imbalanced cross-modal retrieval.

To preserve the modality heterogeneity between the reconstructed representation $\widetilde{t_i^r}$ and v_i^r, we first design $N_k(v_i^r)$ as the list of k-nearest neighbors in the text modality, by using v_i^r as query to rank their distance.

$$N_k(v_i^r) = [t_1, t_2, ..., t_k] \tag{9}$$

where k is a hyper-parameter. It is worth noting that when $k = 0$, it is equivalent to not performing imbalanced data alignment. We will analyze the effect of k in detail in Section 5.5. Since the prototypes are jointly trained with the discrimination loss and invariant loss, the learned prototypes imply the semantic consistence of samples belonging to the same category between different modalities. Therefore, the reconstructed text representation $\widetilde{t_i^r}$ can calculated by the corresponding prototype m_r and $N_k(v_i^r)$:

$$\widetilde{t_i^r} = \psi(m_r, N_k(v_i^r)) \tag{10}$$

Here ψ denotes the modified gated recurrent unit [31] to propagate the prototype information to the nearest text representations dynamically:

$$\begin{aligned} o_z &= tanh(W_o[h_{z-1}, t_z] + b_o) \\ g_z &= \sigma(W_g[h_{z-1}, t_z] + b_g) \\ h_z &= g_z * h_{z-1} + (1 - g_z) * o_z \end{aligned} \tag{11}$$

where $z = 1, 2, ..., k$, h_z denotes the hidden state, $h_0 = m_r$ denotes the corresponding prototype, $h_k = \widetilde{t_i^r}$ represents the reconstructed text representation, σ denotes the sigmoid function, W_o, W_g, b_o, b_g are to-be-learned parameters. o_z is a fused feature which enhances the interaction between h_{z-1} and t_z, and g_z performs as a gate to select the most shared information between h_{z-1} and t_z.

Note that this gating mechanism is timing sensitive, which means that more similar neighbors in the representation space can play a more important role. This is consistent with our intuition, because the closer neighbors should play a greater role. Although the above method could reconstruct the imbalanced data that capture the semantic consistency and modality heterogeneity, there still remain two obvious drawbacks:

- It is well-known that the calculated k-nearest neighbors in the representation space usually deviate from optimal cases, *i.e.*, the k-nearest neighbors may include false items from another category [7, 49]. These false items contain confusing distribution information and propagate errors during the reconstruction process.
- The reconstructed information only retains the similarity relationship from image to text in one direction, while ignoring the similarity relationship from text to image. In fact, there are two opposite directions in image-text similarity calculation due to the modality heterogeneity gap, which have totally different distribution characteristics [24].

To overcome these drawbacks, inspired by [49], we define the k-reciprocal nearest neighbors of v_i^r from the text modality to make the reconstructed representation fully retain the similar relationship between different modalities:

$$\mathcal{R}_k(v_i^r) = [t_i | (t_i \in N_k(v_i^r)) \wedge (v_i^r \in N_k(t_i))] \tag{12}$$

where $N_k(t_i)$ is the k-nearest neighbors in the image modality by using t_i as query to rank their distance, and $\wedge$ is the logical operator 'conjunction'. We can see that an image and a text are called k-reciprocal nearest neighbors, they are both ranked top-k when one of them is taken as the query.

Considering the similarity relationship between items belonging to the same category, we can rewrite this definition as:

$$\mathcal{R}_k(v_i^r) = [t_i | (t_i \in N_k(v_i^r)) \wedge (\Lambda(t_i) \geq \frac{2}{3}k)] \tag{13}$$

where $\Lambda(t_i)$ denotes the number of items in $N_k(t_i)$ that belonging to the same category with v_i^r. Obviously, the k-reciprocal nearest neighbors are more related to v_i^r than k-nearest neighbors, through stricter similarity constraints including image to text direction and text to image direction. Therefore, the reconstructed representation $\widetilde{t_i^r}$ can be obtained as follows:

$$\widetilde{t_i^r} = \psi(m_r, \mathcal{R}_k(v_i^r)) \tag{14}$$

4.3 Training Objective

Our training objective is to preserve the semantic discrimination and modality invariance in the common representation space for cross-modal retrieval. Based on Equations (6) and (7), the discrimination loss can be defined as:

$$\begin{aligned} &\mathcal{L}_{dl} = l(v, M) + l(t, M) \\ &s.t. \quad v \in (\Psi_V \cup \widetilde{\Psi_V}), \quad t \in (\Psi_T \cup \widetilde{\Psi_T}) \end{aligned} \tag{15}$$

where $\widetilde{\Psi_V}$ and $\widetilde{\Psi_T}$ denote the reconstructed representations for imbalanced data. Based on Equations (8), the invariance loss can be defined as:

$$\begin{aligned} &\mathcal{L}_{il} = pl(v, M) + pl(t, M) \\ &s.t. \quad v \in (\Psi_V \cup \widetilde{\Psi_V}), \quad t \in (\Psi_T \cup \widetilde{\Psi_T}) \end{aligned} \tag{16}$$

The final objective function can be defined as:

$$\mathcal{L} = \mathcal{L}_{dl} + \lambda \mathcal{L}_{il} \tag{17}$$

where λ is a hyper-parameter to control the contribution of different components. We will discuss the effect of λ in Section 5.5.

4.4 Inference

In the testing phase, the learned prototypes for known categories can provide inference clues to differentiate unknown category. We implement such inference by setting a threshold ϵ_v for image modality and ϵ_t for text modality. Given a query instance q from the image modality (or text modality), we project it into the common representation space and find the best matching prototype by comparing their similarities. Here we use Euclidean distance for similarity calculation. If the similarity s_q between the query and the best matching prototype is smaller than ϵ_v, the query will be identified as an outlier from unknown category. Obviously, there is a tradeoff in choosing ϵ_v. The detailed analysis of ϵ_v and ϵ_t can be seen in Section 5.4.

Supported by the outlier analysis, we use the mean of the outliers as the new prototype of the unknown category (denoted as m_u), and then pull each outlier to the new prototype, just like the training process. The inferred representation of the query can be calculated as:

$$q_{inf} = \begin{cases} \alpha f(q) + (1 - \alpha) m_u, & \text{if } s_q < \epsilon_v \\ f(q), & \text{otherwise} \end{cases} \tag{18}$$

Table 1: Performance comparison in terms of mAP on three widely-used benchmark datasets for conventional cross-modal retrieval.

Method	Pascal-Sentence			NUS-WIDE-10K			XMediaNet		
	Image→Text	Text→Image	Average	Image→Text	Text→Image	Average	Image→Text	Text→Image	Average
CCA [12]	0.203	0.208	0.206	0.167	0.181	0.174	0.212	0.217	0.215
KCCA [11]	0.488	0.446	0.467	0.351	0.356	0.354	0.252	0.27	0.261
JRL [47]	0.563	0.505	0.534	0.466	0.499	0.483	0.488	0.405	0.447
Corr-AE [6]	0.532	0.521	0.527	0.441	0.494	0.468	0.469	0.507	0.488
CMDN [25]	0.544	0.526	0.535	0.492	0.542	0.517	0.485	0.516	0.501
MCSM [27]	0.598	0.598	0.598	0.522	0.546	0.534	0.540	0.550	0.545
ACMR [34]	0.538	0.544	0.541	0.519	0.542	0.531	0.536	0.519	0.528
CM-GANS [26]	0.603	0.604	0.604	0.536	0.551	0.543	0.567	0.551	0.559
DSCMR [48]	0.668	0.673	0.670	0.555	0.585	0.570	0.641	0.654	0.647
MS^2GAN [38]	0.677	0.670	0.673	0.568	0.574	0.572	0.647	0.656	0.651
PAN	0.686	0.689	0.688	0.590	0.571	0.581	0.669	0.660	0.665

where α is a weight to balance the original representation and the prototype representation, and f is the learned modality-specific transformation function $f_V(\cdot)$ or $f_T(\cdot)$. When the outlier is closer to the prototype of the known categories, a greater penalty should be placed to pull it towards the unknown prototype. Therefore, we define α as:

$$\alpha = \frac{e^{-s_q}}{\sum_{p \in \Psi_O} e^{-s_p}} \tag{19}$$

where Ψ_O denotes the set of outliers from both image and text modality. We use q_{inf} as the final representation to perform cross-modal retrieval.

5 EXPERIMENTS

In the experiments, we first compare our proposed model PAN with ten baseline methods for conventional cross-modal retrieval. We then conduct experiment on robust cross-modal retrieval and test the model performances for handling modality-imbalanced data and queries from unknown category. Finally, we conduct a detailed parameter analysis on the hyper-parameters of PAN.

5.1 Experimental Setup

5.1.1 Datasets. To verify the effectiveness of our proposed method, we conduct experiments on four widely-used benchmark datasets, namely Wikipedia [30], Pascal-Sentence, [29], NUS-WIDE-10K [4] and XMediaNet [27]. The statistics of the four datasets are summarized in Table 2.

5.1.2 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 [13, 34, 48]. Specifically, we compute the mAP scores on the ranked lists of the retrieved results for two different cross-modal retrieval tasks: retrieving text samples using image queries (Image→Text) and retrieving image samples using text queries (Text→Image). The cosine distance is adopted to measure the similarity of features. To calculate mAP, we first evaluate the

Table 2: General statistics of the four datasets used in our experiments, where '/' in the second column denotes the number of train/valid/test set, d_v and d_t are the dimensions of image and text features obtained by VGGNet and Bi-GRU, respectively.

Dataset	Instances	Labels	d_v	d_t
Wikipedia	2173/231/462	10	4096	300
Pascal-Sentence	800/100/100	20	4096	300
NUS-WIDE-10K	8000/1000/1000	10	4096	300
XMediaNet	32000/4000/4000	200	4096	300

average precision (AP) of a set of R retrieved items by:

$$AP = \frac{1}{T} \sum_{r=1}^{R} P_r \times \delta(r) \tag{20}$$

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. The mAP can be calculated by averaging the AP values over all queries.

5.1.3 Implementation Details. For image, we utilize pretrained VGG-19 [32] to extract a 4096-dimensional feature vector from the fc7 layer as the original image feature. For text, we embed each token into 300-dimensional word embedding by GloVe [28] pretrained on the CommonCrawl dataset, and then use a single-layer bidirectional GRU [5] with 512-dimensional hidden states to get the original text feature. To learn a common representation space for image and text modalities, we employ two fully-connected layers with the Rectified Linear Unit (ReLU) [22] active function for each modality. The numbers of the hidden units for the two layers are 2048 and 1024, respectively. The initial prototypes are randomly initialized with the dimension of 1024, and then jointly trained with the common representations. The entire network is optimized by Adam update rule [15] with learning rate 10^{-4} and mini-batch 200.

Table 3: Average mAP scores (mean ± standard deviation) with imbalanced training data on two benchmark datasets.

percentage	Wikipedia				Pascal-Sentence			
	MS^{2}GAN	DAVAE	PAN$_{knn}$	PAN$_{krp}$	MS^{2}GAN	DAVAE	PAN$_{knn}$	PAN$_{krp}$
30%M, 70%I, 0%T	0.425±0.021	0.453±0.016	0.470±0.009	0.477±0.006	0.466±0.025	0.583±0.022	0.655±0.017	0.671±0.014
30%M, 35%I, 35%T	0.439±0.015	0.455±0.009	0.473±0.011	0.481±0.004	0.521±0.017	0.613±0.019	0.647±0.014	0.673±0.015
30%M, 0%I, 70%T	0.417±0.019	0.448±0.018	0.462±0.010	0.473±0.007	0.495±0.024	0.606±0.021	0.642±0.012	0.660±0.013
50%M, 50%I, 0%T	0.452±0.013	0.462±0.011	0.475±0.007	0.480±0.005	0.522±0.016	0.629±0.017	0.659±0.015	0.672±0.013
50%M, 25%I, 25%T	0.461±0.009	0.473±0.010	0.480±0.005	0.491±0.003	0.576±0.020	0.644±0.015	0.661±0.013	0.683±0.009
50%M, 0%T, 50%T	0.433±0.016	0.465±0.021	0.471±0.006	0.475±0.004	0.548±0.019	0.618±0.014	0.652±0.008	0.667±0.016
full data	0.482±0.003	0.485±0.006	0.489±0.002	0.489±0.002	0.664±0.007	0.673±0.010	0.688±0.005	0.688±0.005

Table 4: Average mAP (mean ± standard deviation) scores by directly removing the excessive data on Wikipedia dataset.

percentage	DAVAE	PAN
30%M	0.442± 0.018	0.451± 0.013
50%M	0.455± 0.012	0.461± 0.007
full data	0.485± 0.006	0.489± 0.002

Table 5: Average mAP (mean ± standard deviation) scores by directly removing the excessive data on Pascal-Sentence dataset.

percentage	DAEVE	PAN
30%M	0.562± 0.017	0.589± 0.009
50%M	0.587± 0.011	0.621± 0.008
full data	0.673± 0.010	0.688± 0.005

5.2 Experiments on Conventional Cross-modal Retrieval

5.2.1 Comparison with Representative Methods. To verify the effectiveness of our proposed method for conventional cross-modal retrieval, we conduct experiments on three widely-used benchmark datasets. We compare PAN with ten representative baseline methods, including three shallow learning methods, namely CCA [12], KCCA [11] and JRL [47], and seven deep learning methods, namely Corr-AE [6], CMDN [25], MCSM [27], ACMR [34], CM-GANs [26], DSCMR [48] and MS^{2}GAN [38]. Table 1 reports the mAP scores of our PAN model and the comparative methods.

From the results, we can see that deep learning methods performs obviously better than shallow learning methods, showing the powerful ability of deep neural networks to learn non-linear cross-modal correlations. On the basis of deep neural networks, some methods introduce adversarial learning to generate cross-modal indistinguishable representations, and have achieved the SOTA results currently (*i.e.*, MS^{2}GAN). Compared with these methods, PAN adaptively learns the unified prototypes to explore the cross-modal semantic associations of multi-modal data, achieving the best results on all of the three datasets. Specifically, our PAN outperforms the previous best model, i.e., MS^{2}GAN [38],with improvements 2.4%, 0.9% and 1.4% in terms of average mAP scores on Pascal-Sentence, NUS-WIDE-10K and XMediaNet datasets, respectively.

5.3 Experiments on Imbalanced Training Data

In this section, we conduct the robustness experiment to test the performance of PAN in handling modality-imbalanced training data. We first introduce a dataset split scheme to construct modality-imbalanced training data. Then, we compare our proposed PAN with several baselines.

5.3.1 Dataset Split Scheme. To perform cross-modal retrieval with modality-imbalanced training data, we first introduce a dataset split scheme of the training set. Specifically, we randomly select a certain proportion of the paired multi-modal features from the training set, and then select single-modal features in the remaining set. For example, there are 50% training samples with both image features and text features, 25% samples with only image features and the rest of 25% samples with only text features, we denote this setting by (50%M,25%I,25%T). .

5.3.2 Comparison with Baselines. Since most existing methods can be directly trained on imbalanced data, we first select the existing SOTA cross-modal retrieval method (*i.e.*, MS^{2}GAN [38]) as a baseline method to evaluate the impact of imbalanced data. Then, we compare the proposed PAN with DAVAE [14], which utilizes a variational autoencoder to reconstruct balanced representations with semantic consistency to tackle the imbalanced problem. Note that $k = 0$ is equivalent to not performing imbalanced data alignment, and the effect of k will be analyzed in Figure 3. We repeat each experiment five times, and report the results in Table 3.

From the results, we can see that the existing SOTA method has a significant performance decline in the face of imbalanced data. Moreover, when the proportion of imbalanced data increases, the performance decreases more obviously. We can also see that DAEVE and PAN achieve significant performance improvements by reconstructing balanced representations. Compared with DAEVE, the superiority of our PAN demonstrates the importance of preserving modality heterogeneity during the reconstruction process. In particular, compare the two variants of PAN, *i.e.*, PAN$_{knn}$ and PAN$_{krn}$, we can find that the result of using k-reciprocal neighbors for imbalanced data alignment is always better than k-nearest neighbors by

Table 6: Performance comparison on test queries from both known (Wikipedia) and unknown categories (XmediaNet).

	MS^{2}GAN			PAN		
new class	Image→Text	Text→Image	Average	Image→Text	Text→Image	Average
class1	0.449	0.431	0.440	0.507	0.467	0.487
class2	0.452	0.419	0.435	0.516	0.482	0.499
class3	0.455	0.424	0.439	0.508	0.469	0.488
class4	0.442	0.429	0.436	0.502	0.460	0.481
class5	0.457	0.433	0.445	0.515	0.480	0.497
without new class	0.502	0.462	0.482	0.510	0.468	0.489

Table 7: Performance comparison on test queries from both known (NUS-WIDE) and unknown categories (XmediaNet).

	MS^{2}GAN			PAN		
new class	Image→Text	Text→Image	Average	Image→Text	Text→Image	Average
class1	0.511	0.530	0.520	0.596	0.574	0.585
class2	0.501	0.527	0.514	0.603	0.582	0.592
class3	0.509	0.526	0.517	0.588	0.571	0.580
class4	0.506	0.525	0.515	0.591	0.574	0.582
class5	0.507	0.531	0.519	0.593	0.576	0.584
without new class	0.568	0.574	0.572	0.590	0.571	0.581

Table 8: The tradeoff between acceptance rate AR (%) and rejection rate RR(%) for image modality and text modality with different method. Different rows represent different thresholds.

Image Modality					Text Modality				
	MS^{2}GAN		PAN			MS^{2}GAN		PAN	
ϵ_v	AR	RR	AR	RR	ϵ_t	AR	RR	AR	RR
0.10	100.0	7.5	100.0	100.0	0.10	100.0	0.0	100.0	4.3
0.15	99.7	23.4	98.2	100.0	0.15	93.0	1.8	97.3	17.9
0.20	99.6	39.0	94.3	100.0	0.20	79.5	5.3	94.6	24.8
0.25	99.2	64.7	84.8	100.0	0.25	70.0	13.2	81.3	76.8
0.30	98.7	73.6	74.5	100.0	0.30	69.3	14.0	70.6	79.1
0.35	98.3	84.0	53.4	100.0	0.35	64.7	21.0	66.7	83.2
0.40	97.1	94.3	32.6	100.0	0.40	42.8	26.4	48.2	88.4
0.45	92.5	97.4	11.5	100.0	0.45	11.9	32.7	37.9	89.7
0.50	87.3	100.0	0.4	100.0	0.50	0.0	75.0	13.8	100.0

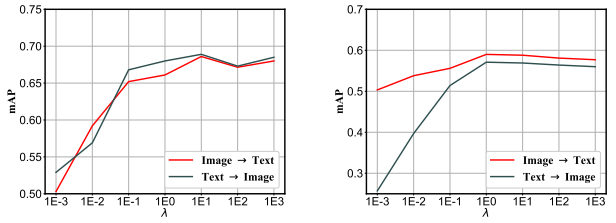
imposing stricter constraints on the nearest neighbor. Furthermore, PAN not only achieves much higher mAP scores but also shows more stable results with smaller variances. This again demonstrates the robustness of PAN in dealing with hybrid-modality training data.

In addition, we further compared the performance of PAN and DAEVE without the usage of excessive data. We report the average mAP scores in Table 4 and Table 5. By comparing Table 3 with Table 4 and Table 5, we can see that the performance of PAN and DAEVE without the usage of excessive data drop significantly in all the evaluations. This shows that exploring imbalanced training data is of great importance to improve the performance in the situation that the imbalanced data is more easier to collect.

5.4 Experiments on Test Queries from Unknown Category

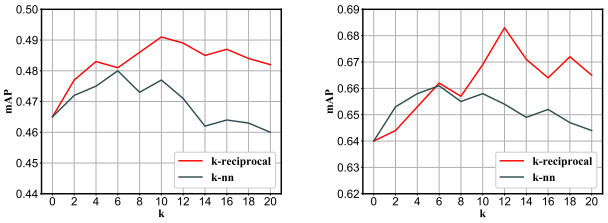
We further conduct experiment to test the robustness of PAN in handling test queries from unknown category. Assume we have trained a cross-modal retrieval system on dataset A, then we use two test sets (test sets of both dataset A and dataset B) to evaluate this retrieval system. The category of test samples in dataset B is not in the category of dataset A and thus can be viewed as unknown category. To prove the robustness of PAN, we design experiments from two perspectives. First, we carry out outlier analysis experiment to evaluate whether the model can recognize the test queries from unknown category. Then, we conduct class-incremental experiment to test the retrieval performance under the situation of partial test data coming from unknown category.

5.4.1 Outlier Analysis. The test samples of dataset B are from unknown category, they should be viewed as outliers and then be rejected by this network. The rejected samples should be inferred to obtain a more reasonable representation for cross-modal retrieval. Meanwhile, the test samples of dataset A should be accepted since they are from the same categories as the training data. We use two measurements, acceptance rate (AR) and rejection rate (RR), to evaluate the performance. AR denotes the percentage of the accepted samples in dataset A, which means how many test samples of dataset A have been accepted. RR denotes the percentage of the rejected samples in dataset B, which means how many test samples of dataset B have been rejected. We adopt the most frequently used threshold-based rejection strategy [45], *i.e.*, if the minimum distance from a sample to all the prototypes is larger than the pre-defined threshold (ϵ_v for image modality and ϵ_t for text modality), it will be accepted, otherwise it will be rejected. Actually, the rejection



(a) Pascal-Sentence (b) NUS-WIDE-10K

Figure 2: Parameter analysis of λ on conventation cross-modal retrieval



(a) Wikipedia, 50%, 25%I, 25%t (b) Pascal-Sentence, 50%, 25%I, 25%t

Figure 3: Parameter analysis of k on cross-modal retrieval with hybrid-modality training data.

and acceptance performances are closely coupled, we can only get a tradeoff between them. For comparison, we also test the outlier analysis performance of MS^{2}GAN, in which the rejection strategies are based on the probabilities produced by the softmax layer. We conduct experiment on NUS-WIDE-10K (dataset A) and XmediaNet (dataset B), and the results are shown in Table 8. Note these results are obtained using different thresholds to give the AR-RR tradeoffs for both image and text modalities.

From the results, we can see that the outlier analysis on the image modality is obviously better than that of the text modality, as the images from known categories have a greater probability of being accepted, while those from unknown category images have a greater probability of being rejected. We can also see that the softmax-based MS^{2}GAN is confused by the XMediaNet test samples, high AR and high RR can not coexist. This indicates that the softmax-based model is not robust in outlier detection. Contrastly, our PAN model can achieve better rejection performance and simultaneously keep satisfactory acceptance rate. For example, while 100% image samples and 83.2% text samples from the XMediaNet dataset being rejected, we can still keep 100% image samples and 66.7% text samples from the NUS-WIDE-10K dataset being accepted. This is a significant advantage compared with softmax-based approach, showing the robustness of our proposed network.

5.4.2 Class-incremental Cross-modal Retrieval. We conduct experiment on Wikipedia, NUS-WIDE-10K and XMediaNet dataset to demonstrate the superiority of PAN for class-incremental learning. We treat test samples from Wikipedia, NUS-WIDE-10K dataset as the known categories data, and choose one category from XMediaNet as the unknown category. We train PAN on the Wikipedia and NUS-WIDE-10K separately, then we feed the test data from both known and unknown categories (which should be learned incrementally) to the trained PAN and obtain their representations in the common space. Based on the outlier analysis experiment, we use the inferred representation based on Equation (18) as the final representation for outliers.

Table 6 and Table 7 show the class-incremental retrieval results compared with softmax-based MS^{2}GAN . From the results, we can see that PAN still keeps high performance when extended to the unknown category, while MS^{2}GAN encounters significant performance decline. In this class-incremental learning process, we did not re-train any part of the network. This further demonstrates the robustness of PAN to test queries from unknown category.

5.5 Parameter Analysis

The parameters are analyzed in this section. We evaluate the influence of λ on conventional cross-modal retrieval, and evaluate k on imbalanced cross-modal retrieval. Figure 2 shows the impact of λ on Pascal-Sentence and NUS-WIDE-10K dataset. It can be seen that the mAP first increase with the growth of λ, and then begins a slow decline after λ surpasses a threshold. The best parameter setting of λ are 10 and 1 on the two datasets. The impacts of the size of k-nearest neighbors and k-reciprocal neighbors are shown in Figure 3. When k is equal to 0, the imbalanced data alignment is not considered, the model has the worst results on all datasets. This confirms that imbalanced training data does impair the performance of cross-modal retrieval. It turns out that if we impose stricter constraints on the nearest neighbor, using k-reciprocal neighbors for imbalanced data alignment is always better than k-nearest neighbors. The results also show that if the value of k is set too large, it will increase the probability of false neighbors belonging to different categories and cause the performance degradation.

6 CONCLUSION

In this paper, we propose a prototype-based adaptive network (PAN) to handle modality-imbalanced training data and test queries from unknown category. We develop a prototype-based representation learning method to jointly learn the common representations across different modalities and the unified prototypes for each category by designing invariance loss and discrimination loss with prototypes as anchors. Furthermore, we propose a prototype propagation strategy to reconstruct imbalanced multi-modal samples, which can preserve the semantic consistency and modality heterogeneity. Experimental results demonstrate the effectiveness of our proposed method in conventional cross-modal retrieval, and further robustness tests show the superiority of our method.

ACKNOWLEDGEMENTS

This work was supported in part by the Ministry of Science & Technology of China under Grants #2020AAA0108401 and #2020AAA0108405, NSFC Grants #11832001 and #71621002, and CAS Strategic Priority Research Program under Grant #XDA27030100.

REFERENCES

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

[2] Jingze Chi and Yuxin Peng. 2018. Dual Adversarial Networks for Zero-shot Cross-media Retrieval.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 663–669.

[3] Jingze Chi and Yuxin Peng. 2019. Zero-shot cross-media embedding learning with dual adversarial distribution network. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 4 (2019), 1173–1187.

[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. 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*. 1–9.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the ACM international conference on Multimedia*. 7–16.

[7] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel. 2015. Person re-identification ranking optimisation by discriminant context information analysis. In *Proceedings of the IEEE International Conference on Computer Vision*. 1305–1313.

[8] 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.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

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

[11] 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.

[12] Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in Statistics*. Springer, 162–190.

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

[14] 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 ACM international conference on Multimedia*. 3283–3291.

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

[16] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data learning of new tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 646–651.

[17] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*. 4247–4255.

[18] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. 2018. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4242–4251.

[19] 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).

[20] Ruoyu Liu, Yao Zhao, Liang Zheng, Shikui Wei, and Yi Yang. 2017. A new evaluation protocol and benchmarking results for extendable cross-media retrieval. *arXiv preprint arXiv:1703.03567* (2017).

[21] Xin Liu, Yiu-ming Cheung, Zhikai Hu, Yi He, and Bineng Zhong. 2020. Adversarial Tri-Fusion Hashing Network for Imbalanced Cross-Modal Retrieval. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2020).

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

[23] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning*.

[24] Kai Niu, Yan Huang, and Liang Wang. 2020. Re-ranking image-text matching by adaptive metric fusion. *Pattern Recognition* (2020), 107351.

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

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

[27] 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.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*. 1532–1543.

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

[30] 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 ACM international conference on Multimedia*. 251–260.

[31] Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Cannim. 2020. Web Table Retrieval using Multimodal Deep Learning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1399–1408.

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

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the 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 ACM international conference on Multimedia*. 154–162.

[35] 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).

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

[37] 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 ACM international conference on Multimedia*. 1437–1445.

[38] 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* (2020), 107335.

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

[40] Lin Wu, Yang Wang, and Ling Shao. 2018. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing* 28, 4 (2018), 1602–1612.

[41] Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2665–2672.

[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 AAAI Conference on Artificial Intelligence*. 2156–2162.

[43] Xing Xu, Huimin Lu, Jingkuan Song, Yang Yang, Heng Tao Shen, and Xuelong Li. 2019. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE Transactions on Cybernetics* 50, 6 (2019), 2400–2413.

[44] Xing Xu, Jingkuan Song, Huimin Lu, Yang Yang, Fumin Shen, and Zi Huang. 2018. Modal-adversarial semantic learning network for extendable cross-modal retrieval. In *Proceedings of the International Conference on Multimedia Retrieval*. 46–54.

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

[46] Yang Yang, De-Chuan Zhan, Xiang-Rong Sheng, and Yuan Jiang. 2018. Semi-Supervised Multi-Modal Learning with Incomplete Modalities.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 2998–3004.

[47] 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.

[48] 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.

[49] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1318–1327.