Bi-CMR: Bidirectional Reinforcement Guided Hashing for Effective Cross-Modal Retrieval

Tieying Li[1], Xiaochun Yang[1*], Bin Wang[1], Chong Xi[1], Hanzhong Zheng[1], Xiangmin Zhou[2]

[1]Northeastern University, China [2]RMIT University, Australia
{tieying, hanzhong}@stumail.neu.edu.cn, {yangxc, binwang}@mail.neu.edu.cn, 15734022139@163.com,
Xiangmin.zhou@rmit.edu.au

Abstract

Cross-modal hashing has attracted considerable attention for large-scale multimodal data. Recent supervised cross-modal hashing methods using multi-label networks utilize the semantics of multi-labels to enhance retrieval accuracy, where label hash codes are learned independently. However, all these methods assume that label annotations reliably reflect the relevance between their corresponding instances, which is not true in real applications. In this paper, we propose a novel framework called Bidirectional Reinforcement Guided Hashing for Effective Cross-Modal Retrieval (Bi-CMR), which exploits a bidirectional learning to relieve the negative impact of this assumption. Specifically, in the forward learning procedure, we highlight the representative labels and learn the reinforced multi-label hash codes by intra-modal semantic information, and further adjust similarity matrix. In the backward learning procedure, the reinforced multi-label hash codes and adjusted similarity matrix are used to guide the matching of instances. We construct two datasets with explicit relevance labels that reflect the semantic relevance of instance pairs based on two benchmark datasets. The Bi-CMR is evaluated by conducting extensive experiments over these two datasets. Experimental results prove the superiority of Bi-CMR over four state-of-the-art methods in terms of effectiveness.

Introduction

With the blooming of multimodal data (e.g., images and texts) in the areas of search engines and social networks, information retrieval across different types of data has attracted wide attention. Accordingly, it gives rise to the emerging real-world application of cross-modal retrieval, which aims to search the semantically relevant instances in all the modalities (e.g., images and texts) given a query of one modality. In order to satisfy the requirements of low storage, high query speed in real-world applications, hashing has gained increasing attention in the field of cross-modal retrieval due to its capability of transforming variant modal instances to uniform binary codes. However, as instances from different modalities are heterogeneous in terms of feature representation and character distribution, exploiting multi-labels of instances from variant modalities to retrieve multimodal data effectively is a big challenge.

*Corresponding Author

Existing cross-modal hashing techniques can be mainly classified into two lines: unsupervised learning (Song et al. 2013; Zhou, Ding, and Guo 2014; Wang et al. 2015; Ding et al. 2016; Hu et al. 2019) and supervised learning (Yang et al. 2017; Jiang and Li 2017; Liong et al. 2017; Li et al. 2018; Zhu et al. 2021). Supervised learning methods exploit labels or the semantic affinities of training data to achieve a performance superior to the unsupervised ones. Early stage supervised learning approaches use label annotations to guide cross-modal hash learning (Liong et al. 2017; Jiang and Li 2017). Multi-label network-based method (Li et al. 2018) and its variant (Xu et al. 2020) improve the supervised learning by integrating a self-supervised semantic network to capture the semantic information from multi-label annotations, and supervise modality-feature learning. All these methods assume that label annotations can reliably reflect the relevance between their corresponding instances. However, this assumption conflicts with human perception in real applications. Consider an example in a benchmark dataset MIRFlickr25K (Huiskes and Lew 2008) for cross-modal retrieval as shown in Figure 1. Given two images with the same labels and a query text shown in the figure, existing methods believe these images are equally relevant to the query. However, it is clear that Figure 1(a) is relevant to the query text, while Figure 1(b) is not, since both the query text and the image in Figure 1 (a) are mainly about "flower", while the image in Figure 1 (b) is mainly about "structure". There is a semantic gap between the low-level annotations and the high-level semantic understanding of images.

To address the problem of effective cross-modal retrieval, we need to learn a new scheme which well narrows the semantic gap between multi-label annotations and instance relevance in the learning phase. A popular scheme of evaluating instance relevance in existing supervised methods is pairwise multi-label similarity matrix (PMLSM) denoted as S: $\{S_{ij}\}$. Here, in the learning phase, S_{ij} indicates if the corresponding instances x_i and x_j are relevant or not(Liong et al. 2017; Jiang and Li 2017; Li et al. 2018). Two instances x_i and x_j are relevant if they share any label annotations, and their relevance is set as $S_{ij} = 1$. Otherwise, they are irrelevant with $S_{ij} = 0$. However, all these PMLSM-based methods highly rely on an accurate similarity matrix S to guide model learning. This kind of similarity measure based on label annotation intersection is also used

*Text query: Beautiful **flowers** bloom in front of the **castle.***

(a)

(b)

	people	flower	structure	tree	…
Text-label:	0	**1**	1	0	0
image-label (a):	0	**1**	1	0	0
image-label (b):	0	1	**1**	0	0

Figure 1: An ignored mismatching example.

as ground truth in the testing phase to measure the relevance of the retrieval results, resulting in wrong evaluations of all PMLSM-based methods. To overcome the above issue, we manually mark pairwise ground truth matrix $G : \{G_{ij}\}$, where $G_{ij} = 1$ if the corresponding instances x_i and x_j are semantically relevant, and $G_{ij} = 0$ otherwise. An instance pair $\langle x_i, y_i \rangle$ is *misjudged* if the ground truth shows $G_{ij} = 0$ but its $S_{ij} = 1$. Statistically, using PMLSM-based methods, 24.97% irrelevant instance pairs are misjudged as relevant for MIRFlickr25K, while around 13.51% pairs are misjudged in NUS-WIDE.

Motivated by the limitation of existing approaches, we propose a Bidirectional Reinforcement Guided Hashing method for Effective Cross-Modal Retrieval (Bi-CMR). The superiority of using Bi-CMR is twofold. First, instances with more common label annotations could be more similar than those with less common labels. Second, for the instances with same representative labels, we could identify the ones that are dissimilar with each other at human perception level. Our Bi-CMR achieves these goals by bidirectional learning. Intra-modal semantic information can be forwarded to reinforce the hash codes of multi-labels and further adjust similarity matrix, while the adjusted similarity matrix is used backward to guide the instance hash codes learning for multimodal instance pairs. Specifically, in the forward learning procedure, a deep self-supervised reconstruction model is proposed over each modality network to learn more accurate hash codes for each intra-model instance, and further assist their multi-label hash code learning. For an instance, we highlight the importance of its representative labels based on its semantics in forward learning. Furthermore, each multi-label annotation vector and its reinforced label hash codes are concatenated to adjust similarity matrix and obtain a new reinforced similarity evaluation, which decreases the false positives. In the backward learning procedure, under the guide of the adjusted similarity matrix, the reinforced multi-label hash codes are used to reduce the mismatching under the situation discussed above and guide the accurate learning of cross-modal instances hash codes. We summarize our contributions as follows:

- We are the first to realize that the assumption "label annotations reliably reflect the instance relevance" conflicts with human perception, thus, the existing relevance measurement based on PMLSM methods are inappropriate. We propose a new evaluation calculation to guide the learning of instance hash codes, which is consistent with human perception.
- We propose a novel bidirectional reinforcement guided hashing method, which reinforces hash code learning in a mutual promotion way. While the semantic correlation of intra-modal information is used forward to reinforce the learning of the multi-label hash codes so that the semantic similarity based on labels can be tuned gradually, and the tuned similarity value is used backward to guide the learning of the cross-modal instance hash codes.
- We manually mark relevant instance pairs over two benchmark datasets to measure retrieval accuracy. Extensive experiments are conducted over these two datasets to evaluate the high effectiveness of Bi-CMR.

Related work

Various hashing methods have been proposed for cross-modal retrieval. Major techniques can be roughly divided into two categories: unsupervised methods and supervised methods. Unsupervised methods (Zhou, Ding, and Guo 2014; Wang et al. 2015; Ding et al. 2016; He et al. 2017; Li et al. 2019) focus on learning hash functions by exploiting the relationship of instances with unlabeled data. Latent semantic sparse hashing (LSSH) (Zhou, Ding, and Guo 2014) uses sparse coding to capture the image structure and models the text modality via matrix factorization. Semantic topic multimodal hashing (STMH) (Wang et al. 2015) uses multimodal hashing and learns the relationship of two modalities in latent semantic space. All these methods ignore the value of the semantic labels, leading to inferior performance.

Supervised methods (Wang et al. 2015; Wu et al. 2015; Lin et al. 2015; Yang et al. 2017; Liong et al. 2017; Jiang and Li 2017; Xu et al. 2017; Li et al. 2018; Ye and Peng 2018; Gu et al. 2019) leverage the semantic labels of image-text pairs as the supervision to guide the hash code learning and boost performance. For example, deep cross-modal hashing (DCMH) (Wang et al. 2015) establishes an end-to-end hashing framework with deep neural network, which conducts feature learning and hash code learning simultaneously. Pairwise Relationship Guided deep hashing (PRDH) (Yang et al. 2017) explores two pairwise constraints in inter-modalities and intra-modalities. Self-Supervised Adversarial hashing (SSAH) (Li et al. 2018) preserves multi-label information to maximize the semantic relevance across modalities. AGAH (Gu et al. 2019) adopts adversarial learning with a multi-label attention map to preserve multi-label information and minimize the semantic gap among modalities. However, all these methods are limited to training data pairs, which lacks generality. Multi-label networks are used to learn semantic features of multi-label annotations (Li et al. 2018; Xu et al. 2020), which achieves good results. However, simply raising the dimensionality of multi-labels does not enrich the semantics of multi-labels. Our Bi-CMR adopts bidirectional reinforcement learning with a self-supervised reconstruction to acquire reinforced representation of multi-labels for effective instance matching.

Problem and Preliminary

Given a database D containing n multimodal data, each of which has the form of $\langle x_i, y_i \rangle$ $(1 \leq i \leq n)$, where x_i is a certain modal instance and y_i=[$y_{i1}, \ldots, y_{iC}$] is its multi-label annotation vector. Given an annotation y_{iv} of modal instance x_i, each annotation $y_{iv} = 1$ if x_i belongs to the v-th class, and $y_{iv} = 0$ otherwise $(1 \leq v \leq C)$. The cross-modal retrieval problem is to find the relevant instances in D to a query instance q.

To simplify the presentation, we focus on the cross-modal retrieval for bi-modal data (i.e., images and texts). Without loss of generality, our task can be easily extended to the scenarios with multiple modalities. In particular, we aim to learn the hash codes for modalities. Let D be a set including an image set D_I and a text set D_T labelled by multi-labels D_L. For an instance x_i in D_I (or D_T), we learn an l-bit hash code $\boldsymbol{h}_i^{I,T} = \{h^{I,T}(x_i) \mid x_i \in D_{I,T}, h^{I,T}(x_i) \in \{-1, 1\}^l\}$. For a multi-label annotation vector y_i, we also learn $\boldsymbol{h}_i^{L} = \{h^{L}(y_i) \mid y_i \in D_L,\ h^{L}(y_i) \in \{-1, 1\}^l\}$. The learned hash functions $h^{I,T}(\cdot)$ are used to generate l-bit hash codes $\boldsymbol{h}_i^{I,T}$ for query and database instances in both modalities. We adopt Hamming distance to determinetong relevance between the hash codes of query and those of database instances. The notations used in this paper and their descriptions are summarized in Table 1 for easy reference.

Table 1: Summary of the main notations.

Not.	Definition and description
D	The database that contains different modal data
x_i	The i-th instance in D
y_i	The multi-label annotation vector of x_i
C	Category number of each multi-label annotation
S_{ij}	Semantic similarity between x_i and x_j
l	The length of learned hash codes
$\boldsymbol{h}^{I,T,L}$	The hash code of images, texts, and multi-labels
$\boldsymbol{f}^{I,T}$	The learned feature vector of images and texts
$dist(\boldsymbol{a}, \boldsymbol{b})$	The hamming distance of hash codes $\boldsymbol{a}$ and $\boldsymbol{b}$
q	The query instance
k	Number of nearest neighbors to retrieve

The Proposed Bi-CMR

This section proposes a bidirectional reinforcement guided hashing framework, Bi-CMR, for the hash code learning.

Framework Overview

Figure 2 depicts our proposed Bi-CMR framework. We use two types of networks: (1) Image, Text Net (Jiang and Li 2017) for learning high-dimensional features and hash codes of image and text modalities; and (2) Multi-label Net (Li et al. 2018; Xu et al. 2020) for generating guide information from multi-label annotations. With the support of these networks, we propose a bidirectional reinforcement module including forward learning and backward learning, which decreases the issue of false positive and false negative. For the forward learning, the semantic correlation of intra-modal information can be used to reinforce the hash codes of multi-labels so that the semantic similarity between every instance pair S_{ij} can be tuned gradually. For the backward learning, the tuned S_{ij} and the reinforced label hash codes are used to guide the learning of the modal hash codes. The bidirectional learning could be run over many iterations. As a result, the hash codes of relevant (irrelevant) instances will be more similar (dissimilar) through bidirectional reinforcement learning, which bridges the semantic gap in cross-modal retrieval.

Forward Learning: Tuning Similarity Matrix

To tune S_{ij}, we need to do the followings: (1) using semantic correlation to reinforce the hash codes of multi-labels, and (2) designing a new formula to calculate S_{ij} so that hash codes of two instances x_i and x_j become more similar (dissimilar) if they are relevant (irrelevant) for each iteration.

Reinforcement from Intra-modality to Label Hash Codes To reinforce multi-label hash codes using semantic correlation, we need to capture the semantic information of instances as much as possible. However, hash mapping from modality feature vectors to hash codes inevitably results in the loss of semantics. Thus, we need to address this problem for ensuring the quality of the reinforcement process. We propose a two-stage stategy to overcome this problem. First, we minimize the semantic loss of intra-modal hash mapping. Then, we reinforce the hash codes of multi-labels.

We develop a deep self-supervised reconstruction (SSR) mechanism to reduce the semantic loss of intra-modality. Extending the idea of reconstruction in (Wang et al. 2014), SSR is a two-layer fully-connected network with an addition of SSR loss on the $Fc2'$ layer as in Figure 2. Intuitively, if no semantic loss occurs in the hashing process, the outputs of network reconstructed from each modal hash code should be similar to its originl feature vector. Thus, we adopt well known *Euclidean norm* as a metric, which can evaluate the change of modal feature vector caused by the reconstruction without any supervision. Let $f^I(x_i)$ and $f^T(x_i)$ (abbr. $\boldsymbol{f}_i^I$ and $\boldsymbol{f}_i^T$) denote the original feature vector of instance x_i^I and x_i^T, respectively. The SSR loss function of image and text modalities is formulated as follows:

$$\mathcal{L}_{ssr} = \sum_{i=1}^{n} (||\boldsymbol{f}_i^I - \boldsymbol{f'}_i^I||_2^2 + ||\boldsymbol{f}_i^T - \boldsymbol{f'}_i^T||_2^2), \qquad (1)$$

where $\boldsymbol{f}_i^I \in \mathbb{R}^{d_I}$ (or $\boldsymbol{f}_i^T \in \mathbb{R}^{d_T}$), $\boldsymbol{f'}_i^I \in \mathbb{R}^{d_I}$ (or $\boldsymbol{f'}_i^T \in \mathbb{R}^{d_T}$) is the reconstructed feature vector generated by the image (or text) SSR mechanism, and d_I (or d_T) is the dimension of image (or text) feature vectors.

We use the intra-modal semantic correlation to reinforce the hash codes of multi-labels. According to the smoothness assumption (Zhu, Lafferty, and Rosenfeld 2005), the instances close to each other are more likely to own common representative labels. We use the semantic correlation of instances with less semantic loss to reinforce the hash codes of multi-labels. Let $a_{ij}^{I,T} = e^{cos(\boldsymbol{h}_i^{I,T}, \boldsymbol{h}_j^{I,T})}$ measure the similarity of two hash codes with the same modal, where $cos(\boldsymbol{a}, \boldsymbol{b})$ is used to measure the semantic correlation of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ since it focuses more on the vectors' differences in

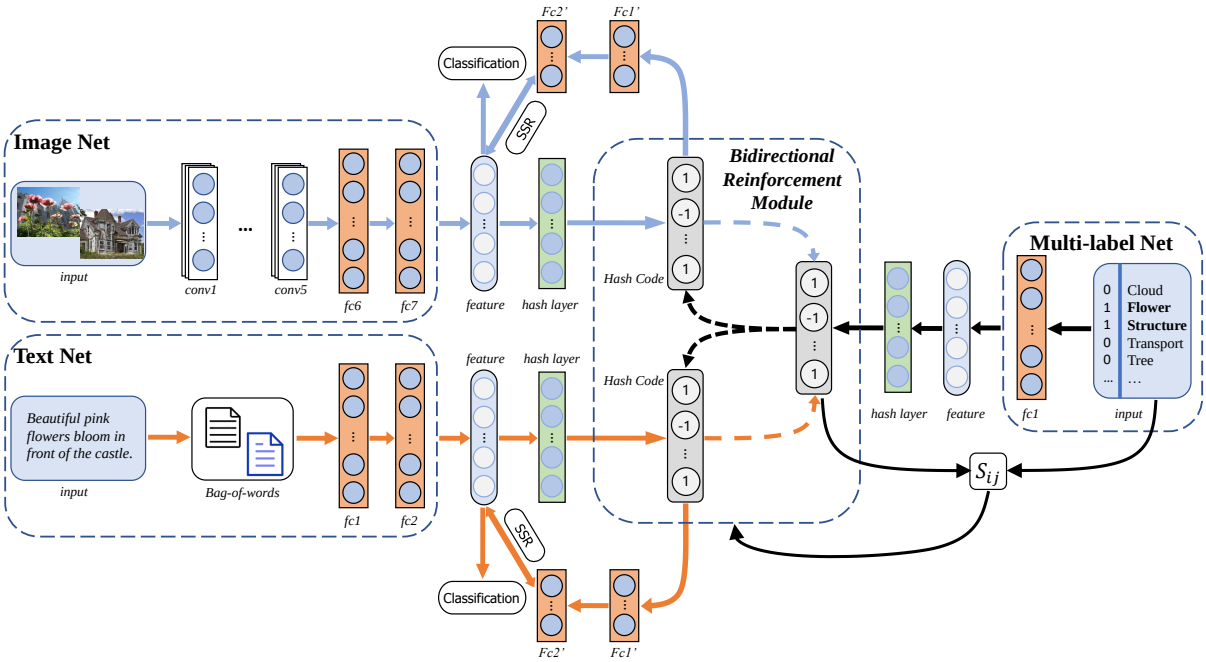


Figure 2: Framework of our proposed Bi-CMR.

distribution, and has the advantage of being stable and accurate for the similarity evaluation of different dimensions. Similar we specify $a_{ij}^{L} = e^{cos(\boldsymbol{h}_i^L, \boldsymbol{h}_j^L)}$. Obviously, if intra-modal instances x_i and x_j have a high degree of similarity as measured by $a_{ij}^{I,T}$, $\boldsymbol{h}_i^L$ and $\boldsymbol{h}_j^L$ should be close to each other. Thus, we minimize the following objective function:

$$\mathcal{L}_{m2l} = \sum_{ij}(\eta * \frac{a_{ij}^{I}}{a_{ij}^{L}} + (1-\eta) * \frac{a_{ij}^{T}}{a_{ij}^{L}}), \tag{2}$$

where $0 < \eta < 1$ is the weighting factor for balancing the contribution of each modality to reinforce the hash codes of multi-labels. The larger $a_{ij}^{I,T}$ is, the larger a_{ij}^{L} will be, while irrelevant instances cause less effect.

Adjusted Similarity Matrix Recall that label annotations are not the only factor for evaluating relevance between two instances. We hope each S_{ij} in the similarity matrix S could better reflect the true relevance of instances x_i and x_j. Besides label annotations, we also consider the semantic correlation of instances to determine S_{ij}. For each iteration in forward learning, we concatenate the multi-label annotation vector y_i of an instance x_i and its learned multi-label hash code $\boldsymbol{h}_i^L$ into one label measure vector $\boldsymbol{L}_i^{'} = concat(y_i, \delta \boldsymbol{h}_i^L)$, where δ is used to control the degree of reinforcement. Then for any two label measure vectors $\boldsymbol{L}_i^{'}$ and $\boldsymbol{L}_j^{'}$, we tune $S_{ij} = 1$ if their cosine similarity $cos(\boldsymbol{L}_i^{'}, \boldsymbol{L}_j^{'}) > \lambda$, and $S_{ij} = 0$ otherwise[1]. Here, λ is the relevance threshold for a dataset, which will be evaluated in experiment section. The tuned similarity matrix $S : \{S_{ij}\}$ considers both multi-labels and the reinforced multi-label hash codes. As such, the instances with more common label annotations could be more similar than those with less common ones, and the semantic correlation of instances are captured as well.

[1]Other similarity or distance functions can also be used to determine the similarity between label measure vectors $\boldsymbol{L}_i^{'}$ and $\boldsymbol{L}_j^{'}$.

Backward Learning: Reinforcement from Label to Cross-modal Instance Pairs

Using the reinforced label hash codes $\boldsymbol{h}^L$ that highlight the representative labels with the semantics from intra-modalities, we can reduce the mismatching of instance pairs. Similar to the design principle of $\mathcal{L}_{m2l}$, the loss function of reinforcement from multi-label to modalities is:

$$\mathcal{L}_{l2m} = \sum_{ij}(\frac{a_{ij}^{L}}{a_{ij}^{I}} + \frac{a_{ij}^{L}}{a_{ij}^{T}} + \frac{a_{ij}^{L}}{a_{ij}^{I\&T}}), \tag{3}$$

where $a_{ij}^{I\&T} = e^{cos(\boldsymbol{h}_i^I, \boldsymbol{h}_j^T)}$ measures the similarity of cross-modal instance hash codes. Note that our reinforcement is applied equally for each intra-modality, thus we do not set the intra-modal weighting factor.

Like (Li et al. 2018), we consider the pairwise loss between an instance hash code $\boldsymbol{h}_i^{I,T}$ and its label hash code $\boldsymbol{h}_i^L$. Two modal instance hash codes $\boldsymbol{h}_i^I$ and $\boldsymbol{h}_j^T$ are associated through their label hash codes. Unlike (Li et al. 2018), we use the adjusted similarity matrix to continuously correct the learning of cross-modal hash codes since the accuracy of similarity evaluation is critical for cross-modal learning. Guided by this matrix, the reinforced multi-label hash codes act as a bridge between modalities, achieving the co-learning with modal instances. Thus, we propose a supervised learning of label hash code to enhance the cross-modal matching of $\boldsymbol{h}_i^T$ and $\boldsymbol{h}_j^I$ base on the tuned similarity matrix S. We use the well known likelihood function to express the probability of S_{ij} under the learned hash codes $\boldsymbol{h}_i^L, \boldsymbol{h}_j^L$ as follows:

$$p(S_{ij}|\boldsymbol{h}_i^L, \boldsymbol{h}_j^L) = \sigma(\Omega_{ij}^L)^{S_{ij}}(1-\sigma(\Omega_{ij}^L))^{1-S_{ij}}, \tag{4}$$

where $\sigma(\Omega_{ij}^L) = \frac{1}{1+e^{-\Omega_{ij}^L}}$ is the sigmoid function, and $\Omega_{ij}^L = \frac{1}{2}\boldsymbol{h}_i^L{}^\top \boldsymbol{h}_j^L$. The relationship between their Hamming distance $dist(\boldsymbol{h}_i^L, \boldsymbol{h}_j^L)$ and inner product $\frac{1}{2}\boldsymbol{h}_i^L{}^\top \boldsymbol{h}_j^L$ is formulated as $dist(\boldsymbol{h}_i^L, \boldsymbol{h}_j^L) = \frac{1}{2}(l - \frac{1}{2}\boldsymbol{h}_i^L{}^\top \boldsymbol{h}_j^L)$. Therefore we can use the inner product to measure the hash code similarity. A larger inner product indicates a bigger probability of S_{ij}, thus $\boldsymbol{h}_i^L$ and $\boldsymbol{h}_j^L$ are classified as similar and vice versa.

Similar to Eq. 4, we aim to maximize the summation of the probabilities $p(S_{ij}|\boldsymbol{h}_i^L, \boldsymbol{h}_j^L)$ and $p(S_{ij}|\boldsymbol{h}_i^L, \boldsymbol{h}_j^{I,T})$. Here, $p(S_{ij}|\boldsymbol{h}_i^L, \boldsymbol{h}_j^{I,T})$ determines whether a label hash code $\boldsymbol{h}_i^L$ and an instance hash code $\boldsymbol{h}_j^{I,T}$ can be classified as similar or not. For ease of computation, we use negative log likelihood to express the correlation loss as follows:

$$\begin{aligned}\mathcal{L}_{cor} = \sum_{i,j=1}^{n} ((&log(1+e^{\Omega_{ij}^L}) - S_{ij}\Omega_{ij}^L) \\ &+(log(1+e^{\Omega_{ij}^{L,I}}) - S_{ij}\Omega_{ij}^{L,I}) \\ &+(log(1+e^{\Omega_{ij}^{L,T}}) - S_{ij}\Omega_{ij}^{L,T})),\end{aligned} \tag{5}$$

where $\Omega_{ij}^{L,I} = \frac{1}{2}\boldsymbol{h}_i^L{}^\top \boldsymbol{h}_j^I$, $\Omega_{ij}^{L,T} = \frac{1}{2}\boldsymbol{h}_i^L{}^\top \boldsymbol{h}_j^T$. As a result, the adjusted similarity matrix can guide the learning of modal hash codes with iterations.

Overall Objective Function

We could further enhance the leaning of instance hash code by incorporating the classification mechanism as follows. Based on the reinforced learned instance hash codes, we classify the instances according to C class categories, and generate a learned multi-label notation vector, denoted $\boldsymbol{I}^i$ (or $\boldsymbol{T}^i$) $\in [0,1]^C$. We minimize the cross entropy (Li et al. 2018) between the $\boldsymbol{I}^i$ (or $\boldsymbol{T}^i$) and $\boldsymbol{y}_i$ as follows:

$$\mathcal{L}_c = \sum_{i=1}^{n} -((log(\boldsymbol{I}^i) + log(\boldsymbol{T}^i))\boldsymbol{y}_i). \tag{6}$$

Then, we have the objective function toward the cross-modal function $h(\cdot)$, where θ is the parameter set of Bi-CMR.

$$\underset{\theta}{argmin}\mathcal{L} = \alpha\mathcal{L}_{ssr} + \beta\mathcal{L}_{m2l} + \gamma\mathcal{L}_{l2m} + \mathcal{L}_{cor} + \mathcal{L}_c. \tag{7}$$

Minimizing $\mathcal{L}$ enables Bi-CMR to learn more accurate hash codes and minimize the distances between the hash codes of semantically similar instances.

Experiments

Experiment Setting

We choose two commonly used datasets, MIRFlickr25K (Huiskes and Lew 2008) and NUS-WIDE10.5K, and manually label relevant pairs in each dataset for evaluation.

MIRFlickr25K[2] is a benchmark dataset collected from Flickr. It consists of $25,000$ image-text pairs selected from 24 categories. We keep $20,015$ text instances that have at least one of the top 20 frequent text tags for our test. We randomly select $2,000$ pairs of image-text couples as the query set and the remaining as the retrieval database. NUS-WIDE10.5K is a dateset created by filtering NUS-WIDE [3]. NUS-WIDE contains $269,648$ image-text pairs, each of which is annotated by one or more labels within 81 concepts. Only the pairs belonging to the 21 most frequent categories are selected for our tests. In total, $195,834$ pairs were selected. We randomly select $10,500$ multi-label image-text pairs and keep them uniformly distributed over 21 label categories. We construct two training sets by randomly selecting $10,000$ from MIRFlickr25K and $4,000$ from NUS-WIDE10.5K. $2,000$ pairs are randomly selected as query set and the remaining as the retrieval database.

To evaluate the effectiveness of all the approaches, we manually mark the ground truth relevant instance pairs for each dataset based on human relevance judgments. We conduct a subjective user study to mark the ground truth as in (Zhou et al. 2017), where the reliability of this user study has been proved. Specifically, five postgraduate students majored in computer science participate in the user study. Each individual is given all the instance pairs in the datasets in a random order. After viewing these pairs, they are asked to give a rating score from 1 to 5 indicating if the instance pair is relevant. Here, higher score indicates more relevance. An instance pair with the rating no smaller than 4 is considered as semantically relevant. For each query instance, we use its manually marked relevant instances as ground truth.

Implementation detail and parameter setting The text modality on MIRFlickr25K and NUS-WIDE10.5K is represented as $1,386$-dimensional and $1,000$-dimensional bag-of-words (BoW) vectors respectively. The dimensions of BoW vectors are decided by the high frequency vocabulary defined in all text annotations. The text BoW vectors are fed into two fully-connected layers with the hidden sizes of $1,024$ and $4,096$ to get the final text features. The hidden sizes of the self-supervised reconstruction (SSR) network are set to $1,024$. We extract image features from two datasets using *fc*7 layer CNN-F network (Chatfield et al. 2014) and each image feature is initially described as a $4,096$-dimensional vector. We also use a $4,096$-*fc* layer to construct $4,096$-dimensional multi-label features for all labels. Following (Song et al. 2015; Gu et al. 2019; Yang et al. 2017), we use $tanh$ as the activation function of hash layer outputs, and $sign$ as a function for generating the final hash codes. The hash layer size for the final hash codes is set as $16, 32$, and 64 respectively.

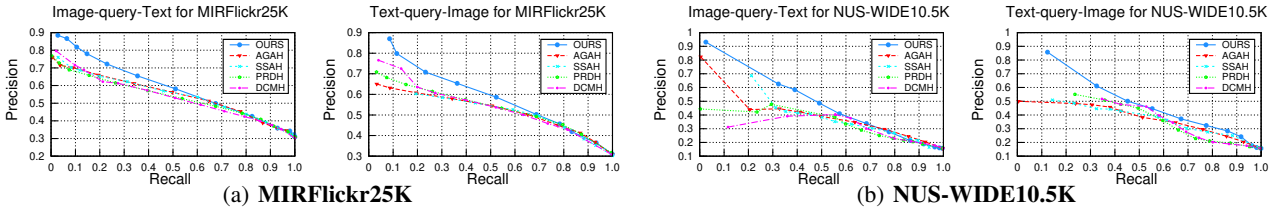
We implement our method in PyTorch, and train our model using the ADAM optimizer with an initialized learning rate of $1e-4$ in image and text network and $1e-3$ in multi-label network with a batch size of 128, respectively. We set the maximum number of epochs as 120 to ensure the convergence. The training time for one epoch does not exceed 2 minutes on a single $RTX2070$ GPU. After every 20 epochs, the learning rate decreases by half.

[2] http://press.liacs.nl/mirflickr/mirdownload.html

[3] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

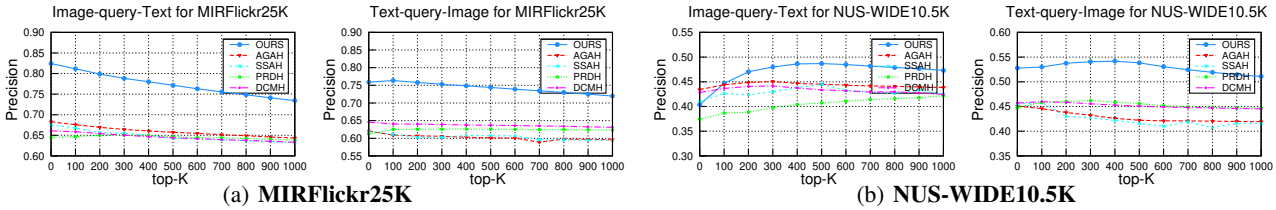
Table 2: MAP comparison results, where the best performance is boldfaced and the runner-up is underlined.

Methods	MIRFlickr25K						NUS-WIDE10.5K					
	IQT			TQI			IQT			TQI		
	16bit	32bit	64bit	16bit	32bit	64bit	16bit	32bit	64bit	16bit	32bit	64bit
DCMH	0.4463	0.4730	0.4878	0.4568	0.4734	<u>0.4835</u>	0.4322	0.4364	0.4394	0.4759	<u>0.4849</u>	<u>0.4884</u>
PRDH	0.4723	0.4815	0.4875	0.4702	0.4724	0.4792	0.4393	0.4315	0.4098	0.4713	0.4837	0.4603
SSAH	0.4760	0.4923	0.4970	<u>0.4854</u>	<u>0.5025</u>	0.4740	0.4645	0.4502	0.4674	0.4867	0.4707	0.4683
AGAH	<u>0.4937</u>	<u>0.4946</u>	<u>0.4975</u>	0.4832	0.4705	0.4808	<u>0.4792</u>	<u>0.4515</u>	<u>0.4707</u>	<u>0.5005</u>	0.4674	0.4699
Bi-CMR	**0.5655**	**0.5822**	**0.5862**	**0.5445**	**0.5553**	**0.5586**	**0.5113**	**0.5118**	**0.5002**	**0.5118**	**0.4971**	**0.5045**



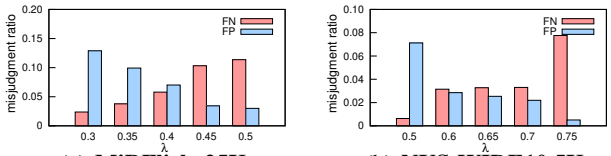
(a) **MIRFlickr25K** (b) **NUS-WIDE10.5K**

Figure 3: Precision-recall curves on two datasets. The baselines are based on CNN-F features (code length = 64).



(a) **MIRFlickr25K** (b) **NUS-WIDE10.5K**

Figure 4: Precision@top-K curves on two datasets. The baselines are based on CNN-F features (code length = 64).



(a) **MiRFlickr25K** (b) **NUS-WIDE10.5K**

Figure 5: Misjudging proportion of false negative (FN) and false positive (FP)(code length = 64).

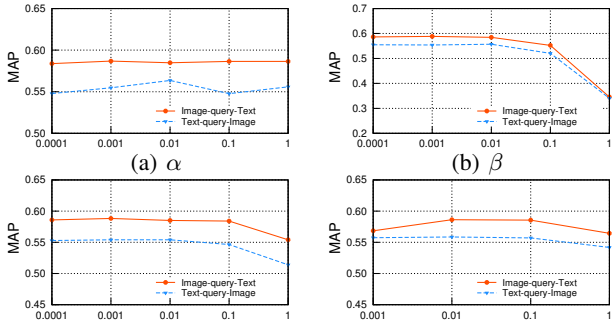
Evaluation metrics and compared methods We adopt two query styles, including *Image query Text* (abbr. IQT) and *Text query Image* (abbr. TQI), where we use images (and texts) in each query set as queries and retrieve texts (and images) from retrieval database.

We adopt two widely used standard evaluation metrics, Mean average precision (MAP) and precision-recall curve, for effeciveness evaluation (Liu et al. 2014). Following AGAH (Gu et al. 2019), we also evaluate the retrieval based on top-K precision curves (Wei et al. 2018), which is a popular metric used in information retrieval.

We compare our Bi-CMR with four state-of-the-art deep hashing cross-modal retrieval methods, including DCMH (Jiang and Li 2017), PRDH (Yang et al. 2017), SSAH (Li et al. 2018), and AGAH (Gu et al. 2019). The source codes of baselines are coded according to their descriptions. Parameters are carefully tuned accordingly. Remind that the existing relevance measurement based on the similarity matrix S conflicts with human perception, we adopt the manually marked relevance matrix G as evaluation metric. Table 3 shows the comparison of different measurements in terms of MAP. We can see that MAPs based on the wrong evaluation are much larger than the real values.

Table 3: MAP on MIRFlickr25K (code length = 64).

Methods / Matrices	DCMH		PRDH		SSAH		AGAH	
	IQT	TQI	IQT	TQI	IQT	TQI	IQT	TQI
$\{S\} : S_{ij}$	0.7501	0.7734	0.7221	0.7514	0.7932	0.8030	0.8075	0.8050
$\{G\} : G_{ij}$	0.4878	0.4835	0.4875	0.4792	0.4970	0.4740	0.4975	0.4808



(a) α (b) β (c) γ (d) δ

Figure 6: Parameter sensitivity analysis on MIRFlickr25K.

Experimental Results and Analysis

Comparison of Effectiveness We compare four existing state-of-the-art cross-modal methods with our proposed Bi-CMR by conducting the cross-modal retrieval over two benchmark datasets in terms of MAP, precision-recall curve and top-K precision curves.

We first report the MAP results in Table 2 under the new similarity evaluation with threshold λ of each dataset. Clearly, our Bi-CMR consistently achieves the best MAP, demonstrating its superiority against all the counter parts. For MIRFLickr25K, Bi-CMR improves the competitors, IQT and TQI, by 16.69% and 12.74% respectively. For NUS-WIDE10.5K, Bi-CMR improves IQT and TQI by 8.77% and 2.69% respectively. The accuracy improvement

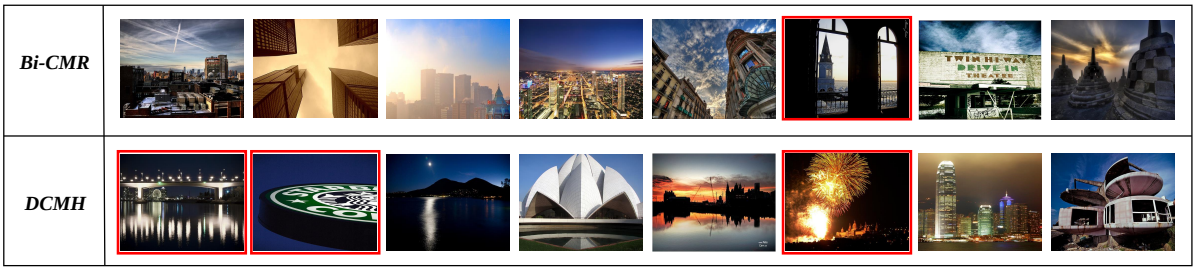


Figure 7: Comparison of ranking results for the text query "*The architecture under the sky with clouds.*".

over MIRFLickr25K is more significant compared with that over NUS-WIDE10.5K. This is because MIRFLickr25K has more multi-label instances, while NUS-WIDE10.5K contains a high proportion of single-label ones.

We also evaluate our Bi-CMR under precision-recall curves and the top-K precision curves. Figures 3 and 4 show the comparison of our Bi-CMR and existing competitors. As we can see that Bi-CMR defeats all the other methods on MIRFlickr25K and NUS-WIDE10.5K under the precision-recall curves and top-K precision curves. In Figures 3, the closer to the top right, the higher the accuracy is, which indicate that our Bi-CMR can reduce false negatives and thus improve recall. As shown above, our approach achieves the best performance in terms of different evaluation metrics. This has further confirmed that our Bi-CMR can retrieve semantically similar instances more accurately. Figure 4(b) shows an exception when performing IQT on NUS-WIDE10.5K dataset. When K is less than 50, the precision of Bi-CMR is lower than AGAH (Gu et al. 2019) and DCMH (Jiang and Li 2017). This is due to the addition of direct inter-modal operations, which restricts these methods to two modalities only. Please notice that the overall curve of Bi-CMR is greater than all the other methods. With the increase of K, it has the most stable precision.

Hyper-parameters Analysis To evaluate the four hyper-parameters in Eq. 7, we construct validation set by randomly choosing $2,000$ data from database. A sensitivity analysis of these hyper-parameters is provided in Figure 6, which indicates that the best choice of α is around $0.01 \sim 0.05$; β and γ is around 0.01, and their excessive disparity will lead to an imbalance in reinforcement; the optimal setting for δ is from 0.01 to 0.1. Since the parameter η mainly depends on the quality of dataset for each modality, and the image quality of both datasets is higher than the text quality in our datasets, we set this parameter to 0.9. To select the threshold λ for new similarity evaluation, the proportion of false negative (FN) and false positive (FP) for multi-labels at different thresholds on each dataset are calculated as shown in Figure 5, which indicates that the best choices for λ is around 0.4 on MIRFlickr25K and 0.7 on the NUS-WIDE10.5K.

Ablation Studies We conduct extensive ablation studies on two datasets. We define five alternatives, Bi-CMR1, Bi-CMR2, Bi-CMR3, Bi-CMR4, and Bi-CMR5, to study the impact of independently training strategy. Here, Bi-CMR1 does not consider forward learning. Bi-CMR2 trains multi-label network without considering $\mathcal{L}_{l2m}$ in backward learning. Bi-CMR3 trains multi-label network without considering $\mathcal{L}_{cor}$ in backward learning. Bi-CMR4 does not adjust S_{ij}. Bi-CMR5 does not consider $\mathcal{L}_{class}$. For a fair comparison, all these variants adopt the same network architecture, settings and the evaluation metric. Table 4 shows the performance comparison of our proposed full Bi-CMR with the first five different ablations for IQT and TQI on two datasets, where we adopt CNN-F network to extract features. We can see that our full Bi-CMR performs best compared with Bi-CMR1, Bi-CMR2, Bi-CMR3 and Bi-CMR4. Removing each component results in slight relative performance degeneration. It reflects the effectiveness of each component of Bi-CMR, and shows the mutual promotion of our method. It would be unable to train if the entire backward is removed, so we split it into Bi-CMR2 and Bi-CMR3. The results also validate that the improvement of Bi-CMR mainly benefits from the bidirectional reinforcement based on multi-label network, which results in higher MAP scores.

Table 4: MAP for ablation analysis (code length = 64).

Methods	MIRFlick25K		NUS-WIDE	
	IQT	TQI	IQT	TQI
Bi-CMR1	0.5810	0.5496	0.4895	0.4832
Bi-CMR2	0.5790	0.5499	0.4813	0.4752
Bi-CMR3	0.4589	0.4597	0.4221	0.4263
Bi-CMR4	0.5731	0.5366	0.4963	0.4931
Bi-CMR5	0.5782	0.5559	0.4892	0.4955
Bi-CMR	0.5862	0.5586	0.5002	0.5045

Case Study We provide certain intuitive retrieval results of Bi-CMR and DCMH on MIRFlickr25K. The top 8 image results retrieved from the whole database are listed in Figure 7, where the incorrect results are highlighted by red boxes. As can be seen, compared with DCMH, Bi-CMR returns less irrelevant images and generates a better result ranking with the irrelevant ones at bottom positions.

Conclusion

In this paper, we propose a novel cross-modal retrieval framework Bi-CMR, which exploits a bidirectional reinforcement to well capture the semantics of instances. First, we propose a new evaluation to guide the learning of instance hash codes to overcome the gap between label annotations and semantic understanding of instances. Then, we propose a novel bidirectional reinforcement guided method for enhancing the hash code learning in a mutual promotion way. We construct two datasets with explicit relevant ground truth based on two benchmark datasets. Extensive experiment results have proved the high effectiveness of Bi-CMR.

References

Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*. BMVA Press.

Ding, G.; Guo, Y.; Zhou, J.; and Gao, Y. 2016. Large-Scale Cross-Modality Search via Collective Matrix Factorization Hashing. *IEEE Trans. Image Process.*, 25(11): 5427–5440.

Gu, W.; Gu, X.; Gu, J.; Li, B.; Xiong, Z.; and Wang, W. 2019. Adversary Guided Asymmetric Hashing for Cross-Modal Retrieval. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR*, 159–167.

He, L.; Xu, X.; Lu, H.; Yang, Y.; Shen, F.; and Shen, H. T. 2017. Unsupervised cross-modal retrieval through adversarial learning. In *2017 IEEE International Conference on Multimedia and Expo, ICME*, 1153–1158.

Hu, M.; Yang, Y.; Shen, F.; Xie, N.; Hong, R.; and Shen, H. T. 2019. Collective Reconstructive Embeddings for Cross-Modal Hashing. *IEEE Trans. Image Process.*, 28(6): 2770–2784.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008*, 39–43.

Jiang, Q.; and Li, W. 2017. Deep Cross-Modal Hashing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 3270–3278.

Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; and Tao, D. 2018. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 4242–4251.

Li, C.; Deng, C.; Wang, L.; Xie, D.; and Liu, X. 2019. Coupled CycleGAN: Unsupervised Hashing Network for Cross-Modal Retrieval. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2019*, 176–183.

Lin, Z.; Ding, G.; Hu, M.; and Wang, J. 2015. Semantics-preserving hashing for cross-view retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, 3864–3872.

Liong, V. E.; Lu, J.; Tan, Y.; and Zhou, J. 2017. Cross-Modal Deep Variational Hashing. In *IEEE International Conference on Computer Vision, ICCV 2017*, 4097–4105.

Liu, W.; Mu, C.; Kumar, S.; and Chang, S. 2014. Discrete Graph Hashing. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, 3419–3427.

Song, D.; Liu, W.; Ji, R.; Meyer, D. A.; and Smith, J. R. 2015. Top Rank Supervised Binary Coding for Visual Search. In *2015 IEEE International Conference on Computer Vision, ICCV*, 1922–1930.

Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; and Shen, H. T. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013*, 785–796.

Wang, D.; Gao, X.; Wang, X.; and He, L. 2015. Semantic Topic Multimodal Hashing for Cross-Media Retrieval. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, 3890–3896.

Wang, X.; Li, Y.; Yang, H.; and Chen, J. 2014. An image retrieval scheme with relevance feedback using feature reconstruction and SVM reclassification. *Neurocomputing*, 127: 214–230.

Wei, T.; Guo, L.; Li, Y.; and Gao, W. 2018. Learning safe multi-label prediction for weakly labeled data. *Mach. Learn.*, 107(4): 703–725.

Wu, B.; Yang, Q.; Zheng, W.; Wang, Y.; and Wang, J. 2015. Quantized Correlation Hashing for Fast Cross-Modal Search. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, 3946–3952.

Xu, X.; Lu, H.; Song, J.; Yang, Y.; Shen, H. T.; and Li, X. 2020. Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval. *IEEE Trans. Cybern.*, 50(6): 2400–2413.

Xu, X.; Shen, F.; Yang, Y.; Shen, H. T.; and Li, X. 2017. Learning Discriminative Binary Codes for Large-scale Cross-modal Retrieval. *IEEE Trans. Image Process.*, 26(5): 2494–2507.

Yang, E.; Deng, C.; Liu, W.; Liu, X.; Tao, D.; and Gao, X. 2017. Pairwise Relationship Guided Deep Hashing for Cross-Modal Retrieval. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017*, 1618–1625.

Ye, Z.; and Peng, Y. 2018. Multi-Scale Correlation for Sequential Cross-modal Hashing Learning. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018*, 852–860.

Zhou, J.; Ding, G.; and Guo, Y. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, 2014*, 415–424.

Zhou, X.; Chen, L.; Zhang, Y.; Qin, D.; Cao, L.; Huang, G.; and Wang, C. 2017. Enhancing online video recommendation using social user interactions. *VLDB J.*, 26(5): 637–656.

Zhu, M.; Shen, D.; Xu, L.; and Wang, X. 2021. Scalable Multi-grained Cross-modal Similarity Query with Interpretability. *Data Sci. Eng.*, 6(3): 280–293.

Zhu, X.; Lafferty, J.; and Rosenfeld, R. 2005. *Semi-supervised learning with graphs*. Ph.D. thesis, Carnegie Mellon University, language technologies institute.

Acknowledgments

The work is partially supported by the National Key Research and Development Program of China (2020YFB1707901), National Natural Science Foundation of China (62072088, 61991404), Ten Thousand Talent Program (ZX20200035), Liaoning Distinguished Professor (XLYC1902057), and ARC Discovery Project (DP200101175).