Cross-Modal Generation and Pair Correlation Alignment Hashing

Weihua Ou, Jiaxin Deng, Lei Zhang, Jianping Gou, *Senior Member, IEEE*, and Quan Zhou, *Member, IEEE*

***Abstract*— Cross-modal hashing is an effective cross-modal retrieval approach because of its low storage and high efficiency. However, most existing methods mainly utilize pre-trained networks to extract modality-specific features, while ignore the position information and lack information interaction between different modalities. To address those problems, in this paper, we propose a novel approach, named cross-modal generation and pair correlation alignment hashing (CMGCAH), which introduces transformer to exploit position information and utilizes cross-modal generative adversarial networks (GAN) to boost cross-modal information interaction. Concretely, a cross-modal interaction network based on conditional generative adversarial network and pair correlation alignment networks are proposed to generate cross-modal common representations. On the other hand, a transformer-based feature extraction network (TFEN) is designed to exploit position information, which can be propagated to text modality and enforce the common representation to be semantically consistent. Experiments are performed on widely used datasets with text-image modalities, and results show that the proposed method achieved competitive performance compared with many existing methods.**

***Index Terms*— Cross-modal hashing, cross-modal generation, correlation alignment, position semantic information, cross-modal interaction.**

I. Introduction

In SOCIAL networks, various modality data, such as images, texts and videos are growing rapidly. To dig useful information from such massive data, cross-modal retrieval [1] has been proposed and has been used in real applications [2], [3]. The main challenge for cross-modal retrieval is how to explore and model the relations between different modalities due to the semantic gap and heterogeneity gap. Although many cross-modal methods [4], [5], [6], [7], [8] have been developed, the requirements of computation and storage are still the main obstacles with the ever-increasing trend of multi-modal data. Thus, cross-modal hashing (CMH) methods, which encode cross-modal data into binary hashing codes, have been extensively studied in the past years [9], [10], [11] because CMH can reduce the computational cost and storage dramatically.

Existing CMH methods mainly includes shallow methods [10], [12], [13], [14], [15] and deep-learning based methods [16], [17], [18], [19], [20], [21]. The shallow methods aim to transform different modalities data into hashing codes with semantic consistency, while they have no advantages to extract the high-level semantic features. With the successes of deep neural network (DNN), many deep cross-modal hashing methods [16], [17], [18], [19], [20], [21] have been proposed in recent years. Compared with shallow architectures, deep cross-modal hashing can learn high-level semantic representations and explore cross-modal correlations. Generally, deep cross-modal hashing methods [16], [17], [18], [19], [20], [21] project the raw features from different modalities into the common space separately and establish the cross-modal relations by pair-wise or label information, while lack cross-modal information interaction. On the other hand, most of them mainly utilize pre-trained network to extract modality-specific features while ignore the position information. In fact, as shown in Fig.1, embedding the position information in images and correlating them with the orientation words (e.g., left, right) in the texts is very important in modelling the relations between different modalities.

To overcome those problems, in this paper, we proposed cross-modal generation and pair correlation alignment hashing (CMGCAH) method, as shown in Fig.1, which includes three modules. The first module is the modality-specific feature extraction module, which is designed to extract modality-specific features for each modality. Different from existing methods, transformer-based feature extraction network (TFEN) is developed to encode the position information of images. The second module is the GAN-based interaction network (GIN) module, which is proposed to boost modality information interaction through adversarial learning. The last module is the feature correlation alignment module, which is proposed to promote the generated cross-modal representations to be semantically consistency. The main contributions of our approach are summarized as follows:

- A Transformer-based image feature extraction model is designed to exploit the position information. Compared with most of the existing methods, the position

Manuscript received 20 October 2021; revised 12 May 2022 and 5 September 2022; accepted 7 November 2022. Date of publication 17 November 2022; date of current version 1 March 2023. This work was supported by the National Natural Science Foundation of China under Grant 62262005, Grant 61976107, Grant 61876093, Grant 61962010, and Grant 61762021. The Associate Editor for this article was H. Lu. *(Corresponding authors: Weihua Ou; Quan Zhou.)*

Weihua Ou and Jiaxin Deng are with the School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550025, China (e-mail: ouweihuahust@gmail.com).

Lei Zhang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610056, China.

Jianping Gou is with the College of Computer and Information Science, College of Software, Southwest University, Chongqing 400715, China (e-mail: cherish.gjp@gmail.com).

Quan Zhou is with the National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: quan.zhou@njupt.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3221787

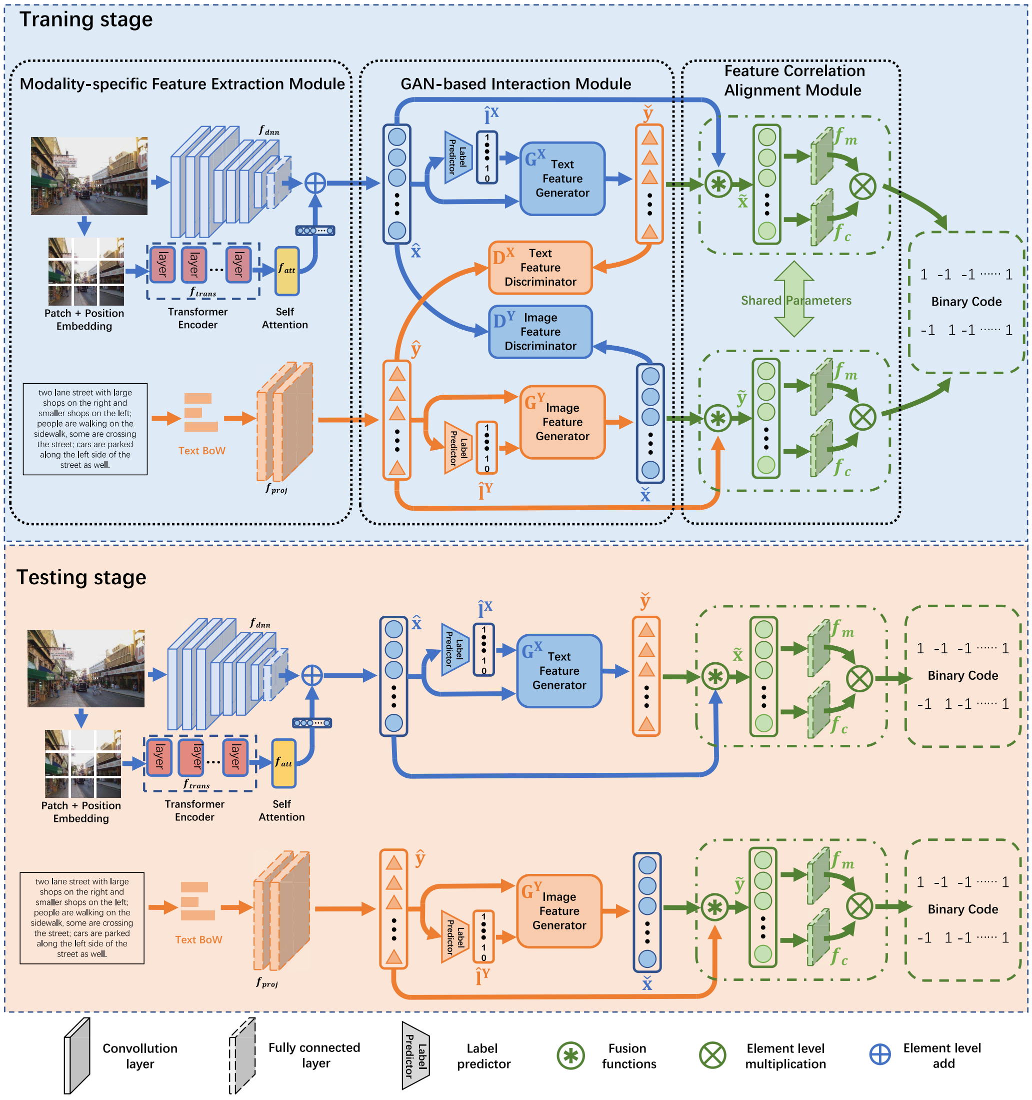

Traning stage
Modality-specific Feature Extraction Module
GAN-based Interaction Module
Feature Correlation Alignment Module
f_{dnn}
Patch + Position Embedding
layer
f_{trans}
Transformer Encoder
f_{att}
Self Attention
two lane street with large shops on the right and smaller shops on the left; people are walking on the sidewalk, some are crossing the street; cars are parked along the left side of the street as well.
Text BoW
f_{proj}
$\hat{\mathbf{x}}$
$\hat{\mathbf{l}}^X$
Label Predictor
G^X Text Feature Generator
$\check{\mathbf{y}}$
D^X Text Feature Discriminator
D^Y Image Feature Discriminator
$\hat{\mathbf{y}}$
G^Y Image Feature Generator
$\hat{\mathbf{l}}^Y$
$\check{\mathbf{x}}$
$\tilde{\mathbf{x}}$
$\tilde{\mathbf{y}}$
f_m
f_c
Shared Parameters
1 -1 -1 1
Binary Code
-1 1 -1 1
Testing stage
Convollution layer
Fully connected layer
Label predictor
Fusion functions
Element level multiplication
Element level add

Fig. 1. Flowchart of the proposed cross-modal generation and pair correlation alignment hashing, which includes three modules, i.e., modality-specific feature extraction module, GAN-based interaction module and feature correlation alignment module. In the feature extraction module, for image modality, a DNN architecture f_{dnn} is used to extract the raw image features, a transformer encoder f_{trans} is designed to embed the position information, followed by a self-guided attention f_{att}. The representation of images $\hat{\mathbf{x}}$ are the combination of raw image features and output of f_{att}. For text modality, a projector network f_{proj} is adopted to extract text feature $\hat{\mathbf{y}}$ based on the bag-of-words representations. GAN-based interaction module utilizes two GANs to promote the modality information interaction. For image modality, the text feature generator G^X conditioned on image features $\hat{\mathbf{x}}$ and predicted label $\hat{I}^X$ is proposed to generate text features $\check{\mathbf{y}}$, the discriminator D^X distinguishes the generated texts features $\check{\mathbf{y}}$ from the corresponding text features $\hat{\mathbf{y}}$. For text modality, the image feature generator G^Y conditioned on text features $\hat{\mathbf{y}}$ and predicted label $\hat{I}^Y$ is designed to generate image features $\check{\mathbf{x}}$, the discriminator D^Y distinguishes the generated image features $\check{\mathbf{x}}$ from the corresponding image features $\hat{\mathbf{x}}$. Feature correlation alignment module takes cross-modal fusion features $\tilde{\mathbf{x}}$ or $\tilde{\mathbf{y}}$ as input and learns a shared cross-modal fusion network. We input image text pairs into the network and train the whole network in the training stage.

information can characterize the fine correlations between different modalities.

- A cross-modal interaction network is developed to boost the modality interaction, which includes conditional generative adversarial networks with label predictor.
- A feature correlation alignment network is proposed to align the features from different modalities, which fuses multi-modal features and utilizes attention mechanism to explore salient correlation.

The rest of this paper is organized as follows. Firstly, we review the related works in Section II, and then present the details of the proposed method in Section III, followed by the algorithm description in Section IV. Finally, we give the experimental results in Section V and conclude this paper in Section VI.

II. Related Works

A. Shallow Cross-Modal Hashing

Many shallow cross-modal hashing methods have been proposed in the past years. This kind of methods mainly focused on two problems. The first problem is how to model the semantic similarity between different modalities, and the second one is how to preserve the semantic similarity in the hashing code space.

For instance, Zhang et al. proposed a large-scale supervised multi-modal hashing with semantic correlation maximization

(SCM) [9], which used labels to construct semantic similarity between different modalities and maximized the correlation in the hashing code space. Ding et al. proposed collective matrix factorization hashing (CMFH) [10] by exploring cross-modal common semantics through matrix factorization. Lin et al. proposed semantics-preserving hashing (SePH) [13], which transformed semantic affinities into a probability distribution and minimized the KL divergence. This method is effective to model the semantic affinities and is robust compared to directly construct the similarity matrix via label. Following that, Wang et al. developed semantic topic multi-modal hashing (STMH) [14], which characterized the semantic similarity using latent semantic topics.

Different from above methods, label consistent matrix factorization hashing (LCMFH) [15] directly used semantic labels to guide the hashing learning and set data with the same label to be the same representation in the semantic space. Considering the real application, Yao et al. proposed online latent semantic hashing (OLSH) [22], which adopted online learning scheme to learn a continuous latent semantic concept space. To narrow the semantic gap, relation learning for cross-modal hashing (SRLCH) [23] mapped class labels of two modalities into subspace, which is used as a bridge for cross-modal data to learn the unified binary codes.

Although much effort has been made, label or pair-wise information only can characterize the manifest relations and many latent relations only can be explored through cross-modal information interaction.

B. Deep Cross-Modal Hashing

Inspired by the success of deep neural networks in computer vision, deep cross-modal hashing methods [16], [17], [18], [19], [20], [21], [24] have been proposed. For example, Jiang and Li [17] integrated the feature learning and hashing codes learning into the same framework with deep neural networks. To exploit inter-modal correlations, Yang et al. proposed pairwise relationship-guided deep hashing (PRDH) [18] by pairwise constraints. Considering the complementarity of different modalities, Lu et al. [25] designed self-weighted fusion strategy to fuse the complementary information into hash codes. Further more, Zhu et el. proposed deep collaborative multi-view hashing (DCMVH) [24], which consists of multiple view-specific networks and a fusion network. Those work explored the common semantic information by projecting their features into the common space individually, and lack information interaction among modalities. Recently, many adversarial cross-modal hashing [5], [26], [27], [28], [29], [30] are proposed. For example, deep adversarial metric learning (DAML) [5] introduces adversarial learning to narrow the gap between modalities. Qiang et al. proposed deep semantic similarity adversarial hashing [28] (DSSAH) for cross-modal retrieval, which introduces an adversarial modality discriminator to establish a common feature space. Liu et al. proposed adversarial tri-fusion hashing network [26] by maximally bridging the semantic gap of the common representations between balanced and imbalanced data. All those methods shown that the information interaction is important for the cross-modal correlations. Different from existing work, we designed transformer-based network to exploit the position information and cross-modal generation interaction network to boost the information interaction between modalities.

III. Proposed Method

A. Problem Statement

Given the training dataset as $\mathcal{D} = (\mathbf{X}, \mathbf{Y}, \mathbf{L})$, where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ are the images, $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$ are the associated pair-wise texts, $\mathbf{L} = \{\mathbf{l}_i\}_{i=1}^n$ is the label matrix, and $\mathbf{l}_i = [l_i^1, l_i^2, \ldots, l_i^d]$ is the associated label vector, where $l_i^j = 1$ ($l_i^j = 0$) represents $(\mathbf{x}_i, \mathbf{y}_i)$ belongs (not) to the j-th class, d is class number.

To perform cross-modal hashing retrieval, we first learn the real representation and then obtain the hashing codes. Specifically, we learn two forward networks, $\mathbf{u}_i = f_x(\mathbf{x}_i; \theta^X)$, $\mathbf{v}_i = f_y(\mathbf{y}_i; \theta^Y)$, $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n]$ and $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ are the real representations of images and texts, θ^Y, θ^Y are the network parameters. Then, we obtain the associated hashing codes of images and texts using an element-wise sign function $\mathbf{b}_i^X = sign(\mathbf{u}_i)$ and $\mathbf{b}_i^Y = sign(\mathbf{v}_i)$, respectively.

B. Modality-Specific Feature Extraction

As shown in Fig.1, we extract the image features and text features, respectively. For the image modality, we adopt pre-trained CNN-F as original network and replace the FC8 layer with a fully connected layer (f_{dnn}) to obtain the features. To encode the position information, an additional transformer encoder (f_{trans}) [31] is adopted to transfer the embedded patches into a set of sequential features, which are fused by a self-guided attention network [32] (f_{att}). Finally, the image features $\hat{\mathbf{x}}_i$ are obtained by combining the raw image features with the output of f_{att}. Those operations can be formulated as below:

$$\hat{\mathbf{x}}_i = f_{dnn}(\mathbf{x}_i) + \lambda f_{att}(f_{trans}(\mathbf{x}_i)), \tag{1}$$

where $\mathbf{x}_i$ is the input image and λ is a hyper-parameter.

For the text modality, we extract the bag-of-words (BoW) representation as the raw text features $\mathbf{y}_i$. Then, the two fully connected layers (f_{proj}) is adopted to project the raw text features $\mathbf{y}_i$ to text features $\hat{\mathbf{y}}_i$. It can be formulated as below:

$$\hat{\mathbf{y}}_i = f_{proj}(\mathbf{y}_i). \tag{2}$$

C. GAN-Based Interaction Network

To boost the information interaction, we utilize GANs to generate cross-modal features from one modality to the other modality. Concretely, two predict networks are design to predict the labels of the given images and texts respectively, then two conditional generative adversarial networks [33] are proposed to promote the information interaction.

For image modality, given the image features $\hat{\mathbf{x}}_i$, we first learn the predict network f_p^X to predict the label of images. This can be formulated as following problem:

$$L_p^X = -\sum_{i=1}^{n} \mathbf{l}_i \cdot \log\left(f_p^X\left(\hat{\mathbf{x}}_i\right)\right) \tag{3}$$

where $\hat{\mathbf{x}}_i$ is the features of the i-th image, n is the samples number.

The text modality is similar to image modality, and the objective function of text label predictor f_p^Y can be formulated as:

$$L_p^Y = -\sum_{i=1}^{n} \mathbf{l}_i \cdot \log\left(f_p^Y\left(\hat{\mathbf{y}}_i\right)\right) \tag{4}$$

where $\hat{\mathbf{y}}_i$ is the feature of the i-th text.

Based on the predicted labels and the image features, we design a conditional generative adversarial network to promote the information interaction. The text feature generator G^X conditioned on image features $\hat{\mathbf{x}}$ and predicted label $\hat{I}^X$ is proposed to generate text features $\check{\mathbf{y}}$, while the discriminator D^X distinguishes the generated texts features $\check{\mathbf{y}}$ from the corresponding text features $\hat{\mathbf{y}}$. The whole objective function L_{adv}^X can be formulated as follows:

$$\begin{aligned} L_{adv}^X = E_{\hat{\mathbf{y}} \sim P_{\hat{\mathbf{y}}}}&\left[\log\ D^X\left(\hat{\mathbf{y}}|\hat{\mathbf{l}}^X\right)\right] \\ &+ E_{\check{\mathbf{y}} \sim P_{\check{\mathbf{y}}}}\left[\log\left(1 - D^X\left(\check{\mathbf{y}}|\hat{\mathbf{l}}^X\right)\right)\right]. \end{aligned} \tag{5}$$

Similarly, for text modality, the image feature generator G^Y conditioned on text features $\hat{\mathbf{y}}$ and predicted label $\hat{I}^Y$ is designed to generate image features $\check{\mathbf{x}}$, on the other hand the discriminator D^Y distinguishes the generated image features $\check{\mathbf{x}}$ from the corresponding image features $\hat{\mathbf{x}}$. The objective function L_{adv}^Y of conditional generative adversarial networks can be formulated as below:

$$\begin{aligned} L_{adv}^Y = E_{\hat{\mathbf{x}} \sim P_{\hat{\mathbf{x}}}}&\left[\log\ D^Y\left(\hat{\mathbf{x}}|\hat{\mathbf{l}}^Y\right)\right] \\ &+ E_{\check{\mathbf{x}} \sim P_{\check{\mathbf{x}}}}\left[\log\left(1 - D^Y\left(\check{\mathbf{x}}|\hat{\mathbf{l}}^Y\right)\right)\right]. \end{aligned} \tag{6}$$

After iterative training, GAN-based interaction network can produce the reconstructed image features and text features. Then, by combining the reconstructed features with the input features, we can obtain the image fusion features and text fusion features by the following formula:

$$\begin{aligned} \widetilde{\mathbf{x}}_i &= f_s(\hat{\mathbf{x}}_i, \check{\mathbf{y}}_i) \\ \widetilde{\mathbf{y}}_i &= f_s(\check{\mathbf{x}}_i, \hat{\mathbf{y}}_i) \end{aligned} \tag{7}$$

where f_s is the fusion function, which is achieved by concatenating text feature vectors and image feature vectors.

D. Feature Correlation Alignment Module

To explore the pair correlation between different modalities, we propose a shared feature correlation alignment network, which fuses features from different modalities and uses attention mechanism to explore semantic correlation. Firstly, the attention module f_m is proposed to get the salient part of the cross-modal features and suppress the unrelated features. The attention operation is defined as follows:

$$m_i = sigmoid\left(f_m\left(\widetilde{\mathbf{x}}_i\right)\right), \tag{8}$$

where f_m is the fully connected layer, m_i is the mask. After that, the cross-modal related features are obtained as follows:

$$\mathbf{u}_i = f_c\left(\widetilde{\mathbf{x}}_i\right) \otimes m_i, \tag{9}$$

where f_c are fully connected layer.

Similarly, for text modality, attention operation and related features are defined as below:

$$\begin{aligned} m_i &= sigmoid\left(f_m\left(\widetilde{\mathbf{y}}_i\right)\right), \\ \mathbf{v}_i &= f_c\left(\widetilde{\mathbf{y}}_i\right) \otimes m_i. \end{aligned} \tag{10}$$

With the above definitions, we explore the most significantly related features from different modalities, and the correlation between the two modalities is exploited implicitly. Therefore, the generated features will help to obtain semantic discriminative and modality-invariant hashing codes.

E. Hash Code Learning

The learned hashing codes should preserve the semantic similarity between instances from different modalities and also should be discriminative semantically. For image modality, we utilize the following loss function, which is motivated by [17]:

$$\begin{aligned} L_s^X = &-\alpha \sum_{i,j=1}^{n}\left(S_{ij}^{XY}\Theta_{ij} - \log\left(1 + e^{\Theta_{ij}}\right)\right) \\ &-\beta \sum_{i,j=1}^{n}\left(S_{ij}^{XX}\Psi_{ij} - \log\left(1 + e^{\Psi_{ij}}\right)\right) \\ &+\gamma\, ||B - U||_F^2, \end{aligned} \tag{11}$$

where $\Theta_{ij} = \frac{1}{2}\mathbf{u}_i^T\mathbf{v}_j$, $\Psi_{ij} = \frac{1}{2}\mathbf{u}_i^T\mathbf{u}_j$. For text modality, the objective function can be similarly defined as follows:

$$\begin{aligned} L_s^Y = &-\alpha \sum_{i,j=1}^{n}\left(S_{ij}^{YX}\Phi_{ij} - \log\left(1 + e^{\Phi_{ij}}\right)\right) \\ &-\beta \sum_{i,j=1}^{n}\left(S_{ij}^{YY}\Gamma_{ij} - \log\left(1 + e^{\Gamma_{ij}}\right)\right) \\ &+\gamma\, ||B - V||_F^2, \end{aligned} \tag{12}$$

where $B = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n]$ are the hashing codes obtained by $\mathbf{b}_i = sign(\mathbf{u}_i + \mathbf{v}_i)$ in the training phase. $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n]$ and $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ are real representations, $\Phi_{ij} = \frac{1}{2}\mathbf{v}_i^T\mathbf{u}_j$, $\Gamma_{ij} = \frac{1}{2}\mathbf{v}_i^T\mathbf{v}_j$. Please note that we only use this to compute hashing codes in the training stage. In the test stage, we get the hashing codes of images and texts by $\mathbf{b}_i^X = sign(\mathbf{u}_i)$ and $\mathbf{b}_i^Y = sign(\mathbf{v}_i)$, respectively.

The first term in (11) is the negative log likelihood of the cross-modal similarities, which is defined as follows:

$$P(S_{ij}^{XY}|\mathbf{u}_i, \mathbf{v}_j) = \begin{cases} \sigma\left(\Theta_{ij}\right) & \text{when} \quad S_{ij}^{XY} = 1 \\ 1 - \sigma\left(\Theta_{ij}\right) & \text{when} \quad S_{ij}^{XY} = 0 \end{cases} \tag{13}$$

where $\sigma\left(\Theta_{ij}\right) = \frac{1}{1+e^{-\Theta_{ij}}}$. $S_{ij}^{XY} = 1$ if image $\mathbf{x}_i$ and text $\mathbf{y}_j$ share the same class label, and otherwise, $S_{ij}^{XY} = 0$, and $S^{XY} = S^{YX}$. Similarly, $S_{ij}^{XX} = 1$ if image $\mathbf{x}_i$ and image $\mathbf{x}_j$ share the same class label, and otherwise, $S_{ij}^{XX} = 0$. The representation $\mathbf{u}_i, \mathbf{v}_j$ are encouraged to be similar when $S_{ij}^{XY} = 1$, and to be different when $S_{ij}^{XY} = 0$.

IV. Algorithm

To sum up, the whole objective function can be rewritten as below:

$$L = (L_{adv}^X + L_s^X + \delta L_p^X) + (L_{adv}^Y + L_s^Y + \delta L_p^Y) \tag{14}$$

where δ is a parameter. We adopt alternative optimization scheme to learn the image network, text network and hashing codes B. Firstly, to optimize the parameters of image feature extraction network θ_F^X, image label predictor θ_P^X, text feature generator θ_G^X, and feature correlation alignment network θ_C, we fix the text network and hashing codes B and optimize following problem:

$$arg \min_{\theta_F^X,\theta_P^X,\theta_G^X,\theta_C} L_{adv}^X + L_s^X + \delta L_p^X. \tag{15}$$

Then, for optimizing the parameters of discriminator model D^X, we fix the text network and image network and optimize following problem:

$$arg\max_{\theta_D^X} L_{adv}^X \tag{16}$$

For the text network, to optimize the parameters of text feature extraction network θ_F^Y, text label predictor θ_P^Y, image feature generator θ_G^Y, and feature correlation alignment network θ_C, we fix image network and B:

$$arg \min_{\theta_F^Y,\theta_P^Y,\theta_G^Y,\theta_C} L_{adv}^Y + L_s^Y + \delta L_p^Y. \tag{17}$$

Then, we optimize the parameters of discriminator D^Y by solving following optimization problem:

$$arg\max_{\theta_D^Y} L_{adv}^Y. \tag{18}$$

Finally, we fix image and text network, and optimize B. The hashing codes can be obtained by $B = sign(U + V)$. The details of the whole training procedure are shown in algorithm 1.

V. Experiments

We carry out experiments on MIRFLICKR-25K, NUS-WIDE and MS-COCO datasets to evaluate the proposed method. We first describe the experiments setting in Section V-A, and then show the experimental results and analysis in Section V-B.

A. Experiments Setting

1) Dataset: The MIRFLICKR-25K dataset [34] contains 25000 instances, each of them includes a image and related textual tags, labeled with a 24-dimensional category vector. Each instance from text modality is represented as a 1386-dimensional BoW vector, and 20015 instances are selected for the experiment following the protocol [35]. Among them, the test set contains 2000 samples, and other samples are regraded as retrieval set. We randomly select 10000 samples from the retrieval set as training set.

The NUS-WIDE dataset [36] consists of 269648 web images. We follow the experimental protocol given in SSAH [35], and use 190421 samples from the 21 most-frequent concepts to conduct the experiment.

The MS-COCO dataset [37] contains 82783 training images and 40504 validation images. We used 87081 images of them labeled with a 91-dimensional category vector. Every image has its corresponding text description, and each of them is represented as a 2000-dimensional BoW vector.

Algorithm 1 Pseudocode of the Proposed Method

```
Require: The training dataset ${\mathbf{x}_i, \mathbf{y}_i, \mathbf{l}_i}_{i=1}^n$, and similarity matrix $S^{XX}$, $S^{XY}$, $S^{YY}$, parameters $\lambda$, $\alpha$, $\beta$, $\gamma$, $\delta$, the learning rate $\eta$, and mini-batch size $k$.

Repeat:
for $iter = 1, 2, \cdots, Iter_x$ do
    Sample $k$ points from training dataset.
    Calculate $\mathbf{u}_i = f_x(\mathbf{x}_i; \theta_F^X, \theta_P^X, \theta_G^X, \theta_C)$ by forward propagation for each sampled point $\mathbf{x}_i$.
    Update parameters of image network $\theta_F^X, \theta_P^X, \theta_G^X$ and feature correlation alignment network $\theta_C$ by equation (15).
end for
Update parameters of discriminator $D^X$ by equation (16).
for $iter = 1, 2, \cdots, Iter_y$ do
    Sample $k$ points from training dataset.
    Calculate $\mathbf{v}_i = f_y(\mathbf{y}_i; \theta_F^Y, \theta_P^Y, \theta_G^Y, \theta_C)$ by forward propagation for each sampled point $\mathbf{y}_i$.
    Update parameters of text network $\theta_F^Y, \theta_P^Y, \theta_G^Y$ and feature correlation alignment network $\theta_C$ by equation (17).
end for
Update parameters of discriminator model $D^Y$ by equation (18).
Update $B$ by $B = sign(U + V)$.
Until: Required iteration number.
```

2) Evaluation: We perform two cross-modal retrieval tasks: retrieving text using image as query (I→T) and retrieving image using text as query (T→I). Mean Average Precision (mAP), defined as below, is used to evaluate the performance.

$$mAP = \frac{1}{m}\sum_{i=1}^{m} AP(q_i),$$

where m is the number of query samples, AP($\cdot$) denotes the average precision, and q_i is the i-th query sample.

3) Implement Details: For the image modality, we adopt ViT and a self-guided attention module to extract the position information features. Firstly, we split the image into 16 image patches and fed them into the ViT network, then a self-guided attention module is used to merge the output of the ViT network, and the output feature dimension is 2048. We then fuse raw image feature and position feature by element-level summation with $\lambda = 0.01$.

For the text modality, the raw text feature dimension of MIRFLICKR-25K, NUS-WIDE and MS-COCO dataset are 1386, 2912 and 2000, respectively. We adopt two fully connected layers to project them to text features. The hidden layer has 4096 nodes, and the output feature dimension is 2048. We set minibatch size $k = 128$ and optimize with the RMSprop optimizer. We start the training with the learning rate $\lambda = 10^{-4.5}$, $\alpha = 0.01$, $\beta = 0.0001$, $\gamma = 0.01$, $\delta = 0.1$.

4) Compared With State-of-the-Arts: We compare the proposed method with several representative methods, including traditional method CCA [38], shallow cross-modal hashing, such as, SCM [9], STMH [14], CMFH [10], SePH [13], NSDH [13], JIMFH [39], and FCMH [40], and deep

TABLE I
mAP RESULTS ON THE MIRFLICKR-25K DATASET

Tasks	Methods	16 bits	32 bits	64 bits
I→T	CCA [38]	0.5571	0.5534	0.5503
	SCM [9]	0.6905	0.6989	0.7049
	STMH [14]	0.6077	0.6215	0.6391
	CMFH [10]	0.6472	0.6579	0.6572
	SePH [13]	0.7141	0.7211	0.7142
	JIMFH [39]	0.6575	0.6738	0.6881
	FCMH [40]	0.7120	0.7322	0.7374
	DCMH [17]	0.7415	0.7491	0.7597
	SSAH [35]	0.7861	0.7938	0.7997
	CPAH [30]	0.7581	0.7619	0.7751
	SAAH [41]	**0.7920**	0.7960	**0.8150**
	OURS	0.7901	**0.8030**	0.8123
T→I	CCA [38]	0.5601	0.5569	0.5533
	SCM [9]	0.7127	0.7207	0.7267
	STMH [14]	0.6041	0.6105	0.6206
	CMFH [10]	0.6252	0.6347	0.6362
	SePH [13]	0.7195	0.7256	0.7225
	JIMFH [39]	0.6937	0.6913	0.7201
	FCMH [40]	0.6883	0.6933	0.7027
	DCMH [17]	0.7668	0.7723	0.7816
	SSAH [35]	0.7802	0.7893	0.7912
	CPAH [30]	0.7734	0.7879	0.7951
	SAAH [41]	**0.7950**	**0.8030**	**0.8060**
	OURS	0.7823	0.7932	0.8045

TABLE II
mAP RESULTS ON THE NUS-WIDE DATASET

Tasks	Methods	16 bits	32 bits	64 bits
I→T	CCA [38]	0.3238	0.3140	0.3064
	SCM [9]	0.4964	0.5216	0.5360
	STMH [14]	0.3993	0.4288	0.4591
	CMFH [10]	0.5362	0.5344	0.5417
	SePH [13]	0.6172	0.6237	0.6183
	JIMFH [39]	0.5021	0.5323	0.5612
	FCMH [40]	0.6013	0.6219	0.6362
	DCMH [17]	0.5820	0.5981	0.6059
	SSAH [35]	0.6079	0.6176	0.6273
	CPAH [30]	0.6157	0.6219	0.6358
	SAAH [41]	**0.6280**	**0.6460**	**0.6560**
	OURS	0.6213	0.6440	0.6462
T→I	CCA [38]	0.3003	0.2986	0.2947
	SCM [9]	0.5170	0.5136	0.5182
	STMH [14]	0.4345	0.4563	0.4589
	CMFH [10]	0.4017	0.4086	0.4254
	SePH [13]	0.5867	0.6123	0.5943
	JIMFH [39]	0.5030	0.5895	0.5809
	FCMH [40]	0.5759	0.5934	0.6057
	DCMH [17]	0.5994	0.5950	0.6011
	SSAH [35]	0.6217	0.6389	0.6687
	CPAH [30]	0.6005	0.6324	0.6370
	SAAH [41]	0.6510	0.6630	0.6590
	OURS	**0.6782**	**0.6801**	**0.6844**

cross-modal hashing, such as, DCMH [17], SSAH [35], CPAH [30], and SAAH [41]. We utilize pre-trained CNN-F network to obtain image features and utilize bag-of-words features as text features for all shallow cross-modal hashing. For the SAAH, we directly cite the reported results from the original paper.

B. Experimental Results

1) Hamming Ranking: Table I reports the mAP results on the MIRFLICKR-25K. As can be seen, compared with the shallow methods, such as CMFH, SePH, JIMFH, FCMH, our method improved average about 9% increase on mAP for two retrieval tasks. Compared with DNN-based methods, such as DCMH, SSAH, CPAH and SAAH, our method also achieved the best performance, with an average increase about 2%. Compared with SAAH, our method obtained competitive results. This is because our method can effectively boost the information interaction and align the correlation between different modalities.

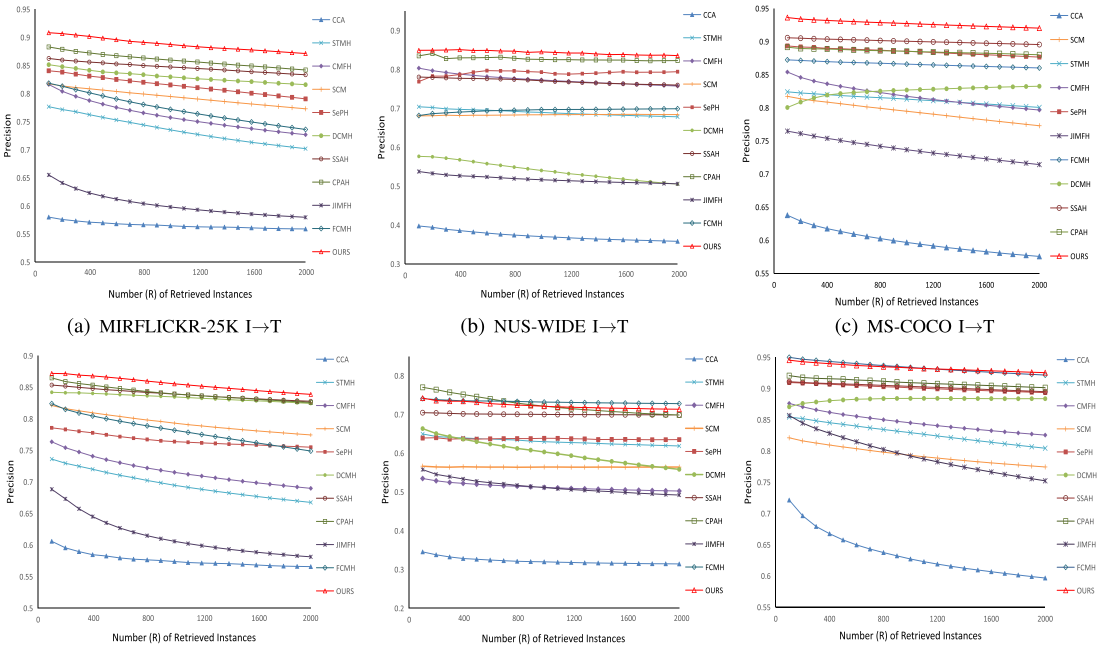
Table II reports the mAP results on the NUS-WIDE dataset, which is large scale and more challenge. From the table, we can see that our method has achieved better performance than that of the other methods. For the task T→I, our method obtained the best performances with an increase about 2% compared to the second method SAAH. This verified that the proposed cross-modal features generation and correlation alignment can effectively promote information interaction between different modalities. Table III reports the mAP results on the MS-COCO dataset. We can see that our method has also achieved better performance than other methods.

TABLE III
mAP RESULTS ON THE MS-COCO DATASET

Tasks	Methods	16 bits	32 bits	64 bits
I→T	CCA [38]	0.5302	0.5152	0.5037
	SCM [9]	0.6991	0.7101	0.7161
	STMH [14]	0.6221	0.6504	0.6533
	CMFH [10]	0.6206	0.6342	0.6351
	SePH [13]	0.7565	0.7799	0.7738
	JIMFH [39]	0.5931	0.5505	0.5603
	FCMH [40]	0.7179	0.7418	0.7596
	DCMH [17]	0.7391	0.7573	0.7662
	SSAH [35]	0.7738	0.7864	0.7937
	CPAH [30]	0.7611	0.7645	0.7701
	SAAH [41]	0.5730	0.5760	0.5710
	OURS	**0.7854**	**0.8026**	**0.8159**
T→I	CCA [38]	0.5344	0.5189	0.5072
	SCM [9]	0.7024	0.7143	0.7248
	STMH [14]	0.6098	0.6097	0.6173
	CMFH [10]	0.6314	0.6523	0.6569
	SePH [13]	0.7716	0.7943	0.7785
	JIMFH [39]	0.5836	0.5634	0.5751
	FCMH [40]	0.7276	0.7516	0.7754
	DCMH [17]	0.7694	0.7862	0.7946
	SSAH [35]	0.7762	0.7853	0.7832
	CPAH [30]	0.7703	0.7731	0.7843
	SAAH [41]	0.5580	0.5510	0.5370
	OURS	**0.7927**	**0.8134**	**0.8188**

Fig.2 shows the precision@top-R curves with length of 64 bits. We set R from 100 to 2000 and calculate the precision with R increases by 100 every time. From the sub-figures, it can be seen that CMGCAH achieved the highest precision than DCMH, SSAH and CPAH with different R values on MIRFLICKR-25K datasets. Especially, when the R is small, the precision scores of CMGCAH is much better than other methods. For NUS-WIDE dataset and MS-COCO dataset, our proposed CMGCAH achieved better results on I→T task. Although the precision of CMGCAH is lower than that of CPAH at the beginning on the task T→I, the precision of CMGCAH decreases slowly with the increase of R value.

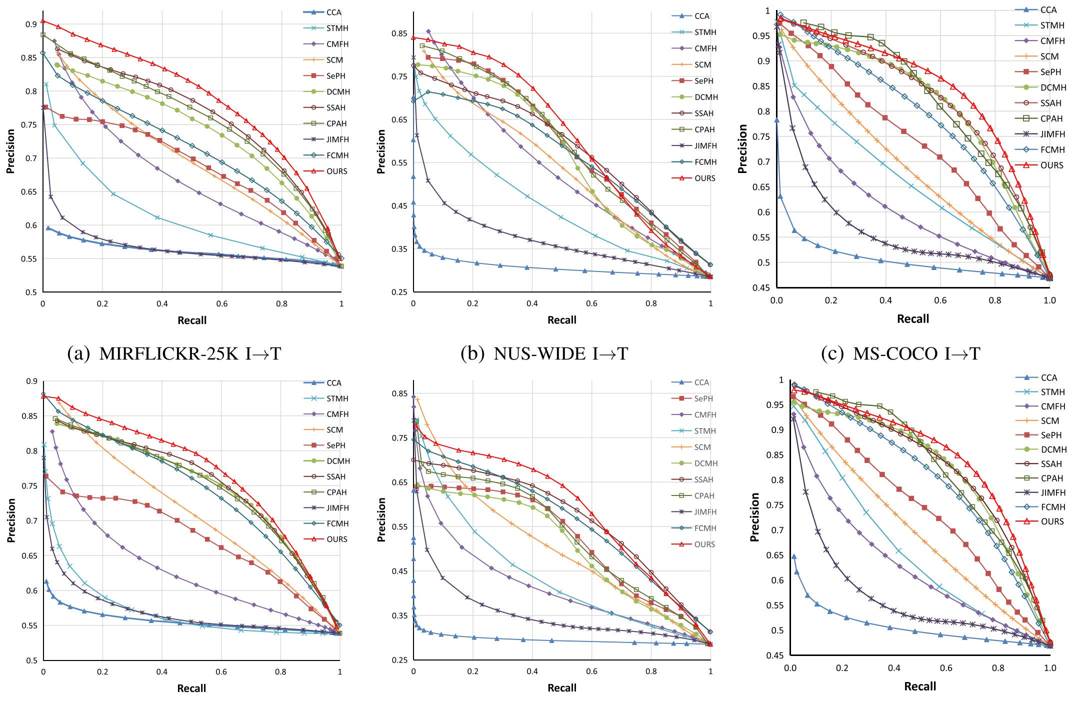
2) Hash Lookup: Fig.3 shows the precision-recall curve with 64 bits on three datasets. As can be seen, CMGCAH is superior to the compared methods. With the increase of recall,



(a) MIRFLICKR-25K I→T (b) NUS-WIDE I→T (c) MS-COCO I→T

(d) MIRFLICKR-25K T→I (e) NUS-WIDE T→I (f) MS-COCO T→I

Fig. 2. Precision@top-R curves on three datasets.



(a) MIRFLICKR-25K I→T (b) NUS-WIDE I→T (c) MS-COCO I→T

(d) MIRFLICKR-25K T→I (e) NUS-WIDE T→I (f) MS-COCO T→I

Fig. 3. Precision-Recall curves on three datasets.

the precision value of CMGCAH decreases more slowly. This demonstrates our approach can effectively boost the information interaction between modalities, and the generated hashing codes has good robustness.

3) Ablation Study: To verify the effectiveness of each module, we study four variants of the proposed model: (a) CMGCAH-a represents feature generation and correlation alignment modules are removed; (b) CMGCAH-b denotes feature attention mechanism is replaced with full connection network; (c) CMGCAH-c denotes feature generation module is removed and directly take features as the input of attention mechanism. (d) CMGCAH-d means the transformer-based feature extraction network is removed. The results are reported in Table IV. From that, we can see that the lowest mAP

TABLE IV
mAP OF CMGCAH, CMGCAH-A, CMGCAH-B, CMGCAH-C AND CMGCAH-D ON MIRFLICKR-25K DATASET WITH CODE LENGTH 64 BITS

Methods	I→T	T→I
CMGCAH-a	0.7893	0.7821
CMGCAH-b	0.7999	0.7893
CMGCAH-c	0.7982	0.7926
CMGCAH-d	0.8110	0.8002
CMGCAH	0.8123	0.8045

result is CMGCAH-a, the structure of which is similar to DCMH, only label information is utilized and lacks information interaction between different modalities. For CMGCAH-b and CMGCAH-c, their mAP results are better than that of CMGCAH-a. Because CMGCAH-b promotes information interaction and ignores the similarity between cross-modal features, while CMGCAH-c improves the similarity utilizing alignment module furtherly. For CMGCAH-d, the transformer-based feature extraction network can extract richer image features, which is beneficial to cross-modal feature generation.

VI. CONCLUSION

This paper studies the cross-modal hashing retrieval focusing on exploring position information from image and boosting the interaction between two modalities. Transformer-based image feature extraction network is designed to exploit position information from image and the GAN-based cross-modal interaction networks are proposed to promote cross-modal information interaction. The comprehensive experiments demonstrate the effectiveness of the proposed method.

REFERENCES

[1] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," 2016, *arXiv:1607.06215*.

[2] Y. Gu and Y. Jie, "Densely-connected multi-magnification hashing for histopathological image retrieval," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1683–1691, Jul. 2019.

[3] L. Zhang, F. Liu, and Z. Zeng, "Combining link and content correlation learning for cross-modal retrieval in social multimedia," in *Human Centered Computing*, Q. Zu and B. Hu, Eds. Berlin, Germany: Springer, 2018, pp. 516–526.

[4] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10394–10403.

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

[6] X. Song, J. Chen, Z. Wu, and Y.-G. Jiang, "Spatial-temporal graphs for cross-modal Text2Video retrieval," *IEEE Trans. Multimedia*, vol. 24, pp. 2914–2923, 2022, doi: 10.1109/TMM.2021.3090595.

[7] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 24, pp. 466–479, 2022.

[8] L. Zhang, L. Chen, C. Zhou, F. Yang, and X. Li, "Exploring graph-structured semantics for cross-modal retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2021, pp. 4277–4286.

[9] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2177–2183.

[10] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2075–2082.

[11] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, May 2018.

[12] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1360–1365.

[13] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3864–3872.

[14] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3890–3896.

[15] D. Wang, X.-B. Gao, X. Wang, and L. He, "Label consistent matrix factorization hashing for large-scale cross-modal similarity search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2466–2479, Oct. 2018.

[16] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1445–1454.

[17] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3232–3240.

[18] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1618–1625.

[19] Y. Shen, L. Liu, L. Shao, and J. Song, "Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4097–4106.

[20] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[21] C. Yan, X. Bai, S. Wang, J. Zhou, and E. R. Hancock, "Cross-modal hashing with semantic deep embedding," *Neurocomputing*, vol. 337, pp. 58–66, Apr. 2019.

[22] T. Yao et al., "Online latent semantic hashing for cross-media retrieval," *Pattern Recognit.*, vol. 89, pp. 1–11, May 2019.

[23] H. T. Shen et al., "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 10, pp. 3351–3365, Oct. 2021.

[24] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, "Deep collaborative multi-view hashing for large-scale image search," *IEEE Trans. Image Process.*, vol. 29, pp. 4643–4655, 2020.

[25] X. Lu, L. Zhu, Z. Cheng, L. Nie, and H. Zhang, "Online multi-modal hashing with dynamic query-adaption," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 715–724.

[26] X. Liu, Y.-M. Cheung, Z. Hu, Y. He, and B. Zhong, "Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 4, pp. 607–619, Aug. 2021.

[27] X. Shen, H. Zhang, L. Li, Z. Zhang, D. Chen, and L. Liu, "Clustering-driven deep adversarial hashing for scalable unsupervised cross-modal retrieval," *Neurocomputing*, vol. 459, pp. 152–164, Oct. 2021.

[28] H. Qiang, Y. Wan, L. Xiang, and X. Meng, "Deep semantic similarity adversarial hashing for cross-modal retrieval," *Neurocomputing*, vol. 400, pp. 24–33, Aug. 2020.

[29] X. Ma, T. Zhang, and C. Xu, "Multi-level correlation adversarial hashing for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3101–3114, Dec. 2020.

[30] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, San Diego, CA, USA, Jun. 2016, pp. 1480–1489.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[34] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr. (MIR)*, 2008, pp. 39–43.

[35] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4242–4251.

[36] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, pp. 1–9.

[37] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 740–755.

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

[39] D. Wang, Q. Wang, L. He, X. Gao, and Y. Tian, "Joint and individual matrix factorization hashing for large-scale cross-modal retrieval," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107479.

[40] Y. Wang, Z.-D. Chen, X. Luo, R. Li, and X.-S. Xu, "Fast cross-modal hashing with global and local similarity embedding," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10064–10077, Oct. 2022, doi: 10.1109/TCYB.2021.3059886.

[41] M. Li, Q. Li, Y. Ma, and D. Yang, "Semantic-guided autoencoder adversarial hashing for large-scale cross-modal retrieval," *Complex Intell. Syst.*, vol. 8, no. 2, pp. 1603–1617, Apr. 2022.

Weihua Ou received the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014. He was Post-Doctoral Research Fellow at the University of Technology Sydney. He is currently a Full Professor with the School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China. So far, he has published over 70 papers in international journals or conferences, such as in IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *PR*. His current research interests include pattern classification and machine learning. He was a PC Member of IJCAI, AAAI, PRCV, and ISAIR; and a Senior Member of CCF. He is an Associate Editor of Cognitive Robotics.

Jiaxin Deng received the master's degree in computer science from Guizhou Normal University, Guiyang, China, in 2021. He is currently pursuing the Ph.D. degree with the Beijing University of Technology, Beijing, China. His research interests include multimedia information retrieval and continual learning.

Lei Zhang is currently pursuing the Ph.D. degree with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include cross-modal retrieval, referring image segmentation, and named entity recognition. He is a Reviewer in ACL Rolling Review.

Jianping Gou (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2012. He was a Post-Doctoral Research Fellow at the University of Sydney. So far, he has published over 100 papers in international journals or conferences, such as in *IJCV*, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *TIST*, IEEE TRANSACTIONS ON CYBERNETICS, and *TKDD*. His current research interests include pattern classification and machine learning. He is a Senior Member of CCF and CSIG. He is an Academic Editor of Scientific Programming and an Editorial Board Member of mathematics.

Quan Zhou (Member, IEEE) received the B.S. degree in electronics and information engineering from the China University of Geosciences, Hubei, China, in 2002, and the M.S. and Ph.D. degrees in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2006 and 2013, respectively. He is currently an Associate Professor with the National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. He has published more than 70 academic articles, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *PR*. His research interests include deep learning, pattern recognition, and computer vision. He is a member of IAPR and a TPC Member of ISAIR and WCSP. He was a reviewer for more than 70 SCI journals. He was the Area Chair of IEEE ICME2019 and PRCV2022. He was a leading Guest Editor of IEEE TRANSACTIONS ON MULTIMEDIA, *PR*, *Computers and Electrical Engineering*, and *Multimedia Tools and Applications*.